A.A. Hendriksen

Betting as an alternative to p-values

Master thesis

Thesis advisor: Prof. Dr. Peter Grünwald

Date master exam: TBD

Mathematisch Instituut, Universiteit Leiden

Abstract

In this thesis, we present the theory of test martingales. Whereas in traditional statistics, the p-value indicates the level of evidence against the null hypothesis, in martingale testing a betting strategy that allows one to make a (virtual) profit is seen as evidence against the null hypothesis.

We extend the concept of test martingales of Shafer et al. (2011) to composite null hypotheses. We refer to these martingales as composite test martingales – they are the main innovation in this thesis. Using composite test martingales, we construct two martingale tests as an alternative for the Student's t-test. These two tests appear to be the first published martingale tests that can differentiate the t-test hypotheses.

The main result of this thesis concerns the Jeffreys Bayesian t-test. It was already known, experimentally, to be robust under optional stopping. Optional stopping refers to the practice of looking at observed experimental data in order to decide whether or not to continue testing. Robustness in this context means that the statistical method preserves its significance level even when optional stopping is employed. We prove the Jeffreys Bayesian t-test to be a martingale test. Therefore, it is robust under optional stopping. In general, under a composite null hypothesis, Bayesian tests are not robust under optional stopping.

Contents

1 Introduction

The "reproducibility crisis" in science has attracted a lot of attention in the past few years. For example, it was featured on the front page of the well-known magazine Economist (2013). An incredible number of scientific results turn out to be false. The fields of psychology and medicine, in particular, have come under scrutiny. Ioannidis (2005) raised awareness of the reproducibility of research in the field of medicine. In 2012, Amgen, an American drug company, attempted to replicate 53 studies in the field of cancer medicine but was able to confirm only 6 cases (Begley and Ellis, 2012). Supported by Kahneman (2014), attention for replication has increased in the field of psychology. In the ensuing replication attempt of 100 published experiments in the field of psychology, less than 50% of the experiments could be replicated (Collaboration et al., 2015).

This lack of reproducible research is caused to some extent by current scientific practices – negative results are often left unpublished, for example, which results in "publication bias". Another part of the problem, however, is caused by the fact that unsuitable statistical methods are used: classical statistical hypothesis tests are often based on p-values, but the p-value does not do what it is supposed to do. In this thesis, we explore a promising alternative statistical hypothesis test: the martingale test.

Whereas in traditional statistics, the p-value indicates the level of evidence against the null hypothesis, in martingale tests a betting strategy that allows one to make a (virtual) profit is seen as evidence against the null hypothesis. This idea has been introduced by Shafer et al. (2011), which builds on classical work by Ville (1939).

The martingale test confronts two important problems in traditional statistics: fragility under optional stopping and the complexity of combining evidence from multiple scientific studies.

Optional stopping refers to the practice of looking at observed experimental data in order to decide whether or not to continue testing. Traditional statistical methods do not allow a scientist to "peek" at the data. In order to derive statistically significant results, they employ a predetermined "stopping rule" and usually stop at a predetermined sample size N. In practice, this means that scientists are tempted to study a few extra data points when the venerable statistical significance of a p-value with $p < 0.05$ has not yet, but almost, been achieved at sample size N – thereby violating the stopping rule. More than 55% of psychologists admitted to using this practice in a survey conducted by John et al. (2012).

Preregistration of scientific experiments can remedy this problem, but is not yet widespread. Moreover, anybody who *has* achieved a statistically significant result without preregistration would be hard pressed to prove that no peeking has occurred.

Martingale tests, on the other hand, are robust under optional stopping: the rate at which true hypotheses are falsely rejected does not excessively increase when a study is stopped upon reaching a positive result. At any point in a scientific investigation, the results may be evaluated and, if deemed statistically significant, the study concluded.

Another strong suit of martingale tests is combining evidence from mul-

tiple studies. "Optional continuation" is the practice of combining evidence of studies that were done because of promising results of previous research on the same subject. In traditional statistics meta-analyses are performed, but combining the evidence (p-values) is usually fraught with difficulty.

In this thesis we introduce the theory of test martingales and construct two martingale tests as an alternative for the Student's t-test. The one-sample t-test is a location test for normally distributed data with unknown scale. It tests whether the mean (or median) of the data is at some predetermined midpoint. Student's t-test is widely used in practice. Because the t-test is scale-invariant, it has a so called *composite* null hypothesis. This prohibits the usual construction of a martingale test. This thesis contains the first published martingale tests that can differentiate the t-test hypotheses.

The second martingale test draws a connection to Bayesian hypothesis testing. On the one hand, it confirms (for this specific case) the widely held conviction in the Bayesian community that Bayesian hypothesis tests are robust under optional stopping, see Rouder (2014). In general, under a composite null hypothesis, these tests are *not* robust under optional stopping – see for example Figure 4 in van der Pas and Grünwald (2014). On the other hand, it lays bare an internal inconsistency in the way standards of evidence are being set for Bayesian hypothesis tests. This is discussed in the conclusion.

Earlier work on sequential analysis was done by Wald (1973) during and after the second world war to boost American industrial production. It has one major drawback: sequential analysis is not possible on hypotheses that are "nested" (overlapping). Since most hypothesis tests in current use employ overlapping hypotheses, the use of sequential analysis would require the introduction of an artificial partition of the null and alternative hypothesis. The artificial separation of hypotheses invites a degree of subjectivity that is insignificant in matters of war, but undesirable in matters of science. Exploration of Wald's sequential testing is therefore outside the scope of this thesis.

We shall first lay the groundwork needed for our results. All mathematical prerequisites are contained in Section 2, as well as a brief overview of classical hypothesis testing. Section 2.2 introduces the statistical hypotheses studied in this thesis and distinguishes between *simple* and *composite* hypotheses – an essential distinction in this thesis.

Section 3 contains a description of test martingales, the mathematical construct for conducting tests for *simple* null hypotheses. This idea is extended to tests with a compound null hypothesis in Section 4. Composite test martingales are the main innovation in this thesis.

Sections 5 and 6 describe the construction of two martingale tests that can be used as alternatives to the t-test. The power of these tests is compared to the traditional Student's t-test in Section 7. We finish with a conclusion and discussion.

2 Preliminaries

This section introduces the mathematical language necessary to prove the main assertions of this thesis. The mathematical prerequisites are not strictly necessary to enjoy Section 3 about test martingales, although the prerequisite knowledge will allow the reader to appreciate the main results by understanding rather than accepting proof by intimidation. We introduce the test hypotheses which we want to be able to distinguish in Section 2.2. This section also introduces the distinction between *simple* and *composite* hypotheses, which is crucially important in this thesis – be sure not to ignore it. Finally, we give a brief overview of classical hypothesis testing.

2.1 Mathematical prerequisites

With exception of the notation, all definitions and results in this section can be found in Durrett (2010).

A collection Σ of subsets of S is called a *σ-algebra* on S if

- $S \in \Sigma$;
- $E \in \Sigma \implies S \setminus E \in \Sigma$;
- $E_1, E_2, \ldots \in \Sigma \implies \bigcup_{n=1}^{\infty} E_n \in \Sigma$.

A *probability space* is a triplet $(\Omega, \mathcal{F}, \mu)$ where Ω is a set, $\mathcal{F}$ is a σ-algebra on Ω, and μ is a probability measure on $\mathcal{F}$. A *random variable* $X : \Omega \to \mathbb{R}$ on $(\Omega, \mathcal{F}, \mu)$ is an $\mathcal{F}$-measurable function. A *random process* is a sequence of random variables $(X_t)_I$ indexed by an index set I. In this thesis, $I = \{1, 2, \ldots\}$ will always denote discrete time.

Let $X = (X_n)_{n=1}^{\infty}$ be a random process *adapted* to the filtration $\mathcal{F} = (\mathcal{F}_n)_{n=1}^{\infty}$ ($\mathcal{F}_n \subseteq F_{n+1}$ for all $n \geq 1$), i.e., X_k is $\mathcal{F}_k$-measurable for all k. Then we call X a *martingale* when X_1 is integrable and for all $n \geq 1$

$$E[X_{n+1} \mid \mathcal{F}_n] = X_n$$

holds. For a *submartingale* or *supermartingale* the equality is replaced by a greater or equal or less than or equal relation respectively. Notably, a submartingale grows and a supermartingale shrinks (in expectation).

Remark 2.1. *In this thesis, Bernoulli-distributed random variables have state space $\{-1, 1\}$ instead of the usual $\{0, 1\}$. For our purposes, 1 and -1 are easier to calculate with. The notation $Y \sim Ber(\theta)$ indicates that Y is Bernoulli-distributed with parameter θ:*

$$P(Y = \quad 1) = \theta$$
$$P(Y = -1) = 1 - \theta.$$

Example 2.2. *Let $X = (X_n)_{n=1}^{\infty}$, $X_n \in \{-1, 1\}$, be independent and identically distributed random variables. Let these random variables be Bernoulli one half distributed, $X_n \sim Ber(\frac{1}{2})$. Then the random process defined by*

$$Y_0 = 0,$$
$$Y_n = \sum_{i=1}^{n} X_n$$

is a martingale. *Note that* $E[Y_0] = 0$ *and the expectation of* X_n *equals zero as well, independent of any previous outcomes. Hence, we have*

$$\begin{aligned} E[Y_{n+1} \mid \sigma(Y_0, Y_1, \ldots, Y_n)] &= Y_n + E[X_{n+1} \mid \sigma(Y_0, Y_1, \ldots, Y_n)] \\ &= Y_n + E[X_{n+1}] \\ &= Y_n. \end{aligned}$$

A *stopping time* τ with respect to a filtration $(\mathcal{F}_n)$ is a random variable taking values in $\{0, 1, 2, \ldots\}$ such that at any time n, $\{\omega \mid \tau(\omega) \leq n\} \in \mathcal{F}_n$.

Let μ, ν be two measures on a measurable space $(X, \mathcal{F})$. If for every $E \in F$ such that $\mu(E) = 0$ we have that $\nu(E) = 0$, we call ν *absolutely continuous* with respect to μ (notation $\nu \ll \mu$).

The measures ν and μ are called *mutually singular*, $\nu \perp \mu$, if there exist disjoint sets E, F such that $\nu(A) = \nu(A \cap E)$ and $\mu(A) = \mu(A \cap E)$ for all sets $A \in \mathcal{F}$. In other words, the measures support different subsets of the whole space.

If the measure ν is absolutely continuous with respect to μ ($\nu \ll \mu$), then there exists a unique $h \in \mathcal{L}^1(X, \mathcal{F}, \mu)$ such that for all $E \in \mathcal{F}$

$$\nu(E) = \int_E h \, d\mu. \tag{2.1}$$

The measurable function h is unique in the sense any other function h' satisfying Equation (2.1) is such that $\mu(\{h \neq h'\}) = 0$. This result is known as the *Radon-Nikodym theorem* and h is referred to as the *Radon-Nikodym derivative* of ν with respect to μ and is written as

$$h = \frac{d\nu}{d\mu}. \tag{2.2}$$

A sequence of random variables $Z_1, \ldots, Z_n$ will be written in long form as much as possible. Sometimes, we will use the notation $(Z_i)_{i=1}^n$ and sometimes we will refer to the sequence as Z^n. This is especially useful when taking averages:

$$\overline{Z^n} = \frac{1}{n} \sum_{i=1}^{n} Z_i. \tag{2.3}$$

2.2 Test hypotheses

In this thesis, it is our primary goal to develop tests which distinguish the hypotheses

H_0**:** The random variables $X_1, X_2, \ldots$ are independent and identically $N(0, \sigma^2)$-distributed with unknown $\sigma > 0$;

H_1**:** The outcomes $X_1, X_2, \ldots$ with $X_i \sim N(\mu, \sigma^2)$ for all $i \in \{1, 2, \ldots\}$ are independent and identically distributed with unknown μ *and* $\sigma > 0$.

Furthermore, we also discuss tests on binary data. These tests distinguish between

H_0': The outcomes $X_1, X_2, \ldots$ are independent and identically Bernoulli-distributed with known parameter $\theta = \frac{1}{2}$, and

H_1': The outcomes $X_1, X_2, \ldots$ with $X_i \sim \text{Ber}(\theta)$ for all $i \in \{1, 2, \ldots\}$ are independent and identically distributed with unknown θ.

We refer to H_0 (and H_0') as the *null hypothesis* and to H_1 (and H_1') as the *alternative hypothesis*. The semantic distinction between the null and alternative hypothesis is explored in Section 2.3. Both hypotheses H_0 and H_1 have an unknown scale parameter σ. In the literature these kind of hypotheses are known as *compound* or *composite* hypotheses, as they cannot be represented by just one probability distribution. In fact, the hypotheses H_0 and H_1 can be symbolically represented by sets

$$H_0 = \{P_{0,\sigma^2} \mid \sigma \in \mathbb{R}, \sigma > 0\},$$
$$H_1 = \{P_{\mu,\sigma^2} \mid \mu, \sigma \in \mathbb{R}, \sigma > 0\},$$

where P_{μ,σ^2} denotes the distribution on an independent and identically distributed sequence of normal random variables with mean μ and variance σ^2. Hence, $P \in H_0$ denotes that P is a distribution on normal random variables with mean 0 and variance σ^2. We refer to hypotheses which contain just one probability measure (such as H_0') as *simple* or *point* hypotheses.

2.3 Classical hypothesis testing

Classical hypothesis testing is also known as the Neyman-Pearson testing paradigm (Rice, 2006). Neyman and Pearson viewed a statistical test as a statistical decision problem between two hypotheses in the sense of Popper (1959). One hypothesis, the null hypothesis, can only be *falsified* or *rejected* in the Neyman-Pearson terminology. The other hypothesis, the alternative hypothesis, serves as just that – an alternative. It is the null hypothesis that is rejected or accepted. In a sense, the theory does not aim to achieve scientific truth, but merely to disprove scientific untruth. Unless otherwise indicated, all definitions in this section can be found in Rice (2006).

We can give a simple definition, which we extend later, of a *statistical test*

$$\delta : \Omega \to \{\text{Accept}, \text{Reject}\} \tag{2.4}$$

as a statistic of some data sample $X_1, \ldots, X_n$ contained in a *sample space* Ω. The test can either accept or reject the null hypothesis. Because the null hypothesis is bestowed the benefit of the doubt, no decision named "Unknown" exists and indecision is mapped to "Accept". The statistical test only rejects the null hypothesis when there is sufficient evidence to do so.

In this paradigm, how can we quantify the quality of a statistical test? Rejecting the null hypothesis when it is true is called a *type I error*. The probability of committing a type I error is known as the *false positive rate* or *significance level* and is usually denoted by α. A rejection by a statistical test is more convincing when it has a low false positive rate. Accepting the null hypothesis when it is false and the alternative hypothesis is true is called a *type II error*. The probability of committing a type II error is called the *false negative rate* and

is denoted by β. The probability that the null hypothesis is rejected when the alternative hypothesis is true is called the *power* of the test and equals $1 - \beta$.

In classical hypothesis testing, the *p-value*

$$p : \Omega \to [0, 1] \tag{2.5}$$

is a statistic of the data which indicates the evidence against the null hypothesis. One may regard the p-value as the smallest significance level at which the null hypothesis would be rejected or, equivalently, as the probability that the result of the hypothesis test is as or more extreme than that actually observed if the null hypothesis were true. Hence, a small p-value indicates more evidence against the null hypothesis and a large p-value indicates less evidence against the null hypothesis. Mathematically, for any significance level $\alpha \in [0, 1]$ and distribution P_0 representing the null hypothesis, a p-value p must at least satisfy (Grünwald, 2017)

$$P_0(p \leq \alpha) \leq \alpha \tag{2.6}$$

and usually satisfies

$$P_0(p \leq \alpha) = \alpha. \tag{2.7}$$

When the null and alternative hypothesis are *simple*, i.e., the null and alternative hypotheses can be represented by a probability measures P_0 and P_1 respectively, then we may define the *likelihood ratio*. Suppose f_0 is the probability density function of P_0 and f_1 is the probability density function of P_1. Then

$$\frac{f_1(X_1, \ldots, X_n)}{f_0(X_1, \ldots, X_n)} \tag{2.8}$$

represents the likelihood ratio for a sample $X_1, \ldots, X_n$. It quantifies the relative odds of the data being produced by the alternative measure compared to the data being produced by the null measure without any reference to a prior belief in the two hypotheses. When the data is discrete instead of continuous, the likelihood ratio can be constructed using the ratio of the probability mass functions.

Hypothesis tests are conducted with a significance level α, which usually equals 0.05 but may be less depending on the scientific field of endeavor. A test may be concluded (the null hypothesis rejected) when the probability of the result under the null hypothesis is less than the required significance level, but how is the significance level determined when the number of outcomes is not known a priori? This is the tail wagging the dog. A *stopping rule* is used to remedy this situation: instead of stopping when the probability of the result surpasses the significance level, the test is concluded when the stopping rule says so. Rejection is decided based on the result of the stopped test. The significance level of a result can be calculated using the stopping rule. Most non-trivial stopping rules require intricate calculation to determine the significance level of results (see for example Armitage et al. (2002) pages 615–623). Hence, classical hypothesis tests are usually stopped after a fixed number of outcomes.

We may now define a p-value based statistical hypothesis test

$$\delta_{\alpha,\tau} : \Omega \to \{\text{Accept, Reject}\}$$
$$\delta_{\alpha,\tau} : (X_1, X_2, \ldots) \mapsto \begin{cases} \text{Reject} & p_\tau(X_1, \ldots, X_\tau) \leq \alpha, \\ \text{Accept} & \text{otherwise} \end{cases}$$

with respect to a stopping rule τ and a significance level α, where Ω is some state space, $(X_i)_{i=1}^{\infty}$ is a random process defined on that state space, and p_τ is a p-value with respect to the stopping rule. Note that the p-value *does depend* on the stopping rule! In practice people report a p-value for a sample as if the sample size had been fixed in advance (John et al., 2012), even if, in reality, the stopping rule was different. So the "p-value" they report is not a valid p-value and the significance level of their hypothesis test is thus wrong. This is the reason that robustness to optional stopping is so important.

The martingale tests developed in this thesis may also be classified as classical since they have a significance level, power, and distinguish between a null and alternative hypothesis. On the other hand, they do not suffer from requiring a fixed stopping rule. Hence we do not refer to our martingale tests as classical.

3 Test martingales: a review

This section introduces the notion of a test martingale. A significant part of this material is from Shafer et al. (2011). We introduce the concept of a test martingale in Section 3.1 with an extensive example. Section 3.2 describes how to construct a test martingale and Section 3.3 explains how a test martingale can be used as a statistical hypothesis test. We show an example of the construction of a test martingale when the alternative hypothesis is composite in Section 3.4. We conclude with a brief comparison of martingale testing to Bayesian hypothesis testing in Section 3.5.

3.1 The test martingale

The notion of a test martingale was introduced by Ville (1939) in 1939 as a tool for testing statistical hypotheses. The work of Shafer et al. (2011) relates test martingales to Bayes factors and p-values. The following is primarily based on Shafer et al. (2011) to whom I owe this beautifully succinct explanation:

> A test martingale is the capital process for a betting strategy that starts with 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.

Mathematically, the definition is even shorter.

Definition 3.1. *A* test martingale $M = (M_n)_{n=0}^{\infty}$ *is a non-negative martingale with initial value* $M_0 = 1$.

Let us consider an example of a test martingale, where in the alternative hypothesis the value of θ is *fixed* instead of an *unknown*:

H_0**:** The outcomes $Z_1, Z_2, \ldots$ are independent and identically Bernoulli-distributed with $\theta = \frac{1}{2}$.

H_1**:** The outcomes $Z_1, Z_2, \ldots$ are independent and identically Bernoulli-distributed with *fixed* $\theta > \frac{1}{2}$.

To disprove H_0, we set up a game where a player, let's call her Alice, starts with capital $Y_0 = 1$. Before each round i, she divides her capital in two by betting a fraction R_i of her money on the outcome $\{Z_i = 1\}$ and a fraction $1 - R_i$ of her money on the outcome $\{Z_i = -1\}$.

The *pay-off* in this game is calibrated to the probabilities of the outcomes. When $Z_i = 1$, Alice's investment in $\{Z_i = 1\}$ doubles and she loses her investment in $\{Z_i = -1\}$. When $Z_i = -1$ she loses the former investment and doubles the latter. Her capital is thus given by the stochastic process

$$Y_n = Y_{n-1}(1 - Z_n + 2R_nZ_n) = \begin{cases} 2R_nY_{n-1} & \text{if } Z_n = 1, \\ 2(1 - R_n)Y_{n-1} & \text{if } Z_n = -1. \end{cases} \tag{3.1}$$

Note that R_n must be $\sigma(Z_1, Z_2, \ldots, Z_{n-1})$-measurable, in other words: R_n must be an "investment" strategy that is possible to execute without looking

into the future. We can easily see that Y_n is a martingale under hypothesis H_0, as

$$\begin{aligned} E[Y_n \mid Y_{n-1}] &= E[Y_{n-1}(1 - Z_n + 2R_nZ_n) \mid Y_{n-1}] \\ &= Y_{n-1}E[1 - Z_n + 2R_nZ_n] \\ &= Y_{n-1}\left[\frac{1}{2}(1 - 1 + 2R_n) - \frac{1}{2}(1 + 1 - 2R_n)\right] \\ &= Y_{n-1} \end{aligned}$$

holds. Alice's initial capital equals one and she never bets more than she has. Her net worth can thus never become negative. Hence, Y is a test martingale for *any* admissible investment strategy $(R_n)_{n=1}^{\infty}$. This is the canonical example of a test martingale.

Although we have just shown that the game is fair under hypothesis H_0, we will show that money is still to be made when one believes that hypothesis H_1 is actually true. Now Alice should decide which strategy (R_n) yields the best return on investment under hypothesis H_1. She should optimize her expected interest rate

$$E\left[\log \frac{Y_n}{Y_0}\right] = E\left[\log Y_n\right].$$

When we look at the expected logarithmic increase per round

$$\begin{aligned} E_\theta[\log Y_n \mid \log Y_{n-1}] &= \log Y_{n-1} + E[\log(1 - Z_n + 2R_nZ_n)] \\ &= \log Y_{n-1} + [\theta \log(2R_n) + (1-\theta)\log(2(1-R_n))] \end{aligned}$$

we notice that the expression on the right hand side is maximized for $R_n = \theta$ since it uniquely solves

$$\begin{aligned} 0 &= \frac{d(\theta \log(2R_n) + (1-\theta)\log(2(1-R_n)))}{dR_n} \\ &= \frac{\theta}{R_n} - \frac{1-\theta}{1-R_n} \\ &= \frac{\theta}{\theta} - \frac{1-\theta}{1-\theta}. \end{aligned}$$

The optimization could also have been solved using Jensen's inequality, but setting the derivative equal to zero suffices here. (Note that the second derivative is negative at $R_n = \theta$). When Alice follows the optimal strategy of fixing R_n to θ, she expects to increase her capital by a factor of

$$\begin{aligned} E_\theta[1 - Z_n + 2R_nZ_n] &= \theta\, 2R_n + 2(1-\theta)(1-R_n) \\ &= 2\theta^2 + 2(1-\theta)^2 \\ &= 4\theta^2 - 4\theta + 2 \\ &= 2 - 4(\theta(1-\theta)), \end{aligned}$$

which always exceeds one unless θ equals one half. Setting $R_n = \theta$ yields the random process

$$Y_n = Y_{n-1}(1 - Z_n + 2\theta Z_n) = Y_{n-1}(1 + Z_n(2\theta - 1)), \tag{3.2}$$

which is a martingale under hypothesis H_0 and is a submartingale under hypothesis H_1. Equivalently, $(Y_i)_{i=1}^{\infty}$ is a fair game when Z_i is evenly distributed, and the most profitable game for Alice when Z_i is positive with probability exactly equal to θ.

This example illustrates the relation between the hypotheses, the pay-off, the outcome process Z, and the strategy R. Its constituent elements can be generalized to any test martingale with a simple null hypothesis $H_0 = \{P_0\}$.

In general, the null hypothesis informs the pay-off

$$W_n = \frac{1}{P_0(Z_n \mid Z_{n-1}, \ldots, Z_1)}, \tag{3.3}$$

where the probability measure P_0 distributes Z according to H_0. In the example, we had $W_n = 2$ for all n since Alice's investments always had a fifty-fifty chance of succeeding. Note that W_n is not necessarily constant in time.

When the outcomes of the process Z are distributed according to the null hypothesis, then by setting the pay-off to (3.3), any betting strategy results in a test martingale (of which no profit is expected). On the other hand, when the outcomes of the process Z are distributed according to the alternative hypothesis, profit generating strategies *do* exist: the resulting process Y may be a submartingale. This holds in general.

In the example, when the alternative hypothesis was true, a betting strategy defined by the alternative hypothesis performed better than any other strategy. This suggests that this strategy is optimal in general, but we do not prove this general case.

We shall see that there are more direct methods of constructing a test martingale.

3.2 The likelihood ratio is a martingale

We have just seen a construction of a test martingale. In this section we show how to construct a test martingale using the likelihood ratio for general simple hypotheses H_0 and H_1. Theorem 3.2, which we will state but not prove, allows one to turn a likelihood ratio into a test martingale. This construction is used in Shafer et al. (2011) and van der Pas and Grünwald (2014), for example.

The following observation on the relation between the likelihood ratio and the Radon-Nikodym derivative is necessary to appreciate Theorem 3.2. Let $X = (X_n)_{n=1}^{\infty}$ be a random process adapted to the filtration $\mathcal{F} = (\mathcal{F}_n)_{n=1}^{\infty}$. Suppose that the null and alternative hypothesis are simple – they can be represented by probability measures P_0 and P_1 on $(X, \mathcal{F})$, respectively. Suppose furthermore that the measures have probability density (or mass) functions f_0 and f_1 respectively and that P_1 is absolutely continuous with respect to P_0 on $\mathcal{F}_n$. Observe that the likelihood ratio on the first n outcomes as defined in Equation (2.8) is a version of the Radon-Nikodym derivative. We have

$$\frac{d\, P_1 \,|_{\mathcal{F}_n}}{d\, P_0 \,|_{\mathcal{F}_n}}(X_1, \ldots, X_n) = \frac{f_1(X_1, \ldots, X_n)}{f_0(X_1, \ldots, X_n)}$$

P_0-almost everywhere since the likelihood ratio satisfies the Radon-Nikodym

property (see Equation (2.1))

$$\int_E \frac{f_1(X_1, \ldots, X_n)}{f_0(X_1, \ldots, X_n)}\, dP_0 = P_1(E)$$

for all $E \in \mathcal{F}_n$.

The following theorem, which can be found in (Sun, 2016), shows that the Radon-Nikodym derivative is a martingale.

Theorem 3.2 (Sun (2016))**.** *Let $X = (X_n)_{n=1}^{\infty}$ be a random process adapted to the filtration $\mathcal{F} = (\mathcal{F}_n)_{n=1}^{\infty}$. Suppose $\mathbb{P}$ and $\mathbb{Q}$ are probability measures on $(X, \mathcal{F})$. Denote with $\mathbb{Q}_n := \mathbb{Q}\,|_{\mathcal{F}_n}$ the restriction of $\mathbb{Q}$ to $\mathcal{F}_n$. Assume $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}$ for all n, i.e., $Q_n \ll \mathbb{P}_n$. Then*

1. *The Radon-Nikodym derivative*

$$M_n := \frac{d\mathbb{Q}_n}{d\mathbb{P}_n} \tag{3.4}$$

   is a martingale on $(X, \mathcal{F}, \mathbb{P})$;

2. *If $\mathbb{Q} \perp \mathbb{P}$, then $M_n \to 0$ a.s. with respect to $\mathbb{P}$ and $M_n \to \infty$ a.s. with respect to $\mathbb{Q}$;*

3. *If $M_n \to 0$ a.s. with respect to $\mathbb{P}$, then $\mathbb{Q} \perp \mathbb{P}$.*

Theorem 3.2 shows that the likelihood ratio

$$Y'_n = Y'_{n-1} \cdot \frac{P_1(Z_n)}{P_0(Z_n)} = \frac{P_1(Z_1, \ldots, Z_n)}{P_0(Z_1, \ldots, Z_n)}$$

with $Y_0 = 1$ is a martingale under P_0, where $P_0(Z_t > 0) = 1/2$ and $P_1(Z_t > 0) = \theta$ with $t \in \mathbb{N}$. Absolutely continuity is guaranteed since P_0 assigns positive probability to all events in $\mathcal{F}_n$ for all $n \geq 1$.

How does this relate to the example of Alice? Are Y_n, as defined in Equation (3.2), and Y'_n equivalent processes? When $Z_n = 1$ we have

$$\begin{aligned} Y'_n &= Y'_{n-1} \cdot 2 \cdot \theta, \\ Y_n &= Y_{n-1}(1 + 2\theta - 1) = Y_{n-1} \cdot 2\theta, \end{aligned}$$

and when $Z_n = -1$ we have

$$\begin{aligned} Y'_n &= Y'_{n-1} \cdot 2 \cdot (1 - \theta), \\ Y_n &= Y_{n-1}(1 - 2\theta + 1) = Y_{n-1} \cdot 2(1 - \theta). \end{aligned}$$

An inductive argument shows that the two processes are indeed equivalent. In the previous section, we have shown that Y generates maximal profit when the alternative hypothesis is true. We conclude that Y'_n is the best process for disproving H_0 under H_1.

More generally we see that Theorem 3.2 allows us to split the *pay-off* and the *strategy*. A singleton hypothesis H_0 gives rise to $\mathbb{P}$, the inverse pay-off, and the alternative hypothesis is represented by $\mathbb{Q}$, the strategy. To return to our example, $\mathbb{Q}$ directly represents Alice's betting strategy R.

The latter two assertions of the theorem characterize the limit behavior of the martingale in terms of the absolute continuity of $\mathbb{P}$ and $\mathbb{Q}$ and vice versa. Note that the fact that $\mathbb{P}_n$ is absolutely continuous with respect to $\mathbb{Q}_n$ for all n does not necessarily imply that $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$ in the limit. Let us return to the example of Alice her game to see a counter-example. Under the null hypothesis H_0 we have, due to the strong law of large numbers[1],

$$P_0\left(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}Z_i = 0 = E_{\frac{1}{2}}[Z_1]\right) = 1.$$

Whereas the alternative hypothesis yields

$$P_1\left(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}Z_i = 2\theta - 1 = E_{\theta}[Z_1]\right) = 1.$$

This shows that there are tail events in $\mathcal{F}$ which have positive probability under P_1 yet no probability mass under P_0, thus showing that in the limit P_1 is not absolutely continuous with respect to P_0. Actually, P_0 and P_1 are mutually singular. By clause (2) of the theorem, we get that almost surely $Y'_n \to \infty$ under P_1 and $Y'_n \to 0$ under P_0 as $n \to \infty$.

We have shown that constructing a test martingale to distinguish between two point hypotheses can be done using the likelihood ratio. Moreover, we have shown that in this simple case, the likelihood ratio corresponds to the *best* investment strategy for Alice, if she really believes that H_1 is true. We have not yet shown how any of this relates to scientific discourse or *actually* distinguishing the two hypotheses.

3.3 The martingale test

The example of Alice has shown that the two hypotheses affect her capital accumulation completely differently. When the null hypothesis is true, she never expects to earn anything. Under the alternative hypothesis, on the other hand, she always expects to increase her capital. This suggests that a test martingale could be used as a statistical hypothesis test. As in classical hypothesis testing, we look at the *false positive rate*, or chance of committing a type I error.

Let us see how the capital accumulation is distributed under the null hypothesis. The following theorem (Dellacherie et al., 1982) is a martingale equivalent of Markov's inequality.

Theorem 3.3. *If M is a non-negative supermartingale then*

$$\mathbb{P}(\sup_n M_n \geq c) \leq \frac{E[M_0]}{c}$$

for $c > 0$.

Theorem 3.3 can be restricted in scope to test martingales to yield the following corollary.

[1]Note that Z_1 is Bernoulli in the sense of Remark 2.1. The expectation is thus zero when $\theta = \frac{1}{2}$.

Corollary 3.4. *In particular, for a test martingale $(M_n)_{n=0}^{\infty}$ we have*

$$\mathbb{P}(\sup_n M_n \geq c) \leq \frac{1}{c} \tag{3.5}$$

and the weaker, but more useful inequality

$$\mathbb{P}(M_\tau \geq c) \leq \frac{1}{c} \tag{3.6}$$

for any stopping time τ.

Equation (3.5) follows from the fact that a test martingale is a non-negative supermartingale with initial value one. Equation (3.6) is a consequence of the fact that the value of a stopped martingale can never exceed the supremum.

Corollary 3.4 allows us to define a statistical hypothesis test based on the test martingale.

Definition 3.5 (Martingale test)**.** *Let $M = (M_i)_{i=1}^{\infty}$ be a test martingale adapted to the filtration $\mathcal{F} = \bigcup_{i=1}^{\infty} \mathcal{F}_n$ on the probability space $(\Omega, \mathcal{F}, P_0)$ with probability measure P_0 corresponding to some null hypothesis. For any significance level α, and any stopping time τ, we define the statistic*

$$\begin{aligned} &\delta_{\alpha,\tau} : \Omega \to \{\textit{Accept, Reject}\} \\ &\delta_{\alpha,\tau} : M(\omega) \mapsto \begin{cases} \textit{Reject} & M_\tau \geq \frac{1}{\alpha}, \\ \textit{Accept} & \textit{otherwise}, \end{cases} \end{aligned}$$

to be a martingale test*.*

Note that Corollary 3.4 guarantees that any martingale test $\delta_{\alpha,\tau}$ has a false positive rate of at most α.

So how does this relate to classical statistics? Suppose Alice wants to use a martingale test to disprove some hypothesis with significance level $\alpha < 0.05$. Suppose she is investigating coin flips. It is widely believed that the coins she is investigating are fair. Yet Alice believes the coins actually have a 60% chance of coming up heads. In order to disprove the null hypothesis (the coins are fair), she sets up the game we have just detailed: she starts with a virtual capital of one Euro and always bets 60% of her capital on heads and 40% on tails. After every coin flip, she may decide whether or not to continue her experiment. As a consequence of Corollary 3.4, she may reject the null hypothesis when her virtual capital exceeds 20 Euro. Her decision has significance level (false positive rate) $\alpha < 0.05$.

Note that the value of the martingale *at the time when you stop* does not depend on the stopping rule: for any n and any stopping rule τ, which leads you to stop at time n, the value of M_n is the same. Compare this with the p-value of a classical hypothesis test

$$\delta'_{\alpha,\tau} : (X_1, X_2, \ldots) \mapsto \begin{cases} \text{Reject} & p_\tau(X_1, \ldots, X_\tau) \leq \alpha, \\ \text{Accept} & \text{otherwise}, \end{cases}$$

which *does* depend on the stopping rule τ: a different stopping rule leads to a different p-value statistic p_τ. Therefore, the stopping rule may not be

modified *during* the experiment, as that would change the p-value as well. In contrast to classical hypothesis testing, a martingale test does not require that the stopping rule be fixed before the experiment starts. Martingale tests retain their false positive rate for *any* stopping rule, due to Corollary 3.4. Therefore, a stopping rule that adjusts to the data (without looking into the future) is still a stopping time in the strict sense of the definition and thus preserves the false positive rate. Even external events, like research funding drying up, are allowed to stop the study if they are truly independent of the experimental outcomes.

Had Alice done the same experiment with classical p-value based testing, she would have had to fix a stopping rule before hand and *stick to it*. We have already noted (John et al., 2012) that more than 55% of psychologists have difficulty adhering to a stopping rule. They report a p-value for a sample as if the sample size had been fixed in advance, even if, in reality, the stopping rule was different. So the "p-value" they report is not a valid p-value. Dienes (2011) explains the dilemma these psychologists are in:

> This puts one in an impossible moral dilemma if, having tested once at the 5% level, an experiment yields a p of 0.06. One cannot reject the null on that number of subjects, yet one cannot accept it either (no matter what the official rules are, would you accept the null for a p of 0.06 when you have predicted an effect?). One cannot publish the data, yet one cannot in good heart bin the data and waste public resources. That would be immoral.

This dilemma simply vanishes when one uses a martingale test. It provides one straightforward solution: collect more data!

Every test martingales has an interpretation as a (conservative) p-value. The statistic

$$\overline{p}_\tau : (M_1, M_2, \ldots) \mapsto \frac{1}{M_\tau} \tag{3.7}$$

of a martingale M with respect to the stopping time τ is a conservative p-value in the sense of Equation (2.6). Due to Corollary 3.4, we have for any significance level α

$$P_0(\overline{p}_\tau(M) \leq \alpha) \leq \alpha$$

with P_0 the probability measure with respect to which M is a martingale. The p-value interpretation allows for combining research results easily. Lemma 3.6 shows how to combine multiple martingales into one. It is followed by an example showing how evidence from martingale tests can be merged in practice.

Lemma 3.6 (Optional continuation)**.** *Suppose we have a countable sequence of test martingales $M^1, M^2, M^3, \ldots$ adapted to filtrations $\mathcal{F}^1, \mathcal{F}^2, \mathcal{F}^3, \ldots$ respectively and a countable sequence of stopping times $\tau_1, \tau_2, \ldots$ on a probability space $(\Omega, \mathcal{F}, \mu)$. We require that*

1. *for all $n, k, i, j \geq 1$ we have that M^n_i is stochastically independent of M^k_j when $n \neq k$;*

2. *the stopping time τ_j is $\sigma(\mathcal{F}^1 \cup \cdots \cup \mathcal{F}^j)$-measurable for all $j \geq 1$;*

3. *for all $i \geq 1$, $\tau_i \geq 1$.*

Then the random process $\overline{M}$ defined by

$$\overline{M}_n = \begin{cases} M^1_n & n \leq \tau_1 \\ M^1_{\tau_1} \cdot M^2_{n-\tau_1} & 0 < n - \tau_1 \leq \tau_1 + \tau_2 \\ \vdots & \vdots \\ \left(\prod_{i=1}^k M^i_{\tau_i}\right) M^{k+1}_{\left(n-\sum_{i=1}^k \tau_i\right)} & 0 < n - \sum_{i=1}^k \tau_i \leq \tau_{k+1} \\ \vdots & \vdots \end{cases}$$

is a test martingale with respect to the filtration $\overline{\mathcal{F}}$ defined by $\overline{\mathcal{F}}_n = \sigma\left(\overline{M}_n\right)$.

Proof. Let us first define the random variable

$$S_k = \sum_{i=1}^k \tau_k$$

with $S_0 = 0$ and let

$$\begin{aligned} E_{0,0} &= \Omega, \\ E_{n,k} &= \{S_k < n \leq S_{k+1}\} \end{aligned}$$

denote the event in which the value of $\overline{M}_n$ comes from the martingale M^{k+1}. Then we may rewrite $\overline{M}$ as

$$\overline{M}_n = \sum_{k=0}^{\infty} \mathbb{1}_{\{E_{n,k}\}} \left(\prod_{i=1}^k M^i_{\tau_i}\right) M^{k+1}_{n-S_k}.$$

Before we verify the martingale property, we write $\overline{M}_{n+1}$ in the more manageable form

$$\begin{aligned} \overline{M}_{n+1} &= \sum_{k=0}^{\infty} \mathbb{1}_{\{E_{n+1,k}\}} \left(\prod_{i=1}^k M^i_{\tau_i}\right) M^{k+1}_{n+1-S_k} \\ &= \sum_{k=0}^{\infty} \mathbb{1}_{\{E_{n,k}\}} \left(\prod_{i=1}^k M^i_{\tau_i}\right) \left[\mathbb{1}_{\{E_{n+1,k}\}} M^{k+1}_{n+1-S_k} + \mathbb{1}_{\{E_{n+1,k+1}\}} M^{k+1}_{\tau_{k+1}} M^{k+2}_1\right]. \end{aligned}$$

We verify the martingale property. We have

$$\begin{aligned} E\left[\overline{M}_{n+1} \mid \overline{\mathcal{F}}_n\right] = \sum_{k=0}^{\infty} \mathbb{1}_{\{E_{n,k}\}} \left(\prod_{i=1}^k M^i_{\tau_i}\right) \Big(&E\left[\mathbb{1}_{\{E_{n+1,k}\}} M^{k+1}_{n+1-S_k} \mid \overline{\mathcal{F}}_n\right] + \\ &E\left[\mathbb{1}_{\{E_{n+1,k+1}\}} M^{k+1}_{\tau_{k+1}} M^{k+2}_1 \mid \overline{\mathcal{F}}_n\right]\Big). \end{aligned}$$

We can extract the $\overline{\mathcal{F}}_n$-measurable random variables from the conditional expectation to obtain

$$\begin{aligned} = \sum_{k=0}^{\infty} \mathbb{1}_{\{E_{n,k}\}} \left(\prod_{i=1}^k M^i_{\tau_i}\right) \Big(&\mathbb{1}_{\{E_{n+1,k}\}} E\left[M^{k+1}_{n+1-S_k} \mid \overline{\mathcal{F}}_n\right] + \\ &\mathbb{1}_{\{E_{n+1,k+1}\}} M^{k+1}_{\tau_{k+1}} E\left[M^{k+2}_1 \mid \overline{\mathcal{F}}_n\right]\Big). \end{aligned}$$

We use the martingale property to determine the conditional expectations

$$= \sum_{k=0}^{\infty} \mathbb{1}_{\{E_{n,k}\}} \left(\prod_{i=1}^{k} M_{\tau_i}^i \right) \left(\mathbb{1}_{\{E_{n+1,k}\}} M_{n-S_k}^{k+1} + \mathbb{1}_{\{E_{n+1,k+1}\}} M_{\tau_{k+1}}^{k+1} \cdot 1 \right).$$

When $S_k < n \leq S_k + \tau_{k+1} < n+1 \leq S_{k+2}$, then $n - S_k = \tau_{k+1}$. So we get

$$= \sum_{k=0}^{\infty} \mathbb{1}_{\{E_{n,k}\}} \left(\prod_{i=1}^{k} M_{\tau_i}^i \right) \left(\mathbb{1}_{\{E_{n+1,k}\}} M_{n-S_k}^{k+1} + \mathbb{1}_{\{E_{n+1,k+1}\}} M_{n-S_k}^{k+1} \right)$$

$$= \sum_{k=0}^{\infty} \mathbb{1}_{\{E_{n,k}\}} \left(\prod_{i=1}^{k} M_{\tau_i}^i \right) M_{n-S_k}^{k+1} \left(\mathbb{1}_{\{E_{n+1,k}\}} + \mathbb{1}_{\{E_{n+1,k+1}\}} \right)$$

Since we have $E_{n,k} = E_{n+1,k} \cup E_{n+1,k+1}$, we get

$$= \sum_{k=0}^{\infty} \mathbb{1}_{\{E_{n,k}\}} \left(\prod_{i=1}^{k} M_{\tau_i}^i \right) M_{n-S_k}^{k+1}$$

$$= \overline{M}_n.$$

Moreover, the initial value $\overline{M}_0$ equals 1 by definition and non-negativity is preserved. □

Suppose two martingale tests are performed. Then we can combine the evidence to reach a stronger conclusion. In fact we can simply multiply the final capital of the test martingales. Suppose Alice has just published her rejection of the null hypothesis (her final martingale capital M_τ exceeds 20). Bob, coincidentally, has also performed the same experiment. He has been less persevering: his final martingale capital $M'_{\tau'}$ slightly exceeds 5. Had Alice and Bob's study actually been one single martingale hypothesis test, then we can combine their results

$$\overline{M}_{\overline{\tau}} = M_\tau \cdot M'_{\tau'} \geq 20 \cdot 5 = 100$$

by multiplying their final martingale capitals to obtain a final capital exceeding 100. The interpretation is that Alice hands over her capital to Bob who uses it as initial capital. A martingale test based on the combined data would have rejected the null hypothesis at the 0.01 significance level. This corresponds to a p-value satisfying $p < 0.01$. Hence, although Bob's study is not significant at the 5% level, his results can be combined with Alice's.

Moreover, this is true even if Bob only decides to carry out this experiment because Alice has good results. Unlike in classical testing (see for example Armitage et al. (2002) pages 615–623), there is no penalty for "conditional" research: collecting more data because previous results look promising.

There is even more good news for Bob. Suppose he was studying these coin flips at the same time as Alice was doing her research. Then we are still allowed to multiply the outcome of his martingale with Alice's[2]. It is only

[2]This is not strictly proven by Lemma 3.6 as it requires that a current stopping time τ_i has no information on any future stopping time τ_{i+k}. If Bob's decision to stop is based on Alice her results, but not vice versa, Lemma 3.6 *does* apply. On the other hand, if Alice and Bob are communicating their results to such extent, one might as well consider their research as one single experiment.

required of Bob and Alice that they do not peek into the future, i.e., that their stopping decisions are true stopping times.

We must remain cautious though: if Bob decides not to publish his results, for instance because his martingale test never rejects the null hypothesis, then Charlie cannot come along and multiply his results with just those of Alice. Once Bob starts his experiment, his results *must* be part of the scientific record.

Another potential pitfall is peeking into the future inadvertently. Coins do not advertise what the next outcome will be. Patients in a medical trial, on the other hand, do. Suppose Bob is conducting a medical trial and half of his patients are either cured or have passed away. His trial is doing well: most of these patients have actually been cured and he has shown the efficacy of his treatment with significance level $\alpha < 0.05$. On the other hand, the remaining patients are not looking good: he expects that most of them will not be cured. Can Bob wrap up his medical trial? No: the fact that he has a reasonably expectation of the remaining patients' health means that he is in some way or another "peeking into the future". Bob has to finish the medical trial with all remaining patients.

We have shown how martingale tests can be used as a statistical hypothesis test. Furthermore, these martingale tests are robust under optional stopping and conveniently enable combining evidence of multiple studies. We have not yet shown how to construct martingale tests when the alternative hypothesis is not simple, but composite.

3.4 A uniform prior for the alternative hypothesis

Previously, we have shown how Alice was able to distinguish between two point hypotheses. In this section, we show how to distinguish between a simple null hypothesis and a composite alternative hypothesis. Now, Alice tests the coins using an alternative hypothesis with θ unknown instead of fixed. Recall

H_0'**:** The outcomes $Z_1, Z_2, \ldots$ are independent and identically Bernoulli-distributed with $\theta = \frac{1}{2}$.

H_1'**:** The outcomes $Z_1, Z_2, \ldots$ are independent and identically Bernoulli-distributed with *unknown* θ.

As we have seen in Section 3.1, Alice's strategy R does not influence whether or not the resulting random process is a martingale. Likewise, Theorem 3.2 shows that any alternative probability distribution $\mathbb{Q}$ which is absolutely continuous with respect to $\mathbb{P}$ for all n results in the successive likelihood ratios defining a martingale. In other words, the alternative hypothesis may employ a Bayesian prior on θ. In fact, *almost any* distribution may be used. Practically, one would prefer a distribution that more closely fits the alternative hypothesis. Also, one should ensure that for all finite n the alternative distribution is absolutely continuous with respect to the null distribution in order to be able to use Theorem 3.2.

We shall use a Bayesian uniform prior on θ to derive a very simple formula

for the test martingale. Let us first define

$$S_0 = 0,$$
$$S_n = \sum_{i=1}^{n} Z_i.$$

After n data points, S_n equals the number of negative data points subtracted from the number of positive data points. This yields

$$\#\{i \mid Z_i = 1, i < n\} = \frac{n - 1 + S_{n-1}}{2},$$
$$\#\{i \mid Z_i = -1, i < n\} = \frac{n - 1 - S_{n-1}}{2}.$$

Using Laplace's rule of succession (Grünwald (2007), page 258), we find that the uniform prior yields the following distribution for the alternative hypothesis

$$P_1(Z_n = 1 \mid Z_1, \ldots, Z_{n-1}) = \frac{\frac{n-1+S_{n-1}}{2} + 1}{n+1} = \frac{n + 1 + Z_n S_{n-1}}{2(n+1)}$$
$$P_1(Z_n = -1 \mid Z_1, \ldots, Z_{n-1}) = \frac{\frac{n-1-S_{n-1}}{2} + 1}{n+1} = \frac{n + 1 + Z_n S_{n-1}}{2(n+1)},$$

with $P_1(Z_1 = \pm 1) = 1/2$. We can also express the conditional likelihood ratio in terms of Z_n

$$\frac{P_1(Z_n \mid Z_1, \ldots, Z_{n-1})}{P_0(Z_n \mid Z_1, \ldots, Z_{n-1})} = \frac{2(n + 1 + Z_n S_{n-1})}{2(n+1)}$$
$$= 1 + Z_n \frac{S_{n-1}}{n+1}.$$

This yields the random process

$$Y_n = \prod_{i=1}^{n} (1 + Z_i \frac{S_{i-1}}{i+1}),$$

which can easily be verified to be a martingale under the null hypothesis by using Theorem 3.2 or by checking the martingale properties directly.

In this section we have shown how to execute a strategy for a compound alternative hypothesis. There were no major obstacles in obtaining the test martingale. The resulting martingale is not necessarily optimal, for instance, using Jeffreys prior on θ results in a test martingale with slightly higher variance. We have opted to use the uniform prior since it is somewhat easier to deal with analytically. We have yet to show how to construct a test martingale for composite *null* hypotheses.

3.5 Relating test martingales and Bayes factors

We shall compare the martingale interpretation of statistical hypothesis tests with Bayesian hypothesis testing . We shall first introduce the Bayesian point of view briefly (Ly et al., 2016).

Suppose again that one is comparing two competing hypotheses. Bayesian statisticians quantify the relative support of the observed data for the one hypothesis over the other with a Bayes factor. Suppose one has a prior belief w on hypothesis H_0 and H_1, and observed data D, then the Bayesian will update his or her belief using the following odds ratio

$$\frac{P(H_1 \mid D)}{P(H_0 \mid D)} = \frac{P(D \mid H_1)}{P(D \mid H_0)} \frac{w(H_1)}{w(H_0)}.$$

The ratio in the middle is independent of one's prior belief and represents the degree to which a Bayesian will update his or her belief based on the observed data. It is called the Bayes factor (Ly et al., 2016).

The astute reader will have noticed that this Bayes factor is in fact equal to the likelihood ratio which we have used to construct test martingales. This implies that *any Bayes factor with a simple null hypothesis is a test martingale*. Drawing this conclusion is not as straightforward in the case of compound null hypotheses, but we will show how this can be done in Section 6. We will use the terms Bayes factor and likelihood ratio interchangeably in the rest of this thesis.

In summary, we have shown how a betting interpretation of probability leads to the definition of the test martingale. We have shown that under absolutely continuity conditions, the likelihood ratio is a test martingale with respect to a simple null hypothesis. We have also seen that, if the null and alternative hypothesis are sufficiently distinguishable, this martingale tends to zero if the null hypothesis is true and to infinity if the alternative hypothesis is true. We have also defined the martingale test. In stark contrast to classical hypothesis tests, this statistical hypothesis test is robust under optional stopping and allows for easily combining evidence from multiple studies. Furthermore, when a compound alternative hypothesis with a Bayesian prior is employed, all properties of the test martingale (and thus the martingale test) are preserved. We have also seen that a Bayes factor is in fact a martingale when the null hypothesis is simple. We now turn our attention toward *composite*, instead of *simple*, null hypotheses.

4 Composite test martingales

In the previous section, we have seen that test martingales are very useful for doing statistical hypothesis tests. Unfortunately, its results have thus far been restricted to cases with a simple null hypotheses. We extend the work of Shafer et al. (2011) in such a way that the results from the previous section carry over to composite null hypotheses. We refer to these martingales as *composite test martingales*. All definitions and results in this section are my own.

There are two interpretations of a compound hypothesis with respect to test martingales:

1. There is a degree of belief (a prior) in the parameters of the null hypothesis. Some parameters may be more likely than others, or (if the prior and probability space permit) all parameters are equally likely in some sense. When data are sampled, the belief in the parameters is updated.

2. One views the hypothesis set as an arsenal of parameters in a game theoretic sense. Suppose Alice designs a test martingale and Bob is her adversary. Then Bob may pick any parameter from the null hypothesis to undermine the martingale property of Alice's test. Hence, Alice must construct a test that is a martingale for *all* parameters in the null hypothesis.

The first interpretation leaves the choice of prior open and is thus inherently subjective. Strictly speaking, it is not even a composite null hypothesis in the strict sense, as there is really only *one* probability distribution that corresponds to the combination of prior and parameter space.

In the second interpretation, the distribution of the sampled data may not be modified. Yet the goal remains to construct a simple null hypothesis. This is achieved by transforming the data in such a way that the transformed data points are identically distributed under any parameter in the null hypothesis. Instead of modifying the probability distribution of the data, we modify the data. In this thesis, we are concerned with the second interpretation exclusively.

Before we are able to give a mathematical definition of a composite test martingale in the second interpretation, we require some groundwork. Let $X = (X_n)_{n=1}^{\infty}$ be a random process adapted to the filtration $\mathcal{F} = (\mathcal{F}_n)_{n=1}^{\infty}$ ($\mathcal{F}_n \subseteq F_{n+1}$ for all $n \geq 1$).

Definition 4.1. *A* pivotal process *with respect to a compound null hypothesis H_0 is a measurable transformation $Z = (Z_n)_{\mathbb{N}}$ of the random process $(X_n)_{\mathbb{N}}$ adapted to the filtration $\mathcal{F}$ such that, for all $\mathbb{P}, \mathbb{P}' \in H_0$ and $A \in \mathcal{G} = \bigvee_{n=1}^{\infty} \sigma(Z_1, \ldots, Z_n)$*

$$\mathbb{P}(A) = \mathbb{P}'(A).$$

This concept is new, as far as we know, and is inspired by the term "pivotal quantity" for a random variable (Shao, 2003). In other words, Z is a transformation of X such that any event on Z has equal measure under *any* probability measure in H_0. All constructions of composite test martingales in this thesis involve a pivotal process.

Example 4.2. *Suppose $X_1, X_2, \ldots$ are independent $N(0, \sigma^2)$-distributed random variables for some value of σ. Then X_1 has equal probability of being positive as it has of being negative. Actually, this is true for any value of σ. So let us define the random variables for $i = 1, 2, \ldots$*

$$Z_i = \begin{cases} 1 & , X_i \geq 0 \\ -1 & , X_i < 0. \end{cases} \tag{4.1}$$

The random process $Z_1, Z_2, \ldots$ is a pivotal process with respect to the hypothesis

$$H = \{P_{0,\sigma^2} \mid \sigma \in \mathbb{R}, \sigma > 0\}.$$

Definition 4.3. *A* composite test martingale *with respect to a compound null hypothesis H_0 is a measurable transformation $M = (M_n)_{\mathbb{N}}$ of the random process $(X_n)_{\mathbb{N}}$ adapted to the filtration $\mathcal{F}$ such that there exists a filtration $\mathcal{G} = \lim_{n\to\infty} \mathcal{G}_n$, where $\mathcal{G}_n \subseteq \mathcal{F}_n$ is a coarsening of $\mathcal{F}_n$ and such that for all $\mathbb{P}_0 \in H_0$,*

1. *M is a non-negative martingale for $\mathbb{P}_0$ adapted to filtration $\mathcal{G}$, and*
2. *$M_0 = 1$.*

A trivial example of a composite test martingale is a process which equals one for any value of $(X_n)_{n=1}^{\infty}$. Obviously, not much can be gained by gambling with such pay-offs. We will find, however, that more interesting examples of composite test martingales exist. One such an example can be obtained looking at the likelihood ratios of the random process $(Z_i)_{i=1}^{\infty}$. We return to this example in Section 5, see Equation (5.1).

Note that Z is not necessarily a martingale with respect to the filtration $\mathcal{F}$. It might be necessary to "throw away some information" to prevent violating the martingale property under some probability measures in H_0.

Once a pivotal process has been found, constructing a composite test martingale is simple. From Theorem 3.2 it follows that with respect to a filtration $\mathcal{G}$ generated by a pivotal process Z,

$$M_n = \left.\frac{dQ}{dP}\right|_{\mathcal{G}_n} \tag{4.2}$$

is a supermartingale under probability measure $P_0 \in H_0$ when Q, P are probability measures on Z with $P \in H_0$. Equation (4.2) defines a martingale when Q is absolutely continuous with respect to P. In this thesis, P and Q will be absolutely continuous for all finite values of n and thus define a martingale.

Note that Theorem 3.3 and Corollary 3.4 apply to composite test martingales under any probability measure $P \in H_0$, thus yielding bounds on the false positive rate.

Although Equation (4.2) does not define the first test martingale for composite null hypotheses ever discovered (Vovk (1993) and Shafer et al. (2011) contain other examples), the construction using pivotal processes seems not to have been described before. The processes defined in Vovk (1993) and Shafer et al. (2011) are named "simultaneous test martingales". Unlike a composite test martingale, a simultaneous test martingale is not "alternative hypothesis agnostic": the construction only works for specific alternative hypotheses. This limitation seems to prevent the construction of a two-sided t-test using a simultaneous test martingale.

The next section provides an example of a composite test martingale.

5 The median martingale test

In this section, we construct a composite martingale test for normal random variables. As in Section 2.2, we want to distinguish between the two hypotheses

H_0**:** The random variables $X_1, X_2, \ldots$ are independent and identically $N(0, \sigma^2)$-distributed with unknown $\sigma > 0$;

H_1**:** The outcomes $X_1, X_2, \ldots$ with $X_i \sim N(\mu, \sigma^2)$ for all $i \in \{1, 2, \ldots\}$ are independent and identically distributed with unknown μ *and* $\sigma > 0$.

We want to design a random process, as a transformation of the random variables X_i, which is a martingale under any distribution in H_0. Hence, this random process must be identically distributed under any of these distributions. We shall see that there exists a trivial transformation for which this holds true.

As we have seen in Example 4.2, the random variables $Z_1, Z_2, \ldots$ defined by

$$Z_i = \begin{cases} 1 & , X_i \geq 0 \\ -1 & , X_i < 0 \end{cases} \tag{5.1}$$

define a pivotal process. We can recast our hypotheses in terms of $Z_1, Z_2, \ldots$ as follows

H_0'**:** The outcomes $Z_1, Z_2, \ldots$ are independent and identically Bernoulli-distributed with $\theta = \frac{1}{2}$.

H_1'**:** The outcomes $Z_1, Z_2, \ldots$ are independent and identically Bernoulli-distributed with *unknown* θ.

There exists a correspondence between hypotheses in H_1 and H_1', since any hypothesis in H_1 can be transformed to a hypothesis in H_1' using

$$\theta = P_{\mu,\sigma^2}(Z_i > 0) = \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\mu - x)^2}{2\sigma^2}}\, dx.$$

Note that the hypotheses on (X_n) imply the hypotheses on (Z_n) but not vice versa. The hypotheses on (Z_n) are substantially more general, since they require only that the underlying random variables X_i are symmetrically, but not necessarily normally, distributed.

Now that we have obtained the pivotal process $Z_1, Z_2, \ldots$ it is easy to construct a test martingale. We may actually use the example from Section 3.4. Recall that we used the uniform Bayesian prior on θ to obtain as alternative probability distribution

$$P_1(Z_1) = \frac{1}{2}$$
$$P_1(Z_n \mid Z_1, \ldots, Z_{n-1}) = \frac{n + 1 + Z_n S_{n-1}}{2(n+1)}.$$

Since the null hypothesis assigns equal probability to 1 and -1, the conditional likelihood ratio equals

$$\begin{aligned}\frac{P_1(Z_n \mid Z_1, \ldots, Z_{n-1})}{P_0(Z_n \mid Z_1, \ldots, Z_{n-1})} &= \frac{2(n+1+Z_n S_{n-1})}{2(n+1)} \\ &= 1 + Z_n \frac{S_{n-1}}{n+1}.\end{aligned}$$

By Theorem 3.2, this yields the composite test martingale

$$M_n = \prod_{i=1}^{n} (1 + Z_i \frac{S_{i-1}}{i+1}).$$

The random process M is a fine example of a composite test martingale. To stress the fact that it is a transformation of the random variable $X_1, X_2, \ldots$ we express M without the Z_i as

$$\begin{aligned}S_n &= \sum_{i=1}^{n} \mathbb{1}_{\{X_i \geq 0\}} - \mathbb{1}_{\{X_i < 0\}}, \\ M_n &= \prod_{i=1}^{n} (1 + (\mathbb{1}_{\{X_i \geq 0\}} - \mathbb{1}_{\{X_i < 0\}}) \frac{S_{i-1}}{i+1}),\end{aligned}$$

with $S_0 = 0$.

For any significance level α, and any stopping time τ, we define the statistic

$$\begin{aligned}&\delta_{\alpha,\tau} : \Omega \to \{\text{Accept, Reject}\} \\ &\delta_{\alpha,\tau} : M(\omega) \mapsto \begin{cases}\text{Reject} & M_\tau \geq \frac{1}{\alpha}, \\ \text{Accept} & \text{otherwise,}\end{cases}\end{aligned}$$

to be the *median martingale test*, which is a martingale test in accordance with Definition 3.5.

We have shown how to construct a martingale test with a compound null hypothesis. This test is actually not the strongest normal location test, since it ignores a lot information present in the data. In the next section, we will see an example of a more powerful martingale test. We examine and compare the power of this test in Section 7.

6 The Jeffreys Bayesian t-test

In this section, we discuss the Jeffreys Bayesian t-test (Jeffreys, 1961). This Bayesian test[3] is used in situations where frequentist statisticians tend to use the Student's t-test. It was popularized by Rouder et al. (2009) who were able to derive a more easily computable formula for the Bayes factor used in the test. The test was previously avoided for reasons of convenience, among others. We have two reasons to discuss this test. First of all, it is only fair that we compare the median martingale test not only to classical hypothesis tests, but also to the most popular available Bayesian test[4]. Furthermore, experimental results (Rouder, 2014) suggest that this Bayesian method is robust under optional stopping. It would be very fortunate if this widely used Bayesian method turns out to be a martingale: it would definitively prove that the Jeffreys Bayesian t-test allows for optional stopping.

As in Section 2.2, we shall distinguish between the two hypotheses

H_0**:** The random variables $X_1, X_2, \ldots$ are independent and identically $N(0, \sigma^2)$-distributed with unknown $\sigma > 0$;

H_1**:** The outcomes $X_1, X_2, \ldots$ with $X_i \sim N(\mu, \sigma^2)$ for all $i \in \{1, 2, \ldots\}$ are independent and identically distributed with unknown μ *and* $\sigma > 0$.

We first construct a scale-invariant pivotal process Z. We then introduce the Jeffreys Bayesian t-test and, most importantly, its Bayes factor: the Jeffreys-Rouder ratio. Our goal is to show that the Jeffreys-Rouder ratio is a martingale with respect to the pivotal process Z. In Section 6.3, we show that the Jeffreys-Rouder ratio is measurable in terms of the pivotal process Z. If it were not, it could not be a martingale with respect to Z. We develop Bayesian null and alternative densities for Z, the pivotal process, in Sections 6.4 and 6.5. We furthermore show that the Bayesian null density equals the pivotal null density. In Section 6.6, we use these results to prove that the Jeffreys-Rouder ratio is a likelihood ratio on the pivotal process Z. Finally, this allows us to conclude that the Jeffreys-Rouder ratio is a martingale.

6.1 A pivotal process

For the median test martingale we have devised a random variable that is scale invariant. We will do so again, but will sacrifice some of the independence properties that the median test martingale possessed. We will, however, retain a lot more of the information in the sample.

We define Z_i as follows:

$$Z_i := \frac{X_i}{|X_1|}. \tag{6.1}$$

Note that $Z_1 = \pm 1$. The rest of the data points are scaled to the first data point. Of course, it is possible that X_1 equals zero. This event has measure zero under the null hypothesis for all values of $\sigma > 0$, so in general we may

[3]The introduction in Ly et al. (2016) is more accessible to the modern reader.

[4]At the time of writing Rouder et al. (2009) has been cited over 900 times.

ignore this possibility. Let $\mathcal{B}$ be the Borel σ-algebra, then we may define the sub σ-algebra on $\mathbb{R} \setminus \{0\}$

$$\mathcal{B}' = \{B \setminus \{0\} \mid B \in \mathcal{B}\}, \tag{6.2}$$

where all occurrences of 0 have been removed. The collection $\mathcal{B}'$ satisfies all requirements of a σ-algebra as outlined in Section 2.1. We posit that X_1 is $\mathcal{B}'$-measurable.

We shall thus consider the random process $X_1, X_2, \ldots$ with respect to the filtration $\mathcal{F}' = \bigcup_{i=1}^{\infty} \mathcal{F}'_i$ with $F'_1 = \mathcal{B}'$ and $\mathcal{F}'_j = \mathcal{B}$ for all other j.

The random process $Z_1, Z_2, \ldots$ as defined above is adapted to the filtration $\mathcal{G} = \bigcup_{i=1}^{\infty} G_i$ with $\mathcal{G}_1 = \sigma(\{\{-1\}, \{1\}\})$ and $\mathcal{G}_j = \mathcal{B}$ for all $j \neq 1$. Lemma 6.1 shows how the induced probability measure on the first n outcomes of the random process Z,

$$\nu^n : \mathcal{G} \to [0, 1], \tag{6.3}$$

is distributed.

Lemma 6.1. *Suppose the random variables $X_1, \ldots, X_n$ are independent and identically $N(0, \sigma^2)$-distributed for some $\sigma > 0$ and are adapted to the filtration $\mathcal{F}'$. Let Z_i be defined as above and let $f_{\sigma^2} : \mathbb{R}^n \to \mathbb{R}$ denote the density of ν^n relative to the product measure $\rho^n = \rho_1 \times \cdots \times \rho_n$, where ρ_1 is the discrete uniform measure on $\{-1, 1\}$ and $\rho_2 \times \cdots \times \rho_n$ is the Lebesgue measure on $\mathbb{R}^{n-1}$. Then*

$$\frac{d\nu^n}{d\rho_1 \times \cdots \times \rho_n} = f_{\sigma^2}(z_1, z_2, \ldots, z_n) = \frac{\left(\sum_{i=1}^n z_i^2\right)^{-n/2} \Gamma(\frac{n}{2})}{\pi^{n/2}}. \tag{6.4}$$

Proof. Since all probability distributions in H_0 are symmetric around zero, Z_1 should equal -1 and 1 with equal probability. Hence ν^1 equals ρ_1. We have

$$f_{\sigma^2}(z_1) = 1. \tag{6.5}$$

Now let us look at the density of $Z_2, \ldots, Z_n$. Let

$$C := 1 + \sum_{i=2}^{n} z_i^2.$$

The following derivation is due to Wouter Koolen and Peter Grünwald. By conditioning f_{σ^2} on all possible values of x_1 and then integrating over x_1, we get

$$\begin{aligned}
f_{\sigma^2}(z_2, \ldots, z_n) &= \int_{\mathbb{R}\setminus\{0\}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x_1^2}{2\sigma^2}} \frac{1}{(2\pi\sigma^2/x_1^2)^{(n-1)/2}} e^{-\frac{x_1^2 \sum_{i=2}^n z_i^2}{2\sigma^2}} dx_1 \\
&= \frac{1}{(2\pi\sigma^2)^{(n-1)/2}} \int_{\mathbb{R}\setminus\{0\}} \frac{1}{\sqrt{2\pi}\sigma} |x_1|^{n-1} e^{-\frac{x_1^2}{2\sigma^2}} e^{-\frac{x_1^2 \sum_{i=2}^n z_i^2}{2\sigma^2}} dx_1 \\
&= \frac{1}{(2\pi\sigma^2)^{(n-1)/2}} \int_{\mathbb{R}\setminus\{0\}} \frac{1}{\sqrt{2\pi}\sigma} |x_1|^{n-1} e^{-\frac{x_1^2 C}{2\sigma^2}} dx_1.
\end{aligned}$$

Let $x = \sqrt{C}x_1$, use substitution to obtain

$$= \frac{1}{(2\pi\sigma^2)^{(n-1)/2}} C^{-n/2} \int_{\mathbb{R}\setminus\{0\}} \frac{1}{\sqrt{2\pi}\sigma} |x|^{n-1} e^{-\frac{x^2}{2\sigma^2}} dx.$$

Using the formula for absolute moments of a normal random variable, we get

$$= \frac{1}{(2\pi\sigma^2)^{(n-1)/2}} C^{-n/2} \sigma^{n-1} \frac{2^{\frac{n-1}{2}} \Gamma(\frac{n}{2})}{\sqrt{\pi}}.$$

This can be greatly simplified to

$$= \frac{C^{-n/2} \Gamma(\frac{n}{2})}{\pi^{n/2}}. \tag{6.6}$$

The conclusion follows from Equations (6.5) and (6.6). □

Note that this density does not depend on σ at all! The random variables Z_i are truly scale-invariant and thus define a pivotal process.

6.2 Jeffreys-Rouder ratio

The Jeffreys Bayesian t-test uses an improper prior on σ for both the null- and alternative hypothesis. It is the standard Jeffreys scale-invariant prior

$$\pi_0(\sigma) = \frac{1}{\sigma}.$$

Instead of providing a prior on μ, Rouder et al. (2009) provide a prior on $\delta = \frac{\mu}{\sigma}$, the effect size. They propose using a Cauchy prior

$$\pi'(\delta) = \frac{1}{\pi(1+\delta^2)}.$$

A prior on δ must neither be too wide, nor too narrow. A proper prior has analytical advantages compared to an improper prior. A more thorough justification of the Cauchy prior is given in Rouder et al. (2009).

We conclude that the hypotheses have priors

$$\pi_0(\sigma) = \frac{1}{\sigma},$$
$$\pi_1(\sigma, \delta) = \frac{1}{\pi\sigma(1+\delta^2)}.$$

The Bayesian marginal densities are thus defined as follows:

$$f_0(x_1, \ldots, x_n) = \int_0^\infty \sigma^{-1} \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{\sum_{i=1}^n x_i^2}{2\sigma^2}} d\sigma, \tag{6.7}$$

$$f_1(x_1, \ldots, x_n) = \int_{-\infty}^\infty \int_0^\infty \frac{1}{\pi\sigma(1+\delta^2)} \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{\sum_{i=1}^n (x_i - \delta\sigma)^2}{2\sigma^2}} d\sigma d\delta. \tag{6.8}$$

The densities defined in Equations (6.7) and (6.8) are quite imposing: solving these integrals is not trivial. Rouder et al. (2009) show[5] that the ratio of the densities, the Bayes factor, can be more economically expressed in terms of t_n, the t-statistic of the data. We have

$$\frac{f_1(x_1, \ldots, x_n)}{f_0(x_1, \ldots, x_n)} = JR(t_n(x_1, \ldots, x_n))$$

with $JR(t)$, the Jeffreys-Rouder ratio, defined as

$$\begin{aligned} JR(t) = & \left(1 + \frac{t^2}{\nu}\right)^{\frac{\nu+1}{2}} \\ & \cdot \int_0^\infty (1+ng)^{-\frac{1}{2}} \left(1 + \frac{t^2}{\nu(1+ng)}\right)^{-\frac{\nu+1}{2}} (2\pi)^{-\frac{1}{2}} g^{-\frac{3}{2}} e^{-\frac{1}{2g}}\, dg, \end{aligned} \tag{6.9}$$

where $\nu = n - 1$ equals the degrees of freedom and n equals the number of data points. We shall refer to this formula as the Jeffreys-Rouder ratio. By not writing "Bayes factor", we avoid the Bayesian connotation and invite the reader to view it as a mathematical object, a blank slate if you will. Although Jeffreys proposed the priors used here, Rouder et al. (2009) worked out the formula shown above. A more general derivation was already published in Liang et al. (2008). We will not derive the Jeffreys-Rouder ratio from Bayesian principles, but will be able to cast it in terms of a martingale interpretation.

6.3 The t-statistic in terms of Z_i

We first show that the Jeffreys-Rouder ratio is $\sigma(Z_1, \ldots, Z_n)$-measurable, with Z_i defined as in Equation (6.1). This should reassure the reader that it is possible for the Jeffreys-Rouder ratio to define a martingale on the random variables $Z_1, \ldots, Z_n$. Because the Jeffreys-Rouder ratio is a function of t_n, we only need to show that t_n is $\sigma(Z_1, \ldots, Z_n)$-measurable.

For the sample average we have[6]

$$\overline{X^n} = |X_1| \overline{Z^n}.$$

For the sample variance, we get

$$\begin{aligned} s_n^2(X^n) &= \frac{1}{n-1} \sum_{i=0}^{n} (X_i - \overline{X^n})^2 \\ &= \frac{1}{n-1} \sum_{i=0}^{n} (X_1 Z_i - X_1 \overline{Z^n})^2 \\ &= \frac{1}{n-1} \sum_{i=0}^{n} X_1^2 (Z_i - \overline{Z^n})^2 \\ &= |X_1|^2 s_n^2(Z^n). \end{aligned}$$

[5] The actual derivation is omitted, as it is here.

[6] If this notation is confusing, see the explanation near Equation (2.3).

So for t_n, we get

$$t_n = \sqrt{n}\frac{\overline{X^n}}{s_n} = \sqrt{n}\frac{\overline{Z^n}}{s_n(Z^n)}.$$

Conclude that t_n is $\sigma(Z_1, Z_2, \ldots, Z_n)$-measurable.

6.4 Bayesian null hypothesis

In this section, we examine the Bayesian null hypothesis. We calculate its density on the pivotal random variables $Z_2, \ldots, Z_n$ and find that under certain conditions it equals the classical null hypothesis density. Finally, we calculate the density on the first data point.

Lemma 6.2. *Under the Bayesian null hypothesis, the density*[7] *of $X_1, \ldots, X_n$ is given by*

$$f_X(x_1, \ldots, x_n) = \left(\sum_{i=1}^n x_i^2\right)^{-n/2} \frac{\Gamma(\frac{n}{2})}{2\pi^{n/2}} \tag{6.10}$$

and the conditional density of $(Z_2, \ldots, Z_n)$ is given by

$$f_Z(z_2, \ldots, z_n \mid x_1) = \frac{\left(\sum_{i=1}^n z_i^2\right)^{-n/2} \Gamma(\frac{n}{2})}{\pi^{n/2}}, \tag{6.11}$$

provided $x_1 \neq 0$.

Proof. Recall from Equation (6.7) that we have the following formula for the Bayesian marginal density

$$f_X(x_1, \ldots, x_n) = \int_0^\infty \sigma^{-1} \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{\sum_{i=1}^n x_i^2}{2\sigma^2}} d\sigma.$$

We can pull the variables $x_1, \ldots, x_n$ out of the integral. Let

$$C := \sum_{i=1}^n x_i^2.$$

Then we get

$$f_X(x_1, \ldots, x_n) = \int_0^\infty \sigma^{-1} \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{C}{2\sigma^2}} d\sigma.$$

Substitute $g = \sigma/\sqrt{C}$, to get

$$\begin{aligned}
&= \int_0^\infty C^{-1/2} g^{-1} \frac{1}{(2\pi C g^2)^{n/2}} e^{-\frac{1}{2g^2}} \sqrt{C} dg \\
&= C^{-n/2} \int_0^\infty g^{-1} \frac{1}{(2\pi g^2)^{n/2}} e^{-\frac{1}{2g^2}} dg \\
&= \left(\sum_{i=1}^n x_i^2\right)^{-n/2} \int_0^\infty g^{-1} \frac{1}{(2\pi g^2)^{n/2}} e^{-\frac{1}{2g^2}} dg.
\end{aligned}$$

[7] f_X is a probability density function with respect to the product Lebesgue measure λ^n.

The integral on the right integrates to

$$= \left(\sum_{i=1}^{n} x_i^2 \right)^{-n/2} \frac{\Gamma(\frac{n}{2})}{2\pi^{n/2}}.$$

This proves (6.10). Let us now look at $f_X(x_2, \ldots, x_n \mid x_1)$. We have

$$\begin{aligned} f_X(x_1, \ldots, x_n \mid x_1) &= \frac{f_X(x_1, \ldots, x_n)}{f_X(x_1)} \\ &= \frac{\left(\sum_{i=1}^{n} x_i^2\right)^{-n/2} \frac{1}{2}\pi^{-n/2}\Gamma(\frac{n}{2})}{|x_1|^{-1}\frac{1}{2}\pi^{-1/2}\Gamma(\frac{1}{2})} \\ &= \frac{\left(\sum_{i=1}^{n} x_i^2\right)^{-n/2} \pi^{-n/2}\Gamma(\frac{n}{2})}{|x_1|^{-1}}. \end{aligned}$$

When we look at the density f_Z of $(Z_2, \ldots, Z_n)$, we may use the multivariate chain rule to obtain

$$\begin{aligned} f_Z(z_2, \ldots, z_n \mid x_1) &= x_1^{n-1} f_X(x_1, x_1 z_2, \ldots, x_1 z_n \mid x_1) \\ &= \frac{x_1^{n-1}\left(\sum_{i=1}^{n} x_i^2\right)^{-n/2} \pi^{-n/2}\Gamma(\frac{n}{2})}{|x_1|^{-1}} \\ &= \left(\sum_{i=1}^{n} \frac{x_i^2}{x_1^2} \right)^{-n/2} \pi^{-n/2}\Gamma(\frac{n}{2}) \\ &= \frac{\left(\sum_{i=1}^{n} z_i^2\right)^{-n/2}\Gamma(\frac{n}{2})}{\pi^{n/2}}, \end{aligned}$$

which shows that (6.11) holds. □

Note that the density depends in no way upon X_1. We conclude that under the Bayesian null hypothesis the random variables $Z_2, \ldots, Z_n$ are independent of X_1.

For the density of the first data point, we have, using Equation (6.10),

$$f_X(x_1) = \frac{1}{2|x_1|}. \tag{6.12}$$

Because the prior on μ (the mean) is improper, the induced density $f_Z(z_1)$ is not defined. Since the density is symmetric, however, we may define an *auxiliary* null density f'_Z on Z that is compatible in its symmetry and is proper (integrates to one). The auxiliary null measure of Z_1 equals ρ_1 (as defined in Lemma 6.1) and the auxiliary null density becomes

$$f'_Z(z_1) = 1. \tag{6.13}$$

Using the independence we have just established, we may derive a density with respect to ρ^n on the full sequence of values of Z

$$\begin{aligned} f'_Z(z_1, z_2, \ldots, z_n) &= f_Z(z_2, \ldots, z_n \mid z_1) f'_Z(x_1) \\ &= f_Z(z_2, \ldots, z_n \mid x_1) f'_Z(z_1) \\ &= \frac{\left(\sum_{i=1}^{n} z_i^2\right)^{-n/2}\Gamma(\frac{n}{2})}{\pi^{n/2}}, \end{aligned} \tag{6.14}$$

provided $z_1 = \pm 1$. Furthermore, and this is really important, we must note that Equations (6.4) and (6.14) are equal. We find that for the classical density f_{σ^2} and the Bayesian density f'_Z

$$f_{\sigma^2}(z_1, z_2, \ldots, z_n) = f'_Z(z_1, z_2, \ldots, z_n) \tag{6.15}$$

holds for all $\sigma > 0$ and $z_1 \in \{-1, 1\}$.

6.5 Bayesian alternative hypothesis

In the previous section, we were able to derive an expression for the conditional density. We are not be able to do this for the Bayesian alternative density. We find that the alternative conditional density is not even truly independent of X_1: it depends on the value of Z_1. This is good enough for our purposes. Furthermore, we shall again calculate the density on the first data point. Finally, we shall construct an auxiliary alternative density.

Lemma 6.3. *Under the Bayesian alternative hypothesis, the random variables $(Z_2, \ldots, Z_n)$ are dependent on X_1 through Z_1 only. Let g_X denote the alternative Bayesian marginal density on X and let g_Z denote the alternative Bayesian marginal density on Z. We have*

$$g_Z(z_2, \ldots, z_n \mid x_1) = g_Z(z_2, \ldots, z_n \mid z_1),$$

provided $x_1 \neq 0$.

Proof. The Bayesian marginal density

$$g_X(x_1, \ldots, x_n) = \int_{-\infty}^{\infty} \int_0^{\infty} \frac{\sigma^{-1}}{\pi(1+\delta^2)} \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{\sum_{i=1}^n (x_i - \delta\sigma)^2}{2\sigma^2}} d\sigma d\delta$$

can be rewritten as

$$g_X(x_1, \ldots, x_n) = \int_{-\infty}^{\infty} \int_0^{\infty} \frac{\sigma^{-1}}{\pi(1+\delta^2)} \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{|x_1|^2 \sum_{i=1}^n (\frac{x_i}{|x_1|} - \frac{\delta\sigma}{|x_1|})^2}{2\sigma^2}} d\sigma d\delta.$$

Now substitute $s = \frac{\sigma}{|x_1|}$, to obtain

$$\begin{aligned} &= \int_{-\infty}^{\infty} \int_0^{\infty} \frac{|x_1| s^{-1}}{\pi(1+\delta^2)} \frac{1}{(2\pi x_1^2 s^2)^{n/2}} e^{-\frac{\sum_{i=1}^n (\frac{x_i}{|x_1|} - \delta s)^2}{2s^2}} |x_1|^{-1} ds\, d\delta \\ &= |x_1|^{-n} \int_{-\infty}^{\infty} \int_0^{\infty} \frac{s^{-1}}{\pi(1+\delta^2)} \frac{1}{(2\pi s^2)^{n/2}} e^{-\frac{\sum_{i=1}^n (\frac{x_i}{|x_1|} - \delta s)^2}{2s^2}} ds\, d\delta. \end{aligned}$$

The conditional density of $(X_2, \ldots, X_n)$ given X_1 is thus expressed as follows:

$$\begin{aligned} g_X(x_2, \ldots, x_n \mid x_1) &= \frac{g_X(x_1, \ldots, x_n)}{g_X(x_1)} \\ &= \frac{|x_1|^{-n} \int_{-\infty}^{\infty} \int_0^{\infty} \frac{s^{-1}}{\pi(1+\delta^2)} \frac{1}{(2\pi s^2)^{n/2}} e^{-\frac{\sum_{i=1}^n (\frac{x_i}{|x_1|} - \delta s)^2}{2s^2}} ds\, d\delta}{|x_1|^{-1} \int_{-\infty}^{\infty} \int_0^{\infty} \frac{s^{-1}}{\pi(1+\delta^2)} \frac{1}{(2\pi s^2)^{1/2}} e^{-\frac{(\frac{x_1}{|x_1|} - \delta s)^2}{2s^2}} ds\, d\delta}. \end{aligned}$$

We want to replace the denominator by a constant, but we still have a factor $\frac{x_1}{|x_1|}$ in the exponent of the denominator. It turns out that we can still do the integral as we have

$$\int_{-\infty}^{\infty}\int_{0}^{\infty} \frac{s^{-1}}{\pi(1+\delta^2)} \frac{1}{(2\pi s^2)^{1/2}} e^{-\frac{(\pm 1-\delta s)^2}{2s^2}} ds\, d\delta = \frac{1}{2}. \tag{6.16}$$

It follows from a symmetry argument that the integral with 1 in the exponent equals the integral with -1 in the exponent. Calculation[8] bears out the equality to one half. Now, from Equation (6.16) we may conclude

$$g_X(x_2,\ldots,x_n \mid x_1) = 2|x_1|^{-(n-1)} \cdot \int_{-\infty}^{\infty}\int_{0}^{\infty} \frac{s^{-1}}{\pi(1+\delta^2)} \frac{1}{(2\pi s^2)^{n/2}} e^{-\frac{\sum_{i=1}^{n}(\frac{x_i}{|x_1|}-\delta s)^2}{2s^2}} ds\, d\delta.$$

Let us determine the density of $(Z_2,\ldots,Z_n)$. We can apply the chain rule to obtain

$$\begin{aligned} g_Z(z_2,\ldots,z_n \mid x_1) &= |x_1|^{n-1} g_X(|x_1|z_2,\ldots,|x_1|z_n \mid x_1) \\ &= |x_1|^{n-1} 2|x_1|^{-(n-1)} \\ &\qquad \cdot \int_{-\infty}^{\infty}\int_{0}^{\infty} \frac{s^{-1}}{\pi(1+\delta^2)} \frac{1}{(2\pi s^2)^{n/2}} e^{-\frac{\sum_{i=1}^{n}(z_i-\delta s)^2}{2s^2}} ds\, d\delta \\ &= 2\int_{-\infty}^{\infty}\int_{0}^{\infty} \frac{s^{-1}}{\pi(1+\delta^2)} \frac{1}{(2\pi s^2)^{n/2}} e^{-\frac{\sum_{i=1}^{n}(z_i-\delta s)^2}{2s^2}} ds\, d\delta \\ &= 2\int_{-\infty}^{\infty}\int_{0}^{\infty} \frac{s^{-1}}{\pi(1+\delta^2)} \frac{1}{(2\pi s^2)^{n/2}} e^{-\frac{(\pm 1-\delta s)^2+\sum_{i=2}^{n}(z_i-\delta s)^2}{2s^2}} ds\, d\delta. \end{aligned}$$

Note that this Equation only depends on x_1 through z_1. The value of z_1 is restricted to ± 1. □

From Equation (6.16) we are able to deduce the density on the first data point

$$g_X(x_1) = \frac{1}{2|x_1|}. \tag{6.17}$$

Note that the Bayesian alternative density equals the density of the Bayesian null hypothesis on the first data point. This will be useful later on. As with the null hypothesis (see Equation (6.13)), we are required to define an *auxiliary* Bayesian alternative density g'_Z for the random variable Z. The auxiliary density with respect to ρ_1 (as defined in Lemma 6.1) equals

$$g'_Z(z_1) = 1. \tag{6.18}$$

We claim that $g_Z(z_2,\ldots,z_n \mid x_1)$ is a probability density, i.e., it integrates to one. The alternative hypothesis has a proper prior on δ and the posterior

[8] Replace the Cauchy prior on δ with a $N(0,g)$ distributed variable where g is distributed like an inverse Chi-squared distribution with one degree of freedom. This simplifies the integration and is equivalent, see Liang et al. (2008).

on σ, $w(\sigma \mid x_1)$, is proper when conditioned on the first data point. We can write

$$g_X(x_2, \ldots, x_n \mid x_1) = \int_0^\infty w(\delta, \sigma \mid x_1) h_{\delta\sigma,\sigma^2}(x_2, \ldots, x_n)\, d\sigma, \tag{6.19}$$

where h_{μ,σ^2} is the standard density of $n-1$ independent and identically $N(\mu, \sigma^2)$-distributed random variables. Since the conditioned prior is proper, the resulting density defines a probability distribution. Moreover, Equation (6.18) shows that

$$g'_Z(z_1, z_2, \ldots, z_n) = g_Z(z_2, \ldots, z_n \mid z_1) g'_Z(z_1) \tag{6.20}$$

is a "proper" probability density with respect to ρ^n as defined in Lemma 6.1.

We conclude this section. We have shown that under the alternative Bayesian hypothesis, the random variables $Z_2, \ldots, Z_n$ depend on X_1 through Z_1. Also, the density on the first data point equals the density of the null hypothesis. Finally, the auxiliary density g'_Z on $Z_1, Z_2, \ldots, Z_n$ is a proper density.

6.6 The Jeffreys-Rouder ratio is a martingale

The Jeffreys-Rouder ratio is a Bayes factor on the random variables $X_1, \ldots, X_n$ where both the null and the alternative hypothesis have an improper prior on the standard deviation. In the previous two sections, we have defined the null and alternative auxiliary Bayesian densities on the random process Z. We shall refer to their ratio as the *auxiliary Bayes factor* on Z. First, we show that the Jeffreys-Rouder ratio is equal to the auxiliary Bayes factor. Then, we show that this defines a likelihood ratio on the pivotal process Z. Finally, we prove that this likelihood ratio is a martingale.

Recall the Jeffreys-Rouder ratio

$$\begin{aligned} JR(t) &= \left(1 + \frac{t^2}{\nu}\right)^{\frac{\nu+1}{2}} \int_0^\infty (1+ng)^{-\frac{1}{2}} \left(1 + \frac{t^2}{\nu(1+ng)}\right)^{-\frac{\nu+1}{2}} (2\pi)^{-\frac{1}{2}} g^{-\frac{3}{2}} e^{-\frac{1}{2g}}\, dg \\ &= \frac{g_X(x_1, \ldots, x_n)}{f_X(x_1, \ldots, x_n)}, \end{aligned}$$

which is defined in terms of the t-statistic and thus can be calculated using both $X_1, \ldots, X_n$ and $Z_1, \ldots, Z_n$.

We first derive that the Jeffreys-Rouder ratio is equal to the auxiliary Bayes factor on $Z_1, \ldots, Z_n$. We may cancel $g_X(x_1)$ and $f_X(x_1)$, since they are equal. The chain rule then allows us to convert to densities on Z. As we have shown in the previous section, we may replace the dependence on x_1 by a dependence on z_1. Finally, we multiply both nominator and denominator by the auxiliary

density on Z_1. We obtain

$$\begin{aligned}
\frac{g_X(x_1,\ldots,x_n)}{f_X(x_1,\ldots,x_n)} &= \frac{g_X(x_2,\ldots,x_n \mid x_1)g_X(x_1)}{f_X(x_2,\ldots,x_n \mid x_1)f_X(x_1)} \\
&= \frac{g_X(x_2,\ldots,x_n \mid x_1)}{f_X(x_2,\ldots,x_n \mid x_1)} \\
&= \frac{|x_1|^{-(n-1)} g_Z(z_2,\ldots,z_n \mid x_1)}{|x_1|^{-(n-1)} f_Z(z_2,\ldots,z_n \mid x_1)} \\
&= \frac{g_Z(z_2,\ldots,z_n \mid x_1)}{f_Z(z_2,\ldots,z_n \mid x_1)} \\
&= \frac{g_Z(z_2,\ldots,z_n \mid z_1)g'_Z(z_1)}{f_Z(z_2,\ldots,z_n \mid z_1)f'_Z(z_1)} \\
&= \frac{g'_Z(z_1,\ldots,z_n)}{f'_Z(z_1,\ldots,z_n)}.
\end{aligned}$$

Now we show that the Jeffreys-Rouder ratio is actually a likelihood ratio on the pivotal process Z. Since $f'_Z(z_1,\ldots,z_n) = f_{\sigma^2}(z_1,\ldots,z_n)$ for all $\sigma > 0$ and $z_1 \in \{-1,1\}$ (see Equation (6.15)), we have

$$\frac{g_X(x_1,\ldots,x_n)}{f_X(x_1,\ldots,x_n)} = \frac{g'_Z(z_1,\ldots,z_n)}{f_{\sigma^2}(z_1,\ldots,z_n)}$$

for all $\sigma > 0$ and $z_1 \in \{-1,1\}$. In other words, the Jeffreys-Rouder ratio is a likelihood ratio on the pivotal process Z.

We have shown that the Jeffreys-Rouder ratio is not only a Bayes factor on the random variables $X_1,\ldots,X_n$, but also a likelihood ratio on the variables $Z_1,\ldots,Z_n$. The following theorem shows that this likelihood ratio is actually a martingale.

Theorem 6.4. *Let $X_1, X_2, \ldots$ be independent and identically $N(0,\sigma^2)$-distributed random variables adapted to the filtration $\mathcal{F}'$ as defined in Section 6.1. Let the random process $Z_1, Z_2, \ldots$ defined by*

$$Z_i = \frac{X_i}{|X_1|}$$

be adapted to the filtration $\mathcal{G} = \bigcup_{i=1}^{\infty} G_i$ with $\mathcal{G}_1 = \sigma(\{\{-1\},\{1\}\})$ and $\mathcal{G}_j = \mathcal{B}$ for all $j \neq 1$.

Then the likelihood ratio

$$\frac{g'_Z(Z_1,Z_2,\ldots,Z_n)}{f'_Z(Z_1,Z_2,\ldots,Z_n)} = \frac{g'_Z(Z_1,Z_2,\ldots,Z_n)}{f_{\sigma^2}(Z_1,Z_2,\ldots,Z_n)} \tag{6.21}$$

is a martingale with respect to the filtration $\mathcal{G}$ for all values of $\sigma > 0$.

Proof. The likelihood ratio is $\mathcal{G}_n$-measurable for all n. The measure defined by the density g'_Z is absolutely continuous with respect to the measure ν^n defined by the density f'_Z on the sigma algebra $\mathcal{G}_n$ for all values of n. First of all, both

measures assign equal weight to the events $\{Z_1 = -1\}$ and $\{Z_1 = -1\}$. Also, we have

$$f_{\sigma^2}(z_1, \ldots, z_n) = \frac{\left(\sum_{i=1}^n z_i^2\right)^{-n/2} \Gamma(\frac{n}{2})}{\pi^{n/2}} > 0$$

for all values of $\sigma > 0$ and $(z_1, \ldots, z_n) \in \{-1, 1\} \times \mathbb{R}^{n-1}$. Therefore, for any set $A \in \mathcal{G}$, we have $\nu^n(A) = 0$ if and only if $\rho^n(A) = 0$. Necessarily, the measure defined by the density g'_Z assigns zero mass to these sets as well.

Both densities integrate to one and hence define a proper probability measure. We may thus use Theorem 3.2 to reach the desired conclusion. □

The previous theorem is the main result of this thesis. It shows that the Jeffreys-Rouder ratio is a martingale. It can be verified that all results carry over when a symmetric prior other than the Cauchy prior is used for the effect size. When an asymmetric prior is used, this might still be the case but this is harder to verify. The distribution on Z_1 would then be an asymmetric Bernoulli distribution.

Notice that Theorem 6.4 does not hold when $X_1 = 0$. In theory, this event has probability zero and we have removed it in the filtration $\mathcal{F}'$. In practice, due to the fact that human measurements have finite precision, this event can definitely occur. Because Bayesian mixtures are invariant under permutation of the data points, one may swap the first data point with a non-zero data point to obtain a valid conclusion.

A martingale test based on the Jeffreys-Rouder ratio can be defined just like we have done at the end of Section 5. We will not do so here.

We discuss the experimental characteristics of the Jeffreys-Rouder Bayesian t-test in the next section. We discuss the implications of this theorem for Bayesian testing in the conclusion and we explore several extensions and generalizations in the discussion section.

7 Results

In this section, we compare the Student's t-test, the median martingale test, and the Jeffreys Bayesian t-test in terms of power (how long it takes to reach a conclusion) and robustness under misspecification (what happens when the data are not normally distributed).

7.1 Power - normal distribution

Differences between martingale and classical hypothesis tests make a fair comparison complicated.

First of all, because the stopping time is decided by the data, a martingale test does not stop after a fixed number of outcomes. Secondly, under the right conditions (as outlined in Theorem 3.2) the martingale test is guaranteed to reject the null hypothesis (it will almost surely require a finite, rather than infinite, number of data points).

The number of data points in a classical t-test, on the other hand, is usually fixed a priori. Based on the expected effect size and the required rate of success, a power analysis decides how many data points are required to reject the null hypothesis. Furthermore, even if the alternative hypothesis is true, there is no guarantee that the Student's t-test will reject the null hypothesis.

We would like to compare the *expected* number of samples required to reject the null hypothesis, but this requires that all tests reject the null hypothesis eventually, i.e., after a finite number of outcomes. Unfortunately, the Student's t-test is not guaranteed to reject the null hypothesis. Therefore, we cannot compare the expected sample size required to reject the null hypothesis. Instead, we measure the rate of rejection after thirty data points for each of the three methods.

Student's t-test We apply the Student's t-test to the samples with *and* without optional stopping. When no optional stopping is performed, the sample is tested once with all 30 data points at significance level $\alpha = 0.05$. When optional stopping *is* performed, the test is repeated for each additional outcome X_n as if the sample size had been fixed to n beforehand. Each of these tests has significance level $\alpha = 0.05$ individually. When any of the 30 tests rejects the null hypothesis, the null hypothesis is considered rejected. This scenario is included to illustrate the vulnerability to optional stopping of Student's t-test.

Martingale tests Both the median martingale test and the Jeffreys Bayesian t-test are performed with and without optional stopping. The tests with optional stopping reject the null hypothesis when the value of the test martingale exceeds 20 after any of the 30 outcomes. The tests without optional stopping only reject the null hypothesis when the value of the test martingale exceeds 20 after all 30 outcomes have been taken into account. We expect that all martingale tests have a significance level of $\alpha < 0.05$ both with and without optional stopping.

The tests are performed on generated data which is $N(\delta, 1)$-distributed for varying values of the effect size δ. For each effect size δ, we have simulated

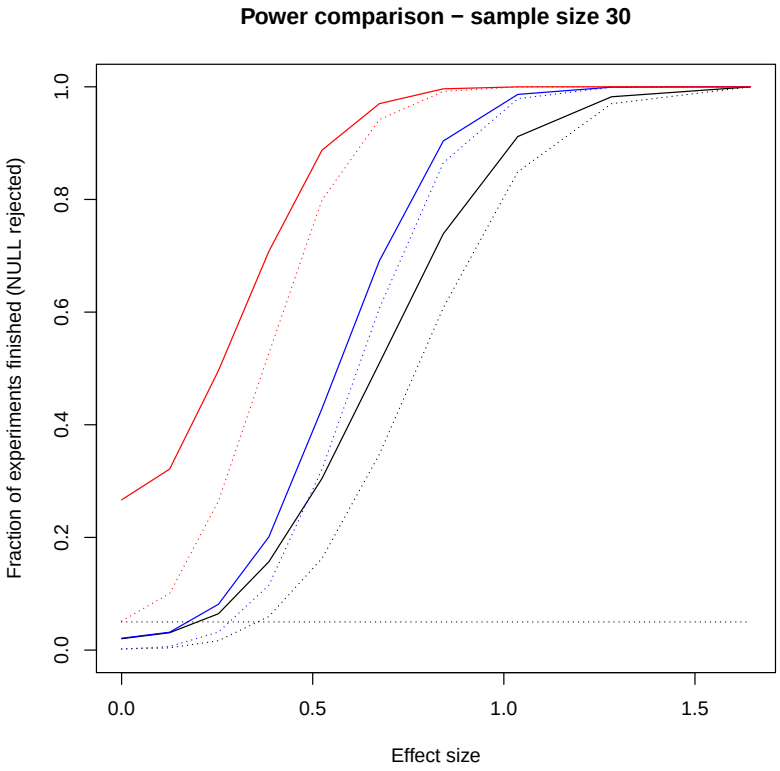

Figure 7.1: A comparison of the rejection rate (power) as the effect size increases. The Student's t-test is red, the median martingale test is black, and the Jeffreys Bayesian t-test is blue. The solid lines use optional stopping. The dotted lines do not. The threshold Bayes factor equals 20. The horizontal dotted line indicates a five percent false positive rate. The data points are normally distributed.

10,000 experiments with 30 data points each. When $\delta = 0$, the generated data correspond to the null hypothesis. We increase the effect size in steps until it reaches approximately 1.5. When $\delta = 1.5$, roughly 95% of the data points are greater than zero.

The results are shown in Figure 7.1. The rejection rate of the tests is plotted against the effect size on the x-axis. When $\delta = 0$, the rejection rate is know as the false positive rate because the generated data really *does* correspond to the null hypothesis. For all $\delta > 0$, the rejection rate equals the *power* (as defined in Section 2.3) with respect to the specific alternative hypothesis that the outcomes $X_1, \ldots, X_{30}$ are $N(\delta, \sigma)$-distributed for some value of $\sigma > 0$.

Notice that under the null hypothesis (effect size equals zero), most tests have a false positive rate of under five percent. As expected, the Student's t-test *with optional stopping* has a significantly higher false positive rate. The normal Student's t-test achieves a perfect five percent false positive rate and the other tests are more conservative.

Unsurprisingly, Student's t-test is the most powerful test by a wide margin. The Jeffreys Bayesian t-test is a distant second, followed by the median martingale test. Furthermore, optional stopping improves power, but not spectacularly. The difference between the solid and dotted lines is the fraction of data points where the martingale dips below 20 after exceeding it. This 'switching' seems to occur less in the Jeffreys Bayesian t-test.

We stress once more that optional stopping using the Student's t-test leads to unreliable conclusions. When it is used in an experiment consisting of thirty data points, it may lead to a false positive rate of almost 30 percent where 5 percent is acceptable.

7.2 Power - Cauchy distribution

When the data are correctly specified (they are normally distributed), the Student's t-test outperforms the other tests. We have checked whether this conclusion still holds when the data are instead Cauchy distributed. The same testing procedure has been followed, but instead of generating normally distributed data, the data follow a Cauchy distribution.

The results, shown in Figure 7.2, are quite different from Figure 7.1. We will ignore the solid red line (Student's t-test with optional stopping). First of all, notice that the false positive rate has not increased for the Jeffreys Bayesian t-test and the Student's t-test. This can possibly be explained by the fact that both the null and alternative hypothesis in these tests expect the data to be normally distributed and are penalized equally for the misspecification. Secondly, since it does not require normally distributed data, the median martingale test is the most powerful in this situation. Finally, the Student's t-test and Jeffreys Bayesian t-test do not achieve perfect accuracy even when the effect size exceeds 6.

We have shown that the Student's t-test is the most powerful test on normally distributed data. We advise that one should employ optional stopping when performing a martingale test to increase power. Furthermore, we see that both the Student's t-test and the Jeffreys Bayesian t-test are robust under misspecification when the data is Cauchy-distributed, although the median martingale test is most powerful on Cauchy-distributed data.

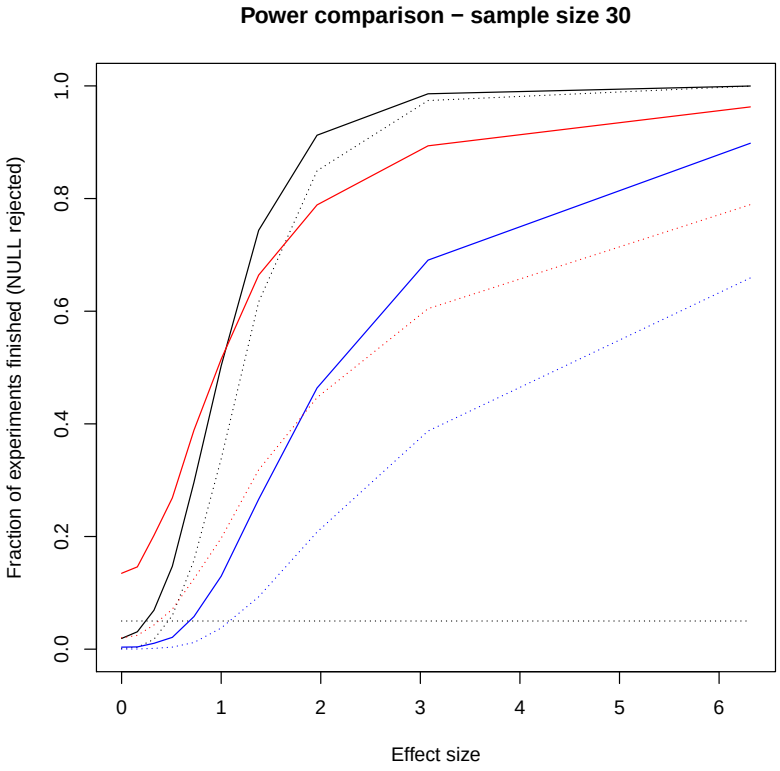

Figure 7.2: A comparison of the rejection rate (power) as the effect size increases. The Student's t-test is red, the median martingale test is black, and the Jeffreys Bayesian t-test is blue. The solid lines use optional stopping. The dotted lines do not. The threshold Bayes factor equals 20. The horizontal dotted line indicates a five percent false positive rate. The data points are Cauchy distributed.

8 Conclusion

In this section, we present the main conclusions of our thesis. The most important result is the existence of martingale tests for the one sample t-test. We may also conclude that the Jeffreys Bayesian t-test admits optional stopping. Finally, we remark that the standards of evidence for hypothesis tests that have both a Bayesian interpretation as well as a martingale interpretation are internally inconsistent.

Shafer et al. (2011) drew a compelling picture of how statistical testing could be made robust under optional stopping. Although they presented a test martingale that worked with respect to a specific compound null and alternative hypothesis, it was not clear at the time that their method could be extended to other statistical decisions involving compound hypotheses such as the t-test. We have shown that at least two such martingales exist for the one-sample t-test. Moreover, their existence suggests a general method of construction which we discuss in Section 9.2.

Simulations (Rouder, 2014) of the Jeffreys Bayesian t-test have already been undertaken to study its robustness under optional stopping. We have shown that the Jeffreys-Rouder ratio may be interpreted as a martingale under any distribution P corresponding to a normal distribution with $\mu = 0$ and some $\sigma > 0$. Hence, any optional sampling strategy that stops when the Bayes factor exceeds some $c > 1$ has a false positive rate of at most $\frac{1}{c}$ under the null hypothesis. Moreover, as can be concluded from Section 7, scientists can and should utilize optional stopping with martingale tests to increase the power of their experiments.

We posit that the current standards of evidence for statistical tests that have both a Bayesian interpretation as well as a martingale interpretation are *internally inconsistent*. Wetzels et al. (2011) provide "a practical comparison of p values, effect sizes, and default Bayes factors as measures of statistical evidence, using 855 recently published t tests in psychology". In this paper, the authors provide a reference table for the interpretation of p-values and Bayes factors. The criteria (see Table 8.1 for a summary) found in Wetzels et al. (2011) do not hold a Bayesian to the same standards of evidence they hold a classical statistician when a martingale test is used. These criteria can also be found in Jeffreys (1961), Kass and Raftery (1995), and Wasserman (2004).

Wetzels et al. (2011) rate a Bayes factor of $3 - 10$ as "substantial evidence for" the alternative hypothesis and a p-value in the range $0.001 - 0.01$ as "substantive evidence against" the null hypothesis (See Table 8.1 for a summary). Although they qualify a high Bayes factor as evidence *in favor of* the alternative hypothesis and a low p-value as evidence *against* the null hypothesis, one will conclude from Table 8.1 that a Bayes factor of 3 provides as much evidence in favor of one's conclusion as does a p-value of 0.01.

Now consider that the Jeffreys-Rouder ratio has both a Bayesian interpretation and a martingale interpretation. Moreover, consider that the conclusion one draws from the Jeffreys Bayesian t-test could be based on both the Bayes factor and the p-value. The interpretation of the Bayes factor and the interpretation of the p-value proposed in Wetzels et al. (2011), however, disagree by an order of magnitude.

For example, suppose an experiment is concluded with a Bayes factor

Table 8.1: This table is a summary of Table 1 in Wetzels et al. (2011). These criteria can also be found in Jeffreys (1961), Kass and Raftery (1995), and Wasserman (2004).

Statistic	Interpretation
p-value	
$< .001$	Decisive evidence against H_0
$0.001 - 0.01$	Substantive evidence against H_0
$0.01 - 0.05$	Positive evidence against H_0
> 0.05	No evidence against H_0
Bayes factor	
> 100	Decisive evidence for H_1
$30 - 100$	Very strong evidence for H_1
$10 - 30$	Strong evidence for H_1
$3 - 10$	Substantial evidence for H_1
$1 - 3$	Anecdotal evidence for H_1
1	No evidence

equal to 3. This corresponds to a p-value of < 0.34 in the martingale interpretation. According to Table 8.1, the Bayes factor offers "substantial evidence" for H_1, while the p-value offers "no evidence" against H_0. Likewise, the Bayes factor need only equal 100 to offer "decisive evidence", whereas the Jeffreys-Rouder ratio must equal at least 1000 to offer the same evidence when interpreted as a p-value. Hence, the standards of evidence for Bayesian tests are not strict enough, if one considers the false positive rate that using those standards of evidence entails.

Reading Wetzels et al. (2011), however, one should note that they find in practice that the Jeffreys Bayesian t-test is more stringent than Student's t-test, as they write:

> Our results showed that when the p value falls in the interval from .01 to .05, there is a 70% chance that the default Bayes factor indicates the evidence for the alternative hypothesis to be only anecdotal or "worth no more than a bare mention" [Bayes factor between 1 and 3];

This can be explained, however, by the fact that the Jeffreys Bayesian t-test is under-powered as compared to Student's t-test (as we have seen in Section 7). Furthermore, we must note (again) that the p-values in these 855 studies might not be "real" p-values since more than 55% of surveyed psychologists (John et al., 2012) admit to reporting a p-value for a sample as if the sample size had been fixed in advance, even if, in reality, the stopping rule was different. Hence, the evidence provided by these p-values might be artificially inflated.

Therefore, we conclude that even if the Bayesian standards of evidence are an improvement in practice, they should still be theoretically consistent with the standards of evidence for classical statistics. When a Bayesian test is a martingale test, its false positive rate is inversely proportional to the Bayes factor. The standards of evidence should be adjusted accordingly.

9 Discussion

This thesis has so far focused on the advantages of martingale tests. This section discusses some limitations of our approach: specifically, we discuss how to "cheat" when using a martingale method. Furthermore, we develop some conjectures as to how test martingales may be constructed for the two-sample t-test and how they may be constructed for composite null hypotheses in general.

This discussion section is incomplete if we fail to remark that there is "no free lunch": although martingale tests are robust under optional stopping, they are necessarily less powerful than classical hypothesis tests. Classical hypothesis tests are designed to be as powerful as possible at some fixed sample size. Martingale tests, on the other hand, must retain their false positive rate at *all* sample sizes.

In Figure 9.1, we see that the false positive rate of the Jeffreys Bayesian t-test (green) slowly increases towards the red line (the 5% significance level), but can never touch it without forgoing the optional stopping property: if it touches the red line, it can never reject the null hypothesis for a sample anymore however convincing the evidence found in the sample may be. Hence, martingale tests are resigned to be less powerful than classical hypothesis tests at any fixed sample size.

9.1 How to subvert a martingale test

One may wonder why we discuss how to cheat the system, rather than prescribe how to work within the confines of its rules. First of all, it is easier to describe what the theory does not allow than it is to isolate the behaviors that are allowed. More importantly, it is the author's belief that the intended effect, more trustworthy science, is more thoroughly achieved by describing how to undermine trustworthy science. Let us explain this by an example.

When a scientist is caught peeking at his p-values, or preregistering a study after collecting the data, which question would he rather answer?

1. "Did you not read that your statistical method does not allow for preregistering the study after collecting the data?"
2. "How come that you have implemented the subversion technique supplied by the authors of your statistical method?"

The first question is simply answered: "I must have overlooked". The second question can most charitably be answered with "But I have developed this subversion mechanism all by myself!". Another reason to focus on ways to subvert the system is that by reducing the number of ways the system *can be* subverted, the trust in *honest* results is bolstered. We present the methods of subversion in order of decreasing obviousness.

9.1.1 Withhold or alter data

Withholding or altering data influences *any* decision process, be it statistical or otherwise. In recent years, we have also seen total fabrications of evidence in science (Levelt et al., 2012). It needs no explanation that any alteration of the

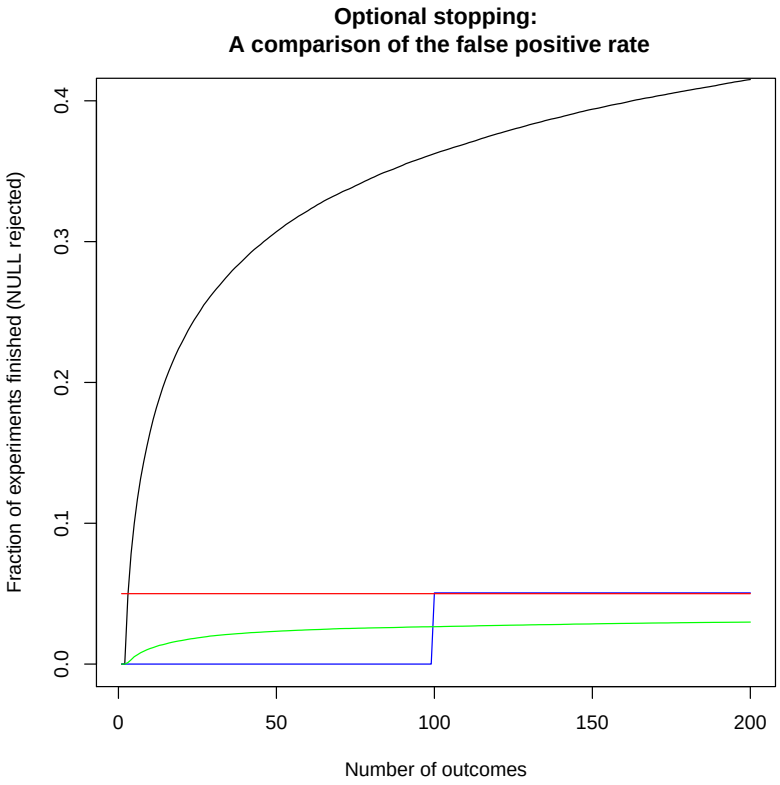

Figure 9.1: A comparison of the false positive rate as the sample size increases. The Student's t-test with optional stopping is black, the Jeffreys Bayesian t-test (with optional stopping) is green, and the Student's t-test stopping after 100 outcomes is blue. The horizontal red line indicates a five percent false positive rate. The data points are $N(0, 1)$-distributed.

observed data flaws the martingale hypothesis tests. This type of fraud can only be avoided by careful scrutiny. Altering or withholding data is widely considered to be scientific fraud – make use of this practice discretely, or risk being caught!

9.1.2 Peek into the future

We have seen in Section 3.3 that research subjects can sometimes give cues of their future performance. In medical trials, one could opt to cut the trial short if the remaining patients in the trial are not looking good. One could also classify the practice of cutting ongoing trials short as withholding data.

One can also influence the result without withholding any data by predicting some external influence. Suppose that one is doing an experiment with sunflowers. Suppose, furthermore, that the sunflowers are heavily influenced by the weather. Then one could, for example, opt to end the experiment and take measurements on the last predicted sunny day before a week of rain. The

martingale theory only works when future performance cannot be predicted. The weather, of course, is readily predicted. This allows a researcher to favorably influence the test results. This practice is plausibly deniable, which is a major advantage to the cheater.

The problem of peeking into the future does not play as large a role in traditional statistics because the stopping rule must be fixed before starting the experiment. Therefore, one cannot respond to the weather or patients' physical appearance during the experiment.

9.1.3 Change the strategy after having seen the data

Another great way to improve an experiment's significance level is changing the betting strategy, R, *after* the data has been collected. The two martingale tests developed in this thesis have been designed to be as objective as possible: they have as little a bias as possible towards any one of the possible alternative hypotheses. If one can plausibly reason that another strategy for the alternative hypothesis is required, one can greatly improve the final virtual capital of the martingale. This can only work if the study is *not* preregistered or if at least the hypothesis test is left unspecified.

A variant of this tactic is also possible. As discussed in Section 7, the performance of the Jeffreys Bayesian t-test is not always better than the median martingale test. One could compute the outcome for both tests and use the outcome that is most advantageous. It is the author's suspicion that this is already common practice for decision problems where multiple statistical tests are available. Without preregistration, choosing the most advantageous statistical test is plausibly deniable. On the other hand, others can easily calculate the results one would have gotten with another test. Always choosing the most favorable test could raise suspicion.

Because a test martingale affords a gambling interpretation, a post-hoc change of the betting strategy should intuitively raise more suspicion than choosing a different traditional test. Let us consider the following scenario:

> When Bob goes to the casino, he usually plays at the roulette table for twenty rounds. If he is unhappy with his gains, he asks the croupier to exactly replay the last twenty rounds so that Bob can improve his strategy.

It takes no genius to see that Bob is cheating. This is exactly what happens when an alternative hypothesis is modified after the fact to suit the data.

9.1.4 Use publication bias to your advantage

The final method we discuss requires extraordinary dedication. One can make use of publication bias to steer the scientific record in a preferred direction. To quote Shafer et al. (2011):

> 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.

One can run several distinct[9] experiments to disprove the same null hypothesis and only publish the positive results. It requires a lot of work, but allows one to prove *any* statement. This is actually the same as withholding data, but because the experiments are artificially separated, one can claim that no data was withheld in any one of the experiments with a positive result and the results of the negative experiments were simply not worthy of publication. If one is required to preregister experiments, this strategy would not withstand scrutiny. The large number of unpublished yet otherwise identical studies would raise suspicion.

To conclude, as a cheater, one should try to avoid preregistering a study. This allows one to positively influence the outcome without resorting to withholding or altering data. If one is forced to preregister the study, one can still identify predictable external factors that can positively influence the outcome of the hypothesis test. If one is careful about avoiding charges of fraud, the use of external factors is most recommended.

9.2 Further work

Much remains to be discovered about composite test martingales. We discuss ways in which our method can be extended or generalized. We first discuss the normal scale test, which is the analog of the t-test for the standard deviation parameter: instead of fixing $\mu = 0$ and leaving the choice of σ open, we fix σ to some value and leave the choice of μ open. We also propose a possible pivotal process for the two-sample t-test, which is perhaps the most popular statistical hypothesis test. It compares two samples from different populations and tests if the population means differ.

Suppose one wants to test the variance of a sample of normal random variables. Suppose furthermore that one wishes to disprove a well-established belief that the standard deviation of these random variables equals α. Then one wishes to distinguish between the hypotheses

H_0**:** The random variables $X_1, X_2, \ldots$ are independent and identically $N(\mu, \alpha^2)$-distributed with unknown μ and known $\alpha > 0$;

H_1**:** The outcomes $X_1, X_2, \ldots$ with $X_i \sim N(\mu, \sigma^2)$ for all $i \in \{1, 2, \ldots\}$ are independent and identically distributed with unknown μ *and* $\sigma > 0$.

We conjecture that these hypotheses afford a discriminating test martingale.

Conjecture 9.1. *Suppose X_i are independent and identically $N(\mu, \sigma^2)$-distributed random variables for some $\mu \in \mathbb{R}, \sigma > 0$ and let the random process $(Z_i)_{\mathbb{N}}$ be defined by*

$$Z_i = X_i - X_1$$

with $Z_1 = 0$. Let f_{μ,σ^2} denote the induced probability density of the random variables $Z_1, \ldots, Z_n$ relative to the product measure $\rho^n = \rho_1 \times \cdots \times \rho_n$, where ρ_1 is the Dirac

[9]The experiments should preferably be performed in different locations and by different people.

measure on $\{0\}$ *and* $\rho_2 \times \cdots \times \rho_n$ *is the Lebesgue measure on* $\mathbb{R}^{n-1}$. *Let* f_0 *denote the Bayesian marginal density on* $Z_1, \ldots, Z_n$ *with Lebesgue prior on* μ *and fixed scale* α *with respect to* ρ^n.

Then for any value of $\mu, \nu \in \mathbb{R}$ *and some fixed* $\alpha > 0$, *we have*

$$f_{\mu,\alpha^2}(z_1, \ldots, z_n) = f_{\nu,\alpha^2}(z_1, \ldots, z_n) \tag{9.1}$$

and

$$f_{\mu,\alpha^2}(z_1, \ldots, z_n) = f_0(z_1, \ldots, z_n). \tag{9.2}$$

Furthermore, Let f_1 *denote the alternative Bayesian marginal density on* $Z_1, \ldots, Z_n$ *with Lebesgue prior on* μ *and any proper prior on* σ *with respect to* $\rho_1 \times \cdots \times \rho_n$. *Then the Bayes factor*

$$\frac{f_1(z_1, \ldots, z_n)}{f_0(z_1, \ldots, z_n)} \tag{9.3}$$

is a test martingale with respect to the null hypothesis.

Now, suppose one wants to perform a two-sample t-test. The competing hypotheses are

H_0**:** The outcomes from the two populations $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ are independent and identically $N(\mu, \sigma^2)$-distributed with unknown μ and unknown $\sigma > 0$;

H_1**:** The outcomes from the one population $X_1, X_2, \ldots$ are independent and identically $N(\mu - \alpha, \sigma^2)$-distributed with unknown μ, α, and $\sigma > 0$. Also, the outcomes from the other population $Y_1, Y_2, \ldots$ are independent and identically $N(\mu + \alpha, \sigma^2)$-distributed.

A Bayes factor has been developed for this test by Rouder et al. (2009) as well. Perhaps, a martingale test exists to distinguish between these hypotheses. In any case, we conjecture that a pivotal process can be constructed with respect to the null hypothesis.

Conjecture 9.2. *Let* $X_1, \ldots, X_n$ *and* $Y_1, \ldots, Y_m$ *be sequences of independent and identically distributed* $N(\mu, \sigma^2)$*-distributed random variables and let* ρ_1 *and* $\rho_3 \times \cdots \times \rho_n$ *be defined as in Conjecture 9.1. We define* ρ_2 *to be the discrete uniform measure on* $\{-1, 1\}$.

Define

$$Z_i = \frac{X_i - Y_1}{|X_2 - Y_1|}$$
$$Z_i' = \frac{Y_i - X_1}{|Y_2 - X_1|}$$

with $Z_1 = Z_1' = 0$, $Z_2 = \pm 1$ *and* $Z_2' = \pm 1$. *We posit that* $(Z_i)_{i=1}^{\infty}$ *and* $(Z_i')_{i=1}^{\infty}$ *are pivotal processes with respect to the null hypothesis.*

The two pivotal processes, if they are truly pivotal, enable one to construct two test martingales. Using Lemma 3.6 one can combine these martingales

into one big martingale. We further believe that using a Lebesgue prior on μ and Jeffreys prior $\pi'(\sigma) = \frac{1}{\sigma}$ on σ allows one to construct an auxiliary Bayesian marginal density that equals the induced density f_{μ,σ^2} on Z and Z' with respect to $\rho_1 \times \rho_2 \times \cdots \rho_n$. These priors are also used for the Bayesian two-sample t-test constructed by Rouder et al. (2009). Although it would be a fortunate coincidence if this martingale equals their Bayes factor, this is not certain at all.

Finally, one would hope that there exists a general proof which shows that the induced density on pivotal process with respect to a parameter θ always equals the Bayesian density with a non-informative prior on θ. The work of Zidek (1969) might prove useful in this regard.

In summary, test martingales unavoidably trade power for robustness under optional stopping. Furthermore, although martingale tests are more difficult to abuse than classical hypothesis tests, methods of subversion do exist. Perhaps publishers should require a martingale test if a study is not preregistered. The lack of power of the martingale test would drive many to adopt preregistration. Finally, we conjecture that extensions to the one sample martingale test exist. If these conjectures are true, they suggest that a general connection between non-informative priors and pivotal processes exists.

10 Bibliography

References

P Armitage, G Berry, and J. N. S. Matthews. *Statistical methods in medical research*. Blackwell Science Ltd, 2002.

C Glenn Begley and Lee M Ellis. Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533, 2012.

Open Science Collaboration et al. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

Claude Dellacherie, Paul-André Meyer, translated, and prepared by J.P. Wilson. *Probabilities and Potential, B Theory of Martingales.* Elsevier, 1982. ISBN 9780080871837.

Zoltan Dienes. Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, 6(3):274–290, 2011.

Rick Durrett. *Probability: theory and examples*. Cambridge university press, 2010.

Economist. Trouble at the lab. *Economist*, 2013. URL `http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble`.

Peter D. Grünwald. *The minimum description length principle*. MIT Press, 2007. ISBN 978-0-262-07281-6.

Peter D. Grünwald. Toetsen als gokken: een redelijk alternatief. *Nieuw Archief voor Wiskunde*, 17(4):236, 2017. URL `www.nieuwarchief.nl/serie5/pdf/naw5-2016-17-4-236.pdf`.

John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, 2005. doi: 10.1371/journal.pmed.0020124. URL `https://doi.org/10.1371/journal.pmed.0020124`.

Harold Jeffreys. *Theory of probability*. Oxford University Press, Oxford, 1961.

Leslie K John, George Loewenstein, and Drazen Prelec. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, 2012.

Daniel Kahneman. A new etiquette for replication. *Social Psychology*, 45(4): 310, 2014.

Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

Willem J. M. Levelt, P. J. D. Drenth, and E. Noort. *Flawed science: The fraudulent research practices of social psychologist Diederik Stapel*. Commissioned by the Tilburg University, University of Amsterdam and the University of Groningen, 2012.

Feng Liang, Rui Paulo, German Molina, Merlise A Clyde, and Jim O Berger. Mixtures of g Priors for Bayesian Variable Selection. *Journal of the American Statistical Association*, 103(481):410–423, 2008. ISSN 0162-1459. doi: 10.1198/016214507000001337.

Alexander Ly, Josine Verhagen, and Eric Jan Wagenmakers. Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, 72:19–32, 2016. ISSN 10960880. doi: 10.1016/j.jmp.2015.06.004.

Karl Popper. *The logic of scientific discovery*. Routledge, 1959.

John Rice. *Mathematical statistics and data analysis*. Nelson Education, 2006.

Jeffrey N. Rouder. Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, 21(2):301–308, 2014. doi: 10.3758/s13423-014-0595-4.

Jeffrey N. Rouder, Paul L. Speckman, Dongchu Sun, Richard D. Morey, and Geoffrey Iverson. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2):225–237, 2009. doi: 10.3758/pbr.16.2.225.

Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test Martingales, Bayes Factors and p-values. *Statistical Science*, 26(1): 84–101, 2011. ISSN 0883-4237. doi: 10.1214/10-STS347. URL `http://projecteuclid.org/euclid.ss/1307626567`.

Jun Shao. *Mathematical Statistics*. Springer texts in statistics, 2003.

Rongfeng Sun. Lecture notes: Non-negative martingales as changes of measure, optional stopping theorem, backwards/reversed martingales., 2016. URL `http://www.math.nus.edu.sg/~matsr/ProbII/Lec4.pdf`.

Stéphanie van der Pas and Peter Grünwald. Almost the best of three worlds: Risk, consistency and optional stopping for the switch criterion in nested model selection. *arXiv preprint arXiv:1408.5724*, 2014.

Jean Ville. *Étude critique de la notion collectif*. PhD thesis, L'université de Paris, 1939.

Vladimir G Vovk. A logic of probability, with application to the foundations of statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 317–351, 1993.

Abraham Wald. *Sequential analysis*. Courier Corporation, 1973.

L Wasserman. *All of statistics: A concise course in statistical inference*. New York: Springer, 2004.

R. Wetzels, D. Matzke, M. D. Lee, J. N. Rouder, G. J. Iverson, and E.-J. Wagenmakers. Statistical Evidence in Experimental Psychology: An Empirical Comparison Using 855 t Tests. *Perspectives on Psychological Science*, 6(3):291–298, 2011. ISSN 1745-6916. doi: 10.1177/1745691611406923.

James V Zidek. A representation of bayes invariant procedures in terms of haar measure. *Annals of the Institute of Statistical Mathematics*, 21(1):291–308, 1969.