Bayesian Analysis (2021) **16**, Number 3, pp. 961–989

Optional Stopping with Bayes Factors: A Categorization and Extension of Folklore Results, with an Application to Invariant Situations

Allard Hendriksen*, Rianne de Heide†, and Peter Grünwald†

Abstract. It is often claimed that Bayesian methods, in particular Bayes factor methods for hypothesis testing, can deal with optional stopping. We first give an overview, using elementary probability theory, of three different mathematical meanings that various authors give to this claim: (1) stopping rule *independence*, (2) posterior *calibration* and (3) (semi-) *frequentist robustness to optional stopping*. We then prove theorems to the effect that these claims do indeed hold in a general measure-theoretic setting. For claims of type (2) and (3), such results are new. By allowing for non-integrable measures based on improper priors, we obtain particularly strong results for the practically important case of models with nuisance parameters satisfying a group invariance (such as location or scale). We also discuss the practical relevance of (1)–(3), and conclude that whether Bayes factor methods actually perform well under optional stopping crucially depends on details of models, priors and the goal of the analysis.

Keywords: Bayesian testing, optional stopping, Bayes factors, group invariance, right Haar prior.

1 Introduction

In recent years, a surprising number of scientific results have failed to hold up to continued scrutiny. Part of this 'replicability crisis' may be caused by practices that ignore the assumptions of traditional (frequentist) statistical methods (John et al., 2012). One of these assumptions is that the experimental protocol should be completely determined upfront. In practice, researchers often adjust the protocol due to unforeseen circumstances or collect data until a point has been proven. This practice, which is referred to as *optional stopping*, can cause true hypotheses to be wrongly rejected much more often than these statistical methods promise.

Bayes factor hypothesis testing has long been advocated as an alternative to traditional testing that can resolve several of its problems; in particular, it was claimed early on that Bayesian methods continue to be valid under optional stopping (Lindley, 1957; Raiffa and Schlaifer, 1961; Edwards et al., 1963). In particular, the latter paper claims that (with Bayesian methods) "it is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time,

*CWI, Amsterdam, allard.hendriksen@cwi.nl
†CWI, Amsterdam and Leiden University, The Netherlands, r.de.heide@cwi.nl; pdg@cwi.nl

 https://doi.org/10.1214/20-BA1234

money, or patience." In light of the replicability crisis, such claims have received much renewed interest (Wagenmakers, 2007; Rouder, 2014; Schönbrodt et al., 2017; Yu et al., 2014; Sanborn and Hills, 2014). But what do they mean mathematically? It turns out that different authors mean quite different things by 'Bayesian methods handle optional stopping'; moreover, such claims are often shown to hold only in an informal sense, or in restricted contexts. Thus, the first goal of the present paper is to give a systematic overview and formalization of such claims in a simple, expository setting and, still in this simple setting, explain their relevance for practice: can we effectively rely on Bayes factor testing to do a good job under optional stopping or not? As we shall see, the answer is subtle. The second goal is to extend the reach of such claims to more general settings, for which they have never been formally verified and for which verification is not always trivial.

Overview In *Section 2*, we give a systematic overview of what we identified to be the three main mathematical senses in which Bayes factor methods can handle optional stopping, which we call *τ-independence, calibration, and (semi-)frequentist*. We first do this in a setting chosen to be as simple as possible — finite sample spaces and strictly positive probabilities — allowing for straightforward statements and proofs of results. In *Section 3*, we explain the practical relevance of these three notions. It turns out that whether or not we can say that 'the Bayes factor method can handle optional stopping' in practice is a subtle matter, depending on the specifics of the given situation: what models are used, what priors, and what is the goal of the analysis. We can thus explain the paradox that there have also been claims in the literature that Bayesian methods *cannot* handle optional stopping in certain cases; such claims were made, for example by Yu et al. (2014); Sanborn and Hills (2014), and also by ourselves (de Heide and Grünwald, 2018). We also briefly discuss *safe tests* (Grünwald et al., 2019) which can be interpreted as a novel method for determining priors that behave better under frequentist optional stopping. The paper has been organized in such a way that these first two sections can be read with only basic knowledge of probability theory and Bayesian statistics. For convenience, we illustrate Section 3 with an informally stated example involving group invariances, so that the reader gets a complete overview of what the later, more mathematical sections are about.

Section 4 extends the statements and results to a much more general setting allowing for a wide range of sample spaces and measures, including measures based on *improper priors*. These are priors that are not integrable, thus not defining standard probability distributions over parameters, and as such they cause technical complications. Such priors are indispensable within the recently popularized *default Bayes factors* for common hypothesis tests (Rouder et al., 2009, 2012; Jamil et al., 2016).

In *Section 5*, we provide stronger results for the case in which both models satisfy the same group invariance. Several (not all) default Bayes factor settings concern such situations; prominent examples are Jeffreys' (1961) Bayesian one- and two-sample t-tests, in which the models are location and location-scale families, respectively. Many more examples are given by Berger and various collaborators (Berger et al., 1998a; Dass and Berger, 2003; Bayarri et al., 2012, 2016). These papers provide compelling arguments

for using the (typically improper) *right Haar prior* on the nuisance parameters in such situations; for example, in Jeffreys' one-sample t-test, one puts a right Haar prior on the variance. In particular, in our restricted context of Bayes factor hypothesis testing, the right Haar prior does not suffer from the *marginalization paradox* (Dawid et al., 1973) that often plagues Bayesian inference based on improper priors. Nevertheless, the right Haar prior is not entirely without problems either (we briefly return to these points in the conclusion).

Haar priors and group invariant models were studied extensively by Eaton (1989); Andersson (1982); Wijsman (1990), whose results this paper depends on considerably. When nuisance parameters (shared by both H_0 and H_1) are of suitable form and the right Haar prior is used, we can strengthen the results of Section 4: they now hold uniformly for all possible values of the nuisance parameters, rather than in the marginal, 'on average' sense we consider in Section 4. However — and this is an important insight — we *cannot take arbitrary stopping rules* if we want to handle optional stopping in this strong sense: our theorems only hold if the stopping rules satisfy a certain intuitive condition, which will hold in many but not all practical cases: the stopping rule must be "invariant" under some group action. For instance, a rule such as 'stop as soon as the Bayes factor is ≥ 20' is allowed, but a rule (in the Jeffreys' one-sample t-test) such as 'stop as soon as $\sum x_i^2 \geq 20$' is not.

Scope and Novelty Our analysis is restricted to Bayesian testing and model selection using the Bayes factor method; we do not make any claims about other types of Bayesian inference. Some of the results we present were already known, at least in simple settings; we refer in each case to the first appearance in the literature that we are aware of. In particular, our results in Section 4.1 are implied by earlier results in the seminal work by Berger and Wolpert (1988) on the likelihood principle; we include them any way since they are a necessary building block for what follows. The real mathematical novelties in the paper are the results on calibration and (semi-) frequentist optional stopping with general sample spaces and improper priors and the results on the group invariance case (Section 4.2–5). These results are truly novel, and — although perhaps not very surprising — they do require substantial additional work not covered by Berger and Wolpert (1988), who are only concerned with τ-independence. In particular, the calibration results require the notion of the 'posterior odds of some particular posterior odds', which need to be defined under arbitrary stopping times. The difficulty here is that, in contrast to the fixed sample sizes where even with continuous-valued data, the Bayes factor and the posterior odds usually have a distribution with full support, with variable stopping times, the support may have 'gaps' at which its density is zero or very near zero. An additional difficulty encountered in the group invariance case is that one has to define filtrations based on maximal invariants, which requires excluding certain measure-zero points from the sample space.

2 The Simple Case

Consider a finite set $\mathcal{X}$ and a sample space $\Omega := \mathcal{X}^T$ where T is some very large (but in this section, still finite) integer. One observes a *sample* $x^\tau \equiv x_1, \ldots, x_\tau$, which is an

initial segment of $x_1, \ldots, x_T \in \mathcal{X}^T$. In the simplest case, $\tau = n$ is a sample size that is fixed in advance; but, more generally τ is a *stopping time* defined by some stopping rule (which may or may not be known to the data analyst), defined formally below.

We consider a hypothesis testing scenario where we wish to distinguish between a null hypothesis H_0 and an alternative hypothesis H_1. Both H_0 and H_1 are sets of distributions on Ω, and they are each represented by unique probability distributions $\bar{P}_0$ and $\bar{P}_1$ respectively. Usually, these are taken to be Bayesian marginal distributions, defined as follows. First one writes, for both $k \in \{0, 1\}$, $H_k = \{P_{\theta|k} \mid \theta \in \Theta_k\}$ with 'parameter spaces' Θ_k; one then defines or assumes some prior probability distributions π_0 and π_1 on Θ_0 and Θ_1, respectively. The Bayesian marginal probability distributions are then the corresponding marginal distributions, i.e. for any set $A \subset \Omega$ they satisfy:

$$\bar{P}_0(A) = \int_{\Theta_0} P_{\theta|0}(A)\, \mathrm{d}\pi_0(\theta) \quad ; \quad \bar{P}_1(A) = \int_{\Theta_1} P_{\theta|1}(A)\, \mathrm{d}\pi_1(\theta). \tag{1}$$

For now we also further assume that for every $n \leq T$, every $x^n \in \mathcal{X}^n$, $\bar{P}_0(X^n = x^n) > 0$ and $\bar{P}_1(X^n = x^n) > 0$ (full support), where here, as below, we use random variable notation, $X^n = x^n$ denoting the event $\{x^n\} \subset \Omega$. We note that there exist approaches to testing and model choice such as testing by nonnegative martingales (Shafer et al., 2011; van der Pas and Grünwald, 2018) and minimum description length (Barron et al., 1998; Grünwald, 2007) in which the $\bar{P}_0$ and $\bar{P}_1$ may be defined in different (yet related) ways. Several of the results below extend to general $\bar{P}_0$ and $\bar{P}_1$; we return to this point at the end of the paper, in Section 6. In all cases, we further assume that we have determined an additional probability mass function π on $\{H_0, H_1\}$, indicating the prior probabilities of the hypotheses. The evidence in favor of H_1 relative to H_0 given data x^τ is now measured either by the *Bayes factor* or the *posterior odds*. We now give the standard definition of these quantities for the case that $\tau = n$, i.e., that the sample size is fixed in advance. First, noting that all conditioning below is on events of strictly positive probability, by Bayes' theorem, we can write for any $A \subset \Omega$,

$$\frac{\pi(H_1 \mid A)}{\pi(H_0 \mid A)} = \frac{P(A \mid H_1)}{P(A \mid H_0)} \cdot \frac{\pi(H_1)}{\pi(H_0)}, \tag{2}$$

where here, as in the remainder of the paper, we use the symbol π to denote not just prior, but also posterior distributions on $\{H_0, H_1\}$. In the case that we observe x^n for fixed n, the event A is of the form $X^n = x^n$. Plugging this into (2), the left-hand side becomes the standard definition of *posterior odds*, and the first factor on the right is called the *Bayes factor*.

2.1 First Sense of Handling Optional Stopping: τ-Independence

Now, in reality we do not necessarily observe $X^n = x^n$ for fixed n but rather $X^\tau = x^\tau$ where τ is a stopping time that may itself depend on (past) data (and that in some cases may in fact be unknown to us). This stopping time may be defined in terms of a *stopping rule* $f : \bigcup_{i \geq 0}^{T} \mathcal{X}^i \to \{\texttt{stop}, \texttt{continue}\}$. $\tau \equiv \tau(x^T)$ is then defined as the random variable which, for any sample $x_1, \ldots, x_T$, outputs the smallest n such that $f(x_1, \ldots, x_n) = \texttt{stop}$.

For any given stopping time τ, any $1 \leq n \leq T$ and sequence of data $x^n = (x_1, \ldots, x_n)$, we say that x^n *is compatible with* τ if it satisfies $X^n = x^n \Rightarrow \tau = n$. We let $\mathcal{X}^\tau \subset \bigcup_{i=1}^{T} \mathcal{X}^i$ be the set of all sequences compatible with τ.

Observations take the form $X^\tau = x^\tau$, which is equivalent to the event $X^n = x^n$; $\tau = n$ for some n and some $x^n \in \mathcal{X}^n$ which of necessity must be compatible with τ. We can thus instantiate (2) to

$$\frac{\pi(H_1 \mid X^n = x^n, \tau = n)}{\pi(H_0 \mid X^n = x^n, \tau = n)} = \frac{P(\tau = n \mid X^n = x^n, H_1) \cdot \pi(H_1 \mid X^n = x^n)}{P(\tau = n \mid X^n = x^n, H_0) \cdot \pi(H_0 \mid X^n = x^n)} = \\ = \frac{\pi(H_1 \mid X^n = x^n)}{\pi(H_0 \mid X^n = x^n)}, \tag{3}$$

where in the first equality we used Bayes' theorem (keeping $X^n = x^n$ on the right of the conditioning bar throughout); the second equality stems from the fact that $X^n = x^n$ logically implies $\tau = n$, since x^n is compatible with τ; the probability $P(\tau = n \mid X^n = x^n, H_j)$ must therefore be 1 for $j = 0, 1$. Combining (3) with Bayes' theorem we get:

$$\overbrace{\frac{\pi(H_1 \mid X^n = x^n, \tau = n)}{\pi(H_0 \mid X^n = x^n, \tau = n)}}^{\gamma(x^n)} = \overbrace{\frac{\bar{P}_1(X^n = x^n)}{\bar{P}_0(X^n = x^n)}}^{\beta(x^n)} \cdot \frac{\pi(H_1)}{\pi(H_0)}, \tag{4}$$

where we introduce the notation $\gamma(x^n)$ for the posterior odds and $\beta(x^n)$ for the Bayes factor based on sample x^n, calculated as if n were fixed in advance.[1]

We see that the stopping rule plays no role in the expression on the right. Thus, we have shown that, for any two stopping times τ_1 and τ_2 that are both compatible with some observed x^n, the posterior odds one arrives at will be the same irrespective of whether x^n came to be observed because τ_1 was used or if x^n came to be observed because τ_2 was used. We say that the posterior odds do not depend on the stopping rule τ and call this property τ-*independence*. Incidentally, this also justifies that we write the posterior odds as $\gamma(x^n)$, a function of x^n alone, without referring to the stopping time τ.

The fact that the posterior odds given x^n do not depend on the stopping rule is the first (and simplest) sense in which Bayesian methods handle optional stopping. It has its roots in the *stopping rule principle*, the general idea that the conclusions obtained from the data by 'reasonable' statistical methods should not depend on the stopping rule used. This principle was probably first formulated by Barnard (1947; 1949); Barnard (1949) very implicitly showed that, under some conditions, Bayesian methods satisfy the stopping rule principle (and hence satisfy τ-independence). Other early sources are Lindley (1957) and Edwards et al. (1963). Lindley gave an informal proof in the context of specific parametric models; in Section 4.1 we show that, under some regularity conditions, the result indeed remains true for general σ-finite $\bar{P}_0$ and $\bar{P}_1$.

[1] A slightly different way to get to (4), which some may find even simpler, is to start with $\bar{P}_0(X^n = x^n, \tau = n) = \bar{P}_0(X^n = x^n)$ (since $X^n = x^n$ implies $\tau = n$), whence $\pi(H_j \mid X^n = x^n, \tau = n) \propto \bar{P}_j(X^n = x^n, \tau = n)\pi(H_j) = \bar{P}_j(X^n = x^n)\pi(H_j)$.

A special case of our result (allowing continuous-valued sample spaces but not general measures) was proven by Raiffa and Schlaifer (1961), and a more general statement about the connection between the 'likelihood principle' and the 'stopping rule principle' which implies our result in Section 4.1 can be found in the seminal work (Berger and Wolpert, 1988), who also provide some historical context. Still, even though not new in itself, we include our result on τ-independence with general sample spaces and measures since it is the basic building block of our later results on calibration and semi-frequentist robustness, which are new.

Finally, we should note that both Raiffa and Schlaifer (1961) and Berger and Wolpert (1988) consider more general stopping rules, which can map to a probability of stopping instead of just $\{\texttt{stop}, \texttt{continue}\}$. Also, they allow the stopping rule itself to be parameterized: one deals with a collection of stopping rules $\{f_\xi : \xi \in \Xi\}$ with corresponding stopping times $\{\tau_\xi : \xi \in \Xi\}$, where the parameter ξ is equipped with a prior such that ξ and H_j are required to be a priori independent. Such extensions are straightforward to incorporate into our development as well (very roughly, the second equality in (3) now follows because, by conditional independence, we must have that $P(\tau_\xi = n \mid X^n = x^n, H_1) = P(\tau_\xi = n \mid X^n = x^n, H_0)$); we will not go into such extensions any further in this paper.

2.2 Second Sense of Handling Optional Stopping: Calibration

An alternative definition of handling optional stopping was introduced by Rouder (2014). Rouder calls $\gamma(x^n)$ the *nominal* posterior odds calculated from an obtained sample x^n, and defines the *observed posterior odds* as

$$\frac{\pi(H_1 \mid \gamma(x^n) = c)}{\pi(H_0 \mid \gamma(x^n) = c)}$$

as the posterior odds given the nominal odds. Rouder first notes that, at least if the sample size is fixed in advance to n, one expects these odds to be equal. For instance, if an obtained sample yields nominal posterior odds of 3-to-1 in favor of the alternative hypothesis, then it must be 3 times as likely that the sample was generated by the alternative probability measure. In the terminology of de Heide and Grünwald (2018), Bayes is *calibrated* for a fixed sample size n. Rouder then goes on to note that, if n is determined by an arbitrary stopping time τ (based for example on optional stopping), then the odds will still be equal — in this sense, Bayesian testing is well-behaved in the calibration sense irrespective of the stopping rule/time. Formally, the requirement that the nominal and observed posterior odds be equal leads us to define the *calibration hypothesis*, which postulates that $c = \frac{P(H_1 \mid \gamma = c)}{P(H_0 \mid \gamma = c)}$ holds for any $c > 0$ that has non-zero probability. For simplicity, for now we only consider the case with equal prior odds for H_0 and H_1 so that $\gamma(x^n) = \beta(x^n)$. Then the calibration hypothesis says that, for arbitrary stopping time τ, for every c such that $\beta(x^\tau) = c$ for some $x^\tau \in \mathcal{X}^\tau$, one has

$$c = \frac{P(\beta(x^\tau) = c \mid H_1)}{P(\beta(x^\tau) = c \mid H_0)}. \tag{5}$$

In the present simple setting, this hypothesis is easily shown to hold, because we can write:

$$\frac{P(\beta(X^\tau) = c \mid H_1)}{P(\beta(X^\tau) = c \mid H_0)} = \frac{\sum_{y \in \mathcal{X}^\tau : \beta(y) = c} P(\{y\} \mid H_1)}{\sum_{y \in \mathcal{X}^\tau ; \beta(y) = c} P(\{y\} \mid H_0)} = \frac{\sum_{y \in \mathcal{X}^\tau : \beta(y) = c} cP(\{y\} \mid H_0)}{\sum_{y \in \mathcal{X}^\tau : \beta(y) = c} P(\{y\} \mid H_0)} = c.$$

Rouder noticed that the calibration hypothesis should hold as a mathematical theorem, without giving an explicit proof; he demonstrated it by computer simulation in a simple parametric setting. Deng et al. (2016) gave a proof for a somewhat more extended setting yet still with proper priors. In Section 4.2 we show that a version of the calibration hypothesis continues to hold for general measures based on improper priors, and in Section 5.4 we extend this further to strong calibration for group invariance settings as discussed below.

We note that this result, too, relies on the priors themselves not depending on the stopping time, an assumption which is violated in several standard default Bayes factor settings. We also note that, if one thinks of one's priors in a default sense — they are practical but not necessarily fully believed — then the practical implications of calibration are limited, as shown experimentally by de Heide and Grünwald (2018). One would really like a stronger form of calibration in which (5) holds under a whole range of distributions in H_0 and H_1, rather than in terms of $\bar{P}_0$ and $\bar{P}_1$ which average over a prior that perhaps does not reflect one's beliefs fully. For the case that H_0 and H_1 share a nuisance parameter g taking values in some set G, one can define this *strong calibration hypothesis* as stating that, for all c with $\beta(x^\tau) = c$ for some $x^\tau \in \mathcal{X}^\tau$, all $g \in G$,

$$c = \frac{P(\beta(x^\tau) = c \mid H_1, g)}{P(\beta(x^\tau) = c \mid H_0, g)}, \tag{6}$$

where β is still defined as above; in particular, when calculating β one does not condition on the parameter having the value g, but when assessing its likelihood as in (6) one does. de Heide and Grünwald (2018) show that the strong calibration hypothesis certainly does *not* hold for general parameters, but they also show by simulations that it does hold in the practically important case with group invariance and right Haar priors (Example 1 provides an illustration). In Section 5.4 we show that in such cases, one can indeed prove that a version of (6) holds.

2.3 Third Sense of Handling Optional Stopping: (Semi-)Frequentist

In classical, Neyman-Pearson style null hypothesis testing, a main concern is to limit the false positive rate of a hypothesis test. If this false positive rate is bounded above by some $\alpha > 0$, then a null hypothesis significance test (NHST) is said to have *significance level* α, and if the significance level is independent of the stopping rule used, we say that the test is *robust under frequentist optional stopping.*

Definition 1. A function $S : \bigcup_{i=m}^{T} \mathcal{X}^i \to \{0, 1\}$ is said to be a frequentist sequential test with significance level α and minimal sample size m that is *robust under optional*

stopping relative to H_0 if for all $P \in H_0$

$$P\left(\exists n, m < n \leq T : S(X^n) = 1\right) \leq \alpha,$$

i.e. the probability that there is an n at which $S(X^n) = 1$ ('the test rejects H_0 when given sample X^n') is bounded by α.

In our present setting, we can take $m = 0$ (larger m become important in Section 4.3), so n runs from 1 to T and it is easy to show that, for any $0 \leq \alpha \leq 1$, we have

$$\bar{P}_0\left(\exists n, 0 < n \leq T : \frac{1}{\beta(x^n)} \leq \alpha\right) \leq \alpha. \tag{7}$$

Proof. For any fixed α and any sequence $x^T = x_1, \ldots, x_T$, let $\tau(x^T)$ be the smallest n such that, for the initial segment x^n of x^T, $\beta(x^n) \geq 1/\alpha$ (if no such n exists we set $\tau(x^T) = T$). Then τ is a stopping time, X^τ is a random variable, and the probability in (7) is equal to the $\bar{P}_0$-probability that $\beta(X^\tau) \geq 1/\alpha$, which by Markov's inequality is bounded by α. □

It follows that, if H_0 is a singleton, then the sequential test S that rejects H_0 (outputs $S(X^n) = 1$) whenever $\beta(x^n) \geq 1/\alpha$ is a frequentist sequential test with significance level α that is robust under optional stopping.

The fact that Bayes factor testing with singleton H_0 handles optional stopping in this frequentist way was noted by Edwards et al. (1963) and also emphasized by Good (1991), among many others. If H_0 is not a singleton, then (7) still holds, so the Bayes factor still handles optional stopping in a mixed frequentist (Type I-error) and Bayesian (marginalizing over prior within H_0) sense. From a frequentist perspective, one may not consider this to be fully satisfactory, and hence we call it 'semi-frequentist'. In some quite special situations though, it turns out that the Bayes factor satisfies the stronger property of being truly robust to optional stopping in the above frequentist sense, i.e. (7) will hold for all $P \in H_0$ and not just 'on average'. This is illustrated in Example 1 below and formalized in Section 5.5.

3 Discussion: Why Should One Care?

Nowadays, even more so than in the past, statistical tests are often performed in an on-line setting, in which data keeps coming in sequentially and one cannot tell in advance at what point the analysis will be stopped and a decision will be made — there may indeed be many such points. Prime examples include group sequential trials (Proschan et al., 2006) and A/B-testing, to which all internet users who visit the sites of the tech giants are subjected. In such on-line settings, it may or may not be a good idea to use Bayesian tests. But can and should they be used? Together with the companion paper (de Heide and Grünwald, 2018) (DHG from now on), the present paper sheds some light on this issue. Let us first highlight a central insight from DHG, which is about the case in which none of the results discussed in the present paper apply: in many practical situations,

many Bayesian statisticians use priors that are *themselves* dependent on parts of the data and/or the sampling plan and stopping time. Examples are Jeffreys prior with the multinomial model and the Gunel-Dickey default priors for 2x2 contingency tables advocated by Jamil et al. (2016). With such priors, final results evidently depend on the stopping rule employed, and even though such methods typically count as 'Bayesian', they do not satisfy τ-independence. The results then become noninterpretable under optional stopping (i.e. stopping using a rule that is not known at the time the prior is decided upon), and as argued by de Heide and Grünwald (2018), the notions of calibration and frequentist optional stopping even become undefined in such a case.

In such situations, one cannot rely on Bayesian methods to be valid under optional stopping in any sense at all; in the present paper we thus focus on the case with priors that are fixed in advance, and that themselves do not depend on the stopping rule or any other aspects of the design. For expository simplicity, we consider the question of whether Bayes factors with such priors are valid under optional stopping in two extreme settings: in the first setting, the goal of the analysis is purely *exploratory* — it should give us some insight in the data and/or suggest novel experiments to gather or novel models to analyze data with. In the second setting we consider the analysis as 'final' and the stakes are much higher — real decisions involving money, health and the like are involved — a typical example would be a Stage 2 clinical trial, which will decide whether a new medication will be put to market or not.

For the first, *exploratory* setting, exact error guarantees might neither be needed at all nor obtainable anyway, so the frequentist sense of handling optional stopping may not be that important. Yet, one would still like to use methods that satisfy some basic *sanity checks* for use under optional stopping. τ-independence is such a check: any method for which it does not hold is simply not suitable for use in a situation in which details of the stopping rule may be unknown. Also calibration can be viewed as such a sanity check: Rouder (2014) introduced it mainly to show that Bayesian posterior odds remain *meaningful* under optional stopping: they still satisfy some key property that they satisfy for fixed sample sizes.

For the second *high stakes* setting, mere sanity and interpretability checks are not enough: most researchers would want more stringent guarantees, for example on Type-I and/or Type-II error control. At the same time, most researchers would acknowledge that their priors are far from perfect, chosen to some extent for purposes of convenience rather than true belief.[2] Such researchers may thus want the desired Type-I error guarantees to hold for all $P \in H_0$, and not just in average over the prior as in (7). Similarly, in the high stakes setting the form of calibration (5) that can be guaranteed for the Bayes factor would be considered too weak, and one would hope for a stronger form of calibration as explained at the end of Section 2.2.

DHG show empirically that for some often-used models and priors, strong calibration can be severely violated under optional stopping. Similarly, it is possible to show that in general, Type-I error guarantees based on Bayes factors simply do not hold simultaneously for all $P \in H_0$ for such models and priors. Thus, one should be cautious using

[2]Even De Finetti and Savage, fathers of subjective Bayesianism, acknowledged this: see Section 5 of DHG.

Bayesian methods in the high stakes setting, despite adhortations such as the quote by Edwards et al. (1963) in the introduction (or similar quotes by e.g. Rouder et al., 2009): these existing papers invariably use τ-independence, calibration or Type-I error control with simple null hypotheses as a motivation to — essentially — use Bayes factor methods in any situation, including presumably high-stakes situations and situations with composite null hypotheses.[3]

Still, and this is equally important for practitioners, while frequentist error control and strong calibration are violated in general, in some important special cases they do hold, namely if the models H_0 and H_1 satisfy a group invariance. We proceed to give an informal illustration of this fact, deferring the mathematical details to Section 5.5.

Example 1. Consider the one-sample t-test as described by Rouder et al. (2009), going back to Jeffreys (1961). The test considers normally distributed data with unknown standard deviation. The test is meant to answer the question whether the data has mean $\mu = 0$ (the null hypothesis) or some other mean (the alternative hypothesis). Following (Rouder et al., 2009), a Cauchy prior density, denoted by $\pi_\delta(\delta)$, is placed on the effect size $\delta = \mu/\sigma$. The unknown standard deviation is a nuisance parameter and is equipped with the improper prior with density $\pi_\sigma(\sigma) = \frac{1}{\sigma}$ under both hypotheses. This is the so-called *right Haar prior* for the variance. This gives the following densities on n outcomes:

$$p_{0,\sigma}(x^n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\left(\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2\right) \quad [\, = p_{1,\sigma,0}(x^n)\,], \tag{8}$$

$$p_{1,\sigma,\delta}(x^n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\left(-\frac{n}{2}\left[\left(\frac{\overline{x}}{\sigma} - \delta\right)^2 + \left(\frac{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2}{\sigma^2}\right)\right]\right),$$

where $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$, so that the corresponding Bayesian marginal densities are given by

$$\bar{p}_0(x^n) = \int_0^\infty p_{0,\sigma}(x^n)\pi_\sigma(\sigma)\,\mathrm{d}\sigma,$$

$$\bar{p}_1(x^n) = \int_0^\infty \int_{-\infty}^\infty p_{1,\sigma,\delta}(x^n)\pi_\delta(\delta)\pi_\sigma(\sigma)\,\mathrm{d}\delta\,\mathrm{d}\sigma = \int_0^\infty p_{1,\sigma}(x^n)\pi_\sigma(\sigma)\,\mathrm{d}\sigma.$$

Our results in Section 5 imply that — under a slight, natural restriction on the stopping rules allowed — the Bayes factor $\bar{p}_1(x^n)/\bar{p}_0(x^n)$ is truly robust to optional stopping in the above frequentist sense. That is, (7) will hold for all $P \in H_0$, i.e. all $\sigma > 0$, and not just 'on average'. Thus, we can give Type I error guarantees irrespective of the true value of σ. Similarly, strong calibration in the sense of Section 2.2 holds for all $P \in \mathcal{H}_0$. The use of a Cauchy prior is not essential in this construction; the result will continue

[3]Since the authors of the present papers are inclined to think frequentist error guarantees are important, we disagree with such claims, as in fact a subset of researchers calling themselves Bayesians would as well. To witness, a large fraction of recent ISBA (Bayesian) meetings is about frequentist properties of Bayesian methods; also the well-known Bayesian authors (Good, 1991 and Edwards et al., 1963) focus on showing that Bayes factor methods achieve a *frequentist Type-I error* guarantee, albeit only for the simple H_0 case.

to hold for any proper prior on δ, including point priors that put all mass on a single value of δ.

As we show in Section 5, these results extend to a variety of settings, namely whenever H_0 and H_1 share a common so-called group invariance. In the t-test example, it is a scale invariance — effectively this means that for all δ, all σ, the distributions of

$$X_1, \ldots, X_n \text{ under } p_{1,\sigma,\delta}, \text{ and } \sigma X_1, \ldots, \sigma X_n \text{ under } p_{1,1,\delta}, \text{ coincide.} \tag{9}$$

For other models, one could have a translation invariance; for the full normal family, one has both translation and scale invariance; for yet other models, one might have a rotation invariance, and so on. Each such invariance is expressed as a *group* — a set equipped with an operation (the group *action*) that satisfies certain axioms. The group corresponding to scale invariance is the set of positive reals, and the group action is scalar multiplication or equivalently division; similarly, the group corresponding to translation invariance is the set of all reals, and the action is addition.

In the general case, one starts with a group G that satisfies certain further restrictions (detailed in Section 5), a model $\{p_{1,g,\theta} : g \in G, \theta \in \Theta\}$ where g represents the invariant parameter (vector) and the parameterization must be such that the analogue of (9) holds. In the example above $g = \sigma$ is the variance and θ is set to $\delta := \mu/\sigma$. One then singles out a special value of θ, say θ_0, one sets $H_0 := \{p_{1,g,\theta_0} : g \in G\}$; within H_1 one puts an arbitrary prior on θ. For every group invariance, there exists a corresponding *right Haar prior* on G; one equips both models with this prior on G. Theorem 8 and 9 imply that in all models constructed this way, we have strong calibration and Type-I error control uniformly for all $g \in G$. While this is hinted at in several papers (e.g. Bayarri et al., 2016; Dass and Berger, 2003) and the special case for the Bayesian t-test was implicitly proven in earlier work by Lai (1976), it seems to never have been proven formally in general before.

Our results thus imply that in some situations (group invariance) with composite null hypotheses, Type-I error control for all $P \in H_0$ under optional stopping is possible with Bayes factors. What about Type-II error control and composite null hypotheses that do *not* satisfy a group structure? This is partially addressed by the *safe testing* approach of Grünwald et al. (2019) (see also Howard et al., 2018 for a related approach). They show that for completely arbitrary H_0 and H_1, for any given prior π_1 on H_1, there exists a corresponding prior π_0 on H_0, the *reverse information projection prior*, so that, for all $P \in H_0$, one has Type-I error guarantees under frequentist *optional continuation*, a weakening of the idea of optional stopping. Further, if one wants to get control of Type-II error guarantees under optional stopping/continuation, one can do so by first choosing another special prior π_1^* on H_1 and picking the corresponding π_0^* on H_0. Essentially, like in 'default' or 'objective' Bayes approaches, one chooses special priors in lieu of a subjective choice; but the priors one ends up with are sometimes quite different from the standard default priors, and, unlike these, allow for frequentist error control under optional stopping.

4 The General Case

Let $(\Omega, \mathcal{F})$ be a measurable space. Fix some $m \geq 0$ and consider a sequence of functions $X_{m+1}, X_{m+2}, \ldots$ on Ω so that each X_n, $n > m$ takes values in some fixed set ('outcome space') $\mathcal{X}$ with associated σ-algebra Σ. When working with proper priors we invariably take $m = 0$ and then we define $X^n := (X_1, X_2, \ldots, X_n)$ and we let $\Sigma^{(n)}$ be the n-fold product algebra of Σ. When working with improper priors it turns out to be useful (more explanation further below) to take $m > 0$ and define an *initial sample* random variable $\langle X^{(m)} \rangle$ on Ω, taking values in some set $\langle \mathcal{X}^m \rangle \subseteq \mathcal{X}^m$ with associated σ-algebra $\langle \Sigma^{(m)} \rangle$. In that case we set, for $n \geq m$, $\langle \mathcal{X}^n \rangle = \{x^n = (x_1, \ldots, x_n) \in \mathcal{X}^n : x^m = (x_1, \ldots, x_m) \in \langle \mathcal{X}^m \rangle\}$, and $X^n := (\langle X^{(m)} \rangle, X_{m+1}, X_{m+2}, \ldots, X_n)$ and we let $\Sigma^{(n)}$ be $\langle \Sigma^{(m)} \rangle \times \prod_{j=m+1}^{n} \Sigma$. In either case, we let $\mathcal{F}_n$ be the σ-algebra (relative to Ω) generated by $(X^n, \Sigma^{(n)})$. Then $(\mathcal{F}_n)_{n=m,m+1,\ldots}$ is a filtration relative to $\mathcal{F}$ and if we equip $(\Omega, \mathcal{F})$ with a distribution P then $\langle X^{(m)} \rangle, X_{m+1}, X_{m+2}, \ldots$ becomes a random process adapted to $\mathcal{F}$. A *stopping time* is now generalized to be a function $\tau : \Omega \to \{m+1, m+2, \ldots\} \cup \{\infty\}$ such that for each $n > m$, the event $\{\tau = n\}$ is $\mathcal{F}_n$-measurable; note that we only consider stopping after m initial outcomes. Again, for a given stopping time τ and sequence of data $x^n = (x_1, \ldots, x_n)$, we say that x^n *is compatible with* τ if it satisfies $X^n = x^n \Rightarrow \tau = n$, i.e. $\{\omega \in \Omega \mid X^n(\omega) = x^n\} \subset \{\omega \in \Omega \mid \tau(\omega) = n\}$.

H_0 and H_1 are now sets of probability distributions on $(\Omega, \mathcal{F})$. Again one writes $H_j = \{P_{\theta|j} \mid \theta \in \Theta_j\}$ where now the parameter sets Θ_j (which, however, could itself be infinite-dimensional) are themselves equipped with suitable σ-algebras.

We will still represent both H_0 and H_1 by unique measures $\bar{P}_0$ and $\bar{P}_1$ respectively, which we now allow to be based on (1) with improper priors π_0 and π_1 that may be infinite measures. As a result $\bar{P}_0$ and $\bar{P}_1$ are positive real measures that may themselves be infinite. We also allow $\mathcal{X}$ to be a general (in particular uncountable) set. Both non-integrability and uncountability cause complications, but these can be overcome if suitable Radon-Nikodym derivatives exist. To ensure this, we will assume that for all $n \geq \max\{m, 1\}$, for all $k \in \{0, 1\}$ and $\theta \in \Theta_k$, $P_{\theta|k}^{(n)}$, $\bar{P}_0^{(n)}$ and $\bar{P}_1^{(n)}$ are all mutually absolutely continuous and that the measures $\bar{P}_1^{(n)}$ and $\bar{P}_0^{(n)}$ are σ-finite. Then there also exists a measure ρ on $(\Omega, \mathcal{F})$ such that, for all such n, $\bar{P}_1^{(n)}$, $\bar{P}_0^{(n)}$ and $\rho^{(n)}$ are all mutually absolutely continuous: we can simply take $\rho^{(n)} = \bar{P}_0^{(n)}$, but in practice, it is often possible and convenient to take ρ such that $\rho^{(n)}$ is the Lebesgue measure on $\mathbb{R}^n$, which is why we explicitly introduce ρ here.

The absolute continuity conditions guarantee that all required Radon-Nikodym derivatives exist. Finally, we assume that the posteriors $\pi_k(\Theta_k \mid x^m)$ (as defined in the standard manner in (12) below; when $m = 0$ these are just the priors) are proper probability measures (i.e. they integrate to 1) for all $x^m \in \langle \mathcal{X}^m \rangle$. This final requirement is the reason why we sometimes need to consider $m > 0$ and nonstandard sample spaces $\langle \mathcal{X}^n \rangle$ in the first place: in practice, one usually starts with the standard setting of a $(\Omega, \mathcal{F})$ where $m = 0$ and all X_i have the same status. In all practical situations with improper priors π_0 and/or π_1 that we know of, there is a smallest finite j and a set $\mathcal{X}^\circ \subset \mathcal{X}^j$ that has measure 0 under all probability distributions in $H_0 \cup H_1$, such that, restricted to the sample space $\mathcal{X}^j \setminus \mathcal{X}^\circ$, the measures $\bar{P}_1^{(j)}$ and $\bar{P}_0^{(j)}$ are σ-finite and

mutually absolutely continuous, and the posteriors $\pi_k(\Theta_k \mid x^j)$ are proper probability measures. One then sets m to equal this j, and sets $\langle \mathcal{X}^m \rangle := \mathcal{X}^m \setminus \mathcal{X}^\circ$, and the required properness will be guaranteed. Our initial sample $\langle X^{(m)} \rangle$ is a variation of what is called (for example, by Bayarri et al. (2012)) a *minimal sample*. Yet, the sample size of a standard minimal sample is itself a random quantity; by restricting $\mathcal{X}^m$ to $\langle \mathcal{X}^m \rangle$, we can take its sample size m to be constant rather than random, which will greatly simplify the treatment of optional stopping with group invariance; see Example 1 and 2 below.

We henceforth refer to the setting now defined (with m and initial space $\langle \mathcal{X}^m \rangle$ satisfying the requirements above) as the *general case*.

We need an analogue of (4) for this general case. If $\bar{P}_0$ and $\bar{P}_1$ are probability measures, then there is still a standard definition of conditional probability distributions $P(H \mid \mathcal{A})$ in terms of conditional expectation for any given σ-algebra $\mathcal{A}$; based on this, we can derive the required analogue in two steps. First, we consider the case that $\tau \equiv n$ for some $n > m$. We know in advance that we observe X^n for a fixed n: the appropriate $\mathcal{A}$ is then $\mathcal{F}_n$, $\pi(H \mid \mathcal{A})(\omega)$ is determined by $X^n(\omega)$ hence can be written as $\pi(H \mid X^n)$, and a straightforward calculation gives that

$$\frac{\pi(H_1 \mid X^n = x^n)}{\pi(H_0 \mid X^n = x^n)} = \left(\left(\frac{d\bar{P}_1^{(n)}/d\rho^{(n)}}{d\bar{P}_0^{(n)}/d\rho^{(n)}} \right)(x^n) \right) \cdot \frac{\pi(H_1)}{\pi(H_0)}, \tag{10}$$

where $(d\bar{P}_1^{(n)}/d\rho^{(n)})$ and $(d\bar{P}_0^{(n)}/d\rho^{(n)})$ are versions of the Radon-Nikodym derivatives defined relative to $\rho^{(n)}$. The second step is now to follow exactly the same steps as in the derivation of (4), replacing $\beta(X^n)$ by (10) wherever appropriate (we omit the details). This yields, for any n such that $\rho(\tau = n) > 0$, and for $\rho^{(n)}$-almost every x^n that is compatible with τ,

$$\overbrace{\frac{\pi(H_1 \mid x^n)}{\pi(H_0 \mid x^n)}}^{\gamma_n} = \frac{\pi(H_1 \mid X^n = x^n, \tau = n)}{\pi(H_0 \mid X^n = x^n, \tau = n)} = \overbrace{\left(\left(\frac{d\bar{P}_1^{(n)}/d\rho^{(n)}}{d\bar{P}_0^{(n)}/d\rho^{(n)}} \right)(x^n) \right)}^{\beta_n} \cdot \frac{\pi(H_1)}{\pi(H_0)}, \tag{11}$$

where here, as below, for $n \geq m$, we abbreviate $\pi(H_k \mid X^n = x^n)$ to $\pi(H_k \mid x^n)$.

The above expression for the posterior is valid if $\bar{P}_0$ and $\bar{P}_1$ are probability measures; we will simply take it as the *definition* of the Bayes factor for the general case. Again this coincides with standard usage for the improper prior case. In particular, let us define the conditional posteriors and Bayes factors given $\langle X^{(m)} \rangle = x^m$ in the standard manner, by the formal application of Bayes' rule, for $k = 0, 1$ and measurable $\Theta_k' \subset \Theta_k$ and $\mathcal{F}$-measurable A,

$$\pi_k(\Theta_k' \mid x^m) := \frac{\int_{\Theta_k'} \frac{dP_{\theta|k}^{(m)}}{d\rho^{(m)}}(x^m) d\pi_k(\theta)}{\int_{\Theta_k} \frac{dP_{\theta|k}^{(m)}}{d\rho^{(m)}}(x^m) d\pi_k(\theta)}, \tag{12}$$

$$\bar{P}_k(A \mid x^m) := \bar{P}_k(A \mid \langle X^{(m)} \rangle = x^m) := \int_{\Theta_k} P_{\theta|k}(A \mid \langle X^{(m)} \rangle = x^m) \, \mathrm{d}\pi_k(\theta \mid x^m), \tag{13}$$

where $P_{\theta|k}(A \mid \langle X^{(m)} \rangle = x^m)$ is defined as the value that (a version of) the conditional probability $P_{\theta|k}(A \mid \mathcal{F}_m)$ takes when $\langle X^{(m)} \rangle = x^m$, and is thus defined up to a set of $\rho^{(m)}$-measure 0.

With these definitions, it is straightforward to derive the following *coherence property*, which automatically holds if the priors are proper, and which in combination with (11) expresses that first updating on x^m and then on $x_{m+1}, \ldots, x_n$ (multiplying posterior odds given x^m with the Bayes factor for n outcomes given $X^m = x^m$, which we denote by $\beta_{n|m}$) has the same result as updating based on the full $x_1, \ldots, x_n$ at once (i.e. multiplying the prior odds with the unconditional Bayes factor β_n for n outcomes):

$$\frac{\pi(H_1 \mid X^n = x^n, \tau = n)}{\pi(H_0 \mid X^n = x^n, \tau = n)} = \overbrace{\left(\frac{d\bar{P}_1^{(n)}(\cdot \mid x^m)}{d\bar{P}_0^{(n)}(\cdot \mid x^m)}(x^n) \right)}^{\beta_{n|m}} \cdot \frac{\pi(H_1 \mid x^m)}{\pi(H_0 \mid x^m)}. \tag{14}$$

4.1 τ-Independence, General Case

The general version of the claim that the posterior odds do not depend on the specific stopping rule that was used is now immediate, since the expression (11) for the Bayes factor does not depend on the stopping time τ.

4.2 Calibration, General Case

We will now show that the calibration hypothesis continues to hold in our general setting. From here onward, we make the further reasonable assumption that for every $x^m \in \langle \mathcal{X}^m \rangle$, $\bar{P}_0(\tau = \infty \mid x^m) = \bar{P}_1(\tau = \infty \mid x^m) = 0$ (the stopping time is almost surely finite), and we define $\mathcal{T}_\tau := \{n \in \mathbb{N}_{>m} \mid \bar{P}_0(\tau = n) > 0\}$.

To prepare further, let $\{B_j \mid j \in \mathcal{T}_\tau\}$ be any collection of positive random variables such that for each $j \in \mathcal{T}_\tau$, B_j is $\mathcal{F}_j$-measurable. We can define the *stopped* random variable B_τ as

$$B_\tau := \sum_{j=0}^{\infty} \mathbb{1}_{\{\tau = j\}} B_j = \sum_{j=m+1}^{\infty} \mathbb{1}_{\{\tau = j\}} B_j, \tag{15}$$

where we note that, under this definition, B_τ is well-defined even if $\mathbf{E}_{\bar{P}_0}[\tau] = \infty$.

We can define the induced measures on the positive real line under the null and alternative hypothesis for any probability measure P on $(\Omega, \mathcal{F})$:

$$P^{[B_\tau]} : \mathcal{B}(\mathbb{R}_{>0}) \to [0, 1] : A \mapsto P\left(B_\tau^{-1}(A)\right), \tag{16}$$

where $\mathcal{B}(\mathbb{R}_{>0})$ denotes the Borel σ-algebra of $\mathbb{R}_{>0}$. Note that, when we refer to $P^{[B_n]}$, this is identical to $P^{[B_\tau]}$ for the stopping time τ which on all of Ω stops at n. The following lemma is crucial for passing from fixed-sample size to stopping-rule based results.

Lemma 1. *Let $\mathcal{T}_\tau$ and $\{B_n \mid n \in \mathcal{T}_\tau\}$ be as above. Consider two probability measures P_0 and P_1 on $(\Omega, \mathcal{F})$. Suppose that for all $n \in \mathcal{T}_\tau$, the following* fixed-sample size calibration property *holds:*

$$\text{for some fixed } c > 0, \;\; {P_0}^{[B_n]}\text{-almost all } b : \frac{P_1(\tau = n)}{P_0(\tau = n)} \cdot \frac{\mathrm{d}{P_1}^{[B_n]}(\cdot \mid \tau = n)}{\mathrm{d}{P_0}^{[B_n]}(\cdot \mid \tau = n)}(b) = c \cdot b. \tag{17}$$

Then we have

$$\text{for } {P_0}^{[B_\tau]}\text{-almost all } b : \frac{\mathrm{d}{P_1}^{[B_\tau]}}{\mathrm{d}{P_0}^{[B_\tau]}}(b) = c \cdot b. \tag{18}$$

The proof is in Section B in the supplementary material (Hendriksen et al., 2020).

In this subsection we apply this lemma to the measures $\bar{P}_k(\cdot \mid x^m)$ for arbitrary fixed $x^m \in \langle \mathcal{X}^m \rangle$, with their induced measures $\bar{P}_0^{[\gamma_\tau]}(\cdot \mid x^m), \bar{P}_1^{[\gamma_\tau]}(\cdot \mid x^m)$ for the *stopped posterior odds* γ_τ. Formally, the posterior odds γ_n as defined in (11) constitute a random variable for each n, and, under our mutual absolute continuity assumption for $\bar{P}_0$ and $\bar{P}_1$, γ_n can be directly written as $\frac{\mathrm{d}\bar{P}_1^{(n)}}{\mathrm{d}\bar{P}_0^{(n)}} \cdot \pi(H_1)/\pi(H_0)$. Since, by definition, the measures $\bar{P}_k(\cdot \mid x^m)$ are probability measures, the Radon-Nikodym derivatives in (17) and (18) are well-defined.

Lemma 2. *We have for all $x^m \in \langle \mathcal{X}^m \rangle$, all $n > m$:*

$$\text{for } \bar{P}_0^{[\gamma_n]}(\cdot \mid x^m)\text{-almost all } b : \frac{\bar{P}_1^{[\gamma_n]}(\tau = n \mid x^m)}{\bar{P}_0^{[\gamma_n]}(\tau = n \mid x^m)} \cdot \frac{\mathrm{d}\bar{P}_1^{[\gamma_n]}(\cdot \mid x^m)}{\mathrm{d}\bar{P}_0^{[\gamma_n]}(\cdot \mid x^m)}(b) = \frac{\pi(H_0 \mid x^m)}{\pi(H_1 \mid x^m)} \cdot b. \tag{19}$$

Combining the two lemmas now immediately gives (20) below, and combining further with (14) and (11) gives (21):

Corollary 3. *In the setting considered above, we have for all $x^m \in \langle \mathcal{X}^m \rangle$:*

$$\text{for } \bar{P}_0^{[\gamma_\tau]}(\cdot \mid x^m)\text{-almost all } b : \frac{\pi(H_1 \mid x^m)}{\pi(H_0 \mid x^m)} \cdot \frac{\mathrm{d}\bar{P}_1^{[\gamma_\tau]}(\cdot \mid x^m)}{\mathrm{d}\bar{P}_0^{[\gamma_\tau]}(\cdot \mid x^m)}(b) = b, \tag{20}$$

and also

$$\text{for } \bar{P}_0^{[\gamma_\tau]}(\cdot \mid x^m)\text{-almost all } b : \frac{\pi(H_1)}{\pi(H_0)} \cdot \frac{\mathrm{d}\bar{P}_1^{[\gamma_\tau]}}{\mathrm{d}\bar{P}_0^{[\gamma_\tau]}}(b) = b. \tag{21}$$

In words, the posterior odds remain calibrated under any stopping rule τ which stops almost surely at times $m < \tau < \infty$.

For discrete and strictly positive measures with prior odds $\pi(H_1)/\pi(H_0) = 1$, we always have $m = 0$, and (20) is equivalent to (5). Note that $\bar{P}_0^{[\gamma_\tau]}(\cdot \mid x^m)$-almost everywhere in (20) is equivalent to $\bar{P}_1^{[\gamma_\tau]}(\cdot \mid x^m)$-almost everywhere because the two measures are assumed to be mutually absolutely continuous.

4.3 (Semi-)Frequentist Optional Stopping

In this section we consider our general setting as in the beginning of Section 4.2, i.e. with the added assumption that the stopping time is a.s. finite, and with $\mathcal{T}_\tau := \{j \in \mathbb{N}_{>m} \mid \bar{P}_0(\tau = j) > 0\}$.

Consider any initial sample $x^m \in \langle \mathcal{X}^m \rangle$ and let $\bar{P}_0 \mid x^m$ and $\bar{P}_1 \mid x^m$ be the conditional Bayes marginal distributions as defined in (13). We first note that, by Markov's inequality, for any nonnegative random variable Z on Ω with, for all $x^m \in \langle \mathcal{X}^m \rangle$, $\mathbf{E}_{\bar{P}_0 | x^m}[Z] \leq 1$, we must have, for $0 \leq \alpha \leq 1$, $\bar{P}_0(Z^{-1} \leq \alpha \mid x^m) \leq \mathbf{E}_{\bar{P}_0|x^m}[Z]/\alpha^{-1} \leq \alpha$.

Proposition 4. *Let τ be any stopping rule satisfying our requirements. Let $\beta_{\tau|m}$ be the stopped Bayes factor given x^m, i.e., in accordance with (15), $\beta_{\tau|m} = \sum_{j=m+1}^{\infty} \mathbb{1}_{\{\tau=j\}} \beta_{j|m}$ with $\beta_{j|m}$ as given by (14). Then $\beta_{\tau|m}$ satisfies, for all $x^m \in \langle \mathcal{X}^m \rangle$, $\mathbf{E}_{\bar{P}_0|x^m}[\beta_{\tau|m}] \leq 1$, so that, by the reasoning above, $\bar{P}_0(\frac{1}{\beta_{\tau|m}} \leq \alpha \mid x^m) \leq \alpha$.*

Proof. We have

$$\mathbf{E}_{\bar{P}_0|x^m}[\gamma_\tau] = \int b \bar{P}_0^{[\gamma_\tau]}(\,\mathrm{d}b \mid x^m) = \int \frac{\mathrm{d}\bar{P}_1^{[\gamma_\tau]}(b \mid x^m)}{\mathrm{d}\bar{P}_0^{[\gamma_\tau]}(b \mid x^m)} \cdot \frac{\pi(H_1 \mid x^m)}{\pi(H_0 \mid x^m)} \, \bar{P}_0^{[\gamma_\tau]}(\,\mathrm{d}b \mid x^m) = \frac{\pi(H_1 \mid x^m)}{\pi(H_0 \mid x^m)},$$

where the first equality follows by definition of expectation, the second follows from Corollary 3, and the third follows from the fact that the integral equals 1.

But now note that

$$\beta_{\tau|m} = \sum_{j=m+1}^{\infty} \mathbb{1}_{\{\tau=j\}} \beta_{j|m} = \sum_{j=m+1}^{\infty} \mathbb{1}_{\{\tau=j\}} \gamma_j \cdot \frac{\pi(H_0 \mid x^m)}{\pi(H_1 \mid x^m)} = \gamma_\tau \cdot \frac{\pi(H_0 \mid x^m)}{\pi(H_1 \mid x^m)},$$

where the second equality follows from (14) together with the first equality in (11). Combining the two equations we get:

$$\mathbf{E}_{\bar{P}_0|x^m}\left[\beta_{\tau|m}\right] = \mathbf{E}_{\bar{P}_0|x^m}\left[\gamma_\tau \cdot \frac{\pi(H_0 \mid x^m)}{\pi(H_1 \mid x^m)}\right] = 1. \qquad \square$$

The desired result now follows by plugging in a particular stopping rule: let $S : \bigcup_{i=m+1}^{\infty} \mathcal{X}^i \to \{0, 1\}$ be the frequentist sequential test defined by setting, for all $n > m$, $x^n \in \langle \mathcal{X}^n \rangle$: $S(x^n) = 1$ if and only if $\beta_{n|m} \geq 1/\alpha$.

Corollary 5. *Let $t^* \in \{m+1, m+2, \ldots\} \cup \{\infty\}$ be the smallest $t^* > m$ for which $\beta_{t|m}^{-1} \leq \alpha$. Then for arbitrarily large T, when applied to the stopping rule $\tau := \min\{T, t^*\}$, we find that*

$$\bar{P}_0(\exists n, m < n \leq T : S(X^n) = 1 \mid x^m) = \bar{P}_0(\exists n, m < n \leq T : \beta_{n|m}^{-1} \leq \alpha \mid x^m) \leq \alpha.$$

The corollary implies that the test S is robust under optional stopping in the frequentist sense relative to H_0 (Definition 1). Note that, just as in the simple case, the setting is really just 'semi-frequentist' whenever H_0 is not a singleton.

5 Optional Stopping with Group Invariance

Whenever the null hypothesis is composite, the previous results only hold under the marginal distribution $\bar{P}_0$ or, in the case of improper priors, under $\bar{P}_0(\cdot \mid X^m = x^m)$. When a group structure can be imposed on the outcome space and (a subset of the) parameters that is joint to H_0 and H_1, stronger results can be derived for calibration and frequentist optional stopping. Invariably, such parameters function as *nuisance parameters* and our results are obtained if we equip them with the so-called *right Haar prior* which is usually improper. Below we show how we then obtain results that simultaneously hold for *all* values of the nuisance parameters. Such cases include many standard testing scenarios such as the (Bayesian variations of the) t-test, as illustrated in the examples below. Note though that our results do not apply to settings with improper priors for which no group structure exists. For example, if $P_{\theta|0}$ expresses that $X_1, X_2, \ldots$ are i.i.d. Poisson(θ), then from an objective Bayes or MDL point of view it makes sense to adopt Jeffreys' prior for the Poisson model; this prior is improper, allows initial sample size $m = 1$, but does not allow for a group structure. For such a prior we can only use the marginal results Corollary 3 and Corollary 5. Group theoretic preliminaries, such as definitions of a (topological) group, the right Haar measure, etcetera can be found in Section B of the supplementary material (Hendriksen et al., 2020).

5.1 Background for Fixed Sample Sizes

Here we prepare for our results by providing some general background on invariant priors for Bayes factors with fixed sample size n on models with nuisance parameters that admit a group structure, introducing the right Haar measure, the corresponding Bayes marginals, and (maximal) invariants. We use these results in Section 5.2 to derive Lemma 7, which gives us a strong version of calibration for fixed n. The setting is extended to variable stopping times in Section 5.3, and then Lemma 7 is used in this extended setting to obtain our strong optional stopping results in Section 5.4 and 5.5.

For now, we assume a sample space $\langle \mathcal{X}^n \rangle$ that is locally compact and Hausdorff, and that is a subset of some product space $\mathcal{X}^n$ where $\mathcal{X}$ is itself locally compact and Hausdorff. This requirement is met, for example, when $\mathcal{X} = \mathbb{R}$ and $\langle \mathcal{X}^n \rangle = \mathcal{X}^n$. In practice, the space $\langle \mathcal{X}^n \rangle$ is invariably a subset of $\mathcal{X}^n$ where some null-set is removed for technical reasons that will become apparent below. We associate $\langle \mathcal{X}^n \rangle$ with its Borel σ-algebra which we denote as $\mathcal{F}_n$. Observations are denoted by the random vector $X^n = (X_1, \ldots, X_n) \in \langle \mathcal{X}^n \rangle$. We thus consider outcomes of fixed sample size, denoting these as $x^n \in \langle \mathcal{X}^n \rangle$, returning to the case with stopping times in Section 5.4 and 5.5.

From now on we let G be a locally compact group G that acts topologically and properly[4] on the right of $\langle \mathcal{X}^n \rangle$. As hinted to before, this proper action requirement sometimes forces the removal from $\mathcal{X}^n$ of some trivial set with measure zero under all hypotheses involved. This is demonstrated at the end of Example 1 below.

[4] A group acts properly on a set Y if the mapping $\psi : Y \times G \mapsto Y \times Y$ defined by $\psi(y, g) = (y \cdot g, y)$ is a proper mapping, i.e. the inverse image of ψ of each compact set in $Y \times Y$ is a compact set in $Y \times G$. (Eaton (1989), Definition 5.1).

Let $P_{0,e}$ and $P_{1,e}$ (notation to become clear below) be two arbitrary probability distributions on $\langle \mathcal{X}^n \rangle$ that are mutually absolutely continuous. We will now generate hypothesis classes H_0 and H_1, both sets of distributions on $\langle \mathcal{X}^n \rangle$ with parameter space G, starting from $P_{0,e}$ and $P_{1,e}$, where $e \in G$ is the group identity element. The group action of G on $\langle \mathcal{X}^n \rangle$ induces a group action on these measures defined by

$$P_{k,g}(A) \coloneqq (P_{k,e} \cdot g)(A) \coloneqq P_{k,e}(A \cdot g^{-1}) = \int \mathbb{1}_{\{A\}}(x \cdot g)\, P_{k,e}(\,\mathrm{d}x) \tag{22}$$

for any set $A \in \mathcal{F}_n$, $k = 0, 1$. When applied to $A = \langle \mathcal{X}^n \rangle$, we get $P_{k,g}(A) = 1$, for all $g \in G$, whence we have created two sets of probability measures parameterized by g, i.e.,

$$H_0 \coloneqq \{P_{0,g} \mid g \in G\} \quad ; \quad H_1 \coloneqq \{P_{1,g} \mid g \in G\}. \tag{23}$$

In this context, $g \in G$, can typically be viewed as nuisance parameter, i.e. a parameter that is not directly of interest, but needs to be accounted for in the analysis. This is illustrated in Example 1 and Example 2 below. The examples also illustrate how to extend this setting to cases where there are more parameters than just $g \in G$ in either H_0 or H_1. We extend the whole setup to our general setting with non-fixed n in Section 5.4.

We use the right Haar measure for G as a prior to define the Bayes marginals:

$$\bar{P}_k(A) = \int_G \int_{\langle \mathcal{X}^n \rangle} \mathbb{1}_{\{A\}}\, \mathrm{d}P_{k,g}\, \nu(\,\mathrm{d}g) \tag{24}$$

for $k = 0, 1$ and $A \in \mathcal{F}_n$. Typically, the right Haar measure is improper so that the Bayes marginals $\bar{P}_k$ are not integrable. Yet, in all cases of interest, they are (a) still σ-finite, and, (b), $\bar{P}_0$, $\bar{P}_1$ and all distributions $P_{k,g}$ with $k = 0, 1$ and $g \in G$ are mutually absolutely continuous; we will henceforth assume that (a) and (b) are the case.

Example 1 (continued)**.** Consider the t-test of Example 1. For consistency with the earlier Example 1, we abbreviate for general measures P on $\langle \mathcal{X}^n \rangle$, $(\,\mathrm{d}P/\,\mathrm{d}\lambda)$ (the density of distribution P relative to Lebesgue measure on $\mathbb{R}^n$) to p. Normally, the one-sample t-test is viewed as a test between $H_0 = \{P_{0,\sigma} \mid \sigma \in \mathbb{R}_{>0}\}$ and $H_1' = \{P_{1,\sigma,\delta} \mid \sigma \in \mathbb{R}_{>0}, \delta \in \mathbb{R}\}$, but we can obviously also view it as test between H_0 and $H_1 = \{P_{1,\sigma}\}$ by integrating out the parameter δ to obtain

$$p_{1,\sigma}(x^n) = \int p_{1,\sigma,\delta}(x^n)\pi_\delta(\delta)\,\mathrm{d}\delta. \tag{25}$$

The nuisance parameter σ can be identified with the group of scale transformations $G = \{c \mid c \in \mathbb{R}_{>0}\}$. We thus let the sample space be $\langle \mathcal{X}^n \rangle = \mathbb{R}^n \setminus \{0\}^n$, i.e., we remove the measure-zero set $\{0\}^n$, such that the group action is proper on the sample space. The group action is defined by $x^n \cdot c = c\, x^n$ for $x^n \in \langle \mathcal{X}^n \rangle, c \in G$. Take $e = 1$ and let, for $k = 0, 1$, $P_{k,e}$ be the distribution with density $p_{k,1}$ as defined in (8) and (25). The measures $P_{0,g}$ and $P_{1,g}$ defined by (22) then turn out to have the densities $p_{0,\sigma}$ and $p_{1,\sigma}$ as defined above, with σ replaced by g. Thus, H_0 and H_1 as defined by (8) and (25) are indeed in the form (23) needed to state our results.

In most standard invariant settings, H_0 and H_1 share the same vector of nuisance parameters, and one can reduce H_0 and H_1 to (23) in the same way as above, by integrating out all other parameters; in the example above, the only non-nuisance parameter was δ. The scenario of Example 1 can be generalized to a surprisingly wide variety of statistical models. In practice we often start with a model $H_1 = \{P_{1,\gamma,\delta} : \gamma \in \Gamma, \theta \in \Theta\}$ that implicitly already contains a group structure, and we single out a special subset $\{P_1, \gamma, \theta_0 : \gamma \in \Gamma\}$; this is what we informally described in Example 1. More generally, we can start with potentially large (or even nonparametric) hypotheses

$$H'_k = \{P_{\theta'|k} : \theta' \in \Theta'_k\} \tag{26}$$

which at first are not related to any group invariance, but which we want to equip with an additional nuisance parameter determined by a group G acting on the data. We can turn this into an instance of the present setting by first choosing, for $k = 0, 1$, a proper prior density π_k on Θ'_k, and defining $P_{k,e}$ to equal the corresponding Bayes marginal, i.e.

$$P_{k,e}(A) := \int P_{\theta'|k}(A)\ \mathrm{d}\pi_k(\theta'). \tag{27}$$

We can then generate $H_k = \{P_{k,g} \mid g \in G\}$ as in (22) and (23). In the example above, H'_1 would be the set of all Gaussians with a single fixed variance σ_0^2 and $\Theta'_1 = \mathbb{R}$ would be the set of all effect sizes δ, and the group G would be scale transformation; but there are many other possibilities. To give but a few examples, Dass and Berger (2003) consider testing the Weibull vs. the log-normal model, the exponential vs. the log-normal, correlations in multivariate Gaussians, and Berger et al. (1998b) consider location-scale families and linear models where H_0 and H_1 differ in their error distribution; another example is when the nuisance parameters comprise an l-dimensional sphere; the right Haar prior is then a uniform probability distribution on this sphere. Importantly, the group G acting on the data induces groups G_k, $k = 0, 1$, acting on the parameter spaces, which depend on the parameterization. In our example, the G_k were equal to G, but, for example, if H_0 is Weibull and H_1 is log-normal, both given in their standard parameterizations, we get $G_0 = \{g_{0,b,c} \mid g_{0,b,c}(\beta, \gamma) = (b\beta^c, \gamma/c), b > 0, c > 0\}$ and $G_1 = \{g_{1,b,c} \mid g_{1,b,c}(\mu, \sigma) = (c\mu + \log(b), c\sigma), b > 0, c > 0\}$. Several more examples are given by Dass (1998).

On the other hand, clearly not all hypothesis sets can be generated using the above approach. For instance, the hypothesis $H'_1 = \{P_{\mu,\sigma} \mid \mu = 1, \sigma > 0\}$ with $P_{\mu,\sigma}$ a Gaussian measure with mean μ and standard deviation σ cannot be represented as in (23). This is due to the fact that for $\sigma, \sigma' > 0, \sigma \neq \sigma'$, no element $g \in \mathbb{R}_{>0}$ exists such that for any measurable set $A \subseteq \langle \mathcal{X}^n \rangle$ the equality $P_{1,\sigma}(A) = P_{1,\sigma'}(A \cdot g^{-1})$ holds. This prevents an equivalent construction of H'_1 in the form of (23).

We now turn to the main ingredient that will be needed to obtain results on optional stopping: the quotient σ-algebra.

Definition 2 (Eaton, 1989, Chapter 2)**.** A group G acting on the right of a set Y induces an equivalence relation: $y_1 \sim y_2$ if and only if there exists $g \in G$ such that $y_1 = y_2 \cdot g$. This equivalence relation partitions the space in *orbits*: $O_y = \{y \cdot g \mid g \in G\}$, the collection of which is called the *quotient space* Y/G. There exists a map, the *natural projection*,

from Y to the quotient space which is defined by $\varphi_Y : Y \to Y/G : y \mapsto \{y \cdot g \mid g \in G\}$, and which we use to define the *quotient σ-algebra*

$$\mathcal{G}_n = \{\varphi^{-1}_{\langle \mathcal{X}^n \rangle}(\varphi_{\langle \mathcal{X}^n \rangle}(A)) \mid A \in \mathcal{F}_n\}. \tag{28}$$

Definition 3 (Eaton, 1989, Chapter 2)**.** A random element U_n on $\langle \mathcal{X}^n \rangle$ is *invariant* if for all $g \in G$, $x^n \in \langle \mathcal{X}^n \rangle$, $U_n(x^n) = U_n(x^n \cdot g)$. The random element U_n is *maximal invariant* if U_n is invariant and for all $y^n \in \langle \mathcal{X}^n \rangle$, $U_n(x^n) = U_n(y^n)$ implies $x^n = y^n \cdot g$ for some $g \in G$.

Thus, U_n is maximal invariant if and only if U_n is constant on each orbit, and takes different values on different orbits; $\varphi_{\langle \mathcal{X}^n \rangle}$ is thus an example of a maximal invariant. Note that any maximal invariant is $\mathcal{G}_n$-measurable. The importance of this quotient σ-algebra $\mathcal{G}_n$ is the following evident fact:

Proposition 6. *For fixed $k \in \{0, 1\}$, every invariant U_n has the same distribution under all $P_{k,g}, g \in G$.*

Chapter 2 of Eaton (1989) provides several methods and examples how to construct a concrete maximal invariant, including the first two given below. Since β_n is invariant under the group action of G (see below), β_n is an example of an invariant, although not necessarily of a maximal invariant.

Example 1 (continued)**.** Consider the setting of the one-sample t-test as described above in Example 1. A maximal invariant for $x^n \in \langle \mathcal{X}^n \rangle$ is $U_n(x^n) = (x_1/|x_1|, x_2/|x_1|, \ldots, x_n/|x_1|)$.

Example 2. A second example, with a group invariance structure on two parameters, is the setting of the two-sample t-test with the right Haar prior (which coincides here with Jeffreys' prior) $\pi(\mu, \sigma) = 1/\sigma$ (see Rouder et al., 2009 for details): the group is $G = \{(a, b) \mid a > 0, b \in \mathbb{R}\}$. Let the sample space be $\langle \mathcal{X}^n \rangle = \mathbb{R}^n \setminus \text{span}(e_n)$, where e_n denotes a vector of ones of length n (this is to exclude the measure-zero line for which the $s(x^n)$ is zero), and define the group action by $x^n \cdot (a, b) = ax^n + be_n$ for $x^n \in \langle \mathcal{X}^n \rangle$. Then (Eaton, 1989, Example 2.15) a maximal invariant for $x^n \in \langle \mathcal{X}^n \rangle$ is $U_n(x^n) = (x^n - \overline{x}e_n)/s(x^n)$, where $\overline{x}$ is the sample mean and $s(x^n) = \left(\sum_{i=1}^n (x_i - \overline{x})^2\right)^{1/2}$.

However, we can also construct a maximal invariant similar to the one in Example 1, which gives a special status to an initial sample:

$$U_n\left(X^n\right) = \left(\frac{X_2 - X_1}{|X_2 - X_1|}, \frac{X_3 - X_1}{|X_2 - X_1|}, \ldots, \frac{X_n - X_1}{|X_2 - X_1|}\right), \quad n \geq 2.$$

5.2 Relatively Invariant Measures and Calibration for Fixed n

Let U_n be a maximal invariant, taking values in the measurable space $(\mathcal{U}_n, \mathcal{G}_n)$. Although we have given more concrete examples above, it follows from the results of Andersson (1982) that, in case we do not know how to construct a U_n, we can always take $U_n =$

$\varphi_{\langle \mathcal{X}^n \rangle}$, the natural projection. Since we assume mutual absolute continuity, the Radon-Nikodym derivative $\frac{\mathrm{d}P_{1,g}^{[U_n]}}{\mathrm{d}P_{0,g}^{[U_n]}}$ must exist and we can apply the following theorem (note it is here that the use of *right* Haar measure is crucial; a different result holds for the left Haar measure):[5]

Theorem (Berger et al., 1998a, Theorem 2.1)**.** *Under our previous definitions of and assumptions on G, $P_{k,g}$, $\bar{P}_k$ let $\beta(x^n) \coloneqq \bar{P}_1(x^n)/\bar{P}_0(x^n)$ be the Bayes factor based on x^n. Let U_n be a maximal invariant as above, with (adopting the notation of (16)) marginal measures $P_{k,g}^{[U_n]}$, for $k = 0, 1$ and $g \in G$. There exists a version of the Radon-Nikodym derivative such that we have for all $g \in G$, all $x^n \in \langle \mathcal{X}^n \rangle$,*

$$\frac{\mathrm{d}P_{1,g}^{[U_n]}}{\mathrm{d}P_{0,g}^{[U_n]}}(U_n(x^n)) = \beta(x^n). \tag{29}$$

As a first consequence of the theorem above, we note (as did Berger et al., 1998a) that the Bayes factor $\beta_n \coloneqq \beta(X^N)$ is $\mathcal{G}_n$-measurable (it is constant on orbits), and thus *it has the same distribution under $P_{0,g}$ and $P_{1,g}$ for all $g \in G$*. The theorem also implies the following crucial lemma:

Lemma 7 (Strong Calibration for Fixed n)**.** *Under the assumptions of the theorem above, let U_n be a maximal invariant and let V_n be a $\mathcal{G}_n$-measurable binary random variable with $P_{0,g}(V_n = 1) > 0$, $P_{1,g}(V_n = 1) > 0$. Adopting the notation of (16), we can choose the Radon-Nikodym derivative $\mathrm{d}P_{1,g}^{[\beta_n]}(\cdot \mid V_n = 1)/\mathrm{d}P_{0,g}^{[\beta_n]}(\cdot \mid V_n = 1)$ so that we have, for all $x^n \in \langle \mathcal{X}^n \rangle$:*

$$\frac{P_{1,g}(V_n = 1)}{P_{0,g}(V_n = 1)} \cdot \frac{\mathrm{d}P_{1,g}^{[\beta_n]}(\cdot \mid V_n = 1)}{\mathrm{d}P_{0,g}^{[\beta_n]}(\cdot \mid V_n = 1)}(\beta_n(x^n)) = \beta_n(x^n), \tag{30}$$

where for the special case with $P_{k,g}(V_n = 1) = 1$, we get $\frac{\mathrm{d}P_{1,g}^{[\beta_n]}}{\mathrm{d}P_{0,g}^{[\beta_n]}}(\beta_n(x^n)) = \beta_n(x^n)$.

5.3 Extending to Our General Setting with Non-Fixed Sample Sizes

We start with the same setting as above: a group G on sample space $\langle \mathcal{X}^n \rangle \subset \mathcal{X}^n$ that acts topologically and properly on the right of $\langle \mathcal{X}^n \rangle$; two distributions $P_{0,e}$ and $P_{1,e}$ on $(\langle \mathcal{X}^n \rangle, \mathcal{F}_n)$ that are used to generate H_0 and H_1, and Bayes marginal measures based on the right Haar measure $\bar{P}_0$ and $\bar{P}_1$, which are both σ-finite. We now denote H_k as $H_k^{(n)}$, $P_{k,e}$ as $P_{k,e}^{(n)}$ and $\bar{P}_k$ as $\bar{P}_k^{(n)}$, all $P \in H_0^{(n)} \cup H_1^{(n)}$ are mutually absolutely continuous.

We now extend this setting to our general random process setting as specified in the beginning of Section 4.2 by further assuming that, for the same group G, for some

[5] This theorem requires that there exists some relatively invariant measure μ on $\langle \mathcal{X}^n \rangle$ such that for $k = 0, 1, g \in G$, the $P_{k,g}$ all have a density relative to μ. Since the Bayes marginal $\bar{P}_0$ based on the right Haar prior is easily seen to be such a relatively invariant measure, the conditions for the theorem apply.

$m > 0$, the above setting is defined for each $n \geq m$. To connect the $H_k^{(n)}$ for all these n, we further assume that there exists a subset $\langle \mathcal{X}^m \rangle \subset \mathcal{X}^m$ that has measure 1 under $P_{k,e}^{(n)}$ (and hence under all $P_{g,e}^{(n)}$) such that for all $n \geq m$:

1. We can write $\langle \mathcal{X}^n \rangle = \{x^n \in \mathcal{X}^n : (x_1, \ldots, x_m) \in \langle \mathcal{X}^m \rangle\}$.
2. For all $x^n \in \langle \mathcal{X}^n \rangle$, the posterior $\nu \mid x^n$ based on the right Haar measure ν is proper.
3. The probability measures $P_{k,e}^{(n)}$ and $P_{k,e}^{(n+1)}$ satisfy Kolmogorov's compatibility condition for a random process.
4. The group action $\cdot$ on the measures $P_{k,e}^{(n)}$ and $P_{k,e}^{(n+1)}$ is compatible, i.e. for every $n > 0$, for every $A \in \mathcal{F}_n$, every $g \in G$, $k \in \{0, 1\}$, we have $P_{k,g}^{(n+1)}(A) = P_{k,g}^{(n)}(A)$.

Requirement 4. simply imposes the condition that the group action considered is the same for all $n \in \mathbb{N}$. As a consequence of 3. and 4., the probability measures $P_{k,g}^{(n)}$ and $P_{k,g}^{(n+1)}$ satisfy Kolmogorov's compatibility condition for all $g \in G$, $k \in \{0, 1\}$ which means that there exists a probability measure $P_{k,g}$ on $(\Omega, \mathcal{F})$ (under which $\langle X^{(m)} \rangle, X_{m+1}, X_{m+2}, \ldots$ is a random process), defined as in the beginning of Section 4, whose marginals for $n \geq m$ coincide with $P_{k,g}^{(n)}$, and there exist measures $\bar{P}_0$ and $\bar{P}_1$ on $(\Omega, \mathcal{F})$ whose marginals for $n \geq m$ coincide with $\bar{P}_0^{(n)}$ and $\bar{P}_1^{(n)}$. We have thus defined a set H_0 and H_1 of hypotheses on $(\Omega, \mathcal{F})$ and the corresponding Bayes marginals $\bar{P}_0$ and $\bar{P}_1$ and are back in our general setting. It is easily verified that the 1- and 2-sample Bayesian t-tests both satisfy all these assumptions: in Example 1, take $m = 1$ and $\langle \mathcal{X}^m \rangle = \mathbb{R} \setminus \{0\}$; in Example 1, take $m = 2$ and $\langle \mathcal{X}^m \rangle = \mathbb{R}^2 \setminus \{(a, a) : a \in \mathbb{R}\}$. The conditions can also be verified for the variety of examples considered by Berger et al. (1998a) and Bayarri et al. (2012). In fact, our initial sample $x^m \in \langle \mathcal{X}^m \rangle$ is a variation of what they call a *minimal sample*; by excluding 'singular' outcomes from $\mathcal{X}^m$ to ensure that the group acts properly on $\langle \mathcal{X}^m \rangle$, we can guarantee that the initial sample is of fixed size. The size of the minimal sample can be larger, on a set of measure 0 under all $P \in H_0 \cup H_1$, e.g. if, in Example 1, $X_1 = X_2$. We chose to ensure a fixed size m since it makes the extension to random processes considerably easier.

In Section 5.1, underneath Example 1 we already outlined how a composite alternative hypothesis can be reduced to a hypothesis with just a free nuisance parameter (or parameter vector) $g \in G$, by putting a proper prior on all other parameters and integrating them out. A similar construction for a single parameter alternative hypothesis in the form of (23) can be applied in the non-fixed sample size case.

5.4 Strong Calibration

Consider the setting, definitions and assumptions of the previous subsection, with the additional assumptions and definitions made in the beginning of Section 4.3, in particular the assumption of a.s. finite stopping time. For simplicity, from now on, we shall also assume equal prior odds, $\pi(H_0) = \pi(H_1) = 1/2$. We will now show a strong calibration theorem for the Bayes factors $\beta_n = (d\bar{P}_0^{(n)})/(d\bar{P}_1^{(n)})(X^n)$ defined in terms of the Bayes marginals $\bar{P}_0$ and $\bar{P}_1$ with the right Haar prior. Thus β_τ is defined as in (15) with β in the role of B.

Theorem 8 (Strong calibration under optional stopping)**.** *Let τ be a stopping time satisfying our requirements, such that additionally, for each $n > m$, the event $\{\tau = n\}$ is $\mathcal{G}_n$-measurable. Then, adopting the notation of (16), for all $g \in G$, for $P_{0,g}^{[\beta_\tau]}$-almost every $b > 0$, we have: $\frac{\mathrm{d}P_{1,g}^{[\beta_\tau]}}{\mathrm{d}P_{0,g}^{[\beta_\tau]}}(b) = b$. That means that the posterior odds remain calibrated under every stopping rule τ adapted to the quotient space filtration $\mathcal{G}_m, \mathcal{G}_{m+1}, \ldots$, under all $P_{0,g}$.*

Proof. Fix some $g \in G$. We simply first apply Lemma 7 with $V_n = \mathbb{1}_{\{\tau=n\}}$, which gives that the premise (17) of Lemma 1 holds with $c = 1$ and β_n in the role of B_n (it is here that we need that τ_n is $\mathcal{G}_n$-measurable, otherwise we could not apply Lemma 7 with the required definition of V_n). We can now use Lemma 1 with $P_{0,g}$ in the role of P_0 to reach the desired conclusion for the chosen g. Since this works for all $g \in G$, the result follows. □

Example 1 (Continued: Admissible and Inadmissible Stopping Rules)**.** We obtain strong calibration for the one-sample t-test with respect to the nuisance parameter σ (see Example 1 above) when the stopping rule is adapted to the quotient filtration $\mathcal{G}_m, \mathcal{G}_{m+1}, \ldots$. Under each $P_{k,g} \in H_k$, the Bayes factors $\beta_m, \beta_{m+1}, \ldots$ define a random process on Ω such that each β_n is $\mathcal{G}_n$-measurable. This means that a stopping time defined in terms of a rule such as 'stop at the smallest t at which $\beta_t > 20$ or $t = 10^6$' is allowed in the result above. Moreover, if the stopping rule is a function of a sequence of maximal invariants, like $x_1/|x_1|, x_2/|x_1|, \ldots$, it is adapted to the filtration $\mathcal{G}_m, \mathcal{G}_{m+1}, \ldots$ and we can likewise apply the result above. On the other hand, this requirement is violated, for example, by a stopping rule that stops when $\sum_{i=1}^{j}(x_i)^2$ exceeds some fixed value, since such a stopping rule explicitly depends on the scale of the sampled data.

5.5 Frequentist Optional Stopping

The special case of the following result for the one-sample Bayesian t-test was proven in the master's thesis (Hendriksen, 2017). Here we extend the result to general group invariances.

Theorem 9 (Frequentist optional stopping for composite null hypotheses with group invariance)**.** *Under the same conditions as in Section 5.4, let τ be a stopping time such that, for each $n > m$, the event $\{\tau = n\}$ is $\mathcal{G}_n$-measurable. Then, adopting the notation of (16), for all $g \in G$, the stopped Bayes factor satisfies $\mathbf{E}_{P_{0,g}}[\beta_\tau] = \int_{\mathbb{R}_{>0}} c\,\mathrm{d}P_{0,g}^{[\beta_\tau]}(c) = 1$, so that, by the reasoning above Proposition 4, we have* for all $g \in G$: $P_{0,g}(\frac{1}{\beta_\tau} \leq \alpha) \leq \alpha$.

Proof. We have

$$\int_{\mathbb{R}_{>0}} c\; dP_{0,g}^{[\beta_\tau]}(c) = \int_{\mathbb{R}_{>0}} \frac{\mathrm{d}P_{1,g}^{[\beta_\tau]}}{\mathrm{d}P_{0,g}^{[\beta_\tau]}}(c) dP_{0,g}^{[\beta_\tau]}(c) = \int_{\mathbb{R}_{>0}} dP_{1,g}^{[\beta_\tau]}(c) = 1,$$

where the first equality follows directly from Theorem 8 and the final equality follows because $P_{1,g}$ is a probability measure, integrating to 1. □

Analogously to Corollary 5, the desired result now follows by plugging in a particular stopping rule: let $S : \bigcup_{i=m}^{\infty} \mathcal{X}^i \to \{0, 1\}$ be the frequentist sequential test defined by setting, for all $n > m$, $x^n \in \langle \mathcal{X}^n \rangle$: $S(x^n) = 1$ if and only if $\beta_n \geq 1/\alpha$.

Corollary 10. *Let $t^* \in \{m+1, m+2, \ldots\} \cup \{\infty\}$ be the smallest $t^* > m$ for which $\beta_{t^*}^{-1} \leq \alpha$. Then for arbitrarily large T, when applied to the stopping rule $\tau := \min\{T, t^*\}$, we find that* for all $g \in G$*:*

$$P_{0,g}(\exists n, m < n \leq T : S(X^n) = 1 \mid x^m) = P_{0,g}(\exists n, m < n \leq T : \beta_n^{-1} \leq \alpha \mid x^m) \leq \alpha.$$

The corollary implies that the test S is robust under optional stopping in the frequentist sense relative to H_0 (Definition 1).

Example 1 (continued)**.** When we choose a stopping rule that is $(\mathcal{G}_m, \mathcal{G}_{m+1}, \ldots)$-measurable, the hypothesis test is robust under (semi-)frequentist optional stopping. This holds for example, for the one- and two-sample t-test (Rouder et al., 2009), Bayesian ANOVA (Rouder et al., 2012), and Bayesian linear regression (Liang et al., 2008). Again, for stopping rules that are not $(\mathcal{G}_m, \mathcal{G}_{m+1}, \ldots)$-measurable, robustness under frequentist optional stopping cannot be guaranteed and could reasonably be presumed to be violated. The violation of robustness under optional stopping is hard to demonstrate experimentally as frequentist Bayes factor tests are usually quite conservative in approaching the asymptotic significance level α.

6 Concluding Remarks

We have identified three types of 'handling optional stopping': τ-independence, calibration and semi-frequentist. We extended the corresponding definitions and results to general sample spaces with potentially improper priors. For the special case of models H_0 and H_1 sharing a nuisance parameter with a group invariance structure, we showed stronger versions of calibration and semi-frequentist robustness to optional stopping. Some final remarks are in order. First, one of the remarkable properties of the right Haar prior is that, under some additional conditions on $P_{0,g}$ and $P_{1,g}$ in (22), $\beta_m = \beta(x^m) = 1$ for all $x^m \in \langle \mathcal{X}^m \rangle$, implying that equal prior odds lead to equal posterior odds after a minimal sample, no matter what the minimal sample is Berger et al. (1998b). One might conjecture that our results rely on this property, but this is not the case: in general, one can have $\beta(x^m) \neq 1$, yet our results still hold. For example, in the Bayesian t-test, Example 1, $m = 1$ and $\beta(x^1) = 1$ can be guaranteed only if the prior π_δ on δ is symmetric around 0; but our calibration and frequentist robustness results hold irrespective of whether it is symmetric or not.

As a second remark, it is worth noting that — as is immediate from the proofs — all our group-invariance results continue to hold in the setting with H_k' as in (26), and the definition of the Bayes marginal $P_{k,e}$ relative to θ' as in (27) replaced by a probability measure on $(\Omega, \mathcal{F})$ that is not necessarily of the Bayes marginal form. The results work for any probability measure; in particular one can take the alternatives for the Bayes marginal with proper prior that are considered in the minimum description length and sequential prediction literature (Barron et al., 1998; Grünwald, 2007) under the name

of *universal distribution* relative to $\{P_{\theta'} \mid \theta' \in \Theta'\}$; examples include the prequential or 'switch' distributions considered by van der Pas and Grünwald (2018).

As a third remark, a sizeable fraction of Bayesian statisticians is wary of using improper priors at all. An important (though not the only) reason is that their use often leads to some form of the *marginalization paradox* described by Dawid et al. (1973). It is thus useful to stress that in the context of Bayes factor hypothesis testing, the right Haar prior is immune at least to this particular paradox. In an informal nutshell, the marginalization paradox occurs if the following happens: (a) the Bayes posterior $\pi(\zeta \mid X^n)$ for the quantity of interest ζ based on prior $\pi(\zeta, g)$ with improper marginal on g, only depends on the data X^n through the maximal invariant U_n, i.e. $\pi(\zeta \mid X^n) = f(U_n(X^n))$ for some function f, yet (b) there exists no prior π' on ζ such that the corresponding posterior $\pi'(\zeta \mid U_n(X^n)) = f(U_n(X^n))$. In words, the result of Bayesian updating based on the full data X^n only depends on the maximal invariant U^n; but Bayesian updating directly based on U^n can never give the same result — a paradox indeed. While in general, this can happen even if g is equipped with the right Haar prior [Case 1, page 199] (Dawid et al., 1973), Berger et al.'s Theorem 2.1 (reproduced in Section 5.2 in our paper) implies that it does not occur in the context of Bayes factor testing, where $\zeta \in \{H_0, H_1\}$, and H_0 and H_1 are null and alternatives satisfying the requirements of Section 5. Berger's theorem expresses that for all values of the nuisance parameter $g \in G$, the likelihood ratio $\mathrm{d}P_{1,g}^{[U_n]} / \mathrm{d}P_{0,g}^{[U_n]}(U_n(X^n))$ based on $U_n(X^n)$ is equal to the Bayes factor based on X^n with the right Haar prior on g, so that the paradox cannot occur.

Finally, as pointed out by a referee, even though our use of the right Haar prior avoids the marginalization paradox, there are still some issues with its use. First, not all priors on parameters of interest work well in combination with right Haar priors: in our running Example 1, with light-tailed priors on δ such as a normal or a point prior, for fixed n, the Bayes factor does not go to 0 as the data become increasingly extreme in the sense that their empirical variance goes to 0 but their mean does not. This phenomenon of *information inconsistency* is avoided by placing a heavy-tailed prior on δ, as advocated by e.g. Rouder et al. (2009) and Jeffreys (1961). Many Bayesians, who think that at least proper priors should always have an interpretation in terms of degrees-of-belief, would object to any method (such as using right Haar priors) that induces such a restriction on 'reasonable' priors. On the other hand, 'objective' Bayesians (Berger, 2006) would not see any problems here, and pragmatic Bayesians may not care about information inconsistency. A similar issue is that, like all improper priors, right Haar priors are defined only up to a constant proportionality factor, and the Bayes factor approach only makes sense if these factors are chosen to be the same for both models. Yet again, for 'objective' Bayesians this would not count as an issue. Still, there is one issue that is even problematic from an objective Bayes standpoint: in some cases, the same models H_0 and H_1 can be generated by many different groups with different right Haar priors (Eaton and Sudderth, 2002; Berger et al., 2008). Sometimes different groups lead to the same Bayes factor, sometimes they do not. Our results continue to hold in that case, but in practice, it would not be clear which Bayes factor one would want to choose. We are currently working on an in-depth analysis of this problem. Preliminary results

suggest that in such cases, there may in fact be a unique preferred choice for the Bayes factor, but more work is needed to establish this with certainty.

Supplementary Material

Supplementary Material (DOI: 10.1214/20-BA1234SUPP; .pdf). The paper ends with supplementary material (Hendriksen et al., 2020), comprising Section A containing basic background material about groups, and Section B containing all longer mathematical proofs.

References

Andersson, S. (1982). "Distributions of maximal invariants using quotient measures." *The Annals of Statistics*, 10(3): 955–961. MR0663446. 963, 980

Barnard, G. A. (1947). "Review of *Sequential Analysis* by Abraham Wald." *Journal of the American Statistical Association*, 42(240). MR0044466. doi: https://doi.org/10.1177/0008068319510401. 965

Barnard, G. A. (1949). "Statistical inference." *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(2): 115–149. MR0034975. 965

Barron, A., Rissanen, J., and Yu, B. (1998). "The minimum description length principle in coding and modeling." *IEEE Transactions on Information Theory*, 44(6): 2743–2760. MR1658898. doi: https://doi.org/10.1109/18.720554. 964, 984

Bayarri, M. J., Benjamin, D. J., Berger, J. O., and Sellke, T. M. (2016). "Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses." *Journal of Mathematical Psychology*, 72: 90–103. MR3506028. doi: https://doi.org/10.1016/j.jmp.2015.12.007. 962, 971

Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012). "Criteria for Bayesian model choice with application to variable selection." *The Annals of Statistics*, 40(3): 1550–1577. MR3015035. doi: https://doi.org/10.1214/12-AOS1013. 962, 973, 982

Berger, J. (2006). "The case for objective Bayesian analysis." *Bayesian Analysis*, 1(3): 385–402. MR2221271. doi: https://doi.org/10.1214/06-BA115. 985

Berger, J. O., Pericchi, L. R., and Varshavsky, J. A. (1998a). "Bayes factors and marginal distributions in invariant situations." *Sankhyā: The Indian Journal of Statistics, Series A*, 307–321. MR1718789. 962, 981, 982

Berger, J. O., Pericchi, L. R., and Varshavsky, J. A. (1998b). "Bayes factors and marginal distributions in invariant situations." *Sankhyā: The Indian Journal of Statistics, Series A*, 307–321. 979, 984

Berger, J. O., Sun, D., et al. (2008). "Objective priors for the bivariate normal model." *The Annals of Statistics*, 36(2): 963–982. MR2396821. doi: https://doi.org/10.1214/07-AOS501. 985

Berger, J. O. and Wolpert, R. L. (1988). *The Likelihood Principle*. Hayward, CA: Institute of Mathematical Statistics, 2nd edition. MR0773665. 963, 966

Dass, S. C. (1998). "Unified Bayesian and conditional frequentist testing procedures." Ph.D. thesis, University of Michigan. MR2698532. 979

Dass, S. C. and Berger, J. O. (2003). "Unified conditional frequentist and Bayesian testing of composite hypotheses." *Scandinavian Journal of Statistics*, 30(1): 193–210. MR1965102. doi: https://doi.org/10.1111/1467-9469.00326. 962, 971, 979

Dawid, A. P., Stone, M., and Zidek, J. V. (1973). "Marginalization paradoxes in Bayesian and structural inference." *Journal of the Royal Statistical Society: Series B (Methodological)*, 35(2): 189–213. MR0365805. 963, 985

Deng, A., Lu, J., and Chen, S. (2016). "Continuous monitoring of A/B tests without pain: Optional stopping in Bayesian testing." In *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*, 243–252. IEEE. MR3578198. doi: https://doi.org/10.1109/MCS.2016.2602089. 967

Eaton, M. L. (1989). "Group Invariance Applications in Statistics." *Regional Conference Series in Probability and Statistics*, 1: i–133. MR1089423. 963, 977, 979, 980

Eaton, M. L. and Sudderth, W. D. (2002). "Group invariant inference and right Haar measure." *Journal of Statistical Planning and Inference*, 103(1-2): 87–99. MR1896985. doi: https://doi.org/10.1016/S0378-3758(01)00199-9. 985

Edwards, W., Lindman, H., and Savage, L. J. (1963). "Bayesian statistical inference for psychological research." *Psychological Review*, 70(3): 193–242. 961, 965, 968, 970

Good, I. J. (1991). "C383. A comment concerning optional stopping." *Journal of Statistical Computation and Simulation*, 39(3): 191–192. MR1917915. doi: https://doi.org/10.1080/00949650211421. 968, 970

Grünwald, P., de Heide, R., and Koolen, W. (2019). "Safe testing." *arXiv preprint arXiv:1906.07801*. MR3108509. doi: https://doi.org/10.1007/978-3-642-39091-3_21. 962, 971

Grünwald, P. D. (2007). *The Minimum Description Length Principle*. Cambridge, MA: MIT Press. 964, 984

de Heide, R. and Grünwald, P. (2018). "Why optional stopping is a problem for Bayesians." *arXiv preprint arXiv:1708.08278*. 962, 966, 967, 968, 969

Hendriksen, A., de Heide, R., and Grünwald, P. (2020). "Supplement to: Optional Stopping with Bayes Factors: a categorization and extension of folklore results, with an application to invariant situations." *Bayesian Analysis*. 975, 977

Hendriksen, A. A. (2017). "Betting as an alternative to p-values." Master's thesis, Leiden University, Dept. of Mathematics. 983

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2018). "Uniform, nonparametric, non-asymptotic confidence sequences." *arXiv preprint arXiv:1810.08240*. 971

Jamil, T., Ly, A., Morey, R. D., Love, J., Marsman, M., and Wagenmakers, E.-J. (2016). "Default "Gunel and Dickey" Bayes factors for contingency tables." *Behavior Research Methods*, 49(2): 638–652. 962, 969

Jeffreys, H. (1961). *Theory of Probability*. Oxford, England: Oxford. MR0187257. 962, 970, 985

John, L. K., Loewenstein, G., and Prelec, D. (2012). "Measuring the prevalence of questionable research practices with incentives for truth telling." *Psychological Science*. 961

Lai, T. L. (1976). "On confidence sequences." *The Annals of Statistics*, 4(2): 265–280. MR0395103. 971, 989

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). "Mixtures of g priors for Bayesian variable selection." *Journal of the American Statistical Association*, 103(481): 410–423. MR2420243. doi: https://doi.org/10.1198/016214507000001337. 984

Lindley, D. V. (1957). "A statistical paradox." *Biometrika*, 44(1/2): 187–192. MR0087273. doi: https://doi.org/10.1093/biomet/44.1-2.179. 961, 965

van der Pas, S. and Grünwald, P. D. (2018). "Almost the best of three worlds: risk, consistency and optional stopping for the switch criterion in nested model selection." *Statistica Sinica*, 28(1): 229–253. MR3752259. 964, 985

Proschan, M. A., Lan, K. G., and Wittes, J. T. (2006). *Statistical Monitoring of Clinical Trials: A Unified Approach*. Springer Science & Business Media. 968

Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Cambridge, MA: Harvard University Press. MR0117844. 961, 966

Rouder, J. N. (2014). "Optional stopping: No problem for Bayesians." *Psychonomic Bulletin & Review*, 21(2): 301–308. 962, 966, 969

Rouder, J. N., Morey, R. D., Speckman, P. L., and Province, J. M. (2012). "Default Bayes factors for ANOVA designs." *Journal of Mathematical Psychology*, 56(5): 356–374. MR2983394. doi: https://doi.org/10.1016/j.jmp.2012.08.001. 962, 984

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. (2009). "Bayesian t tests for accepting and rejecting the null hypothesis." *Psychonomic Bulletin & Review*, 16(2): 225–237. 962, 970, 980, 984, 985

Sanborn, A. N. and Hills, T. T. (2014). "The frequentist implications of optional stopping on Bayesian hypothesis tests." *Psychonomic Bulletin & Review*, 21(2): 283–300. 962

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., and Perugini, M. (2017). "Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences." *Psychological Methods*, 22(2): 322–339. 962

Shafer, G., Shen, A., Vereshchagin, N., and Vovk, V. (2011). "Test martingales, Bayes

factors and p-values." *Statistical Science*, 26(1): 84–101. MR2849911. doi: https://doi.org/10.1214/10-STS347. 964

Wagenmakers, E.-J. (2007). "A practical solution to the pervasive problems of p values." *Psychonomic Bulletin & Review*, 14(5): 779–804. 962

Wijsman, R. A. (1990). *Invariant Measures on Groups and Their Use in Statistics*. Institute of Mathematical Statistics. MR1218397. 963

Yu, E. C., Sprenger, A. M., Thomas, R. P., and Dougherty, M. R. (2014). "When decision heuristics and science collide." *Psychonomic Bulletin & Review*, 21(2): 268–282. 962

Acknowledgments

We are grateful to Wouter Koolen, for extremely useful conversations that helped with the math, to Jeff Rouder, for providing inspiration and insights that sparked off this research, to Aaditya Ramdas, who brought Lai (1976) to our attention, and to an anonymous referee who raised an important issue concerning potential ambiguities in the use of right Haar priors.