Anytime-Valid Inference For Multinomial Count Data

Michael Lindon
Netflix
michael.s.lindon@gmail.com

Alan Malek
alan.malek@gmail.com

Abstract

Many experiments compare count outcomes among treatment groups. Examples include the number of successful signups in conversion rate experiments or the number of errors produced by software versions in canary tests. Observations typically arrive in a sequence and practitioners wish to continuously monitor their experiments, sequentially testing hypotheses while maintaining Type I error probabilities under optional stopping and continuation. These goals are frequently complicated in practice by non-stationary time dynamics. We provide practical solutions through sequential tests of multinomial hypotheses, hypotheses about many inhomogeneous Bernoulli processes and hypotheses about many time-inhomogeneous Poisson counting processes. For estimation, we further provide confidence sequences for multinomial probability vectors, all contrasts among probabilities of inhomogeneous Bernoulli processes and all contrasts among intensities of time-inhomogeneous Poisson counting processes. Together, these provide an "anytime-valid" inference framework for a wide variety of experiments dealing with count outcomes, which we illustrate with several industry applications.

1 Introduction

Many experiments compare *count outcomes* among treatment groups (arms). For example, in online conversion experiments we seek the treatment group that achieves the greatest number of conversions (signups, purchases etc...). In software experiments, we seek the software version that produces the fewest number of errors [Lindon et al., 2022, Kuo and Yang, 1996]. These experiments are frequently complicated in practice by time-variability over the course of the experiment caused by uncontrollable external factors. The Bernoulli success probability of conversion is often time-varying due to uncontrollable product changes and market conditions. The instantaneous rate of software errors is time-varying due to time-varying usage patterns of the software application among users. Fortunately, the time-variability is often common to all arms of the experiment. Therefore, the count outcomes among arms should be in some sense comparable under the null hypothesis that all arms behave equally.

In testing these hypotheses, it is desirable if a statistical test can both (i) detect large treatment effects early, and (ii) detect small treatment effects eventually. If there is a large negative effect then it is important to end the experiment as soon as possible to reduce risk to the experimental units and/or company. However, it is also important to detect small negative effects, as these can accumulate over time. Classical fixed-n tests cannot satisfy both of these objectives, as there is only one opportunity to perform the test. Performing the test earlier achieves objective (i), but doesn't achieve objective (ii) as it may be underpowered. Performing the test later achieves objective (ii), but doesn't achieve objective (i) and treatments may affect experimental units negatively for lengthy amounts of time. For this reason we provide sequential tests that allow experiments to be continuously monitored. We have included a discussion of fixed-n testing vs sequential testing and corresponding literature review in the appendix.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Sequential testing in experiments with Bernoulli outcomes is complicated in practice because failure outcomes are typically not directly observed. In many fixed-n approaches, Bernoulli fails are often *inferred* from the absence of successes within an arbitrary window of time. This is not possible for real-time sequential testing when the raw data are streams of Bernoulli successes with associated timestamps. We are able to circumvent this issue by basing the sequential test solely on the counts of successes accumulated for each arm so far. This has the advantage of also removing the time-varying components in the success probabilities.

We begin by providing an anytime-valid inference approach for multinomial families based on a mixture martingale, which is widely applicable to experiments dealing with count outcomes. We build upon this test to provide sequential tests of time-inhomogeneous Bernoulli processes and time-inhomogeneous Poisson counting processes. The former is widely applicable to industry A/B experiments with Bernoulli outcomes, which we illustrate with a payments funnel experiment. The latter is widely applicable to industry A/B experiments which observe point processes in time, which we illustrate with an A/B test used to monitor the safe roll-out of a new software application version. Our results are general for an arbitrary number of arms.

2 A Sequential Multinomial Test

2.1 Theory

Our development of sequential tests for different kinds of count data begins with a sequential test for multinomial observations. The construction follows the following sequence of steps common in the literature [Shafer et al., 2011, Waudby-Smith and Ramdas, 2020, Howard et al., 2021]. Step 1: Define a relevant test statistic. Step 2: Show that the test statistic is a test martingale - a nonnegative supermartingale under the null hypothesis. Step 3: Use martingale inequalities to bound the frequentist Type I error probability uniformly over time below a desired level u. Step 4: Invert the sequential test based on the test martingale to obtain a confidence sequence with a time-uniform coverage guarantee of at least $1 - u$. Extensions to other kinds of count data, including Bernoulli, Binomial and Poisson counting processes, are then obtained by recognizing relationships that exist to the multinomial distribution.

Our test martingale is obtained through a mixture martingale construction, integrating the multinomial likelihood with respect to a Dirichlet mixture distribution [Kaufmann and Koolen, 2021]. The likelihood mixture is also the statistic used in mixture sequential probability ratio tests (mSPRT) [Wald, 1945]. Bayesians will immediately recognize this statistic as a Bayes factor [Jeffreys, 1935, Kass and Raftery, 1995], interpreting the Dirichlet mixture distribution as a prior distribution, which could be used for computing posterior probabilities. However, not wishing to confuse the reader, we emphasize that we are not developing a Bayesian procedure. We aim to develop a procedure for "anytime-valid" inference, providing u-level sequential tests and $1 - u$ confidence sequences. We will borrow some language from Bayesian analysis, but the frequentist Type I error and coverage guarantees *do not* depend on the particular choice of Dirichlet distribution. To avoid confusion, we encourage the reader to view the Bayes factor simply as a test statistic.

The test statistic can also be viewed as an *e-process* [Grünwald et al., 2021, Pérez-Ortiz et al., 2022, Hendriksen et al., 2021]. A review of these constructions and their relationships is discussed in Ramdas et al. [2022]. The purpose of this paper is not to study the theoretical properties of these constructions [Wald and Wolfowitz, 1948], but rather to develop anytime-valid tests for several challenging applications in modern online experimentation dealing with count outcomes.

For step 1, consider a sequence $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \ldots$ of independent $\mathrm{Multinomial}(1, \boldsymbol{\theta})$ random variables with $\boldsymbol{\theta} \in \triangle^d$, the $d-1$ simplex. We use bold typeset to denote vectors. Under the null hypothesis, M_0, it is assumed that

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots | M_0 \overset{\text{i.i.d.}}{\sim} \mathrm{Multinomial}(1, \boldsymbol{\theta}_0). \tag{1}$$

To construct a model over alternatives, M_1, we consider the marginal distribution obtained from the following joint model

$$\begin{aligned} \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots | \boldsymbol{\theta}, M_1 &\overset{\text{i.i.d.}}{\sim} \mathrm{Multinomial}(1, \boldsymbol{\theta}), \\ \boldsymbol{\theta} | M_1 &\sim \text{Dirichlet}(\boldsymbol{\alpha}_0). \end{aligned} \tag{2}$$

The following expressions reduce to simple forms when a uniform distribution over the simplex is used, achieved by setting $\alpha_{0,i} = 1$ for $i = 1, \ldots, d$, which may be a good default value. The Type I error and coverage guarantees are unaffected by this choice, but the expected stopping time of the sequential test is influenced. The optimal choice of this parameter depends upon the value of $\boldsymbol{\theta}$ under the alternative, which is unknown. Stopping times under large departures from the null are reduced by using the uniform distribution. Whereas stopping times under very small departures from the null are reduced by concentrating the Dirichlet distribution about $\boldsymbol{\theta}_0$, which can be achieved with the choice $\alpha_{0,i} = k\theta_{0,i}$ for a concentration parameter $k \in \mathbb{R}^+$. In general, the average stopping time is reduced by matching the Dirichlet distribution to the distribution of effects upon which these sequential tests will be performed.

Let $S_i^n = \sum_{j=1}^n x_{j,i}$ and $\boldsymbol{S}_n = (S_1^n, \ldots, S_d^n) \in \mathbb{R}^d$. In addition, let $|\boldsymbol{v}| = \sum_i v_i$ denote the element-wise sum of a vector $\boldsymbol{v}$, $\boldsymbol{v}^{\boldsymbol{w}} = \prod_i v_i^{w_i}$ to denote element-wise exponentiation of two vectors $\boldsymbol{v}$ and $\boldsymbol{w}$, and Beta to denote the multivariate Beta function $\mathrm{Beta}(\boldsymbol{v}) := (\prod_i \Gamma(v_i))/\Gamma(\sum_i v_i)$. The resulting Bayes factor comparing models M_1 to M_0 is given by

$$BF_{10}(\boldsymbol{x}_{1:n}) = \frac{\mathrm{Beta}(\boldsymbol{\alpha}_0 + \boldsymbol{S}_n)}{\mathrm{Beta}(\boldsymbol{\alpha}_0)} \frac{1}{\boldsymbol{\theta}_0^{\boldsymbol{S}_n}}. \tag{3}$$

which appears as early as Good [1967] (derivation in Appendix A.4). In a Bayesian analysis, the Bayes factor multiplied by the prior odds gives the posterior odds of M_1 over M_0. In this work, we take the prior odds to be unity so that the terms Bayes factor and posterior odds can be used interchangeably.

In sequential applications, it often makes sense to compute Equation (3) recursively. Let $O_n(\boldsymbol{\theta}_0)$ denote the posterior odds at n, then

$$O_n(\boldsymbol{\theta}_0) = \frac{\mathrm{Beta}(\boldsymbol{\alpha}_{n-1} + \boldsymbol{x}_n)}{\mathrm{Beta}(\boldsymbol{\alpha}_{n-1})} \frac{1}{\boldsymbol{\theta}_0^{\boldsymbol{x}_n}} O_{n-1}(\boldsymbol{\theta}_0), \tag{4}$$

where $\boldsymbol{\alpha}_n = \boldsymbol{\alpha}_{n-1} + \boldsymbol{x}_n$ and $O_0(\boldsymbol{\theta}_0) = p(M_1)/p(M_0) = 1$. When $\alpha_{0,i}$ are integer valued, Equation (4) can be expressed in terms of the posterior mean $\mathbb{E}[\theta_i | \boldsymbol{x}_{1:n-1}]$ for each θ_i as

$$O_n(\boldsymbol{\theta}_0) = \prod_{i=1}^d \left(\frac{\mathbb{E}[\theta_i | \boldsymbol{x}_{1:n-1}]}{\theta_{0,i}} \right)^{x_{n,i}} O_{n-1}(\boldsymbol{\theta}_0). \tag{5}$$

Details are provided in Appendix A.5. Equation (5) yields an intuitive betting interpretation [Shafer, 2021]. It represents our bet against the null hypothesis $\boldsymbol{\theta}_0$ using a $\boldsymbol{\theta}$ learned from the data. In this case, our estimate of $\boldsymbol{\theta}$ based on the data is the posterior mean. If a uniform prior is used, our initial bet on each outcome $x_{1,i}$ is simply $1/d$, and our bet on each outcome $x_{n,i}$ is $(1 + S_i^{n-1})/(n + d)$. The dependence of $O_n(\boldsymbol{\theta}_0)$ on the observed data $\boldsymbol{x}_{1:n}$ is implicit in this notation, yet the null value $\boldsymbol{\theta}_0$ being tested is made explicit to aid the discussion of confidence sequences in Theorem 2.4.

Step 2 in our construction is to demonstrate that $O_n(\boldsymbol{\theta}_0)$ is a nonnegative supermartingale under the null hypothesis M_0.

Theorem 2.1. *Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ be a sequence of independent $\mathrm{Multinomial}(1, \boldsymbol{\theta})$ random variables with the filtration $\mathcal{F}_{n-1} = \sigma(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n-1})$ and consider the sequence of posterior odds $O_n(\boldsymbol{\theta}_0)$ defined in Equation* (4) *with $O_0(\boldsymbol{\theta}_0) = 1$. Then*

$$\mathbb{E}_{M_0}[O_n(\boldsymbol{\theta}_0) | \mathcal{F}_{t-1}] = O_{n-1}(\boldsymbol{\theta}_0). \tag{6}$$

The proof is found in Appendix A.6. Theorem 2.1 states that $O_n(\boldsymbol{\theta}_0)$ is a nonnegative martingale under the null hypothesis with respect to the canonical filtration.

Step 3 is to use the posterior odds to construct a test martingale.

Theorem 2.2. *Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ be a sequence of independent $\mathrm{Multinomial}(1, \boldsymbol{\theta})$ random variables and consider the sequence of posterior odds $O_n(\boldsymbol{\theta}_0)$ defined in Equation* (4) *with $O_0(\boldsymbol{\theta}_0) = 1$. Then*

$$\mathbb{P}_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \left(\exists n \in \mathbb{N} : O_n(\boldsymbol{\theta}_0) \geq 1/u \right) \leq u \tag{7}$$

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

The proof is provided in Appendix A.7. The time-uniform bound presented in Theorem 2.2 controls the deviations of a stochastic process for all t simultaneously and is essential for proving the correctness of sequential tests and verifying the optional stopping and optional continuation properties. It provides a valid stopping rule: reject the null at time $\tau = \inf\{n \in \mathbb{N} : O_n(\boldsymbol{\theta}_0) \geq 1/u\}$. Simply stated, a practitioner who rejects the null hypothesis as soon as the posterior odds become larger than $1/u$ faces a frequentist Type I error with probability of at most u. Shafer et al. [2011], Johari et al. [2021] bring this idea back to more familiar territory by constructing a *sequential p-value* by tracking the running supremum of the posterior odds and taking its inverse, or equivalently

$$p_0 = 1 \text{ and}$$
$$p_n = \min(p_{n-1}, 1/O_n(\boldsymbol{\theta}_0)).$$

It follows from this definition and equation (7) that

$$\mathbb{P}_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\left(\exists n \in \mathbb{N} : p_n \leq u\right) \leq u, \tag{8}$$

which is an easily digestible generalization of a fixed-n p-value to sequential settings. Instead of holding only at some pre-specified $n \in \mathbb{N}$, this guarantee holds *for all* $n \in \mathbb{N}$. This construction is shown in Figure 1. A simulation empirically demonstrating the control of false positives under continuous monitoring relative to a χ^2 test is shown in Appendix A.9.

Before completing Step 4, it is useful to show that this sequential test is not trivial. For this test to have utility it must possess the ability to control not only Type I errors, as in theorem 2.2, but also Type II errors. This is provided by the following theorem.

Theorem 2.3. *Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ be a sequence of independent* $\text{Multinomial}(1, \boldsymbol{\theta})$ *random variables and consider the sequence of posterior odds $O_n(\boldsymbol{\theta}_0)$ defined in Equation* (4) *with $O_0(\boldsymbol{\theta}_0) = 1$. If $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, then*

$$\frac{1}{n}\log O_n(\boldsymbol{\theta}_0) \to D_{KL}(\boldsymbol{\theta}||\boldsymbol{\theta}_0) \qquad a.s., \tag{9}$$

where $D_{KL}(\boldsymbol{\theta}||\boldsymbol{\theta}_0)$ is the Kullback-Leibler divergence of the true multinomial distribution with true parameter $\boldsymbol{\theta}$ from the multinomial distribution under the null hypothesis with null parameter $\boldsymbol{\theta}_0$.

The proof is given in Appendix A.8. Theorem 2.3 states that if the null hypothesis is not true, with $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, then the Bayes factor will diverge to infinity and exceed the $1/u$ threshold in Theorem 2.2 (a.s.), or equivalently that the sequential p-value converges to zero and falls below the u threshold (a.s.). In other words, this test is guaranteed to reject the null almost surely if the null is incorrect, which is considered to be *asymptotically power 1* by Robbins [1970]. This result follows simply from the posterior consistency of Bayes factors. We now state step 4 of the construction.

Theorem 2.4. *Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ be a sequence of independent* $\text{Multinomial}(1, \boldsymbol{\theta})$ *random variables and consider the sequence of posterior odds for testing the null hypothesis $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, $O_n(\boldsymbol{\theta}_0)$, defined in Equation* (4) *with $O_0(\boldsymbol{\theta}_0) = 1$. Let $C_n(u) = \{\boldsymbol{\theta} \in \triangle^d : O_n(\boldsymbol{\theta}) < 1/u\}$ denote the set of parameter vectors that would not be rejected by the test at the u level, then*

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\boldsymbol{\theta} \in C_n(u) \textit{ for all } n \in \mathbb{N}\right) \geq 1 - u \tag{10}$$

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

A simple corollary of theorem 2.4 is that $\mathbb{P}_{\boldsymbol{\theta}}\left(\boldsymbol{\theta} \in \bigcap_{n=1}^{\infty} C_n(u)\right) \geq 1 - u$. This result provides a confidence statement for sequentially estimating the true parameter vector $\boldsymbol{\theta}$ as the experiment progresses. The confidence set $C_n(u)$ for $\boldsymbol{\theta}$ is a convex subset of $\triangle^d$, with convexity following from the concavity of the multinomial log-likelihood. Confidence intervals on the individual elements of $\boldsymbol{\theta}$ can be obtained by projecting $C_n(u)$ onto the coordinate axes in the following manner.

Corollary 2.5. *For $C_n(u)$ as in Theorem 2.4, let*

$$j^+_{n,i}(u) = \sup\{\theta_i : \boldsymbol{\theta} \in C_n(u)\},$$
$$j^-_{n,i}(u) = \inf\{\theta_i : \boldsymbol{\theta} \in C_n(u)\},$$

then

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\forall i : \theta_i \in \bigcap_{n=1}^{\infty}[j^-_{n,i}(u), j^+_{n,i}(u)]\right) \geq 1 - u. \tag{11}$$

$j^+_{n,i}(u)$ can be computed by solving the following convex optimization program

$$\begin{aligned} \max \quad & \theta_i \\ \text{s.t.} \quad & c + \log u \leq \sum_i S_i^n \log \theta_i \\ & \sum_i \theta_i = 1 \end{aligned} \tag{12}$$

where $c = \log \mathrm{Beta}(\boldsymbol{\alpha}_0 + \boldsymbol{S}_n) - \log \mathrm{Beta}(\boldsymbol{\alpha}_0)$. The constraints in the optimization program simply define $C_n(u)$. Similarly, $j^-_{n,i}(u)$ is obtained by minimizing θ_i over this set.

Lastly, the confidence set in Theorem 2.4 is equal to the Bayesian *support set* [Wagenmakers et al., 2022].

Corollary 2.6. *The Bayesian support set, defined as* $\{\boldsymbol{\theta} \in \triangle^d : p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n}, M_1) \geq up(\boldsymbol{\theta}|M_1)\}$ *has time-uniform frequentist coverage of at least* $1 - u$.

The proof is a simple application of the Savage-Dickey density ratio $O_n(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|M_1)/p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n}, M_1)$ [Dickey, 1971].

2.2 Application: Sample Ratio Mismatch Tests

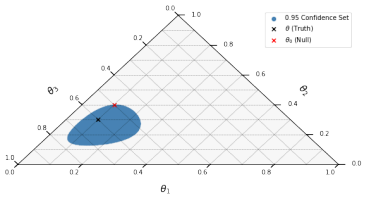
Comparing the counts of experimental units assigned to each arm of a multi-arm experiment can often detect bias and errors in the experiment. Most online controlled experiments follow simply randomized designs, whereby a new unit is randomly assigned to one of d arms according to a vector of pre-specified probabilities $\boldsymbol{\theta}$. The assignment outcome for a new unit is independent of other units (individualistic), unit-level covariates, potential outcomes (unconfounded) and can therefore be summarized as an independent $\mathrm{Multinomial}(1, \boldsymbol{\theta})$ random variable. Under these assumptions, the assignment mechanism can be considered ignorable when performing inference on causal estimands such as the average treatment effect [Imbens and Rubin, 2015].

Although simple in theory, the systems that perform assignments quickly grow in complexity as the number of concurrent experiments increases [Tang et al., 2010]. This increased complexity increases the risk of introducing bugs that cause departures from the intended assignment mechanism, breaking the assumption of ignorability and rendering causal estimates invalid. Zhao et al. [2016] provide an account of an incorrect hashing algorithm introducing bias into the assignment mechanism.

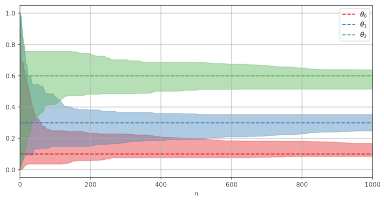
After assignment and measurement, data passes through processing and cleansing pipelines before analysis. If incorrect cleansing logic is applied, there is a risk that specific observations may be selectively removed, introducing a "missing not at random" missing data mechanism, rendering causal estimates invalid [Rubin, 1976]. Fabijan et al. [2019] describes an experiment in which units from the treatment arm were unintentionally removed, with the probability of their removal depending on their observed outcomes.

An arm with a surprisingly low or high number of units is usually symptomatic of an implementation error in the experiment and is colloquially referred to as a *sample ratio mismatch* (SRM) [Fabijan et al., 2019]. These errors can be caught by comparing the counts of experimental units in each arm against the intended assignment probabilities. It is now considered good practice to validate the experiment setup by performing a χ^2 test, comparing the observed counts against the expected counts under the intended assignment mechanism [Chen et al., 2018]. However, the χ^2 test is an example of a *fixed-*n test, providing statistical guarantees when performed *once*. Due to this limitation, it is typically performed after data collection and before analysis. To reveal a bug that renders an expensive experiment invalid only after the experiment is finished would be less than ideal. Ideally, SRMs are detected as early as possible so that the implementation error can be corrected before more units enter the experiment. This necessitates sequential testing. We provide a sequential multinomial test for testing a point null based on the counts of each outcome.

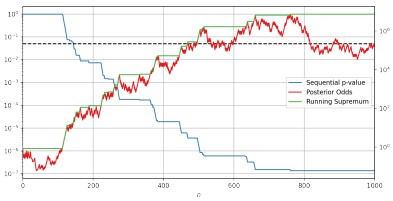
Figure 1 shows the application of the sequential multinomial test to a representative experiment exhibiting an SRM. In this example the intended assignment probabilities are $\boldsymbol{\theta}_0 = (0.1, 0.4, 0.5)$. These propensity scores are assumed to be known by the analyst and will subsequently be used to perform inference about average treatment effects. Due to a bug, either in randomization or data processing, some units are preferentially assigned to certain arms based on their covariates, or data is



(a) Joint Confidence Set at $n = 144$



(b) Simultaneous Marginal Confidence Sequences



(c) Sequential p-value and Bayes Factor

Figure 1: (a) The confidence set $C_{144}(0.05)$ at $n = 144$ as defined in Theorem 2.4. The true $\boldsymbol{\theta} = (0.1, 0.3, 0.6)$ is marked with a black cross, and the null hypothesis $\boldsymbol{\theta}_0 = (0.1, 0.4, 0.5)$ is denoted with a red cross. (b) Simultaneous 0.95 confidence sequences that cover the individual elements of $\boldsymbol{\theta} = (0.1, 0.3, 0.6)$ obtained from Corollary 2.5 and computed via the optimization program in Equation (12). (c) (Left axis) Sequential p-value (blue) defined in Equation (8). Critical value $u = 0.05$ (dashed-black). $\boldsymbol{\theta} = (0.1, 0.3, 0.6)$, $\boldsymbol{\theta}_0 = (0.1, 0.4, 0.5)$. $p_n < 0.05$ for all $n \geq 144$. (Right axis) The posterior odds defined in Equation (3) (red), with the running supremum (green).

being lost not at random, resulting in observations arriving with probabilities $\boldsymbol{\theta} = (0.1, 0.3, 0.6)$. The sequential p-value is less than 0.05 for all $n \geq 144$. This is the smallest n for which $\boldsymbol{\theta}_0 \notin C_n(0.05)$.

3 Inhomogeneous Bernoulli Processes

Wald [1947] described an experiment to test the null hypothesis that two guns are equally precise. Two guns are fired at a target simultaneously for many attempts, and their success is recorded. Each attempt is modelled as a Bernoulli trial and, under the null hypothesis, the probability of each gun hitting the target on the same attempt is equal. However, the success probability varies across attempts due to gusty wind conditions. As the guns are shot simultaneously, the wind condition affects each gun equally. If the null is rejected, choosing the gun that obtained the greatest number of successes seems reasonable.

3.1 Theory

Suppose a new experimental unit is randomly assigned to one of d experiment treatment groups (arms) at time t, according to assignment probabilities $\boldsymbol{\rho} \in \triangle^d$, and a Bernoulli outcome is observed. The Bernoulli probability for arm i at time t is parameterized by $p_i(t) = e^{\mu(t)} e^{\delta_i}$ so that the time-varying effect is multiplicative and common to all arms. The improvement of arm j over arm i at any time is then $p_j(t)/p_i(t) = \exp(\delta_j - \delta_i)$, and the difference on the log-scale is simply $\delta_j - \delta_i$. Suppose Bernoulli failures are ignored and arms are compared only through their counts of Bernoulli successes. The (conditional) probability that the next Bernoulli success comes from arm i is

$$\theta_i = \frac{\rho_i e^{\delta_i}}{\sum_{j=1}^{d} \rho_j e^{\delta_j}}, \tag{13}$$

which is independent of the time-varying effect. The arm from which the next Bernoulli success arrives is, therefore, a $\mathrm{Multinomial}(1, \boldsymbol{\theta})$ random variable, and the counts of Bernoulli successes for

each arm are independent of the time-varying nuisance parameter $\mu(t)$. Framing the problem this way allows the sequential multinomial test to perform inference on $\boldsymbol{\delta}$. Simple hypotheses about $\boldsymbol{\delta}$ can therefore be translated into testing simple hypotheses about $\boldsymbol{\theta}$. Equality among success probabilities can be tested by simply testing the null multinomial hypothesis $\boldsymbol{\theta}_0 = \boldsymbol{\rho}$. The individual components δ_i are not identifiable, as adding a constant to each element results in the same $\boldsymbol{\theta}$, yet contrasts of the form $\sum_i a_i \delta_i$ for $\sum_i a_i = 0$ are identifiable.

Let $\sigma_{\boldsymbol{\rho}} : \mathbb{R}^d \to \triangle^d$ denote a generalization of the softmax function to include $\boldsymbol{\rho}$, with $\sigma_{\boldsymbol{\rho}}(\boldsymbol{\delta})_i$ equal to the right hand side of Equation (13).

Theorem 3.1. *Let $K_n(u) = \sigma_{\boldsymbol{\rho}}^{-1}(C_n(u))$, then*

$$\mathbb{P}[\boldsymbol{\delta} \in K_n(u) \textit{ for all } n \in \mathbb{N}] \geq 1 - u \tag{14}$$

The proof is a direct consequence of Theorem 2.4. The following corollary yields simultaneous confidence sequences for all contrasts.

Corollary 3.2. *Let $K_n(u) = \sigma_{\boldsymbol{\rho}}^{-1}(C_n(u))$ and $\mathcal{A}^d = \{\boldsymbol{a} \in \mathbb{R}^d : \sum_i a_i = 0\}$ denote the set of all d-dimensional contrasts. For all $\boldsymbol{a} \in \mathcal{A}^d$ define*

$$l^+_{n,\boldsymbol{a}}(u) = \sup\{\sum_i a_i \delta_i : \boldsymbol{\delta} \in K_n(u)\} \textit{ and}$$
$$l^-_{n,\boldsymbol{a}}(u) = \inf\{\sum_i a_i \delta_i : \boldsymbol{\delta} \in K_n(u)\}.$$

Then

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\forall \boldsymbol{a} \in \mathcal{A}^d : \sum_i a_i \delta_i \in \bigcap_{n=1}^{\infty} [l^-_{n,a}(u), l^+_{n,a}(u)]\right) \geq 1 - u.$$

The upper bound $l^+_{n,a}(u)$ is the solution to the convex optimization

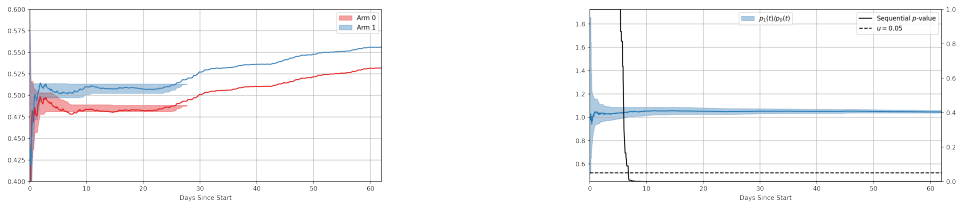
$$\begin{aligned} \max \quad & \sum_i a_i \delta_i \\ \text{s.t.} \quad & c \leq \sum_i S_i^n \left(\delta_i + \log \rho_i - \log \sum_j \rho_j e^{\delta_j}\right), \end{aligned} \tag{15}$$

where $c = \log \mathrm{Beta}(\boldsymbol{\alpha}_0 + \boldsymbol{S}_n) - \log \mathrm{Beta}(\boldsymbol{\alpha}_0) + \log u$. Convexity follows from the log-sum-exponential function. The lower bound $l^-_{n,a}(u)$ is the solution to the corresponding minimization problem. This is visualized in Figure 2.

A hypothesis can be rejected at the u level as soon as the set that it defines fails to intersect with the confidence set $K_n(u)$. Note that $K_n(u_1) \subset K_n(u_2)$ for $u_1 > u_2$. To obtain a sequential p-value, we seek the largest u such that the null is not rejected. That is, we seek the smallest set $K_n(u)$ such that there is a non-empty intersection with the subset of $\mathbb{R}^d$ defined by the hypothesis. This too can be achieved by a convex optimization program. One can simply maximize u over the feasible set defined by the intersection of the $K_n(u)$ and the set defined by the hypothesis. Suppose one wishes to test the hypothesis $\delta_0 \geq \delta_1$ and $\delta_0 \geq \delta_2$. The sequential p-value at time n is the inverse of the solution to the following convex program

$$\begin{aligned} \min \quad & q \\ \text{s.t.} \quad & c \leq \log(q) + \sum_i S_i^n \left(\delta_i + \log \rho_i - \log \sum_j \rho_j e^{\delta_j}\right) \\ & \delta_0 \geq \delta_1 \\ & \delta_0 \geq \delta_2 \end{aligned} \tag{16}$$

where $c = \log \mathrm{Beta}(\boldsymbol{\alpha}_0 + \boldsymbol{S}_n) - \log \mathrm{Beta}(\boldsymbol{\alpha}_0)$. A simulation study presenting sequential p-values and confidence sequences on $\boldsymbol{\delta}$ is given in Appendix A.11



(a) (b)

Figure 2: (a) Assuming p_1 and p_0 are stationary and estimated using multinomial confidence sequences from from Corollary 2.5. (b) Assuming $p_1(t)$ and $p_0(t)$ are dynamic but with $p_1(t)/p_0(t) = \exp(\delta_1 - \delta_0)$ and estimating this quotient using the confidence sequences from equation Corollary 3.2. (Left axis) Confidence sequences are visualized with shaded regions and MLE estimates are visualized with solid lines. (Right axis) sequential p-value.

3.2 Application: Conversion Rate Optimization

Many modern experiments are designed to increase conversion rate and drive additional revenue. The term *conversion* is used to describe a Bernoulli trial such as user signup or purchase. The typical setup is that there are d different *experiences* (such as different versions of a signup page), and each experience i is assumed to have some conversion probability p_i. When a new user visits the experience, they either convert or they do not, and a successful or unsuccessful conversion event is logged. One might think that it would be easy to estimate p_i sequentially using the multinomial confidence sequences of section 2 and the counts of successful and unsuccessful conversions of arm i. However, one soon runs into problems, as illustrated below.

The following case study comes from a signup funnel experiment at Netflix. It was hypothesized that adding additional methods of payments might increase the overall number of conversions. Alternatively, it could also be plausible that additional payment methods may simply cannibalize conversions from other payment methods, not increasing the overall number of conversions at all. An A/B test was created to test these hypotheses. The control group received the standard set of payment methods, while the treatment group were offered the standard set plus additional payment methods. To begin, let's first incorrectly assume that the success probabilities are constant. Figure 2 shows the application of the multinomial confidence sequences from Corollary 2.5 to separately estimate the conversion probabilities for arms 0 and 1. Note that the running intersection of anytime-valid confidence intervals becomes the empty set, and the MLE exits the confidence sequence. An empty intersection would be a rare event (with probability less than α) if the assumption of constant probabilities in Section 2 were true. Instead, it indicates that the conversion probabilities are not constant but time-varying, invalidating a commonly made assumption in conversion rate experimentation.

Many online conversion experiments share similarities with Wald's gun-shooting example. The success probability of a conversion is often time-varying due to external factors such as the day of the week, recent product launches, or new promotions. These external factors likely affect all arms of the experiment equally, so it is reasonable to expect the time variation to be common to all. Moreover, it is often the case in these experiments that an unsuccessful signup (a 0 or "fail" Bernoulli outcome) is not directly observed. When a successful signup occurs, an event is typically logged, but if no signup occurs, then an unsuccessful signup must be inferred from the absence of a successful signup (typically through an ad-hoc definition such as "an experimental unit failed to convert within T units of time since allocation"). This definition of conversion can make the analysis sensitive to the arbitrary choice of time window, and can present challenges to a sequential analysis if one must wait T units of time before the first Bernoulli outcome is realized.

Instead, we propose using the assumptions of section 3, which only requires the counts of successful signups. The assumption of constant conversion probabilities is relaxed to a constant multiplicative difference between them, which is typically more realistic in practice. The confidence sequence on the multiplicative constant is shown in 2 (b). A winning arm can be declared in approximately one

week instead of nine, toward which all subsequent visitors can be directed, dramatically increasing signups relative to the fixed-n experiment.

4 Inhomogeneous Poisson Counting Processes

4.1 Theory

A counting process is a stochastic process $\{N(t) : t \geq 0\}$ satisfying $N(0) = 0$, $N(t) \in \mathbb{N}_0$ and $N(s) \leq N(t)$ for $s \leq t$. The inhomogeneous Poisson counting process is defined by an *intensity function* $\lambda : \mathbb{R} \to \mathbb{R}^+$ that is locally integrable, $\int_B \lambda(t)dt \leq \infty$ for all bounded Borel measurable sets $B \in \mathbb{R}$, defining a measure $\Lambda(B) = \int_B \lambda(t)dt$ [Kingman, 1992]. For any collection of disjoint Borel measurable sets $B_1, B_2, \ldots$ the inhomogeneous Poisson counting process has the property that $N(B_i)$ are independent $\mathrm{Poisson}(\Lambda(B_i))$ random variables. The inhomogeneous Poisson counting process can be defined in terms of an inhomogeneous Poisson point process on the real line by simply counting the number of points in a set. For our applications, these points correspond to times of events. At any time t the probability density of the time-difference s to the next point is given by $g(s) = \lambda(t+s)\exp(-\int_0^s \lambda(t+s)ds)$. The independent increments property implies a *memoryless* property of the process: the counts in the next time interval or waiting time until the next point are independent of the process history. This is a necessary property to establish the following theorem.

Theorem 4.1. *Consider d inhomogeneous Poisson point processes indexed by $i \in \{1, \ldots, d\}$ where process i has intensity functions $\lambda_i(t) = \rho_i e^{\delta_i}\lambda(t)$. Let each point produced by process i be marked with the corresponding process index i. At any time t, such as immediately after the previous point, the probability that the next point has mark i is given by*

$$\theta_i = \frac{\rho_i e^{\delta_i}}{\sum_{j=1}^d \rho_j e^{\delta_j}}. \tag{17}$$

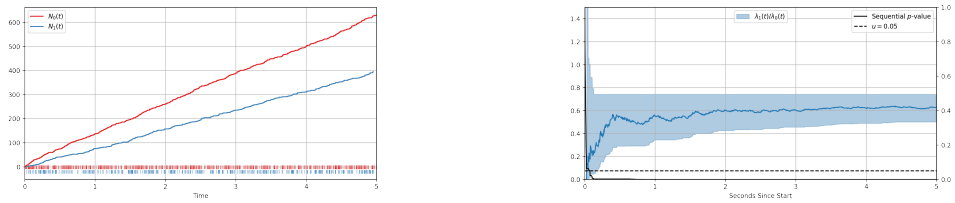
This gives the probability that the next point in time is from process i. The proof is given in the Appendix A.10. Theorem 4.1 states that the sequence of marks can be considered a sequence of $\mathrm{Multinomial}(1, \boldsymbol{\theta})$ random variables, allowing the sequential multinomial test to perform inference on $\boldsymbol{\delta}$. For example, a sequential test of equality among d time-inhomogeneous Poisson point processes ($\lambda_i(t) = \lambda_j(t)$ for all pairs i and j) can be obtained from the sequential multinomial test of the hypothesis $\boldsymbol{\theta}_0 = (1/d, \ldots, 1/d)$. Once again, the total counts in each arm is a statistic that is independent of the time-varying nuisance parameter $\lambda(t)$. A simulation study presenting sequential p-values and confidence sequences on $\boldsymbol{\delta}$ is given in Appendix A.12

4.2 Application: Software Canary Testing

When continuously deploying new software to users, engineers often adopt the practice of *canary testing* [Lindon et al., 2022, Schermann et al., 2018]. A canary test is a controlled experiment in which users are randomly assigned to the current software or a newer release candidate. The experimenter's goal is to study the performance of the release candidate in a production environment before releasing it globally, essentially acting as a quality control gate before full deployment. If the release candidate performs significantly worse, it is blocked, and developers must resolve the offending issues. This strategy helps to prevent bugs from reaching all users. However, performance regressions are still experienced by the users in the experiment. Reducing the risk for these users motivates the experiment's stoppage as soon as a performance regression is detected. Optional stopping requires performance to be measured in real-time and requires sequential testing methodology.

Performance regressions are measured in terms of the rates of *events*. These events could correspond to errors, failures, or any occurrence of interest sent to a central logging service by each instance of the software. The data naturally forms a marked 1-dimensional point process in time, recording the timestamp and type of event. The instantaneous rate is expected to be time-varying due to varying traffic and usage patterns.

The following example is taken from a canary test Netflix. This acts as a quality control gate in the software delivery process when releasing new versions of the client application. A regression that would have affected approximately 60% of all devices was detected with this methodology in less than one second. *Successful play starts* (SPS) are carefully monitored in this experiment between



(a) (b)

Figure 3: (a) Ruglplot shows the timestamps of SPS events being received, while solid lines show the counting processes for arms 0 and 1. (b) (Left axis) Shaded region shows the confidence sequence for the quotient of inhomogeneous Poisson process intensity functions $\lambda_1(t)/\lambda_0(t)$ obtained from Corollary 3.2. Blue solid line shows the MLE of $\lambda_1(t)/\lambda_0(t) = e^{\delta_1 - \delta_0}$. (Right axis) Black solid line shows sequential p-value.

the existing software and the release candidate. An SPS event is sent by the streaming application to the central logging system whenever a requested title successfully begins playback. If significantly fewer SPS events are being received from the treatment group running the release candidate, then it indicates there is an issue with the new software preventing some streams from starting. Figure 3 shows that the confidence sequence on $\lambda_1(t)/\lambda_0(t)$ falls below 1.0 in less than a second, indicating that the instantaneous rate of SPS events for arm 1 is less than the instantaneous rate of SPS events for arm 0. In this case, the canary test was aborted and the offending bug was identified, preventing a serious regression from being released to all users.

5 Conclusion

The contributions of this paper provide an "anytime-valid" approach to inference in several important applications dealing with count data. The anytime-valid guarantees permit experiments to be continuously monitored and enables optional stopping, which can greatly reduce both the time required for experiments to complete and the potential harm to units in the experiment. We first introduced a sequential test for Multinomial hypotheses using a mixture martingale construction, which has already proven to be an effective solution for rapidly detecting sample ratio mismatches in online controlled experiments. We then used this result to develop a sequential test of equality and contrasts in inhomogeneous Bernoulli processes, which has practically demonstrated dramatic speedups in decision-making in conversion experiments. Contrary to many widely used models, our approach does not assume that the conversion probabilities are constant, which is often violated in the real world. Moreover, only successful conversions, such as signups, are observed in practice. This is different from Bernoulli outcome models in which it is assumed that both successes and failures are observed. Our proposed sequential test enjoys the added convenience of only requiring successful Bernoulli outcomes to be observed. Lastly, we used the sequential Multinomial test to develop a sequential test for equality and contrasts in time-inhomogeneous Poisson counting processes. These play an important role in the monitoring of systems, such as data pipelines and software usage.

The confidence sequences provided in this paper allow inference to be made at any time and data-dependent stopping rules for ending an experiment. This in of itself can dramatically speed up the time to reach conclusions with valid statistical guarantees. An obvious extension to this work is to combine these confidence sequences with a strategy for adapting the assignment probabilities with the goal of assigning fewer units to suboptimal arms. The confidence sequences presented here can be used to construct an adaptive algorithm that identifies the best arm with tunably high probability through the least upper confidence bound (LUCB) algorithm, as has been done successfully in Howard and Ramdas [2022].

References

Michael Lindon, Chris Sanden, and Vaché Shirikian. Rapid regression detection in software deployments through sequential testing. In *Proceedings of the 28th ACM SIGKDD Conference on*

Knowledge Discovery and Data Mining, KDD '22, page 3336–3346, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3539099. URL https://doi.org/10.1145/3534678.3539099.

Lynn Kuo and Tae Young Yang. Bayesian computation for nonhomogeneous poisson processes in software reliability. *Journal of the American Statistical Association*, 91(434):763–773, 1996. ISSN 01621459. URL http://www.jstor.org/stable/2291671.

Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test martingales, bayes factors and p -values. *Statist. Sci.*, 26(1):84–101, 02 2011. doi: 10.1214/10-STS347. URL https://doi.org/10.1214/10-STS347.

Ian Waudby-Smith and Aaditya Ramdas. Confidence sequences for sampling without replacement. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20204–20214. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/e96c7de8f6390b1e6c71556e4e0a4959-Paper.pdf.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055 – 1080, 2021. doi: 10.1214/20-AOS1991. URL https://doi.org/10.1214/20-AOS1991.

Emilie Kaufmann and Wouter M. Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 22(246):1–44, 2021. URL http://jmlr.org/papers/v22/18-798.html.

A. Wald. Sequential tests of statistical hypotheses. *Ann. Math. Statist.*, 16(2):117–186, 06 1945. doi: 10.1214/aoms/1177731118. URL https://doi.org/10.1214/aoms/1177731118.

Harold Jeffreys. Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(2):203–222, 1935. doi: 10.1017/S030500410001330X.

Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995. ISSN 01621459. URL http://www.jstor.org/stable/2291091.

Peter Grünwald, Rianne de Heide, and Wouter Koolen. Safe testing, 2021.

Muriel Felipe Pérez-Ortiz, Tyron Lardy, Rianne de Heide, and Peter Grünwald. E-statistics, group invariance and anytime valid testing, 2022. URL https://arxiv.org/abs/2208.07610.

Allard Hendriksen, Rianne de Heide, and Peter Grünwald. Optional Stopping with Bayes Factors: A Categorization and Extension of Folklore Results, with an Application to Invariant Situations. *Bayesian Analysis*, 16(3):961 – 989, 2021. doi: 10.1214/20-BA1234. URL https://doi.org/10.1214/20-BA1234.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference, 2022. URL https://arxiv.org/abs/2210.01948.

A. Wald and J. Wolfowitz. Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, 19(3):326–339, 1948. ISSN 00034851. URL http://www.jstor.org/stable/2235638.

I. J. Good. A bayesian significance test for multinomial distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 29(3):399–431, 1967. ISSN 00359246. URL http://www.jstor.org/stable/2984384.

Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(2):407–431, 2021. doi: https://doi.org/10.1111/rssa.12647. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssa.12647.

Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. Always valid inference: Continuous monitoring of a/b tests. *Operations Research*, 2021. doi: 10.1287/opre.2021.2135. URL https://doi.org/10.1287/opre.2021.2135.

Herbert Robbins. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409, 1970. ISSN 00034851. URL http://www.jstor.org/stable/2239848.

Eric-Jan Wagenmakers, Quentin F. Gronau, Fabian Dablander, and Alexander Etz. The support interval. *Erkenntnis*, 87(2):589–601, Apr 2022. ISSN 1572-8420. doi: 10.1007/s10670-019-00209-z. URL https://doi.org/10.1007/s10670-019-00209-z.

James M. Dickey. The Weighted Likelihood Ratio, Linear Hypotheses on Normal Location Parameters. *The Annals of Mathematical Statistics*, 42(1):204 – 223, 1971. doi: 10.1214/aoms/1177693507. URL https://doi.org/10.1214/aoms/1177693507.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

Diane Tang, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer. Overlapping experiment infrastructure: More, better, faster experimentation. In *Proceedings 16th Conference on Knowledge Discovery and Data Mining*, pages 17–26, Washington, DC, 2010.

Zhenyu Zhao, Miao Chen, Don Matheson, and Maria Stone. Online experimentation diagnosis and troubleshooting beyond aa validation. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 498–507, 2016. doi: 10.1109/DSAA.2016.61.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 12 1976. ISSN 0006-3444. doi: 10.1093/biomet/63.3.581. URL https://doi.org/10.1093/biomet/63.3.581.

Aleksander Fabijan, Jayant Gupchup, Somit Gupta, Jeff Omhover, Wen Qin, Lukas Vermeer, and Pavel Dmitriev. Diagnosing sample ratio mismatch in online controlled experiments: A taxonomy and rules of thumb for practitioners. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2156–2164, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016. doi: 10.1145/3292500.3330722. URL https://doi.org/10.1145/3292500.3330722.

Nanyu Chen, Min Liu, and Ya Xu. Automatic detection and diagnosis of biased online experiments, 2018.

A. Wald. *Sequential analysis*. J. Wiley & sons, Incorporated, 1947. URL https://books.google.com/books?id=0nREAAAAIAAJ.

J.F.C. Kingman. *Poisson Processes*. Oxford Studies in Probability. Clarendon Press, 1992. ISBN 9780191591242. URL https://books.google.com/books?id=VEiM-OtwDHkC.

Gerald Schermann, Jürgen Cito, Philipp Leitner, Uwe Zdun, and Harald C. Gall. We're doing it live: A multi-method empirical study on continuous experimentation. *Information and Software Technology*, 99:41–57, 2018. ISSN 0950-5849. doi: https://doi.org/10.1016/j.infsof.2018.02.010. URL https://www.sciencedirect.com/science/article/pii/S0950584917302136.

Steven R. Howard and Aaditya Ramdas. Sequential estimation of quantiles with applications to A/B testing and best-arm identification. *Bernoulli*, 28(3):1704 – 1728, 2022. doi: 10.3150/21-BEJ1388. URL https://doi.org/10.3150/21-BEJ1388.

P. Armitage. *Interim Analyses in Clinical Trials*. Multiple Comparisons, Selection and Applications in Biometry. CRC Press, 1993. ISBN 9780824788957.

Jerzy Neyman, Egon Sharpe Pearson, and Karl Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933. doi: 10.1098/rsta.1933.0009. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsta.1933.0009.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527 – 535, 1952. doi: bams/1183517370. URL https://doi.org/.

G. Casella and R.L. Berger. *Statistical Inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning, 2002. ISBN 9780534243128. URL https://books.google.com/books?id=0x_vAAAAMAAJ.

F. J. Anscombe. Fixed-sample-size analysis of sequential observations. *Biometrics*, 10(1):89–100, 1954. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/3001665.

Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. Peeking at a/b tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 1517–1525, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874. doi: 10.1145/3097983.3097992. URL https://doi.org/10.1145/3097983.3097992.

P. Armitage, C. K. McPherson, and B. C. Rowe. Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A (General)*, 132(2):235–244, 1969. ISSN 00359238. URL http://www.jstor.org/stable/2343787.

Joseph B. Kadane, Mark J. Schervish, and Teddy Seidenfeld. Reasoning to a foregone conclusion. *Journal of the American Statistical Association*, 91(435):1228–1235, 1996. ISSN 01621459. URL http://www.jstor.org/stable/2291741.

Jean Ville. *Étude critique de la notion de collectif*. 1939. URL http://eudml.org/doc/192893.

James O. Berger, Lawrence D. Brown, and Robert L. Wolpert. A unified conditional frequentist and bayesian test for fixed and sequential simple hypothesis testing. *Ann. Statist.*, 22(4):1787–1807, 12 1994. doi: 10.1214/aos/1176325757. URL https://doi.org/10.1214/aos/1176325757.

J. O. Berger, B. Boukai, and Y. Wang. Unified frequentist and bayesian testing of a precise hypothesis. *Statist. Sci.*, 12(3):133–160, 09 1997. doi: 10.1214/ss/1030037904. URL https://doi.org/10.1214/ss/1030037904.

James O. Berger, Benzion Boukai, and Yinping Wang. Simultaneous bayesian-frequentist sequential testing of nested hypotheses. *Biometrika*, 86(1):79–92, 1999. ISSN 00063444. URL http://www.jstor.org/stable/2673538.

Jerome Cornfield. A bayesian test of some classical hypotheses, with applications to sequential clinical trials. *Journal of the American Statistical Association*, 61(315):577–594, 1966. ISSN 01621459. URL http://www.jstor.org/stable/2282772.

D. A. Darling and Herbert Robbins. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences*, 58(1):66–68, 1967. ISSN 0027-8424. doi: 10.1073/pnas.58.1.66. URL https://www.pnas.org/content/58/1/66.

Ian Waudby-Smith, David Arbour, Ritwik Sinha, Edward H. Kennedy, and Aaditya Ramdas. Doubly robust confidence sequences for sequential causal inference, 2021.

C. Jennison and B.W. Turnbull. *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC Interdisciplinary Statistics. CRC Press, 1999. ISBN 9781584888581. URL https://books.google.com/books?id=qBrpTcAYtNQC.

Ivair R Silva and Martin Kulldorff. Continuous versus group sequential analysis for post-market drug and vaccine safety surveillance. *Biometrics*, 71(3):851–858, 2015.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17(none):257 – 317, 2020. doi: 10.1214/18-PS321. URL https://doi.org/10.1214/18-PS321.

P. A. W. Lewis and G. S. Shedler. Simulation of nonhomogeneous poisson processes with degree-two exponential polynomial rate function. *Operations Research*, 27(5):1026–1040, 1979. doi: 10.1287/opre.27.5.1026. URL https://doi.org/10.1287/opre.27.5.1026.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Our work is motivated by three important sequential applications in online experimentation. Testing proper randomization via sample ratio mismatch testing. Testing hypotheses among time-varying probabilities in conversion rate optimization and testing hypotheses among time-varying poisson process intensities in software canary experiments. We have provided viable sequential testing solutions to all of three of these applications via sequential multinomial tests, sequential tests of inhomogeneous Bernoulli processes, and sequential tests of time-inhomogeneous Poisson counting processes.
   (b) Did you describe the limitations of your work? [Yes] We have explained in the introductory section that we do not study the theoretical properties such as optimality our proposed tests. Our tests are specific instances of a generic construction known as the method of mixture martingales and mixture sequential probability ratio tests. The optimality properties of these tests are well established and we have linked to the relevant literature. Our work focuses on providing specific solutions to some important applications using these constructions.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] All proofs include the assumptions, and we have even presented examples of what can go wrong when these assumptions do not hold - see the application of multinomial confidence sequences when the probabilities are time varying in section 3
   (b) Did you include complete proofs of all theoretical results? [Yes] All proofs are in the appendix
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We have several simulations in the appendix, which all include clear instructions on how they were created
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] We include two real-world case studies of how this methodology is employed with practical value at a leading internet streaming company. A draft of this paper has been reviewed internally by said company and cleared for submission
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

A.1 Limitations of Fixed-n Testing

To quote Armitage [1993], "the classical theory of experimental design deals predominantly with experiments of predetermined size, presumably because the pioneers of the subject, particularly R. A. Fisher, worked in agricultural research, where the outcome of a field trial is not available until long after the experiment has been designed and started.". The author points out that many popular statistical tests are of the *fixed-n* or *fixed-horizon* kind, which operate just once on a complete dataset in a Neyman-Pearson type testing framework [Neyman et al., 1933]. The early development of these tests was driven by the wide variety of applications in which all observations arrive at the same time or when the experimenter is simply handed a complete dataset [Robbins, 1952]. Following this mode of one-time statistical analysis, these tests have been specifically optimized to maximize power at analysis time subject to a Type I error configuration. The classical solution to minimizing experiment cost is then to find a statistical test that provides the same Type I/II error guarantees at a smaller sample size, such as seeking the uniformly most powerful unbiased test within a particular class [Casella and Berger, 2002]. In many modern applications, however, data typically arrive in streams rather than in sets; observations often arrive in a sequence instead of simultaneously. Therefore, it makes sense that statistical tests optimized for a one-time analysis of a complete collection of observations may not be optimal in experiments where observations arrive sequentially.

There are practical difficulties in using fixed-n tests in experiments where observations arrive sequentially. Consider these problems first from the perspective of hypothesis *testing*. The biggest drawback of using a fixed-n test in a sequential application is that it can only be performed *once*. Fixed-n tests are fine in applications where all observations arrive simultaneously or when the experimenter is handed a complete dataset, as there is only one possible opportunity to perform the test. However, if observations arrive in a sequence, the experimenter is presented with many opportunities to perform the test. Perform the test too early, and the Type II error probability will be high, resulting in many small effects being undetected. Perform the test too late, and the experiment may be more costly than is strictly necessary. Sample size calculations also fail to remedy these issues for the following reasons. Closed-form sample size expressions may not exist beyond trivial textbook models, the required inputs are frequently unknown, and most problematically, sample size calculations require the specification of a minimum detectable effect (MDE). The problem with specifying an MDE is that the experimenter, unwilling to sacrifice power even for small effect sizes, typically specifies these to be conservatively low, resulting in quadratic growth of the required sample size and a more costly experiment than strictly necessary (in particular, relative to a sequential design). For instance, consider one-sided "no-harm" testing applications where the goal is to detect (possibly small) adverse effects. Specifying a small MDE causes the required sample size to be large, resulting in large adverse effects remaining undetected for lengthy amounts of time and prolonged harm to the experimental units.

The latter example highlights a further practical difficulty with using fixed-n tests in sequential applications from the *estimation* perspective. The experimenter is often curious about the current performance of each arm, stemming from the concern that an arm might have a substantial adverse effect on those assigned to it. To address this concern, the experimenter might wish to estimate the effect by computing a confidence interval. Unfortunately, the $1 - u$ confidence statement obtained by inverting an u-level fixed-n test only holds at a fixed-n: it is only a one-time guarantee. The experimenter cannot hope to "monitor" the effect of each arm by computing multiple fixed-n confidence intervals spaced out over different times, as their intersection does not have any coverage guarantee.

Instead of stopping the experiment at a predetermined sample size, it is more natural and useful for the stopping rule to be data-dependent in sequential applications. That is, to perform *optional stopping*. Stopping a test based on whether the data observed so far contains strong evidence for or against a hypothesis removes the need to perform a troublesome sample size calculation and allows the experiment to be terminated adaptively. If conclusions seem unclear at a chosen analysis time, such as confidence statements being insufficiently tight, then allowing the test to run for longer is also useful. That is, to perform *optional continuation*. Unfortunately, the Type I error and confidence guarantees from fixed-n tests are not preserved under *optional stopping or continuation*.

Many experimenters do not specify a fixed sample size in advance simply because they have not made up their minds about the requirements of the experiment or the available resources [Anscombe, 1954]. This can lead to an invalid practice of "peeking" where a fixed-n test is used to define a stopping rule, and estimands are monitored continuously by repeated fixed-n confidence statements, a procedure that does not possess the statistical guarantees that an experimenter might naively expect [Johari et al., 2017]. Repeated applications of fixed-n tests on accumulating sets of data result in ever-increasing Type I error probabilities [Armitage et al., 1969]. A stopping rule configured to stop sampling when a hypothesis is rejected by a fixed-n test is guaranteed to reject the null, allowing experimenters to sample to a foregone conclusion [Anscombe, 1954, Kadane et al., 1996]. Consequently, the intersection of fixed-n confidence sets is guaranteed to converge to the empty set.

A.2 Solutions via Sequential Testing

Sequential designs remain the preferred form of scientific inquiry by many, so these experimenters would benefit greatly from the development of new statistical tests that support the desired operations of continuous monitoring with optional stopping and continuation. The solution presented here generalizes the frequentist guarantees already familiar to many by extending results to hold *for all n* instead of a *fixed-*n. We compute a *sequential* p*-value* such that the probability of this being less than u *for any* $n \in \mathbb{N}$ is less than u. Similarly, we compute *confidence sequences*: a countable collection of sets such that the probability the estimand is covered *by all* sets, and hence their intersection, is greater than $1 - u$. We provide a review of the sequential testing literature in Appendix A.3.

There are numerous advantages to this approach. Confidence sets and p-values remain valid at all times, which enables experimenters to "check in" and *continuously monitor* the progress of their experiments. The ability to perform optional stopping allows developers to build a layer of automated stopping logic on top of experiments, reducing risk by quickly eliminating poorly-performing arms and terminating as soon as hypotheses have been rejected. This removes the need for human supervision and helps scale the number of experiments performed by automating their orchestration. This approach also appeals to both Bayesians and Frequentists: despite presenting the frequentist properties, it is fundamentally built upon a Bayes factor. Confidence sequences are constructed for the *vector* of parameters for all arms, providing simultaneous confidence sequences for all contrasts among arms in contrast to pairwise comparison tests, which require multiple testing corrections. Lastly, our approach is applicable to situations in which there is time-variability common to all arms, as we construct a test statistic that removes the time-varying components.

A.3 Review of Sequential Testing

The earliest work is often attributed to Ville [1939] with the introduction of a *test martingale*. This object is a nonnegative supermartingale under the null hypothesis, and one can use martingale inequalities to construct sequential designs that control Type I error. Wald [1945] introduced the mixture sequential probability ratio test (mSPRT) for testing composite vs simple null hypotheses. The mSPRT can be viewed as a Bayes factor by interpreting the weight function used to integrate the likelihood ratio as a Bayesian prior over the alternative. Testing in a Bayesian framework via Bayes factors is attributed to Jeffreys [Jeffreys, 1935, Kass and Raftery, 1995]. Proofs of the validity of Bayes factors with nuisance parameters under varied interpretations of optional stopping are provided by Hendriksen et al. [2021]. The use of Bayes factors for sequential testing exists, therefore, both in a purely Bayesian framework from computing posterior probabilities over hypotheses and in Wald's mSPRT framework for obtaining frequentist error probabilities. Similarities and differences between both approaches are discussed at length in Berger et al. [1994, 1997, 1999]. Bayes factors have been used in the design of sequential clinical trials by Cornfield [1966]. Test martingales can be interpreted as Bayes factors and the inverse of the running supremum can be used to construct sequentially valid p-values [Shafer et al., 2011].

Johari et al. [2021] use Wald's mSPRT to construct *anytime-valid* inference for the difference in two Gaussian means with known variance by using the inverse of the running supremum of the mSPRT test martingale to construct a sequential p-value, and use the duality between p-values and confidence sets to construct confidence sequences for the difference. The "anytime-valid" namesake explicitly refers to the fact that this test is safe under optional stopping, in the sense that we may reject the null at the u-level as soon as the sequential p-value falls below u without violating the Type I error guarantees. Similarly, confidence intervals have a $1 - u$ coverage guarantee at any time, allowing

the progress of a statistical test to be continuously monitored and making it robust to the human temptation to peek at results [Johari et al., 2017]. In contrast to our proposal for count data, however, this method requires Gaussian approximations based on central limit theorem arguments and requires a plugin estimators for unknown parameters. Although this method works well in practice, these approximations may not be justified at lower sample sizes, and so the sequential properties of this method may not be strictly guaranteed outside of Gaussian families. Confidence sequences appear as early as Darling and Robbins [1967]. Robbins [1970] showed that it is always possible to disprove a null hypothesis by sequentially collecting data until the null is rejected at the u-level by a fixed-n frequentist test, regardless of the chosen value of u. This result follows as a consequence of the law of iterated logarithm. The anytime-valid approach through the use of sequential p-values and confidence sequences has been greatly extended by Howard et al. [2021], providing univariate nonparametric and nonasymptotic confidence sequences for broad classes of random variables. Confidence sequences for doubly robust causal estimands are presented in Waudby-Smith et al. [2021]. Confidence sequences for sampling without replacement are provided in Waudby-Smith and Ramdas [2020].

Between fixed and anytime-valid/sequential testing are group sequential testing (GST) methods [Jennison and Turnbull, 1999]. In GST a finite and fixed number of analyses are planned as part of the design and are performed upon reaching the pre-specified sequence of sample sizes. There is the opportunity to reject the null hypothesis or continue to the next round at each analysis. GST only partially solves the optional stopping requirement and fails to solve the optional continuation requirement. Optional stopping is only partially solved because analyses can occur at pre-determined sample sizes, when practically the requirement is to perform analyses at pre-determined *times*. Suppose an experimenter wishes to perform analyses every day for a month. Due to varying traffic, there is no guarantee that the pre-determined sample sizes align with every day of the month. Optional continuation is not permitted as one loses the ability to collect more data beyond the final analysis. Silva and Kulldorff [2015] show that for every group sequential test, there exists a fully sequential test that is uniformly better in its ability to stop sooner.

A.4 Derivation of Equation (3) (Bayes Factor)

The Bayes factor is defined as

$$BF_{01}(x_{1:n}) = \frac{p(\boldsymbol{x}_{1:n}|M_1)}{p(\boldsymbol{x}_{1:n}|M_0)} = \frac{\int p(\boldsymbol{x}_{1:n}|\boldsymbol{\theta}, M_1)p(\boldsymbol{\theta}|M_1)d\boldsymbol{\theta}}{p(\boldsymbol{x}_{1:n}|M_0)}. \tag{18}$$

Under the assumptions for M_0 and M_1 expressed in Equation (1) and Equation (2), we can show

$$\begin{aligned}
p(\boldsymbol{x}_{1:n}|M_1) =& \int p(\boldsymbol{x}_{1:n}|\boldsymbol{\theta}, M_1)p(\boldsymbol{\theta}|M_1)d\boldsymbol{\theta} \\
=& \int \frac{\Gamma\left(\sum_{ij} x_{i,j}+1\right)}{\prod_j \Gamma\left(\sum_i x_{i,j}+1\right)} \prod_j \theta_j^{\sum_i x_{i,j}} \frac{\Gamma\left(\sum_j \alpha_{0,j}\right)}{\prod_j \Gamma(\alpha_{0,j})} \prod_j \theta_j^{\alpha_{0,j}-1} d\boldsymbol{\theta} \\
=& \frac{\Gamma\left(\sum_{ij} x_{i,j}+1\right)}{\prod_j \Gamma\left(\sum_i x_{i,j}+1\right)} \frac{\Gamma\left(\sum_j \alpha_{0,j}\right)}{\prod_j \Gamma(\alpha_{0,j})} \int \prod_j \theta_j^{\sum_i x_{i,j}+\alpha_{0,j}-1} d\boldsymbol{\theta} \\
=& \frac{\Gamma\left(\sum_{ij} x_{i,j}+1\right)}{\prod_j \Gamma\left(\sum_i x_{i,j}+1\right)} \frac{\Gamma\left(\sum_j \alpha_{0,j}\right)}{\prod_j \Gamma(\alpha_{0,j})} \frac{\prod_j \Gamma\left(\sum_i x_{i,j}+\alpha_{0,j}\right)}{\Gamma\left(\sum_{ij} x_{i,j}+\sum_j \alpha_{0,j}\right)} \text{ and} \\
p(\boldsymbol{x}_{1:n}|M_0) =& \frac{\Gamma\left(\sum_{ij} x_{i,j}+1\right)}{\prod_j \Gamma\left(\sum_i x_{i,j}+1\right)} \prod_j \theta_{0,j}^{\sum_i x_{i,j}}.
\end{aligned} \tag{19}$$

The result follows from cancelling terms in numerator and denominator:

$$BF_{01}(x_{1:n}) = \frac{p(\boldsymbol{x}_{1:n}|M_1)}{p(\boldsymbol{x}_{1:n}|M_0)} = \frac{\Gamma(\sum_{j=1}^d \alpha_{0,j})}{\Gamma(\sum_{j=1}^d \alpha_{0,j} + \sum_{i=1}^n x_{i,j})} \frac{\prod_{j=1}^d \Gamma(\alpha_{0,j} + \sum_{i=1}^n x_{i,j})}{\prod_{j=1}^d \Gamma(\alpha_{0,j})} \frac{1}{\prod_{j=1}^d \theta_{0,j}^{\sum_{i=1}^n x_{i,j}}}. \tag{20}$$

A.5 Derivation of Equation (4) (Sequential Posterior Odds Updating)

The posterior odds in favor of M_1 to M_0 after observing $\boldsymbol{x}_{1:n}$ is defined as

$$\begin{aligned}
\frac{p(M_1|\boldsymbol{x}_{1:n})}{p(M_0|\boldsymbol{x}_{1:n})} &= \frac{\int p(\boldsymbol{x}_{1:n}|\boldsymbol{\theta}, M_1)p(\boldsymbol{\theta}, M_1)d\boldsymbol{\theta}}{p(\boldsymbol{x}_{1:n}|M_0)} \frac{p(M_1)}{p(M_0)}, \\
&= \frac{p(\boldsymbol{x}_{1:n}|M_1)}{p(\boldsymbol{x}_{1:n}|M_0)} \frac{p(M_1)}{p(M_0)}, \\
&= \frac{\prod_{i=1}^n p(\boldsymbol{x}_i|\boldsymbol{x}_{1:i-1}|M_1)}{\prod_{i=1}^n p(\boldsymbol{x}_i|\boldsymbol{x}_{1:i-1}|M_0)} \frac{p(M_1)}{p(M_0)}, \\
&= \frac{p(\boldsymbol{x}_n|\boldsymbol{x}_{1:n-1}, M_1)}{p(\boldsymbol{x}_n|\boldsymbol{x}_{1:n-1}, M_0)} \frac{p(M_1|\boldsymbol{x}_{1:n-1})}{p(M_0|\boldsymbol{x}_{1:n-1})}, \\
&= \frac{\int p(\boldsymbol{x}_n|\boldsymbol{\theta}, \boldsymbol{x}_{1:n-1}, M_1)p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n-1}, M_1)d\boldsymbol{\theta}}{p(\boldsymbol{x}_n|\boldsymbol{x}_{1:n-1}, M_0)} \frac{p(M_1|\boldsymbol{x}_{1:n-1})}{p(M_0|\boldsymbol{x}_{1:n-1})},
\end{aligned}$$

where the last expression stresses the recursive definition of the posterior odds factor in terms of products of posterior predictive densities. The posterior distribution of $\boldsymbol{\theta}|\boldsymbol{x}_{1:n}, M_1 \sim \text{Dirichlet}(\boldsymbol{\alpha}_n)$ where $\boldsymbol{\alpha}_n = \boldsymbol{\alpha}_{n-1} + \boldsymbol{x}_n$ with $\boldsymbol{\alpha}_0$ the initial prior parameter choice. The posterior predictive densities are easily computed as

$$p(\boldsymbol{x}_n|\boldsymbol{x}_{1:n-1}, M_1) = \frac{\Gamma(\sum_i x_{n,i} + 1)}{\prod_i \Gamma(x_{n,i} + 1)} \frac{\Gamma(\sum_i \alpha_{n-1,i})}{\prod_i \Gamma(\alpha_{n-1,i})} \frac{\prod_i \Gamma(\alpha_{n-1,i} + x_{n,i})}{\Gamma(\sum_i \alpha_{n-1,i} + x_{n,i})}$$

and

$$p(\boldsymbol{x}_n|\boldsymbol{x}_{1:n-1}, M_0) = \frac{\Gamma(\sum_i x_{n,i} + 1)}{\prod_i \Gamma(x_{n,i} + 1)} \prod \theta_{0,i}^{x_{n,i}}.$$

It follows that

$$\frac{p(M_1|\boldsymbol{x}_{1:n})}{p(M_0|\boldsymbol{x}_{1:n})} = \frac{\Gamma(\sum_i \alpha_{n-1,i})}{\Gamma(\sum_i \alpha_{n-1,i} + x_{n,i})} \frac{\prod_i \Gamma(\alpha_{n-1,i} + x_{n,i})}{\prod_i \Gamma(\alpha_{n-1,i})} \frac{1}{\prod_i \theta_{0,i}^{x_{n,i}}} \frac{p(M_1|\boldsymbol{x}_{1:n-1})}{p(M_0|\boldsymbol{x}_{1:n-1})}, \tag{21}$$

$$\tag{22}$$

where

$$\boldsymbol{\alpha}_n = \boldsymbol{\alpha}_{n-1} + \boldsymbol{x}_n. \tag{23}$$

To see the alternative "betting" form of the posterior odds when $\alpha_{0,i}$ are integer valued, note that $x_{n,i}$ is equal to 1 for a single index $i \in \{1, \ldots, d\}$ and is equal to 0 for the remaining indices. The previous expression can then be written

$$\frac{p(M_1|\boldsymbol{x}_{1:n})}{p(M_0|\boldsymbol{x}_{1:n})} = \frac{\Gamma(\sum_i \alpha_{n-1,i})}{\Gamma(1 + \sum_i \alpha_{n-1,i})} \prod_{i=1}^d \left(\frac{\Gamma(\alpha_{n-1,i} + 1)}{\Gamma(\alpha_{n-1,i})} \right)^{x_{n,i}} \frac{1}{\prod_i \theta_{0,i}^{x_{n,i}}} \frac{p(M_1|\boldsymbol{x}_{1:n-1})}{p(M_0|\boldsymbol{x}_{1:n-1})}, \tag{24}$$

$$= \frac{1}{\sum_i \alpha_{n-1,i}} \prod_{i=1}^d \alpha_{n-1,i}^{x_{n,i}} \frac{1}{\prod_i \theta_{0,i}^{x_{n,i}}} \frac{p(M_1|\boldsymbol{x}_{1:n-1})}{p(M_0|\boldsymbol{x}_{1:n-1})}, \tag{25}$$

$$= \prod_{i=1}^d \left(\frac{\alpha_{n-1,i}}{\sum_j \alpha_{n-1,j}} \right)^{x_{n,i}} \frac{1}{\prod_i \theta_{0,i}^{x_{n,i}}} \frac{p(M_1|\boldsymbol{x}_{1:n-1})}{p(M_0|\boldsymbol{x}_{1:n-1})}, \tag{26}$$

$$\tag{27}$$

where the first term can be recognized as the mean of a Dirichlet distribution with parameter $\boldsymbol{\alpha}_{n-1}$. Tidying up terms yields

$$\frac{p(M_1|\boldsymbol{x}_{1:n})}{p(M_0|\boldsymbol{x}_{1:n})} = \prod_{i=1}^d \left(\frac{\alpha_{n-1,i}}{\theta_{0,i} \sum_j \alpha_{n-1,j}} \right)^{x_{n,i}} \frac{p(M_1|\boldsymbol{x}_{1:n-1})}{p(M_0|\boldsymbol{x}_{1:n-1})}, \tag{28}$$

$$\tag{29}$$

A.6 Proof of Theorem 2.1 (Martingale Property of Posterior Odds)

Proof.

$$
\begin{aligned}
E_{M_0}[O_{n+1}(\boldsymbol{\theta}_0)|\mathcal{F}_n] &= \int \frac{p(\boldsymbol{x}_{n+1}|\boldsymbol{x}_{1:n}, M_1)}{p(\boldsymbol{x}_{n+1}|\boldsymbol{x}_{1:n}, M_0)} O_n(\boldsymbol{\theta}_0) p(\boldsymbol{x}_{n+1}|\boldsymbol{x}_{1:n}, M_0) d\boldsymbol{x}_{n+1} \\
&= O_n(\boldsymbol{\theta}_0) \int p(\boldsymbol{x}_{n+1}|\boldsymbol{x}_{1:n}, M_1) d\boldsymbol{x}_{n+1} \\
&= O_n(\boldsymbol{\theta}_0),
\end{aligned}
$$

where $\mathcal{F}_n = \sigma(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$. □

A.7 Proof of Theorem 2.2 (Construction of Test-Martingale)

First, the following lemma is required

Lemma A.1. *(Ville's Maximal Inequality) If Z_n is a nonnegative supermartingale with respect to the filtration $\mathcal{F}_n$, then*

$$
\mathbb{P}[\exists n \in \mathbb{N}_0 : Z_n \geq u] \leq \frac{Z_0}{u} \tag{30}
$$

Proof. See Ville [1939], Howard et al. [2020] □

Theorem 2.1 shows that $O_n(\boldsymbol{\theta}_0)$ is a nonnegative martingale (and therefore also a supermartingale) under the null hypothesis with initial value $O_0(\boldsymbol{\theta}_0) = 1$. The result then follows immediately from Lemma A.1.

A.8 Proof of Theorem 2.3 (Asymptotic Properties of Bayes Factors)

From Equation (4),

$$
\begin{aligned}
\log O_n(\boldsymbol{\theta}_0) &= \log \mathrm{Beta}(\boldsymbol{\alpha}_0 + \boldsymbol{S}_n) - \log \mathrm{Beta}(\boldsymbol{\alpha}_0) - \sum_i S_i^n \log \boldsymbol{\theta}_{0,i} \\
&= \sum_i \log \Gamma(\alpha_{0,i} + S_i^n) - \log \Gamma(|\boldsymbol{\alpha}_0 + \boldsymbol{S}^n|) + \\
&\quad \log \Gamma(|\boldsymbol{\alpha}_0|) - \sum_i \log \Gamma(\alpha_{0,i}) \\
&\quad - \sum_i S_i^n \log \boldsymbol{\theta}_{0,i}.
\end{aligned} \tag{31}
$$

Using Stirlings approximation $\log \Gamma(z) = z \log z - z + o(\log z)$

$$
\begin{aligned}
\log O_n(\boldsymbol{\theta}_0) &= \sum_i (\alpha_{0,i} + S_i^n) \log(\alpha_{0,i} + S_i^n) - (\alpha_{0,i} + S_i^n) \\
&\quad - (|\boldsymbol{\alpha}_0| + n) \log(|\boldsymbol{\alpha}_0| + n) + (|\boldsymbol{\alpha}_0| + n) + \\
&\quad - \sum_i S_i^n \log \theta_{0,i} + o(\log n) \\
&= \sum_i (\alpha_{0,i} + S_i^n) \log \left(\frac{\alpha_{0,i} + S_i^n}{(|\boldsymbol{\alpha}_0| + n)} \right) \\
&\quad - \sum_i S_i^n \log \theta_{0,i} + o(\log n) \\
&= \sum_i S_i^n \log \left(\frac{\alpha_{0,i} + S_i^n}{(|\boldsymbol{\alpha}_0| + n)} \frac{1}{\theta_{0,i}} \right) + o(\log n) \\
\frac{1}{n} \log O_n(\boldsymbol{\theta}_0) &= \sum_i \frac{S_i^n}{n} \log \left(\frac{\alpha_{0,i} + S_i^n}{(|\boldsymbol{\alpha}_0| + n)} \frac{1}{\theta_{0,i}} \right) + o\left(\frac{\log n}{n} \right).
\end{aligned} \tag{32}
$$

$\frac{S_i^n}{n}$ and $\frac{\alpha_{0,i}+S_i^n}{(|\boldsymbol{\alpha}_0|+n)}$ converge to θ_i almost surely by the strong law of large numbers. It follows that

$$\frac{1}{n}\log O_n(\boldsymbol{\theta}_0) \overset{\text{a.s.}}{\to} \sum_i \theta_i \log\left(\frac{\theta_i}{\theta_{0,i}}\right), \tag{33}$$

by Slutsky's theorem and the continuous mapping theorem, which can be recognized as the Kullback-Leibler divergence of a $\text{Multinomial}(1, \boldsymbol{\theta})$ distribution from a $\text{Multinomial}(1, \boldsymbol{\theta}_0)$ distribution.

A.9 Type I Error Probability Simulation

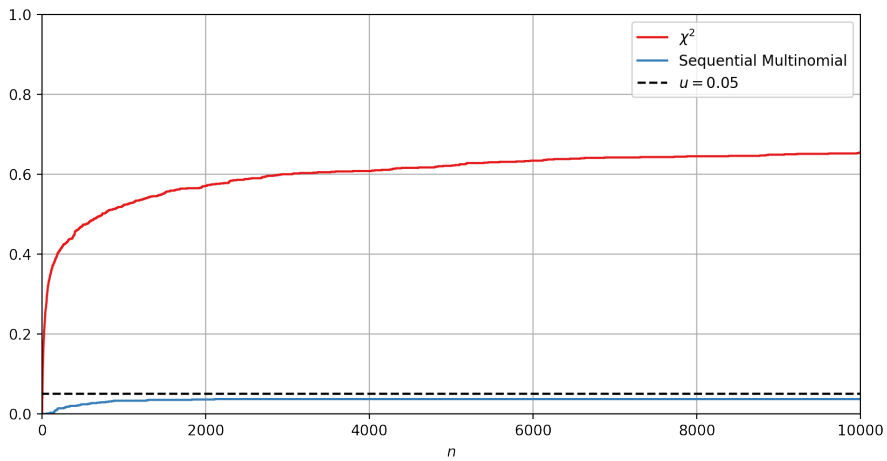


Figure 4: Estimated probability of falsely rejecting the null by sample size n under stopping rules (red) when the χ^2 p-value falls below 0.05 (blue) when the sequential p-value from Equation (8) falls below 0.05. Estimates based on 10000 simulations. Null rejected incorrectly 654 and 37 times by χ^2 and sequential multinomial tests respectively.

A.10 Proof of Theorem 4.1

The following two lemmas are required prove Theorem 4.1. Consider a poisson point process the intensity function $\lambda(t)$. The probability density over the event time t_i conditional on the previous event having arrived at time t_{i-1} is given by

$$p(t_i|t_{i-1}) = \lambda(t_i)e^{-\int_{t_{i-1}}^{t_i}\lambda(s)ds}1_{(t_{i-1},\infty]}(t_i). \tag{34}$$

This is used to prove the following lemma

Lemma A.2. *Let t_{i-1} denote the time of the previously observed event. Suppose the current time is $T > t_{i-1}$. The probability density over the next event time t_i conditional on no observation having occured in $(t_{i-1}, T]$ is given by*

$$p(t_i|t_i > T) = \lambda(t_i)e^{-\int_{T}^{t_i}\lambda(s)ds}1_{(T,\infty]}(t_i). \tag{35}$$

Proof. The probability of no event taking place $(t_{i-1}, T]$ is given by the $\text{Poisson}(\Lambda(t_{i-1}, T])$ distribution

$$\mathbb{P}[N(t_{i-1}, T] = 0] = e^{-\int_{t_{i-1}}^{T}\lambda(s)ds}. \tag{36}$$

The conditional density for t_i given $t_i > T$ is obtained by conditioning on $N(t_{i-1}, T] = 0$ as follows

$$\begin{aligned} p(t_i|t_i > T) &= \frac{\lambda(t_i)e^{-\int_{t_{i-1}}^{t_i}\lambda(s)ds}1_{(T,\infty]}(t_i)}{e^{-\int_{t_{i-1}}^{T}\lambda(s)ds}} \\ &= \lambda(t_i)e^{-\int_{T}^{t_i}\lambda(s)ds}1_{(T,\infty]}(t_i). \end{aligned} \tag{37}$$

□

The following lemma asks, given two inhomogeneous Poisson point processes 0 and 1, what is the probability that the next event comes from 0?

Lemma A.3. *Consider two inhomogeneous Poisson point processes with intensities $\lambda_0(t) = e^{\delta_0}\lambda(t)$ and $\lambda_1 = e^{\delta_1}\lambda(t)$. Let the current time be denoted T. The probability that the next event is from process 1 is given by*

$$\frac{e^{\delta_1}}{e^{\delta_0} + e^{\delta_1}}. \tag{38}$$

Proof. Let τ_0 and τ_1 denote the next event times from process 0 and 1 respectively. From Lemma A.2

$$p(\tau_1|\tau_1 > T) = \lambda_1(\tau_1)e^{-\Lambda_1(T,\tau_1]}1_{(T,\infty]}(\tau_1)$$
$$p(\tau_0|\tau_0 > T) = \lambda_0(\tau_0)e^{-\Lambda_0(T,\tau_0]}1_{(T,\infty]}(\tau_0),$$

where $\Lambda_i(T, \tau_i] = \int_T^{\tau_i} \lambda_i(s)ds$ The probability that the next event is from process 1 is given by

$$\begin{aligned}
\mathbb{P}[\tau_1 < \tau_0|\tau_0, \tau_1 > T] &= \int_T^{\infty} \lambda_0(\tau_0)e^{-\Lambda_0(T,\tau_0]} \int_T^{\tau_0} \lambda_1(\tau_1)e^{-\Lambda_1(T,\tau_1]}d\tau_1 d\tau_0 \\
&= \int_T^{\infty} \lambda_0(\tau_0)e^{-\Lambda_0(T,\tau_0]} \left(1 - e^{-\Lambda_1(T,\tau_0]}\right) d\tau_0 \\
&= 1 - \int_T^{\infty} \lambda_0(\tau_0)e^{-\int_T^{\tau_0} \lambda_0(s)+\lambda_1(s)}d\tau_0 \\
&= 1 - \frac{e^{\delta_0}}{(e^{\delta_0} + e^{\delta_1})} \int_T^{\infty} (e^{\delta_0} + e^{\delta_1})\lambda(\tau_0)e^{-(e^{\delta_0}+e^{\delta_1})\int_T^{\tau_0} \lambda(s)}d\tau_0 \\
&= 1 - \frac{e^{\delta_0}}{(e^{\delta_0} + e^{\delta_1})} \\
&= \frac{e^{\delta_1}}{(e^{\delta_0} + e^{\delta_1})}.
\end{aligned}$$

□

The proof of Theorem 4.1 can now be given using Lemma A.3 and the superposition property [Kingman, 1992] of Poisson processes.

Proof. Consider d inhomogeneous Poisson point processes with intensity functions $\lambda_i(t) = \rho_i e^{\delta_i}\lambda(t)$ for $i \in \{1, 2, \ldots, d\}$. Choose any process of interest j, with corresponding intensity $\rho_j e^{\delta_j}\lambda(t)$. Let the union of all timestamps from the other $i \neq j$ processes be combined into a new "not j" process. By the superposition property of the Poisson process, the combined timestamps form a new Poisson process with intensity function $\lambda_u(t) = \sum_{i\neq j} \rho_i e^{\delta_i}\lambda(t)$. This reduces the problem to a comparison of two inhomogeneous Poisson point processes. From Lemma A.3, the probability that the next event corresponds to process j is then

$$\frac{\rho_j e^{\delta_j}}{\rho_j e^{\delta_j} + \sum_{i\neq j} \rho_i e^{\delta_i}} = \frac{\rho_j e^{\delta_j}}{\sum_i \rho_i e^{\delta_i}}. \tag{39}$$

□

A.11 Simulation: Time Inhomogeneous Bernoulli Processes

Consider the following example. An experimental unit i is randomly assigned to one of 3 arms with probabilities $\boldsymbol{\rho} = (0.1, 0.3, 0.6)$. Let $g(i)$ map the unit to the arm index to which it is assigned. A Bernoulli success is observed for unit i with probability $e^{\mu(i)}e^{\delta_{g(i)}}$ with $\boldsymbol{\delta} = (\log 0.2, \log 0.3, \log 0.4)$ and $\mu(i) = \frac{1}{2}\sin(\frac{7\pi i}{n}) + \frac{1}{2}$. Confidence sequences for contrasts $\delta_2 - \delta_1$ and $\delta_2 - \delta_0$, obtained through Equation (15), are shown in Figure 5. The sequential p-value for testing the hypothesis $\delta_0 - \delta_1 \geq 0$ and $\delta_0 - \delta_2 \geq 0$, obtained through Equation (16), are shown using the right axis of Figure 5. The

p-value is less than $u = 0.05$ for all $n \geq 573$. This is the smallest n for which the joint confidence set over these contrasts fails to intersect with the set defined by the hypothesis (the lower left quadrant) as shown in Figure 6.

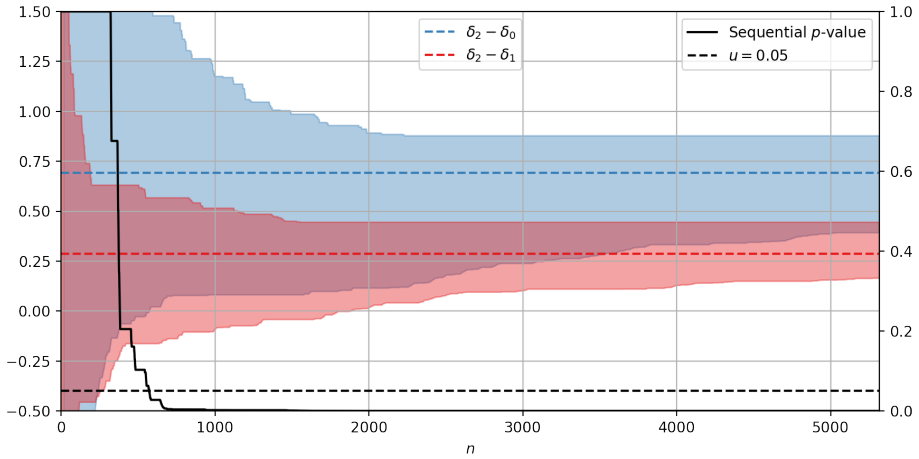


Figure 5: (Left axis) 0.95 Simultaneous confidence sequences for $\delta_2 - \delta_1 = \log(0.4) - \log(0.3) \approx 0.29$ and $\delta_2 - \delta_0 = \log(0.4) - \log(0.2) \approx 0.69$ provided by Corollary 3.2 and obtained via the solution to Equation (15). The confidence sequences for $\delta_2 - \delta_0$ and $\delta_2 - \delta_1$ are completely positive for $n \geq 573$ and $n \geq 1882$ respectively, after which one can conclude with probability $1 - u$ that arm 2 is optimal. (Right axis) Sequential p-value for testing the null hypothesis $\delta_0 \geq \delta_1$ and $\delta_0 \geq \delta_2$ obtained by solving Equation (16). The p-value is less than critical value $u = 0.05$ for all $n \geq 573$.

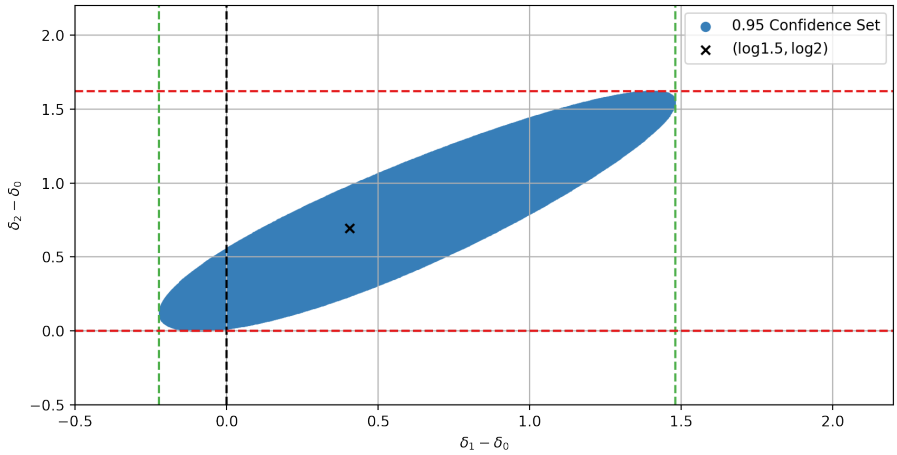


Figure 6: 0.95 joint confidence set for $\delta_2 - \delta_0$ and $\delta_1 - \delta_0$ at $n = 573$. (Black cross) True parameter values $(\log 3/2, \log 2)$. (Red dashed) $l^-_{573,\boldsymbol{b}}$ and $l^+_{573,\boldsymbol{b}}$, (Green dashed) $l^-_{573,\boldsymbol{c}}$ and $l^+_{573,\boldsymbol{c}}$ as in Corollary 3.2 with $\boldsymbol{b} = (-1, 0, 1)$ and $\boldsymbol{c} = (0, -1, 1)$.

A.12 Simulation: Time Inhomogeneous Poisson Processes

Consider the following example with only two arms, such as a canary test designed to test if a new software version produces more errors. Units are assigned to arms with probability $\boldsymbol{\rho} = (0.8, 0.2)$. $\lambda_1(t)$ can be expressed in terms of $\lambda_0(t)$ as $\lambda_1(t) = \rho e^{\delta} \lambda_0(t)$ with $\rho = \frac{\rho_1}{\rho_0}$ and $e^{\delta} = e^{\delta_1 - \delta_0}$. Let $\delta = 1.5$ and $\lambda_0(t) = 2000\text{sigmoid}(\sin(10\pi t) + 8t - 4)$. Point process realizations are obtained by thinning a homogeneous Poisson point process with rate 2000 [Lewis and Shedler, 1979]. Figure 7 shows the point process realizations, intensities, and counting processes for each arm. Figure 8 shows the confidence sequence for δ and the sequential p-value for testing equality $\delta = 0$ ($\boldsymbol{\theta} = \boldsymbol{\rho}$).

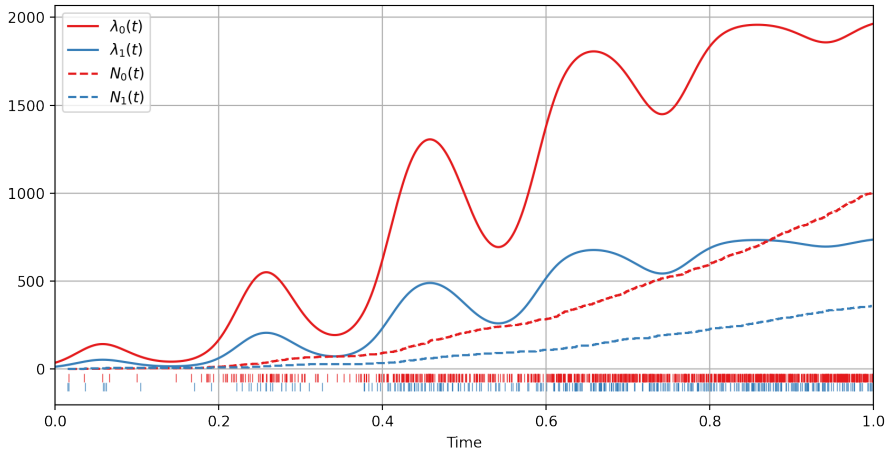


Figure 7: Inhomogeneous Poisson point process intensities $\lambda_0(t) = 2000\text{sigmoid}(\sin(10\pi t) + 8t - 4)$ and $\lambda_1(t) = \frac{1}{4} e^{\frac{3}{2}} \lambda_0(t)$. Associated counting processes $N_0(t)$ and $N_1(t)$. Point process realizations (rug-plots).

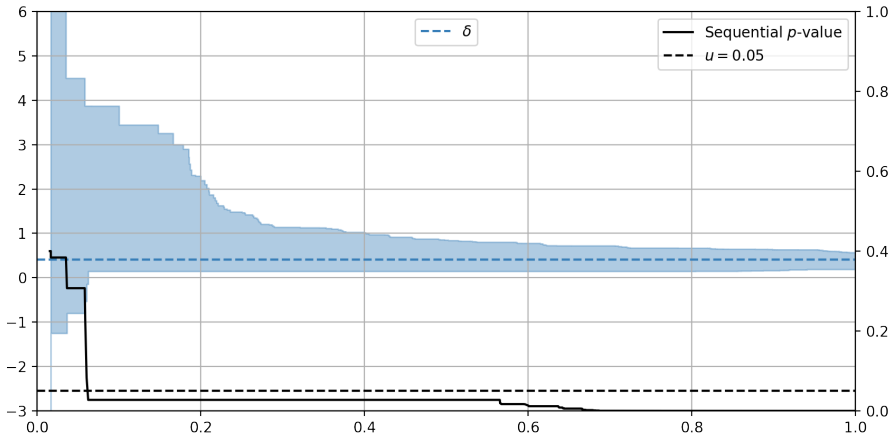


Figure 8: (Left axis) 0.95 *continuous-time* Confidence sequence for $\delta = 1.5$. (Right axis) Sequential p-value for testing equality i.e. $\delta = 0 \Rightarrow \boldsymbol{\theta} = \boldsymbol{\rho}$.