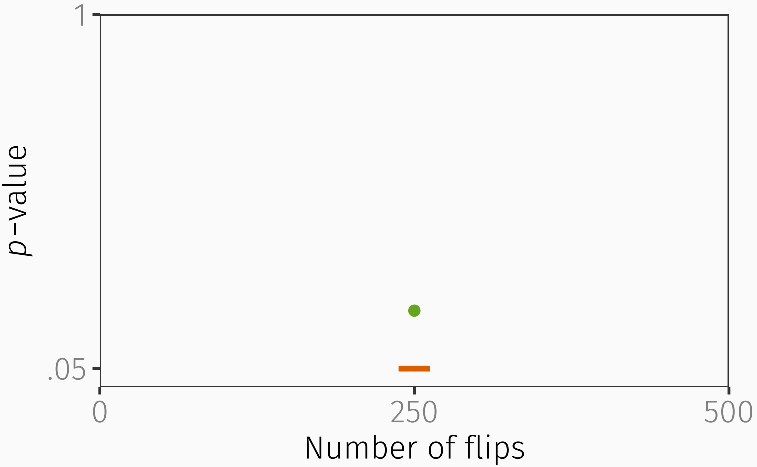
Nonparametric generalizations of the sequential probability ratio test

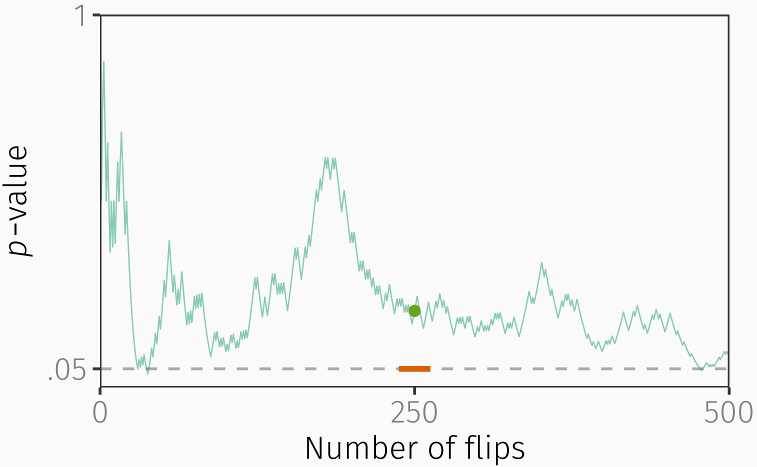
Steve Howard
Joint work with Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon

November 4, 2019

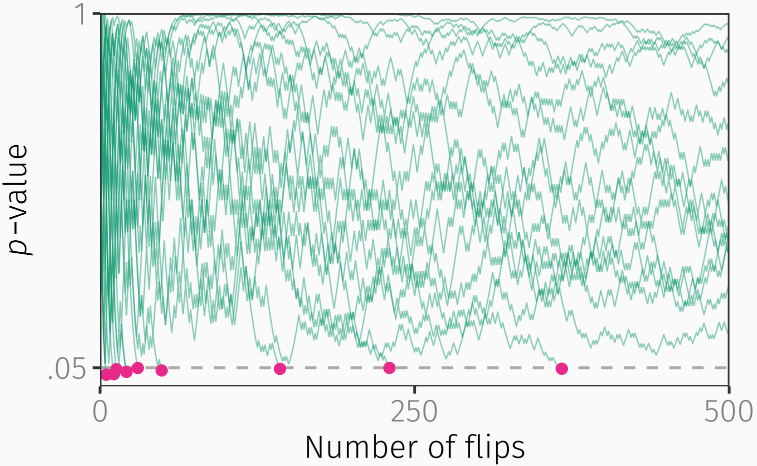
One p-value from a fair coin



One path of p-values from a fair coin

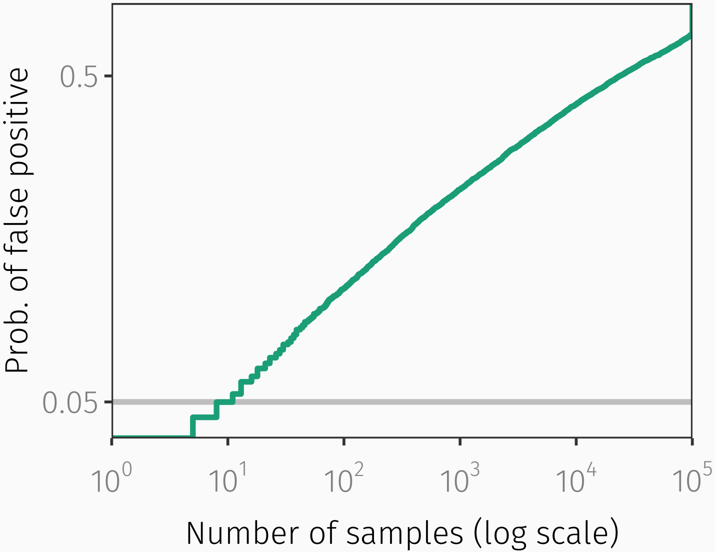


Continuous monitoring of fixed-sample p-values inflates the false positive rate.

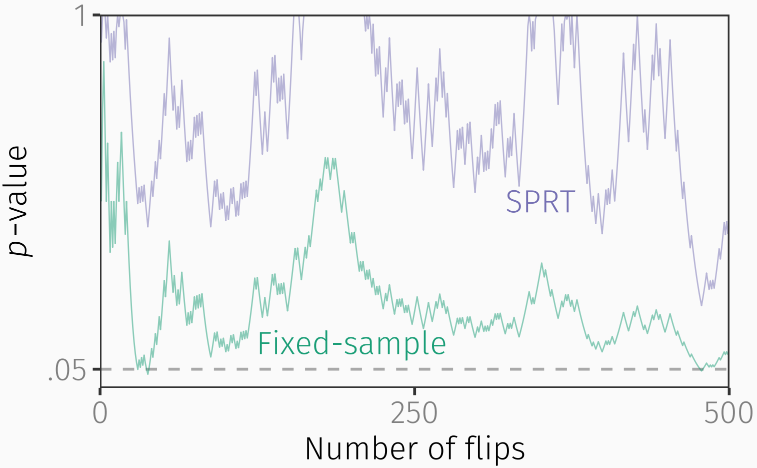


Here, with a fair coin, 10 out of 25 paths reach significance.

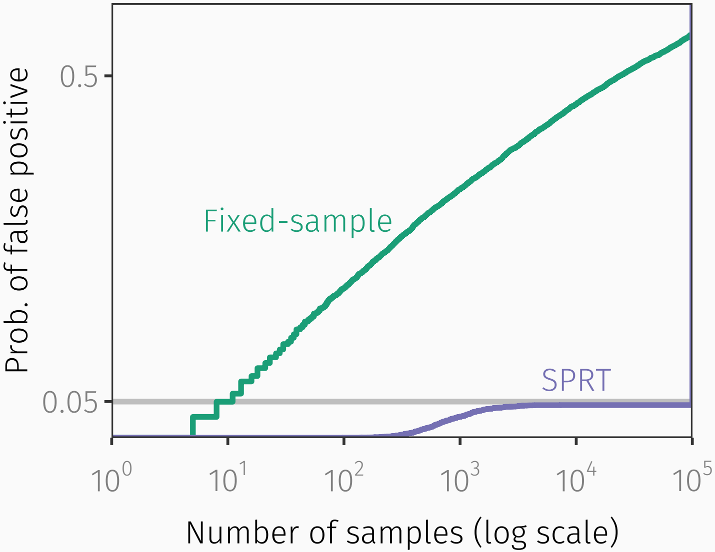
The false positive rate grows arbitrarily large with enough flips.



The sequential probability ratio test (SPRT) yields more conservative p-values.



The SPRT controls false positives uniformly over time.



The SPRT is based on a likelihood ratio.

Write S_t for the number of heads after t flips of a coin with bias π.

The SPRT is based on a likelihood ratio.

Write S_t for the number of heads after t flips of a coin with bias π.

Testing $H_0 : \pi = 1/2$ against $H_1 : \pi = \pi_1$, the likelihood ratio is

$$L_t = \frac{\pi_1^{S_t}(1-\pi_1)^{t-S_t}}{(1/2)^t}.$$

The SPRT is based on a likelihood ratio.

Write S_t for the number of heads after t flips of a coin with bias π.

Testing $H_0 : \pi = 1/2$ against $H_1 : \pi = \pi_1$, the likelihood ratio is

$$L_t = \frac{\pi_1^{S_t}(1-\pi_1)^{t-S_t}}{(1/2)^t}.$$

For fixed t, Neyman-Pearson says: compute $\mathbb{P}(L_t \geq x)$.

- E.g., $\mathbb{P}(L_t \geq \text{observed value})$ is a p-value, and
- if $\mathbb{P}(L_t \geq x_{0.05}) = 0.05$, then $x_{0.05}$ is a critical value.

The SPRT is based on a likelihood ratio.

Write S_t for the number of heads after t flips of a coin with bias π.

Testing $H_0 : \pi = 1/2$ against $H_1 : \pi = \pi_1$, the likelihood ratio is

$$L_t = \frac{\pi_1^{S_t}(1 - \pi_1)^{t - S_t}}{(1/2)^t}.$$

For fixed t, Neyman-Pearson says: compute $\mathbb{P}(L_t \geq x)$.

- E.g., $\mathbb{P}(L_t \geq \text{observed value})$ is a p-value, and
- if $\mathbb{P}(L_t \geq x_{0.05}) = 0.05$, then $x_{0.05}$ is a critical value.

But we care about $\mathbb{P}(L_t \geq x \text{ for some t})$. *That's what the SPRT controls.*

The likelihood ratio is a nonnegative supermartingale.

Fact: under H_0, the likelihood ratio is a nonnegative supermartingale, i.e., for each $t \geq 1$,

1. $L_t \geq 0$, and
2. $\mathbb{E}\left(L_t \mid L_1, \ldots, L_{t-1} \right) \leq L_{t-1}$, where we take $L_0 = 1$.

The likelihood ratio is a nonnegative supermartingale.

Fact: under H_0, the likelihood ratio is a nonnegative supermartingale, i.e., for each $t \geq 1$,

1. $L_t \geq 0$, and
2. $\mathbb{E}\left(L_t \mid L_1, \ldots, L_{t-1} \right) \leq L_{t-1}$, where we take $L_0 = 1$.

Ville's inequality: for any $x > 0$, $\mathbb{P}(L_t \geq x \text{ for some t}) \leq 1/x$.

The likelihood ratio is a nonnegative supermartingale.

Fact: under H_0, the likelihood ratio is a nonnegative supermartingale, i.e., for each $t \geq 1$,

1. $L_t \geq 0$, and
2. $\mathbb{E}\left(L_t \mid L_1, \ldots, L_{t-1} \right) \leq L_{t-1}$, where we take $L_0 = 1$.

Ville's inequality: for any $x > 0$, $\mathbb{P}(L_t \geq x \text{ for some t}) \leq 1/x$.

- So $1/L_t$ is an *always-valid p-value* [Johari 2015], and
- $x = 20$ is a critical value for a 0.05-level sequential test.

The SPRT’s uniform false positive control depends only on the supermartingale property of the likelihood ratio.

The Bernoulli SPRT works for any bounded distribution.

Is the process $L_t = \pi_1^{S_t}(1 - \pi_1)^{t-S_t}/(1/2)^t$ a supermartingale under any other circumstances?

The Bernoulli SPRT works for any bounded distribution.

Is the process $L_t = \pi_1^{S_t}(1 - \pi_1)^{t-S_t}/(1/2)^t$ a supermartingale under any other circumstances?

Yes: L_t is a supermartingale whenever S_t is a sum of independent, $[0, 1]$-valued observations with mean $1/2$.

The Bernoulli SPRT works for any bounded distribution.

Is the process $L_t = \pi_1^{S_t}(1 - \pi_1)^{t-S_t}/(1/2)^t$ a supermartingale under any other circumstances?

Yes: L_t is a supermartingale whenever S_t is a sum of independent, $[0, 1]$-valued observations with mean $1/2$.

- This follows from a bound the moment-generating function, ala exponential concentration bounds (e.g., Hoeffding, 1963).

The Bernoulli SPRT works for any bounded distribution.

Is the process $L_t = \pi_1^{S_t}(1-\pi_1)^{t-S_t}/(1/2)^t$ a supermartingale under any other circumstances?

Yes: L_t is a supermartingale whenever S_t is a sum of independent, $[0,1]$-valued observations with mean 1/2.

- This follows from a bound the moment-generating function, ala exponential concentration bounds (e.g., Hoeffding, 1963).

So the Bernoulli SPRT is actually a valid sequential test for the mean of *any* bounded distribution.

The SPRT is derived in a parametric setting, but in fact holds over a nonparametric class of distributions.

The SPRT is derived in a parametric setting, but in fact holds over a nonparametric class of distributions.

Exponential concentration results provide the link between parametric and nonparametric SPRTs.

Many exponential concentration results lead to “nonparametric SPRTs”.

- Bennett (1962) $\Rightarrow$ nonparametric Poisson SPRT
- Hoeffding (1963) $\Rightarrow$ nonparametric Bernoulli and Gaussian SPRTs
- Tropp (2011,2012) $\Rightarrow$ nonparametric Poisson and Gaussian *matrix* SPRTs
- de la Peña (1999) $\Rightarrow$ self-normalized Gaussian SPRT for symmetric distributions
- ...

Empirical-Bernstein sequential test

Theorem (H., Ramdas, McAuliffe, Sekhon 2019+)

Suppose $S_t = \sum_{i=1}^{t}(X_i - \mathbb{E}X_i)$ *where the* X_i *are independent and* $[a, a+b]$*-valued. Let* $(\widehat{X}_i)$ *be any predictable sequence. Then for any* $\lambda > 0$,

$$L_t = \exp\{\lambda S_t - \psi(\lambda)V_t\}$$

is a nonnegative supermartingale, where

$$\psi(\lambda) = \frac{-\log(1 - b\lambda) - b\lambda}{b^2},$$

$$V_t = \sum_{i=1}^{t}(X_i - \widehat{X}_i)^2.$$

Empirical-Bernstein sequential test

Theorem (H., Ramdas, McAuliffe, Sekhon 2019+)

Suppose $S_t = \sum_{i=1}^t (X_i - \mathbb{E}X_i)$ where the X_i are independent and $[a, a+b]$-valued. Let $(\widehat{X}_i)$ be any predictable sequence. Then for any $\lambda > 0$,

$$L_t = \exp\{\lambda S_t - \psi(\lambda) V_t\}$$

is a nonnegative supermartingale, where

$$\psi(\lambda) = \frac{-\log(1 - b\lambda) - b\lambda}{b^2},$$

$$V_t = \sum_{i=1}^t (X_i - \widehat{X}_i)^2.$$

Empirical-Bernstein sequential test

Theorem (H., Ramdas, McAuliffe, Sekhon 2019+)

Suppose $S_t = \sum_{i=1}^{t}(X_i - \mathbb{E}X_i)$ *where the* X_i *are independent and* $[a, a+b]$*-valued. Let* $(\widehat{X}_i)$ *be any predictable sequence. Then for any* $\lambda > 0$,

$$L_t = \exp\left\{\lambda S_t - \psi(\lambda) V_t\right\}$$

is a nonnegative supermartingale, where

$$\psi(\lambda) = \frac{-\log(1 - b\lambda) - b\lambda}{b^2},$$

$$V_t = \sum_{i=1}^{t}(X_i - \widehat{X}_i)^2.$$

Empirical-Bernstein sequential test

Theorem (H., Ramdas, McAuliffe, Sekhon 2019+)

Suppose $S_t = \sum_{i=1}^{t}(X_i - \mathbb{E}X_i)$ *where the* X_i *are independent and* $[a, a+b]$*-valued. Let* $(\widehat{X}_i)$ *be any predictable sequence. Then for any* $\lambda > 0$*,*

$$L_t = \exp\left\{\lambda S_t - \psi(\lambda) V_t\right\}$$

is a nonnegative supermartingale, where

$$\psi(\lambda) = \frac{-\log(1 - b\lambda) - b\lambda}{b^2},$$

$$V_t = \sum_{i=1}^{t}(X_i - \widehat{X}_i)^2.$$

Empirical-Bernstein sequential test

Theorem (H., Ramdas, McAuliffe, Sekhon 2019+)

Suppose $S_t = \sum_{i=1}^t (X_i - \mathbb{E}X_i)$ *where the* X_i *are independent and* $[a, a+b]$*-valued. Let* $(\widehat{X}_i)$ *be any predictable sequence. Then for any* $\lambda > 0$,

$$L_t = \exp\left\{\lambda S_t - \psi(\lambda) V_t\right\}$$

is a nonnegative supermartingale, where

$$\psi(\lambda) = \frac{-\log(1 - b\lambda) - b\lambda}{b^2},$$

$$V_t = \sum_{i=1}^t (X_i - \widehat{X}_i)^2.$$

Empirical-Bernstein sequential test

Theorem (H., Ramdas, McAuliffe, Sekhon 2019+)

Suppose $S_t = \sum_{i=1}^t (X_i - \mathbb{E}X_i)$ *where the* X_i *are independent and* $[a, a+b]$*-valued. Let* $(\widehat{X}_i)$ *be any predictable sequence. Then for any* $\lambda > 0$*,*

$$L_t = \exp\left\{\lambda S_t - \psi(\lambda) V_t\right\}$$

is a nonnegative supermartingale, where

$$\psi(\lambda) = \frac{-\log(1 - b\lambda) - b\lambda}{b^2},$$

$$V_t = \sum_{i=1}^t (X_i - \widehat{X}_i)^2.$$

Average treatment effect: setup

Unit i has fixed potential outcomes $Y_i(0), Y_i(1)$, for $i = 1, 2, \ldots$

Average treatment effect: setup

Unit i has fixed potential outcomes $Y_i(0), Y_i(1)$, for $i = 1, 2, \ldots$

We assign unit i randomly to treatment or control, and observe $Y_i(1)$ or $Y_i(0)$ accordingly. [Neyman 1923, Rubin 1974]

Average treatment effect: setup

Unit i has fixed potential outcomes $Y_i(0), Y_i(1)$, for $i = 1, 2, \ldots$

We assign unit i randomly to treatment or control, and observe $Y_i(1)$ or $Y_i(0)$ accordingly. [Neyman 1923, Rubin 1974]

Assumption: no interference

Average treatment effect: setup

Unit i has fixed potential outcomes $Y_i(0), Y_i(1)$, for $i = 1, 2, \ldots$

We assign unit i randomly to treatment or control, and observe $Y_i(1)$ or $Y_i(0)$ accordingly. [Neyman 1923, Rubin 1974]

Assumption: no interference

Our goal: after observing units $1, \ldots, t$, we'd like to estimate

$$\text{ATE}_t := \frac{1}{t}\sum_{i=1}^{t}[Y_i(1) - Y_i(0)].$$

Average treatment effect: setup

Unit i has fixed potential outcomes $Y_i(0), Y_i(1)$, for $i = 1, 2, \ldots$

We assign unit i randomly to treatment or control, and observe $Y_i(1)$ or $Y_i(0)$ accordingly. [Neyman 1923, Rubin 1974]

Assumption: no interference

Our goal: after observing units $1, \ldots, t$, we'd like to estimate

$$\mathrm{ATE}_t := \frac{1}{t} \sum_{i=1}^{t} [Y_i(1) - Y_i(0)].$$

Assumption: $Y_i(k) \in [a, a+b]$ for $k = 0, 1$, all i.

Average treatment effect: theorem

For each unit i, we construct an estimator X_i of the individual treatment effect $Y_i(1) - Y_i(0)$ with two key properties:

1. **Unbiased**: $\mathbb{E}X_i = Y_i(1) - Y_i(0)$
2. **Variance** of X_i depends on **prediction errors** $(Y_i(1) - \widehat{Y}_i(1))^2$ and $(Y_i(0) - \widehat{Y}_i(0))^2$.

Average treatment effect: theorem

For each unit i, we construct an estimator X_i of the individual treatment effect $Y_i(1) - Y_i(0)$ with two key properties:

1. **Unbiased**: $\mathbb{E}X_i = Y_i(1) - Y_i(0)$
2. **Variance** of X_i depends on **prediction errors** $(Y_i(1) - \widehat{Y}_i(1))^2$ and $(Y_i(0) - \widehat{Y}_i(0))^2$.

Theorem (H., Ramdas, McAuliffe, Sekhon 2019+)

Assume no interference and $Y_t(k) \in [a, a+b]$ for all k, t. Let $S_t = \sum_{i=1}^t X_i - t \cdot \mathrm{ATE}_t$. Then for any $\lambda > 0$,

$$L_t = \exp\left\{ \lambda S_t - \psi(\lambda) \sum_{i=1}^{t} (X_i - \widehat{X}_i)^2 \right\}$$

is a nonnegative supermartingale.

Thank you!

- `stevehoward@berkeley.edu`
- Exponential line-crossing inequalities:
  `https://arxiv.org/abs/1808.03204`
- Uniform, nonparametric, non-asymptotic confidence sequences:
  `https://arxiv.org/abs/1810.08240`
- Sequential estimation of quantiles with applications to A/B-testing and best-arm identification: `https://arxiv.org/abs/1906.09712`
- Implementations of many uniform boundaries and confidence sequences:
  `https://github.com/gostevehoward/confseq`
- Slides: `stevehoward.org`