UNIVERSITAT DE BARCELONA
Facultat de Matemàtiques
i Informàtica

GRAU DE MATEMÀTIQUES

Treball final de grau

Sequential hypothesis testing in online controlled experiments

Autor: Arnau Pérez Reverte

Director: Dr. Josep Vives i Santa Eulàlia
Realitzat a: Facultat de Matemàtiques i Informàtica

Barcelona, January 15, 2024

Contents

Abstract

This thesis explores the sequential analysis paradigm in the field of mathematical statistics, where sample size is not predetermined, allowing for adaptive decision-making. The first chapter outlines the theory's foundations, particularly in sequential hypothesis testing, introducing the properties of two of the most relevant sequential tests: the *Sequential Probability Ratio Test* (SPRT) and the *mixture Sequential Probability Ratio Test* (mSPRT). The second chapter focuses on applying sequential hypothesis testing to online controlled experimentation, using *Always Valid Inference*. This alternative offers a statistically rigorous and efficient solution, potentially outperforming fixed-horizon methods. Empirical evidence of simulated real-world case-scenarios supports the proposed methodology's advantages in online controlled experimentation.

Resum

Aquesta tesi explora el paradigma de l'anàlisi seqüencial dins del camp de l'estadística matemàtica, on la mida de la mostra no està predeterminada, permetent així una presa de decisions adaptativa. El primer capítol explica els fonaments d'aquesta teoria; en particular, sobre els tests d'hipòtesi seqüencials, enunciant les propietats de dos dels tests seqüencials més rellevants: el *Sequential Probability Ratio Test* (SPRT) i el *mixture Sequential Probability Ratio Test* (mSPRT). El segon capítol es centra en l'aplicació dels tests d'hipòtesi seqüencials a l'experimentació controlada en línia, utilitzant *Always Valid Inference*. Aquesta alternativa ofereix una solució estadísticament rigorosa i eficient, amb el potencial de superar els mètodes clàssics amb mida de mostra fixa. Evidència empírica obtinguda mitjançant la simulació de casuístiques reals recolza les avantatges de la metodologia proposada en l'experimentació controlada en línia.

2020 Mathematics Subject Classification: 62F03, 62L05, 62L10, 62L15, 62P30

Key-words and phrases: sequential analysis, sequential hypothesis testing, SPRT, mSPRT, online controlled experiments, A/B testing, Always Valid Inference

Acknowledgements

First and foremost, I would like to thank my supervisor Dr. Josep Vives for his support and guidance in the development of my thesis. I want to also thank my loving girlfriend Ainhoa and all of my family for their unconditional support; all the friends I have made during these four years at the faculty, specially Òscar and Arnau; and also all of my friends back in Terrassa.

Lastly, this body of work is dedicated to my late grandmother who sadly passed away during the last days of writing:

> *Àvia, tota la meva educació la tinc gràcies a tu. Aquest treball està dedicat a tu ja que simbolitza el final de la meva carrera, que és allò que tu sempre volies arribar a veure. Espero que estiguis orgullosa de la persona que vas criar de la mateixa manera que jo estic orgullós de poder dir que vas ser la meva àvia.*
>
> *Fins que ens retrobem al Cel.*

Modern mathematics teaching has settled the theory of mathematical statistics as the one first developed during early XX^{th} century by prominent figures like Fisher, Neyman and Pearson. While their contributions have proven to be exceptionally useful, oftentimes the dominion of this theory leaves no room for other ideas which can be more suitable for certain problems. This is nothing new in the field of statistics, where the actual interpretation of probability is already known to be the subject of a controversial discussion.

In the first chapter of this body of work we introduce the theory of sequential analysis, which is defined to be the statistical theory corresponding to a context where the sample size is not fixed beforehand. Whether it is estimation or testing that we want to perform, observations are gathered sequentially, so at each step we can decide to stop our inference procedure according to a predefined rule, or else we can continue sampling. The development of this field was motivated by the efficient decision-making that this framework indeed offers. This is specially the case for sequential hypothesis testing, which was the foundation stone of sequential analysis, and which traditionally found its applications in quality control and clinical trials, where quickness in decision is most valuable. Nonetheless, the emergence and wide availability of data in the last decade has brought sequential testing to new grounds.

Controlled experiments are held nowadays in almost the entirety of the Internet. Every social network, e-commerce site or online product optimizes its features in order to maximize user engagement or return-on-investment, using randomization and statistical methods to detect significant effects. It's clear that identifying these features as quickly as possible is crucial for the successful development of the product. However, classic statistical methods, and fixed-horizon hypothesis testing in particular, define a really strict framework which followed unrigorously can truncate the statistical validity of the results. Moreover, these tools might not even result optimal for the problem in question.

Thereupon, the second chapter focuses on introducing sequential hypothesis testing to online controlled experimentation via *Always Valid Inference* [6], which adapts the sequential methodology to the intricacies that the controlled experimentation framework presents. We will show that the proposed alternative provides an statistically rigorous and efficient solution, which can even outshine fixed-horizon in certain casuistics. Finally, these results are furthermore backed with empirical evidence through the simulation of real-word case-scenarios using The R Programming Language.

Chapter 1

Sequential analysis

1.1 Sequential statistical models

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Suppose that we sequentially observe *i.i.d.* realizations of random variables $X_1, X_2, \ldots$ taking values on the measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ and we wish to perform inference over the distribution of X_1. For each new observation X_n, we gather an n-dimensional random vector

$$\mathbf{X}_n = (X_1, \ldots, X_n) : (\Omega, \mathcal{F}) \to (\mathcal{X}_n, \mathcal{A}_n)$$

known as our *sample* of *size* $n \in I \subseteq \mathbb{N}$, where

$$\mathcal{X}_n = \prod_{i=1}^{n} \mathcal{X}, \quad \mathcal{A}_n = \mathcal{B}(\mathcal{X}_n) = \mathcal{B}(\mathcal{X})^n.$$

In contrast to classical statistical theory, we can decide to stop the sampling process at any stage, so I may be non-finite, or else we can decide a maximum sample size beforehand. We assume that the law of X_1 is contained in a family of probability distributions $\mathcal{P}$.

Definition 1.1 (**Sequential statistical model**)**.** A *sequential statistical model* is a triple $((\mathcal{X}_n, \mathcal{A}_n)_{n\in I}, \mathcal{P}, I)$ where

- $(\mathcal{X}_n, \mathcal{A}_n)_{n\in I}$ is a family of measurable spaces, known as the *sequential sample space*
- $\mathcal{P}$ is a family of probability distributions on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$
- $I \subseteq \mathbb{N}$ is a non-empty set representing all possible sample sizes

During this work we will focus on the study of parametric sequential statistical models:

Definition 1.2 (**Parametric sequential statistical model**)**.** A sequential statistical model is said to be *parametric* if $\mathcal{P} = \{\mathbb{P}_\theta,\ \theta \in \Theta\}$, where the non-empty $\Theta \subseteq \mathbb{R}^d$ is known as the *parameter space*.

Observe that our growing sample generates a natural filtration of the probability space $\mathbb{F} = (\mathcal{F}_n)_{n\in I}$, where $\mathcal{F}_n = \sigma(\mathbf{X}_n)$ represents the information available with our sample of size n. Hence, the terminal decision to stop sampling and terminate our inference procedure shall be made using only the observed information $\mathcal{F}_n$. Recall by the theory of stochastic processes that this action can be encapsulated by the following definition:

Definition 1.3 (**Stopping time**)**.** A random variable $N : \Omega \to I$ is said to be a *stopping time* with respect to a filtration $\mathbb{F} = (\mathcal{F}_n)_{n\in I}$ if $\{N = n\} \in \mathcal{F}_n$ for all $n \in I$.

From now on, we fix the filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$.

1.2 Sequential hypothesis testing

Consider a (parametric) sequential statistical model. Suppose that we have a partition of the parameter space in $\Theta = \Theta_0 \cup \Theta_1$, and we wish to know whether $\theta \in \Theta_0$ or else $\theta \in \Theta_1$. Equivalently, we wish to test between the two hypothesis

$$\begin{cases} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \in \Theta_1 \end{cases}$$

where H_0 is known as the *null hypothesis* and H_1 as the *alternative hypothesis*.

Definition 1.4 (**Simple and composite hypothesis**)**.** A hypothesis H_i is said to be *simple* if $\Theta_i = \{\theta_i\}$. We say that H_i is *composite* otherwise.

The sequential analysis literature highlights two main approaches to testing hypothesis sequentially, which can only be used depending on whether H_i are simple or composite.

Given that our realizations are observed in a sequential fashion, then at every stage $n \in I$ we can make use of all the available information up to that point $\mathcal{F}_n$ to make a decision. The first (and classic) approach doesn't apply restrictions on the type of H_i, and consists in choosing at each sampling stage whether to

1. Stop the experiment and either accept or reject H_0 according to some rule.
2. Continue sampling.

Definition 1.5 (**Sequential hypothesis test**)**.** Given a sequential statistical model $((\mathcal{X}_n, \mathcal{A}_n)_{n\in I}, \mathcal{P}, I)$ and a partition of the parameter space $\Theta = \Theta_0 \cup \Theta_1$, a *sequential hypothesis test* for the hypotheses $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$ is a pair $\delta = (N, d)$ where

- $N : \Omega \to I$ is a stopping time with respect to $\mathbb{F}$, the *random sample size*
- $d = (d_n)_{n\in I}$ is a family of $\mathcal{F}_n$-measurable functions $d_n : \mathcal{X}_n \to \Delta$, the *decision rule*

where $\Delta := \{0, 1\}$, with $d_n = 1$ meaning that we reject H_0 (accept H_1) and $d_n = 0$ that we accept H_0.

The sampling process shall continue until N is first observed, meaning that we have gathered enough evidence to either accept or reject the null hypothesis H_0 following the corresponding decision rule at stage $N = n$, d_n.

The second, more modern approach, builds upon last definition but lies its foundations in decision theory ([6]). In this case, we require H_0 to be simple. Suppose then that, under the null hypothesis, sampling observations costs nothing, so our preferred action is to observe *ad infinitum*, and hence accept H_0. However, if the alternative hypothesis happens to be true, our sampling costs a fixed amount, so we want to stop sampling as soon as possible and reject H_0 ([14]).

Definition 1.6 (**Open-ended sequential hypothesis test**)**.** A sequential hypothesis test is said to be *open-ended* if it rejects a simple null hypothesis in finite time:

$$d_n \equiv d = \mathbb{1}_{\{N<\infty\}} \text{ for all } n \in I.$$

Notice then how this approach can be seen as more "detection-oriented", meaning that the (simple) null hypothesis maintains its role of being the default hypothesis indefinitely, being compared against a (possibly composite) set of alternatives, and we wish to only stop our procedure in case the alternative hypothesis is true.

In either case, we will more commonly refer to sequential hypothesis tests as simply *sequential tests*.

Sequential tests being defined by a stopping time and a terminal decision implies that comparison is fundamentally achieved by the following magnitudes:

Definition 1.7 (**Expected sample size**)**.** Given a sequential hypothesis test $\delta = (N, d)$, the expected value $\mathbb{E}_\theta(N)$ for any $\theta \in \Theta$ is known as the *expected sample size* (ESS).

Higher moments of N are also relevant for the study of the random sample size; however, we will mostly focus on the ESS, and in occasions we will also treat its second order moment.

Moreover, upon test termination two types of error can be committed:

Definition 1.8 (**Type-I error**)**.** The *Type-I error* function of a sequential hypothesis test $\delta = (N, d)$ is the probability of false rejection of the null hypothesis, as a function of $\theta \in \Theta_0$:

$$\mathbb{P}_\theta(d_N = 1) \text{ for } \theta \in \Theta_0.$$

Definition 1.9 (**Type-II error**)**.** The *Type-II error* function of a sequential test $\delta = (N, d)$ is the probability of incorrectly accepting the null hypothesis, as a function of $\theta \in \Theta_1$:

$$\mathbb{P}_\theta(d_N = 0) \text{ for } \theta \in \Theta_1.$$

Inherited by fixed-horizon hypothesis testing theory, we will sometimes make use of the following terminology:

Definition 1.10 (**Test size**)**.** We say a sequential hypothesis test $\delta = (N, d)$ is of *size* $\alpha \in (0, 1)$ if

$$\alpha = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(d_N = 1)$$

Definition 1.11 (**Power of a sequential test**)**.** The *power* function of a sequential hypothesis test $\delta = (N, d)$ is the probability of correctly rejecting the null hypothesis, as a function of $\theta \in \Theta_1$:

$$\mathbb{P}_\theta(d_N = 1) = 1 - \mathbb{P}_\theta(d_N = 0), \text{ for } \theta \in \Theta_1.$$

Finally, it's worth noting that open-ended tests, by being more "detection-oriented", satisfy the property of being *nested*: the family $(\delta(\alpha))_{\alpha \in (0,1)}$ of open-ended tests of size α satisfies that $N(\alpha)$ is *a.s.* non-increasing in α, and $d(\alpha)$ is *a.s.* non-decreasing in α. In other words, "less conservative rules necessarily terminate faster and make more rejections" ([7]).

1.3 The Sequential Probability Ratio Test (SPRT)

Sequential hypothesis testing and sequential analysis as a whole was first introduced by Wald [21] with the *Sequential Probability Ratio Test* (SPRT), which was designed to test a simple null hypothesis against a simple alternative.

Formally, let $((\mathcal{X}_n, \mathcal{A}_n)_{n\in I}, \mathcal{P}, I)$ be a sequential statistical model such that

$$\Theta = \{0, 1\}, \quad \mathcal{P} = \{\mathbb{P}_0, \mathbb{P}_1\}$$

where $\mathbb{P}_0, \mathbb{P}_1$ are two distinct probability measures corresponding to probability density functions f_0, f_1, respectively, and suppose we test for

$$\begin{cases} H_0 : \theta = 0 \\ H_1 : \theta = 1 \end{cases}.$$

The fundamental tool upon which most sequential tests are developed is the likelihood ratio of a sample:

Definition 1.12 (Likelihood ratio). Given a sequential statistical model, the *likelihood ratio* (LR) at sampling stage $n \in I$ for testing the simple hypotheses $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$ is the function of the sample

$$\begin{aligned} \Lambda_n : \mathcal{X}_n &\longrightarrow \mathbb{R} \\ \mathbf{X}_n &\mapsto \prod_{i=1}^{n} \frac{f_{\theta_1}(X_i)}{f_{\theta_0}(X_i)}. \end{aligned}$$

Given an n-dimensional sample $\mathbf{X}_n$, the LR compares the two proposed statistical models at stage $n \in I$ by taking the ratio of the joint probability of the sample under both cases. Indeed, if H_0 is true, then the LR becomes small as the experiment develops; otherwise, the LR becomes large. Wald then, naturally, proposed the first sequential test as the one which consider this likelihood ratio at each sampling stage and only stops the experiment in case its value becomes too large or too small.

Definition 1.13 (SPRT). The *Sequential Probability Ratio Test* (SPRT) with boundaries (B, A) for testing the simple hypothesis $H_0 : \theta = 0$ and $H_1 : \theta = 1$ is the sequential test $\delta = (N, d)$ defined by

$$\begin{aligned} N &= \inf\{n \in I : \Lambda_n \notin (B, A)\} \\ d &= (d_n)_{n\in I}, \text{ with } d_n(\mathbf{X}_n) = \begin{cases} 0 & \text{if } \Lambda_n \leq B \\ 1 & \text{if } \Lambda_n \geq A \end{cases} \end{aligned}$$

where $A, B \in \mathbb{R}$ such that $0 < B < 1 < A < \infty$.

Hence, the SPRT proposes to continue sampling until the likelihood ratio leaves the interval (B, A) for the first time. The magnitudes of B, A then will mean how tolerant we are with decanting upon the truth of the null or the alternative hypothesis.

For theoretical reasons, sometimes it will be convenient to work with the *log-likelihood ratio* (LLR). Let $Z_k := \log\left(f_1(X_k)/f_0(X_k)\right)$, then

$$\lambda_n(\mathbf{X}_n) := \log \Lambda_n(\mathbf{X}_n) = \sum_{i=1}^{n} \log\left(\frac{f_1(X_i)}{f_0(X_i)}\right) = \sum_{i=1}^{n} Z_i,$$

which translates the SPRT to the sequential test $\delta^* = (N^*, d^*)$ with

$$\begin{aligned} N^* &= \inf\{n \in I : \lambda_n \notin (b, a)\} \\ d^* &= (d_n^*)_{n\in I}, \text{ with } d_n^*(\mathbf{X}_n) = \begin{cases} 0 & \text{if } \lambda_n \le b \\ 1 & \text{if } \lambda_n \ge a \end{cases} \end{aligned}$$

where $a := \log A, b := \log B$, satisfying $b < 0 < a$.

1.3.1 Type-I and Type-II error control

Observe that given that both the null and the alternative hypothesis are simple, the Type-I error and Type-II error functions are constant. Hence, the SPRT boundaries (B, A) can be chosen in such a way that the Type-I error and Type-II error functions are fixed to predefined values $\alpha, \beta \in (0, 1)$. Assuming that $\mathbb{P}_\theta(N < \infty) = 1$, we will derive relations between the error functions and B, A.

For each sample size $n \in I$, consider the subset

$$B_n := \{B < \Lambda_k < A \;\forall\; k = 1, 2, \ldots, n-1, \quad \Lambda_n \ge A\} \subset \mathcal{X}_n$$

Then,

$$\begin{aligned} \mathbb{P}_0(d_N = 1) = \mathbb{P}_0(\Lambda_N \ge A) &= \sum_{n=1}^{\infty} \mathbb{P}_0(N = n, \Lambda_n \ge A) \\ &= \sum_{n=1}^{\infty} \int_{B_n} f_0(\xi_1, \ldots, \xi_n)\, d\xi_1, \ldots, d\xi_n \\ &= \sum_{n=1}^{\infty} \int_{B_n} \frac{f_0(\xi_1, \ldots, \xi_n)}{f_1(\xi_1, \ldots, \xi_n)} f_1(\xi_1, \ldots, \xi_n)\, d\xi_1, \ldots, d\xi_n \\ &= \sum_{n=1}^{\infty} \mathbb{E}_1(\Lambda_n^{-1} \mathbb{1}_{\{N=n,\, \Lambda_n \ge A\}}) = \mathbb{E}_1(\Lambda_N^{-1} \mathbb{1}_{\{\Lambda_N \ge A\}}) \le \frac{\mathbb{P}_1(\Lambda_N \ge A)}{A} = \frac{1 - \mathbb{P}_1(d_N = 0)}{A} \end{aligned} \tag{1.1}$$

without equality since Λ_N may be strictly greater than A. We can analogously derive the following bound for β :

$$\mathbb{P}_1(d_N = 0) = \mathbb{P}_1(\Lambda_N \leq B) \leq B\mathbb{P}_0(\Lambda_N \leq B) = B(1 - \mathbb{P}_0(d_N = 1)).$$

Hence, treating the inequalities as an approximation,

$$\mathbb{P}_0(d_N = 1) \approxeq \frac{1 - \mathbb{P}_1(d_N = 0)}{A}, \quad \mathbb{P}_1(d_N = 0) \approxeq B(1 - \mathbb{P}_0(d_N = 1)). \tag{1.2}$$

Setting A, B as a function of α, β we obtain the *Wald boundaries*

$$A \coloneqq \frac{1-\beta}{\alpha}, \quad B \coloneqq \frac{\beta}{1-\alpha}.$$

Moreover, observe that if we solve Equation 1.2 for the error functions we also obtain

$$\mathbb{P}_0(d_N = 1) \approxeq \frac{1 - B}{A - B}, \quad \mathbb{P}_1(d_N = 0) \approxeq B\frac{A-1}{A-B}. \tag{1.3}$$

1.3.2 Properties of N

We shall now study the ESS and other properties of the random sample size N of the SPRT. We first present the following definition:

Definition 1.14 (**Exponentially bounded r.v.**). A non-negative random variable M is said to be *exponentially bounded* (EB) if there exist constants $C > 0$ and $0 < \rho < 1$ such that $\mathbb{P}(M > m) \leq C\rho^m$ for all $m \geq 1$.

Observe that the EB property implies that M has finite moment-generating function. In particular, $\mathbb{P}(M < \infty) = 1$ and $\mathbb{E}(M^k) < \infty$ for all $k = 1, 2, \ldots$. Indeed, we have

$$\mathbb{P}(M < \infty) = \sum_{i=1}^{\infty} \mathbb{P}(M = i) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(M = i) = \lim_{n\to\infty} \mathbb{P}(M \leq n) \Longrightarrow$$
$$0 \leq \mathbb{P}(M = \infty) = \lim_{n\to\infty} \mathbb{P}(M > n) \leq \lim_{n\to\infty} C\rho^n = 0$$

since $0 < \rho < 1$. Moreover,

$$\mathbb{E}(e^{tM}) = \sum_{n=1}^{\infty} e^{tn}\mathbb{P}(M = n) \leq \sum_{n=1}^{\infty} e^{tn}\mathbb{P}(M > n - 1)$$
$$\leq \sum_{n=1}^{\infty} e^{tn} C\rho^{n-1} \leq Ce^t \sum_{n=1}^{\infty} (\rho e^t)^{n-1}$$

which converges for $t < -\log \rho$.

The following result will prove that the random sample size of the SPRT is, in fact, exponentially bounded.

Lemma 1.15 (**Stein's lemma**). *Let $(X_n)_{n\in I}$ be i.i.d. random variables with $\mathbb{P}(X_1 = 0) < 1$. Let $\lambda_n := \sum_{j=1}^{n} X_j$, $n = 1, 2, \ldots$, let $a, b \in \mathbb{R}$ such that $b < 0 < a$, and define*

$$M := \min\{n \geq 1 : \lambda_n \notin (b, a)\}$$

Then M is EB.

Proof. If $\mathbb{P}(X_1 = 0) < 1$, then there is $x > 0$ such that either $\mathbb{P}(X_1 \geq x) > 0$ or $\mathbb{P}(X_1 \leq -x) > 0$. Without loss of generality, assume that $\mathbb{P}(X_1 \geq x) = \varepsilon > 0$. Let $m \in \mathbb{N}$ such that $mx > a - b$. Then

$$\mathbb{P}(\lambda_m \geq a - b) \geq \mathbb{P}(\lambda_m \geq mx) \geq \mathbb{P}(X_1 \geq x, \ldots, X_m \geq x) = \varepsilon^m,$$

and hence for all $k \geq 1$,

$$\mathbb{P}(M > mk) = \mathbb{P}(b < \lambda_n < a, \quad n = 1, \ldots, mk) \leq (1 - \varepsilon^m)^k.$$

For any n, let k be such that $mk < n \leq (k + 1)m$. Then

$$\mathbb{P}(M > n) \leq \mathbb{P}(M > km) \leq (1 - \varepsilon^m)^k \leq (1 - \varepsilon^m)^{\frac{n}{m}-1} = \frac{1}{1 - \varepsilon^m}(1 - \varepsilon^m)^{\frac{n}{m}} = C\rho^n$$

where $C := 1/(1 - \varepsilon^m)$ and $\rho := (1 - \varepsilon^m)^{1/m}$. □

Consider the SPRT defined by the LLR δ^*. Observe that if f_0 and f_1 are distinct *a.e.*, then $\mathbb{P}_i(Z_0 \neq 0) = 1$, for $i = 0, 1$. Hence, applying Lemma 1.15 we obtain that the random sample size N^* of the SPRT is exponentially bounded, implying that the sequential test terminates *w.p.* 1 and all moments of N^* exist (and hence, the same applies for the classic SPRT).

Given that $\mathbb{E}_i(N) < \infty$ for $i = 0, 1$, we can try to derive an approximate expression for the expected sample size in terms of the predefined Type-I and Type-II error rates. We will make use of the following proposition:

Proposition 1.16 (**Wald identities**).

1. *Let $(X_n)_{n\in I}$ be i.i.d. such that $\mu := \mathbb{E}(X_1) < \infty$, and let N be a stopping time such that $\mathbb{E}\,|N| < \infty$. Then,*

$$\mathbb{E}(S_N) = \mu\mathbb{E}(N).$$

2. *Let* $(X_n)_{n\in I}$ *be i.i.d. such that* $\mathbb{E}\left|X_1^2\right| < \infty$, *with* $\mathbb{E}(X_1) = 0$ *and* $\sigma^2 = \mathbb{V}(X_1)$. *Let* N *be a stopping time such that* $\mathbb{E}\,|N| < \infty$. *Then,*

$$\mathbb{E}(S_N^2) = \sigma^2 \mathbb{E}(N).$$

Proof.

1. We suppose initially that $X_n \geq 0$ for all n. Observe that $\{N \geq n\} = \left(\bigcup_{j=1}^{n-1}\{N = j\}\right)^c$ is independent of $X_n, X_{n+1}, \ldots$ since N is a stopping time. Then, by monotone convergence

$$\begin{aligned}\mathbb{E}(S_N) = \mathbb{E}\left(\sum_{n=1}^{N} X_n\right) = \mathbb{E}\left(\sum_{n=1}^{\infty} X_n \mathbb{1}_{\{N\geq n\}}\right) = \sum_{n=1}^{\infty} \mathbb{E}(X_n \mathbb{1}_{\{N\geq n\}}) \\ = \sum_{n=1}^{\infty} \mathbb{E}(X_n)\mathbb{E}(\mathbb{1}_{\{N\geq n\}}) = \mu \sum_{n=1}^{\infty} \mathbb{P}(N \geq n) = \mu\mathbb{E}(N).\end{aligned}$$

   For the general case, we apply the same procedure to each term of the decomposition

$$\sum_{n=1}^{N} X_n = \sum_{n=1}^{N} \max(X_n, 0) - \sum_{n=1}^{N} - \min(X_n, 0).$$

2. Recall that $(M_n)_{n\in I}$ with $M_n \coloneqq S_n^2 - n\sigma^2$ is a martingale. Hence, $(M_{N\wedge n})_{n\in I}$ is also a martingale and

$$0 = \mathbb{E}(M_{N\wedge n}) = \mathbb{E}\left(S_{N\wedge n}^2 - (N \wedge n)\sigma^2\right) = \mathbb{E}(S_{N\wedge n}^2) - \sigma^2\mathbb{E}(N \wedge n). \quad (1.4)$$

   On one hand, we have $\mathbb{E}(N \wedge n) \uparrow \mathbb{E}(N)$ as $n \to \infty$. On the other, observe that since $\mathbb{E}(X_1) = 0$, then $(S_n)_{n\in I}$ is a martingale, and hence, $(S_{N\wedge n})_{n\in I}$ is a martingale. Observe that the latter satisfies, by (1.4),

$$\mathbb{E}(S_{N\wedge n}^2) = \sigma^2\mathbb{E}(N \wedge n) \leq \sigma^2\mathbb{E}(N) < \infty$$

   for all $n \in I$. Therefore, $S_{N\wedge n}$ converges *a.s.* and in L^2 to S_N. It's easy to see that convergence in L^2 implies convergence of the second moment.

□

Notice that upon observing N^*, the LLR λ_{N^*} can be approximated as a two-valued random variable taking the values $\lambda_{N^*} \leq b$ or $\lambda_{N^*} \geq a$. Therefore,

$$\mathbb{E}_\theta(\lambda_{N^*}) \approx b\mathbb{P}_\theta(\lambda_{N^*} \leq b) + a\mathbb{P}_\theta(\lambda_{N^*} \geq a) = b\mathbb{P}_\theta(\Lambda_N \leq B) + a\mathbb{P}_\theta(\Lambda_N \geq A).$$

Together with Proposition 1.16 and the approximations given by Equation 1.3, we obtain

$$\mathbb{E}_0(N) = \mu_0^{-1}\left(\alpha \log\left(\frac{1-\beta}{\alpha}\right) + (1-\alpha)\log\left(\frac{\beta}{1-\alpha}\right)\right) \tag{1.5}$$

$$\mathbb{E}_1(N) = \mu_1^{-1}\left((1-\beta)\log\left(\frac{1-\beta}{\alpha}\right) + \beta\log\left(\frac{\beta}{1-\alpha}\right)\right) \tag{1.6}$$

where $\mu_i := \mathbb{E}_i(X)$ for $i = 0, 1$.

Further higher order moments of N can be approximated by differentiating its characteristic function ([4]) and are beyond the scope of this work.

1.3.3 Optimality of the SPRT

The most remarkable property of the SPRT, first proved by Wald and Wolfowitz [22], is that in the case of *i.i.d.* observations and finite ESS under both hypothesis, the SPRT minimizes the expected sample size among all tests of the same size and power, including those of fixed sample size.

Recall that under fixed-horizon testing, given a fixed sample size and a desired test size, we can always find a test which is of the most power among all tests of the same characteristics (UMP). However, if we wish to achieve a particular level of power, we are forced to increase the sample size until we reach the desired value. Observe then that this tradeoff of increasing the sample size until both error rates are controlled comes out more naturally under the sequential paradigm, where we decide to bound first both error rates and later minimize the ESS under both hypothesis. Moreover, it's worth noting that the SPRT does not make use of any knowledge about the distribution of the likelihood ratio.

The proof of optimality relies on the following proposition:

Proposition 1.17 (**Wald's likelihood ratio identity**). *Let* $((\mathcal{X}_n, \mathcal{A}_n)_{n\in I}, \mathcal{P}, I)$ *be sequential statistical model with* $\mathcal{P} = \{\mathbb{P}_0, \mathbb{P}_1\}$. *Suppose that*

$$\mathbb{E}_1(Y_n) = \mathbb{E}_0(Y_n\Lambda_n) \tag{1.7}$$

for any $\mathcal{F}_n$*-measurable random variable* Y_n. *Then, for any stopping time* N *and non-negative random variable* Y *such that* $Y\mathbb{1}_{\{N=n\}}$ *is* $\mathcal{F}_n$*-measurable for all* n,

$$\mathbb{E}_1(Y\mathbb{1}_{\{N<\infty\}}) = \mathbb{E}_0(Y\Lambda_N\mathbb{1}_{\{N<\infty\}}).$$

In particular, if $Y = \mathbb{1}_A$ *for some* $A \in \mathcal{F}$

$$\mathbb{P}_1(A,\ N < \infty) = \mathbb{E}_0(\Lambda_N \mathbb{1}_{\{A,\ N<\infty\}}).$$

Proof. The proof repeats (1.1) with the assumption of the property of the random variable Y and (1.7) used to justify the second equality:

$$\mathbb{E}_1(Y\mathbb{1}_{\{N<\infty\}}) = \sum_{n=1}^{\infty} \mathbb{E}_1(Y\mathbb{1}_{\{N=n\}}) = \sum_{n=1}^{\infty} \mathbb{E}_0(Y\Lambda_n\mathbb{1}_{\{N=n\}}) = \mathbb{E}_0(Y\Lambda_N\mathbb{1}_{\{N<\infty\}}).$$

□

Formally, let $\mathcal{C}(\alpha,\beta)$ be the class of two simple hypothesis, sequential or fixed-horizon tests with Type-I and Type-II error probabilities at most α and β, respectively, for given $0 < \alpha, \beta < 1$, and with $\mathbb{E}_i(N) < \infty$, $i = 0, 1$. Then the following theorem holds:

Theorem 1.18 (Wald-Wolfowitz). *Let the observations* $(X_n)_{n\in I}$ *be i.i.d. with density* f_0 *under* H_0 *and with density* f_1 *under* H_1*, where* $f_0 \not\equiv f_1$ *a.e. Assume that* $\alpha+\beta < 1$*. If the bounds* (b, a) *can be selected in such a way that Type-I and Type-II error are at most* α *and* β*, respectively, then the SPRT* $\delta^* = (N^*, d^*)$ *is optimal in the class* $\mathcal{C}(\alpha,\beta)$ *under the criteria*

$$\inf_{\delta\in\mathcal{C}(\alpha,\beta)} \mathbb{E}_0(N) = \mathbb{E}_0(N^*) \quad \textit{and} \quad \inf_{\delta\in\mathcal{C}(\alpha,\beta)} \mathbb{E}_1(N) = \mathbb{E}_1(N^*).$$

Proof. It will be sufficient for this work to follow the proof by Siegmund [14] which asserts that the ESS approximations given in (1.5) are approximately minimal in the class $\mathcal{C}(\alpha,\beta)$. A rather complete proof involves several concepts out of the main scope of this work and can be found in Wald and Wolfowitz [22] or Ferguson [3].

Recall first that for any random variable Y we have $\mathbb{E}(\exp\{Y\}) \geq \exp\{\mathbb{E}(Y)\}$: indeed, since $\exp\{x\} \geq 1+x$ for all $x \in \mathbb{R}$, then $\mathbb{E}\left(\exp\{Y - \mathbb{E}(Y)\}\right) \geq 1+\mathbb{E}(Y-\mathbb{E}(Y)) = 1$, and the result follows.

Let $\delta = (N, d) \in \mathcal{C}(\alpha,\beta)$ be arbitrary. If we make use of Proposition 1.17, we have that

$$\begin{aligned}
\alpha = \mathbb{P}_0(d_N = 1) &= \mathbb{E}_1\left(\Lambda_N^{-1}\mathbb{1}_{\{d_N=1\}}\right) \\
&= E_1\left(\exp\{-\log\Lambda_N\}\middle|\, d_N = 1\right)\mathbb{P}_1(d_N = 1) \\
&\geq \exp\{-E_1\left(\log\Lambda_N\middle|\, d_N = 1\right)\}(1-\beta) \\
&= \exp\{-\mathbb{E}_1\left(\log\Lambda_N\mathbb{1}_{\{d_N=1\}}\right)/(1-\beta)\}(1-\beta).
\end{aligned}$$

Hence, taking logarithms yields

$$(1-\beta)\log\left(\frac{\alpha}{1-\beta}\right) \geq -\mathbb{E}_1(\log \Lambda_N \mathbb{1}_{\{d_N=1\}}) \tag{1.8}$$

Analogously, we obtain

$$\beta \log\left(\frac{1-\alpha}{\beta}\right) \geq \mathbb{E}_1(\log \Lambda_N \mathbb{1}_{\{d_N=0\}}) \tag{1.9}$$

Adding (1.8) and (1.9) together, and using the first identity from Proposition 1.16, we obtain

$$(1-\beta)\log\left(\frac{\alpha}{1-\beta}\right) + \beta \log\left(\frac{1-\alpha}{\beta}\right) \geq -\mathbb{E}_1(\log \Lambda_N) = -\mu_1 \mathbb{E}_1(N),$$

which is equivalent to the approximation given by (1.5). Proceeding similarly we obtain a lower bound equivalent to the other approximation.

□

1.4 The mixture Sequential Probability Ratio Test (mSPRT)

Let $((\mathcal{X}_n, \mathcal{A}_n)_{n\in I}, \mathcal{P}, I)$ be a sequential statistical model. For the rest of this chapter we will assume that $\mathcal{P} = \{\mathbb{P}_\theta,\ \theta \in \Theta\}$ with the density f_θ belonging to the natural exponential type family of probability distributions with $\Theta \subset \mathbb{R}$, i.e., we assume that

$$f_\theta(x) = h(x)\exp\{\theta x - \psi(\theta)\}$$

for all $\theta \in \Theta$, where $h : \mathcal{X} \to \mathbb{R}$ is a sufficient statistic and $\psi : \Theta \to \mathbb{R}$ is a strictly convex function known as the *log-partition function*. Moreover, we also assume that $h(x) = 1$.

Suppose we wish to test

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}.$$

For each sampling stage $n \in I$, recall the likelihood ratio between f_{θ_0} and f_θ for arbitrary $\theta \in \Theta_1$:

$$\Lambda_n^\theta(\mathbf{X}_n) = \prod_{i=1}^n \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} = \exp\{(\theta - \theta_0)S_n - n(\psi(\theta) - \psi(\theta_0))\}$$

where $S_n \coloneqq \sum_{i=1}^n X_i$. Given a probability density function $\pi : \Theta \to \mathbb{R}$ known as the *mixing distribution* we can then consider the *mixture likelihood ratio* (mLR), defined as

$$\bar{\Lambda}_n^\pi(\mathbf{X}_n) = \int_\Theta \Lambda_n^\theta(\mathbf{X}_n)\pi(\theta)\, d\theta.$$

Definition 1.19 (mSPRT). The *mixture Sequential Probability Ratio Test* (mSPRT) with boundary $a > 0$ and mixing distribution $\pi : \Theta \to \mathbb{R}$ for testing the simple null hypothesis $H_0 : \theta = \theta_0$ against the composite alternative $H_1 : \theta \neq \theta_0$ is the open-ended test $\delta = (N, d)$ such that

$$N = \inf\{n \in I : \bar{\Lambda}_n^\pi(\mathbf{X}_n) \geq a\}$$

Hence, the mSPRT runs indefinitely under the assumption that H_0 is true until the mLR first excesses the fixed boundary a, in which case we reject H_0 in favor of H_1. It's also worth noting that the pdf π is sometimes regarded as a Bayesian *prior* for the parameter θ over the values of Θ_1, meaning that if θ were to be treated as a random variable, then π would encapsulate all information known about θ assuming H_1 to be true before observing the data $\mathbf{X}_n$ at sampling stage $n \in I$ ([2]).

1.4.1 Type-I error control

With the mSPRT being an open-ended test, we shall prove that Type-I error can indeed be bounded at any predefined level and for any choice of mixing distribution. Formally, we wish to find $a := a(\alpha) > 0$ such that

$$\mathbb{P}_{\theta_0}(d = 1) = \mathbb{P}_{\theta_0}(N < \infty) = \mathbb{P}_{\theta_0}(\bar{\Lambda}_n^{\pi} \geq a) \leq \alpha$$

for any $\alpha \in (0, 1)$ and pdf $\pi : \Theta \to \mathbb{R}$.

In this case, notice that since $f_{\theta_0} \not\equiv f_\theta$ *a.e.* for all $\theta \in \Theta_1$, then $(\Lambda_n^\theta)_{n\in I}$ is a martingale under the null hypothesis. Indeed, for every $n \in I$ observe that

$$\begin{aligned}
\mathbb{E}_{\theta_0}|\Lambda_n^\theta| &= \int_{\mathcal{X}_n} \Lambda_n^\theta \, f_{\theta_0}(\xi_1) \ldots f_{\theta_0}(\xi_n) \, d\xi_1 \ldots d\xi_n \\
&= \int_{\mathcal{X}_n} f_\theta(\xi_1) \ldots f_\theta(\xi_n) \, d\xi_1 \ldots d\xi_n \\
&= 1
\end{aligned}$$

Moreover,

$$\begin{aligned}
E_{\theta_0}(\Lambda_{n+1}^\theta|\mathcal{F}_n) &= E_{\theta_0}\left(\frac{f_\theta(X_{n+1})}{f_{\theta_0}(X_{n+1})}\Lambda_n^\theta \middle| \mathcal{F}_n\right) = \Lambda_n^\theta E_{\theta_0}\left(\frac{f_\theta(X_{n+1})}{f_{\theta_0}(X_{n+1})}\middle| \mathcal{F}_n\right) \\
&= \Lambda_n^\theta \mathbb{E}_{\theta_0}\left(\frac{f_\theta(X_{n+1})}{f_{\theta_0}(X_{n+1})}\right) = \Lambda_n^\theta \int_{\mathcal{X}} \frac{f_\theta(\xi)}{f_{\theta_0}(\xi)} f_{\theta_0}(\xi) \, d\xi \\
&= \Lambda_n^\theta \int_{\mathcal{X}} f_\theta(\xi) \, d\xi = \Lambda_n^\theta.
\end{aligned}$$

By Fubini's theorem it follows that $(\bar{\Lambda}_n^\pi)_{n\in I}$ is also a martingale under the same assumptions. Recall now Doob's First Martingale Inequality:

Theorem 1.20. (Doob's First Martingale Inequality) *Let $M = (M_n)_{n\geq 0}$ be a martingale. Let $M_n^* = \sup_{0\leq j\leq n} |M_j|$. Then*

$$\mathbb{P}\left(M_n^* \geq a\right) \leq \frac{\mathbb{E}(|M_n|)}{a}$$

Proof. Let $T = \min\{j : |M_j| \geq a\}$. Since the absolute value $\varphi(x) = |x|$ is a convex function and is increasing on $\mathbb{R}_+$, then it follows that $(|M_n|)_{n\geq 0}$ is a submartingale. Since $\{T \leq n, |M_T| \geq a\} = \{M_n^* \geq a\}$, then

$$\mathbb{P}(M_n^* \geq a) = \mathbb{P}(T \leq n, |M_T| \geq a) \leq \mathbb{E}\left(\frac{|M_T|}{a}\mathbb{1}_{\{T\leq n\}}\right)$$

and since $M_T = M_{T\wedge n}$ on $\{T \leq n\}$,

$$\mathbb{P}(M_n^* \geq a) \leq \frac{1}{a}\mathbb{E}\left(|M_{T\wedge n}|\mathbb{1}_{\{T\leq n\}}\right) \leq \frac{\mathbb{E}\left(|M_{T\wedge n}|\right)}{a} \leq \frac{\mathbb{E}\left(|M_n|\right)}{a}.$$

□

Provided that $(\bar{\Lambda}_n^{\pi})_{n\in I}$ is a martingale under H_0, we can apply Theorem 1.20 to obtain $\mathbb{P}_{\theta_0}(\bar{\Lambda}_n^{\pi} \geq a) \leq 1/a$ for all $a > 0$. Hence, by setting $a := 1/\alpha$ we obtain

$$\mathbb{P}_{\theta_0}(N < \infty) = \mathbb{P}_{\theta_0}(\bar{\Lambda}_n^{\pi} \geq \alpha^{-1}) \leq \alpha \tag{1.10}$$

for all $n \in I$.

1.4.2 Expected sample size of the mSPRT

This last section will expose an approximation to the ESS of the mSPRT, as well as some auxiliary results that we will use in further chapters.

Our main result, which we will prove, was introduced by Pollak and Siegmund [10, Theorem 1], and gives an asymptotic approximation for the expected sample size of the mSPRT in the limit as $a \to \infty$ ($\alpha \to 0$ by last section) when $\theta \neq \theta_0$. For the rest of this section, $\delta = (N, d)$ will denote an arbitrary mSPRT, and we will denote its mixing distribution by g to avoid confusion. Moreover, denote by

$$I(\theta, \theta_0) := (\theta - \theta_0)\psi'(\theta) - (\psi(\theta) - \psi(\theta_0)).$$

The proof of the theorem relies on several lemmas which are described later. To simplify statements, all of our auxiliary results are proved under the same assumptions (and same notation) as the theorem:

Theorem 1.21 (**Pollak and Siegmund**). *Suppose $\theta \neq \theta_0$ such that g exists in a neighborhood of θ and is positive and continuous at θ. Then as $a \to \infty$,*

$$\mathbb{E}_{\theta}(N) \approxeq [2\log a + \log(\log a / I(\theta)) - \log(2\pi g(\theta)^2/\psi''(\theta)) - 1]/2I(\theta) + o(1) \tag{1.11}$$

Proof. Put $\mu := \mathbb{E}_{\theta}(X_1)$ and $\sigma^2 := \mathbb{V}_{\theta}(X_1)$. Then, X_1 has finite moment-generating function and by differentiating it, it's easy to see that $\psi'(\theta) = \mu$ and $\psi''(\theta) = \sigma^2$. We can center X_1 to obtain $\psi'(\theta_0) = \mathbb{E}_{\theta_0}(X_1) = 0$, and with its distribution being from the exponential type family we can assume without loss of generality that $\theta_0 = 0$ and $\psi(\theta_0) = 0$. Under this framework,

$$\bar{\Lambda}_n^{g} := \bar{\Lambda}_n^{g}(\mathbf{X}_n) = \int_{\Theta} \exp\{\eta S_n - n\psi(\eta)\}\, g(\eta) d\eta$$

and denote $I(\theta) \coloneqq I(\theta, 0)$. Assume also (again, without loss of generality) that $\theta > 0$, and hence, by the strict convexity of the log-partition function ψ, we have $\mu = \psi'(\theta) > 0$.

By Lemma 1.22 we are assured that $\mathbb{E}_\theta(N) < \infty$; then, by the definition of N,

$$\log a \leq \theta S_N - N\psi(\theta) - \frac{1}{2}\log N \tag{1.12}$$

$$+ \log\left(N^{1/2}\int_\Theta \exp\{(\eta-\theta)S_N - N(\psi(\eta)-\psi(\theta))\}g(\eta)\,d\eta\right) \tag{1.13}$$

Let $0 < \varepsilon < 1$ and $0 < \delta_1 < \delta_2$ be arbitrary. Also, let $n_1 = (1-\varepsilon)\log a/I(\theta)$ and $A = \{N > n_1,\ \max_{n\geq n_1}\left|n^{-1}S_n - \mu\right| < \delta_1\}$. Then,

$$\mathbb{P}_\theta(A^c) \leq \mathbb{P}_\theta(N \leq n_1) + \mathbb{P}_\theta\left(\max_{n\geq n_1}\left|n^{-1}S_n - \mu\right| \geq \delta_1\right)$$

and hence by Lemma 1.23 and Lemma 1.24, there exists $\lambda > 0$ such that

$$\mathbb{P}_\theta(A^c) = O(a^{-\lambda}) \text{ as } a \to \infty. \tag{1.14}$$

By Proposition 1.16,

$$\begin{aligned}\int_A (\theta S_N - N\psi(\theta))f_\theta(x)\,dx &= \int_A (\theta(S_N-\mu N) + NI(\theta))f_\theta(x)\,dx \\ &= I(\theta)\mathbb{E}_\theta(N) - \theta\int_{A^c}(S_N-\mu N)f_\theta(x)\,dx - I(\theta)\int_{A^c} N f_\theta(x)\,dx.\end{aligned} \tag{1.15}$$

Then by Schwarz's inequality, Wald's Lemma for Squared Sums and Proposition 1.26 we obtain

$$0 < \int_{A^c} N f_\theta(x)\,dx \leq \left(\mathbb{E}_\theta(N^2)\mathbb{P}_\theta(A^c)\right)^{1/2} = o(1) \tag{1.16}$$

and

$$\begin{aligned}\int_{A^c}(S_N-\mu N)f_\theta(x)\,dx &\leq \left(\mathbb{E}_\theta(S_N-\mu N)^2\mathbb{P}_\theta(A^c)\right)^{1/2} \\ &= \left(\mathbb{E}_\theta(N)\sigma^2\mathbb{P}_\theta(A^c)\right)^{1/2} = o(1)\end{aligned} \tag{1.17}$$

as $a \to \infty$. Hence, combining (1.16) and (1.17) with (1.15) we obtain

$$\int_A (\theta S_N - N\psi(\theta))f_\theta(x)\,dx = I(\theta)\mathbb{E}_\theta(N) + o(1) \tag{1.18}$$

Now, let $0 < \gamma < 1$ be arbitrary, and fix $\delta_2 < \min\{\mu, I(\theta)/\theta\}$ such that $g(\eta)$ is defined for all $|\eta - \theta| < \delta_2$. Then,

$$g(\theta)(1-\gamma) \le g(\eta) \le g(\theta)(1+\gamma) \tag{1.19}$$

and

$$\frac{1}{2}(\eta-\theta)^2\sigma^2(1-\gamma) \le \psi(\eta) - \psi(\theta) - (\eta-\theta)\mu \le \frac{1}{2}(\eta-\theta)^2\sigma^2(1+\gamma) \tag{1.20}$$

Recall the term (1.13): if we consider the integral over the values of $|\eta - \theta| < \delta_2$, then using (1.19) and (1.20) we obtain

$$\begin{aligned}
&N^{1/2}\int_{|\eta-\theta|<\delta_2} \exp\{(\eta-\theta)S_N - N(\psi(\eta)-\psi(\theta))\}g(\eta)\,d\eta \\
&\quad\le (2\pi)^{1/2}\exp\left\{\frac{(S_N-\mu N)^2}{2\sigma^2(1-\gamma)N}\right\}N^{1/2} \\
&\qquad\times\int_{|\eta-\theta|<\delta_2}\phi\left((\sigma^2(1-\gamma)N)^{1/2}\left(\eta-\theta-\left(\frac{S_N-\mu N}{\sigma^2(1-\gamma)N}\right)\right)\right)g(\eta)\,d\eta \\
&\quad\le \left(\frac{2\pi}{\sigma^2(1-\gamma)}\right)^{1/2}\exp\left\{\frac{(S_N-\mu N)^2}{2\sigma^2(1-\gamma)N}\right\}g(\theta)(1+\gamma)
\end{aligned} \tag{1.21}$$

with $\phi(x)$ being the density of the standard normal distribution.

We know by Lemma 1.25 that for sufficiently small δ_1, δ_2 on A

$$N^{1/2}\int_{|\eta-\theta|<\delta_2}\exp\{(\eta-\theta)S_N - N(\psi(\eta)-\psi(\theta))\}g(\eta)\,d\eta \le \varepsilon(a)$$

where $\varepsilon(a)$ is nonrandom and $\varepsilon(a) \to 0$ as $a \to \infty$. Hence using (1.21),

$$\begin{aligned}
&\int_A \log\left(N^{1/2}\int_{|\eta-\theta|<\delta_2}\exp\{(\eta-\theta)S_N - N(\psi(\eta)-\psi(\theta))\}g(\eta)\,d\eta\right)f_\theta(x)\,dx \\
&\quad\le \frac{1}{2}\log\left(\frac{2\pi}{\sigma^2(1-\gamma)}\right) + \log\left(g(\theta)(1+\gamma)\right) \\
&\qquad + \left(2\sigma^2(1-\gamma)\right)^{-1}\int_A \frac{(S_N-\mu N)^2}{N}f_\theta(x)\,dx + o(1).
\end{aligned} \tag{1.22}$$

Moreover, by the definition of A and using Wald's Lemma for Squared Sums and Proposition 1.26 it follows that the term

$$\begin{aligned}
\int_A \frac{(S_N-\mu N)^2}{N}f_\theta(x)\,dx &\le {n_1}^{-1}\mathbb{E}_\theta(S_N-\mu N)^2 \\
&= \sigma^2 n_1^{-1}\mathbb{E}_\theta(N) = \sigma^2(1-\varepsilon)^{-1} + o(1).
\end{aligned} \tag{1.23}$$

Then, if we integrate the whole inequality (1.12) with respect to $\mathbb{P}_\theta$ on A, using (1.14), (1.18), (1.22) and (1.23) we obtain as $a \to \infty$

$$I(\theta)\mathbb{E}_\theta(N) \geq \log a + \frac{1}{2}\log n_1 - \frac{1}{2}\log\left(\frac{2\pi(g(\theta)(1+\gamma))^2}{\sigma^2(1-\gamma)}\right) - (2(1-\gamma)(1-\varepsilon))^{-1} + o(1). \tag{1.24}$$

Finally, since ε and γ are arbitrary, if we neglect the excess over the boundary

$$\mathbb{E}_\theta\left(\log \bar{\Lambda}_N^g\right) - \log a,$$

our result follows. □

Lemma 1.22. $\mathbb{E}_\theta(N) < \infty$.

Proof. Let $\tau = \min\{N, n\} - 1$, so $\tau + 1$ is a stopping time *w.r.t.* $\mathbb{F}$. Let $|\eta - \theta| < \delta$ be a neighbourhood where g is defined. From the definition of N we know that

$$\log a \geq \theta S_\tau - \psi(\theta)\tau + \log\left(\int_{|\eta-\theta|<\delta} \exp\{(\eta-\theta)S_\tau - \tau(\psi(\eta)-\psi(\theta))\}g(\eta)\,d\eta\right) \tag{1.25}$$

Using the Taylor's expansion around θ

$$\psi(\eta) = \psi(\theta) + (\eta-\theta)\psi'(\theta) + \frac{1}{2}(\eta-\theta)^2\psi''(\eta)$$

and restricting δ such that $|\eta-\theta|\,\psi''(\eta) \leq 1$ for all $|\eta-\theta| < \delta$,

$$\log a > \theta(S_\tau - \mu\tau) + I(\theta)\tau + \log\left(\int_{|\eta-\theta|<\delta} \exp\{(\eta-\theta)(S_\tau-\mu\tau)\}g(\eta)\,d\eta\right) - \delta_\tau \tag{1.26}$$

and hence by Jensen's inequality

$$\log a \geq \theta(S_\tau - \mu\tau) + (I(\theta)-\delta)\tau - \log\left(\int_{|\eta-\theta|<\delta} g(\eta)\,d\eta\right) + (c-\theta)(S_\tau-\mu\tau), \tag{1.27}$$

with

$$c := \frac{\int_{|\eta-\theta|<\delta}\eta g(\eta)\,d\eta}{\int_{|\eta-\theta|<\delta} g(\eta)\,d\eta} = \int_{|\eta-\theta|<\delta}\eta\,d\eta.$$

Therefore,

$$\begin{aligned}(I(\theta)-\delta)\tau \leq & -c(S_{\tau+1}-\mu(\tau+1))+c\,|X_{\tau+1}-\mu| \\ & +\log a-\log\left(\int_{|\eta-\theta|<\delta} g(\eta)\,d\eta\right).\end{aligned} \tag{1.28}$$

By Proposition 1.16 and Schwarz's inequality, $\mathbb{E}_\theta(S_{\tau+1}-\mu(\tau+1))=0$ and

$$\mathbb{E}_\theta(|X_{\tau+1}-\mu|) \leq \left(\mathbb{E}_\theta\left(\sum_{k=1}^{\tau+1}(X_k-\mu)^2\right)\right)^{1/2} = \left(\sigma^2\mathbb{E}_\theta(\tau+1)\right)^{1/2}$$

Hence, if we pick δ sufficiently small such that $I(\theta)-\delta>0$, taking expectations on (1.28) yields an upper bound for $\mathbb{E}_\theta(\tau)$ as $n\to\infty$.

□

Lemma 1.23. *For all $\delta>0$ there exists $0<\lambda<\infty$, $0<\alpha<\infty$ such that*

$$\mathbb{P}_\theta\left(\max_{n\geq r}\left|\frac{S_n}{n}-\mu\right|\geq\delta\right)\leq\alpha\exp\{-\lambda r\}$$

Proof. We have

$$\mathbb{P}_\theta\left(\max_{n\geq r}\left|\frac{S_n}{n}-\mu\right|\geq\delta\right)\leq\sum_{n=r}^{\infty}\mathbb{P}_\theta\left(\left|\frac{S_n}{n}-\mu\right|\geq\delta\right)$$

therefore, for all $\xi>\theta>0$,

$$\begin{aligned}\mathbb{P}_\theta\left(S_n-n\mu\geq n\delta\right) &= \int_{\{S_n-n\mu\geq n\delta\}}\exp\{(\theta-\xi)S_n-n(\psi(\theta)-\psi(\xi))\}f_\xi(x)\,dx \\ &\leq \exp\{-n((\mu+\delta)(\xi-\theta)-(\psi(\theta)-\psi(\xi)))\}\mathbb{P}_\xi\left(S_n-n\mu\geq n\delta\right).\end{aligned}$$

Now, since $\mu=\psi'(\theta)$, then $\psi(\xi)-\psi(\theta)\approxeq\mu(\xi-\theta)$ as $\xi\downarrow\theta$. Hence, if we pick ξ sufficiently close to θ, there exists $\lambda_1:=\lambda_1(\delta,\theta)>0$ such that

$$\mathbb{P}_\theta\left(S_n-n\mu\geq n\delta\right)\leq\exp\{-\lambda_1 n\}.$$

A similar argument yields

$$\mathbb{P}_\theta\left(S_n-n\mu\leq -n\delta\right)\leq\exp\{-\lambda_2 n\}$$

for some $\lambda_2:=\lambda_2(\delta,\theta)>0$. Therefore, there exists $\lambda:=\lambda(\delta,\theta)>0$ such that

$$\mathbb{P}_\theta\left(|S_n-n\mu|\geq n\delta\right)\leq\exp\{-\lambda n\}. \tag{1.29}$$

The result follows from summing over $n\geq r$. □

Lemma 1.24. *Let* $0 < \varepsilon < 1$ *and* $n_1 = (1-\varepsilon)\log a/I(\theta)$. *There exists* $\lambda > 0$ *such that*

$$\mathbb{P}_\theta(N \le n_1) = O(a^{-\lambda}) \text{ as } a \to \infty.$$

Proof. For all $x > 0$

$$\begin{aligned}\mathbb{P}_\theta(N \le n_1) &\le \mathbb{P}_\theta\left(S_{n_1} - \mu n_1 \ge x(\sigma^2 n_1)^{1/2}\right) \\ &\quad + \int_{\{N \le n_1,\ S_{n_1} - \mu n_1 < x(\sigma^2 n_1)^{1/2}\}} \exp\{\theta S_{n_1} - n_1\psi(\theta)\} f_0(x)\, dx \\ &\le \mathbb{P}_\theta\left(S_{n_1} - \mu n_1 \ge x(\sigma^2 n_1)^{1/2}\right) \\ &\quad + \exp\{I(\theta)n_1 + \theta x(\sigma^2 n_1)^{1/2}\}\mathbb{P}_0(N \le n_1).\end{aligned} \tag{1.30}$$

We know by (1.10) that $\mathbb{P}_0(N < \infty) \le a^{-1}$, then for $x = \varepsilon(I(\theta)\log a)^{1/2}/2\theta\sigma$, equation (1.30) yields an upper bound which is $O(a^{-\varepsilon/2})$. The result then follows by using (1.29). □

Lemma 1.25. *Given* $\delta_2 < \min(\mu, I(\theta)/\theta)$, *for all* δ_1 *satisfying that for all* $\eta \ge \theta + \delta_2$

$$\psi(\eta) \ge \psi(\theta) + (\eta - \theta)(\psi'(\theta) + 2\delta_1) \tag{1.31}$$

and

$$4\delta_1 < \delta_2 \inf_{0 \le x \le \theta} \psi''(x), \tag{1.32}$$

then on A,

$$\int_{|\eta-\theta|>\delta_2} N^{1/2} \exp\{(\eta-\theta)S_N - N(\psi(\eta) - \psi(\theta))\} g(\eta)\, d\eta \le \varepsilon(a) \tag{1.33}$$

where $\varepsilon(a)$ *is a nonrandom quantity such that* $\varepsilon(a) \to 0$ *as* $a \to \infty$.

Proof. By the mean value theorem, there exists $\xi \in (\theta, \eta)$ such that

$$\psi(\eta) = \psi(\theta) + (\eta - \theta)\psi'(\xi).$$

By the strict convexity of ψ, ψ' is strictly increasing and hence $\xi := \xi(\eta)$ is an increasing function of η for $\eta > \theta$. Split the integration (1.33) into the domains $\{\eta \le 0\}$, $\{0 < \eta < \theta - \delta_2\}$ and $\{\eta > \theta + \delta_2\}$, and denote these integrals by I_1, I_2 and I_3, respectively. Then on A for sufficiently large a

$$I_1 \le N^{1/2} \exp\{-(\theta(\mu - \delta_1) - \psi(\theta))N\} \le {n_1}^{1/2} \exp\{-n_1(I(\theta) - \delta_1\theta)\}. \tag{1.34}$$

Moreover, using (1.31), on A for large a,

$$\begin{aligned} I_3 &\leq N^{1/2} \int_{\{\eta \geq \theta + \delta_2\}} \exp\{(\eta - \theta)(\mu + \delta_1)N - N(\eta - \theta)(\mu + 2\delta_1)\} g(\eta) d\eta \\ &\leq n_1^{1/2} \exp\{-\delta_1 \delta_2 n_1\}. \end{aligned} \tag{1.35}$$

If we also develop a two-term Taylor series expansion and use (1.32), under the same previous assumptions

$$\begin{aligned} I_2 &\leq N^{1/2} \int_{\{0 < \eta < \theta - \delta_2\}} \exp\{N\,|\eta - \theta|\,(\delta_1 - \frac{1}{2}\,|\eta - \theta|\,\psi''(\xi))\} g(\eta)\; d\eta \\ &\leq \int_{\{0 < \eta < \theta - \delta_2\}} \exp\{N\,|\eta - \theta|\,(\delta_1 - 2\delta_1)\} g(\eta)\; d\eta \\ &\leq {n_1}^{1/2} \exp\{-\delta_1 \delta_2 n_1\}. \end{aligned} \tag{1.36}$$

The result follows from (1.34), (1.36) and (1.35). □

We now highlight our last auxiliary result since it characterizes the run-length of the mSPRT, and it will be of use in the next chapter.

Proposition 1.26. *$N \xrightarrow{\mathbb{P}} \log a / I(\theta)$ and $N \xrightarrow{L^p} \log a / I(\theta)$ for $p = 1, 2$.*

Proof. Convergence in probability follows easily from (1.13), (1.14), (1.21) and Lemma 1.25. Hence,

$$\liminf_{a \to \infty} \mathbb{E}_\theta(N^p)/(\log a)^p \geq (I(\theta))^{-p} \text{ for } p = 1, 2.$$

Reasoning as in the proof of Lemma 1.22 we obtain equation (1.28) with τ replaced by $N - 1$. If we apply Proposition 1.16, we obtain

$$\liminf_{a \to \infty} \mathbb{E}_\theta(N^p)/(\log a)^p \leq (I(\theta))^{-p} \text{ for } p = 1, 2.$$

□

Note that the ESS given by (1.11) can be equivalently expressed as

$$\mathbb{E}_\theta(N) \approxeq \frac{1}{2I(\theta)}[2 \log a + \log \log a] + O(1).$$

Further work along the same line of research as of Theorem 1.21 shows the mSPRT with arbitrary mixing distribution to be asymptotically second-order optimal. Formally, the result is as follows, and its proof was developed by Pollak [11, Theorem 2]:

Theorem 1.27. *Suppose $g : [b, c] \to \mathbb{R}$ is continuous and positive, with $\theta_0 < b < c < \infty$ and $[b, c] \subseteq \Theta_1$. Suppose $\theta \neq \theta_0$. Then,*

$$\inf_{\{N:\, \mathbb{P}_{\theta_0}(N<\infty)\leq a^{-1}\}} \sup_{b\leq\theta\leq c} \mathbb{E}_\theta(N) = \frac{1}{2I(\theta)}[2\log a + \log\log a] + O(1)$$

as $a \to \infty$.

Chapter 2

Application to online controlled experiments

Controlled experimentation has recently experienced a major rise on its industrial use on fields other than traditional ones like pharmacy or auditing, mostly due to the increase in the generation and availability of large amounts of data.

The simplest (but most used) example of controlled experiment is the one know as the *A/B test*. An A/B test consists of the deployment of two different *variants* of a product, A and B, into two equitable groups of users, know as the *control* and *treatment* groups, respectively. The control variant is usually considered as the default or already existing version of the product, and the treatment variant is rather considered as a new iteration we wish to contrast.

Each user in each group makes use of their respective version of the product and generates a quantitative measure or metric of our interest, which we record. The objective is to assess whether variant B generates a significant increase (or decrease) in the value of our metric with respect to variant A, or else both variants are the same. In other words, we say that we want to check whether an *effect* exists when modifying features from variant A to variant B. If such an effect exists, we then wish to *detect* it.

It's easy to see that our problem can be formalized as a hypothesis test, where we test the simple null hypothesis that no difference between the two groups exists versus the composite alternative that the groups are indeed different. In the case of regular metric values, the data generating process could be assumed to be normally distributed, so we would test the difference between the means of the two groups. For binary data, we would test for the difference in rates.

Our focus on A/B testing is put on online controlled experimentation, with online connoting experiments deployed on digital interfaces: for example, website visitors which are randomized to two different variants of the site in a persistent manner (meaning that they will continue receiving the same variant after ending the session), where difference in variants may simply be aesthetic changes to the user interface (UI), and we may record relevant metrics such as time spent on the site by the user or whether they clicked on a certain button or not.

The largest digital companies like Google, Facebook, Netflix or Microsoft run thousands of A/B tests every year involving sample sizes of millions of users with the objective to assess changes on UI, relevance algorithms or customer support systems. A relevant example on the power of these procedures is described in Kohavi et al. [8, Chapter 2], where a simple change on the ad headlines display of Microsoft's search engine Bing led to a $100M annual return-on-investment. Most of these big companies have built fully-fledged internal tools for running controlled experiments of this kind, known as *experimentation engines* or *experimentation platforms* ([20], [15]). However, the widespread of A/B testing has led to the creation of companies specialized in the development of tools for online controlled experimentation, where Optimizely, Statsig or Eppo are some of the most popular.

A convenient commercial experimentation platform may provide a simple, easy-to-use dashboard to track ongoing A/B tests, and it may do so without advanced statistical training requirements over its users. Classical hypothesis testing methods are used to compute the experiment results, which are reported to the user via parameters like p-values and confidence intervals, due to the simple decision rule that these define. In most cases, these platforms also allow their users to continuously monitor their tests as new observations are received, in what is known as *fixed-horizon testing*: the user fixes a sample size N and a significance level α at the start of the experiment, and after each observation received X_n a p-value p_n is computed using the n-dimensional sample, rejecting H_0 after reaching N observations if and only if $p_N \leq \alpha$ ([6]). Note that the p_n for $n < N$ are computed only to allow the user to observe the development of the experiment.

Continuous monitoring finds its value in the ability to detect true effects as quickly as possible, or else stopping the experiment earlier in case no effect is ever noticed. Nonetheless, commercial platforms were found to be also allowing their users to dynamically modify the experiment's sample size based on the reported data. This practice, known as *peeking* by the experimentation community, truncates the statistical validity of the test, inflating Type-I error to levels beyond the predefined significance level α.

Johari et al. [6] explores statistical methods to allow the continuous monitoring of A/B tests, providing valid and efficient inference on the simple interface of an experimentation platform. In order to do so, the authors highlight the following objectives:

1. Maintain a simple reporting interface: inference shall be made using traditional A/B testing parameters like p-values or confidence intervals.

2. Type-I error shall be controlled under any stopping time with respect to the observed data.

3. Efficiently trade-off runtime and detection: if the user is willing to wait until the p-value drops below its predefined significance level α, the resulting decision rule shall efficiently trade-off the overall experiment runtime and the test's detection (power). Moreover, this trade-off shall be obtained with no previous knowledge on the user's preferences.

This paper is considered to be the main contributor to the introduction of sequential analysis to the A/B testing space, presenting this paradigm as the key to allow the continuous monitoring of controlled experiments while maintaining statistical rigour.

2.1 Formalization of A/B tests

Consider an online controlled experiment, and let $(X_n)_{n\in I}$ and $(Y_n)_{n\in I}$ be the two independent sequences of *i.i.d.* observations corresponding to the metric data generated by the visitors of the control and treatment groups, respectively. As discussed previously, we can assume that these observations are normally or Bernoulli distributed. However, during this work we will focus on the first case.

Suppose then that $X_n \sim \mathcal{N}(\mu_0,\ \sigma^2)$ and $Y_n \sim \mathcal{N}(\mu_1,\ \sigma^2)$, where $\mu_0, \mu_1 \in \mathbb{R}$ are unknown and $\sigma^2 \in \mathbb{R}$ is common and assumed to be known. Our parameter of interest, however, is the difference of these means.

Definition 2.1 (**Effect size**)**.** The difference between the group means $\theta = \mu_1 - \mu_0$ is know as the *effect size*.

Given that the deployment of both variants is assumed to be equitable, we can make the simplification that visitors arrive as a sequence of *i.i.d.* pairs $(W_n)_{n\in I}$, with $W_n = (X_n, Y_n)$. We can then consider the difference $Z_n \coloneqq Y_n - X_n$, and hence $Z_n \sim \mathcal{N}(\theta, 2\sigma^2)$.

In this context, at each sample stage $n \in I$ we gather an n-dimensional sample $\mathbf{Z}_n = (Z_1, \ldots, Z_n)$, and hence we can fix the following sequential statistical model $((\mathcal{X}_n, \mathcal{A}_n)_{n\in I}, \mathcal{P}, I)$:

1. $(\mathcal{X}_n, \mathcal{A}_n) = (\mathbb{R}^n,\ \mathcal{B}(\mathbb{R})^n)$ for all $n \in I$.
2. $\mathcal{P} = \{\mathbb{P}_\theta,\ \theta \in \Theta\}$, where $\mathbb{P}_\theta$ is the probability measure corresponding to the density $f_\theta(x) = \frac{1}{\sqrt{2}\sigma}\phi\left(\frac{x-\theta}{\sqrt{2}\sigma}\right)$ and $\Theta = \mathbb{R}$.
3. $I = \mathbb{N}$, allowing the possibility that the experiment runs indefinitely.

Under this framework, we can formalize our online A/B test as the contrast

$$\begin{cases} H_0 : \theta = 0 \\ H_1 : \theta \neq 0 \end{cases}.$$

Note then that detecting that a true effect exists depends on the power of the test that we use.

2.2 Always Valid Inference

Having formalized our A/B testing problem as a sequential hypothesis test, we would now be ready to propose a suitable statistical method to assess it. However, the authors emphasize the importance of maintaining a simple inference framework through the use of p-values.

Recall fixed-horizon testing theory: for any $n \in \mathbb{N}$, there exists a family of uniformly most powerful (UMP) tests parameterized by α, which maximize power uniformly over all $\theta \neq 0$ while maintaining a test size of α. Furthermore, these tests reject H_0 if a test statistic $\tau_n \geq k(\alpha)$, where $k(\alpha)$ is a particular threshold which only depends on the test size α. Equivalently, the same decision rule can be obtained by computing a p-value as

$$p_n = \inf\{\alpha : \tau_n \geq k(\alpha)\} \tag{2.1}$$

rejecting H_0 if and only if $p_n \leq \alpha$.

As introduced earlier, commercial platforms traditionally allowed continuous monitoring of experiments via fixed-horizon testing. Formally, the user commits to a sample size $N \in \mathbb{N}$ and a significance level $\alpha \in (0, 1)$ in advance, generating a sequence of p-values $(p_n)_{n=1}^N$ computed according to (2.1) using $\mathbf{Z}_n$, and rejecting the null hypothesis in case $p_N \leq \alpha$. Observe that this approach provides a simple

and user-friendly decision rule that controls Type-I error and maximizes power for any choice of (N, α), with the only drawback that N must be estimated prior to the experiment in order to achieve a desired level of detection, which is commonly known as the *Minimum Detectable Effect* (MDE). This estimate is then highly sensible to the MDE and the magnitude of the effect size θ.

Observe that the property that for each $n \in \mathbb{N}$, the family of UMP tests, and consequently, the decision rules defined by p_n, controls Type-I error, can be equivalently expressed as the p_n being *superuniform* under the null hypothesis, i. e.

$$\forall x \in [0, 1],\ \mathbb{P}_0(p_n \leq x) \leq x.$$

This first definition then aims to generalize our sequence of UMP-generated p-values obtained via fixed-horizon testing:

Definition 2.2 (**Fixed-horizon p-value process**)**.** A *fixed-horizon p-value process* is any sequence $(p_n)_{n \in I}$ of $[0, 1]$-valued, $\mathcal{F}_n$-measurable random variables p_n such that for all $n \in I$, p_n is *superuniform* under the null hypothesis.

But recall that our objective is to allow the user to stop the experiment whenever they want, following any data-dependent rule, and with the resulting p-value controlling Type-I error.

Definition 2.3 (**Always valid p-value process**)**.** We say that a fixed-horizon p-value process $(p_n)_{n \in I}$ is *always valid* if for any (possibly infinite) stopping time T it holds

$$\forall x \in [0, 1],\ \mathbb{P}_0(p_T \leq x) \leq x.$$

Having adapted p-values to our particular problem, our next step is to find a way to construct always valid p-values using sequential tests. Recall that our A/B testing problem involves testing a simple null hypothesis against a composite alternative, with our contrast assessing whether a true effect exists (in other words, to detect) when comparing a default version against an alternative variation. Moreover, we are comfortable allowing the experiment to run indefinitely, only stopping in case an effect is detected, or in case the user decides to. Hence, it is natural to approach this sequential testing problem using open-ended tests.

The following theorem states a natural correspondence between open-ended tests and always valid p-value processes.

Theorem 2.4.

1. *Let $\delta(\alpha) = (N(\alpha), d(\alpha))$ be an open-ended test of size $\alpha \in (0,1)$. Then*

$$p_n = \inf\{\alpha : \ N(\alpha) \leq n, \ d(\alpha) = 1\} \tag{2.2}$$

   defines an always valid p-value process.

2. *For any always valid p-value process $(p_n)_{n=1}^{\infty}$, an open-ended test $\tilde{\delta}(\alpha) = (\tilde{N}(\alpha), \tilde{d}(\alpha))$ of size $\alpha \in (0,1)$ is obtained from $(p_n)_{n=1}^{\infty}$ as follows:*

$$\tilde{N}(\alpha) = \inf\{n : p_n \leq \alpha\}, \ \tilde{d}(\alpha) = \mathbb{1}_{\{\tilde{N}(\alpha) < \infty\}} \tag{2.3}$$

3. *Let $\delta(\alpha) = (N(\alpha), d(\alpha))$ be any open-ended test where $N = \infty$ whenever $d = 0$. If $(p_n)_{n=1}^{\infty}$ is derived as in (2.2), then the construction (2.3) recovers the original open-ended test: $\tilde{\delta}(\alpha) = \delta(\alpha)$.*

Proof. Let N be a stopping time with respect to $\mathbb{F}$. Nestedness of open-ended tests implies that, for any $s \in [0,1]$, $\varepsilon > 0$:

$$\{p_N \leq s\} \subset \{N(s+\varepsilon) \leq N, \ d(s+\varepsilon) = 1\} \subset \{d(s+\varepsilon) = 1\}.$$

Therefore, $\mathbb{P}_0(p_N \leq s) \leq \mathbb{P}_0(d(s+\varepsilon) = 1) \leq s + \varepsilon$, and hence the result follows letting $\varepsilon \to 0$. Conversely, it is immediate from the definition that the tests are nested. For any $\varepsilon > 0$

$$\mathbb{P}_0(d(\alpha) = 1) = \mathbb{P}_0(N(\alpha) < \infty) \leq \mathbb{P}_0(p_{N(\alpha)} \leq \alpha + \varepsilon) \leq \alpha + \varepsilon$$

where the last inequality follows from the definition of always valid p-value process. The result follows letting $\varepsilon \to 0$. □

2.3 Power and run-time trade-off

In the last section we introduced always valid p-value processes to provide inference in a simple interface while controlling Type-I error under any data-dependent rule that the user might take. Moreover, we have seen that these processes naturally correspond to open-ended tests. Recall that these procedures provide power one, meaning that, for the hypothetical case of a user that is willing to wait forever, any true effect is assured to be detected.

However, in practice, no user is ever willing to wait forever, so power must be traded-off in favor of a shorter run-time. Therefore, the objective of this section is to pick an open-ended test that will allow any user to trade-off detection and run-time efficiently, regardless (and without prior knowledge) of their priorities.

We can naturally characterize user behaviour in the following way:

Definition 2.5 (**User**). A *user* is a pair (M, α) such that $M \in \mathbb{N}$ is the *maximum sample size* and $\alpha \in (0, 1)$ is the *significance level*.

Let $(p_n)_{n=1}^{\infty}$ be an always valid p-value process corresponding to an A/B test, and let $(M,\ \alpha)$ be an arbitrary user. As the experiment develops, the user will stop the first time $p_n \leq \alpha$ for some n, rejecting H_0 in favor of H_1, or upon reaching the M-th observation, accepting H_0 in that case. Hence, we can define the (M, α) *user's decision rule* $(N(M, \alpha),\ d(M, \alpha))$ as

$$N(M, \alpha) \coloneqq \min\{N(\alpha), M\},\ d(M, \alpha) \coloneqq \mathbb{1}_{\{N(\alpha) \leq M\}}$$

where $\delta(\alpha) = (N(\alpha), d(\alpha))$ is the corresponding open-ended test constructed using our always valid p-value process according to Theorem 2.4.

With user behaviour well defined, we shall now formalize efficiency in our context of power and run-time trade-off. Let $(N(M, \alpha),\ d(M, \alpha))$ an arbitrary (M, α) user decision rule. We will consider the two following functions:

Definition 2.6 (**Power profile**). The *power profile* associated to $(N(M, \alpha),\ d(M, \alpha))$ is the power function

$$\nu(\theta;\ M, \alpha) = \mathbb{P}_\theta(d = 1) \text{ for } \theta \neq 0.$$

Definition 2.7 (**Relative run-length profile**). The *relative run-length profile* associated to $(N(M, \alpha),\ d(M, \alpha))$ is the function

$$\rho(\theta;\ M, \alpha) = \mathbb{E}_\theta(N)/M \text{ for } \theta \neq 0.$$

Any (M, α) user wishes to pick a decision rule which maximizes its power profile and minimizes its relative run-length profile, with perfect efficiency implying $\rho(\theta;\ M, \alpha) = 0$ and $\nu(\theta;\ M, \alpha) = 1$ for $\theta \neq 0$. As discussed beforehand, perfect efficiency will never be attainable in practice, so we shall study an open-ended test that optimizes our criteria as best as possible.

We will focus on the family of open-ended tests given by the mSPRT. Indeed, given an arbitrary mixing distribution π, consider $(\delta^\pi(\alpha))_{\alpha \in (0,1)}$ the family of mSPRT indexed by α, with $\delta^\pi(\alpha) = (N^\pi(\alpha), d^\pi(\alpha))$. We know by Proposition 1.26 that the run-length of the mSPRT when $\theta \neq 0$ in the limit as $\alpha \to 0$ is given by

$$\begin{aligned} N^\pi(\alpha)/\log(1/\alpha) \xrightarrow{\mathbb{P}} I(\theta)^{-1} &= \{\theta\psi'(\theta) - (\psi(\theta) - \psi(0))\}^{-1} \\ &= \frac{4\sigma^2}{\theta^2}. \end{aligned} \tag{2.4}$$

Observe then that we can exploit this characterization to compare the maximum sample size M with the magnitude of the significance level α for any given user (M, α), in a context where α is expected to be small. Hence, we can then simplify our problem to optimize for three distinct types of users. Denote from now on $(N^{\pi}(M,\alpha),\ d^{\pi}(M,\alpha))$ the (M,α) user decision rule corresponding to the mSPRT with mixing distribution π.

2.3.1 "Aggressive" users

This type of user is characterized by its choice of a rather large significance level relative to their maximum sample size ($M \gg \log(1/\alpha)$), meaning that they are willing to wait longer despite not being too restrictive about their Type-I error control. It seems clear then that, in this case, the user receives almost the entirety of the power one provided by the mSPRT:

Proposition 2.8. *Let $\rho(\theta;\ M,\alpha)$ and $\nu(\theta;\ M,\alpha)$ be the relative run-length and power profiles, respectively, associated with $(N^{\pi}(M,\alpha),\ d^{\pi}(M,\alpha))$. If $\alpha \to 0$ and $M \to \infty$ such that $M/\log(1/\alpha) \to \infty$, then $\rho(\theta;\ M,\alpha) \to 0$ and $\nu(\theta;\ M,\alpha) \to 1$ for $\theta \neq 0$.*

Proof. Follows immediately from (2.4). □

2.3.2 "Conservative" users

In contrast to the last case, "Conservative" users are those which choose a small significance level relative to their maximum sample size ($M \ll \log(1/\alpha)$), and hence choose to be really demanding about Type-I error control but are not willing to wait long enough to payoff this restriction. Naturally, in this case any feasible user decision rule $(N(M,\alpha),\ d(M,\alpha))$ performs as well as the mSPRT:

Proposition 2.9. *Let $(N(M,\alpha),\ d(M,\alpha))$ be any feasible user decision rule, and let $\nu(\theta;\ M,\alpha)$ be its corresponding power profile. Given that $\alpha \to 0$, $M \to \infty$ such that $M/\log(1/\alpha) \to 0$, we have $\nu(\theta;\ M,\alpha) \to 0$ for $\theta \neq 0$.*

Proof. Fix $\theta \neq 0$, and assume for contradiction that there is some $\beta < 1$ such that there exists some feasible user decision rule $(N^*(M,\alpha),\ d^*(M,\alpha))$ such that

$$\mathbb{P}_{\theta}\left(d^*(M,\alpha) = 0\right) \leq \beta \text{ as } \alpha \to 0, M \to \infty.$$

Provided that $\mathbb{P}_0\left(d^*(M,\alpha) = 1\right) \leq \alpha$, using a lower bound from Hoeffding [5] implies that there exists some κ such that

$$\mathbb{E}_{\theta}(N^*) \geq \kappa \log(1/\alpha)(1 + o(1)).$$

Given that $M/\log(1/\alpha) \to 0$ in the limit, this lower bound implies that N^* exceeds the maximum sample size with positive probability. □

2.3.3 "Goldilocks" users

The non-trivial and limiting user case is that which finds and equilibrium between the maximum run-length and the significance level ($M \sim \log(1/\alpha)$). To study this case, given any family of open-ended tests $(\delta(\alpha))_{\alpha\in(0,1)}$, we shall define a measure of worst-case efficiency over $\theta \neq 0$ for an arbitrary (M, α) user.

Definition 2.10 (**Relative efficiency**)**.** Let $\delta(\alpha) = (N(\alpha), d(\alpha))$ be an open-ended test of size α. Let $\rho(\theta;\ M, \alpha)$ and $\nu(\theta;\ M, \alpha)$ be the relative run-length and power profiles associated with the user decision rule $(N(M, \alpha),\ d(M, \alpha))$ corresponding to an arbitrary user (M, α). The *relative efficiency* of the test at (M, α) is

$$\varphi(M, \alpha) = \inf_{\delta^*\in\Delta(M,\alpha)} \inf_{\theta\neq 0} \frac{\rho(\theta)}{\rho(\theta;\ M, \alpha)}$$

where $\Delta(M, \alpha) = \{\delta^*(\alpha) :\ N^* \leq M,\ \mathbb{P}_0(d^* = 1) \leq \alpha,\ \forall\ \theta \neq 0\ \nu(\theta) \geq \nu(\theta;\ M, \alpha)\}$.

Observe that this definition of efficiency makes a comparison (as a function of every user (M, α)) of the worst-case scenario of the relative run-length of an open-ended test from a particular family with that of all the open-ended tests that provide at least the same power for all $\theta \neq 0$.

Under this definition, our main result then shows that when $M \sim \log(1/\alpha)$, the relative efficiency of the mSPRT with any mixing distribution approaches one in the limit.

Theorem 2.11. *Let $\varphi(M, \alpha)$ be the relative efficiency of the mSPRT with mixing distribution π, $\delta^\pi = (N^\pi(\alpha), d^\pi(\alpha))$. If $\alpha \to 0$, $M \to \infty$ such that $M = O(\log(1/\alpha))$, we have $\varphi(M, \alpha) \to 1$.*

Proof. To establish asymptotic efficiency, given (M, α), it is sufficient to find some θ_1, where for every feasible test $\delta^* = (N^*, d^*)$ with $\nu^*(\theta_1) \geq \nu(\theta_1;\ M, \alpha)$, we have that $\rho^*(\theta_1) \geq \rho(\theta_1;\ M, \alpha)(1 + o(1))$.

By [9], a normal approximation holds asymptotically for $\mathbb{P}_\theta(d(M, \alpha) = 0)$; in particular, if we fix any θ,

$$\mathbb{P}_\theta\left(d(M, \alpha) = 0\right) = \Phi\left(\log(1/\alpha)^{1/2} B(M, \alpha, \theta)\right)(1 + o(1))$$

with $B(M,\alpha,\theta) = \left(\frac{2\sigma^2 I(\theta)^3}{\theta^2}\right)^{1/2}\left(\frac{M}{\log(1/\alpha)} - I(\theta)^{-1}\right)$, and Φ denoting the standard normal distribution cdf. On the other hand, standard results on the log partition function ψ imply that for fixed (M,α),

$$\log(1/\alpha)^{1/2} B(M,\alpha,\theta) \sim \eta_2 \log(1/2)^{1/2}\left(\frac{M\theta^2}{\log(1/\alpha)} - \eta_3\right) \text{ as } \theta \to 0$$

with $\eta_2, \eta_3 > 0$. Combining the two results, for

$$\theta_1 = \sqrt{\frac{\log(1/\alpha)}{M}\left(\sqrt{\frac{2}{\eta_2}} + \eta_3\right)}$$

we can see that eventually

$$\mathbb{P}_{\theta_1}(\delta(M,\alpha) = 0) \leq \Phi\left(\sqrt{2\log(1/\alpha)}\right) =: \beta_1,$$

in other words, that the mSPRT has power at least $1-\beta_1$ at θ_1 in the limit.

Suppose now that $\delta^* = (N^*, d^*)$ is another test that achieves $1-\beta_1$ power at θ_1. If α is sufficiently small such that $0 < \alpha + \beta < 1$, making use of a lower bound of the ESS described in [5], we can show that for any $\theta \in (0, \theta_1)$,

$$\begin{aligned}\mathbb{E}_\theta(N^*) &\geq \frac{|\log(\alpha+\beta_1)| - \frac{1}{4\sigma^2}\theta_1^2 |\log(\alpha+\beta_1)|^{1/2}}{\max\{I(\theta),\ I(\theta,\theta_1)\}} \\ &= I(\theta)^{-1}\log(1/\alpha)(1+o(1)).\end{aligned} \tag{2.5}$$

By continuity, the result holds at θ_1 aswell. Comparing expression (2.5) with (2.4) gives the desired inequality on the relative run-times at θ_1. □

2.4 Empirics and comparison to fixed-horizon testing

In this last section we seek to simulate the mSPRT with arbitrary parameters in order to verify some of the properties described during this work. Moreover, we expose an empirical comparison of this procedure with its fixed-horizon alternative in A/B testing, following the methodology described in Johari et al. [6, Section 5.6] and Stenberg [18, Section 3.1].

The simulations are performed in the framework of The R Programming Language [12] for the purpose of reproducible research, and using the package **mixtureSPRT**

by Stenberg [17]. This package provides an implementation of the mSPRT as a function of arbitrary control and treatment group samples, as well as parameters like the significance level or the mixing distribution variance. The backend of the function is implemented in C++ for a significant speed-up in computation, specially towards averaging values across multiple simulations. Moreover, our code is adapted to the **Future** package by Bengtsson [1] to allow concurrency in R. The code can be found attached to the thesis in the repository.

As described in Section 2.1, the control and treatment data consists of two independent sequences of *i.i.d.* observations $(X_n)_{n\in I}$ and $(Y_n)_{n\in I}$, with $X_n \sim \mathcal{N}(\mu_0,\ \sigma^2)$ and $Y_n \sim \mathcal{N}(\mu_1,\ \sigma^2)$, leading to the difference $Z_n := Y_n - X_n$, with $Z_n \sim \mathcal{N}(\theta, 2\sigma^2)$ (θ unknown and σ^2 known). For our simulations, we simplify to $\mu_0 = 0$ and $\mu_1 = \theta$, and let the population variance be fixed to $\sigma^2 = 1/2$. To account for variability, we propose a normal prior distribution over the effect size $\theta \sim \mathcal{N}(\mu_\theta,\ \sigma_\theta^2)$, which in principle is different from the mixing distribution. This way, and as Figure 2.1 represents, sampling a single instance of our experiment data requires to sample first θ from the normal prior.

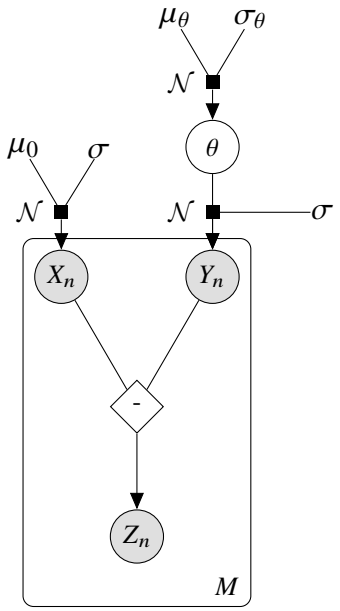

Figure 2.1: Plate notation representation of the data-generating process

Formally, we define a single simulation to require only the following parameters:

- (M, α), a user with maximum sample size M and significance level α
- μ_θ, the mean of the normal prior over the effect size θ
- σ_θ, the standard deviation of the normal prior over the effect size θ

Providing these parameters, the script first generates a sequence of M *i.i.d.* observations $(Z_n)_{n=1}^M$ which will be used by the mSPRT. In this case, we choose our mixing distribution π to be $\mathcal{N}(0, \tau^2)$, since in that case, the mLR has a closed analytical form ([18, Equation 10]):

$$\bar{\Lambda}_n^\pi = \sqrt{\frac{2\sigma^2}{2\sigma^2 + n\tau^2}} \exp\left\{\frac{\tau^2 n^2 (\bar{Z}_n)^2}{4\sigma^2(2\sigma^2 + n\tau^2)}\right\} \tag{2.6}$$

where $\bar{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i$. The mixing distribution variance τ^2 is not a parameter of the simulation since it is chosen to be optimal following [6, Theorem 3].

Our objective with these simulation parameters is to provide a reference for the power profile $\nu(\theta;\ M, \alpha)$ and relative run-length profile $\rho(\theta;\ M, \alpha)$ of an arbitrary user (M, α), while also being able to control the distribution of the true effect size θ. We choose to estimate their mean, or expected value over the values of $\theta \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2)$, which can be approximated via *Monte Carlo simulation*: we perform B independent simulations with the same parameters $(M, \alpha, \mu_\theta, \sigma_\theta)$ and then estimate these quantities by their sample average:

$$\hat{\nu}(M, \alpha) \coloneqq \mathbb{E}_{\theta\sim\mathcal{N}(\mu_\theta,\sigma_\theta^2)}(\nu(\theta;\ M, \alpha)) \approx \frac{1}{B}\sum_{n=1}^{B} \mathbb{1}_{\{N\le M\}}$$

$$\hat{\rho}(M, \alpha) \coloneqq \mathbb{E}_{\theta\sim\mathcal{N}(\mu_\theta,\sigma_\theta^2)}(\rho(\theta;\ M, \alpha)) \approx \frac{1}{M}\frac{1}{B}\sum_{n=1}^{B} \min\{N, M\}$$

By the Law of Large Numbers, a larger value of B will yield more confident point estimates. Hence, for our computations we fix $B = 10000$. Typical, real-word scenarios give values of $\alpha \in \{0.1, 0.05, 0.01\}$ for the significance level, and expect small effect effect sizes of the 5%, 1% and 0.5%; we focus on $\mu_\theta = 0.05$. Moreover, we experiment with values for the prior variance $\sigma_\theta^2 \in \{0.1, 0.01, 0.001\}$.

We first study the average power profile $\hat{\nu}(M, \alpha)$ and the average run-length profile $\hat{\rho}(M, \alpha)$ as a function of M, for different fixed values of the parameters α and σ_θ:

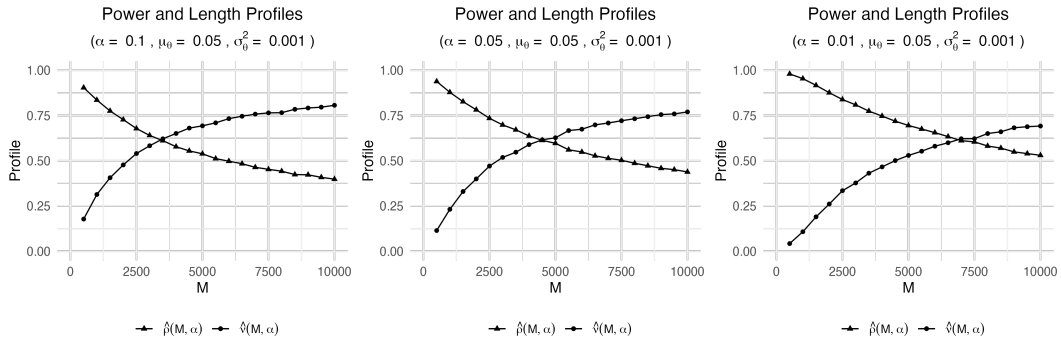

Figure 2.2: $\hat{\rho}(M, \alpha), \hat{\nu}(M, \alpha)$ as a function of M ($\alpha = 0.1, 0.05, 0.01$)

We can clearly observe that power increases as the user's willingness to wait longer increases, as it leaves room for the mLR to make more rejections, and hence, the run-length decreases as a consequence. Moreover, we observe that the steepness of the power function (and hence, of the run-length profile) is inversely proportional to the magnitude of α, as expected by the nestedness property of open-ended tests.

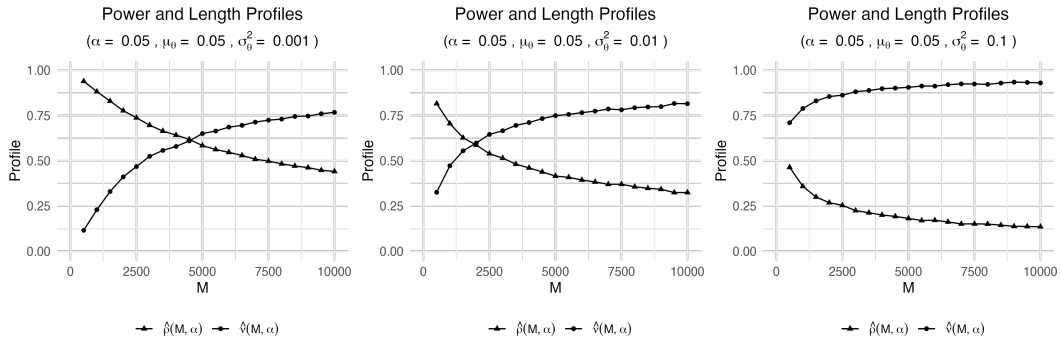

Figure 2.3: $\hat{\rho}(M, \alpha), \hat{\nu}(M, \alpha)$ as a function of M ($\sigma_\theta^2 = 0.001, 0.01, 0.1$)

When it comes to the prior variance σ_θ^2, we can observe that a higher variability around μ_θ increases significantly the test's detection and hence, decreases the overall run-length.

We now present the real-world scenario where sequential methods shine over their fixed-horizon counterparts. As described in Section 2.2, fixed-horizon testing forces us to fix a sample size prior to the execution of our experiment. Standard methods provide calculators for this sample size as a function of the desired test power and a MDE, among other parameters. This MDE, which represents a prior estimate of

the magnitude of the true effect size, remains crucial in the actual run-length of the experiment, since a miss-specification of its value can turn out in a insufficient sample size in the higher setting, or else in a waste of time (or even cost) in the other case. Our sequential method then oversees the specification of this parameter, and moreover, adapts our run-length to the true effect size.

For the following simulations, we focus in the casuistry where our MDE estimate for our fixed-horizon test is smaller than the actual true effect: in particular, we fix this MDE value to be 50%, 80% and 100% of the true effect. For each case, we perform B simulations with specific values of our parameters $(M, \alpha, \mu_\theta, \sigma_\theta)$ that we know yield a particular average power profile value $\hat{\nu}(M, \alpha)$. Moreover, given the values for $(\text{MDE}, \sigma, \alpha, \hat{\nu}(M, \alpha))$ we can compute the fixed-horizon sample size n^* such that the one-sample, UMP t-test of size α achieves $\hat{\nu}(M, \alpha)$ power when testing $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$, given that we estimate $\theta = \text{MDE}$. To do so, we use the sample size calculator included in the **stats** standard package of R. We then study the distribution of $\hat{\rho}(M, \alpha)/\hat{\rho}_f(M, \alpha)$, where $\hat{\rho}_f(M, \alpha) := n^*/M$.

The following plots correspond to an approximate power profile of $\hat{\nu} \approx 0.7536$, achieved by the parameters $(M, \alpha, \mu_\theta, \sigma_\theta) = (5500, 0.05, 0.05, 0.1)$:

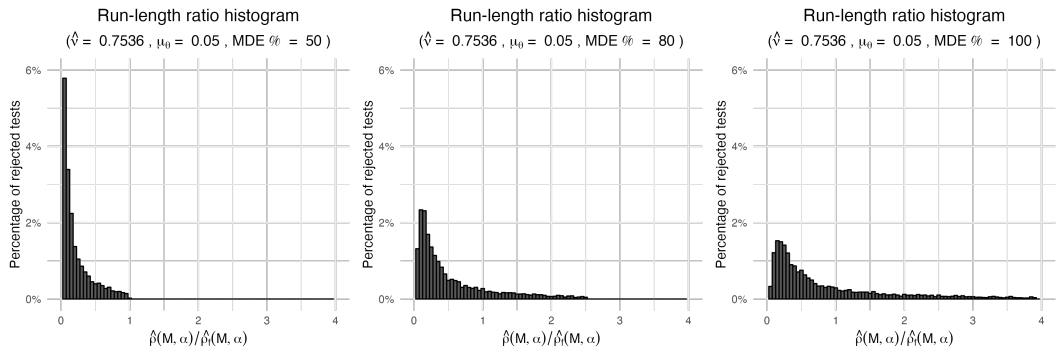

Figure 2.4: $\hat{\rho}(M, \alpha)/\hat{\rho}_f(M, \alpha)$ distribution (MDE % = 50, 80, 100)

It easy to see that the mSPRT can save, in real-time, any MDE underestimation that we may have committed in a fixed-horizon setting: when our MDE is 50% of the true effect size, all sequential tests that provide detection take less observations than fixed-horizon. This efficiency decreases as we set the MDE closer to the actual true effect size; however, in all three cases, the run-length ratio is less than 1 in at least half of all rejected tests.

Chapter 3

Conclusions and further research

To conclude this body of work I would like to look back on everything achieved before and during its development; and I would also like to give a glimpse into what this thesis means for my future personal research.

Back in 2022 I already had the idea of applying hypothesis testing in what could be the calibration process of a sensor, focusing on the aspect that observations are gathered in a sequential fashion; however, I had no clue on the existence of sequential analysis as a field. Then, on June 2023, I stumbled upon Philip B. Stark's excellent "Notes on Applied Statistics" repository ([16]), where I was first introduced to the SPRT. Further research led me to a Spotify Engineering blog-post discussing the importance of the application of sequential testing to online controlled experimentation ([13]), where I discovered the Always Valid Inference paper ([6]) that makes up the second chapter of this work.

Writing this thesis has been a nourishing yet challenging experience. Indeed, being able to research something of my personal interest like sequential hypothesis testing to make up my Bachelor's thesis has been an honour to me. This work has displayed the potential of the sequential paradigm for testing hypotheses, which is rather unknown to most people educated on statistics. For that same reason, research on this field has been quite difficult. The initial work by Wald dates back to the 1940s, and the more modern literature concerning open-ended tests was only developed in the 1970s. There hasn't been any major wide-spread of this field nor well-known literature that summarizes the most important advances in this theory. In addition to that, some of the proofs and concepts developed rely on advanced topics in the theory of stochastic processes, like renewal theory or optimal stopping theory, which I hadn't been introduced during my degree. Nonetheless, sequential testing is now considered state-of-the-art in the online experimentation industry, with most experimentation engines adopting this framework in their product ([13], [19], [20]),

specially the mSPRT via Always Valid Inference. Therefore, the difficulties found during the writing of this work are paid-off by seeing how relevant these techniques are.

The development of this project has already provided me a solid foundation on the theory of sequential analysis. Nonetheless, I look forward to learn more advanced and modern topics, like the extension to continuous-time stochastic processes or a fully Bayesian treatment; but most importantly, research other applications. Indeed, I remain interested in the employment of sequential hypothesis testing in online change-point detection, which has found its own subset of literature in statistical quality control. This already has its successful applications in industrial settings; however, I look forward to investigate other potential use-cases in data science. With the advent of streaming data, I anticipate the sequential analysis literature to flourish in the following decades, driven by the necessity to perform efficient decision-making while maintaining statistical rigour.

Bibliography

[1] H. Bengtsson. "A Unifying Framework for Parallel and Distributed Processing in R using Futures". In: *The R Journal* 13.2 (2021), pp. 208–227.

[2] J. M. Bernardo. "Una introducció a l'estadística bayesiana". In: *Butlletí de la Societat Catalana de Matemàtiques* 17.1 (2002), pp. 7–64.

[3] T. S. Ferguson. *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, July 10, 2014.

[4] B. K. Ghosh and P. K. Sen, eds. *Handbook of sequential analysis*. Statistics 118. New York: Dekker, 1991.

[5] W. Hoeffding. "Lower Bounds for the Expected Sample Size and the Average Risk of a Sequential Procedure". In: *The Annals of Mathematical Statistics* 31.2 (June 1960). Publisher: Institute of Mathematical Statistics, pp. 352–368.

[6] R. Johari, L. Pekelis, and D. J. Walsh. *Always Valid Inference: Bringing Sequential Analysis to A/B Testing*. July 16, 2019. arXiv: `1512.04922[math, stat]`.

[7] R. Johari et al. "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. New York, NY, USA: Association for Computing Machinery, Aug. 13, 2017, pp. 1517–1525.

[8] R. Kohavi, D. Tang, and Y. Xu. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press, Feb. 14, 2020.

[9] T. L. Lai and D. Siegmund. "A Nonlinear Renewal Theory with Applications to Sequential Analysis I". In: *The Annals of Statistics* 5.5 (Sept. 1977). Publisher: Institute of Mathematical Statistics, pp. 946–954.

[10] M. Pollak and D. Siegmund. "Approximations to the Expected Sample Size of Certain Sequential Tests". In: *The Annals of Statistics* 3.6 (Nov. 1975). Publisher: Institute of Mathematical Statistics, pp. 1267–1282.

[11] M. Pollak. "Optimality and Almost Optimality of Mixture Stopping Rules". In: *The Annals of Statistics* 6.4 (1978). Publisher: Institute of Mathematical Statistics, pp. 910–916.

[12] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2022. url: https://www.R-project.org/.

[13] M. Schultzberg and S. Ankargren. *Choosing a Sequential Testing Framework - Comparisons and Discussions*. Mar. 21, 2023. url: https://engineering.atspotify.com/2023/03/choosing-sequential-testing-framework-comparisons-and-discussions/ (visited on 01/14/2024).

[14] D. Siegmund. *Sequential Analysis*. Springer Series in Statistics. New York, NY: Springer New York, 1985.

[15] T. Singer. *Coming Soon: Confidence - An Experimentation Platform from Spotify*. Spotify Engineering. Aug. 3, 2023. url: https://engineering.atspotify.com/2023/08/coming-soon-confidence-an-experimentation-platform-from-spotify/ (visited on 12/23/2023).

[16] P. B. Stark. *Notes on Applied Statistics*. Dec. 4, 2023. url: https://github.com/pbstark/StatNotes (visited on 01/14/2024).

[17] E. Stenberg. *mixtureSPRT: Mixture Sequential Probability Ratio Test*. R package version 1.0. 2019.

[18] E. Stenberg. "SEQUENTIAL A/B TESTING USING PRE-EXPERIMENT DATA". MA thesis. Uppsala University, Department of Statistics, 2019.

[19] M. Stewart. *Sequential Testing on Statsig*. Oct. 18, 2023. url: https://www.statsig.com/blog/sequential-testing-on-statsig (visited on 01/13/2024).

[20] *Under the Hood of Uber's Experimentation Platform*. Uber Blog. Aug. 28, 2018. url: https://www.uber.com/en-ES/blog/xp/ (visited on 12/23/2023).

[21] A. Wald. "Sequential Tests of Statistical Hypotheses". In: *The Annals of Mathematical Statistics* 16.2 (June 1945), pp. 117–186.

[22] A. Wald and J. Wolfowitz. "Optimum Character of the Sequential Probability Ratio Test". In: *The Annals of Mathematical Statistics* 19.3 (Sept. 1948). Publisher: Institute of Mathematical Statistics, pp. 326–339.