University of Piraeus
School of Finance and Statistics
Department of Statistics and Insurance Science

Postgraduate Program in Applied Statistics

«Sequential Analysis Models and Applications»

Alexandros L. Stefanidis

MSc Dissertation
submitted to the Department of Statistics and Insurance Science
of the University of Piraeus in partial fulfilment of the requirements
for the degree of Master of Science in Applied Statistics

Piraeus
September 2015

This thesis was approved unanimously by the Three-Member Commission of Inquiry appointed by the Special General Assembly of the department of Statistics and Insurance Science of the University of Piraeus at no. …….. meeting according to the Bylaws of the Postgraduate Program in Applied Statistics

The Committee members were:
- Assistant Professor M. Boutsikas (Supervisor)
- Associate Professor D. Antzoulakos
- Professor C. Iliopoulos

The approval of the thesis by the Department of Statistics and Insurance Science of the University of Piraeus does not imply acceptance of the author’s opinion.

Abstract

The purpose of this thesis is to review the basic methods of sequential analysis. In particular, we present the measure-theoretic approach for proving Wald's equations and the fundamental equation, via the martingale theory and the theory of stopping times. We study the exact and asymptotic properties of the techniques for constructing fixed width confidence intervals for the mean of the normal distribution and their extension to p-dimensional confidence regions. A complete description of the sequential probability ratio test is given along with Wald's approximation for estimating the operating characteristic function and the average sample number. Finally, applications of the sequential methodology are presented in order to define the analytical and geometric aspect of the cumulative sum statistical algorithm and the approximation of the average run length. We used the R software package for getting all the numerical results.

Keywords

Martingales, stopping times, sequential procedure, SPRT, CUSUM, average run length

Contents

List of Tables

List of Figures

List of Abbreviations

r.v.	random variable(s)
i.i.d.	independent identically distributed
p.d.f.	probability density function
c.d.f.	cumulative distribution function
m.g.f.	moment generating function
d.f.	distribution function
s.t.	stopping time
c.i.	confidence interval
a.s.	almost surely
s.p.	stochastic process
o.s.t.	Optimal Stopping theory
K-L	Kullback-Leibler
AR(1)	Fist order autroregreesive
OC	Operating Characteristic
ASN	Average Sample Number
SPRT	Sequential Probability Ratio Test
CUSUM	Cumulative Sum
ADD	Average Detection Delay
SADD	Supremum Average Detection Delay
ESADD	Essential Supremum Average Detection Delay
PFA	Probability of False Alarm

Introduction

Sequential analysis is the branch of statistics that deals with problems where the sample size is not fixed but changes in each experiment. According to Ghosh (1991) the first ideas of sequential analysis can be found in the celebrated work of Huyghens, Bernoulli and De Moivre which referred to the gambler's ruin. Initial applications of the sequential methods emerged in quality control and particularly in double sampling plans (Dodge & Romig, 1929) and the Shewhart control charts (1931). These were the most significant areas in which the sample size was not fixed in advance but was actually a random variable. The purpose of sequential analysis is to do statistical inference while using the smallest possible sample and the properties of the random variable that represents it.

Wald (1947) was the first to construct a sequential method for hypotheses testing, equivalent to the uniformly most powerful Neyman-Pearson test (1933). Wald defined to accept the null hypothesis $\mathcal{H}_0$ when the likelihood ratio becomes smaller than a constant e.g. A, to reject $\mathcal{H}_0$ when the ratio becomes greater than a constant B and to continue sampling otherwise. The attempt to compare this test with the classical Neyman-Pearson one, led Wald and Wolfowitz (1948) to the discovery of optimal stopping theory (*o.s.t.*). O.s.t. investigates the form of the random variables which optimize certain conditions, under the sequential testing framework. Significant applications of o.s.t. can be found in game theory (Ferguson, 2007), financial mathematics (Shiryaev, 2013) and quickest change detection (Veeravalli, 2012).

Non-fixed sample size problems can be found in the area of sequential estimation (Mukhopadhyay, 2009), i.e. the construction of confidence interval with a priori fixed width, the problem of finding a mean estimator with a priori bounded risk and the theory of stochastic approximation (Ghosh, 1997). Also, Wald's revolutionary contribution to the theory of stochastic processes gave birth to equations that include random sums of random variables. We will see later on that Wald's equations can be used to estimate the average run length of the cumulative sum statistical algorithm (Page, 1954). Finally, several applications of sequential analysis can be found in clinical trials, signal processing and multiarmed bandit problems (Lai, 2001).

CHAPTER 1

Martingales and Stopping Times

1.1 Basic notions

In order to prove Wald's equations we will borrow some basic results from the theory of stochastic processes and martingale theory. This chapter is the probabilistic part of sequential statistics and includes the machinery for estimating the time of a process, as we will see in the following chapters.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $X: \Omega \to \mathbb{R}$ a r.v. and $(\mathbb{R}, \mathcal{B}, \mu_X)$ the induced space, where $\mathcal{B}$ is the Borel σ-algebra of $\mathbb{R}$.

Definition 1.1.1. *Let $\mathcal{C}$ be a class of subsets of Ω. The σ-algebra generated by $\mathcal{C}$ is the smallest σ-algebra which includes all the elements of $\mathcal{C}$. We denote $\sigma(\mathcal{C})$.*

Definition 1.1.2. *We call σ-algebra generated by the r.v. X the set $\sigma(X) = \{X^{-1}(B): B \in \mathcal{B}\}$. Similarly we expand the notion of a generated σ-algebra by the r.v. $X_1, X_2, \dots, X_n$ according to the relation $\sigma(X_1, X_2, \dots, X_n) := \sigma(\bigcup_{i=1}^{n} \sigma(X_i))$.*

An instant consequence is that X is $\mathcal{F}$-measurable if-f $\sigma(X) \subset \mathcal{F}$. We remind that generally $\mathbb{E}(X \cdot I_A) = \int_A X \, d\mathbb{P}, \forall A \in \mathcal{F}$, but $\mathbb{E}\big(X \cdot I_{[X \in B]}\big) = \int_B x \, f_X(x) dx$ when the r.v. X is continuous and $\sum_{x \in B} x \, \mathbb{P}(X = x)$ when X is discrete. We will examine the theorems related to continuous r.v. The same results can be proven for the case of discrete r.v. and can be found in various notes (e.g. Walsh, 2014). The following is an extension of the conditional expectation definition.

Definition 1.1.3. *Let $\mathcal{G}$ be a σ-algebra such as $\mathcal{G} \subset \mathcal{F}$. If $\mathbb{E}|X| < \infty$ then we define as expectation of X given $\mathcal{G}$, a r.v. Z such as:*

(I1) *Z is $\mathcal{G}$-measurable*

(I2) $\mathbb{E}(Z \cdot I_A) = \mathbb{E}(X \cdot I_A)$, $\forall A \in \mathcal{G}$

We denote $Z = \mathbb{E}(X|\mathcal{G})$. We can prove with the use of the Radon-Nikodym derivative, that Z can be defined when X is a r.v. that satisfies $\mathbb{E}(|X|) < \infty$. Also, if Z' is another r.v. that satisfies (I1) and (I2) then $Z' = Z$ a.s.

We will not use the notation *a.s.* for the equality of conditional expectations. We now present some of the properties that will be used in the following proofs.

Definition 1.1.4. *Two σ-algebras $\mathcal{G}$ and $\mathcal{G}'$ will be called independent if* $\forall A \in \mathcal{G}$ *and* $\forall B \in \mathcal{G}'$*:* $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

A r.v. X, is independent of $\mathcal{G}'$ if the σ-algebras $\mathcal{G} = \sigma(X)$ and $\mathcal{G}'$ are independent.

Theorem 1.1.1. *If X and Y are r.v. such as* $\mathbb{E}|X|, \mathbb{E}|Y| < \infty$*, then:*

(i) $\mathbb{E}(\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(X)$

(ii) If X is $\mathcal{G}$-measurable, then $\mathbb{E}(X|\mathcal{G}) = X$

(iii) $\mathbb{E}(aX + bY|\mathcal{G}) = a\mathbb{E}(X|\mathcal{G}) + b\mathbb{E}(Y|\mathcal{G})$ *,* $\forall a, b \in \mathbb{R}$

(iv) If the r.v.. X is independent of $\mathcal{G}$ then $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X)$

(v) If X is $\mathcal{G}$-measurable then $\mathbb{E}(XY|\mathcal{G}) = X\mathbb{E}(Y|\mathcal{G})$

Proof. (i) It follows from property (I2) of the definition 1.1.3 setting $A = \Omega$.

(ii) Since X is measurable, we apply property (I2) for $Z = X$.

(iii). Let $ab \neq 0$. We will show initially that $aX + bY$ is measurable. Indeed, since the Borel σ-algebra is generated by intervals of the form $(-\infty, x]$, where $x \in \mathbb{R}$, it suffices to show that $\{aX + bY \leq x\} \in \mathcal{G}$. We have that

$$aX + bY < x \Rightarrow aX < x - bY$$

And because of the density of the rational numbers in $\mathbb{R}$, $\exists\ q \in \mathbb{Q}$ such that

$$aX < q < x - bY \Rightarrow \{aX + bY < x\} = \bigcup_{q \in \mathbb{Q}} \left\{X < \frac{q}{a}\right\} \cap \left\{Y > \frac{x - q}{b}\right\} \in \mathcal{G}$$

since X and Y are $\mathcal{G}$-measurable r.v. Using the additivity property of the integrals, the required property is valid. For the rest of the values of a and b, working in a similar manner one can prove that X and Y are $\mathcal{G}$-measureable r.v.

(iv) The r.v. $g\colon \Omega \to \mathbb{R}$ such as $g(\omega) = \mathbb{E}(X)$ is obviously $\mathcal{G}$-measurable. Let $A \in \mathcal{G}$. Using independency we find that

$$\mathbb{E}(X \cdot I_A) = \mathbb{E}(X)\mathbb{E}(I_A) = \mathbb{E}(\mathbb{E}(X) \cdot I_A))$$

according to (iii).

(v) The proof can be found in Walsh (2004).

We can see the σ-algebra as a set of elements which includes all the «information» for the r.v. X. Thus, during an experiment and taking stepwise the random sample $X_1, X_2, \ldots$, $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ shows the information we gathered from $X_1, X_2, \ldots, X_n$, and includes all the information of the past r.v., so that $\mathcal{F}_{n-1} \subset \mathcal{F}_n$, $\forall n \in \mathbb{N}$. Practically, we can find this situation in stochastic games. The r.v. X_n shows a player's fortune at the n-th step. Then the event of winning or being excluded from the game at that particular step depends exclusively upon the r.v. $X_1, X_2, \ldots, X_n$ which will be known at that time.

Definition 1.1.5. *A sequence of σ-algebras $(\mathcal{F}_n)_{n\geq 1}$ is called filtration if $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}$.*

Definition 1.1.6. *The stochastic process (s.p.) $(X_n)_{n\geq 1}$ will be called adapted to the filtration $(\mathcal{F}_n)_{n\geq 1}$ if the r.v. X_n are $\mathcal{F}_n$-measurables* $\forall n \in \mathbb{N}$.

It can be easily seen that every s.p. $(X_n)_{n\geq 1}$ is adapted to the filtration $(\mathcal{F}_n)_{n\geq 1}$, where $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. This filtration is called natural filtration of $(X_n)_{n\geq 1}$.

Definition 1.1.7. *The s.p. $(X_n)_{n\geq 1}$ with $\mathbb{E}|X_n| < \infty$ will be called martingale with respect to the filtration $(\mathcal{F}_n)_{n\geq 1}$ if it is adapted and $(X_{n+1}|\mathcal{F}_n) = X_n$, $\forall n \in \mathbb{N}$. We denote $(X_n, \mathcal{F}_n)_{n\geq 1}$.*

Example 1.1.1. If $(X_n, \mathcal{F}_n)_{n\geq 1}$ is a martingale, then all the r.v.. X_n have the same expectation since

$$\mathbb{E}(X_{n+1}) = \mathbb{E}(\mathbb{E}(X_{n+1}|\mathcal{F}_n)) = \mathbb{E}(X_n)$$

and thus recursively,

$$\mathbb{E}(X_n) = \mathbb{E}(X_1)\,, \forall n \in \mathbb{N} \quad \square$$

1.2 Stopping times

In statistical analysis and in the area of martingales, it is vital to define the termination of sequential sampling. The termination depends upon the decision rule, but also the variable that shows the step at which we will stop.

Definition 1.2.1. *Let $(\mathcal{F}_n)_{n\geq 1}$ be a filtration. A r.v. $T: \Omega \to \mathbb{N} \cup \{+\infty\}$ will be called stopping time (s.t.) with respect to the filtration $(\mathcal{F}_n)_{n\geq 1}$, if* $\{T = n\} \in \mathcal{F}_n$ $\forall n \in \mathbb{N}$.

Example 1.2.1. The r.v.

$$T = \inf\{n \geq 1: S_n = a \;\; or \;\; S_n = b\}\,,\ \inf\emptyset = \infty$$

where $S_n = \sum_{i=1}^{n} X_i$, is a stopping time (s.t.) with respect to the natural filtration $(\mathcal{F}_n)_{n\geq 1}$ of the r.v. $X_1, X_2, \ldots, X_n$, since

$$\{T = n\} = \{S_1 \notin \{a,b\}\} \cap \ldots \cap \{S_{n-1} \notin \{a,b\}\} \cap \{S_n \in \{a,b\}\} \in \mathcal{F}_n. \quad \square$$

Example 1.2.2. We denote $x \wedge y = \min\{x, y\}$. If T_1 and T_2 are s.t. with respect to the filtration $(\mathcal{F}_n)_{n\geq 1}$, then the r.v. $T_1 \wedge T_2$ is a s.t. Indeed

$$\{T_1 \wedge T_2 = n\} = (\{T_1 = n\} \cap \{T_2 \geq n\}) \cup (\{T_1 \geq n\} \cap \{T_2 = n\}) \in \mathcal{F}_n$$

since

$$\{T_i \geq n\} = \Omega \setminus \{\cup_{j=1}^{n-1} \{T_i = j\}\} \in \mathcal{F}_n$$

We will show that if the sequence $(X_n, \mathcal{F}_n)_{n\geq 1}$ is a martingale with respect to the natural filtration, then $(X_{T\wedge n}, \mathcal{F}_{T\wedge n})_{n\geq 1}$ is also a martingale, where $\mathbb{P}(T < \infty) = 1$ (X_T must be well defined). The r.v. $X_{T\wedge n}$ is $\mathcal{F}_n$-measurable if it is $\mathcal{F}_{T\wedge n}$-measurable and $T \wedge n \leq n$. We can see practically $X_{T\wedge n}$ for $T = 4$ as $X_1, X_2, X_3, X_4, X_4, X_4, \ldots$. Thus

$$|X_{T\wedge n}| \leq \max_{1\leq i\leq n}\{|X_i|\} \leq \sum_{i=1}^{n} |X_i| \Rightarrow \mathbb{E}|X_{T\wedge n}| < \infty.$$

We observe that

$$X_{T\wedge(n+1)} = X_{T\wedge n} + (X_{n+1} - X_n) I_{\{T>n\}}$$

and therefore

$$\begin{aligned}
\mathbb{E}(X_{T\wedge(n+1)} | \mathcal{F}_n) &= X_{T\wedge n} + \mathbb{E}\left((X_{n+1} - X_n) I_{\{T>n\}} | \mathcal{F}_n\right) \\
&= X_{T\wedge n} + I_{\{T>n\}}\, \mathbb{E}((X_{n+1} - X_n) | \mathcal{F}_n) \\
&= X_{T\wedge n} + I_{\{T>n\}} (\mathbb{E}(X_{n+1} | \mathcal{F}_n) - X_n) \\
&= X_{T\wedge n} + I_{\{T>n\}} \cdot 0 \\
&= X_{T\wedge n}
\end{aligned}$$

For the first and second equality we used the fact that $I_{\{T>n\}}$ is $\mathcal{F}_n$-measurable, the properties (ii) and (v) of the theorem 1.1.1 and that $(X_n, \mathcal{F}_n)_{n\geq 1}$ is a martingale. □

There are many types of s.t. in statistics. Two of the most commonly used are the *hitting stopping times* (see Example 1.2.1) and the *exit times* (see Paragraph 3.2). Initially we would like the process we examine, to stop after a finite number of steps with probability 1. So we would like to examine s.t. such that $\mathbb{P}(T < \infty) = 1$ (*regular times*). At this point it is worth mentioning that when $\mathbb{E}(T) < \infty$ then $\mathbb{P}(T < \infty) = 1$. We can see that for the sets $A_n = \{T \geq n\}$

$$A_{n+1} \subset A_n \text{ και } \{T = \infty\} = \bigcap_{n=1}^{\infty} A_n$$

and therefore from the Markov inequality

$$\mathbb{P}(A_n) \leq \frac{1}{n}\mathbb{E}(T) \Rightarrow \mathbb{P}(T = \infty) \leq \lim_{n \to \infty} \mathbb{P}(A_n) = 0$$

The inverse is not generally true. The expectation of a hitting time of a symmetric random walk is infinite (e.g. see Ghosh 1997, pg. 27).

1.3 The optional stopping theorem

The fundamental way that the stopping times are related to the martingales theory can be given from the next theorem which is called *the optional stopping theorem.*

Theorem 1.3.1. *Let* $(X_n, \mathcal{F}_n)_{n\geq 1}$ *be a martingale and* T *a s.t.. If the following conditions are true*

(i) $\mathbb{P}(T < \infty) = 1$

(ii) $\mathbb{E}|X_T| < \infty$

(iii) $\lim_{n\to\infty} \mathbb{E}\left(X_n I_{\{T>n\}}\right) = 0$

then

$$\mathbb{E}(X_T) = \mathbb{E}(X_1).$$

The next theorem shows a more useful form of the optional stopping theorem.

Theorem 1.3.2. *Let* $(X_n, \mathcal{F}_n)_{n\geq 1}$ *be a martingale and* T *a s.t Then* $\mathbb{E}(X_T) = \mathbb{E}(X_1)$ *if one of the following conditions is true*

(i) $\exists\, c \in \mathbb{N}: T \leq c$ *σ.β.*

(ii) $\mathbb{E}(T) < \infty$ *and* $\exists\, c \in \mathbb{R}$ *such that* $\mathbb{E}(|X_{n+1} - X_n|\, |\mathcal{F}_n) \leq c$ *,* $\forall n < T$.

The proof is based on the use of $(X_{T\wedge n}, \mathcal{F}_{T\wedge n})_{n\geq 1}$ from Example 1.2.2. Writing $X_{T\wedge n}$ as a sum $X_{T\wedge n} = X_1 + \sum_{n=1}^{T\wedge n}(X_{n+1} - X_n)$ and using the dominated convergence and the monotone convergence theorem (e.g. see Cheliotis, 2014, pg. 26), it can be proven that

$$\mathbb{E}(X_1) = \lim_{n\to \infty} \mathbb{E}(X_{T\wedge n}) = \mathbb{E}(X_T)$$

We will see a similar and detailed proof for Wald's second equation.

1.4 Wald's equations

From now on we will make extended use of compound sums of r.v., i.e. sums where the number of terms is a r.v. A typical example is the r.v. that shows the profit of a player in a game, which plays until lose all his profit (*gambler's ruin*). One of the most important results is the fundamental equation of sequential analysis and Wald's two equations.

Theorem 1.4.1. *Let $(X_n)_{n\geq 1}$ be i.i.d. r.v. such that $|X_i| < \infty$, μ be the mean value and σ the variance. Let T be a s.t. such as $\mathbb{E}(T) < \infty$. Then for the compound r.v. $S_T = \sum_{i=1}^{T} X_i$ the following are true:*

(i) $\mathbb{E}(S_T) = \mu \cdot \mathbb{E}(T)$

(ii) $\mathrm{var}(S_T - T\mu) = \sigma^2 \mathbb{E}(T)$, if additionally $\mathrm{var}(X_n) < \infty$.

Proof. (i) Let $Y_n = S_n - n\mu$ and $(\mathcal{F}_n)_{n\geq 1}$ be the natural filtration of the r.v. $X_1, X_2, \dots, X_n$. We will show that the sequence $(Y_n, \mathcal{F}_n)_{n\geq 1}$ is a martingale.

$$Y_{n+1} = S_{n+1} - (n+1)\mu = S_n + X_{n+1} - (n+1)\mu$$

and thus

$$\mathbb{E}(Y_{n+1}|\mathcal{F}_n) = \mathbb{E}(S_n|\mathcal{F}_n) + \mathbb{E}(X_{n+1}|\mathcal{F}_n) - (n+1)\mathbb{E}(\mu \cdot I_\Omega|\mathcal{F}_n) \tag{1.1}$$

We observe that: S_n is $\mathcal{F}_n$-measurable and according to property (ii) of Theorem 1.1.1 we get that

$$\mathbb{E}(S_n|\mathcal{F}_n) = S_n \tag{1.2}$$

The r.v. X_{n+1} is independent from $X_1, X_2, \dots, X_n$ independent of $\mathcal{F}_n$, therefore from property (iv) of Theorem 1.1.1

$$\mathbb{E}(X_{n+1}|\mathcal{F}_n) = \mathbb{E}(X_{n+1}) = \mu \tag{1.3}$$

Finally from property (iii) of Theorem 1.1.1

$$\mathbb{E}(\mu \cdot I_\Omega|\mathcal{F}_n) = \mu \cdot \mathbb{E}(I_\Omega|\mathcal{F}_n) = \mu \tag{1.4}$$

From (1.1)-(1.4) we conclude that $(Y_{n+1}|\mathcal{F}_n) = Y_n$, and so the sequence $(Y_n, \mathcal{F}_n)_{n\geq 1}$ is a martingale. Also:

$$\mathbb{E}(|Y_{n+1} - Y_n| \, |\mathcal{F}_n) = \mathbb{E}(|X_{n+1} - \mu| \, |\mathcal{F}_n) \leq 2\mathbb{E}|X_1| < \infty$$

and applying Theorem 1.2.2:

$$\mathbb{E}(Y_T) = \mathbb{E}(Y_1) = \mathbb{E}(S_1 - \mu) = 0$$

from where it follows that

$$\mathbb{E}(S_T) = \mu \cdot \mathbb{E}(T).$$

(ii) Without loss of generality, let $\mathbb{E}(X_i) = 0$. Set $Z_n = S_n^2 - n\sigma^2$ and $(\mathcal{F}_n)_{n\geq 1}$ be the same as (i). We will firstly show that $(Z_n, \mathcal{F}_n)_{n\geq 1}$ is a martingale. The r.v. Z_n are $\mathcal{F}_n$-measurable, $\mathbb{E}|Z_n| < \infty$ and also

$$\begin{aligned} Z_{n+1} &= S_{n+1}^2 - (n+1)\sigma^2 = (S_n + X_{n+1})^2 - (n+1)\sigma^2 \\ &= S_n^2 + X_{n+1}^2 + 2X_{n+1}S_n - (n+1)\sigma^2 \end{aligned}$$

We can see that $\mathbb{E}(S_n^2|\mathcal{F}_n) = S_n^2$ because S_n^2 and $\mathcal{F}_n$-measurable and because of independency

$$\mathbb{E}(X_{n+1}^2|\mathcal{F}_n) = \mathbb{E}(X_{n+1}^2) = \sigma^2.$$

Also

$$\mathbb{E}(X_{n+1}S_n|\mathcal{F}_n) = \mathbb{E}(X_{n+1}|\mathcal{F}_n)S_n = \mathbb{E}(X_{n+1})S_n = 0$$

So $\mathbb{E}(Z_{n+1}|\mathcal{F}_n) = S_n^2 + \sigma^2 - (n+1)\sigma^2 = Z_n$. $(Z_n, \mathcal{F}_n)_{n\geq 1}$ is a martingale and thus $(Z_{T\wedge n}, \mathcal{F}_{T\wedge n})_{n\geq 1}$ is also a martingale according to Example 1.2.2. From Example 1.1.1

$$\mathbb{E}(Z_{T\wedge n}) = \mathbb{E}(Z_1) = 0$$

this implies

$$\mathbb{E}(S_{T\wedge n}^2) = \sigma^2 \mathbb{E}(T \wedge n) \tag{1.5}$$

$T \wedge n$ is an increasing sequence of r.v. and $T \wedge n \xrightarrow{n\to\infty} T$. Applying the monotone convergence theorem we get that

$$\lim_{n\to\infty} \mathbb{E}(T \wedge n) = \mathbb{E}(T)$$

In order to enter the limit $n \to \infty$ on the left part of the inequality (1.5) we will use the dominated convergence theorem

$$\begin{aligned} S_{T\wedge n}^2 \leq S_T^2 = \left(\sum_{i=1}^{T} X_i\right)^2 &= \sum_{i=1}^{T} X_i^2 + 2\sum_{j<i}^{T} X_i X_j \\ &= \sum_{i=1}^{\infty} X_i^2 I_{\{T\geq i\}} + 2\sum_{j<i}^{\infty} X_i X_j I_{\{T\geq i\}} I_{\{T\geq j\}} \end{aligned} \tag{1.6}$$

We can see that the r.v. $I_{\{T\geq i\}}$ and $X_j I_{\{T\geq i\}} I_{\{T\geq j\}}$ are $\mathcal{F}_{i-1}$-measurable. Because the r.v. X_i is independent of the σ-algebra $\mathcal{F}_{i-1}$ it follows that

$$\mathbb{E}(S_T^2) = \sigma^2 \sum_{i=1}^{\infty} \mathbb{P}(T \geq i) = \sigma^2\, \mathbb{E}(T) < \infty \tag{1.7}$$

From (1.6) and (1.7) and the dominated convergence theorem we get the required equation. □

1.5 The Fundamental Equation

The following theorem connects the r.v. T, with the compound sum S_T and the moment generating function of the r.v. X_i and is referred in bibliography as *the fundamental equation of sequential analysis* or *Wald's fundamental equation.*

Theorem 1.5.1. *Let* $(X_n)_{n\geq 1}$ *be i.i.d r.v. with m.g.f.* $M(t) = \mathbb{E}(e^{tX})$. *If the following are true*

(i) $\exists\ t_0 \neq 0$ *such as* $1 \leq M(t_0) < \infty$

(ii) $\exists\ c \in \mathbb{R}_+$ *and s.t.* T *with* $\mathbb{E}(T) < \infty$ *and* $|S_n| \leq c\ , \forall n < T$

then

$$\mathbb{E}(e^{t_0 S_T} M(t_0)^{-T}) = 1.$$

Proof. Set $Y_n = \frac{e^{t_0 S_n}}{M(t_0)^n}$ and let $(\mathcal{F}_n)_{n\geq 1}$ be the natural filtration. We will show that the sequence $(Y_n, \mathcal{F}_n)_{n\geq 1}$ is a martingale. Indeed

$$\begin{aligned}\mathbb{E}(Y_{n+1}|\mathcal{F}_n) &= \mathbb{E}\left(\frac{e^{t_0 S_{n+1}}}{M(t_0)^{n+1}}|\mathcal{F}_n\right) = \mathbb{E}\left(\frac{e^{t_0 S_n}}{M(t_0)^n}\cdot\frac{e^{t_0 X_{n+1}}}{M(t_0)}|\mathcal{F}_n\right)\\ &= Y_n\cdot\mathbb{E}\left(\frac{e^{t_0 X_{n+1}}}{M(t_0)}|\mathcal{F}_n\right) = Y_n M(t_0)^{-1}\mathbb{E}(e^{t_0 X_{n+1}}) = Y_n\end{aligned}$$

and from the properties (ii) and (v) of Theorem 1.1.1. It is true that

$$\mathbb{E}(|Y_{n+1} - Y_n|\mathcal{F}_n) = Y_n\cdot\mathbb{E}\left(\left|\frac{e^{t_0 X_{n+1}}}{M(t_0)} - 1\right|\right) \leq \frac{Y_n}{M(t_0)}\mathbb{E}(e^{t_0 X_{n+1}} + M(t_0)) = 2Y_n$$

From (ii) $\forall n < T$: $Y_n = \frac{e^{t_0 S_n}}{M(t_0)^n} \leq \frac{e^{t_0|S_n|}}{M(t_0)^n} \leq e^{t_0 c}$. Applying Theorem 1.3.2 we get that $\mathbb{E}(Y_T) = \mathbb{E}(Y_1) \Rightarrow \mathbb{E}(e^{t_0 S_T} M(t_0)^{-T}) = 1.$ □

In order to simplify the use of the above theorem, it suffices to find proper relations for the m.g.f. $M(t)$ so that there exist a $t_0 \neq 0$ with $M(t_0) = 1$. Taking twice the derivative of the m.g.f.

$$M''(t) = \mathbb{E}(X^2 e^{tX}) \tag{1.8}$$

If $\mathbb{P}(X > 0) > 0$, then $\exists\ \varepsilon > 0$ and $\delta \in (0,1)$ such as $\mathbb{P}(X > \varepsilon) = \delta$. Thus from (1.8) $M''(t) > 0$, the curve γ of the m.g.f. is convex. So $y = 1$ can only intersect the curve in two points. We would like γ not to be tangent to the line $y = 1$, at the point with abscissa $t = 0$. So we want $M'(0) = \mu \neq 0$. There are two cases related to the sign of the mean value μ. If $\mu < 0$

($\mu > 0$) we demand $\lim_{t \to +\infty} M(t) = +\infty$ ($\lim_{t \to -\infty} M(t) = +\infty$) so that $y = 1$ intersects γ in at least on point different from $t = 0$. To get this limit we want the moment generating function to be bounded by a function of t which diverges to infinity as $t \to +\infty$. If $\mathbb{P}(X > 0) > 0$ then

$$\forall t > 0 \colon M(t) = \mathbb{E}(e^{tX}) \geq e^{t\varepsilon}\delta$$

and letting t go to infinity we get the required limit for γ. Similarly one can prove the equivalent limit when $\mu > 0$. We conclude to the following result.

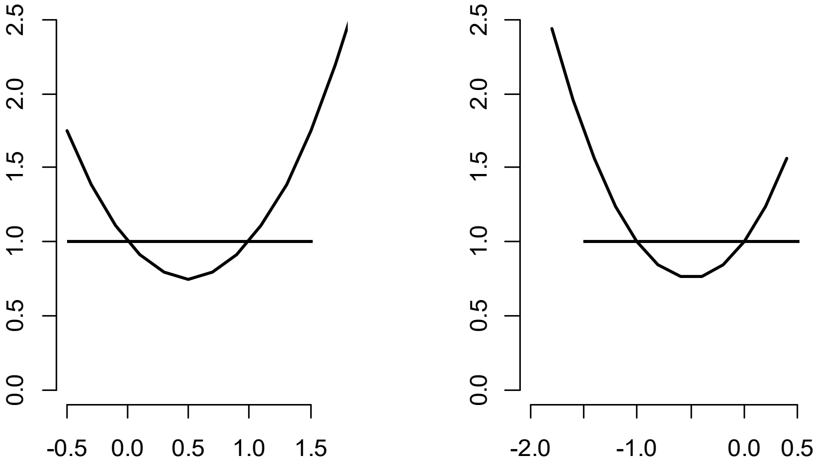


Figure 1.5.1 *Plot of the moment generating function when the r.v. has negative and positive mean value*

Proposition 1.5.1. *If for a r.v. X (i)* $\mathbb{P}(X > 0) > 0$, *(ii)* $\mathbb{P}(X < 0) > 0$, *(iii)* $\mathbb{E}(X) \neq 0$ *and (iv)* $M(t) < \infty$, *then* $\exists!\, t_0 \in \mathbb{R} \setminus \{0\} : M(t_0) = 1$.

The comparison of the above statement with the fundamental equation of sequential analysis is generally useful for finding probabilities of the form $\mathbb{P}(S_T > k)$ where T is an exit time and k a constant. Also, there are applications related to the approximation of the curves OC and ASN, to the approximation of the ARL function of the CUSUM algorithm and to problems that can be found in renewal theory (e.g. see Ghosh 1997, pg. 50-55).

1.6 Optimal stopping theory

We have seen so far how the expectation of a stopping time can be related to the expectation of a compound sum. In many statistical problems, we know the type of the s.t. and try to investigate its properties using Wald's equation. However, one can define an inverse problem: given several properties of the s.t. find its form.

Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. r.v., $(\mathcal{F}_n)_{n\geq 1}$ be the natural filtration and T a s.t. with respect to the filtration such as

$$\mathbb{P}(1 \leq T < \tau) = 1 \tag{1.9}$$

where $\tau \in [0 + \infty)$. Denote Δ_τ the set of all the s.t. with respect to the above filtration that satisfy (1.9). If $(Y_n)_{n\geq 1}$ is a stochastic process of the space $(\Omega, \mathcal{F}, \mathbb{P})$, then the problem of o.s.t. is to compute the quantity

$$V = \sup_{T\in\Delta_\tau} \mathbb{E}(Y_T) \tag{1.10}$$

and find the formula of T which maximizes it. Obviously we could define the same problem for the lower bound in relation (1.10). Shiryaev (2013) presents the methodology for finding a solution of (1.10) with the use of martingale theory. We will see in Chapter 4 some problems of o.s.t. that include statistical algorithms.

CHAPTER 2

Sequential Interval Estimation

2.1 Calculating the optimal sample size

In this paragraph we will prove that the problem of constructing a 100(1-a)% confidence interval for the mean value of the normal distribution, that has an a priori fixed width, cannot be solved using any fixed sample-size methodology.

Denote $\mathbf{X} = (X_1, \dots, X_n)$ the random sample of size n, $\mathbf{x} = (x_1, .. x_n)$ one realization and $\delta(\mathbf{X})$ an estimator of the mean value μ of the sample distribution. Then the *0-1 loss function* (see Mukhopadhyay, 2009, pg. 16) for the specific problem can be written as

$$W(\delta(\mathbf{X}), \mu) = \begin{cases} 0, & |\delta(\mathbf{X}) - \mu| \leq d \\ 1, & |\delta(\mathbf{X}) - \mu| > d \end{cases} \tag{2.1}$$

where $d > 0$ is the fixed width of the c.i. We will show that

$$\nexists \delta(\mathbf{X}) : \mathbb{P}(|\delta(\mathbf{X}) - \mu| \leq d) \geq 1 - a \tag{2.2}$$

The next theorem can be found in Lehman (1951).

Theorem 2.1.1 *Suppose that for the i.i.d. r.v.*

$$X_1, \dots, X_n \sim \sigma^{-1} f\left(\frac{x - \theta}{\sigma}\right),$$

where $\theta \in \mathbb{R}$, $\sigma \in \mathbb{R}_+$, are unknown parameters. Assume that the loss function is given by the relation $W(\delta(\mathbf{X}), \theta) = H(|\delta(\mathbf{X}) - \theta|)$, where $H(|y|)$ is an increasing function of $|y|$. If $M = \sup\{H(|y|): y \in \mathbb{R}\}$, then $\forall L < M$ there does not exist $\delta(\mathbf{X})$ such as

$$\sup\{\mathbb{E}[W(\delta(\mathbf{X}), \theta)]: (\theta, \sigma) \in (\mathbb{R}, \mathbb{R}_+)\} \leq L.$$

We observe that for the loss function (2.1)

$$\mathbb{E}[W(\delta(\mathbf{X}), \theta)] = \mathbb{P}(|\delta(\mathbf{X}) - \mu| > d)$$

Using $M = 1, L = a$, and according to Theorem 2.1.1 we conclude that

$$\nexists \delta(\mathbf{X}) : \mathbb{P}(|\delta(\mathbf{X}) - \mu| > d) \leq a$$

and thus (2.2) is valid.

Taking into account the above statement, we would like to estimate n and the interval $I_n = [\bar{X}_n - d, \bar{X}_n + d]$ for which

$$\mathbb{P}(\mu \in I_n) \geq 1 - a \tag{2.3}$$

We will examine the case of the normal distribution. The general theory for other distributions can be found in Ghosh (1997).

Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2), 0 < \sigma < \infty$. Assume initially that the variance in unknown. After calculations we get that

$$\mathbb{P}(\mu \in I_n) = \mathbb{P}(|\bar{X}_n - \mu| \leq d) = 2\Phi\left(\frac{d\sqrt{n}}{\sigma}\right) - 1 \tag{2.4}$$

In order to solve the inequality (2.3) using (2.4), it suffices to express a with the use of the function Φ. Therefore we have to find a x so that

$$1 - a = 2\Phi(x) - 1$$

and thus

$$x = \Phi^{-1}\left(1 - \frac{a}{2}\right) = z_{a/2} \tag{2.5}$$

From (2.3), (2.4) and (2.5):

$$2\Phi\left(\frac{d\sqrt{n}}{\sigma}\right) - 1 \geq 2\Phi(z_{a/2}) - 1$$

we conclude that

$$n \geq \frac{\sigma^2}{d^2} z_{a/2}^2 = k \tag{2.6}$$

According to the above results we now have that

$$\mathbb{P}(\mu \in I_n) \geq 1 - a \Leftrightarrow n \geq k \tag{2.7}$$

Since the variance σ^2 was known for calculating k, then k will be the optimal sample size for the construction of the confidence interval.

2.2 Stein's method

The constatnt k in the relation (2.7) can be estimated using the sample variance s^2 when σ^2 is unknown. The $a/2$ upper quantile point of the normal distribution will be replaced with the equivalent point $t^2_{n_0-1;a/2} \equiv t^2_{n_0-1}$ of Student's distribution, where the degrees of freedom are reduced by 1, and so we take an estimator $\hat{k}$ of k. Estimating the variance using an initial sample size n_0, n_0 can satisfy (2.7). Otherwise we chose $\hat{k} - n_0$ observations and the final c.i. will be $[\bar{X}_{\hat{k}} - d, \bar{X}_{\hat{k}} + d]$. Stein (1949) proposed the following estimator of the sample size:

$$T = \max\left\{n_0, \left\lfloor \frac{t^2_{n_0-1} S^2_{n_0}}{d^2} \right\rfloor + 1\right\} \tag{2.8}$$

where $\lfloor x \rfloor$ is the biggest integer which is smaller or equal to x. We can see that the r.v. T depends upon the sample variance. Thus, if $(\mathcal{F}_n)_{n\geq 1}$ is the natural filtration, then the r.v. T is a s.t. Clearly $\mathbb{P}(T < \infty)$.

In order to prove that using Stein's method (1949) we can solve the problem stated in Paragraph (2.1), we will need the following result.

Theorem 2.2.1. *Let $X_1, \dots, X_n$ be a random sample from normal distribution. The r.v. $\bar{X}_n = \sum_{i=1}^{n} X_i$ and $S^2_n = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ are stochastically independent.*

We also conclude that the r.v. $\bar{X}_n$ and $I(T = n)$ are independent $\forall n \geq n_0$ (see Mukhopadhyay, 2009, pg. 101). Therefore:

$$\begin{aligned}\mathbb{P}(\mu \in I_T) &= \sum_{n=n_0}^{\infty} \mathbb{P}(\{|\bar{X}_n - \mu| \leq d\} \cap \{T = n\}) \\ &= \sum_{n=n_0}^{\infty} \mathbb{P}(|\bar{X}_n - \mu| \leq d)\, \mathbb{P}(T = n) \\ &= \mathbb{E}\left[2\Phi\left(\frac{\sqrt{T} d}{\sigma}\right) - 1\right]\end{aligned} \tag{2.9}$$

When the width of the c.i. is decreasing, i.e. $d \to 0$, we will need more observations for its construction and so we will refer to an asymptotic property. In the following theorem we state the properties of the method, with respect to Chow and Robbins's (1965) definition.

Theorem 2.2.2. *For the stopping time T and for every μ and σ^2, the following are true:*

(i) $\lim_{d\to 0} \mathbb{E}\left(\frac{T}{k}\right) > 1$ *(asymptotic inefficiency)*

(ii) $\mathbb{P}(\mu \in I_T) \geq 1 - a$ *(exact consistency)*

(iii) $\lim_{d\to 0} \mathbb{P}(\mu \in I_T) = 1 - a$ *(asymptotic consistency)*

Proof: (i). From the definition (2.8) of the r.v. T we have that

$$\frac{t_{n_0-1}^2 S_{n_0}^2}{d^2} \leq T \leq n_0 + \frac{t_{n_0-1}^2 S_{n_0}^2}{d^2} \tag{2.10}$$

and taking the expectation

$$\frac{t_{n_0-1}^2 \sigma^2}{d^2} \leq \mathbb{E}(T) \leq n_0 + \frac{t_{n_0-1}^2 \sigma^2}{d^2}$$

Dividing by k the following relation, where k was defined in (2.6), and letting d tend to zero

$$\lim_{d\to 0} \mathbb{E}\left(\frac{T}{k}\right) = \frac{t_{n_0-1}^2}{z_{a/2}^2} \tag{2.11}$$

We can use the Cornish-Fisher approximation (Johnson, 1970) for the quantile points, and we can express the $a/2$ quantile point of Student's distribution as a function of the equivalent quantile point of the standard normal distribution:

$$t_{n_0-1} = z_{a/2} + \frac{z_{a/2}(z_{a/2}^2 + 1)}{4(n_0 - 1)} + O(n_0^{-2})$$

So we get $\frac{t_{n_0-1}}{z_{a/2}} > 1$ and using (2.11) we prove (i).

(ii). From the inequality (2.10) and the relation (2.9) it follows that

$$\begin{aligned}\mathbb{P}(\mu \in I_T) &\geq \mathbb{E}\left[2\Phi\left(\tfrac{t_{n_0-1}S_{n_0}}{\sigma}\right) - 1\right] \\ &= \mathbb{E}\left[\mathbb{P}\left(|Z| \leq t_{n_0-1}S_{n_0}\sigma^{-1} | S_{n_0}\right)\right], \ Z\sim N(0,1) \\ &= \mathbb{E}\left[\mathbb{E}\left(I(|Z| \leq t_{n_0-1}S_{n_0}\sigma^{-1}) | S_{n_0}\right)\right], \\ &= \mathbb{E}\left[I(|Z| \leq t_{n_0-1}S_{n_0}\sigma^{-1})\right] = \mathbb{P}\left(|Z| \leq t_{n_0-1}S_{n_0}\sigma^{-1}\right) \\ &= \mathbb{P}\left(Y \leq t_{n_0-1}\right), \ Y\sim t(n_0 - 1) \\ &= 1 - a\end{aligned} \tag{2.12}$$

(iii). It is true that $\sqrt{T}d \xrightarrow{d\to 0} t_{n_0-1} S_{n_0}$, $\sqrt{T}d \leq 2t_{n_0-1}S_{n_0-1} = W$ for small values of d, according to inequality (2.10), and $\mathbb{E}[|W|] < \infty$. This implies that using the dominated convergence theorem we get

$$\mathbb{E}\left(2\Phi\left(\frac{\sqrt{T}d}{\sigma}\right)-1\right)\overset{d\to 0}{-}\mathbb{E}\left(2\Phi\left(\frac{t_{n_0-1}S_{n_0}}{\sigma}\right)-1\right)=1-a.$$

where the last equality followed from (2.12). □

2.3 Distribution of the sample size

We get the following results related to the sample size. Particularly, we compute the probabilities

$$\begin{aligned}\mathbb{P}(T=n_0)&=\mathbb{P}\left(\left\lfloor\frac{t_{n_0-1}^2S_{n_0}^2}{d^2}\right\rfloor+1\leq n_0\right)=\mathbb{P}\left(0\leq\frac{t_{n_0-1}^2S_{n_0}^2}{d^2}\leq n_0\right)\\&=\mathbb{P}\left(0<Y\leq\frac{n_0(n_0-1)d^2}{\sigma^2t_{n_0-1}^2}\right)\end{aligned}$$

where $Y{\sim}\chi_{n_0-1}^2$. Similarly for every $m\in\mathbb{N}$:

$$\begin{aligned}\mathbb{P}(T=n_0+m)&=\mathbb{P}\left(\left\lfloor\frac{t_{n_0-1}^2S_{n_0}^2}{d^2}\right\rfloor=n_0+m-1\right)\\&=\mathbb{P}\left(n_0+m-1<\frac{t_{n_0-1}^2S_{n_0}^2}{d^2}\leq n_0+m\right)\\&=\mathbb{P}\left(\frac{(n_0+m-1)(n_0-1)d^2}{\sigma^2t_{n_0-1}^2}<Y\leq\frac{(n_0+m)(n_0-1)d^2}{\sigma^2t_{n_0-1}^2}\right),\end{aligned}$$

where $Y{\sim}\chi_{n_0-1}^2$. If we set

$$g_{n_0}(y)=\left\{2^{(n_0-1)/2}\Gamma\left(\tfrac{n_0-1}{2}\right)\right\}^{-1}\exp(-y/2)y^{(n_0-3)/2}\ ,\ y>0$$

and $c_m=\frac{(n_0+m)(n_0-1)d^2}{\sigma^2t_{n_0-1}^2}$, $k=0,1,2,...$, we can then write the above probabilities as

$$\mathbb{P}(T=n_0)=\int_0^{c_0}g_{n_0}(y)dy\ ,\ \text{και}\ \ \mathbb{P}(T=n_0+m)=\int_{c_{m-1}}^{c_m}g_{n_0}(y)dy.$$

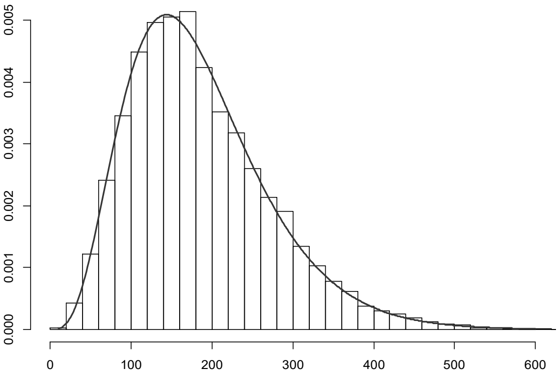


Figure 2.3.1. *Histogram of the stopping time T*

It is clear that the simulated values of the r.v. T are almost the same with the theoretical ones. The code required to construct the histogram, can be found in Appendix III. Specifically, we applied the following: We created a loop of $n = 10^4$ repetitions where for each repetition we generated $n_0 = 10$ random numbers from $N(1, 3^2)$, calculated the r.v. T for $\alpha = 0.05$ and stored its value in a matrix T_s. In order to calculate the probabilities $\mathbb{P}(T = j)$ we considered two cases: If $j = n_0$ and $j = n_0 + k$, where the values of k range from 1 to 590. Using the above formulas we stored the probabilities in the matrix *Prob* and calculated the theoretical values.

2.4 Purely sequential method

Anscombe (1952) was the first to introduce a purely sequential method for the construction of c.i. with a priori fixed width. The s.t. will now be

$$T = \min\left\{n \geq n_0 \,:\, n \geq \frac{z_{a/2}^2 S_n^2}{d^2}\right\} \tag{2.13}$$

where $n_0 \geq 2$ is the initial sample size of the procedure. The final c.i. will be $I = [\bar{X}_T - d, \bar{X}_T + d]$. The following results, related to the sequential method, can be found in Woodroofe (1977) and Chow and Robbins (1965).

Theorem 2.4.1 *For the stopping time T, the following are true:*

(i) $\mathbb{E}(T) - k = g(1) + o(1)$ *, for* $n_0 \geq 4$

(ii) $\mathbb{P}(\mu \in I_T) = 1 - a + \frac{z_{a/2}^2}{2k}\left(2g(1) - 1 - z_{a/2}^2\right) f\left(z_{a/2}^2; 1\right) + o(C^{-1})$

(iii) $\lim_{d \to 0} \mathbb{E}\left(\frac{T}{k}\right) = 1$

(iv) $\lim_{d \to 0} \mathbb{P}(\mu \in I_T) = 1 - a$

Where $g(x) = \frac{1}{2} - \frac{1}{x} - \frac{1}{x}\sum_{n=1}^{\infty}\frac{1}{n}\mathbb{E}(\max\{0, \chi_{nx}^2 - 2nx\})$*,* χ_{nx}^2 *is a r.v. having* χ^2 *distribution with* nx *degrees of freedom and* f *is the p.d.f. of the* χ_1^2 *distribution.*

Since $g(1) < 0$, using (i) and (ii) we have that $\mathbb{E}(T) < k$ and $\mathbb{P}(\mu \in I_T) < 1 - a$. Therefore the sequential method does not satisfy the property of exact consistency. We can use Wald's equation to find an upper bound for the s.t. T. From the definition:

$$T - 1 < n_0 - 1 + \frac{z_{a/2}^2 S_{T-1}^2}{d^2}$$

and

$$(T - n_0)(T - 2) < \frac{z_{a/2}^2}{d^2}\sum_{i=1}^{T-1}(X_i - \bar{X}_{T-1})^2 < \frac{z_{a/2}^2}{d^2}\sum_{i=1}^{T}(X_i - \mu)^2 = \frac{z_{a/2}^2}{d^2}\sum_{i=1}^{T}Y_i^2.$$

Thus

$$T^2 - (n_0 + 2)T < (T - n_0)(T - 2) < \frac{z_{a/2}^2}{d^2}\sum_{i=1}^{T}Y_i^2$$

$$\mathbb{E}(T^2) - (n_0 + 2)\mathbb{E}(T) < \frac{z_{\alpha/2}^2}{d^2}\sigma^2\mathbb{E}(T) = k\mathbb{E}(T).$$

But since $0 < \sigma^2 = \mathbb{E}(T^2) - \left(\mathbb{E}(T)\right)^2$, we end up with the inequality

$$\mathbb{E}(T) < k + n_0 + 2 \text{ , } k = \frac{z_{\alpha/2}^2}{d^2}\sigma^2.$$

2.5 Extension in *p*-dimensions

We will extend the methods for constructing c.i. with a priori known width, in the case of random vectors.

Let $\mathbf{X}_1$, $\mathbf{X}_2$, …,$\mathbf{X}_n \sim N_p(\boldsymbol{\mu}, \Sigma)$, $\boldsymbol{\mu}$ is a known vector and Σ is the symmetric and positive definite variance-covariance matrix. We denote

$$\bar{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i \text{ , } S_n = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}}_n)(\mathbf{X}_i - \bar{\mathbf{X}}_n)^t,$$

where λ_i are the eigenvalues of Σ and $\lambda_{(p)}$ is its largest eigenvalue. In order to prove some basic results we will need the following proposition

Proposition 2.5.1. *The following are true* (e.g. see. Johnson 2007, pg. 78, 163)*:*

(i) $\max_{\mathbf{x}\neq\boldsymbol{O}} \frac{\mathbf{x}^t\Sigma\mathbf{x}}{\mathbf{x}^t\mathbf{x}} = \lambda_{(p)}$

(ii) Σ *is diagonalizable*

(iii) $n(\overline{\mathbf{X}}_n - \boldsymbol{\mu})^t\Sigma^{-1}(\overline{\mathbf{X}}_n - \boldsymbol{\mu}) \sim \chi_p^2$

We need to find a sample size so that the spherical region of radius d that we will construct, will have confidence coefficient at least $1-\alpha$, $\alpha \in (0,1)$. Let

$$\Pi_n = \{\mathbf{x} \in \mathbb{R}^p : (\overline{\mathbf{X}}_n - \mathbf{x})^t(\overline{\mathbf{X}}_n - \mathbf{x}) \leq d^2\}$$

From (ii): $\exists \mathrm{L}, \mathrm{D} \in \mathcal{M}_{p\times p}(\mathbb{R})$, where L is orthogonal and $\mathrm{D} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ such as $\Sigma = \mathrm{L^tDL}$. Setting $\mathbf{Y} = \mathrm{L}(\overline{\mathbf{X}}_n - \boldsymbol{\mu})$ we get the following:

$$(\overline{\mathbf{X}}_n - \boldsymbol{\mu})^t\Sigma^{-1}(\overline{\mathbf{X}}_n - \boldsymbol{\mu}) = \mathbf{Y}^t\mathrm{D}^{-1}\mathbf{Y} \leq \frac{\mathbf{Y}^t\mathbf{Y}}{\lambda_{(p)}} = \frac{(\overline{\mathbf{X}}_n - \boldsymbol{\mu})^t(\overline{\mathbf{X}}_n - \boldsymbol{\mu})}{\lambda_{(p)}}$$

and therefore:

$$\mathbb{P}(\boldsymbol{\mu} \in \Pi_n) \geq \mathbb{P}\big(\lambda_{(p)}(\overline{\mathbf{X}}_n - \boldsymbol{\mu})^t\Sigma^{-1}(\overline{\mathbf{X}}_n - \boldsymbol{\mu}) \leq d^2\big)$$

$$= \mathbb{P}\left(n(\overline{\mathbf{X}}_n - \boldsymbol{\mu})^t\Sigma^{-1}(\overline{\mathbf{X}}_n - \boldsymbol{\mu}) \leq \frac{nd^2}{\lambda_{(p)}}\right) = F\left(\frac{nd^2}{\lambda_{(p)}}\right)$$

where F is the c.d.f. of χ_p^2. Let a be the upper α-quantile point of the χ_p^2 distribution. Thus, $F(a) = 1 - \alpha$. We concluded that

$$\mathbb{P}(\boldsymbol{\mu} \in \Pi_n) \geq 1 - \alpha \Leftrightarrow n \geq \frac{a\lambda_{(p)}}{d^2} = C$$

To construct a spherical region with confidence coefficient $1-\alpha$, the sample size must be at least $n = \left\lfloor \frac{a\lambda_{(p)}}{d^2} \right\rfloor + 1$.

2.5.1 Healy's method

Healy (1956) proposed a method similar to Stein's (1949) two stage procedure. We take an initial sample $\mathbf{X}_1$, $\mathbf{X}_2$, …,$\mathbf{X}_m$, where $m \geq p+1$. Then calculate $\overline{\mathbf{X}}_m$, S_m and the maximum eigenvalue $\lambda_{(p)m}$ of the matrix S_m. We define the stopping time

$$T = \max\left\{m, \left\lfloor \frac{u\lambda_{(p)}}{d^2} \right\rfloor + 1\right\}$$

where $u = T_\alpha^2 = \frac{p(m-1)}{m-p} F_{p,m-p,\alpha}$ is the upper α-quantile point $T_{p,m-p}^2$ of Hotelling's distribution and $F_{p,m-p,\alpha}$ is the upper α-quantile point of $F_{p,m-p}$ distribution. If $T > m$ then we collect $T - m$ random vectors and construct the final $100(1-\alpha)\%$ confidence region

$$\Pi_T = \{\mathbf{x} \in \mathbb{R}^p : (\overline{\mathbf{X}}_T - \mathbf{x})^t (\overline{\mathbf{X}}_T - \mathbf{x}) \le d^2\}.$$

Theorem 2.5.1. *For every* $\boldsymbol{\mu}$, Σ, p, m, d *and* α *the following are true*

(i) $T(\overline{\mathbf{X}}_T - \boldsymbol{\mu})^t S_m^{-1} (\overline{\mathbf{X}}_T - \boldsymbol{\mu}) \sim T^2(p, m-p)$

(ii) $\mathbb{P}(\boldsymbol{\mu} \in \Pi_T) \ge 1 - \alpha$ *(exact consistency)*

Proof. (i). The proof is based on the independence of the random vector $\overline{\mathbf{X}}_n$ with $I(T = n)$ (e.g. see. Mukhopadhyay, 2009).

(ii) We compute that

$$\begin{aligned}\mathbb{P}((\overline{\mathbf{X}}_T - \boldsymbol{\mu})^t (\overline{\mathbf{X}}_T - \boldsymbol{\mu}) \le d^2) &\ge \mathbb{P}\left((\overline{\mathbf{X}}_T - \boldsymbol{\mu})^t S_m^{-1} (\overline{\mathbf{X}}_T - \boldsymbol{\mu}) \le \frac{d^2}{\lambda_{(p)m}}\right) \\ &= \mathbb{P}\left(T(\overline{\mathbf{X}}_T - \boldsymbol{\mu})^t S_m^{-1} (\overline{\mathbf{X}}_T - \boldsymbol{\mu}) \le \frac{Td^2}{\lambda_{(p)m}}\right) \\ &\ge \mathbb{P}(T(\overline{\mathbf{X}}_T - \boldsymbol{\mu})^t S_m^{-1} (\overline{\mathbf{X}}_T - \boldsymbol{\mu}) \le u) = 1 - a\end{aligned}$$

where the first inequality follows from exactly the same way as for the matrix Σ and the second inequality is valid because $\frac{Td^2}{\lambda_{(p)m}} \ge u$, according to the definition of the r.v. T and (i). □

We proved that Healy's method (1956) satisfies the property of exact consistency.

2.5.2 Srivastava's method

Assume now that we collect an initial sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m$ of size $m \ge p + 1$. Srivastava (1967) defined the stopping time

$$T = \min\left\{n \ge m : n \ge \left(\frac{a}{d^2}\right) \lambda_{(p)n}\right\} \tag{2.14}$$

where a is the upper α-quantile point of the χ_p^2 distribution. Each time we collect a random vector, compute the matrix S_n, its maximum eigenvalue and examine whether or not the condition $n \ge \left(\frac{a}{d^2}\right) \lambda_{(p)n}$ is true. The final region will be Π_T.

Theorem 2.5.2. *For the stopping time (2.14) and for every* $\boldsymbol{\mu}$, Σ, p, m, α *we have that:*

(i) $\mathbb{P}(T < \infty) = 1$

(ii) $\lim_{d \to 0} \mathbb{E}\left({}^{T}\!/_{C}\right) = 1$ *(asymptotic efficiency)*

(iii) $\lim_{d\to 0}\mathbb{P}(\boldsymbol{\mu}\in\Pi_n)=1-\alpha$ *(asymptotic consistency)*

Srivastava's (1967) method inherits the first order asymptotic properties but does not have Healy's exact consistency. The proof of the theorem is based on a similar methodology used in the univariate case (see Mukhopadhyay, 2009, pg. 288-289).

2.6 Comparing methods via simulation

In order to examine the exact and asymptotic properties of the sequential and Stein's method and also the methods for constructing confidence regions with fixed radius, we will use Monte Carlo simulation for the estimation of the mean number of the sample size, when $\alpha = 0.05$.

For Stein's method (Appendix IV, Program 1) we create a loop of $n = 10^3$ repetitions and each time generate $n_0 = 10$ random numbers from $N(1,3^2)$, calculate the value of the r.v. T and store the value in a matrix T_s. If $T \le n_0$, the final c.i. will be $I = [\bar{X}_T - d, \bar{X}_T + d]$. But if $T > n_0$ we generate $T - n_0$ more random numbers from $N(1,3^2)$ and calculate $\bar{X}_T$. Having now the final sample and the c.i. I, we check if $\mu \in I$. If it does so, we increase the value of a counter variable *s* by one, to use it for the calculation of the probability $p = \mathbb{P}(\mu \in I)$. We repeat the above procedure for six values of the radius d. The *matrix mat.Stein* will include the values of the optimal sample size k, the mean value $\bar{T}$, the standard deviation $SD(T)$, the ratio $\bar{T}/k$ and the estimated probability, p for various values of d.

For the sequential method (see Appendix IV, Program 2) we use a loop of $n = 10^3$ repetitions in order to get values of the r.v. T. Each time we generate $n_0 = 10$ random numbers from $N(1,3^2)$ and compute the quantity $k_0 = \frac{z_{\alpha/2}^2 s_{n_0}^2}{d^2}$. To find the minimum i that satisfies the condition $i \ge k_0$ (see the definition of the r.v. T) we use the while loop, generating each time a random number from $N(1,3^2)$ and calculating the new k_0, until $i \ge k_0$. Having estimated the s.t. $T = i$, we examine if $\mu \in I$, similarly as Stein's method. All the above are repeated for 6 values of the radius d. The final matrix *mat.Seq* will include similar quantities to the matrix *mat.Stein.*

As far as Healy's method is concerned (see Appendix IV, Program 3) we apply the following: We calculate initially the maximum eigenvalue of the matrix $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 2 \end{bmatrix}$ and the optimal sample size for a given value of d. We construct a loop of $n = 10^3$ repetitions, where at each repetition we generate $m = 10$ random vectors from the distribution $N(\boldsymbol{\mu},\Sigma)$, where $\boldsymbol{\mu} = [1 \quad 2]$, calculate the maximum eigenvalue of the sample variance matrix S and then the r.v. T according to the definition. If $T \le m$ the confidence region will be $\Pi = \{\mathbf{x} \in \mathbb{R}^p : (\bar{\mathbf{X}}_T - \mathbf{x})^t(\bar{\mathbf{X}}_T - \mathbf{x}) \le d^2\}$. But if $T > m$, we generate $T - m$ more vectors, calculate the new $\bar{\mathbf{X}}_T$ and the region Π. We check if $\boldsymbol{\mu} \in \Pi$, increasing the value of a counter variable *sum* so as to estimate $\mathbb{P}(\boldsymbol{\mu} \in \Pi)$. All the above are repeated for 6 values of the radius d. The matrix *mat.Healy* will have a similar form to the univariate case.

Finally, for Srivastava's method (see Appendix IV, Program 4) we compute initially the maximum eigenvalue of the matrix Σ and the value of the optimal sample size C. In a loop of $n = 10^3$ repetitions we apply the following: we generate $m = 10$ random vectors from $N(\boldsymbol{\mu}, \Sigma)$ and calculate the maximum eigenvalue of the sample covariance matrix S. Inside a while loop with condition $N < \frac{\alpha}{d^2}\lambda_n$, we generate each time a random vector from $N(\boldsymbol{\mu}, \Sigma)$ and compute the maximum eigenvalue of the new sample matrix S. At the end of the loop, we will have taken an estimation of the s.t. T. Similarly as the previous method, we use a counter *sum* to estimate the probability $\mathbb{P}(\boldsymbol{\mu} \in \Pi)$ and store the results in the matrix *mat.Sriva* for 6 values of d. Therefore, we take the following tables.

TABLE 2.6.1
Exact and asymptotic interval estimation (p=1)

		Stein's Method				*Purely Sequential Method*			
d	k	$\overline{T}$	$SD(T)$	$\overline{T}/k$	$\Pr(\mu \in I_T)$	$\overline{T}$	$SD(T)$	$\overline{T}/k$	$\Pr(\mu \in I_T)$
1	34.6	46.8	21.6	1.354	0.950	32.9	9.6	0.952	0.940
0.75	61.5	81.5	37.2	1.326	0.952	60.2	12.1	0.979	0.948
0.5	138.3	183.5	82.7	1.327	0.948	136.8	17.5	0.989	0.941
0.25	553.2	757.4	348.3	1.369	0.952	551.3	33.8	0.997	0.940
0.15	1537.6	2034.6	971.3	1.324	0.951	1538.1	56.7	1.001	0.951
0.05	13829.3	18445.1	8906.1	1.334	0.944	13830.0	163.4	1.000	0.959

TABLE 2.6.2
Exact and asymptotic region estimation (*p=2*)

		Healy's Method				*Srivastava's Method*			
d	C	$\overline{T}$	$SD(T)$	$\overline{T}/k$	$\Pr(\mu \in I_T)$	$\overline{T}$	$SD(T)$	$\overline{T}/k$	$\Pr(\mu \in I_T)$
1	13.2	24.9	10.64	1.880	0.994	14.0	3.74	1.062	0.985
0.75	23.5	43.1	18.27	1.835	0.995	23.1	6.76	0.982	0.975
0.5	52.9	94.6	41.62	1.788	0.989	52.0	10.92	0.982	0.971
0.25	221.6	390.9	165.75	1.847	0.993	211.1	21.08	0.998	0.977
0.15	587.7	1060.1	460.51	1.804	0.987	588.2	34.17	1.001	0.985
0.05	5289.5	9590.7	4049.95	1.813	0.990	5288.9	101.79	1.000	0.980

From the results of the sequential method, we can see that the standard deviation of the r.v. T is clearly smaller (and so it is a more reliable estimate of the sample k). According to the table, we can verify all the asymptotic properties of the two methods. For Stein's procedure we see that the estimation of the sample size is always greater than k. Also, for all values of d the probability $p = \mathbb{P}(\mu \in I_T)$ is approximately 0.95. For the sequential method, the expected sample size is smaller than k for large values of d and asymptotically $\bar{T} \approx k$. The probability p is approaching the value 0.95 (Theorem 2.4.1 (ii)-(iv)), starting from smaller values than 0.95.

Generally we conclude that for the sequential method we need less observations, the method gives a more reliable estimate of k, but it does not have the property of exact consistency. On the contrary, Stein's method needs more observations for estimating k, but can be used to solve the fundamental problem of constructing a c.i. with constant width, for any value of the width.

For the spherical regions, we have similar results with the univariate case. Srivastava's method however, does not satisfy exactly the property of asymptotic consistency, for the given values of the radius of the region. Also, it is worthy of mentioning that Stein's method satisfies the property of asymptotic consistency in comparison with Healy's method.

Several mixtures of univariate methods have been created in order to optimize the speed of terminating the sequential procedure and also obtain asymptotic properties. For instance, the *accelerated* method and the *three stage procedure,* are based on an initial estimate of ρk, where $\rho \in (0,1)$. It can be proven that when ρ is near zero, the two methods will have the same performance as Stein's method, but for values near one they are similar to the purely sequential method.

We can verify (see Mukhopadhyay, 2009, pg.130-131) that the accelerated method is faster than the purely sequential method, keeps the asymptotic properties, but does not satisfy the exact consistency property. Similar results are true for the three stage procedure. In the next table we present the properties of the six methods where the second order asymptotic consistency is defined as

$$\lim_{d \to 0} \mathbb{E}(T - k) < \infty.$$

TABLE 2.6.3
Properties of sequential interval estimation methods

	Property			
	Exact	*Asymptotic*		
Method	*Consistency*	*Consistency*	*First Order Efficiency*	*Second Order Efficiency*
Stein	✓	✓	×	×
Purely Sequential	×	✓	✓	✓
Accelerated	×	✓	✓	✓
Three Stage	×	✓	✓	✓
Healy	✓	×	×	×
Srivastava	×	✓	✓	✓

We mention that the methods presented in this paragraph, can be expanded for finding an estimator of the mean of a normal distribution whose risk is bounded by a fixed upper bound. One will need to use a necessary and sufficient condition of the form (2.1.7).

CHAPTER 3

Sequential Probability Ratio Test

3.1 Introduction

Assume that $(X_n)_{n\geq 1}$ are i.i.d. r.v. $X_i \sim f(x|\theta)$ with mean value θ and variance $0 < \sigma < \infty$. Initially we will examine the simple hypothesis testing

$$\mathcal{H}_0 \colon \theta = \theta_0 \text{ against } \mathcal{H}_1 \colon \theta = \theta_1 \; (\theta_1 > \theta_0) \tag{3.1}$$

where $\theta_0, \theta_1 \in \Theta \subset \mathbb{R}$. The sample size, according to the classical theory of Neyman-Pearson, (1933) is constant and a priori known. However, we will now take the observations stepwise and each time we will choose to reject or accept $\mathcal{H}_0$. Denote the likelihood ratio

$$R_n := \frac{L(\theta_1|X_1, \dots, X_n)}{L(\theta_0|X_1, \dots, X_n)} = \prod_{i=1}^{n} \frac{f(X_i|\theta_1)}{f(X_i|\theta_0)}.$$

Large values of R_n will give evidence to reject $\mathcal{H}_0$, and small values to accept $\mathcal{H}_0$. It suffices therefore to consider two bounds which will define the termination rule. Let $A, B \in \mathbb{R}_+$ with $A < 1 < B$. The *sequential probability ratio test (SPRT)* is based on the following recursive process: if at the *n*-th step

$$A < R_n < B \tag{3.2}$$

then we take another observation X_{n+1}, calculate R_{n+1} and check again the condition (3.2). We will stop sampling if $R_n \leq A$, and accept $\mathcal{H}_0$, or stop if $R_n \geq B$ and reject $\mathcal{H}_0$. We can use a more convenient form instead of (3.2) which will enable us to use the results from the second chapter. Taking the logarithm at each part of (3.2), the following equivalent condition occurs

$$a < S_n < b \tag{3.3}$$

where $a = \log A$, $b = \log B$, $S_n = \sum_{i=1}^{n} Z_i$ and $Z_i = \log\left(\frac{f(X_i|\theta_1)}{f(X_i|\theta_0)}\right)$.

TABLE 3.1.1

n-th step of the SPRT

Condition	*Decision*
$S_n \le a$	*accept the null hypothesis*
$S_n \ge b$	*reject the null hypothesis*
$a < S_n < b$	*resample*

Typical questions that arise are the following: How can we compute the bounds A and B; Does SPRT terminate after a finite number of steps? What are the pros and cons of the SPRT in comparison with the uniformly most powerful Neyman-Pearson test? How can we generalize the existing theory to composite hypotheses testing?

3.2 Finite termination

Define the r.v.

$$T := \inf\{n \ge 1 : S_n \notin (a,b)\} \ , \ \inf\emptyset = \infty \tag{3.4}$$

T is a s.t. Indeed, if we consider the natural filtration of the r.v. $X_1, X_2, \ldots$ we get that:

$$\{T = n\} = \{S_1 \in (a,b)\} \cap \ldots \cap \{S_{n-1} \in (a,b)\} \cap \{S_n \notin (a,b)\} \in \mathcal{F}_n$$

since the r.v. S_i are $\mathcal{F}_i$-measurable. We can verify that the r.v. T is bounded a.s., i.e. $\mathbb{P}(T < \infty) = 1$, if we use the Central Limit Theorem:

$$\mathbb{P}(a < S_n < b) \approx \Phi\left(\frac{b - n\mu}{\sigma\sqrt{n}}\right) - \Phi\left(\frac{a - n\mu}{\sigma\sqrt{n}}\right)$$

Letting n approach ∞, the last quantity tends to zero. However, we could have stronger results for the r.v. T, taking into account the following theorem (see Govindarajulu, 1974, pg. 23-25).

Theorem 3.2.1. *If the i.i.d. r.v.* $(Z_n)_{n\ge1}$ *satisfy the condition* $\mathbb{P}(Z_i = 0) < 1$*, then there are constants* $q \in (0,1)$ *and* $c \in \mathbb{R}_+$ *such as* $\mathbb{P}(T > n) \le cq^n$*,* $\forall n \ge 1$.

Proof. Since $\mathbb{P}(Z_i = 0) < 1$, we assume without loss of generality, that $\mathbb{P}(Z_i > 0) > 0$ (similar results can be proved if $\mathbb{P}(Z_i < 0) > 0$).

$$\exists \varepsilon, \delta > 0 \text{ so that } \mathbb{P}(Z > \varepsilon) \geq \delta \tag{3.5}$$

Let $r \in \mathbb{N}$ be such that

$$r\varepsilon > \log(B/A) \tag{3.6}$$

and let $n = kr$, where $k \in \mathbb{N}$ is a constant. Define the sums

$$S_1 = \sum_{i=1}^{r} Z_i, \qquad S_2 = \sum_{i=r}^{2r} Z_i, \dots, \qquad S_k = \sum_{i=(k-1)r}^{kr} Z_i.$$

Since the r.v. Z_i are i.i.d., then S_i will also be i.i.d. and

$$\mathbb{P}\left(|S_1| > \log\left(\frac{B}{A}\right)\right) = \mathbb{P}\left(\left|\sum_{i=1}^{r} Z_i\right| > \log\left(\frac{B}{A}\right)\right) \geq \mathbb{P}\left(\sum_{i=1}^{r} Z_i > \log\left(\frac{B}{A}\right)\right)$$
$$\geq \mathbb{P}\left(\sum_{i=1}^{r} Z_i > r\varepsilon\right) \geq \prod_{i=1}^{r} \mathbb{P}(Z_i > \varepsilon) \geq \delta^r$$

where we used (3.6) for the second inequality. In order to terminate after kr steps, all the variables $|S_i|$ must be smaller than $\log(B/A)$. Hence

$$\mathbb{P}(T > kr) \leq \prod_{j=1}^{k} \mathbb{P}\left(|S_j| < \log\left(\frac{B}{A}\right)\right) \leq (1 - \delta^r)^k.$$

Setting $c = 1/(1 - \delta^r)$ and $q = (1 - \delta^r)^{1/r}$, the required inequality occurs after simple calculations. □

The above property shows that the r.v. T has a positive positively skewed distribution.

Corollary 3.2.1. *For the stopping time T we have that $\mathbb{E}(T) < \infty$.*

Proof. In the case of the SPRT, the event $\{f(X_i|\theta_1) = f(X_i|\theta_0)\}$ has probability zero, since $\theta_1 \neq \theta_0$. Therefore $\mathbb{P}(Z_i = 0) < 1$ and using Theorem 3.2.1:

$$\mathbb{E}(T) = \sum_{n=1}^{\infty} \mathbb{P}(T \geq n) = \mathbb{P}(T = 1) + \sum_{n=1}^{\infty} \mathbb{P}(T > n) \leq c_0 + c \sum_{n=1}^{\infty} q^n = \frac{c'}{(1-q)} < \infty. \quad \square$$

Since we proved $\mathbb{E}(T) < \infty$, it follows that $\mathbb{P}(T < \infty) = 1$. We mention that using the above inequality one can prove that all the moments of the r.v. T are finite.

3.3 Calculating bounds

Consider now the decision rule $d: \mathcal{X} \to \{0,1\}$, defined by the condition

$$d = \begin{cases} 0, & S_T \leq a \\ 1, & S_T \geq b \end{cases} \tag{3.7}$$

Then SPRT is uniquely defined by the pair (T, d). Clearly, when the test refers to fixed sample size n, then $T \equiv n$. Denote $\mathbb{P}_\theta(\cdot)$ the probability of an event, when the parameter of the p.d.f. is θ. For the type I and II errors we get that:

$$\alpha = \mathbb{P}(I) = \mathbb{P}_{\theta_0}(d = 1)$$

$$\beta = \mathbb{P}(II) = \mathbb{P}_{\theta_1}(d = 0)$$

We can find conditions for the values of A and B, following the fixed sample-size methodology.

Theorem 3.3.1. *For the bounds A and B of the SPRT*

$$A \geq \frac{\beta}{1-\alpha} \text{ and } B \leq \frac{1-\beta}{\alpha}$$

The numbers $\tilde{A} = \frac{\beta}{1-\alpha}$ *and* $\tilde{B} = \frac{1-\beta}{\alpha}$ *will be called Wald's approximations.*

Proof. Let $\mathcal{A}_n^i \subset \mathbb{R}^n$ be the set of all points for which we reject the hypothesis $\mathcal{H}_i$, where $i = 0,1$ at the *n*-th step. We can verify that $\forall n \neq m : \mathcal{A}_n^i \cap \mathcal{A}_m^i = \emptyset$. Hence

$$\begin{aligned} \mathbb{P}_{\theta_0}(d = 1) &= \sum_{1 \leq n < \infty} \mathbb{P}_{\theta_0}(d = 1 | T = n) \\ &= \sum_{1 \leq n < \infty} \int_{\mathcal{A}_n^0} L(\theta_0 | \mathbf{x}) d\mathbf{x} \leq B^{-1} \sum_{1 \leq n < \infty} \int_{\mathcal{A}_n^0} L(\theta_1 | \mathbf{x}) d\mathbf{x} = B^{-1}(1 - \beta) \end{aligned}$$

and $B \leq \frac{1-\beta}{\alpha}$.

Similarly

$$\begin{aligned} \mathbb{P}_{\theta_1}(d = 0) &= \sum_{1 \leq n < \infty} \mathbb{P}_{\theta_1}(d = 0 | T = n) \\ &= \sum_{1 \leq n < \infty} \int_{\mathcal{A}_n^1} L(\theta_1 | \mathbf{x}) d\mathbf{x} \leq A \sum_{1 \leq n < \infty} \int_{\mathcal{A}_n^1} L(\theta_0 | \mathbf{x}) d\mathbf{x} = A(1 - a) \end{aligned}$$

then $A \geq \frac{\beta}{1-\alpha}$. □

Wald's approximations do not guarantee that the type I and II errors will be equal to the given values α and β. If α' and β' are the error probabilities we get, then using Theorem 3.3.1

$$\begin{array}{c} \dfrac{\beta'}{(1-\alpha')} \le \dfrac{\beta}{(1-\alpha)} \\ \dfrac{1-\beta'}{\alpha'} \ge \dfrac{1-\beta}{\alpha} \end{array} \Rightarrow \begin{array}{c} (1-\alpha)\beta' \le \beta(1-\alpha') \\ (1-\beta)\alpha' \le (1-\beta')\alpha \end{array} \Rightarrow \alpha' + \beta' \le \alpha + \beta$$

This actually means that at least one of the inequalities $\alpha' \le \alpha$ or $\beta' \le \beta$ must be true. So from the application of SPRT, at most one error probability will be greater than the one we wanted. Also, we usually choose small enough values for α and β. Without loss of generality, we assume that $\alpha + \beta < 1$. This implies $B > 1 > A$ and we verify that the smaller the distance between θ_0 and θ_1, so the closer the likelihood ratio to 1, the more observations we will need to take a decision. We will see a similar result with the use of the Kullback-Leibler divergence (see Nowak, 2010, pg. 1-3).

Example 3.3.1. Let $(X_n)_{n\ge1}$ be i.i.d. r.v. for which $X_i \sim N(\theta, \sigma^2)$, the mean value θ is unknown and the variance σ^2 is a known constant. The p.d.f. is

$$f(x|\theta,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x-\theta)^2\right\}, -\infty < x < +\infty$$

We compute that $Z_i = \log\left[\frac{f(X|\theta_1,\sigma^2)}{f(X|\theta_0,\sigma^2)}\right] = \frac{\theta_1-\theta_0}{\sigma^2}\left(X - \frac{\theta_1+\theta_0}{2}\right)$.

In order to test hypothesis (3.1), under the condition (3.3) we get that

$$C_A + nD < \sum_{i=1}^{n} X_i < C_B + nD$$

where $C_A = (\theta_1 - \theta_0)^{-1}\sigma^2 \log(A)$, $C_B = (\theta_1 - \theta_0)^{-1}\sigma^2 \log(B)$ and $D = (\theta_0 + \theta_1)/2$

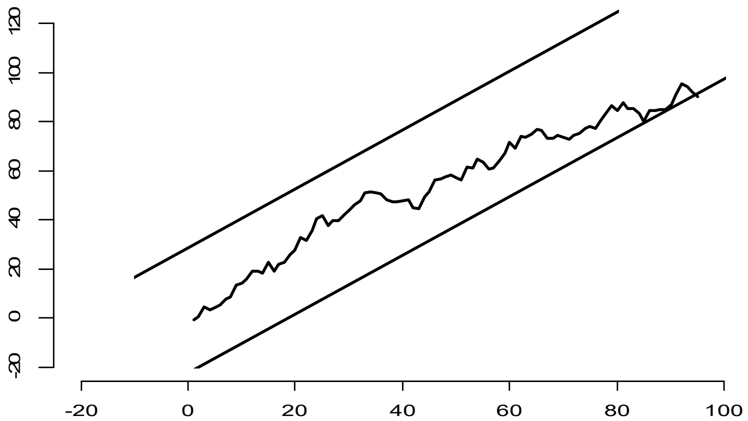


Figure 3.3.1 *Simulation under* $\mathcal{H}_0$ *for* $\theta_0 = 1$, $\theta_1 = 1.4$, $\sigma = 1$, $\alpha = 0.05$ *and* $\beta = 0.1$.

According to the figure, we will make a decision for or against $\mathcal{H}_0$ when the random walk $Y_n = \sum_{i=1}^{n} X_i$ crosses one of the two parallel lines $\varepsilon_A = C_A + nD$ or $\varepsilon_B = C_B + nD$. To construct Figure 3.3.1 (see Appendix III) we set $A = \frac{\beta}{1-\alpha}$ and $= \frac{1-\beta}{\alpha}$ **.** Initially we calculate C_A, C_B and D for the given values of θ_1 and θ_0. While the sum Y_n satisfies the condition $C_A + nD < Y_n < C_B + nD$, we generate a random number from $N(\theta_0, \sigma)$, compute the new sum Y_{n+1} and check whether or not the above inequality is true for $n+1$. In that way, we draw the values Y_n that lie between the two lines, until a point comes out of the two parallel lines.

3.4 Optimal property

The following theorem was proved by Wald and Wolfowitz (1948) and is characterized as the «optimal property of the SPRT» because it guarantees that the SPRT is faster than the classical Neyman-Pearson test.

Theorem 3.4.1. *Suppose that we want to do a test of the form (3.1) and let* (T, d) *be the SPRT as it was defined in (3.3), with* $\mathbb{P}_{\theta_0}(d = 1) = \alpha$*,* $\mathbb{P}_{\theta_1}(d = 0) = \beta$ *and* $\alpha + \beta < 1$*. If* (T', d') *is another sequential test that satisfies*

(i) $\mathbb{E}_{\theta_0}(T'), \mathbb{E}_{\theta_1}(T') < \infty$

(ii) $\mathbb{P}_{\theta_0}(d' = 1) \leq \alpha$ *and* $\mathbb{P}_{\theta_1}(d' = 0) \leq \beta$

then $\mathbb{E}_{\theta_0}(T') \geq \mathbb{E}_{\theta_0}(T)$ and $\mathbb{E}_{\theta_1}(T') \geq \mathbb{E}_{\theta_1}(T)$.

The proof of the theorem is based on statistical decision theory and also o.s.t.

Assume that θ is a r.v. that takes two values θ_0, θ_1 with a priori distribution $\pi = \mathbb{P}(\theta = \theta_0)$ and $\mathbb{P}(\theta = \theta_1) = 1 - \pi$. Every observation we take will have a cost $c > 0$, the false rejection of $\mathcal{H}_0$ will have cost $c_0 > 0$ and the false rejection of $\mathcal{H}_1$ will have cost $c_1 > 0$. The loss function (see Shiryaev, 2007, pg 165) will then be

$$L(\theta, d) = \begin{cases} c_0, & (\theta = \theta_{0,} d = 1) \\ c_1, & (\theta = \theta_{1,} d = 0) \\ 0, & (\theta = \theta_i, d = i, i = 0,1) \end{cases}$$

The expectation of L is:

$$\mathbb{E}(L(\theta, d)) = c_0 \mathbb{P}(\{\theta = \theta_0\} \cap \{d = 1\}) + c_1 \mathbb{P}(\{\theta = \theta_1\} \cap \{d = 0\})$$

$$= c_0 \pi \, \mathbb{P}(d = 1 | \theta = \theta_0) + c_1 (1 - \pi) \, \mathbb{P}(d = 0 | \theta = \theta_1)$$

according to the Bayes formula. The risk of the decision rule δ will be equal to

$$\rho_\pi(\delta) = \mathbb{E}_\pi(L(\theta, d)) + c\mathbb{E}_\pi(T)$$

Denote $\pi_n = \mathbb{P}(\theta = \theta_0 | X_1, \dots, X_n)$ and $\Delta := \{\delta = (T, d) : T \in \mathcal{T}_n \text{ and } d \in \mathcal{D}\}$, where $\mathcal{T}_n$ is the set of all regular stopping times with respect to the filtration $\mathcal{F}_n$ and $\mathcal{D}$ is the set of the decision rules. It can be proved (Shiryaev, 2007, pg 165-170) that the rule δ which minimizes the risk $\rho_{\pi_n}(\delta)$, consists of the stopping rule of the form (3.7), and so it is the SPRT.

3.5 OC and ASN functions

Two basic measures have been created to evaluate the performance of the sequential tests. The first one is described by a function showing the reliability of the test, similar to the power function of the fixed sample-size tests. The second one refers to the speed of a test, which is defined by the expected number of observations needed to take a decision. Thus, we create two curves that will accompany every sequential procedure (T, d).

Definition 3.5.1. *We define the operating characteristic curve (OC) as the function* $Q(\theta) = \mathbb{P}_\theta(d = 0)$.

Clearly $Q(\theta_0) = 1 - \alpha$, $Q(\theta_1) = \beta$ and $Q(\theta) = 1 - \pi(\theta)$, where $\pi(\theta)$ is the power function. We will examine the case of simple hypotheses testing. In order to find an analytical expression of the function Q, we will use Theorem (1.5.1). If the r.v. $Z_i = \log\frac{f(X_i|\theta_1)}{f(X_i|\theta_0)}$ satisfies the four conditions of Proposition (1.5.1), then $\exists\ t_0 \neq 0$ such as $\mathbb{E}_\theta(e^{-t_0 S_T}) = 1$. Using this relation and due to the uniqueness of t_0, we verify that it is a function of θ. Assuming that the r.v. S_T does not exceed too much the bounds α and β, i.e. $S_T \approx a$ or $S_T \approx b$

$$\begin{aligned} 1 = \mathbb{E}_\theta(e^{-t_0 S_T}) &= \mathbb{E}_\theta(e^{-t_0 S_T} | S_T \leq a) + \mathbb{E}_\theta(e^{-t_0 S_T} | S_T \geq b) \\ &\approx e^{-t_0 a}\mathbb{P}_\theta(S_T \leq a) + e^{-t_0 b}\mathbb{P}_\theta(S_T \geq b) \\ &= e^{-t_0 a}Q(\theta) + e^{-t_0 b}\big(1 - Q(\theta)\big) \end{aligned}$$

and so $Q(\theta) \approx \frac{e^{-t_0 b} - 1}{e^{-t_0 b} - e^{-t_0 a}}$, for $t_0 \neq 0$. We extend the function Q to the whole $\mathbb{R}$, in order to be continuous:

$$Q(\theta) \approx \begin{cases} \dfrac{e^{-t_0(\theta)b} - 1}{e^{-t_0(\theta)b} - e^{-t_0(\theta)a}}, & t_0(\theta) \in \mathbb{R} \setminus \{0\} \\ \dfrac{b}{b - a}, & t_0(\theta) = 0 \end{cases} \tag{3.8}$$

Similarly, we extend the function $t_0 = t_0(\theta)$ (t_0 can be solved with respect to θ, e.g. see formula 3.12) to $\mathbb{R}$, setting its value equal to its limit, when $t = 0$.

Definition 3.5.2. *We define as average sample number (ASN) the expectation* $\mathbb{E}_\theta(T)$ *as a function of* θ.

Since T is s.t. with $(T) < \infty$, there are two cases:

(i) If $\mathbb{E}_\theta(Z_1) \neq 0$, applying Wald's first equation

$$\mathbb{E}_\theta(T) = \frac{\mathbb{E}_\theta(S_T)}{\mathbb{E}_\theta(Z_1)} \approx \frac{aQ(\theta) + b\big(1 - Q(\theta)\big)}{\mathbb{E}_\theta(Z_1)} \tag{3.9}$$

(ii) If $\mathbb{E}_\theta(Z_1) = 0$, applying Wald's second equation

$$\mathbb{E}_\theta(T) = \frac{\mathbb{E}_\theta(S_T^2)}{\mathbb{E}_\theta(Z_1^2)} \approx \frac{a^2 Q(\theta) + b^2\big(1 - Q(\theta)\big)}{\mathbb{E}_\theta(Z_1^2)} \tag{3.10}$$

The function ASN will be continuous in $[\theta_0, \theta_1]$ and will have at least one root $\tilde{\theta} \in (\theta_0, \theta_1)$, according to the properties of the K-L divergence (see Appendix I).

Even if for the simple hypotheses testing we used only two points of the parametric space Θ, we will construct the functions OC and ASN for various values of the parametric space (see Table 3.6.2), since there are sequential composite tests (see Wald, 1947) which can be induced, using proper transformations, to tests of the form (3.1). Many of these tests fall into the group of *sequential invariant tests* (see Baseville, 1993, pg. 148). However, we will not examine these tests since

(i) the sequential composite tests require a priori knowledge of the data

(ii) they do not have any optimal property to compare them with equivalent fixed sample-size tests

(iii) the transformations do not sustain the independence of the variables

(iv) they are computationally complex

3.6 Comparing methods via simulation

In this chapter we will examine the hypothesis testing (3.1) using simulated data when $\theta_0 = 1, \theta_1 = 1.4, \sigma = 2$ and $\alpha = 0.05$ (also $\beta = 0.1$ in Paragraphs 3.6.2 and 3.6.3).

3.6.1 Comparing SPRT with the Neyman-Pearson test

In order to compare two sequential tests, we will use SPRT's optimal property. If for a given sample size n and type I error, we find the type II error of the Neyman-Pearson test, then using SPRT with parameters these two errors, the expected sample sizes should be smaller than n.

Assume that we take a sample of size n from the normal distribution $N(\theta, \sigma^2)$. Hence for testing hypothesis (3.1), the uniformly most powerful Neyman-Pearson test will have as rejection region (e.g. see Hogg, 1970, pg. 243)

$$R = \{\mathbf{X} \in \mathcal{X}: T(\mathbf{X}) > z_a\}$$

where $T(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \theta_0)}{\sigma}$. The type II error will be given by the formula

$$\mathbb{P}(II) = \Phi\left(\frac{\sqrt{n}(\theta_0 - \theta_1)}{\sigma} + z_\alpha\right)$$

from which we get the value of β.

To illustrate the above results (see Appendix IV, Program 5) we apply the following: For the estimation of the quantities $\mathbb{E}_{\theta_0}(T)$, $\mathbb{P}(II)$ and $\mathbb{E}_{\theta_1}(T)$, $\mathbb{P}(I)$ of the SPRT we use two loops (using the while command) where we examine if the sum S_n exceeds the values of a and b (see Table 3.6.1), generating random numbers from $N(\theta_0, \sigma^2)$ and $N(\theta_1, \sigma^2)$ respectively. Outside of these loops we check if there is a false rejection or acceptance of the null hypothesis, increasing the values of the counter variables. After the termination of the while loops, we get one value for T_{θ_0} and T_{θ_1}. Putting all these inside a loop of $N = 10^4$ repetitions, we store the results in a row of a matrix named *matrixSPRT*. However, these will be valid for particular values of α and β. As mentioned above, for given values of α and n (n has 7 values) we compute $\beta = \mathbb{P}(II)$, and with these α and β calculate each time the bounds a and b for the SPRT.

TABLE 3.6.1

Comparing SPRT with the Neyman-Pearson test

	Neyman-Pearson					*SPRT*				
n	α	β	$\alpha+\beta$	$\overline{T}_{\theta_0}$	$SE(\overline{T}_{\theta_0})$	$\overline{T}_{\theta_1}$	$SE(\overline{T}_{\theta_1})$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}+\hat{\beta}$
30	0,05	0,709	0,759	14,04	0,0019	22,74	0,0024	0,0565	0,625	0,681
50	0,05	0,591	0,641	22,58	0,0029	37,97	0,0036	0,0511	0,520	0,571
100	0,05	0,361	0,411	45,03	0,0046	72,92	0,0056	0,0479	0,319	0,367
150	0,05	0,211	0,261	71,05	0,0064	101,80	0,0073	0,0475	0,192	0,239
200	0,05	0,118	0,168	99,07	0,0081	122,36	0,0085	0,0479	0,106	0,154
250	0,05	0,065	0,115	127,82	0,0094	134,10	0,0094	0,0445	0,056	0,100
300	0,05	0,034	0,084	157,82	0,0107	144,75	0,0107	0,0469	0,029	0,076

We can see that for all values of n, the quantities $\overline{T}_{\theta_0}$ and $\overline{T}_{\theta_1}$ are smaller than n. Also, at each case, the sum $\hat{\alpha} + \hat{\beta}$ is smaller than $\alpha + \beta$, which verifies the theoretical results of Paragraph 3.3. Finally, from the output of the R software, we get the following two histograms for $\alpha = 0.05, \beta$ =0.034 and $n = 10^4$

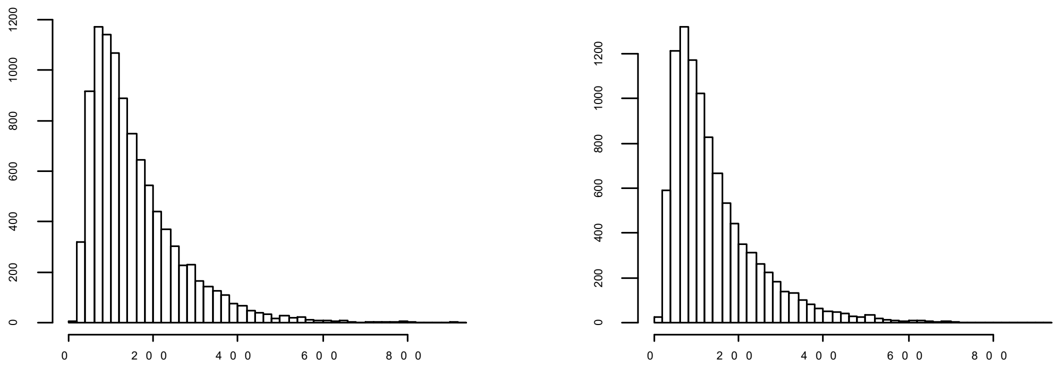


Figure 3.6.1 *Histograms of the r.v. T under $\mathcal{H}_0$ and $\mathcal{H}_1$ respectively*

The shape of the histograms complies with the property of Theorem 3.2.1.

We will give two characteristic examples, appearing in the bibliography, so as to compare the functions OC and ASN.

3.6.2 Case of independent variables

Let $(X_n)_{n\geq 1}$ be i.i.d. r.v. with $X_i \sim N(\theta, \sigma^2)$. We will find the functions OC and ASN, for given error probabilities α and β, to test hypothesis (3.1). From Example (3.3.1), we get

$$Z_i = \log\frac{f(X_i|\theta_1)}{f(X_i|\theta_0)} = -\frac{(X_i-\theta_1)^2}{2\sigma^2} + \frac{(X_i-\theta_0)^2}{2\sigma^2} = \frac{\theta_0-\theta_1}{\sigma^2}\left(X_i - \frac{\theta_0+\theta_1}{2}\right)$$

and taking the expectation

$$\mathbb{E}_\theta(Z) = \frac{\theta_1-\theta_0}{\sigma^2}\left(\theta - \frac{\theta_0+\theta_1}{2}\right) \tag{3.11}$$

$$\mathbb{E}_\theta(Z^2) = k^2(\sigma^2+\theta^2) + k^2\tilde{\theta}(\tilde{\theta}-2\theta)$$

where $k = (\theta_1-\theta_0)/\sigma^2$ and $\tilde{\theta} = (\theta_1+\theta_0)/2$. It is true that $0 < F_X(\tilde{\theta}) < 1$, where F_X is the c.d.f. of X and so $\mathbb{P}(Z<0) > 0$ and $\mathbb{P}(Z>0)>0$. Clearly, $\theta_0 < \tilde{\theta} < \theta_1$, which can be verified from the properties of the K-L divergence. The m.g.f. $M_Z(t)$ is well defined since

$$M_Z(t) = e^{-k\tilde{\theta}t}M_X(kt) < \infty.$$

We have proved so far that the r.v. Z satisfies the conditions of Proposition (1.5.1), when the parameter θ of the normal distribution differs from $\tilde{\theta}$. This implies that there is a unique $t_0^* \neq 0$ such as $M_Z(t_0^*) = 1$ or $M_Z(-t_0) = 1$, where $t_0 = -t_0^*$. In order to apply the fundamental equation, it suffices to show that Z satisfies condition (ii) of Theorem (1.5.1). From the definition of the r.v. T

$$\forall n < T: a < S_n < b \Rightarrow |S_n| \leq \max\{|a|, |b|\}$$

and thus

$$\mathbb{E}(e^{-t_0 Z}) = 1 \Rightarrow t_0(\theta) = \frac{2}{\theta_1 - \theta_0}\left(\theta - \frac{\theta_0 + \theta_1}{2}\right) \tag{3.12}$$

where we used the m.g.f. of X, which is $M_X(t) = \exp\{\theta\tau + \sigma^2 t^2/2\}$. When $\theta = \tilde{\theta}$, we set $t_0(\tilde{\theta}) = 0$.

We note that in order to solve the equation $(e^{-t_0 Z}) = 1$, in the general case where the relation between t_0 and θ is not linear, one does not have to use arithmetic methods because of the uniqueness of t_0. We generate values of t_0 in a set of the form $[-l_1, 0)\cup(0, l_2]$, where $l_1, l_2 > 0$. Then we get values of θ, and substituting to relation (3.8), we find values of $Q(\theta)$ and thus $\mathbb{E}_\theta(T)$ from (3.9).

3.6.3 Case of correlated variables

This case appears in control charts for correlated r.v. However, we will see in the particular case (AR(1) model) that the logarithm of likelihood can be presented in a simple manner. Let $(X_n)_{n\geq 1}$ be a Gaussian AR(1) process with $\mathbb{E}(X_n) = \theta$ and $\mathrm{var}(X_n) = \sigma^2$, so X_n will be given by the formula $X_{n+1} = \rho X_n + \gamma + \varepsilon_n$, where $X_0 = 0$, $|\rho| < 1$, $\gamma \in \mathbb{R}$, $\varepsilon_n \sim N(0, \sigma_\varepsilon^2)$ and the distribution of $X_1, X_2, \dots, X_n$ is the multivariate normal distribution

$$f(\mathbf{x}|\boldsymbol{\theta}, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\theta})^t \Sigma^{-1}(\mathbf{x} - \boldsymbol{\theta})\right\}$$

where we denoted $\mathbf{x} = (x_1, \dots, x_n)^t$ and $\boldsymbol{\theta} = (\theta, \dots, \theta)^t$. We can prove that $\gamma = (1-\rho)\theta$, $\sigma_\varepsilon^2 = (1-\rho^2)\sigma^2$ and that the variance-covariance matrix is

$$\Sigma = \begin{bmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ \rho & 1 & \cdots & \rho^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \cdots & 1 \end{bmatrix}$$

We will test hypothesis (3.1). In order to apply the SPRT we calculate the sum

$$\begin{aligned} S_n = \log(R_n) &= \log\left(\frac{f(X_1, \dots, X_n|\theta_1)}{f(X_1, \dots, X_n|\theta_0)}\right) \\ &= \log\left(\frac{f(X_1, \dots, X_{n-1}|\theta_1) f(X_n|X_1, \dots, X_{n-1}; \theta_1)}{f(X_1, \dots, X_{n-1}|\theta_0) f(X_n|X_1, \dots, X_{n-1}; \theta_0)}\right) \end{aligned}$$

$$= \log\left(\frac{f(X_1,\dots,X_{n-1}|\theta_1)}{f(X_1,\dots,X_{n-1}|\theta_0)}\right) + \log\left(\frac{f(X_n|X_1,\dots,X_{n-1};\theta_1)}{f(X_n|X_1,\dots,X_{n-1};\theta_0)}\right)$$

$$= S_{n-1} + \lambda_n \tag{3.13}$$

It suffices to find λ_n. If $\mathbf{X_1} = (X_1,\dots,X_{n-1})^t$ and $\mathbf{X_2} = X_n$, then for the conditional distribution $\mathbf{X_2}|\mathbf{X_1}$ we know that (see Johnson, 2007, pg. 156-163)

$$\mathbf{X_2}|\mathbf{X_1} \sim N(\theta + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{x_1} - \boldsymbol{\theta_1}), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$$

where Σ_{ij} are given by $\Sigma_{12} = \sigma^2[\rho^{n-1} \quad \rho^{n-2} \quad \cdots \quad \rho]^{\,t}$, $\Sigma_{21} = \sigma^2[\rho^{n-1} \quad \rho^{n-2} \quad \cdots \quad \rho]$,

$$\Sigma_{11} = \begin{bmatrix} 1 & \rho & \cdots & \rho^{n-2} \\ \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}, \ \Sigma_{11}^{-1} = \frac{1}{(1-\rho^2)\sigma^2}\begin{bmatrix} 1 & -\rho & \cdots & 0 \\ -\rho & 1+\rho^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix},$$

$\boldsymbol{\theta_1}$ is the mean vector of $\mathbf{X_1}$ and $\mathbf{x_1}$ is its realization. After calculations we get

$$\Sigma_{21}\Sigma_{11}^{-1} = [0,0,\dots,\rho]$$

and

$$\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} = [0,0,\dots,\rho] \cdot (\sigma^2[\rho^{n-1},\rho^{n-2},\dots,\rho]^t) = \sigma^2\rho^2$$

which implies

$$X_n|X_1,\dots,X_{n-1} \sim N(\theta + \rho(X_{n-1} - \theta), \sigma^2(1-\rho^2))$$

From (3.13):

$$\lambda_n = -\frac{1}{2(1-\rho^2)\sigma^2}[(X_n - \theta_1 - \rho(X_{n-1} - \theta_1))^2 - (X_n - \theta_0 - \rho(X_{n-1} - \theta_0))^2]$$

$$= \frac{(\theta_1 - \theta_0)}{(1+\rho)\sigma^2}\left[\left((1-\rho)\theta + \varepsilon_n\right) - (1-\rho)\frac{(\theta_0 + \theta_1)}{2}\right] \tag{3.14}$$

3.6.4 Comparing OC and ASN functions

We will now compare Wald's approximation for calculating the ASN and OC functions, with simulated values. Also, we will compute the values of the functions for the AR(1) model, for three values of ρ (see Appendix IV).

We will follow the next steps (see Appendix IV, Program 6 and 7) in order to create proper algorithms: As far as the calculation of the ASN and OC is concerned, using Wald's approximation, we create a loop of 11 repetitions for various values of θ, when $\alpha = 0.05, \beta =$

0.1, $\theta_0 = 1$, $\theta_1 = 1.4$, $\sigma = 2$, $a = \log\left(\frac{\beta}{1-\alpha}\right)$, $b = \log\left(\frac{1-\beta}{\alpha}\right)$. At each repetition we find $t_0(\theta)$ from formula (3.12), then $Q(\theta)$ from (3.8) and finally ASN from formulas (3.9) and (3.10), checking if $\theta = \frac{\theta_0+\theta_1}{2}$. These values will be stored in the matrix *A.wald*. To find simulated values, inside a while loop we examine if the sum S_n exceeds the bounds a and b (see Table 3.1.1). The counting variable *count* will be used to estimate ASN whereas the counter q to estimate OC. These repetitions will be done 11 times, saving all the values in the matrix *A.sim* that has 11 rows.

For the AR(1) model (see Appendix IV, Program 7), the methodology is used in similar manner. Inside a while loop, we check when the sum S_n exceeds the interval (a, b). The values of this sum occur by generating random numbers from $N(0, \sigma_\varepsilon^2)$ and substituting to formula (3.14). Finally, all the above are enclosed in a loop for three values of ρ.

TABLE 3.6.2
Comparing OC and ASN values via Wald's approximation and Monte Carlo simulation

		Independent random variables				*AR(1) Model*					
		Wald's approximation		*Simulation*		$\rho = 0.1$		$\rho = 0.5$		$\rho = 0.9$	
t_0	θ	*OC*	*ASN*	*OC*	*ASN*	*OC*	*ASN*	*OC*	*ASN*	*OC*	*ASN*
-1.0	1.00	0.950	99.71	0.963	105.59	0.957	131.018	0.957	314.391	0.945	1961.19
-0.8	1.04	0.916	113.69	0.911	121.19	0.911	145.756	0.930	346.230	0.912	2236.76
-0.6	1.08	0.863	128.87	0.877	141.02	0.872	171.881	0.841	401.662	0.877	2438.38
-0.4	1.12	0.786	143.74	0.784	154.93	0.806	186.359	0.789	466.843	0.799	2707.46
-0.2	1.16	0.683	155.88	0.694	170.70	0.689	205.640	0.700	484.150	0.688	2943.03
0.0	1.20	0.562	162.68	0.567	171.77	0.530	213.436	0.547	529.469	0.549	3154.75
0.2	1.24	0.436	162.60	0.397	177.08	0.422	218.780	0.413	528.497	0.436	3093.65
0.4	1.28	0.319	156.07	0.303	184.22	0.326	206.288	0.319	493.541	0.323	3089.11
0.6	1.32	0.224	145.09	0.220	169.72	0.233	189.738	0.210	450.108	0.211	2715.06
0.8	1.36	0.151	132.04	0.149	154.95	0.131	168.969	0.159	401.801	0.142	2524.26
1.0	1.40	0.100	118.81	0.104	126.14	0.091	154.751	0.074	372.405	0.095	2220.86

According to the values for the independent variables case, the existence of differences is due to the error of Wald's approximation for the r.v. S_T ($S_T \approx a \ or \ S_T \approx b$). Actually Wald did not consider the quantities $|S_T - a|$ and $|S_T - b|$ but assumed that the boundaries were absorbing.

We verify that when the values of ρ are far from 0.1, then we will need more observations to decide for or against $\mathcal{H}_0$. This implies that the type of correlation between the r.v. will play an important role in the computation of ASN. This can be proven analytically taking the expectation in formula (3.14)

$$\mathbb{E}(\lambda_n) = \frac{(1-\rho)}{(1+\rho)} \frac{(\theta_1 - \theta_0)}{\sigma^2} \left(\theta - \frac{\theta_0 + \theta_1}{2}\right) \tag{3.15}$$

and letting ρ approach 1 or -1. Then the increment λ_n will become smaller and we will need more observations or become larger and need less observations. Also, it can be proven (Baseville, 1993, pg. 142-143) that the function ASN of the AR(1) model will be the product of $(1+\rho)/(1-\rho)$ with the equivalent function ASN of the i.i.d. case.

CHAPTER 4

Cumulative Sum Algorithm

4.1 Overview of statistical detection algorithms

Change detection algorithms were initially used in statistical quality control. According to Ghosh (1991), Shewhart (1931) was the first to develop a sequential algorithm known as *Shewhart's control chart*. Generally, control charts are used for the surveillance of procedures with respect to a characteristic. Usually, the quantity examined is the mean value of the distribution of a characteristic. A statistical function is viewed in a control chart, along with two parallel lines, indicating that if the value of the function exceeds these lines, then there is evidence for a mean shift. Many change detection algorithms were created to deal with the detection of mean, variance or both. They were used in sectors beyond quality control, such as statistical signal processing, bio-surveillance, monitoring navigation systems, e.t.c. We need to explain though some basic differences. There are two types of statistical detection. In the first case we have gathered all the data and then search for any change in distribution of the characteristic. This procedure is called *offline.* In the second case, we collect the observations one by one and detect each time for a change. This is called *online* detection and is one of the celebrated applications of sequential statistics. Active research areas deal with the problem of *quickest change detection*, where we search for the statistical functions and the decision rules which satisfy certain optimality conditions. The purpose is to detect abrupt changes in the mean, as soon as possible.

To describe the whole process explicitly, we will borrow some basic notions from statistical quality control. Initially, we examine if there is a mean shift. We set a threshold (or two, if we want to check any decrease in mean) and when the value of the statistical function exceeds this threshold, we then have evidence of an "alarm" that the value has changed according to the classical terminology (e.g. see Antzoulakos, 2012). We note that there is also the case were we have an alarm but no actual change in mean happened. This indicates that we have a *false alarm* (or type II error). Also, it is possible that the mean has changed and there is no alarm (type II error). In that case, probably the algorithm will detect later on the change but there is going to be a delay in the change detection. For instance, let us see the following diagram.

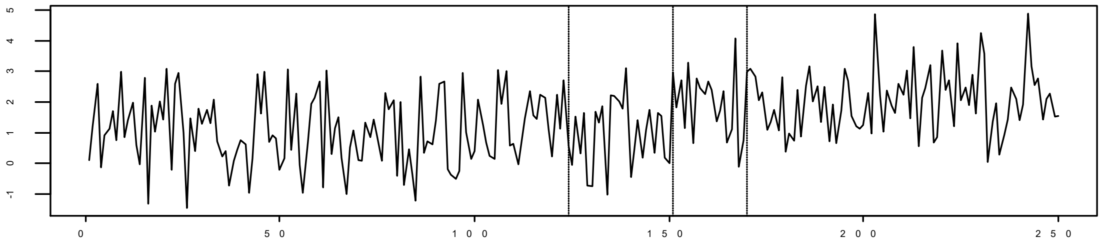


Figure 4.1 *Change in the mean of the distribution*

The first 150 points were generated from $N(1,1)$ whereas the rest 100 from $N(2,1)$. If the parameters of the algorithm were tuned so that the algorithm detects a change as soon as possible, we could then have an indication of an alarm for mean shift, at the point from which the first line passes through, instead of the second line (namely, since there is no change in distribution, there will be a false alarm). If now the algorithm was tuned to have a small number of false alarms, then it would not detect a change quickly, resulting in a delay (e.g. third line). The purpose of sequential analysis is to minimize the *mean detection delay* while the *mean time between false alarms* is constant. We will see in the following paragraphs that the whole martingale theory and the SPRT, can be used to create the CUSUM algorithm. CUSUM is widely used in many sciences and can be found as an option in statistical packages such as Minitab and Statgraphics.

4.2 CUSUM algorithm

4.2.1 Intuitive approach

The *cumulative sum (CUSUM*) algorithm was proposed by Page (1954) and was actually announced seven years later after Wald's (1947) contribution on SPRT. CUSUM, in comparison with Shewhart's control chart, takes into account the previous observations using a sum, as we will see later on.

Our experiment will have the following form: Each time we collect a sample of size one from the normal distribution $N(\theta,\sigma^2)$ and test hypothesis (3.1). Assume that θ_0, θ_1 and the variance σ^2 are known. We set

$$S_n=\sum_{i=1}^{n}(X_i-\theta_0).$$

While $\mathcal{H}_0$ is true, we expect that the points (n,S_n) will be placed near the xx' axis, since $\mathbb{E}(S_n)=0$. If the mean value is $\theta_1>\theta_0$, at the point with abscissa τ_0

$$\mathbb{E}(S_n)=\mathbb{E}\left(\sum_{i=1}^{\tau_0-1}(X_i-\theta_0)+\sum_{i=\tau_0}^{n}(X_i-\theta_0)\right)=(\theta_1-\theta_0)(n-\tau_0)$$

This implies that the points (n, S_n) for which $n > \tau_0$ will be near the line with slope $\lambda = (\theta_1 - \theta_0)$. We note that the point τ_0 will be called *change point*.

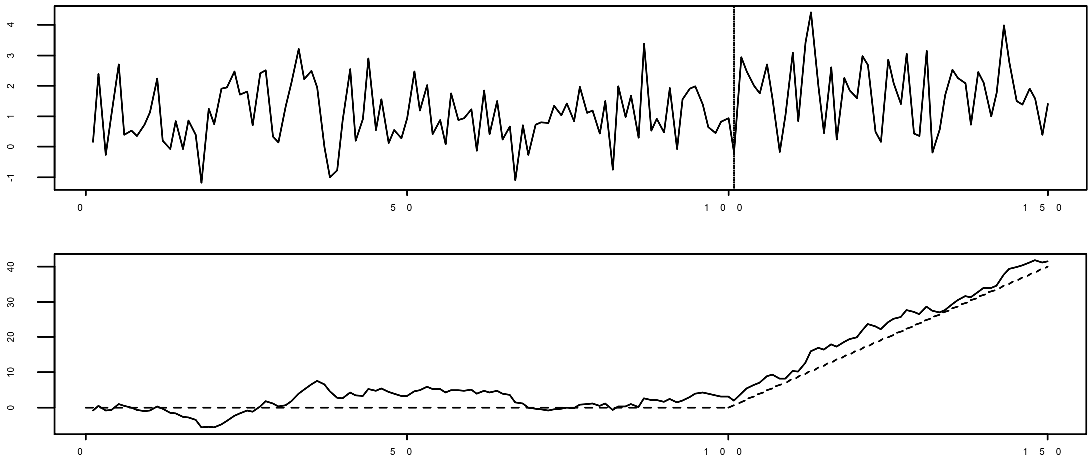


Figure 4.2.1 (n, S_n) *points when there is mean shift at the change point* $\tau_0 = 100$.

Let us now see how we can apply similar examples with the use of the likelihood ratio R_n. Denote as usual with $Z = \log\left(\frac{f(X|\theta_1)}{f(X|\theta_0)}\right)$. According to the properties of the K-L divergence (see Appendix (I)) since

$$\mathbb{E}_{\theta_0}(Z) < 0 \text{ και } \mathbb{E}_{\theta_1}(Z) > 0$$

we may expect, that after a mean shift of the distribution f, the sum $S_n = \log(R_n) = \sum_{i=1}^{n} Z_i$ would start to increase. Let τ_0 be the change point we observe, using the statistical function. From Example (3.3.1) we have that $Z_i = \frac{\theta_1 - \theta_0}{\sigma^2}\left(X_i - \frac{\theta_1 + \theta_0}{2}\right)$.

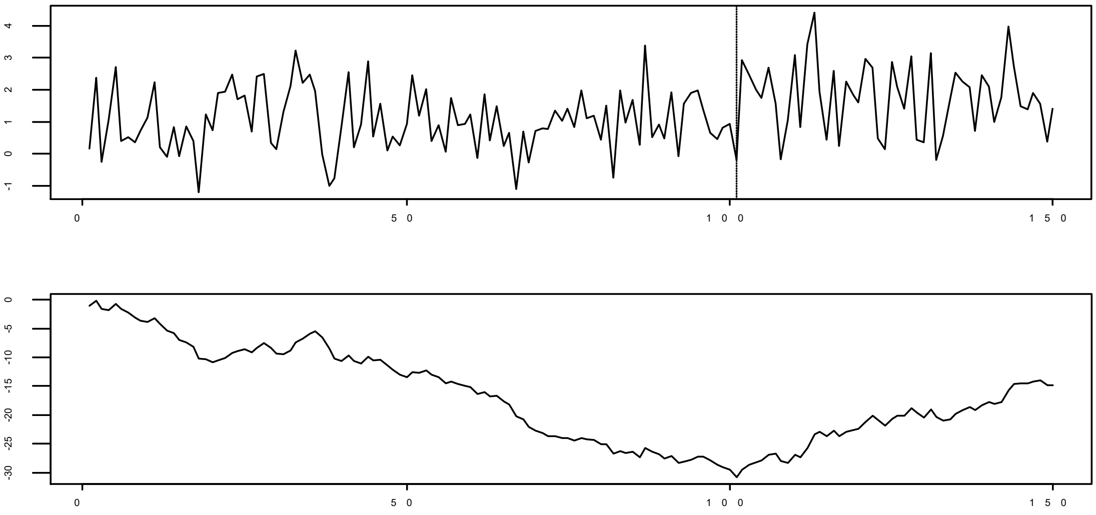


Figure 4.2.2 (n, S_n) *points when there is a mean shift, using the log-likelihood ratio*

Page (1954) proposed the use of a threshold $h > 0$, which will be set by the experimenter, in a way that each time the quantity

$$G_n = S_n - \min_{1 \le k < n} S_k$$

becomes greater or equal to h, we will have an alarm for a mean shift. We denote the stopping time

$$T = \inf\{n \in \mathbb{N}: G_n \geq h\} \ , \ \inf\emptyset = \infty.$$

To illustrate the algorithm using the R software, we set $h = 7$ (the way we compute the constant h, will be explained in Paragraph 4.3.3). The first 100 points were generated from $N(\theta_0, \sigma^2)$, with $\theta_0 = 1$ and $\sigma = 1$, whereas the remaining 50 from $N(\theta_1, \sigma^2)$ with $\theta_1 = 1.8$. We calculate the sum $S_n = \sum_{i=1}^{n} Z_i$, and store its values in matrix S and the values $S_n - \min_{1 \le k < n} S_k$ in a matrix G. Clearly, we will get an estimate of T when an element of G exceeds for the first time the value $h = 7$. We store all the values i for which $G[i] > h$ in a matrix and find its minimum element, which will be T.

```
 # R code (CUSUM algorithm)
 set.seed(5);theta0 <-1 ; theta1 <-1.8;sigma<-1
 S<-matrix();X1<-rnorm(100,mean=theta0,sd=1);X2<-rnorm(50,mean=theta1,sd=1)
 X<-c(X1,X2)
 S[1] <- (theta1-theta0)/sigma^2*(X[1]-(theta1+theta0)/2)
 for(i in 1:99) {
 S[i+1]<-S[i]+(theta1-theta0)/sigma^2*(X[i+1]-(theta1+theta0)/2)
 }
 for(i in 100:149) {
 S[i+1] <- S[i]+(theta1-theta0)/sigma^2*(X[i+1]-(theta1+theta0)/2)
 }
 par(mar=c(3,3,2,2));index<-1:150
 plot(index,X, type="l", lwd=1, pch=16,ylab="",xlab="")
 abline(v=101,lty=3)
 h<-7;k<-1;m<-c();G<-c();G[1]<-0
 for(i in 2:length(S)) {
  G[i]<-S[i]-min(S[seq(1,i-1,1)])
  if(G[i]>h) {
  m[k]<-i
  k<-k+1
  }
 }
 plot(G,type="l");abline(h=h,lty=3);abline(v=min(m),lty=3)
 T<-min(m)
 for(i in 1:length(S)){
  if( S[i]==min(S[seq(1,T-1,1)]) ) {
   t0hat<-i+1
  }
 }
 print(t0hat)
[1]102
```

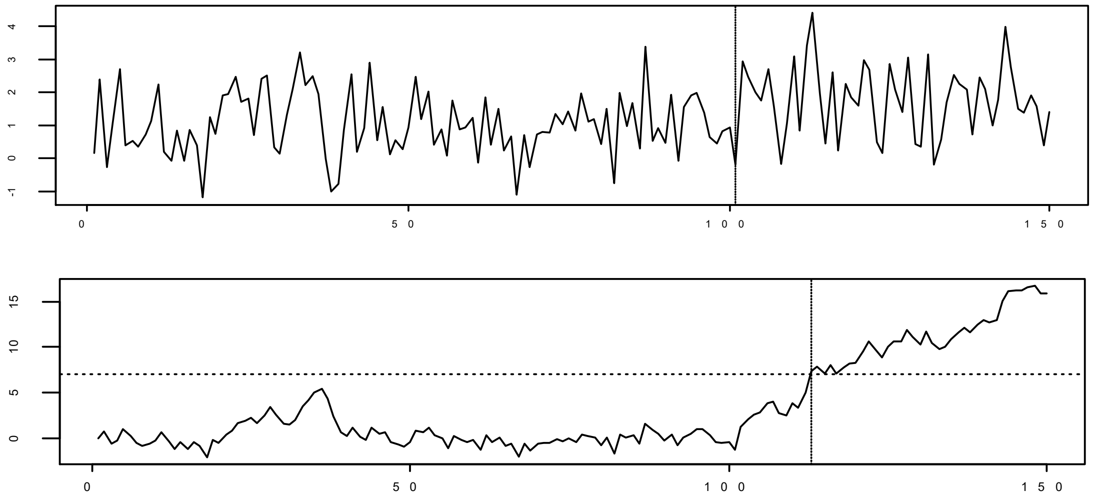


Figure 4.2.3 (n, G_n) *points when there is a mean shift. The horizontal line indicates the threshold* h

In order to estimate τ_0 according to Figure 4.2.3, we have to compute the quantity

$$\hat{\tau}_0 = \underset{1 \leq k < T}{\operatorname{argmin}} S_k + 1$$

At this particular example, we estimated that the time at which there is a change in distribution, is $\hat{\tau}_0 = 102$, since the minimum value of the sum S_k can be achieved when $k = 101$. Therefore, the alarm has been delayed only for $\hat{\tau}_0 - \tau_0 = 102 - 101 = 1$ observation.

4.2.2 The CUSUM algorithm as a repeated SPRT

The second way that Page (1954) defined CUSUM was using SPRT. The CUSUM algorithm used for detecting mean shift (increase), can be seen as a repeated SPRT, each time using zero as the lower threshold and h as the upper one, in the following manner: every time the cumulative sum S_n becomes smaller than zero, the algorithm restarts by setting S_n equal to zero. So CUSUM can be described by the following formula

$$S_n = (S_{n-1} + Z_n)^+$$

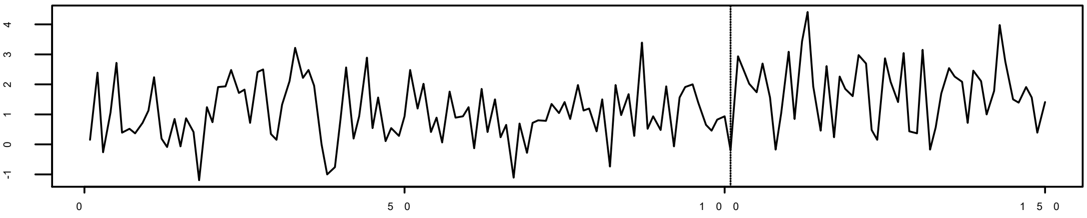
where $S_0 = 0$, $(x)^+ = \max(x, 0)$, and the r.v. Z_i were defined in Paragraph 3.3. The s.t. will now be

$$T = \inf \{n \in \mathbb{N}: S_n \geq h\} \ , \ \inf \emptyset = \infty. \tag{4.1}$$

Similar to the previous program, we generated the first 100 points from $N(\theta_0, \sigma^2)$, where $\theta_0 = 1$ and $\sigma = 1$, whereas the remaining 50 from $N(\theta_1, \sigma^2)$ where $\theta_1 = 1.8$. We use $h = 7$

as the threshold. While the sum S_n lies in the interval $(0, h)$ we continue to calculate its new value. If it becomes smaller or equal to zero, we set $S_n = 0$ and store the point n to the matrix L (the number of elements of the matrix L will be equal to the number of times the algorithm restarted). If $S_n \geq h$, we store n to the matrix k.

```
# R Code (CUSUM as repeated SPRT
set.seed(5)
 theta0 <-1;theta1<-1.8;sigma<-1
 S <-c();Z<-c();L<-c();k<-c();j<-1;l<-1;Z[1]<-0;h <-7
 X1<-rnorm(100,mean=theta0,sd=1);X2<-rnorm(50,mean=theta1,sd=1);X<-c(X1,X2)
 S[1]<-0;
 Z[1]<-(theta1-theta0)/sigma^2*(X[1]-(theta1+theta0)/2)
 S[2]<-S[1]+Z[1]
 for(i in 2:149) {
  if((S[i]>0) & (S[i]<h)) {
   Z[i]<-(theta1-theta0)/sigma^2*(X[i]-(theta1+theta0)/2)
   S[i+1]<-S[i]+Z[i]
  }
  if(S[i]<=0) {
   S[i]<-0
   Z[i]<-(theta1-theta0)/sigma^2*(X[i]-(theta1+theta0)/2)
   S[i+1]<-S[i]+Z[i]
   L[j] <- i
   j <-j+1
  }
  if(S[i]>=h) {
   k[l]<-i
   l<-l+1
   Z[i]<-(theta1-theta0)/sigma^2*(X[i]-(theta1+theta0)/2)
   S[i+1]<-S[i]+Z[i]
  }
 }
 par(mar=c(3,3,2,2)); plot(S,type="l",ylab="",xlab="")
 abline(v=k[1],lty=3);abline(h=h,lty=3)
 for(i in 1:length(S)){
  if( S[i]==min(S[seq(1,T-1,1)]) ) {
   t0hat<-i+1
  }
 }
 print(t0hat);print(length(L)) # changepoint and number of times  the algorithm
restarted
[1] 103
[1]48
```



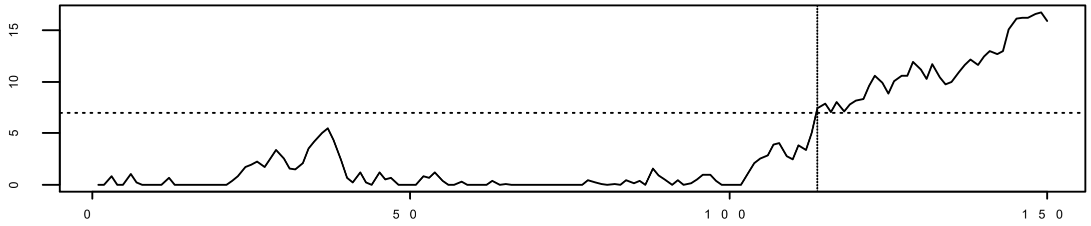


Figure 4.2.4 (n, S_n) points *when there is a mean shift, using SPRT*

From the output of the R software, we have that the change point $\hat{\tau}_0$ is equal to 103 and the number of times the CUSUM algorithm restarted is 48.

4.2.3 V-mask method

Suppose that we want to examine the hypothesis H_0 against two alternatives

$$H_{-1}: \theta = \theta_0 - \delta\sigma \text{ - } H_0: \theta = \theta_0 \text{ - } H_1: \theta = \theta_0 + \delta\sigma \tag{4.2}$$

where $\delta > 0$. Barnard (1959) had proposed a geometric way to detect a mean shift: we plot the cumulative sums $S_n = \sum_{i=1}^{n}(X_i - \theta)$, $1 \leq n \leq m$. At the m-th point, which is the last one, we place a point O at distance d, and from the point O we draw two lines. Each line will have an angle φ with the $x'x$ axis. If any of the points exceeds the upper line then we will have evidence that the mean value increased from its initial value μ_0 . If any point exceeds the lower line, then the mean value will have decreased. In that way, we can check all the points, starting from the last one.

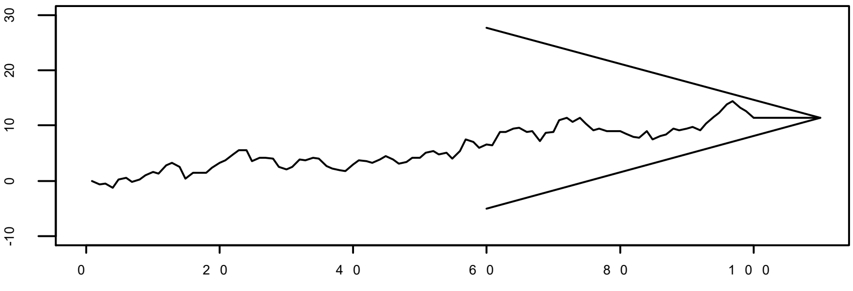


Figure 4.2.5 *Description of CUSUM as a V-mask.*

The parameters φ and d will be related to the error probabilities I and II and the parameter δ. Johnson (1962) proposed a test similar to Armitage's (1950) test. This test consists of the simultaneous use of two SPRT for testing hypothesis H_0 against H_{-1} and H_0 against H_1.

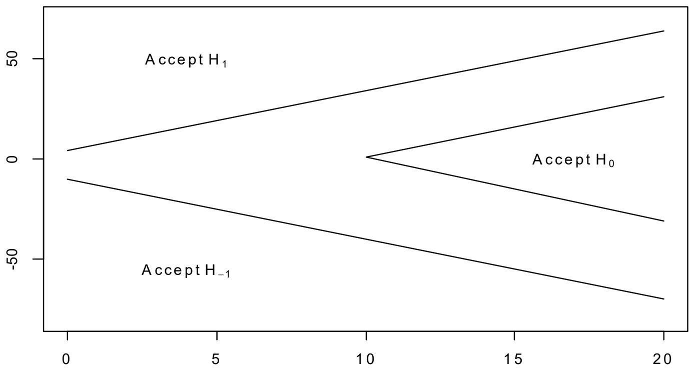


Figure 4.2.6 *Armitage's test*

If we see Figure 4.2.5 inversely, we can verify that it can be interpreted as Armitage's two sided test, under the condition that we will never accept H_0 $(\beta = 0)$ and taking the observations, starting from the last one

$$Y_1 = \sigma^{-1}(X_n - \theta_0),\;\; Y_2 = \sigma^{-1}(X_{n-1} - \theta_0), \ldots\;\; Y_n = \sigma^{-1}(X_1 - \theta_0).$$

Then the likelihood ratio of the first test will be equal to

$$R_n = \frac{\left(\frac{1}{2\pi}\right)^{n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(Y_i + \delta)^2\right)}{\left(\frac{1}{2\pi}\right)^{n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} Y_i^2\right)} = \exp\left(\delta \sum_{i=1}^{n} Y_i + \frac{n\delta^2}{2}\right)$$

Applying the methodology developed in Chapter 3, after calculations, we get that we will accept H_{-1} if $\sum_{i=1}^{n} Y_i \le -\frac{n\delta}{2} - \frac{1}{\delta}\log\left(\frac{1-\beta}{\alpha}\right)$, and when $\beta = 0$ we will have $\sum_{i=1}^{n} Y_i \le -\frac{n\delta}{2} + \frac{1}{\delta}\log(\alpha)$. Similarly we will accept H_1 when $\sum_{i=1}^{n} Y_i \ge \frac{n\delta}{2} - \frac{1}{\delta}\log(\alpha)$. The lines $\varepsilon_1{:}\, y_1 = -\frac{x\delta}{2} + \frac{1}{\delta}\log(a)$ and $\varepsilon_2{:}\, y_2 = \frac{x\delta}{2} - \frac{1}{\delta}\log(a)$ will intersect at the point $\left(\frac{2}{\delta^2}\log(a)\,, 0\right)$, and so if we denote d the distance from the point $(0{,}0)$ and φ the angle of the line ε_1 with the $x'x$ axis, we have $\tan\varphi = \frac{\delta}{2}$. We concluded that Barnard's (1959) method, can be illustrated using tools of the sequential methodology. Many statistical packages include options for the CUSUM algorithm and the V-mask.

4.2.4 H-K method

Page (1954) had used another form of the CUSUM algorithm, to test the two sided hypotheses (4.2). We define the cumulative sums

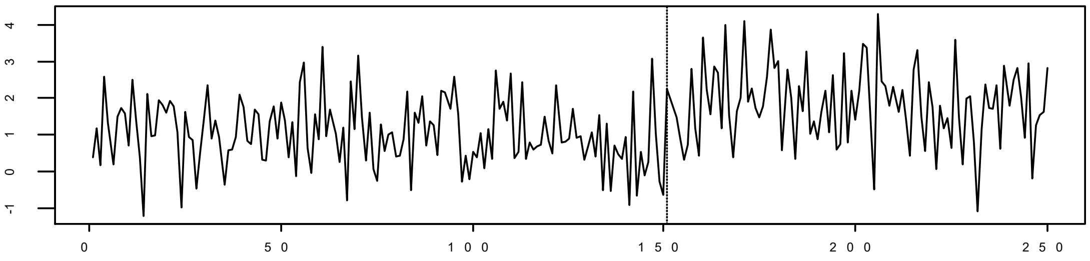
$$C_n^+ = \max\{0, X_n - (\theta_0 + K) + C_{n-1}^+\}\,,\;\; C_0^+ = 0$$

$$C_n^- = \min\{0, X_n - (\theta_0 - K) + C_{n-1}^-\}\,, \quad C_0^- = 0$$

where $K = k\sigma$ and $k = \frac{\delta}{2}$. Suppose that $H = h\sigma$. We will accept $\mathcal{H}_1$ if $C_n^+ > H$ and reject $\mathcal{H}_{-1}$ if $C_n^- < -H$. According to the bibliography, the parameters K and H will be called *reference value* and *decision interval* respectively. In the graph, we present both sums C_n^+ and C_n^-.

The first 150 points were generated from the distribution $N(\theta_0, \sigma^2)$, for $\theta_0 = 1$ and $\sigma = 1$, whereas the remaining 100 from $N(\theta_1, \sigma^2)$, $\theta_1 = 1.8$. We used $H = 5$, $K = 0.5$, and calculated the sums C_n^+ and C_n^- from the above formulas. If $C_n^- < -H$ and $C_n^+ > H$, we store n to the matrix f .Clearly, the first element of the matrix, which is its minimum, is the first point out of $(-H, H)$. We estimated that $n = 165$.

```
 # R code (H-K method)
 set.seed(1)
 theta0<-1;theta1<-1.8;sigma<-1;K<-k*sigma;delta<-1;k<-delta/2;h<-5;H<-h*sigma
 Cnplus<-c();Cnminus<-c();X<-c()
 i<-1;d<-0
 X0<-rnorm(150,mean=theta0,sd=sigma); X1<-rnorm(100,mean=theta1,sd=sigma)
 X<-c(X0,X1);plot(X,type="l");abline(v=151,lty=3)
 Cnplus[1]<-0;Cnminus[1]<-0;j<-1;f<-c()
 for(i in 1:length(X)) {
 Cnplus[i+1]<-max(0,X[i]-(theta0+K)+Cnplus[i])
 Cnminus[i+1]<-min(0,X[i]-(theta0-K)+Cnminus[i])
  if((Cnplus[i+1]>H) | (Cnminus[i+1]<(-H))) {
  f[j]<-i+1
  j<-j+1
  }
 }
 par(mar=c(3,3,2,2))
 plot(Cnplus,type="l",main="",ylim=c(-8,35),xlab="",ylab="")
 lines(Cnminus,type="l");
 abline(h=H,lty=2);abline(h=-H,lty=2);abline(v=min(f),lty=3)
 print(min(f)) # first point out of the interval (-H,H)
[1]165
```



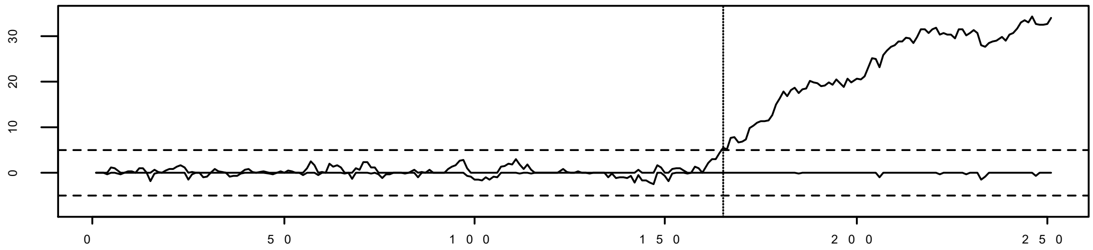


Figure 4.2.7 *H-K method for* $k = 0.5$ *and* $h = 5$*. The horizontal lines are the thresholds H and* $-H$

This form of the CUSUM algorithm is the default form included in the statistical packages and is analogous to Shewhart's two sided control chart.

4.3 ARL function

We will use some basic definitions equivalent to the ASN from Chapter 3, which will be then applied to assess the performance of statistical algorithms. Let τ be a r.v. that shows the number of points in a control chart until the first alarm.

Definition 4.3.1 *We will call average run length (ARL) of a statistical detection algorithm, the expectation of the r.v.* τ*:*

$$\mathrm{ARL} = \mathbb{E}_\theta(\tau).$$

Note that under the framework of the test (3.1) we distinguish two cases:

(i) When $\theta = \theta_0$ the quantity $\mathrm{ARL}_0 = \mathbb{E}_{\theta_0}(\tau)$ will express the expected number of points until a false alarm. Hence, we could say that this is the *mean time between two false alarms*.

(ii) When $\theta = \theta_1$ the quantity $\mathrm{ARL}_1 = \mathbb{E}_{\theta_1}(\tau)$ will express the expected number until a mean-shift detection. Clearly, this quantity is the same as the *mean detection delay.*

We will use all the above in Paragraph 4.3.3, to choose the value of the parameter h. Generally, for a statistical detection algorithm, we want to have large ARL_0 and small ARL_1. We will develop analytical methods to compute ARL_θ for every value of θ. Consider from now on that CUSUM is expressed as a repeated SPRT (see Paragraph 4.2.2).

Denote:

z: the initial value of the CUSUM's statistical function

$P(z)$: the probability that a run starting from z ends at the lower threshold

$N(z)$: the average run length, with z as the initial value of the run

$N_1(z)$: the average run length, when the run crosses 0

$N_2(z)$: the average run lengh when the run crosses h

$L(z)$: the average run length of the CUSUM algorithm

τ_k: the r.v. showing the number of SPRT used (see Paragraph 4.2.2, where $\tau_k = 48$), when the sum S_n crosses the lower bound 0

$T_{\varepsilon,h}$: the stopping time (3.4) with lower threshold ε and upper threshold h

One can verify that τ_k has the geometric distribution with parameter $P(0)$ and thus

$$\mathbb{E}(\tau_k) = \frac{P(0)}{1 - P(0)},$$

and

$$\begin{aligned} \text{ARL} = L(0) &= \frac{P(0)}{1 - P(0)} \cdot N_1(0) + N_2(0) \\ &= \frac{P(0)N_1(0) + (1 - P(0))N_2(0)}{1 - P(0)} = \frac{N(0)}{1 - P(0)} \end{aligned} \tag{4.3}$$

There are many methods to compute the quantities $N(0)$ and $P(0)$. The functions $P(z)$ and $N(z)$ satisfy the Fredholm equations of type two (Page, 1954)

$$P(z) = \int_{-\infty}^{-z} f(x)dx + \int_0^h P(x)f(x - z)dx$$

and

$$N(z) = 1 + \int_0^h N(x)f(x - z)dx$$

where f is the p.d.f. of the r.v. $Z = \log\left(\frac{f(X|\theta_1)}{f(X|\theta_0)}\right)$.

Even if these equations cannot be solved analytically, one can use arithmetic approximations and particularly the method of *Gaussian quadrature,* under the assumption that f satisfies certain regularity conditions (see Baseville, 1993, pg. 169). It can be proven that this method gives a precise approximation of the ARL (see Granjon, 2012, pg. 13) but requires many computations.

4.3.1 Wald's and Siegmund's approximation

We will now use all the results from Wald's theory discussed in Chapter 3. According to the notation

$$L(0) = \frac{\mathbb{E}(T_{0,h})}{1 - P(0)} \tag{4.4}$$

and when $\theta \neq \tilde{\theta}$ ($\tilde{\theta}$ is such that $\mathbb{E}_{\tilde{\theta}}(Z) = 0$) we will have

$$\mathbb{E}_\theta(T_{\varepsilon,h}) \approx \frac{\varepsilon \mathbb{P}(S_T \leq \varepsilon) + h\big(1 - \mathbb{P}(S_T \leq \varepsilon)\big)}{\mathbb{E}_\theta(Z)}$$

$$\mathbb{P}(S_T \leq \varepsilon) \approx \frac{e^{-t_0(\theta)h} - 1}{e^{-t_0(\theta)h} - e^{-t_0(\theta)\varepsilon}} \tag{4.5}$$

where t_0 is the solution of the equation

$$\mathbb{E}_\theta\big(e^{-t_0(\theta)Z}\big) = 1.$$

Because a substitution of formula (4.5) to formula (4.4), for $\varepsilon = 0$, will give an undefined quantity ($P(0) = \mathbb{P}(S_T \leq 0) = 1$), we will take the limit as ε tends to zero. Substituting formula (4.5) to formula $(T_{\varepsilon,h})$ we get that

$$\mathbb{E}_\theta(T_{\varepsilon,h}) \approx \frac{1}{\mathbb{E}_\theta(Z)}\left(h + \varepsilon \frac{1 - e^{-t_0(\theta)h}}{1 - e^{t_0(\theta)\varepsilon}}\right)$$

and when $\varepsilon \to 0$,

$$\mathbb{E}_\theta(T_{0,h}) \approx \frac{1}{\mathbb{E}_\theta(Z)}\left(h + \frac{e^{-t_0(\theta)h}}{t_0(\theta)} - \frac{1}{t_0(\theta)}\right).$$

If $\theta = \tilde{\theta}$,

$$\mathbb{E}_{\tilde{\theta}}(T_{\varepsilon,h}) \approx \frac{\varepsilon^2 \mathbb{P}(S_T \leq \varepsilon) + h^2\big(1 - \mathbb{P}(S_T \leq \varepsilon)\big)}{\mathbb{E}_{\tilde{\theta}}(Z^2)}$$

and

$$\mathbb{P}(S_T \leq \varepsilon) \approx \frac{h}{h + \varepsilon}.$$

Substituting $\mathbb{P}(S_T \leq \varepsilon)$ to formula of $\mathbb{E}_{\tilde{\theta}}(T_{\varepsilon,h})$ and letting $\varepsilon \to 0$,

$$\mathbb{E}_{\tilde{\theta}}(T_{0,h}) = \frac{h^2}{\mathbb{E}_{\tilde{\theta}}(Z^2)}$$

Therefore we concluded that

$$\mathrm{ARL}_\theta \approx \begin{cases} \dfrac{1}{\mathbb{E}_\theta(Z)}\left(h + \dfrac{e^{-t_0(\theta)h}}{t_0(\theta)} - \dfrac{1}{t_0(\theta)}\right), & \theta \neq \tilde{\theta} \\ \dfrac{h^2}{\mathbb{E}_\theta(Z^2)}, & \theta = \tilde{\theta}. \end{cases}$$

Siegmund (1985) had proposed a similar way. Wald (1947) actually did not take into account the differences $|S_T - h|$ and $|S_T - 0|$ while Siegmund, using the *diffusion approximation theory,* estimated them.

Let us denote

$$\rho_+ = \mathbb{E}_\theta(S_T - h|S_T - h \geq 0)$$

$$\rho_- = \mathbb{E}_\theta(S_T|S_T \leq 0)$$

It can then be proved that (see Baseville, 1993, pg. 174-175)

$$\mathrm{ARL}_\theta \approx \begin{cases} \dfrac{1}{\mathbb{E}_\theta(Z)}\left(h + \rho_+ - \rho_- + \dfrac{e^{-t_0(\theta)(h+\rho_+-\rho_-)}}{t_0(\theta)} - \dfrac{1}{t_0(\theta)}\right), & \theta \neq \tilde{\theta} \\ \dfrac{(h + \rho_+ - \rho_-)^2}{\mathbb{E}_\theta(Z^2)}, & \theta = \tilde{\theta} \end{cases}$$

The above formula is actually Wald's approximation where we replaced h with $h + \rho_+ - \rho_-$. In the case of normal distribution

$$\rho_+ - \rho_- = -\frac{2}{\pi}\int_0^{+\infty} x^{-2}\log\left[\frac{2}{x^2}\left(1 - e^{-\frac{1-x^2}{2}}\right)\right]dx \approx 2 \cdot 0.583 = 1.166$$

4.3.2 Comparing ARL functions

We will now examine the different formulas of the ARL function, for various values of θ. Consider the CUSUM algorithm, as a repeated SPRT test, for $\theta_0 = 1$, $\theta_1 = \theta_0 + \delta\sigma$, $\sigma = 1$, $\delta = 1$ and $h = 4$. From Chapter 3 we can compute that

$$\mathbb{E}_\theta(Z) = \theta - 1.5,\ \ \mathbb{E}_{1.5}(Z^2) = 1,\ \ t_0(\theta) = 2\theta - 3.$$

Hence the ARL function is given from the following formulas

$$\mathrm{ARL}_\theta \approx \begin{cases} \dfrac{1}{(\theta - 1.5)}\left(h + \dfrac{e^{-(2\theta-3)h}}{2\theta - 3} - \dfrac{1}{2\theta - 3}\right), & \theta \neq 1.5 \\ h^2, \ \ \theta = 1.5 \end{cases}$$

$$\mathrm{ARL}_\theta \approx \begin{cases} \dfrac{1}{(\theta-1.5)}\left(h+1.166+\dfrac{e^{-(2\theta-3)(h+1.66)}}{2\theta-3}-\dfrac{1}{2\theta-3}\right), & \theta \neq 1.5 \\ (h+1.166)^2, \quad \theta = 1.5 \end{cases}$$

from Wald's and Siegmund's approximation, respectively.

Using R (see Appendix IV, Program 8), we can see the differences between the two approximations of the ARL function, with respect to θ, and compare them with the simulated values. We create a while loop, for various values of θ. The values of the statistical functions occur simply by examining whether $\theta = 1.5$ or $\theta \neq 1.5$.

For the simulated values, we create a while loop, checking if the sum $S_n = \sum_{i=1}^{n} Z_i$ becomes less than $h = 4$. If $S_n \leq 0$ for any values of S_n, we set $S_n = 0$, according to the definition of the CUSUM algorithm.

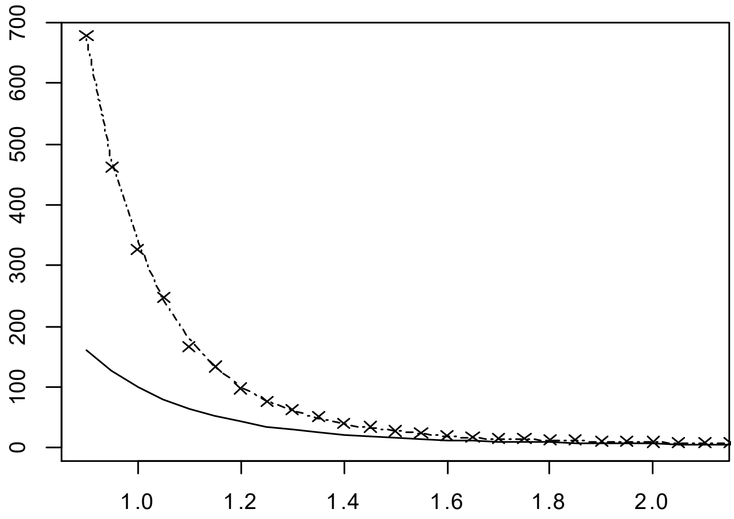


Figure 4.3.1 *The ARL function for various values of θ, via simulation (x), Siegmund's approximation (dashed line) and Wald's approximation*

We can verify that Siegmund's approximation is better than Wald's, especially for values of θ smaller than $\tilde{\theta}$.

We will now compare CUSUM with Shewhart's control chart for individual observations and known variance.

It can be proved that the ARL function for the two-sided Shewhart control chart, with respect to δ, is given by the formula (e.g. see Antzoulakos, 2010, pg. 36)

$$\mathrm{ARL}(\delta) = \frac{1}{1-\Phi(\delta+3)+\Phi(\delta-3)}$$

while for the two-sided CUSUM algorithm (see Granjon, 2012, pg. 15)

$$\frac{1}{\mathrm{ARL}(\delta)} = \frac{1}{\mathrm{ARL}_{+}(\delta)} + \frac{1}{\mathrm{ARL}_{-}(\delta)}$$

where $\mathrm{ARL}_{+}(\delta)$ is the average run length for detecting a mean shift (increase) of order $|\delta|\sigma$ and $\mathrm{ARL}_{-}(\delta)$ for detecting a mean decrease of order $|\delta|\sigma$. The formula for $\mathrm{ARL}_{-}(\delta)$ is similar to $\mathrm{ARL}_{+}(\delta)$, which was computed analytically, and can be found using the proper $t_0 = t_0(\theta)$.

In order to have an ARL_0 equal to 370.4, the parameters were calculated using Siegmund's approximation. Furthermore, we found that $h = 4.76713$ and $\delta = 1$. Initially, we got the value of ARL_0 for the two sided CUSUM test, with the use of a function. The function has h as a parameter and can be used in a loop to detect the values of h for which $\mathrm{ARL}_0 > 370.3$ and $\mathrm{ARL}_0 < 370.5$. Having collected two values, we can verify that the required h is 4.76713.

```
 # R code (ARL0)
 # ARL0~370.4 and δ=1
 theta0<-1;theta<-theta0
 ARLf<-function(h) {
 ARLplus<-1/(theta-1.5)*(h+1.166+exp(-(2*theta-3)*(h+1.166))/(2*theta-3)-
1/(2*theta-3))
 ARLminus<-1/(0.5-theta)*(h+1.166+exp(-(1-2*theta)*(h+1.166))/(1-2*theta)-
1/(1-2*theta))
 w <- 1/ARLplus + 1/ARLminus
 return(1/w)
  }
 for (i in seq(1,5,0.00025)) {
   if (ARLf(i)>370.3 & ARLf(i)<370.5) {
   print(ARLf(i));print(i) }
 }
[1] 370.3517
[1] 4.767
[1] 370.4458
[1] 4.76725
```

Now, in order to find ARL for different values of δ, we use the above formulas inside two loops. One for the CUSUM algorithm and one for Shewhart's chart.

```
 # R code (ARL computation)
 h<-4.76713;sigma<-1;theta0<-1
 ARLcusum<-c();ARLplus<-c();ARLminus<-c();j<-1
 for(delta in seq(0,4,0.25)){
 theta<-theta0+delta*sigma
 ARLplus[j]<-ifelse(theta==1.5,(h+1.166)^2,1/(theta-1.5)*(h+1.166+exp(-
(2*theta-3)*(h+1.166))/(2*theta-3)-1/(2*theta-3)))
 ARLminus[j]<-ifelse(theta==0.5,((h+1.166)^2)/2,1/(0.5-
theta)*(h+1.166+exp(-(1-2*theta)*(h+1.166))/(1-2*theta)-1/(1-2*theta)))
 ARLcusum[j]<-1/(1/ARLplus[j]+1/ARLminus[j])
 j<-j+1
 }
 ARLshew<-c();k<-1  # ARL shew
```

```
for(delta in seq(0,4,0.25)){
ARLshew[k]<-1/(1-pnorm(delta+3,mean=0,sd=1)+pnorm(delta-3,mean=0,sd=1))
k<-k+1
}
cbind(seq(0,4,0.25),ARLcusum,ARLshew)
```

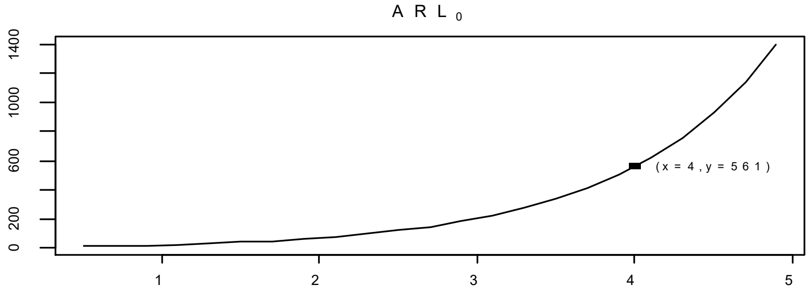
TABLE 4.3.1
Comparison of the ARL function between Shewhart chart and CUSUM

δ	*CUSUM*	*Shewhart*	δ	*CUSUM*	*Shewhart*
0,00	370,40	370,40	2,00	3,73	6,30
0,25	121,36	281,15	2,25	3,23	4,41
0,50	35,18	155,22	2,50	2,84	3,24
0,75	16,14	81,22	2,75	2,54	2,49
1,00	9,87	43,89	3,00	2,29	2,00
1,25	7,02	24,96	3,25	2,09	1,67
1,50	5,43	14,97	3,50	1,92	1,45
1,75	4,43	9,47	3,75	1,78	1,29

We can see that for small values of δ, the CUSUM algorithm has smaller average length, in comparison with the Shewhart chart. This means that it detects a mean shift faster. The opposite happens for values of δ greater than 2.75, but the difference is very small. We conclude, that the most efficient algorithm is CUSUM, whereas for large shifts (greater or equal to 3σ) the Shewhart control chart is slightly better.

4.3.3 Tuning parameters

As we have seen so far, we need to estimate the parameters θ_0, θ_1, σ^2 and h in order to use CUSUM. When θ_0 and σ^2 are unknown, we use an initial sample size and calculate the sample mean and sample variance. When θ_1 is unknown, the choice of δ depends on our a-priori knowledge of the data. So, the experimenter should decide what will the order of mean shift be, using all the previous observations.



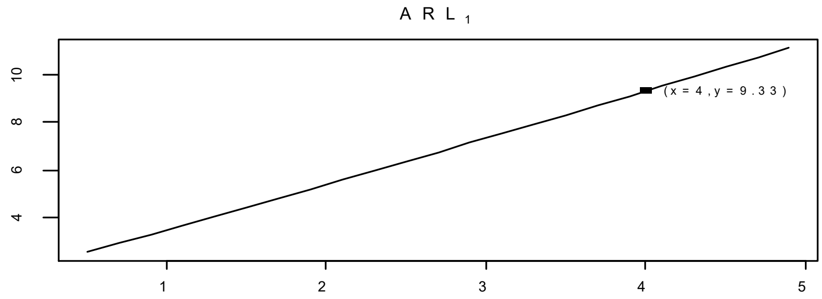


Figure 4.3.2 *ARL as a function of h, according to Siegmund's approximation*

We can get useful information about the CUSUM algorithm, from the above graphs (see Granjon, 2012, pg. 14). One can verify that CUSUM with parameter $h = 4$, will have an average of one false alarm in every 561 observations. Also, it detects a mean shift of order 1σ with a mean delay of 9 points. From these two diagrams, we conclude that the smaller the h the more sensitive to detection the algorithm will be. We will have small values of ARL_0 and thus many false alarms. If on the other hand we choose a large h, then the algorithm will give a small number of false alarms, but there will be a delay in detecting the mean shift. This analogy can be tuned each time by the experimenter.

4.4 Optimal property of the CUSUM

In this paragraph we will present some problems from o.s.t. We note that when τ_0 is the change point, we will denote $\mathbb{P}_n$ as the probability measure with p.d.f. given by the relation

$$f_{\theta_1}(x_1, \dots, x_n) = \prod_{k=1}^{\tau_0 - 1} f_{\theta_0}(x_k) \cdot \prod_{k=\tau_0}^{n} f_{\theta_1}(x_k)$$

and with $\mathbb{P}_\infty$ the probability measure if there is no change in distribution.

A trivial question that one could make is the following: since for the CUSUM we used SPRT which has an optimal property, will CUSUM have also an optimal property, concerning its detection speed? The answer is positive and was firstly introduced for the asymptotic case by Lorden (1971). While for the SPRT we had the quantities a and β which presented the type I and II errors and were used as initial conditions, we now have ARL_0 and ARL_1 instead. To evaluate CUSUM's performance, we will examine the mean time between false alarms and the average detection delay, as mentioned in Paragraph 4.1. These two quantities are actually ARL_0 and ARL_1.

Denote τ_0 the actual change point of the distribution, T the s.t. of the statistical algorithm and $\varDelta_\gamma = \{T: \mathbb{E}_\infty(T) \geq \gamma\}$ where $\gamma > 1$.

Lorden (1971) had proposed a measure to assess the detection delay of a statistical algorithm. Namely, the *essential supremum average detection delay* (ESADD) which is defined as

$$\mathrm{ESADD}(T) = \sup_{\tau_0 \geq 1} \text{ess sup } \mathbb{E}_{\theta_1}\left[(T - \tau_0 + 1)^+ | \mathcal{F}_{\tau_0 - 1}\right]$$

where $\mathcal{F}_{\tau_0 - 1}$ is the natural filtration and the expectation is taken under the p.d.f. f_{θ_1}. The problem is to find a s.t. $T \in \Delta_\gamma$ which will minimize $\mathrm{ESADD}(T)$. It was proved initially (Lorden, 1971) that the s.t. (4.1), i.e. the CUSUM algorithm, satisfies this condition asymptotically, i.e. $\gamma \to \infty$ and that the formula connecting the *mean detection delay* with the *mean time between false alarms* is asymptotically:

$$\mathrm{ESADD}(T) \approx \frac{\log(\gamma)}{\mathrm{D}(f_1 || f_0)}$$

Later on, (Moustakides, 1986) proved that the CUSUM algorithm minimizes the quantity $\mathrm{ESADD}(T)$ for every $\gamma > 1$.

We will now mention two other problems of finding performance measures, that usually appear in the bibliography related to sequential analysis. Their solutions however, are algorithms different from CUSUM.

(i) Pollack (1985) defined a measure for the evaluation of the detection delay, as the *supremum average detection delay* (SADD)

$$\mathrm{SADD}(\tau) = \sup_{\tau_0 \geq 1} \mathbb{E}_{\tau_0}(T - \tau_0 | T \geq \tau_0)$$

We search again for a s.t. $T \in \Delta_\gamma$ that minimizes SADD. The r.v. T which is a solution to this problem is the s.t. defined by Shiryaev-Roberts (see Veeravalli, 2012, pg. 26).

(ii) When the change point τ_0 is a r.v. with non negative values, we could define as *average detection delay* (*ADD*) the quantity given by the formula

$$\mathrm{ADD}(\tau) = \mathbb{E}((T - \tau_0)^+) = \sum_{n=1}^{\infty} \pi_n \mathbb{E}_n (T - \tau_0)^+$$

where $\pi_n = \mathbb{P}(\tau_0 = n)$. We denote as PFA the *probability of false alarm* (*PFA*)

$$\mathrm{PFA}(T) = \mathbb{P}(T < \tau_0) = \sum_{n=1}^{\infty} \pi_n \, \mathbb{P}_n(T < \tau_0)$$

Assume that $\Delta_a = \{T : \mathrm{PFA}(T) \leq \alpha\}$, $a \in (0,1)$. One can prove that when the r.v. τ_0 is geometrically distributed, then the s.t. $T \in \Delta_a$ which will minimize PFA, is Shiryaev's algorithm (see Veeravalli, 2012, pg. 11-14)

The problems stated so far, are actually problems of finding the stopping time which will minimize a formula, while satisfying initial conditions. The field of o.s.t. is the suitable tool for constructing s.t. that appear in problems of statistical detection theory.

Summary

We saw how the theory of martingales contributes to proving the basic equations of sequential analysis. Wald’s equations are the mathematical tools for finding the solution of various problems in statistics, such as the estimation of the upper bound for the mean sample size of the c.i. with fixed width, the construction of the ASN function for testing statistical hypotheses, the estimation of the ARL function needed to examine the CUSUM and also for many other problems including the gambler’s ruin (Ghosh, 1991).

Observing the numerical results, we concluded that the sequential statistical methods require small sample size, e.g. the purely sequential method for c.i. construction, the SPRT in comparison with the Neyman-Pearson test and the CUSUM algorithm compared to the Shewhart control chart.

One of the most important results of sequential analysis was the definition of the SPRT’s optimal property, since it was the first step needed for the creation of o.s.t. The tools of o.s.t. can be used for solving fundamental problems in the field of statistical detection, classifying algorithms in accordance with their optimal properties.

Modern applications of sequential analysis include problems of defining quickest change detection algorithms but for correlated r.v. and also the investigation of their properties. Additionally, when the properties are not inherited, an asymptotic approach is recommended. For instance (Baseville, 1993) there are some variations of the CUSUM algorithm, Bayes type algorithms and their extension. A full presentation of the properties of novel statistical algorithms for the non-i.i.d. case is given by Veeravalli (2012).

Appendix

(I) Kullback-Leibler divergence

Definition. *Assume two continuous r.v. with p.d.f.* f_1 *and* f_2 *respectively. We define as K-L the quantity*

$$D_{KL}(f_1||f_2) = \int_{-\infty}^{\infty} f_1(x)\log\left(\frac{f_1(x)}{f_2(x)}\right)dx$$

Proposition. *For the K-L divergence* $D_{KL}(f_1||f_2) \geq 0$.

Proof. Ισχύει ότι

$$\begin{aligned} D_{KL}(f_1||f_2) &= \int_{-\infty}^{\infty} f_1(x)\log\left(\frac{f_1(x)}{f_2(x)}\right)dx \\ &= \mathbb{E}_{f_1}\left[-\log\left(\frac{f_2(x)}{f_1(x)}\right)\right] \geq -\log\left[\mathbb{E}_{f_1}\left(\frac{f_2(x)}{f_1(x)}\right)\right] \\ &= -\log\left[\int_{-\infty}^{\infty}\frac{f_2(x)}{f_1(x)}f_1(x)dx\right] = -\log(1) = 0, \end{aligned}$$

where in the second identity we used the Jensen inequality. □

Let $Z = \log\left(\frac{f(x|\theta_1)}{f(x|\theta_0)}\right)$, with $\theta_1 > \theta_0$. We observe that

$$\begin{aligned} \mathbb{E}_{\theta_0}(Z) &= \int_{-\infty}^{\infty} f(x|\theta_0)\log\left(\frac{f(x|\theta_1)}{f(x|\theta_0)}\right)dx \\ &= -\int_{-\infty}^{\infty} f(x|\theta_0)\log\left(\frac{f(x|\theta_0)}{f(x|\theta_1)}\right)dx = -D_{KL}(f_0||f_1) < 0. \end{aligned}$$

Similarly we can prove that $\mathbb{E}_{\theta_1}(Z) > 0$.

(II) Matrix theory

In order to calculate Σ_{11} we will find first the inverse of the matrix Σ

$$\Sigma = \begin{bmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ \rho & 1 & \cdots & \rho^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \cdots & 1 \end{bmatrix}$$

Proposition. $\det \Sigma = (1-\rho^2)^n$

Proof. For $n = 2$ we have that $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ and thus $\det \Sigma = 1 - \rho^2$. Suppose the required formula is valid for $n = k$. Then if

$$\Sigma^k = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{k-1} & \rho^k \\ \rho & 1 & \rho^3 & \cdots & \rho^{k-2} & \rho^{k-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho^{k-1} & \rho^{k-2} & \rho^{k-3} & \cdots & 1 & \rho \\ \rho^k & \rho^{k-1} & \rho^{k-2} & \cdots & \rho & 1 \end{bmatrix}$$

and we multiply by $-\rho$ the penultimate line and add it to the last one, we obtain the matrix

$$\begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{k-1} & \rho^k \\ \rho & 1 & \rho^3 & \cdots & \rho^{k-2} & \rho^{k-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho^{k-1} & \rho^{k-2} & \rho^{k-3} & \cdots & 1 & \rho \\ 0 & 0 & 0 & \cdots & 0 & 1-\rho^2 \end{bmatrix}$$

and expanding the determinant of the matrix with respect to the last row, we get

$$(1-\rho^2)\det \Sigma^{k-1} = (1-\rho^2)(1-\rho^2)^k = (1-\rho^2)^{k+1}$$

and thus the required formula is valid according to induction. □

In order to find the inverse of the matrix Σ^k we will work initially in dimensions two, three and four. The inverse matrices which derive after simple calculations are

$$\Sigma_2^{-k} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}, \Sigma_3^{-k} = \frac{1}{(1-\rho^2)^2}\begin{bmatrix} 1 & -\rho & 0 \\ -\rho & 1+\rho^2 & -\rho \\ 0 & -\rho & 1 \end{bmatrix}$$

$$\Sigma_4^{-k} = \frac{1}{(1-\rho^2)^3}\begin{bmatrix} 1 & -\rho & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 \\ 0 & -\rho & 1+\rho^2 & -\rho \\ 0 & 0 & -\rho & 1 \end{bmatrix}$$

Therefore for k we define

$$\Sigma_k^{-k} = \frac{1}{(1-\rho^2)^{k-1}}\begin{bmatrix} 1 & -\rho & \cdots & 0 \\ -\rho & 1+\rho^2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & -\rho \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

We conclude that $\Sigma^k \Sigma_k^{-k} = \mathrm{I}$ and due to the uniqueness of the inverse matrix $(\Sigma^k)^{-1} = \Sigma_k^{-k}$. Thus from the above we get that the formula for finding Σ_{11}^{-1}, stated in paragraph 3.6.3, is true.

(III) R code for figures

Figure 1.5.1

```
 # R Code (MG function)
 install(plotrix)
 library(plotrix)
 par(mar=c(4,4,2,2))
 par(mfrow=c(1,2))
 x <- seq(-0.5,2,0.2)
 y <- x*(x-1)+1
 pl.1 <- plot(x,y,bty="n",type="l", col="black", bty="n",lwd=2, ylab="",cex=0.9,
xlab="",main="", ylim=c(0,2.5),xlim=c(-0.5,1.7))
 ablineclip(a=1,b=0, lty=1, lwd=2, col="black", x1=-0.5, x2=1.5)
 x1 <- seq(-1.8,0.5,0.2)
 y1 <- x1*(x1+1)+1
 pl.2 <- plot(x1,y1,bty="n",type="l", col="black", bty="n",lwd=2, ylab="",cex=0.9,
xlab="",main="", ylim=c(0,2.5), xlim=c(-2, 0.5))
 ablineclip(a=1,b=0, lty=1, lwd=2, col="black", x1=-1.5, x2=0.5)
```

Figure 2.3.1

```
 # R code (Histogram)
 set.seed(1)
 n <- 10000;n0 <- 10;mu <- 1;sigma <- 3;a <- 0.05;
 t <- qt(1-a/2,df=n0-1)  # ποσοστιαίο σημείο της κατανομής Student
 Ts <- matrix()          # πίνακας με τα μεγέθη δειγμάτων T
 j <- 1;d <- 0.5
 k <- qnorm(1-a/2,mean=0,sd=1)^2*sigma^2/d^2
  for(i in 1:n) {
    X1 <- rnorm(n0, mean=mu, sd=sigma)
    Ts[i] <- max(n0, floor(t^2*sd(X1)^2/d^2)+1)
   }
 Prob <- matrix();f <-1;l <-1
 for(j in n0:600) {
  if(j==n0) {
  Prob[l] <- pchisq(n0*(n0-1)*d^2/(sigma^2*t^2), df=n0-1)
  } else {
  Prob[l] <- pchisq( (n0+f)*(n0-1)*d^2/(sigma^2*t^2),df=n0-1) - pchisq( (n0+f-
1)*(n0-1)*d^2/(sigma^2*t^2),df=n0-1)
  f <- f+1
+  }
+  l<-l+1
+ }
> G <- n0:600;par(mar=c(3,4,3,2));par(mfrow=c(1,1))
> Histo <- hist(Ts,
main="",prob=T,xlab="",ylab="",ylim=c(0,0.005),xlim=c(0,600),breaks=30)
> lines(x=G,y=Prob,col="gray21",lty=1, lwd=3))
```

Figure 3.3.1

```
 library(MASS) ;library(plotrix)
 set.seed(2)
 a=0.05;b=0.1;sigma=2;theta0=1;theta1=1.4;A <- b/(1-a); B <- (1-b)/a
 C1 <- sigma^2*log(A)/(theta1-theta0);C2 <- sigma^2*log(B)/(theta1-theta0)
 D <-  (theta1+theta0)/2
 Mat.H0 <- matrix();x0 <-rnorm(1,mean=theta0, sd=sigma)
 S <- x0;i <- 1;
 Mat.H0[i] <- S
 while(S>(C1+i*D) & S<(C2+i*D)) {
  x <- rnorm(1, mean=theta0, sd=sigma);
  S <- S+x;
  i <- i+1
  Mat.H0[i] <- S
 }
 par(mar=c(3,5,4,2))
 plot(Mat.H0,type="l", col="black", bty="n",lwd=2,ylab="",cex=0.9, xlab="",main="",
ylim=c(-20,120),xlim=c(-20, 110))
 ablineclip(C1,D, lty=1, lwd=2, col="black", x1=-5 ,x2=100, y1=-20)
 ablineclip(C2,D, lty=1, lwd=2, col="black", x1=-10, x2=80)
```

Figure 4.1.1

```
 set.seed(2)
 X<-rnorm(150,1,1);Y<-rnorm(100,2,1)
 Z <-c(X,Y);par(mar=c(3,3,2,2))
 index=1:250
 plot(index,Z, type="l", lwd=1,ylab="",xlab="")
 abline(v=151,lty=3);abline(v=124,lty=3)
 abline(v=170,lty=3)
```

Figure 4.2.1

```
 set.seed(5)
 theta0 <-1;theta1 <-1.8;sigma<-1
 S<-matrix();X1<-rnorm(100,mean=theta0,sd=1);X2<-
rnorm(50,mean=theta1,sd=1);X<-c(X1,X2)
 S[1] <- X[1]-theta0
 for(i in 1:99) {
 S[i+1]<-S[i]+(X[i+1]-theta0)
 }
 for(i in 100:149) {
 S[i+1] <- S[i]+(X[i+1]-theta0)
 }
 par(mar=c(3,3,2,2));par(mfrow=c(1,1))
 plot(X,type="l",lwd=1,ylab="",xlab="")
 abline(v=101,lty=3)
 plot(S, type="l", lwd=1,ylab="",xlab="")
 segments(x0=0,y0=0,x1=100,y1=0,lty=2,lwd=1)
 segments(x0=100,y0=0,x1=150,y1=(theta1-theta0)*(150-100),lty=2,lwd=1)
```

Figure 4.2.2

```
 set.seed(5)
 theta0 <-1 ; theta1 <-1.8;sigma<-1
 S<-matrix();X1<-rnorm(100,mean=theta0,sd=1);X2<-
rnorm(50,mean=theta1,sd=1);X<-c(X1,X2)
 S[1] <- (theta1-theta0)/sigma^2*(X[1]-(theta1+theta0)/2)
 for(i in 1:99) {
 S[i+1]<-S[i]+(theta1-theta0)/sigma^2*(X[i+1]-(theta1+theta0)/2)
 }
 for(i in 100:149) {
 S[i+1] <- S[i]+(theta1-theta0)/sigma^2*(X[i+1]-(theta1+theta0)/2)
 }
 par(mar=c(3,3,2,2))
 index<-1:150
 plot(index,X, type="l", lwd=1, pch=16,ylab="",xlab="")
 abline(v=101,lty=3)
 plot(S, type="l", lwd=1,ylab="",xlab="")
```

Figure 4.2.4

```
 set.seed(1)
 theta0 <-1;sigma<-1
 X<-rnorm(100,theta0,sigma);S<-matrix();S[1]<-0
 for(i in 1:99) {
 S[i+1] <-S[i]+(X[i]-theta0)
 }
 par(mar=c(3,3,2,2))
 plot(S,type="l",xlim=c(0,110),ylim=c(-10,30),xlab="",ylab="")
 segments(x0=100,y0=S[100],x1=110,y1=S[100])
 segments(x0=110,y0=S[100],x1=60,y1=-5)
 segments(x0=110,y0=S[100],x1=60,y1=S[100]+abs(S[100]+5))
```

Figure 4.2.5

```
 x <-0;y <-0;par(mar=c(3,3,2,2))
 plot(x,y,type="l",xlim=c(0,20),ylim=c(-80,70),ylab="",xlab="")
 segments(x0=0,y0=4,x1=20,y1=64);segments(x0=10,y0=1,x1=20,y1=31)
 segments(x0=10,y0=1,x1=20,y1=-31);segments(x0=0,y0=-10,x1=20,y1=-70)
 text(x=4,y=50,expression(paste("Accept"," ",H[1])))
 text(x=4,y=-55,expression(paste("Accept"," ",H[-1])))
 text(x=17,y=0,expression(paste("Accept"," ",H[0])))
```

Figure 4.3.2

```
 # ARL0
 theta0<-1;theta<-theta0;ARL<-matrix();h<-seq(0.5,5,0.2)
 par(mar=c(3,3,3,2))
 ARLf<-function(h) {
 w<-1/(theta-1.5)*(h+1.66+exp(-(2*theta-3)*(h+1.66))/(2*theta-3)-
1/(2*theta-3))
 return(w)
 }
 ARL<-1/(theta-1.5)*(h+1.66+exp(-(2*theta-3)*(h+1.66))/(2*theta-3)-
1/(2*theta-3))
 plot(h,ARL,type="l",main=expression(paste(ARL[0])))
 points(x=4,ARLf(4),pch=15);print(ARLf(4))
[1] 560.9773
 text(x=4.5,y=ARLf(4),cex=0.8,expression(paste("(x=4,y=561)")))
 # ARL1
 theta1<-2;theta<-theta1;ARL<-matrix();h<-seq(0.5,5,0.2)
 par(mar=c(3,3,3,2))
 ARLf<-function(h) {
 w<-1/(theta-1.5)*(h+1.66+exp(-(2*theta-3)*(h+1.66))/(2*theta-3)-
1/(2*theta-3))
 return(w)
 }
 ARL<-1/(theta-1.5)*(h+1.66+exp(-(2*theta-3)*(h+1.66))/(2*theta-3)-
1/(2*theta-3))
 plot(h,ARL,type="l",main=expression(paste(ARL[1])))
 points(x=4,ARLf(4),pch=15);print(ARLf(4))
[1] 9.326965
 text(x=4.5,y=ARLf(4),cex=0.8,expression(paste("(x=4,y=9.33)")))
```

(IV) R code for simulations

Program 1

```
 # Stein’s method
 set.seed(1)
 n <- 1000;n0 <- 10;mu <- 1;sigma <- 3;a <- 0.05;j <- 1
 t <- qt(1-a/2,df=n0-1);mat.Stein <- matrix(,6,5)
 Ts <- matrix()  # πίνακας με τα μεγέθη δειγμάτων T

 for(d in c(1, 0.75 ,0.5, 0.25, 0.15, 0.05)) {  # radius of c.i.
 s<-0;k <- qnorm(1-a/2,mean=0,sd=1)^2*sigma^2/d^2

  for(i in 1:n) {
    X1 <- rnorm(n0, mean=mu, sd=sigma)
    Ts[i] <- max(n0, floor(t^2*sd(X1)^2/d^2)+1)
    if(Ts[i]>n0) {
     X2 <- rnorm(Ts[i]-n0,mean=mu,sd=sigma) # choice of observ.
     X3 <- c(X1,X2) # πίνακας με το τελικό δείγμα
    }
    else {
     X3 <- X1
    }
    if(mu>=mean(X3)-d & mu<=mean(X3)+d) {
     s <- s+1
   }
  }
  mat.Stein [j,] <- c(k, mean(Ts), sd(Ts), mean(Ts)/k, s/n)
  j <- j+1
 }
 print(mat.Stein)
```

Program 2

```
# Purely Sequential Method
set.seed(1)
n <- 1000;n0 <- 10;mu <- 1;sigma <- 3;a <- 0.05;l <-1
Ta <- matrix() # matrix with T values
mat.Seq <- matrix(,6,5); # final table of algorithm
X <- matrix() # table with the final sample

for(d in c(1, 0.75 , 0.5, 0.25, 0.15, 0.05)) { # radius of c.i.
  k <- qnorm(1-a/2,mean=0,sd=1)^2*sigma^2/d^2
  s <- 0
 for(j in 1:n) {
   i <- n0
   X <- rnorm(n0, mean=mu, sd=sigma)
   k0 <- qnorm(1-a/2, mean=0, sd=1)^2*sd(X)^2/d^2
   while(i<k0) {  # βρόχος για εκτίμηση του Ta
     i <- i+1
     X[i] <- rnorm(1, mean=mu, sd=sigma)
     k0 <- qnorm(1-a/2, mean=0, sd=1)^2*sd(X)^2/d^2
   }
   Ta[j] <- i
   if(mu>=mean(X)-d & mu<=mean(X)+d) {
     s <- s+1
  }
 }
 mat.Seq[l,] <- c(k, mean(Ta), sd(Ta), mean(Ta)/k, s/n)
 l <- l+1
}
print(mat.Seq)
```

Program 3

```
# Healy’s Method
 set.seed(1);library(MASS)
 p=2;m=10;Mean=c(1,2);S=matrix(c(1,0.5,0.5,2),nrow=2,ncol=2)
;alpha=0.05;n=1000
 u<-p*(m-1)/(m-p)*qf(1-alpha, df1=p, df2=m-p);
mat.Healy <- matrix(,ncol=5,nrow=6);j<-1
 lambda_S <- max(eigen(S)$values)   # eigenvalue of Σ matrix

 for(d in c(1,0.75, 0.5, 0.25, 0.15, 0.05)) {   # radius of region
   C <- qchisq(1-alpha, df=p)*lambda_S/d^2
   sum=0;XTbar <- c();T <- c();X3<-c()
   for(i in 1:n) {

     X1 <- mvrnorm(m, mu=Mean, Sigma=S)
     Sn <- cov(X1)  # sample covariance matrix
     lambda <- max(eigen(Sn)$values)
     T[i] <- max(m,floor(u*lambda/d^2)+1)

     if(T[i]>m) {
       X2 <-  mvrnorm(T[i]-m, mu=Mean, Sigma=S)
       X3 <- rbind(X1,X2)
      } else {
       X3 <- X1
     }
     XTbar <- c(mean(X3[,1]),mean(X3[,2]))   # mean value vector
     if (t(XTbar-Mean)%*%(XTbar-Mean)<=d^2){   # Confidence region
       sum=sum+1
     }
   }
  mat.Healy[j,] <- c(C, mean(T), sd(T), mean(T)/C, sum/n)
  j <- j+1
 }

 print(mat.Healy)
```

Program 4

```
 # Srivastava’s Method
 set.seed(1)
 library(MASS)
 p=2;m=10;Mean=c(1,2);S=matrix(c(1,0.5,0.5,2),nrow=2,ncol=2);
alpha=0.05;n=1000
 u<-p*(m-1)/(m-p)*qf(1-alpha, df1=p, df2=m-p);mat.Sriva <-
matrix(,ncol=5,nrow=6);j<-1
 a<- qchisq(1-alpha, df=p);
 lambda_S <- max(eigen(S)$values)

 for(d in c(1,0.75, 0.5, 0.25, 0.15, 0.05)) {   # radius of region
   C <- qchisq(1-alpha, df=p)*lambda_S/d^2  # optimal sample
   sum=0;XTbar <- c();T <- c();X<-c()
   for(i in 1:n) {
     X <- mvrnorm(m, mu=Mean, Sigma=S)
     Sn <- cov(X)  # sample covariance matrix
     lambda_n <- max(eigen(Sn)$values)
     N <- m       # Αρχική τιμή του N
     while(N<a/d^2*lambda_n){
        N <- N+1
        Xnew <- mvrnorm(1, mu=Mean, Sigma=S)
        X <- rbind(X,Xnew)
        Sn <- cov(X)
        lambda_n <- max(eigen(Sn)$values)
     }
     T[i] <- N    # estimation of sample size
     XTbar <- c(mean(X[,1]),mean(X[,2]))    # mean vector
     if (t(XTbar-Mean)%*%(XTbar-Mean)<=d^2){    # confidence region
      sum=sum+1
     }
   }
  mat.Sriva[j,] <- c(C, mean(T), sd(T), mean(T)/C, sum/n)
  j <- j+1
 }
 print(mat.Sriva)
```

Program 5

```
set.seed(1)
sigma<-2;theta0<-1;theta1<-1.4;alpha<-0.05;n<-200;beta<-matrix();j<-1;N<-10000;
for(n in c(30,50,100,150,200,250,300)) {
beta[j]<-pnorm((theta0-theta1)/sqrt(sigma^2/n)+qnorm(1-alpha,mean=0,sd=1),mean=0,sd=1)
j<-j+1
}
matrixNP<-matrix(ncol=3,nrow=length(beta))  # final table N-P
matrixNP[,1]<-c(30,50,100,150,200,250,300)
matrixNP[,2]<-rep(alpha,length(beta));matrixNP[,3]<-beta
matrixSPRT<-matrix(nrow=length(beta),ncol=8)
for(j in 1:length(beta)) {
 A <- beta[j]/(1-alpha); B <- (1-beta[j])/alpha
 a<-log(A);b<-log(B)
  mat.H0<-matrix();mat.H1<-matrix()  # tables with T0 and T1
  k<-0;l<-0   # counters for Pr(I) and Pr(II)
  for(i in 1:N) {
    X0<-rnorm(1,mean=theta0,sd=sigma)
    Z0<-(theta1-theta0)/sigma^2*(X0-(theta0+theta1)/2)
    Sn.0<-Z0  # initial value of sum Sn under H0
    countH0<-1  # counter for mean value E[T0]
    X1<-rnorm(1,mean=theta1,sd=sigma)
    Z1<-(theta1-theta0)/sigma^2*(X1-(theta0+theta1)/2)
    Sn.1<-Z1  # initial value of Sn under H1
    countH1<-1  # counter for the mean value E[T1]
   while((a<Sn.0) & (Sn.0<b)) {  # loop for estimation Pr[I] and E[T0]
     X0<-rnorm(1,mean=theta0,sd=sigma)
     Z0<-(theta1-theta0)/sigma^2*(X0-(theta0+theta1)/2)
     Sn.0<-Sn.0+Z0
     countH0<-countH0+1
   }
   while((a<Sn.1) & (Sn.1<b)) {  # loop for estim. of Pr[II] and E[T1]
    X1<-rnorm(1,mean=theta1,sd=sigma)
     Z1<-(theta1-theta0)/sigma^2*(X1-(theta0+theta1)/2)
     Sn.1<-Sn.1+Z1
     countH1<-countH1+1



   }
  if(Sn.0>=b) {
    k<-k+1
  }
  if(Sn.1<=a) {
    l<-l+1
  }
  mat.H0[i]<-countH0
  mat.H1[i]<-countH1
 }
matrixSPRT[j,]<-c(mean(mat.H0),sd(mat.H0)/length(mat.H0)
,mean(mat.H1),sd(mat.H1)/length(mat.H1),k/N,l/N,k/N+l/N,alpha+beta[j])
}
totalmatrix<-cbind(matrixNP,matrixSPRT)  # final table
colnames(totalmatrix)<-c("n","alpha","beta_NP","Mean[T;H0]", "s.e(T0_bar)","Mean[T;H1]",
"s.e(T1_bar)","Pr[I]","Pr[II]", "Pr[I]+Pr[II]","a+b")
print(totalmatrix)
par(mfrow=c(1,2));par(mar=c(3,3,2,2))
```

```
 hist(mat.H0,breaks=40,xlab="",ylab="",main="");hist(mat.H1,breaks=40,xlab="",
ylab="",main="")
```

Program 6

```
set.seed(1) # Wald’s approximation
 alpha <- 0.05;beta <- 0.1;a <- log(beta/(1-alpha));b <- log((1-beta)/alpha)
 theta0 <- 1;theta1 <- 1.4;sigma <- 2;k <- (theta1-theta0)/sigma^2;theta.hat <-
(theta1+theta0)/2
 i <- 1;A.wald <- matrix(nrow=length(seq(1,1.4,0.04)), ncol=4, byrow=TRUE)
 Q <- c();ASN <-c();EZ1 <-c();t0<-c();EZ1.sq <- c()
 for(theta in seq(1,1.4,0.04)) {
  if (theta!=theta.hat) {
   t0[i] <- 2/(theta1-theta0)*(theta-theta.hat)
   Q[i] <- (exp(1)^(-t0[i]*b)-1 )/(exp(1)^(-t0[i]*b) - exp(1)^(-t0[i]*a))
   EZ1[i] <- (theta1-theta0)/sigma^2*(theta-theta.hat)
   ASN[i] <- (a*Q[i]+b*(1-Q[i]))/EZ1[i]
   } else {
   t0[i]<-0
   Q[i] <- b/(b-a)
   EZ1.sq[i] <- k^2*(sigma^2+theta^2)+k^2*theta.hat*(theta.hat-2*theta)
   ASN[i] <- (a^2*Q[i]+b^2*(1-Q[i]))/EZ1.sq[i]
   }
  A.wald[i,] <- c(t0[i],theta,Q[i], ASN[i])
  i <- i+1
 }
 j<-1
 ASN.sim <- c();oc.sim <-c(); # ASN and OC simulation

 for(theta.sim in seq(1,1.4,0.04)) {
  count <- c();q <- c();n <- 1000
  for(i in 1:n) {
   count[i] <- 1;q[i] <- 0
   X <- rnorm(1,mean=theta.sim, sd=sigma)
   Sx <- log(dnorm(X,mean=theta1,sd=sigma)/dnorm(X, mean=theta0, sd=sigma))
   while(Sx>a & Sx<b) {
    X1 <- rnorm(1, mean=theta.sim, sd=sigma)
    Sx <- Sx +log(dnorm(X1, mean=theta1, sd=sigma)/dnorm(X1, mean=theta0,
sd=sigma))
    count[i] <- count[i]+1
   }
   if(Sx<=a) {
    q[i] <- q[i]+1
   }
  }
  ASN.sim[j] <- mean(count)
  oc.sim[j] <- mean(q)
  j <- j+1
 }
 A.sim<-cbind(oc.sim,ASN.sim); A<-cbind(A.wald,A.sim); print(A) # Final Table
```

Program 7

```
 # ASN for the model AR(1) for three values of ρ
 set.seed(1);alpha <- 0.05;beta <- 0.1;theta0 <- 1;theta1 <- 1.4;sigma <-2
 a<- log(beta/(1-alpha));b<- log((1-beta)/alpha);n <- 1000
 mat.ASN.p <- matrix(,nrow=length(seq(1,1.4,0.04)),ncol=4,byrow=TRUE)
 mat.ASN.p[,1] <-seq(1,1.4,0.04)
 mat.OC.p <- matrix(,nrow=length(seq(1,1.4,0.04)),ncol=3,byrow=TRUE);k <-2
 for(p in c(0.1,0.5,0.9)) {
  j <-1;ASN <- c();OC <- c()
  for(theta in seq(1,1.4,0.04)) {
   sigmaE <- sqrt((1-p^2))*sigma; oc <- c();count <- c()
   for(i in 1:n) {
    oc[i] <- 0
    En <- rnorm(1,0,sigmaE)
    count[i] <- 2
    S <- (theta1-theta0)/((1+p)*sigma^2)*(((1-p)*theta+En) - (1-
p)*(theta1+theta0)/2 )
    while (S>a & S<b) {
     En <- rnorm(1,0,sigmaE)
     S <- S+(theta1-theta0)/((1+p)*sigma^2)*(((1-p)*theta+En) - (1-
p)*(theta1+theta0)/2 )
     count[i] <- count[i]+1
    }
    if(S<a) {
    oc[i] <- oc[i]+1
    }
   }
   ASN[j] <- mean(count)
   OC[j] <- mean(oc)
   j <- j+1
  }
 mat.ASN.p[,k] <- ASN;mat.OC.p[,k-1] <- OC
 k <- k+1
 }
 totalmatrix<-cbind(mat.ASN.p,mat.OC.p);print(totalmatrix) # Final Table
```

Program 8

```
 set.seed(1)
 theta0 <-1;theta1<-2;sigma<-1;n<-1000
 h<-4;l<-1;N<-seq(0.9,2.5,0.05)
 ARLsim<-matrix();ARLwald<-matrix()
 ARLsieg<-matrix();ARLmat<-matrix(,nrow=length(N),ncol=4);meanEZ<-matrix()
 for(theta in N) {
   k<-matrix()
  for(j in 1:n) {
   S<-matrix();Z<-matrix()
   X<-rnorm(1,mean=theta,sd=sigma)
   S[1]<-0;Z[1]<-(theta1-theta0)/sigma^2*(X-(theta1+theta0)/2)
   S[2]<-S[1]+Z[1]
   i<-2
   while(S[i]<h) {
    if(S[i]<=0) {
    S[i]<-0;
    }
   X<-rnorm(1,mean=theta,sd=sigma)
   Z[i]<-(theta1-theta0)/sigma^2*(X-(theta1+theta0)/2)
   S[i+1]<-S[i]+Z[i]
   i<-i+1
   }
  k[j]<-i
  }
  ARLsim[l] <- mean(k)
  ARLwald[l]<-ifelse(theta==1.5,h^2,1/(theta-1.5)*(h+exp(-(2*theta-3)*h)/(2*theta-
3)-1/(2*theta-3)))
  ARLsieg[l]<-ifelse(theta==1.5,(h+1.166)^2,1/(theta-1.5)*(h+1.166+exp(-(2*theta-
3)*(h+1.166))/(2*theta-3)-1/(2*theta-3)))
  meanEZ[l] <-(theta1-theta0)/sigma^2*(theta-(theta1+theta0)/2)
 l<-l+1
 }
 ARLmat[,1]<-meanEZ;ARLmat[,2]<-ARLsim;ARLmat[,3]<-ARLwald;ARLmat[,4]<-ARLsieg;
 par(mar=c(3,3,2,2))
 plot(N,ARLsieg,type="l",lty=4,xlab="",xlim=c(0.9,2.1),ylab="",main="")
 points(N,ARLsim,pch=4)
 lines(N,ARLwald,type="l",lty=1)
```

Bibliography

Greek

D. Antzoulakos (2010) *Statistical Quality Control*, University of Piraeus, Lecture Notes.

D. Cheliotis (2014) *Introduction to stochastic calculus*, Athens University, Lecture Notes.

Foreign

F. Anscombe (1952) Large sample theory of sequential estimation, *Proc. Camb. Phil. Soc.*, Vol.49, pp 600-607.

P. Armitage (1950) Sequential analysis with more than two alternative hypothesis and its relation to discriminant function analysis, *J.Roy.Statistist.Soc*. Ser. B, Vol 12, No 1, pp 137-144.

G.A. Barnard (1959) Control charts and stochastic processes, *J. Roy. Statistist. Soc*., Ser. B, Vol 21, pp 239-271.

M. Baseville & I.Nikiforov (1993), *Detection of abrupt changes:theory and applications*, Prentice Hall.

Y.Chow & H. Robbins (1965), On the asymptotic theory of fixed width sequential confidence intervals for the mean, *Ann. Math. Statist*., Vol 36, pp 457-462.

H.Dodge & H. Romig (1929), A method of sampling inspection, *Bell Syst. Tech. J.* Vol 8, pp 613-631.

T. Ferguson (2000) Optimal stopping and applications, unpublished manuscript
http://www.math.ucla.edu/~tom/Stopping/Contents.html.

Z. Govindarajulu (1974) *Sequential Statistical Procedures*, Academic Press Inc.

M. Ghosh (1991) *Handbook of Sequential Analysis*, CRC Press.

M.Ghosh & N.Mukhopadhyay (1997) *Sequential Estimation*, Wiley

P. Granjon (2012) The CUSUM algorithm, a small review, unpublished manuscript.

W.C. Healy (1956) Two sample procedures in simultaneous estimation, *Ann. Math. Statist.*, Vol 27, pp 687-702.

R. Hogg & A. Craig (1970) *Introduction to mathematical statistics*, Macmillan Co.

R.A. Johnson (2007) *Applied Multivariate Analysis*, Pearson

N. Johnson & F. Leone (1962) Cumulative sum control charts: mathematical principles applied to their construction and use, *Industrial Quality Control*, Vol 18, pp 15-21.

N. Johnson & S. Kotz (1970) Continuous univariate distributions-2, Wiley: New York, pp 102.

T.L. Lai (2001) Sequential analysis: some classical problems and new challenges, *Statistica Sinica*, 11(2):303-408.

E. Lehman (1951) Notes on the theory of estimation, *University of California Press: Berkeley.*

G. Lorden (1971) Procedures for reacting to a change in distribution , *Ann. Math. Statist.*, Vol 42, pp 1897-1908.

G. Moustakides (1986) Optimal procedures for detecting changes in distribution, *Annals Statistics*, Vol 14, pp 1379-1397.

N. Mukhopadhyay (2009) *Sequential methods and their applications*, Chapman and Hall/CRC..

J. Neyman & E. Pearson (1933) On the problem of the most efficient tests of statistical hypotheses, *Philosophical Transactions of the Royal Society of London*, Ser.A, pp 289-337.

R. Nowak (2010) Statistical Signal Processing, Lecture notes, ECE 830, Lecture 7
http://nowak.ece.wisc.edu/ece830/ece830_lecture7.pdf

E.S. Page (1954) Continuous inspection schemes, *Biometrika* Vol 41, No 1, pp 100-115.

M. Pollack (1985) Optimal detection of a change in distribution, *Annals Statistics*, Vol 13, pp 206-227.

W.A. Shewhart (1931) Economic control of manufactured product, Van Nostrand, New York.

A.N. Shiryaev (2007) *Optimal stopping rules*, Springer-Verlag, New York, NY.

A.N. Shiryaev & M.V. Zhitlukhin (2013*)* Optimal stopping problems for Brownian motion with drift and disorder: application to mathematical finance and engineering, Steklov Mathematical Institute, Moscow.

D. Siegmund (1985) Corrected diffusion approximation and their applications, Proccedings of the Berkeley Conference in honor of Jerzy Neyman and Jack Kiefer, Vol II, pp 599-617.

M.S. Srivastava (1967) On fixed width confidence bounds for regression parameters and mean vector, *J. Roy. Statist. Soc.*, Series B, Vol 29, pp 132-140.

C. Stein (1949) Some problems in sequential estimation. *Econometrica,* 17, pp 77-78.

V.V. Veeravalli & T. Banerjee (2012) *Quickest change detection*, ECE Department and Coordinated Science Laboratory.

A. Wald & J. Wolfowitz (1948) Optimum character of the sequential probability ratio test, *Ann. Math. Statist.* Vol 19, pp 326-339.

A. Wald (1947) *Sequential analysis*, Wiley, New York.

J. Walsh (2014) Elementary introduction to martingales, unpublished manuscript.
https://www.math.ubc.ca/~walsh/marts.pdf

M. Woodroofe (1977) Second order approximation for sequential point and interval estimation, *Ann. Statistist.*, Vol 5, pp 984-995.