F.W.N. Boesten

Simple hypothesis testing

Bachelor thesis, 20 juni 2013

Thesis advisor: Prof. Dr. Peter Grünwald

e-mail: frank.boesten@live.nl

Mathematisch Instituut, Universiteit Leiden

1. Introduction

Thousands of researchers all over the world make use of statistical methods to determine whether a hypothesis is correct or false. Shockingly, there are a lot of problems associated with the most common methods of hypothesis testing [1]. Often, the researchers who use these methods are not even aware of them, and therefore invalid conclusions are sometimes drawn. There are two solutions to this problem: First, statisticians could put more effort in explaining problems of the tests researchers use. Second, statisticians could develop better tests.

This thesis focuses on the second solution, and it focuses specifically on testing a simple versus a simple hypothesis. The most common method of testing simple versus simple hypothesis testing is the Neyman-Pearson test. This test is briefly explained in section 2.

Section 3 and 4 are about two tests that have been proposed as an alternative to the Neyman-Pearson test: the robust P-value method and the sequential likelihood ratio test. These tests are not commonly used in practice, because there are a number of problems associated with them, which are described at the end of each section. To examine whether these problems can be fixed somehow, the main question of this thesis is to examine the differences between these two tests. To examine this difference, the second test is generalized to the *generalized sequential likelihood ratio test*. This general test is described in chapter 5. The robust P-value test and the generalized sequential likelihood ratio test are very alike, but there still is an interesting difference, which is also described in chapter 5. The general test may be a practical alternative for the Neyman-Pearson test in testing simple versus simple hypotheses, but further research is needed to find out whether this is the case.

This thesis focuses primarily on testing simple versus simple hypotheses, but chapter 6 gives an introduction to testing complex hypotheses sequentially. Testing complex versus simple hypothesis is more common in practice, but also more difficult, therefore a detailed description of this type of testing lies beyond the scope of this thesis. Directions for future research are described in section 7.

In the remainder of this introduction, some basic statistical definitions are briefly explained. People who are familiar with statistics can choose to skip this part.

1.1. **Basic statistical definitions.** In simple hypothesis testing, we always compare two hypotheses. In most cases, one of these hypotheses is the *null hypothesis*, H_0. This is the hypothesis the scientist usually wants to disprove. It is a hypothesis which corresponds to a general or default position. This is usually translated in a hypothesis which assumes that there is no relationship between two measured phenomena or that a potential treatment has no effect.

The other hypothesis is the one the scientist usually wants to prove. We call this hypothesis the alternative hypothesis, H_1. This is usually translated in a hypothesis which assumes that there is a relationship between two measured phenomena or that a potential treatment does have an effect.

Informal definition 1. (simple hypothesis) *A simple hypothesis is a probability distribution for which all parameters of the distribution are specified.*

For example: the hypothesis that a coin is fair, e.g. the probability of throwing heads is equal to $\frac{1}{2}$, is a simple hypothesis, because the parameter of the binomial distribution, p, is specified to be $\frac{1}{2}$. Another example is of a simple hypothesis is "Our medicine does not have an effect on the blood pressure." We probably know how the blood pressure is distributed in a normal population, therefore we exactly know how the data should be distributed if the medicine does not have an effect.

Informal definition 2. (composite hypothesis) *A composite hypothesis is a probability distribution for which the parameters of the distribution are not completely specified.*

For example: the hypothesis that a coin is unfair, e.g. the probability of throwing heads is unequal to $\frac{1}{2}$, is a composite hypothesis. We do not know what the parameter of the binomial distribution, the chance of success p, is exactly. We only know it is unequal to $\frac{1}{2}$. Another example is: Our medicine has an effect on the blood pressure. This hypothesis does not tell us how the data are distributed; it only tells us they are not distributed as they would be in a normal population. Therefore, the parameters of this hypotheses are not completely specified.

Definition 1. (probability measure) *A probability measure is a measure P on a set Ω with the property that $P(\Omega) = 1$. We equip P with the Borel σ-algebra, which means that all sets $A \subset \Omega$ that we will be interested in are measurable, if $P(A)$ is well defined. This is usually written as a 'probability triple': (Ω, σ, P), where σ denotes the Borel σ-algebra in our case.*

Intuitively, everybody knows what a probability is. However, we need to define it mathematically to be able to prove theorems. Therefore this formal definition is needed. It 'measures' the probability of an event. For example: Let P be the probability measure under assumption of the hypothesis that a coin is fair. Let A denote the event that the coin turnes up head. Then the probability measure of the event A, $P(A)$, is equal to $\frac{1}{2}$.

From now on, we shall denote by P the probability measure under assumption of the simple hypothesis H_0, and by Q the probability measure under assumption of the simple hypothesis H_1.

Definition 2. (random variable) *A random variable is a measurable function from $\Omega \to \mathbf{R}$.*

Because we use the Borel σ-algebra, all random variables we are interested in will be well-defined.

Definition 3. (sample space) *Let χ denote the range of possible outcomes for one event. The sample space Ω is the set of possible sequences of data we can observe. In our setting, $\Omega = \chi^n$ or $\Omega = \chi^\infty$. If $\Omega = \chi^n$, and each element of Ω is written as $\omega = (x_1, x_2, \cdots, x_n)$, then we define the random variables $X_1, \cdots, X_n$: $X_1 = x_1, \cdots X_n = x_n$. If $\Omega = \chi^\infty$ and each element of Ω is written as $(x_1, x_2, \cdots)$, then we define the random variables $X_1, X_2, \cdots$ as $X_1 = x_1, X_2 = x_2, \cdots$.*

For example: A coin is tossed and 1 denotes the event that the coin turns up head, 0 denotes the event the coin turns up tails. Then the range for one individual outcome is $\chi = \{0,1\}$. Our sample space can either be finite or infinite. If it is finite, we have: $\Omega = \chi^n = \{0,1\}^n$. If it is infinite, we have: $\Omega = \chi^\infty = \{0,1\}^\infty$.

From now on, we assume that our X_i $(i \in S)$ are independent and identically distributed. This assumption is often made in statistical hypothesis testing.

Definition 4. (cylindrical point) *Let χ^∞ be an infinite dimensional sample space and denote the elements of χ^∞ by $\{x_i\}_{i\in\mathbf{N}}$. For any set of n given real numbers $a_1, \cdots, a_n$, we shall denote by $C(a_1, \cdots a_n)$ the subset of χ^∞ for which $x_1 = a_1, \cdots, x_n = a_n$. $C(a_1, \cdots a_n)$ is called a cylindrical point of order n.*

Of course, a researcher will never continue to sample on forever. While an infinite sequence perhaps could be observed, the researcher chooses to stop after a certain amount of observations. After he stops, the researcher knows the infinite sequence of observations is an element of the cylindric point of these initial observations.

Definition 5. (hypothesis test) *A test, T, is a function from the sample space $\Omega \to \{0,1\}$. If 0 is the outcome of this function, H_0 is accepted. Else it is rejected, and H_1 is accepted.*

Definition 6. (type I error) *A type I error occurs when a test rejects H_0 while H_0 is true.*

Definition 7. (significance level) *The probability of a type I error for a certain test is called the significance level. This is usually denoted by $\alpha = P(T = 1)$.*

The significance level for a certain test is chosen by the researcher who conducts the test. Its value depends on how certain you want to be that if the test rejects the null hypothesis, H_0 is indeed correct. Of course, researchers will want to choose α as low as possible. However, lowering α usually means you have to make more observations, which costs more money, or it means increasing β (see below), which is also undesirable. Researchers therefore always have to balance between the cost and the uncertainty of their test.

The value of α researchers usually use depends on the field of study. In the field of psychology for example, $\alpha = 0.05$ is usually used, while in physics, the value $\alpha = 0.001$ is more common.

Definition 8. (type II error) *A type II error occurs when a test accepts H_0 while H_0 is false. This probability is usually denoted by $\beta = Q(T = 0)$.*

Of course, researchers want β to be as low as possible, but again, lowering β costs money. It is beginning to become more and more common to use $\beta = 0.20$ [6].

Definition 9. (power) *The power of a test ($1 - \beta$) is the probability of rejecting H_0 when H_0 is false.*

The probability of a type I error (α) and the probability of a type II error (β) for a certain test, T, give information about how 'good' a test is. If a researchers can choose between two tests T and T' with type I and II errors respectively α, β and α^*, β^*, then if $\alpha < \alpha^*$ and $\beta < \beta^*$, a researcher will always choose the first test.

Definition 10. (stopping rule) *A stopping rule is a function $S : \cup_{i>0}\, \chi^i \to \{0,1\}$, where 0 denotes that the researcher stops making observations, and 1 denotes that the researcher makes another observation.*

Examples of stopping rules are: ″Stop when the head of the coin has ended upwards 4 times.″ or ″Stop when 100 observations have been made.″

Informal definition 3. (optional stopping) *Changing the stopping rule while making observations is called optional stopping.*

A test is sensitive to optional stopping if the statistical analysis is no longer valid when the stopping rule is changed. It is insensitive to optional stopping when it does not matter whether we change our stopping rule while making observations. Of course, it is better if a test is insensitive to optional stopping, but usually this means you have to make more observations, which costs more money.

2. Neyman-Pearson hypothesis test

This most common method of hypothesis testing is based on the work of Fisher, Neyman and Pearson and dates from the beginning of the 20th century. It is widely used in all sorts of research. It is also the method which is taught to high-school students and millions of social science students worldwide attending courses in statistics.

An important property of the test is that we have to determine our entire sampling plan in advance. This means that beforehand, we define a stopping rule, and we cannot change it while making observations. The test is therefore sensitive to optional stopping. After observing the data, the likelihood ratio is calculated, and the test tells you whether you should accept or reject H_0.

2.1. **test design.** We fix our stopping rule and significance level α in advance. We denote by x the data which is observed after applying the stopping rule. The test is defined as follows:

(1) If $P(x) = 0$, reject H_0 with type I error probability equal to zero
(2) If $Q(x) = 0$, accept H_0 with type II error probability equal to zero
(3) If $\frac{P(x)}{Q(x)} > c$, accept H_0
(4) If $\frac{P(x)}{Q(x)} \leq c$, accept H_1

c is called the rejection constant, and is determined such that the type I error of the test is less than α. The correct value of c depends on the distribution of P.

This test has an interesting property regarding its power, which is called the Neyman-Pearson lemma:

Theorem 1. *Suppose that H_0 and H_1 are two simple hypotheses and that the test that rejects H_0 whenever the likelihood ratio is less than c and significance level α. Then* any other *test for which the significance level is less than or equal to α has power less than or equal to that of the likelihood ratio test.*

Proof of Theorem 1. Denote the elements of Ω by x and the random variable defined on this sample space by X. A hypothesis test amounts to using a decision function $d(x)$, where $d(x) = 0$ if H_0 is accepted and $d(x) = 1$ if H_0 is rejected.

Since $d(X)$ is a Bernoulli random variable, $E(d(X)) = p$ (note that the small p denotes something different than the large P). The significance level of the test is thus $\alpha = P(d(X) = 1) = E_0(d(X))$, and the power is $Q(d(X) = 0) = E_1(d(X))$. Here E_0 and E_1 respectively denote the expectation under assumption of H_0 and H_1.

Now let $d(x)$ correspond to the likelihood ratio test: $d(x) = 1$ if $P(x) < cQ(x)$ and $E_0(d(x)) = \alpha$. Let $d^*(x)$ be the decision function of another test satisfying $E_0(d^*(X)) \leq E_0(d(X)) = \alpha$. We will show that $E_1(d^*(X)) \leq E_1(d(X))$, which proves the theorem.

The following inequality holds:

$$d^*(x)\,[cQ(x) - P(x)] \leq d(x)\,[cQ(x) - P(x)]\,,$$

because if $d(x) = 1$, $cQ(x) - P(x) > 0$ and if $d(x) = 0$, $cQ(x) - P(x) \leq 0$. Now integrating (or summing) both sides of the inequality above with respect to x gives:

$$cE_1(d^*(X)) - E_0(d^*(X)) \leq cE_1(d(X)) - E_0(d(X))$$

and thus

$$E_0(d(X)) - E_0(d^*(X)) \leq c\left[E_1(d(X)) - E_1(d^*(X))\right].$$

And hence we have proved that $E_1(d^*(X)) \leq E_1(d(X))$, as the left-hand side is nonnegative by assumption.

2.2. **Example for normally distributed data.** Consider a psychologist who wants to conduct an experiment to find out whether students from the faculty of mathematics are on average highly gifted (the average IQ is 130) or as gifted as the general population. To do this, he takes a sample of N math students and measures their IQ. It is well known that the IQ in the general population is normally distributed with mean 100 and standard deviation 15. The psychologist assumes the standard deviation of the IQ among math students is also 15. He chooses $\alpha = 0.05$ as significance level for his test.

We then have the following mathematical model: $(X_1, \cdots, X_N) = X$ is a random sample from a normal distribution having variance 15^2. We consider the following simple hypothesis: $H_0 : \mu = 100 = \mu_0$ and $H_1 : \mu = 130 = \mu_1$. To evaluate the test, we first need to determine our rejection constant c. Therefore, we calculate the likelihood ratio:

$$\frac{P(X)}{Q(X)} = \frac{\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(X_i - \mu_0)^2)}{\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(X_i - \mu_0)^2)} = \exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}((X_i - \mu_0)^2 - (X_i - \mu_1)^2)$$

The only part dependent on the data is the sum $\sum_{i=1}^{N}((X_i - \mu_0)^2 - (X_i - \mu_1)^2$. Expanding the squares reduces this to:

$$2N\overline{X}(\mu_0 - \mu_1) + N\mu_0^2 - N\mu_1^2,$$

where only $2N\overline{X}(\mu_0 - \mu_1)$ is dependent on the data. As $\mu_0 - \mu_1 < 0$, the likelihood ratio is small if $\overline{X}$ is large. We want to find a x_0 such that $P(\overline{X} > x_0) = \alpha$ if H_0 is true.

Under assumption of H_0, we know that $\overline{X}$ is distributed with mean μ_0 and standard deviation $\frac{\sigma}{\sqrt{n}}$. Since

$$P(\overline{X} > x_0) = P(\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} > \frac{x_0 - \mu_0}{\sigma/\sqrt{n}})$$

we know that $\frac{x_0 - \mu_0}{\sigma/\sqrt{n}} = z(\alpha)$, and hence we need to choose $x_0 = \frac{z(\alpha)\sigma}{\sqrt{n}} + \mu_0$. Filling in this expression in our likelihood ratio gives us the value of c.

Note that for $\lim n \to \infty$, the value of x_0 approaches μ_0 arbitrarily close. This means that for large sample sizes, the null hypothesis is rejected for very small deviations from 100 for the mean of the sample size. This is quite remarkable, as this would mean that if the sample size is large enough, the test would reject H_0 if the mean of the IQ of the math students is closer to 100 than 130. For example, it is possible that the test rejects H_0 while a mean of 101 is measured. This is quite bizarre.

2.3. **Advantages and disadvantages.** The example shows that for normally distributed data, H_0 can be rejected while the data actually favors H_0. It can easily be shown that this is also true for data which are not normally distributed. A researcher using this method of hypothesis testing therefore should be very careful in interpreting the outcome of the test. However, many researchers who use this method of hypothesis testing, are not aware of this property. Research conducted by Freeman (1993) [2] shows that of a sample of doctors, dentists and medical students, only 20% of them are aware of this disadvantage. This property is therefore a huge disadvantage of this test.

Another large disadvantage, is that in real life, it is often impossible stick to your predefined stopping rule. First of all, if you conduct an experiment on for example 100 persons, you do not know whether they are all going to show up on the experiment, whether mistakes will be made which make the observations invalid, etc. Researchers will therefore often 'cheat': they alter their stopping rule, while it makes the statistical analysis invalid. Of course, this is not desirable.

Third, it can occur that after examining 100 observations, it turns out the evidence is not conclusive, but does give a strong indication that for example your medicine does have an effect. In Neyman-Pearson hypothesis testing, you are not allowed to conduct a number of additional observations to find out whether conclusive evidence can be found. If you would like to do more research, you have to start all over again and cannot use the data that you have already collected, which is a huge waste of money.

The above disadvantages and a number of other disadvantages of this method are described in more detail by van der Pas [1].

3. The robust P-value test

The robust P-value method of simple hypothesis testing has been suggested as a new form of hypothesis testing, as it has a number of attractive properties with regard to the problems which arise with the method of Neyman-Pearson hypothesis testing. It dates from the end of the 20th century and is not widely known or used among researchers. It is almost equivalent to the likelihood ratio test: we again use the likelihood ratio of the distributions under assumption of H_0 and H_1.

There are however two major differences with the Neyman-Pearson hypothesis test. First, we do not have to determine our stopping rule in advance. The test is thus insensitive to optional stopping. Second, the value of the rejection constant is independent of the sample size.

3.1. **Test Design.** When making use of the Robust P-value method, you do not know in advance what your sample size is. We denote this undetermined value of the sample size n (which is thus a random variable). Let $x^m = x_1 \cdots x_n$ be a sequence of m outcomes from a sample space Ω. Choose a fixed value $\alpha^* \in (0, 1]$ (α^* is the rejection constant). Then after making m observations:

(1) If $P(x^m) = 0$, reject H_0 with type I error probability equal to zero
(2) If $Q(x^m) = 0$, accept H_0 with type II error probability equal to zero
(3) If $\frac{P(x^m)}{Q(x^m)} < \alpha^*$, reject H_0
(4) If $\frac{P(x^m)}{Q(x^m)} \geq \alpha^*$, continue sampling or stop doing observations (and accept H_0)

Van der pas (2010) [1] has proved theorem 2, which relates the type I error to α^* for a fixed number of observations. This proof relies on Markov's inequality:

Corollary 1. (Markov's inequality) *Let Y be a random variable with $P(Y \geq 0) = 1$ for which $E(Y)$ exists and let $c > 0$, then: $P(Y \geq c) \leq \frac{E(Y)}{c}$.*

Proof of Corollary 1. This is the proof for the discrete case, but the continuous case is entirely analogous.

$$E(Y) = \sum_y yP(y) = \sum_{y<c} yP(y) + \sum_{y \geq c} yP(y)$$

Because it holds that $P(Y \geq 0) = 1$, all terms in both sums are nonnegative. Thus:

$$E(Y) \geq \sum_{y \geq c} yP(y) \geq \sum_{y \geq c} cP(y) = cP(Y \geq c)$$

.

Theorem 2. *Assume for all $x^n \in \chi^n$ that $Q(x^n) \neq 0$ and choose $\alpha^* \in (0, 1]$. Further, let n be a fixed number. If H_0 is rejected by the robust P-value test, the type I error probability is less than α^*.*

Proof of Theorem 2. Let $Y = \frac{P(X^n)}{Q(X^n)}$. Note that Y cannot take on negative values, because both Q and P can only take on values between zero and one. As $Q(x^n) \neq 0$ for all $x^n \in \chi^n$, we know the expected value of Y exists. Therefore, we can apply Markov's inequality on Y:

$$P(\frac{P(X^n)}{Q(X^n)} \leq \alpha^*) = P(\frac{Q(X^n)}{P(X^n)} \geq \frac{1}{\alpha^*}) \leq \frac{E(\frac{Q(X^n)}{P(X^n)})}{\frac{1}{\alpha^*}} = \alpha^* \sum_{x^n} P(x^n) \cdot \frac{Q(x^n)}{P(x^n)} = \alpha^* \sum_{x^n} Q(x^n) = \alpha$$

This theorem tells us that if the sample size n is fixed in advance, you know the type I error is smaller than α^*. However, we do not want do fix our n in advance, because the robust P-value method is supposed to be insensitive to optional stopping.

It is possible to proof that the type I error is smaller than α^* if you do not fix your sample size in advance. One way of proving this, is by making use of Martingales:

Definition 11. (discrete-time martingale) *A discrete-time martingale is a sequence of random variables $X_1, X_2, \cdots$ that satisfies for any time n:*

$E(|X_n|) < \infty$
$E(X_{n+1}|X_1 = x_1, \cdots, X_n = x_n) = x_n$

The setting of making observations without knowing the sample size in advance, can be translated to a Martingale as follows: define $M_n = \frac{Q(X_1,\cdots,X_n)}{P(X_1,\cdots,X_n)}$. $\{M_n\}_{n\in\mathbf{N}}$ is a Martingale where the n'th element of the Martingale denotes the inverse of the likelihood ratio after making n observations. (Why we need to take the inverse of the likelihood becomes clear later.) Hence, if after making n observations we have that $\frac{1}{M_n} < \alpha^*$, the null hypothesis is rejected.

We now prove that $\{M_n\}_{n\in\mathbf{N}}$ is indeed a Martingale.

For the first condition, we need to assume that for all $\omega \in \cup_{i\in\mathbf{N}} \chi^i : P(\omega) \neq 0, Q(\omega) \neq 0$. In practice, it rarely occurs that a researcher formulates a hypothesis which places a zero probability on any event in the sample space. Hence it is not a problem that we have to make this assumption.

Denote by $E_P(X)$ the expected value under assumption of the null hypothesis. We want to prove that $E_P(M_{n+1}|M_1 = m_1, \cdots, M_n = m_n) = m_n$. Because the variables $X_1, X_2, \cdots$ are identically and independently distributed, the value of M_{n+1} only depends on M_n. Hence we know that $E_P(M_{n+1}|M_1 = m_1, \cdots, M_n = m_n) = E_P(M_{n+1}|M_n = m_n)$. Now we have:

$$E_P(M_{n+1}|M_n = m_n) = \sum_{x^n \in \chi^n} \frac{Q(x^n)}{P(x^n)} P(x^n) m_n = m_n \sum_{x^n \in \chi^n} Q(x^n) = m_n$$

Hence we have proven that $\{M_n\}_{n\in\mathbf{N}}$ is a Martingale. Note that we needed that M_n is the inverse of the likelihood after n observations for this last step.

Now we can make use of theorems that apply to Martingales. One of these is Doob's optional sampling theorem:

Theorem 3. (Doob's optional sampling theorem) *Let $\{M_s\}_{s\in\mathbf{N}}$ be a Martingale and define $M^*_\infty = \sup_{s\in\mathbf{N}} M_s$. Then for each $\delta \in [0,1]$: $P(\frac{1}{M^*_\infty} \leq \delta) \leq \delta$*

Proof of Theorem 3. The proof of this theorem lies beyond the scope of this thesis. For a proof of the theorem, see reference 3.

It is now clear why we needed to introduce Martingales. The theorem tells us that:

$$\forall \alpha^* \in [0,1] : P(\frac{1}{\sup_n \frac{Q(X^n)}{P(X^n)}} \leq \alpha^*) \leq \alpha^*,$$

which is equivalent to:

$$\forall \alpha^* \in [0,1] : P(\sup_n \frac{P(X^n)}{Q(X^n)} \leq \alpha^*) \leq \alpha^*.$$

This last statements proves that the robust P-value test is insensitive to optional stopping, as it proves that regardless of what stopping rule you use, you know that the chance on a type I error is smaller than α^*

3.2. **Example for normally distributed data.** Consider the same example which is described in section 2.2 (the psychologist who wants to determine whether math students are highly gifted). The likelihood is determined in exactly the same manner. The test procedure is therefore as follows. If

$$\exp(-\frac{1}{2\sigma^2}(2n\overline{X}(\mu_0 - \mu_1) + n\mu_0^2 - n\mu_1^2)) < \alpha,$$

reject H_0. Else, continue sampling or stop doing observations and accept H_0.

We again examine what happens when the sample size goes to infinity. We therefore rewrite the above expression:

$$\exp(-\frac{n}{2\sigma^2}(2\overline{X}(\mu_0 - \mu_1) + \mu_0^2 - \mu_1^2)) < \alpha.$$

We see that if $(2\overline{X}(\mu_0 - \mu_1) + \mu_0^2 - \mu_1^2) < 0$, the expression goes to zero for large n (and H_0 is accepted), and if it is smaller than zero, the expression goes to infinity for large n. Hence we are interested in the case where the expression is zero. Equating the expression to zero gives:

$$\overline{X} = \frac{1}{2}(\mu_0 + \mu_1),$$

Intuitively this sounds correct. For large sample sizes, the test simply accepts the hypothesis to which the mean of the data lies the closest.

3.3. **Advantages and disadvantages.** As the rejection constant is independent of the sample size, the test is insensitive to optional stopping. This means we do not have to determine our sample size beforehand, and can for example decide to make additional observations if our first observations do not give us decisive information. This is one major advantage of the robust P-value method compared to the Neyman-Pearson test.

A second advantage is that for large sample sizes, the test does what it should do: it favors the hypothesis to which the mean of the observations lies the closest.

Why don't researchers use this test instead of the Neyman-Pearson hypothesis test? Theoretically, the test has huge advantages over the Neyman-Pearson test.

One answer to this question is that conducting a robust P-value method costs a lot more money than a Neyman-Pearson test.

Typically, a researcher receives funds to make for example 100 observations. After making these observations, the researcher conducts a statistical test. If we choose for example $\alpha = 0.05$, the data likelihood of the data under assumption of H_1 must be 20 times greater than under assumption of H_0. This is a lot sterner than if we do the same with the Neyman-Pearson test. A researcher would therefore only use the robust P-value method, when he wants to be able to change his stopping rule while making the observations. And as this comes at a great cost, a researcher will not do this very often.

Another disadvantage of this test, is that it is not possible to say anything about your type II error in advance. As a researcher, before spending a lot of money to collect data, you want to know whether your investment is going to be worthwhile. If your β is large, the probability that you waste a lot of money is large. Therefore, this is also a large disadvantage.

4. The sequential likelihood ratio test

The sequential likelihood ratio test has been proposed by Wald in 1945 [4] as another alternative to regular hypothesis testing. Again, the sample size n is considered a random variable instead of a constant, and again we use the likelihood ratio. We make only one observation at a time and after each observation, the sequential test tells us whether we should make another observation, reject H_0 or accept H_0.

The goal for Wald was to construct a test which needs less data than the Neyman-Pearson hypothesis test. When conducting a Neyman-Pearson hypothesis test, it is for example possible that after doing 50 of the 100 observations, you already know which hypothesis is going to be preferred by the data, but you have to make the other 50 additional observations, as your statistical analysis is invalid otherwise. He therefore constructed a test which has the possibility to stop after each observation.

4.1. **Test Design.** Let $(x_1, \cdots, x_m) = x^m$ be a sequence of m observations. Let p, q respectively be the density function of an individual observation X_i under assumption of P, Q. We then have: $P(x^m) = \prod_{i=1}^m p(x_i)$, $Q(x^m) = \prod_{i=1}^m q(x_i)$ The test is then defined as follows. After making m observations:

(1) If $P(x^m) = 0$, reject H_0 with type I error probability equal to zero
(2) If $Q(x^m) = 0$, accept H_0 with type II error probability equal to zero
(3) If $\frac{Q(x^m)}{P(x^m)} \geq A$, accept H_1
(4) If $\frac{Q(x^m)}{P(x^m)} \leq B$, accept H_0
(5) If $B < \frac{Q(x^m)}{P(x^m)} < A$, make an additional observation

Theorem 4. *If $A, B \neq 0$, the sequential likelihood ratio test terminates with probability* 1.

Proof of Theorem 4. After making n observations, we calculate the likelihood ratio:

$$\frac{Q(x_1, \cdots, x_n)}{P(x_1, \cdots, x_n)} = \frac{q(x_1) \times q(x_2) \cdots q(x_n)}{p(x_1) \times p(x_2) \cdots p(x_n)},$$

We continue to make observations when:

$$B < \frac{q(x_1) \times q(x_2) \cdots q(x_n)}{p(x_1) \times p(x_2) \cdots p(x_n)} < A.$$

We transform the equation on log-scale, and with $\log(\frac{q(x_i)}{p(x_i)}) = z_i$ and get:

$$\log(B) < \sum_{i=1}^{n} z_i < \log(A).$$

Note that $\{Z_i\}$ $i = 1, \cdots, n$ is again a sequence of independent identically distributed random variables. Denote by m the smallest integer for which either $\sum_{i=1}^m Z_i > \log(A)$ or $\sum_{i=1}^m Z_i < \log(B)$. If we prove that $P(m = \infty) = 0$, then we have proved that the test terminates with probability 1.

Let $c = |\log(A)| + |\log(B)|$. If we prove that the probability is zero that $(\sum_{i=1}^k Z_i)^2 < c^2$ holds for all $k \in \mathbf{N}$, then we have proved that $P(m = \infty) = 0$.

As we know that $\{Z_i\}_{i \in \mathbf{N}}$ is a sequence of independent random variables having the same distribution, we know that the expected value of $(\sum_{i=1}^{j} Z_i)^2$ converges to ∞ as $j \to \infty$. Hence we know there exists a positive integer r such that the expected value of $(\sum_{i=1}^{j} Z_i)^2$ is bigger than $4c^2$. For this value r, we know that $P[(\sum_{i=1}^{r} Z_i)^2 \geq 4c^2] = \epsilon$ with $\epsilon \in (0,1]$.

Note that if we prove that the chance is zero that $(\sum_{i=1}^{lr} Z_i)^2 < c^2$ for all $l \in \mathbf{N}$, then we have proved that the chance is zero that $(\sum_{i=1}^{k} z_i)^2 < c^2$ holds for all $k \in \mathbf{N}$. We know that:

$$P[(\sum_{i=1}^{lr} Z_i)^2 < c^2 \text{ for all } l \in \mathbf{N})] = \prod_{j=1}^{\infty} P[(\sum_{i=1}^{lr} Z_i)^2 < c^2 \text{ holds for } l = j \mid \text{ it holds for all } l \in \{1, \cdots j-1\}]$$

We first examine a single term of the right-side expression, therefore let $j \in \mathbf{N}$. We know that $(\sum_{i=1}^{(l-1)r} Z_i)^2 < c^2$. We also know that $P[(\sum_{i=1+(l-1)r}^{lr} Z_i)^2 \geq 4c^2] = \epsilon$. Consider $(\sum_{i=1}^{lr} Z_i)^2 = (\sum_{i=1+(l-1)r}^{lr} Z_i + \sum_{i=1}^{(l-1)r} Z_i)^2$. We know that $\sum_{i=1}^{(l-1)r} Z_i \in (-c, c)$ and we know with probability ϵ that $(\sum_{i=1+(l-1)r}^{lr} Z_i) \in (\infty, -2c] \cup [2c, \infty)$. For every $a \in (-c, c)$ and every $b \in (\infty, -2c] \cup [2c, \infty)$ it holds that $|a + b| > c$. Hence, we know with probability at least ϵ that $(\sum_{i=1}^{lr} Z_i)^2 \geq c^2$.

Now we know that $P[(\sum_{i=1}^{lr} Z_i)^2 < c^2 \text{ holds for } l = j \mid \text{ it holds for all } l \in \{1, \cdots j-1\}] \leq 1 - \epsilon$ for every $j \in \mathbf{N}$, and hence we can finally conclude:

$$P(m = \infty) \leq \prod_{j=1}^{\infty} 1 - \epsilon = 0$$

Theorem 5. *If we take $A = \frac{1-\beta^*}{\alpha^*}$ and $B = \frac{\beta^*}{1-\alpha^*}$, then we know that the following inequality holds for the type I and type II error: $\alpha + \beta \leq \alpha^* + \beta^*$.*

Proof of Theorem 5. Let $\{x_m\}_{m \in \mathbf{N}}$ be an infinite sequence of observations. Our sample space is the set of all infinite sequences. When for a cylindric point $C(x_1, \cdots, x_n)$ we have $\frac{Q(a_1,\cdots,a_n)}{P(a_1,\cdots,a_n)} \geq A$, and there exists no $m < n$ such that $\frac{Q(a_1,\cdots,a_m)}{P(a_1,\cdots,a_m)} \geq A$ or $\frac{Q(a_1,\cdots,a_m)}{P(a_1,\cdots,a_m)} \leq B$, we will call it a type 1 cylindric point. These are the sequences of points for which H_1 is accepted. Similarly, we define cylindric points of type 0.

Let S_0, S_1 resp. be the set of all cylindric points of type 0,1. Theorem 4 tells us that $P(S_0 \cup S_1) = 1$ and $Q(S_0 \cup S_1) = 1$, which simply means that under assumption of both hypotheses the test will terminate.

Since for each sample $(x_1, \cdots, x_n)$ for which $c(x_1, \cdots, x_n)$ is an element of S_1 the inequality $\frac{Q(x_1,\cdots,x_n)}{P(x_1,\cdots,x_n)} \geq A$ holds, we know that $\frac{Q(S_1)}{P(S_1)} \geq A$. Similarly $\frac{Q(S_0)}{P(S_0)} \leq B$.

Of course $P(S_1) = \alpha$ and $Q(S_0) = \beta$, and as S_0 and S_1 are disjoint, it follows that $P(S_0) = 1 - \alpha$ and $Q(S_1) = 1 - \beta$. Substituting this in the inequalities gives:

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

If we take $A = \frac{1-\beta^*}{\alpha^*}$ and $B = \frac{\beta^*}{1-\alpha^*}$, divide the first equation by $(1-\beta)(1-\beta^*)$, multiply the second by $(1-\alpha^*)(1-\alpha)$, and add the resulting inequalities, we get:

$$\alpha + \beta \leq \alpha^* + \beta^*.$$

Theorem 5 only tells us that if we take $A = \frac{1-\beta^*}{\alpha^*}$ and $B = \frac{\beta^*}{1-\alpha^*}$, we then have $\alpha + \beta \leq \alpha^* + \beta^*$. This only tells us that we know that either $\alpha \leq \alpha^*$ or $\beta \leq \beta^*$. However, in his paper Wald writes about this choice of A and B: *"The probability α on an error of the first kind cannot exceed α^* and the probability β of an error of the second kind cannot exceed β, except by a very small quantity which can be neglected for practical purposes"* [4, p133]. The mathematical proof of this statement however lies beyond the scope of this thesis. Therefore we simply assume that for all practical purposes, by this choice of A and B, we know that $\alpha \leq \alpha^*$ and $\beta \leq \beta^*$.

For Wald, the goal of this test was to save the number of observations. In his paper, he states the following about the number of observations necessary to conduct this test: *"The sequential probability ratio test frequently results in a saving of about 50% in the number of observations as compared with the current most powerful test."* [4, p119], where with 'the current most powerful test', he refers to the Neyman-Pearson test. Wald shows the above statement is true for testing the mean of a normally distributed variate, for values of $\alpha \in [0.01, 0.05]$, $\beta \in [0.01, 0.05]$, under assumption of H_0 and H_1.

A couple of years after his first paper on sequential likelihood ratio tests in 1945, Wald an Wolfowitz also proved the Wald-Wolfowitz theorem. It is the equivalent of the Neyman-Pearson lemma, but then for the sequential likelihood ratio test:

Theorem 6. (Wald-Wolfowitz theorem) *Let S_0 be any sequential probability ratio test for deciding between two simple alternatives H_0 and H_1, and S_1 another test for the same purpose. We define $(i, j = 0, 1) : \alpha_i(S_j) =$ probability, under S_j, of rejecting H_i when it is true; $E_i^j(N) =$ expected number of observations to reach a decision under test S_j when the hypothesis H_i is true. (It is assumed E_i^1 exists.) Then if $\alpha_i(S_1) \leq \alpha_i(S_0)$, it follows $E_i^0(N) \leq E_i^1(N)$.*

Proof of Theorem 6. The proof of this theorem lies beyond the scope of this thesis. For a proof of the theorem, see reference 5.

The theorem states that of all sequential tests with the same power the sequential probability ratio test requires on the average the fewest observations.

4.2. **Example for normally distributed data.** We again consider the same example as we examined in sections 2.2 and 3.2. The likelihood ratio is again given by:

$$\exp(-\frac{1}{2\sigma^2}(2n\overline{X}(\mu_1 - \mu_0) + n\mu_1 - n\mu_0))$$

If we let $\beta = 0.20$ (which is a value regularly used in sciences,[6]), we will reject our null hypothesis when the likelihood ratio is greater than $\frac{1-\beta}{\alpha} = 16$. We will accept our null hypothesis when the likelihood ratio is smaller than $\frac{\beta}{1-\alpha} = \frac{4}{19}$. Else, we continue sampling.

This means we reject H_0 when the data is 16 times more likely to occur under assumption of H_1 than under H_0. We accept H_0 when it is about 5 times more likely than the data occurs under assumption of H_0 than under H_1.

4.3. **Advantages and disadvantages.** As the rejection constants are independent of the sample size, the test shows a lot of similarity with the robust P-value test. It is almost as stern as the robust P-value test, as it also demands that the likelihood of the data under assumption of H_1 must be a lot greater than the likelihood of the data under assumption of H_0. In the case of $\alpha = 0.05$ and $\beta = 0.20$, the null hypothesis is rejected when the data is 16 times more likely to occur under assumption of H_1, compared to 20 times in the robust P-value test for $\alpha = 0.05$.

To compare this with the Neyman-Pearson test, we consider the example already discussed in section 2.2. If we take a sample size of 100, with $\alpha = 0.05$, $\sigma = 15$, $\mu_0 = 100$, then H_0 is rejected if $\overline{X} > x_0 = 102.5$. If we insert this value in the expression for the likelihood ratio, we get: $\frac{P(X)}{Q(X)} = e^{2900} \approx 10^{1260}$. This means that it is possible that H_0 is rejected while the likelihood ratio greatly favors H_0. We now see why the robust P-value and Wald's test are a lot more stern than the Neyman-Pearson test, as they both only reject a hypothesis if the likelihood favors that hypothesis.

One of the reasons why the robust P-value test is not used in practice, is that it needs a lot more data (and thus costs a lot more money) than the Neyman-Pearson test. However, Wald has shown that the Sequential Likelihood Ratio Test frequently results in a 50% saving compared to the Neyman-Pearson test. Therefore, in theory the Likelihood Ratio Test is superior to the robust P-value test and Neyman-Pearson test, as it saves observations and is stern.

Of course the major disadvantage is that you have to make your observations one by one. In practice this is often not possible. Also, although the test in general reduces the number of observations, it does not tell you how many observations you are going to have to make, which is sometimes difficult to deal with in practice.

5. Generalized Sequential Likelihood Ratio Test

In section 3 we saw that in theory, the robust P-value test is an improvement of the Neyman-Pearson test, as it is sterner. However, the test generally needs too many data to carry out the test. In section 4 we saw that the sequential likelihood ratio test by Wald is sterner, but also saves a lot of observations compared to the Neyman-Pearson test. It is therefore superior to both other tests, but it has one major disadvantage: we have to make our observations one by one.

In this section, we generalize the sequential likelihood ratio test to a test where we do not have to make our observations one by one. Hopefully, this results in a test which is better than all previously named tests. We will see that the robust P-value test and generalized sequential likelihood ratio test are very alike.

5.1. **Test Design.** Beforehand, we decide how many observations we want to make, which we denote with n_1. If we put the n_1 data points in the test, it either terminates or tells us to make more observations. If it does not terminate, we decide to make n_2 additional observations, etc. We therefore have a possibly infinite sequence $(n_1, n_2, \cdots)$. Denote the i'th set of data points by $(x_{i_1}, \cdots, x_{i_{n_i}}) = X^i$. After doing i sets of observations:

(1) If $P(X^1, \cdots, X^i) = 0$, reject H_0 with type I error probability equal to zero
(2) If $Q(X^1, \cdots, X^i) = 0$, accept H_0 with type II error probability equal to zero
(3) If $\frac{Q(X^1,\cdots,X^i)}{P(X^1,\cdots,X^i)} \geq A$, accept H_1
(4) If $\frac{Q(X^1,\cdots,X^i)}{P(X^1,\cdots,X^i)} \leq B$, accept H_0
(5) If $B < \frac{Q(X^1,\cdots,X^i)}{P(X^1,\cdots,X^i)} < A$, choose how many additional observations you want to make

Theorem 7. *If $A, B \neq 0$, the generalized sequential likelihood ratio test terminates with probability* 1.

Proof of Theorem 7. After each i'th set of data points, we calculate the likelihood ratio:

$$\frac{Q(X^1, X^2, \cdots, X^i)}{P(X^1, X^2, \cdots, X^i)} = \frac{q(x_{1_1}) \times \cdots \times q(x_{i_{n_i}})}{p(x_{1_1}) \times \cdots \times p(x_{i_{n_i}})}$$

We continue to make observations when:

$$B < \frac{q(x_{1_1}) \times \cdots \times q(x_{i_{n_i}})}{p(x_{1_1}) \times \cdots \times p(x_{i_{n_i}})} < A.$$

We transform the equation on log-scale, and with $\log(\frac{q(x_{i_j})}{p(x_{i_j})}) = z_{i_j}$ and get:

$$\log(B) < \sum_{k=1}^{i} \sum_{h=1}^{n_k} z_{k_h} < \log(A).$$

Note that $\left\{Z_{i_j}\right\}$ is again a sequence of independent identically distributed random variables. Denote by m the smallest index of the sequence $(n_1, n_2, \cdots)$ for which either $\sum_{k=1}^{m} \sum_{h=1}^{n_k} Z_{k_h} > \log(A)$ or $\sum_{k=1}^{m} \sum_{h=1}^{n_k} Z_{k_h} < \log(B)$. If we prove that $p(m = \infty) = 0$, then we have proved that the test terminates with probability

1.

Let $c = |\log(A)| + |\log(B)|$. If we prove that the probability is zero that $(\sum_{k=1}^{i} \sum_{h=1}^{n_k} Z_{k_h})^2 < c^2$ holds for all i, then we have proved that $p(m = \infty) = 0$.

As we know that $\left\{ z_{i_j} \right\}$ is a sequence of independent random variables having the same distribution, we know that the expected value of $(\sum_{k=1}^{i} \sum_{h=1}^{n_k} Z_{k_h})^2$ converges to ∞ as $i \to \infty$. Hence we know there exists a positive integer r such that the expected value of the square over the sum over r elements of the sequence $\left\{ z_{i_j} \right\}$ is bigger than $4c^2$. Then certainly we know for this r that $P[(\sum_{k=1+(l-1)r}^{lr} \sum_{h=1}^{n_k} Z_{k_h})^2 \geq 4c^2] = \epsilon$ with $\epsilon \in (0, 1]$ for each $l \in \mathbf{N}$.

Note that if we prove that the probability is zero that $(\sum_{k=1}^{lr} \sum_{h=1}^{n_k} Z_{k_h})^2 < c^2$ holds for all $l \in \mathbf{N}$, then we have proved that the chance is zero that $(\sum_{k=1}^{i} \sum_{l=1}^{n_k} Z_{k_h})^2 < c^2$ holds for all $i \in \mathbf{N}$. We know that:

$$P[(\sum_{k=1}^{lr} \sum_{h=1}^{n_k} Z_{k_h})^2 < c^2 \text{ holds for all } l \in \mathbf{N}] =$$

$$\prod_{j=1}^{\infty} P[(\sum_{k=1}^{lr} \sum_{l=1}^{n_k} Z_{k_h})^2 < c^2 \text{ holds for l=j} \mid \text{it holds for all } l \in \{1, \cdots, j-1\}].$$

We first examine a single term of the right-side expression, therefore let $j \in \mathbf{N}$. We then know that $\sum_{k=1}^{(j-1)r} \sum_{h=1}^{n_k} Z_{i_h} < c^2$. We also know that $P[\sum_{i=1+(j-1)r}^{jr} \sum_{h=1}^{n_k} Z_{i_h} \geq 4c^2] = \epsilon$. Consider $(\sum_{k=1}^{jr} \sum_{h=1}^{n_k} Z_{k_h})^2 = (\sum_{i=1+(j-1)r}^{jr} \sum_{h=1}^{n_k} Z_{i_h} + \sum_{k=1}^{(j-1)r} \sum_{h=1}^{n_k} Z_{i_h})^2$. We know that $\sum_{k=1}^{(j-1)r} \sum_{h=1}^{n_k} Z_{i_h} \in (-c, c)$ and we know with probability ϵ that $\sum_{i=1+(j-1)r}^{jr} \sum_{h=1}^{n_k} Z_{i_h} \in (-\infty, -2c] \cup [2c, \infty)$. For every $a \in (-c, c)$ and every $b \in (-\infty, -2c] \cup [2c, \infty)$ it holds that $|a + b| > c$. Hence we know with probability at least ϵ that $\sum_{k=1}^{jr} \sum_{h=1}^{n_k} Z_{i_h} \geq c^2$.

Now we know that $P[(\sum_{k=1}^{lr} \sum_{l=1}^{n_k} Z_{k_h})^2 < c^2$ holds for $l = j \mid$ it holds for all $l \in \{1, \cdots, j-1\}] \leq 1 - \epsilon$ for every $j \in \mathbf{N}$, and we can finally conclude:

$$P(m = \infty) \leq \prod_{j=1}^{\infty} 1 - \epsilon = 0.$$

Theorem 8. *If we take $A = \frac{1-\beta^*}{\alpha^*}$ and $B = \frac{\beta^*}{1-\alpha^*}$, then we know that the following inequality holds for the type I and type II error: $\alpha + \beta \leq \alpha^* + \beta^*$*

Proof of Theorem 8. Again, let $(n_1, n_2, \cdots)$ be the sequence of numbers where the i'th number represents the number additional observations made after the $i - 1$'th iteration of the test. When for a cylindric point $C(X^1, \cdots, X^i)$ we have that $\frac{Q(X^1,\cdots,X^i)}{P(X^1,\cdots,X^i)} \geq A$, and there exists no integer $j < i$ such that $\frac{Q(X^1,\cdots,X^j)}{P(X^1,\cdots,X^j)} \geq A$ or $\frac{Q(X^1,\cdots,X^j)}{P(X^1,\cdots,X^j)} \leq B$, we call it a cylindric point of type 1. Similarly, we define cylindric points of type 0.

Let $S_0, S_1 \subset \bigcup_{j \in \mathbf{N}} (\chi^{\sum_{i=1}^{j} n_i})$ resp. be the set of all cylindric points of type 0,1, which by definition consists only of samples of length $n \in \{n_1, n_1 + n_2, n_1 + n_2 +$

$n_3, \cdots\}$ number of observations. Theorem 7 tells us that $P(S_0 \cup S_1) = 1$ and $Q(S_0 \cup S_1) = 1$, which simply means that under assumption of both hypotheses the test will terminate.

Since for each sample $(X^{n_1}, \cdots, X^{n_i})$ for which $C(x^{n_1}, \cdots, x^{n_i})$ is an element of S_1 the inequality $\frac{Q(x^{n_1}, \cdots, x^{n_i})}{P(X^{n_1}, \cdots, X^{n_i})} \geq A$ holds, we know that $\frac{Q(S_1)}{P(S_1)} \geq A$. Similarly $\frac{Q(S_0)}{P(S_0)} \leq B$.

Of course $P(S_1) = \alpha$ and $Q(S_0) = \beta$, and as S_0 and S_1 are disjoint, it follows that $P(S_0) = 1 - \alpha$ and $Q(S_1) = 1 - \beta$. Substituting this in the inequalities gives:

$$A \leq \frac{1-\beta}{\alpha} \quad , \quad B \geq \frac{\beta}{1-\alpha}$$

If we take $A = \frac{1-\beta^*}{\alpha^*}$ and $B = \frac{\beta^*}{1-\alpha^*}$, divide the first equation by $(1-\beta)(1-\beta^*)$, multiply the second by $(1-\alpha^*)(1-\alpha)$, and add the resulting inequalities, we get:

$$\alpha + \beta \leq \alpha^* + \beta^*$$

Theorem 8 only tells us that $\alpha + \beta \leq \alpha^* + \beta^*$. In section 4, we stated that for the normal sequential likelihood ratio test, for practical purposes, we can assume that $\alpha \leq \alpha^*$ and $\beta \leq \beta^*$. As the proof of this statement for the non-general case lies beyond the scope of this thesis, we were also unable to check this statement for the general case. Intuitively however, it would be very strange if the statement is not true for the general case, because of the great similarity of the proofs of the above theorems between the general and non-general case. For now, we therefore assume that for all practical purposes, we can simply assume that by this choice of A and B, we know that $\alpha \leq \alpha^*$ and $\beta \leq \beta^*$

5.2. **Are the robust P-value method and sequential likelihood ratio test special cases of this general test?** It is simple to see that the sequential likelihood ratio test is a special case of the generalized sequential likelihood ratio test. And on first sight, it also looks like the robust P-value method is a special case of the general sequential test: if you fill in $B = 0$, then the test design of the general sequential test is exactly the same as the robust P-value test design. This would mean that a robust P-value test corresponds to a general sequential test with type II error chosen to be zero.

However, it is not that simple. When conducting a general sequential test, you choose a simple stopping rule: you stop when the likelihood falls in some predetermined region. The only thing you can change while making observations, is the number of observations you are going to make in the next iteration of the test. When conducting a robust P-value test, you know nothing about your stopping rule in advance, except that you know you are going to stop making observations when the likelihood is smaller than α^*. But one thing is for certain: most researchers will stop making observations at one time, they generally do not plan to go on forever.

Thus, when you choose to conduct a general sequential test with $B = 0$, because of the stopping rule, you would accept that it is possible to sample on forever, while someone conducting a robust p-value test generally does not want to do that. You can only say that the general sequential test and robust p-value test are the same in one very special case: namely the case where the researcher is willing to make infinitely many observations. As this generally is not the case, we can conclude that the robust p-value test and the general sequential test are fundamentally different.

5.3. **Advantages and disadvantages.** The general sequential test may be a good alternative for the Neyman-Pearson test. First, the test is sterner than the Neyman-Pearson test. Second, the number of observations required to conduct the test may also be acceptable in some cases. In chapter 4 we saw that for the special case of making your observations one by one, the test generally results in a 50% saving of observations compared to the Neyman-Pearson test. Further research could focus on whether the number of observations required to conduct the general test is acceptable in other cases. If it is, it could be an alternative for the Neyman-Pearson test.

6. Short introduction to testing complex hypotheses sequentially

The extension of the Neyman-Pearson test from simple to composite hypotheses is quite simple. If you want to test the hypothesis $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$, you simply calculate the p-value of the observed data (the probability of obtaining data at least as extreme as the observed data). If the p-value is smaller the predetermined significance level α, you reject the null hypothesis. The Neyman-Pearson lemma (theorem 1) shows that for the simple case, the Neyman-Pearson test is optimal. One very important property of testing complex hypotheses with the Neyman-Pearson test, is that is also optimal, as in: there exists no test with equal or less significance level, which has higher power. As optimality of this extension for complex hypotheses has been proven for the Neyman-Pearson test, no more mathematical research is needed for these type of tests. (Of course there are still a lot wrong with the interpretation of the results of these tests, but mathematically there is nothing wrong with them.)

However, it is a lot more difficult to extend Wald's sequential test from simple to composite hypotheses, and after 65 years of research, no general test has been found. Research has mainly focused on testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta \geq \theta_1 (> \theta_0)$. The difficulty is to find a test which is optimal, and in the sequential context this means: finding a test which minimizes the expected number of observations.

Wald himself proposed the following way of extending his test to complex hypotheses: If you want to test $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$ with type I and type II errors α and β, use the the test for the simple case with $H : \theta = \theta_0$ and $K : \theta = \theta_1$. This test has minimum expected sample size at $\theta = \theta_0$ and $\theta = \theta_1$ by the Wald-Wolfowitz theorem (see theorem 6), but for other θ, its maximum expected sample size over other θ can be considerably larger than the optimal fixed sample size. Hence Wald's suggestion does not provide an optimal test for all θ.

So far, the best extension of Wald's sequential test (as far as the author of this thesis has been able to check), has been obtained by Lai (1988) [7]. He proposed to use a stopping rule of the form

$$\hat{N} = \inf\{n : \max[\sum_{i=1}^{n} \log \frac{f_{\hat{\theta}_n}(X_i)}{f_{\theta_1}(X_i)}, \sum_{i=1}^{n} \log \frac{f_{\hat{\theta}_n}(X_i)}{f_{\theta_0}(X_i)}] \geq g(cn)\}.$$

This test is designed for the parameter of an exponential family of distributions, and cost c per observation. f_θ is the distribution of an individual observation under the assumption that the parameter is θ, and $\hat{\theta}_n$ is the maximum likelihood estimator of θ after n observations. In the test discussed so far, we have not yet encountered a cost per observation, c. It simply is the cost (in terms of money for example) of making an additional observation, and in some hypothesis tests this can be added to the inference procedure. The function g satisfies $g(t) \sim \log t^{-1}$ as $t \to 0$ and $g(t) \to 0$ as $t \to \infty$.

Lai has proven that this test is nearly optimal, and he also gave a closed-form approximation of the function g [7]. However, Lay proved this optimality result only for testing the parameter of an exponential family of distributions.

7. Directions for future research

There are three things further research could focus on:

First, it would be interesting to see how the generalized sequential likelihood ratio test works in practice. For the normal sequential likelihood ratio test, we know it results in a 50 % saving of observations compared to the Neyman-Pearson test. We however do not know how this works out for the generalized sequential test. A first step would be to examine after how many observations researchers generally are able to analyze there data. For a researcher making 30 observations a day, in most cases it would be very inefficient to analyze the data after each individual observation. However, it would probably not be too much trouble to analyze the data once a day. Research could then focus on how the generalized likelihood ratio test performs compared to the Neyman-Pearson test under these circumstances.

Second, research could focus on examining the three methods for the non-simple case. Testing simple versus simple hypotheses in practice is rare, simple versus a composite hypothesis or a composite versus a composite hypothesis is a lot more common in practice.

Third, in this thesis, we were unable to give a proof for relation between the type I & II errors and the rejection constants for the generalized sequential likelihood ratio test. This part still needs to be proven.

8. References

[1] Pas, S.L. (2010), Much ado about the p-value, *www.math.leidenuniv.nl/theses*

[2] Freeman, P.R. (1993), The role of p-values in analysing trial results, *Statistics in Medicine*, 12, 1443-1452

[3] J.L. Doob (1971) What is a martingale? Amer. Math. Monthly, volume 78, number 5, 451-463.

[4] Wald, A. (1945), Sequential tests of statistical hypotheses, *The Annals of Mathematical Statistics*, volume 16, number 2, 117-186

[5] A. Wald and J. Wolfowitz (1949), Optimum character of the sequential probability ratio test, *The Annals of Mathematical Statistics*, volume 19, number 3, 326-339

[6] Moore, D.S., McCabe, G.P., Craig, B.A. (2009), Introduction to the Practice of Statistics, *Purdue university*, sixth edition, 405

[7] T.L. Lai (1988), Nearly Optimal Sequential Tests of Composite Hypotheses, *The Annals of Mathematical Statistics*, volume 16, number 2, 856-886