Safe tests for 2 x 2 contingency tables and the Cochran-Mantel-Haenszel test

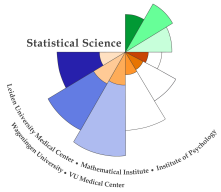
Rosanne J Turner (s1028227)

Thesis advisor: Prof. Dr. P.D. Grünwald

MASTER THESIS

Defended on September 23, 2019

Specialization: Data Science



STATISTICAL SCIENCE
FOR THE LIFE AND BEHAVIOURAL SCIENCES

Contents

Abstract **4**

Notation **5**

1 Introduction **7**
1.1 Hypothesis testing . . . 8
1.2 The safe testing framework . . . 12
1.3 The GROW-criterion . . . 14
1.4 2x2 contingency tables . . . 17
1.5 Thesis outline . . . 19

2 Finding the GROW S-value for 2x2 tables **21**
2.1 Finding the GROW S-value for a *subset* of the alternative hypothesis . . . 21
2.2 BFGS for finding the JIPr . . . 22
2.3 Comparing BFGS results to a theoretical minimum . . . 25
2.4 Extending BFGS to composite hypotheses . . . 27

3 The *simple* S-value **29**
3.1 The simple S-value for one-sided testing when $n_a = n_b$ . . . 29
3.2 The simple S-value for two-sided testing when $n_a = n_b$ . . . 30
3.3 The simple S-value when $n_a \neq n_b$ . . . 34
3.4 Choices for the restricted alternative hypothesis . . . 36

4 The KL-divergence based S-value **43**
4.1 Reverse information projection . . . 43
4.2 JIPr for the KL-based S-value . . . 44
4.3 Power of the S-values . . . 46

5 Optional stopping with the simple S-value **49**
5.1 Robustness under the frequentist interpretation of optional stopping . . . 49
5.2 The simple S-value is robust under optional stopping . . . 50
5.3 The *expected* sample size under optional stopping . . . 51

6 Real-world examples **53**
6.1 Doll data . . . 53
6.2 Canine therapy data . . . 54
6.3 ECMO data . . . 57

7 A safe version of the Cochran-Mantel-Haenszel test **59**

7.1 Motivation for a *stratified* version of the 2x2 contingency table test . . . 59
7.2 Defining hypotheses and likelihoods for the safe CMH-test . . . 60
7.3 Finding the JIPr for the default safe CMH-test . . . 61
7.4 The simple S-value for the CMH-test . . . 64
7.5 Power of the safe CMH test . . . 66

8 Conclusion and future work **69**
8.1 Reparameterization . . . 70
8.2 Uniformly most powerful safe tests . . . 70
8.3 S-values under the null hypothesis . . . 71

Bibliography **73**

A Figures, Tables and derivations **77**
A.1 Chapter 3 . . . 77
A.2 Chapter 4 . . . 78
A.3 Chapter 7 . . . 79

B R package: preliminary manual **83**
B.0 General functions . . . 83
jipr . . . 83
jiprCMH . . . 85
S test . . . 85
STestCMH . . . 86
B.1 Scenario 1: n known, L known . . . 88
FindDeltaForLDefault . . . 88
FindDeltaForLKL . . . 88
FindDeltaForLCMH . . . 89
B.2 Scenario 2: δ known, L known . . . 91
FindNForLDefault . . . 91
FindNForLKL . . . 92
FindNForLCMH . . . 93
B.3 Scenario 3: n known, α known . . . 94
FindDeltaUmpDefault . . . 94
FindDeltaUmpKL . . . 95
FindDeltaUmpCMH . . . 96
B.4 Scenario 4: δ° known, α known . . . 97
FindNForPowerDefault . . . 97
FindNForPowerKL . . . 98
FindNForPowerCMH . . . 99
B.5 List of default tables . . . 101

Abstract

Recently, a new theory of hypothesis testing was introduced: safe testing. Within the safe testing framework, random variables called S-values are used for hypothesis testing. S-values can be interpreted as both conservative p-values *and* Bayes factors. Further, they allow for *optional continuation*: S-values from multiple studies can be multiplied while retaining a type-I error guarantee, and some S-values are even robust under the frequentist interpretation of optional stopping. For this thesis, I developed safe tests for two classical frequentist hypothesis tests: the 2x2 contingency table test and its stratified equivalent, the Cochran-Mantel-Haenszel test. These tests were designed to be GROW (growth-rate optimal in the worst case) for certain subsets of the alternative hypothesis. Two versions of the tests were presented: a version that provides the GROW S-value for a restricted alternative hypothesis based on a minimal absolute difference between group means, and a version that is based on the Kullback-Leibler divergence between the alternative and null hypothesis. For the 'minimal absolute difference' version, an analytically computable 'simple' S-value turned out to exist, which is robust under optional stopping. I showed that when using this safe test for optional stopping, the expected sample size needed to achieve a desired power can be lower than when using Fisher's exact test. No 'simple' definition could be found for the Kullback-Leibler version: this GROW safe test has to be found through numerical optimization. Nevertheless, the Kullback-Leibler version could still be preferred in some cases: it was shown to gain higher power for certain data-generating distributions compared to the simple S-value. Both S-values were implemented in an R package: the `safe2x2` package.

Acknowledgements

Many thanks to Peter Grünwald for giving me the opportunity to work on this challenging thesis project. I am also grateful that Tim van Erven and Stéphanie van der Pas were willing to read this thesis and take place in the graduation committee for the thesis defence. I want to thank Alexander Ly, Rianne de Heide, Judith ter Schure, Dmitrii Ostrovskii and Muriel Perez for all the discussions about the safe testing project. I would like to thank Alexander in Judith in particular; they have worked simultaneously on developing an `R` package for a safe version of the t-test, and I have enjoyed our collaboration during this development proces. Lastly, many thanks to the CWI and all colleagues from the Machine Learning group for providing such a welcoming and pleasant working environment.

Notation

Table 0.1: Frequently occurring notations and the page where they are introduced.

symbol	description	page
P	A probability *distribution*	10
p	A probability *mass function*	10
$\mathcal{H}_0$	The null hypothesis: set of distributions on some sample space	10
$\mathcal{H}_1$	The alternative hypothesis: set of distributions on some sample space	10
Θ_0	The null hypothesis parameter space	10
Θ_1	The alternative hypothesis parameter space	10
α	A significance level and the probability of making a type-I error	10
β	The probability of a type-II error	10
B_{10}	A Bayes factor representing the posterior odds of the alternative hypothesis against the null hypothesis	11
p_{W_j}	Bayes marginal with prior W_j on Θ_j, the parameter space for $\mathcal{H}_j$	11
$w_j(\theta)$	Prior mass on a $\theta \in \Theta_j$	11
$W(\Theta_j)$	All possible prior distributions on Θ_j	16
y^n	Observed sequence of outcomes (0 or 1) on a sample of length n	17
x^n	Fixed group assignment (a or b) for each observation in a sample of length n	17
n_a	The number of observations in group a in a 2x2 contingency table test	17
n_b	The number of observations in group b in a 2x2 contingency table test	17
n_{a1}	The number of ones observed in group a in a 2x2 contingency table test	17
n_{b1}	The number of ones observed in group b in a 2x2 contingency table test	17
θ_a	For 2x2 tables: the probability of observing a 1 in group a under the alternative hypothesis	17
θ_b	For 2x2 tables: the probability of observing a 1 in group b under the alternative hypothesis	17
θ_0	For 2x2 tables: the probability of observing a 1 under the null hypothesis	17
θ_1	For 2x2 tables: (θ_a, θ_b)	17
$D(P\|\|Q)$	Kullback Leibler (KL) divergence between distributions P and Q	15
S^*	A GROW S-value	15
$\Theta_1(\delta)$	The restricted alternative hypothesis parameter space relative to some distance measure δ	21
$\bar{\Theta}(\delta)$	Subset of all distributions on the boundary of $\Theta_1(\delta)$: the distributions that will yield the worst case growth rate	22
ϕ	Adjustment made to the simple S-value in the case where $n_a \neq n_b$	35
δ^*	In the case of optimizing for power: the divergence measure used to construct the S-value	38
δ°	In the case of optimizing for power: the divergence measure used to define the resrtricted alternative hypothesis one wants to optimize for	38
S^*_δ	A GROW S-value constructed with $\bar{\Theta}_1(\delta) \subset \Theta_1$	38
M	The number of Monte-Carlo iterations per unique combination of parameters of interest	41

Chapter 1

Introduction

"Research waste" is a major problem for science: a significant fraction of scientific studies do not contribute fully to our knowledge, either through mistakes in the design or analysis of the studies [Chalmers and Glasziou, 2009]. To satisfy assumptions of classical frequentist hypothesis tests, study designs, including the sample size, have to be determined upfront [Fisher, 1938]. However, it has been shown that researchers do not always adhere to this in practice, potentially leading to misleading results [John et al., 2012]. This requirement of classical hypothesis tests also limits the flexibility of the researcher: it is not possible to *engage in optional continuation, i.e. keep collecting data after the initial sample size has been reached, and to add these new data to the initial sample, as this makes the results invalid.* In addition, current techniques for combining evidence collected in multiple studies are susceptible to various forms of *bias*, also leading to a loss of a type-I error guarantee [Ter Schure and Grünwald, 2019].

This is why Grünwald et al. [2019] (from now on referred to as GHK) recently developed a new theory of hypothesis testing: safe testing. In this new method, random variables called 'S-values' are constructed, a name chosen to emphasize that these S-values *could provide an alternative to the classical p-values.* These S-values have some favourable properties that *improve the interpretability* of hypothesis testing, for example through interpretation as a gambling problem. Further, they *improve the flexibility* of hypothesis testing and study designs, through allowing for optional continuation and sometimes even optional stopping, as will be elaborated on throughout this thesis.

Contribution

The primary aim of this thesis was to develop a safe version of one of the classical frequentist hypothesis tests: the 2x2 contingency table test. Two versions of this safe test were previously proposed by GHK: one where the researcher aims to test for a minimal *substantively relevant difference* between the two groups, and one based on *Kullback-Leibler divergence* between the alternative and null hypothesis. GHK provided a basic mathematical formulation for both S-values, and a first implementation for small sample sizes ($\leq$ 20) with equal subgroup sizes. In this thesis, I aimed to develop fast implementations of both versions, for any sample size, with the goal of releasing an R package with the implementations. For the version based on the minimal substantively relevant difference, I defined a *generalized* definition of the safe test, providing the *best safe test* for every sample size: the simple S-value. This simple S-value is analytically computable, resulting in fast computation. Further, this S-value is not only valid under optional continuation, but also under optional stopping (which will be defined in Chapter 5); it was illustrated that the actual expected sample size needed to achieve a certain power can even be lower than for Fisher's

exact test with this simple S-value. Further, an extension to the stratified version of the 2x2 contingency table test, the Cochran-Mantel-Haenszel test (CMH test) was developed as well.

An R package was developed with implementations of these two safe tests: the `safe2x2` package. Several scenarios possibly encountered during hypothesis testing were included in the package, for example the case where one is interested in a certain minimal difference between group means, or where one wants to optimize power at a maximum sample size that resources are available for. A preliminary version of the manual of this package is attached to Appendix B of this thesis. Simultaneously to the development of the `safe2x2` package, CWI-colleagues Alexander Ly and Judith ter Schure have been working on developing an R package for a safe version of the t-test. The final version of the `safe2x2` package will be released soon in collaboration with Ly and Ter Schure to ensure the user experience is similar while using either of these packages. The release of both packages will be accompanied by a paper highlighting the most important findings while designing both safe tests.

All experiments described throughout this thesis were performed in RStudio on a personal computer with a CPU with average power. Hence, it can be assumed displayed running times for algorithms will give a realistic impression for users of the `safe2x2` package. The code used for generating all the figures and tables in the thesis can be found on my online repository on GitHub[1], accompanied by information on the used versions of R and R packages. The preliminary code of the package will be displayed there as well, until the package has been released officially.

In the remainder of this introduction, first, the idea of and need for hypothesis testing and the three most well-known approaches towards hypothesis testing are introduced. Next, the safe testing framework is introduced formally and the added value in comparison to existing methods of hypothesis testing is discussed. In the third section of this introduction, optimal S-values named 'GROW S-values' are defined. In the fourth section, the *null* and *alternative* hypothesis for the 2x2 contingency table test are defined and existing methods for analyzing 2x2 tables are discussed. Lastly, in the fifth section, the outline of the rest of this thesis is summarized.

1.1 Hypothesis testing

Why hypothesis testing?

One could state that the final goal of all scientific research is to learn something new about the true nature and structure of the world. Researchers collect data and try to learn something from it by finding the patterns and structures that underly those data. But when can one conclude that someone has learned something about the world from just a small sample of data? In other words, when can one infer the patterns or qualities observed in their sample to the entire population? When available data consist of only a fraction of the entire world or population, observed results could have been influenced by the way the data was selected, measured, and by chance. The latter is where hypothesis testing comes in.

In hypothesis testing, one assesses how compatible the data are with an assumed 'null hypothesis'. One could then regard the null hypothesis as incompatible with the data if one is confident it cannot

[1] www.github.com/rosanneturner. Would this repository become (unexpectedly) unavailable in the near future, I can be reached through the CWI (Rosanne dot Turner at cwi dot nl) for requests for a copy of the code.

be true based on the data collected: when the outcome of the hypothesis test is *significant*. This concept has been introduced in a statistical context by John Venn in 1888 [McCloskey and Ziliak, 1996], an English mathematician working on the problem of determining whether physical features were really different between various groups of English students:

> "*(...) differences in the above tables are permanent and significant, in the sense that we may be tolerably confident that if we took another similar batch we should find a similar difference; and which are merely transient and insignificant, in the sense that another similar batch is about as likely as not to reverse the conclusion we have obtained.*" - Venn and Galton [1888], p. 148.

In other words, Venn viewed observations in samples as significant if they were 'permanent': if one would sample new batches from the same population, one would find similar observations. On the other hand, if we are not so sure we would obtain the same observations when collecting another sample, the observations are 'transient' and hence insignificant. This implies that when a hypothesis test is significant, it should provide evidence that we have found *something permanent*: an underlying structure in the data that is presumably true for the whole population.

When the outcome of the hypothesis test is significant and null hypothesis is thus incompatible with the data, an *alternative hypothesis* can be 'accepted'. Depending on the kind of hypothesis test one uses, the alternative hypothesis can be specified explicitly, or the null hypothesis is tested on its own. For example, in linear regression, one could test the null hypothesis that a coefficient is 0, i.e. there is no correlation between the predictor and the outcome variable. When one then finds this null hypothesis to be incompatible with the data, one can accept the alternative hypothesis that the coefficient is unequal to 0, and that there is an underlying relation between the predictor and outcome. One can imagine that the conclusions one draws from a hypothesis test, i.e. the things one can learn from the data, can be heavily influenced by the choice of/ to specify an alternative hypothesis; this will be elaborated on in the next section.

Current approaches towards hypothesis testing and their limitations

There are lots of ways in which one could assess the compatibility between a null hypothesis and collected data. Currently, three different approaches or 'schools' of hypothesis testing are dominant in research practice: Fisherian, Neyman-Pearsonian and Jeffreys' approaches to testing [Berger, 2003], which all three have distinct approaches towards this task. These three schools and difficulties that could arise from their approaches and assumptions are discussed in the the following sections.

We now first need to introduce *p-values*. In practice, in the biomedical and economic sciences, usually a mix of Fisherian and Neyman-Pearsonian hypothesis testing is used; both approaches work with p-values, and reporting of p-values is often requested by journals in these fields [Woolston, 2015]. Some notation is now introduced first; frequently used notations are also summarized in Table 0.1. Let our sample space for this section be defined as $\mathcal{Z}^n$ with some finite n; data collected are a sample $(z_1, \ldots, z_n) = z^n \in \mathcal{Z}^n$. Throughout this manuscript, P will denote a probability distribution, and p will denote a probability mass function, with $P(Z^n = z^n) = p(z^n)$ denoting the

probability of z^n. We define our null hypothesis $\mathcal{H}_0$ and our alternative hypothesis $\mathcal{H}_1$ (notation adapted from GHK):

$$\mathcal{H}_0 = \{P_\theta : \theta \in \Theta_0\} \quad ; \quad \mathcal{H}_1 = \{P_\theta : \theta \in \Theta_1\}. \tag{1.1.1}$$

For constructing the *p-value*, for some $\theta \in \Theta_0 \cup \Theta_1$, data collected are presumed to be generated by the distribution: $Z^n \sim P_\theta$. A random variable, the "test statistic", T is then defined as a function of the data. The more extreme this test statistic is, the lower the probability of observing such a test statistic under $\mathcal{H}_0$ should be, and the more evidence there is against $\mathcal{H}_0$ [Berger, 2003]. For a singleton $\mathcal{H}_0$ ($\Theta_0 = \{\theta_0\}$), the p-value based on z^n is then defined as:

$$\text{pval}(z^n) = P_{\theta_0}(T(Z^n) \geq T(z^n)). \tag{1.1.2}$$

Thus, the p-value reflects the the probability of observing a result as extreme or more extreme when data are generated under the null hypothesis. One can the decide that the data are incompatible with the null hypothesis, i.e. to reject the null hypothesis when one judges the p-value to be sufficiently small.

A Fisherian would state that a p-value can be an indicator of the strength of evidence against the null hypothesis [Berger, 2003], i.e. the lower a p-value, the greater the incompatibility between the data and the null hypothesis; $\mathcal{H}_1$ is not mentioned and need not be specified. This is quite different from the Neyman-Pearsionian view of testing. There, the desired probability of a type-I error (α), the probability of declaring that the null hypothesis is incompatible with the data ('rejecting' the null hypothesis) while it is true, is specified upfront. Further, an alternative hypothesis has to be specified upfront as well, and the probability of *not rejecting* this alternative hypothesis while data are in fact generated by some distribution from the alternative hypothesis is reported: this is the type-II error (β). The *power* of the test for data generated under a distribution $P_\theta : \theta \in \Theta_1$ is then defined as $1 - \beta_\theta$: the probability of rejecting the null hypothesis when data are generated under the distribution from the alternative hypothesis.

With Neyman-Pearsonian testing, the null hypothesis is rejected if the p-value is smaller than or equal to α. Only the decision to reject or accept the null hypothesis is reported, along with the specified α and β. As a consequence of this, variation of the data within the rejection range is not reflected in the reported test outcome, and a frequent criticism of Neyman-Pearsonian testing is that not enough information is gained from the data [Berger, 2003]. As mentioned before, in current scientific practice, a mix of Fisherian and Neyman-Pearsonian testing is used: researchers in their methods predetermine some α and β, but then report the exact p-values instead of just the decision to reject, and often regard test results as stronger when they have lower p-values.

P-values, or p-value tests in the Neyman-Pearsonian approach, are less informative than most researchers that currently use them think. Recently, 800 researchers even signed a comment in the impactful journal Nature, proposing to not use p-values in hypothesis testing any more [Amrhein et al., 2019]. Researchers often misinterpret the p-value as the probability that the null hypothesis is true given the test statistic, instead of the probability of observing such a test statistic (or a more extreme test statistic) under the null hypothesis. A well-known and startling illustration (brought to our attention via Van der Pas [2010]) is the study by Wulff et al. [1987], where in a survey under Danish doctors, 52% of the respondents believed the p-value represented the probability of the null hypothesis being true (and vice versa, one minus the p-value representing the probability of the alternative hypothesis being true). A more recent study under anesthesiologists revealed that, 30

years later, p-value interpretation is still difficult; in this survey, 30% of respondents interpreted the p-value as the probability of the null hypothesis being true, and only 15% were able to define p-values correctly [Schober et al., 2017].

This misinterpretation of p-values can lead to huge overestimation of the evidence against the null hypothesis [Berger, 2003]. In practice, the majority of hypothesis tests conducted concern tests where the null hypothesis is (at least approximately) true, i.e. 'wrong ideas' were tested. Assuming that approximately ten percent of alternative hypotheses are initially true, the 'positive predictive value' of a p-value below 0.05 can decrease to 0.31, with 69 percent of significant test results coming from tests where the null hypothesis was true [Sterne and Smith, 2001]. This is a lot higher than the"5 percent chance of the null hypothesis being true" a large fraction of researchers believe the critical value for the p-value to guarantee [Wulff et al., 1987].

Another important criticism of p-values is that a p-value lower than a certain critical value does not have to implicate *substantive* significance; with large sample sizes, small differences from the null hypothesis can result in a high test statistic [Deming, 1982, McCloskey and Ziliak, 1996]. For example, this is relevant for researchers working with large datasets, applying drag-net methods for finding relations between variables, e.g. while building regression models. But what does one have to conclude, when one has found a coefficient to be significantly different from 0 in such a model, but with such a small value that it does not have any relevance for the researcher's question? It would be preferable to think about the effect size of interest before conducting any hypothesis tests, and to have this property incorporated in the hypothesis test.

Jeffreys' approach to testing already provides some solutions to the challenges presented above for p-value hypothesis testing. With Jeffreys' approach, the explicit formulation of a null hypothesis and an alternative hypothesis are required, and both hypotheses are incorporated in calculating the hypothesis test's outcome [Berger, 2003]. Jeffreys' approach to testing uses *Bayes marginal distributions* (marginals). We had already defined Θ_1 and Θ_0 for our alternative and null hypothesis. The Bayes marginal distribution P_{W_j} provides a weighted average of every distribution $P_\theta : \theta \in \Theta_j$. Formally, W_j is a (proper) probability distribution on Θ_j, often called the *prior*, with mass or density function w_j. If w_j is a density function, then the Bayes marginal density is defined as follows:

$$p_{W_j}(z^n) = \int_\theta w_j(\theta) p_\theta(z^n) d\theta, \tag{1.1.3}$$

and if w_j is a mass function, the integral is replaced by a sum.

One can observe that *prior beliefs* one has about the null and alternative hypothesis in the form of the prior W_j have to be specified. This way, one could choose to assign higher weights to parameters one believes to be more likely to truly generate the data. Hence, the test outcome could be influenced by beliefs one holds ('expert opinions') prior to an experiment. This is the major criticism Fisherians and Neyman-Pearsonians have against the Bayesian approach: the design of the hypothesis test is not necessarily *objective* [Berger, 2003]. This potential challenge raised *objective Bayesianism*: a group of statisticians who believed priors should be as non-informative as possible [Berger et al., 2015]. Examples of such priors are the uniform prior, the well-known Jeffreys prior, designed by Jefrreys, and the right-Haar prior for models satisfying a group-invariance [Berger et al., 2015]. However, different research groups still have different opinions on which objective

priors perform best, and there yet is no standard choice for such a prior [Berger et al., 2015].

We can use the Bayes marginals to express the *posterior odds* of the alternative hypothesis against the null hypothesis being true given the data using Bayes' rule:

$$\frac{p(\Theta_1|z^n)}{p(\Theta_0|z^n)} = \frac{p_{W_1}(z^n)}{p_{W_0}(z^n)} \frac{\pi(\Theta_1)/p(z^n)}{\pi(\Theta_0)/p(z^n)}. \tag{1.1.4}$$

Hence, if we assume that both hypotheses are equally likely a priori ($\pi(\Theta_1) = \pi(\Theta_0) = \frac{1}{2}$), this expression can be simplified to the ratio of the marginals under the alternative hypothesis Θ_1 and the null hypothesis Θ_0. This is the *Bayes factor* (B_{10})[2] or likelihood ratio Jeffreys' proposed to use for hypothesis testing [Kass and Raftery, 1995]:

$$B_{10}(z^n) = \frac{p_{W_1}(z^n)}{p_{W_0}(z^n)}. \tag{1.1.5}$$

The Bayes factor might provide a more intuitive expression than the p-value based tests presented above, as the evidence for the alternative hypothesis is directly compared to the evidence for the null hypothesis. A Bayes factor B_{10} bigger than 1 indicates that there is more evidence for the alternative hypothesis, and a Bayes factor smaller than 1 indicates that there is more evidence for the null, and that the null hypothesis should not be rejected [Berger, 2003]. Here, the incorrect interpretation of the p-value as a hypothesis test outcome reflecting the probability of the alternative- or null hypothesis being true *does* apply: the bigger the Bayes factor is, the greater the evidence for $\mathcal{H}_1$ being true given the data.

In conclusion, the three current schools of hypothesis testing differ greatly with regard to test design and to the information test outcomes provide. Because of the difference in nature of reported statistics, one can imagine that combining or comparing evidence across multiple studies using different test statistics is difficult: how does one compare a Neyman-Pearsonian test outcome ($\mathcal{H}_0$ was rejected with critical value $\alpha = 0.05$), a Fisherian one ($p = 0.01$) and a Bayes factor ($B_{10}(z^n) = 100$)? The safe testing framework will now be introduced, which can be interpreted in the Bayesian, Fisherian and Neyman-Pearsonian contexts, while also providing possibilities for optional stopping and optional continuation.

1.2 The safe testing framework

Safe testing, recently introduced by GHK, is based on *S-values*. These are nonnegative random variables that satisfy, for all probability distributions under the null hypothesis:[3]

$$\forall\, P \in \mathcal{H}_0 : \quad \mathbb{E}_P[S(Z^n)] \leq 1. \tag{1.2.1}$$

In the following section, it will be motivated how these S-values can be composed, why S-values are useful for hypothesis testing, and how S-values can be interpreted in both the frequentist and

[2] Would we want to express the posterior odds of the null hypothesis against the alternative hypothesis, we could use $B_{10}^{-1} = B_{01}$.

[3] Throughout the manuscript, $S(Z^n)$ and $S(z^n)$ will be abbreviated to S in some of the formulas with the purpose of improving readability.

Bayesian contexts.

S-values in the frequentist context

With respect to the frequentist context, a *safe test* $T_\alpha(S)$ based on an S-value is defined as:

$$T_\alpha(S) = \begin{cases} \text{reject}_0 \text{ if } S \geq \frac{1}{\alpha} \\ \text{not reject}_0 \text{ if } S < \frac{1}{\alpha}, \end{cases} \tag{1.2.2}$$

for *significance level* $0 \leq \alpha \leq 1$. Because of their definition, safe tests have a type-I error guarantee at level α:

$$\begin{aligned} \forall P \in \mathcal{H}_0 : \; P(T_\alpha(S) = \text{reject}_0) &= P\left(S \geq \frac{1}{\alpha}\right) \\ &\leq \frac{\mathbb{E}_P[S]}{\alpha^{-1}} \\ &= \alpha \, \mathbb{E}_P[S] \\ &\leq \alpha, \end{aligned} \tag{1.2.3}$$

where the first inequality follows from Markov's inequality [Ross, 2010]. Since we now have that:

$$P(T_\alpha(S) = \text{reject}_0) = P\left(\frac{1}{S} \leq \alpha\right) \leq \alpha, \tag{1.2.4}$$

$\frac{1}{S}$ is by definition a *conservative p-value* [GHK]. Hence, it can be concluded that all S-values can be converted to and interpreted as p-values, but the converse is not necessarily true. It does not hold for arbitrary p-values pval for all $P \in \mathcal{H}_0$ that $\mathbb{E}_P[\frac{1}{\text{pval}(Z^n)}] \leq 1$. This is illustrated in Section 1.4 for Fisher's exact test.

As a result, S-values are overall *more conservative* than p-values: $\frac{1}{S}$ will often yield higher values than a p-value from a frequentist test conducted on the same data. As will be illustrated in Chapter 4, this results in a slightly smaller power of the safe tests compared to frequentist tests when performed on datasets of the same size. However, in Chapter 5 it will be shown that because of the properties of some special safe test that allows for *optional stopping*, the amount of data needed to reach a certain power can even be expected to be lower than the amount of data needed with the frequentist test in certain cases.

Bayes factors can be S-values

Let us consider the expected value of Bayes factors under a null hypothesis $P_0 \in \mathcal{H}_0$ (for now, we consider the case where $n = 1$ and thus $Z^n = Z$):

$$\mathbb{E}_{P_0}[B_{10}(Z)] = \int p_0(z) \frac{p_{W_1}(z)}{p_{W_0}(z)} dz. \tag{1.2.5}$$

We can observe that for singleton $\mathcal{H}_0 = \{P_0\}$, we must have that W_0 puts all mass on P_0, and $p_0(z) = p_{W_0}(z)$, The expression then simplifies to:

$$\mathbb{E}_{P_0}[B_{10}(Z)] = \int p_{W_1}(z) dz = 1. \tag{1.2.6}$$

It can be concluded that at least Bayes factors with a singleton null hypothesis are S-values. Such 'special' Bayes factors are also *test martingales*; test statistics that are robust under optional stopping, and have already been studied extensively [Shafer et al., 2011]. As an important extension of this, GHK have shown that, altough for composite $\mathcal{H}_0$ not all Bayes factors are S-values, and not all S-values are Bayes factors, some special Bayes factors for composite $\mathcal{H}_0$ *are* S-values. Further, the 'best' S-values that we can compose are Bayes factors with some very special priors, which will be elaborated on in Section 1.3.

Safe tests are still safe under optional continuation

In summary, the best S-values that one can compose are some special Bayes factors, that also can be interpreted as conservative p-values. Because of this, both Bayesian and frequentist statisticians can use S-values, and we will now show that S-values from various experiments even could be combined [GHK].

Let us consider a researcher collecting a first set of data $Z^{(1)}$ and calculating the corresponding S-value $S^{(1)}$. Based on this S-value and some side information (e.g. she still has budget for the project, extra data samples suddenly have become available), the researcher decides to collect a second set of data $Z^{(2)}$ and to calculate an S-value $S^{(2)}$. This process continues until the final sample $Z^{(k)}$ is collected and the researcher decides to stop the experiments because the resulting S-value $S^{(k)}$ is high or low enough, or new side information emerges (for example, the researcher's budget is insufficient). Multiplying all obtained S-values into a 'summarizing' statistic $S^{\pi} = \prod_{i=1}^{k} S^{(i)}$ would now *still result in an S-value*. Since it holds that $\forall S^{(i)} : \mathbb{E}_{P_0}\left[S^{(i)}\right] \leq 1$, it can be derived that the expectation of the product of all S-values also has an expectation smaller than or equal to 1: no matter what rule or side-information was used for constructing S^{π}, $\mathbb{E}_{P_0}\left[S^{\pi}\right] \leq 1$. It then follows that for S^{π}, we also have a type-I error guarantee. A formal proof can be found in GHK.

An intuitive analogy can be presented in the form of a *gambling problem* (adapted from GHK). Imagine entering a casino with a start capital €1 in one's pocket and playing a slot machine. When d euros are inserted in the machine, it returns Sd euros, with S based on the result from spinning the slot machine. Hence, if $S < 1$, you lose money, and if $S > 1$, you win money. The $S^{(1)}d$ euros are then reinserted in the machine, and $S^{(2)}S^{(1)}d$ euros are returned, and so on. It is common knowledge that the slot machine will sample from a distribution where the expected gain is smaller then 1: the casino does not actually want most customers to make money, and, despite the beliefs of many casino enthusiasts, adhering to certain superstitious beliefs (e.g. "continue playing until you have seen a strawberry three times in a row") is known to not increase one's expected profit. Hence, even if the capital is reinvested very often, we do not expect to yield a profit: when we finally leave the casino with our yielded €$1 \times \prod_{i=1}^{k} S^{(i)}$, this amount is expected to be less than 1.

1.3 The GROW-criterion

One can imagine that many random variables satisfy the criterion to be an S-value (see (1.2.1)). However, not all choices for S-values are sensible, as some might provide a very low probability of actually rejecting the null hypothesis when it is false, i.e. they have a *low power*. For example consider the S-value $\forall z^n \in \mathcal{Z}^n : \; S(z^n) = 1$; it satisfies (1.2.1), but $\mathcal{H}_0$ will never be rejected. Or, alternatively, consider testing whether a Bernoulli distribution has a mean θ equal to 0.5 versus a mean *not equal to* 0.5 with a Bayes factor: one could then design an S-value where the majority of the weight in W_1 is put on one extreme value $\theta^* \in \Theta_1$, for example $\theta^* = 0.99$. The S-value will

then remain low, and we will fail to reject the null hypothesis, when the data are really generated under a Bernoulli distribution with mean lower than 0.5, for example when $\theta = 0.01$.

Thus, a criterion for 'good' S-values seems needed: we want our 'capital' S^π to *grow fast* if $\mathcal{H}_1$ is true. To ensure a certain expected growth of our S-value (under the gambling interpretation) or *power* (under the frequentist interpretation), it is desirable to put a *lower bound* on the performance of the S-value under all distributions in the alternative hypothesis. GHK have argued that this performance of the S-value is best measured as *expected capital growth* (also named *doubling rate*, [Cover and Thomas, 1990]), i.e. for $P \in \mathcal{H}_1$: $\mathbb{E}_P[\log(S(Z^n))]$. This performance measure provides an intuition of how to interpret the S-value: for a large series of experiments, S^π grows exponentially at the rate of the expected capital growth [Cover and Thomas, 1990].

It will now be illustrated through an example adapted from GHK why it is further desirable to optimize S-values with respect to expected capital growth. Consider the following S-value, based on Neyman-pearsonian hypothesis testing with a certain p-value pval:

$$S(z^n) = \begin{cases} 0 \text{ if } \text{pval}(z^n) \geq \alpha \\ \frac{1}{\alpha} \text{ if } \text{pval}(z^n) \leq \alpha. \end{cases} \tag{1.3.1}$$

If pval is indeed a (strict or conservative) p-value, then $\forall P \in \mathcal{H}_0 : \ \mathbb{E}_P[S(Z^n)] = \frac{1}{\alpha} P(\text{pval}(Z^n) \leq \alpha)) \leq \frac{1}{\alpha}\alpha = 1$, so S is indeed an S-value. In the optional continuation setting, one can observe that for this test, at least for some $P \in \mathcal{H}_1$, there is a probability of obtaining $S(z^n) = 0$ and thus losing all one's capital, resulting in a power of 0 in subsequent experiments. However, it has just been argued that the type-I error guarantee S-values possess in the context of optional continuation is one of their major advantages.

Hence, we might call an S-value 'good' if the expected value of the safe test $\mathbb{E}_{P_\theta}[f(S(Z^n))]$ is sufficiently large and greater than 0 for any $\theta \in \Theta_1$, with $f : \mathbb{R} \to \mathbb{R}$ an increasing function. Using the identity function here would not prevent the occurrence of situations as described above, where a substantial probability of getting $S = 0$ under the alternative hypothesis exists. However, using the logarithm would prevent this, as S-values that are close to 0 for certain sequences of data would then yield very big negative expected values. Throughout this thesis, the natural logarithm will be used.

Based on such reasoning, GHK introduced GROW (*growth-rate optimal in worst-case*) S-values. $\mathcal{H}_1$ and $\mathcal{H}_0$ are required to be non-overlapping in the simplest form of the definition. Assuming we believe $\mathcal{H}_1$ is true, we do not know which distribution in $\mathcal{H}_1$ really generates the data, but we want to design an S-value that gives good results even in the *worst case*, i.e. the distribution in $\mathcal{H}_1$ most similar to $\mathcal{H}_0$. In this *worst case*, we want the S-value to perform *optimally*. The GROW S-value S^* is hence defined as the S-value that has the highest expected capital growth when the data are generated under the worst case scenario in $\mathcal{H}_1$ [GHK]:

$$S^* = \underset{\text{S: S is an S-value}}{\arg\max} \min_{P \in \mathcal{H}_1} \mathbb{E}_P[\log(S)]. \tag{1.3.2}$$

One can observe that the best function of the data *that is an S-value* has to be picked in the equation above. In Section 1.2.5, it was already mentioned that not all Bayes factors with composite $\mathcal{H}_0$ are S-values. However, it turns out that the GROW S-value can be constructed as a Bayes factor between P_{W_1} and P_{W_0} with two very special priors W_1 and W_0.

First note that we have that the expectation of the logarithm of an S-value S with $S(z^n) = \frac{p_{\theta_1}(z^n)}{p_{\theta_0}(z^n)}$ under P_{θ_1} is equal to the Kullback-Leibler divergence between P_{θ_1} and P_{θ_0} [Cover and Thomas, 1990]:

$$\mathbb{E}_{P_{\theta_1}}\left[\log(S(Z^n))\right] = \mathbb{E}_{P_{\theta_1}}\left[\log\left(\frac{p_{\theta_1}(Z^n)}{p_{\theta_0(Z^n)}}\right)\right] = D\left(P_{\theta_1}||P_{\theta_0}\right). \tag{1.3.3}$$

Let us define the set of all possible priors on Θ_j as $W(\Theta_j)$. GHK have shown that, in the case of composite $\mathcal{H}_0$, the P_{W_0} providing the *best* S-value under a certain fixed $P_{W_1'}$ can be found through finding the P_{W_0} that minimizes the Kullback-Leibler divergence between this fixed $P_{W_1'}$ and the set of all possible marginal distributions from $\mathcal{H}_0$. This is called the *Reverse Information Projection (RIPr)* of $P_{W_1'}$ on $\{P_{W_0} : W_0 \in W(\Theta_0)\}$, which is achieved by finding the $W_0^* \in W(\Theta_0)$ that achieves:

$$\min_{P\in\{P_{W_0}:W_0\in W(\Theta_0)\}} D(P_{W_1'}||P) = \min_{W_0\in W(\Theta_0)} D(P_{W_1'}||P_{W_0}). \tag{1.3.4}$$

The resulting S-value $S = \frac{P_{W_1'}}{P_{W_0^*}}$ is the S-value with the best expected capital growth that could be designed with that particular fixed $P_{W_1'}$, in the sense that [GHK]:

$$\min_{W_0\in W(\Theta_0)} D(P_{W_1'}||P_{W_0}) = \max_{\text{S: S is an S-value}} \mathbb{E}_{P_{W_1'}}\left[\log(S)\right]. \tag{1.3.5}$$

To find the GROW S-value, the marginal distribution in the alternative hypothesis having the smallest expected growth rate, even when paired with its 'best' (RIPr) marginal distribution from the null hypothesis to compose the S-value with, now has to be found (see (1.3.2)). GHK have shown that, under mild regularity conditions on $\mathcal{H}_0$ and $\mathcal{H}_1$, the S^* achieving (1.3.2) can be obtained through minimizing the Kullback-Leibler divergence between the two hypotheses over both sets of priors $W(\Theta_1)$ and $W(\Theta_0)$ simultaneously. This is called the *Joint Information Projection (JIPr)* of $\{P_{W_0} : W_0 \in W(\Theta_0)\}$ and $\{P_{W_1} : W_1 \in W(\Theta_1)\}$ onto each other, which can be regarded as an extension of the RIPr in (1.3.4):

$$S^*(z^n) = \frac{p_{W_1^*}(z^n)}{p_{W_0^*}(z^n)} \quad \text{with} \quad \min_{W_1\in W(\Theta_1)} \min_{W_0\in W(\Theta_0)} D(P_{W_1}||P_{W_0}) = D(P_{W_1^*}||P_{W_0^*}). \tag{1.3.6}$$

The regularity conditions needed for S^* to be the same in (1.3.2) and (1.3.6) invariably hold for the 2x2 table hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ considered in this thesis [GHK, p.20].

GHK have further shown that the distribution $P_{W_1^*}$ is the marginal distribution in $\mathcal{H}_1$ with the *worst-case expected capital growth, would we use this S-value for testing*, i.e. it holds that:

$$\forall\, \theta_1 \in \Theta_1 \;:\quad E_{P_{\theta_1}}\left[\log(S^*(Y^n))\right] \geq E_{P_{W_1^*}}\left[\log(S^*(Y^n)\right], \tag{1.3.7}$$

so we have a guarantee on the *performance* of the S-value: we can expect that when the data are in reality generated by any distribution $P_{\theta_1} : \theta_1 \in \Theta_1$ the expected capital growth will at least be $L := D(P_{W_1^*}||P_{W_0^*})$. This property of the GROW S-value will be used for testing if the S-values found through numerical approximation of the JIPr really satisfy the GROW criterion throughout this thesis, for example in Sections 2.4 and 3.2.

1.4 2x2 contingency tables

Now, the test we aim to develop a safe version for in this thesis will be introduced: the 2x2 contingency table test. Contingency tables summarize some outcomes $Y \in \{0,1\}$ in groups $X \in \{a,b\}$, i.e. a sequence of outcomes Y^n for n samples is observed, and the groups for those samples are fixed and defined through X^n. The 2x2 contingency table then depicts the counts of samples where $y_i = 1$ for $x_i = a$, $y_i = 1$ for $x_i = b$, and so forth, as illustrated in Table 1.1 (adapted from GHK). It is useful to analyze a contingency table, for example, when one aims to assess whether occurrences of the event Y differ between group a and b.

Table 1.1: 2x2 Contingency Table

	0	1	sum
a	n_{a0}	n_{a1}	n_a
b	n_{b0}	n_{b1}	n_b
sum	n_0	n_1	n

Under the null hypothesis, the occurrence of events is equally likely in group a and b: the number of ones n_1 is assumed to be independent of the groups. Hence, we can define the null hypothesis as $\mathcal{H}_0 = \{P_{\theta_0} : \theta_0 \in [0,1]\}$, with $P_{\theta_0} = \text{Bernoulli}(\theta_0)$. It then follows that the probability of observing a particular *sequence* $y^n \in \mathcal{Y}^n$ is:

$$p_{\theta_0}(y^n) = \theta_0^{n_1}(1-\theta_0)^{n0} \tag{1.4.1}$$

Let us now define the probability of observing a 1 in group a as θ_a, and the probability of observing a 1 in group b as θ_b. Under the alternative hypothesis, we assume there are two Bernoulli distributions: one for group a, and one for group b, with $\theta_a \neq \theta_b$. Thus, the alternative hypothesis can be defined as $\mathcal{H}_1 = \{P_{\theta_1} = P_{\theta_a,\theta_b} : (\theta_a,\theta_b) \in \Theta_1; \theta_a \neq \theta_b\}$, $\Theta_1 = [0,1]^2$. P_{θ_1} is the probability of observing a *sequence* $y^n \in Y^n$ given the group x^n [GHK]:

$$p_{\theta_1}(y^n|x^n) = \theta_a^{n_{a1}}(1-\theta_a)^{n_{a0}}\theta_b^{n_{b1}}(1-\theta_b)^{n_{b0}}. \tag{1.4.2}$$

Throughout this manuscript, $p_{\theta_1}(y^n|x^n)$ will be abbreviated as $p_{\theta_1}(y^n)$: we regard x^n (and thus n_a and n_b) as fixed.

For designing the GROW S-value in Section 2, the above distributions will be used to calculate the Bayes marginal mass function values for the null and alternative hypothesis. In the remainder of this section, two standard and established methods for analyzing 2x2 contingency tables will be introduced, as they will be later used for comparison in terms of power with the safe tests.

Fisher's exact test for 2x2 contingency tables

In the frequentist approach to hypothesis testing (Fisherian or Neyman-Pearsonian), Fisher's exact test is often used for the analysis of 2x2 contingency tables. In Fisher's exact test, one conditions on n_a, n_b and n_1. Then, all counts in all cells except for one are fixed, and under the null hypothesis the counts in the free cell follow a hypergeometric distribution [Fisher, 1935b]. The probability of observing n_{a1} ones in group a when given n_a, n_b and n_1 according to the hypergeometric distribution

is [Ross, 2010]:

$$h(n_{a1}) \coloneqq P_{HG}(N_{a1} = n_{a1}) = \frac{\binom{n_a}{n_{a1}}\binom{n_b}{n_{b1}}}{\binom{n}{n_1}} = \frac{\binom{n_a}{n_{a1}}\binom{n_b}{n_1 - n_{a1}}}{\binom{n}{n_1}}. \tag{1.4.3}$$

The two-sided p-value for Fisher's exact test is then calculated by summing all hypergeometric probabilities for all possible values of the cell count that are *equal to or smaller than* the probability of the observed count [Fisher, 1935a]. This will be denoted as f:

$$f(n_{a1}) \coloneqq \sum_{i:h(i)\leq h(n_{a1})} h(i) \text{ with } i \in \{0, \ldots, n_a\}. \tag{1.4.4}$$

In other words, Fisher's exact test represents the chance of observing a result *as or more extreme* as the observed result under the null hypothesis, i.e. Fisher's exact test is a p-value (see (1.1.2)).

As described in Section 1.2, a p-value can only be used to define a safe test when $\forall P \in \mathcal{H}_0 : \mathbb{E}_P[\frac{1}{\text{pval}}] \leq 1$. It can be shown that Fisher's exact test does not define an S-value in this way, i.e. the expected value of $\frac{1}{f}$ is bigger than 1 under the null hypothesis. We can put an upper bound on $f(n_{a1})$, so we can put a lower bound on the expectation of $\frac{1}{f}$. If we consider the case where $n_a = n_b$, the hypergeometric distribution is symmetric around $n_{a1} = \frac{1}{2}n_1$. It then follows that:

$$f(n_{a1}) \leq 2\sum_{i=0}^{n_{a1}} h(i) \quad \text{for } n_{a1} \leq \left\lceil \frac{1}{2}n_1 \right\rceil - 1 \tag{1.4.5}$$

$$\leq 2(n_{a1}+1)h(n_{a1}). \tag{1.4.6}$$

Since the hypergeometric distribution is symmetric around $n_{a1} = \frac{1}{2}n_1$, so is the value of $f(n_{a1})$. Hence:

$$\mathbb{E}_{P_{HG}}\left[\frac{1}{f(N_{a1})}\right] = \sum_{n_{a1}=0}^{n_1} h(n_{a1})\frac{1}{f(n_{a1})} \tag{1.4.7}$$

$$\geq 2\sum_{n_{a1}=0}^{\lceil \frac{1}{2}n_1 \rceil - 1} h(n_{a1})\frac{1}{2(n_{a1}+1)h(n_{a1})} \tag{1.4.8}$$

$$= \sum_{n_{a1}=0}^{\lceil \frac{1}{2}n_1 \rceil - 1} \frac{1}{n_{a1}+1} \tag{1.4.9}$$

$$> 1, \tag{1.4.10}$$

and the expectation diverges as $n_1 \to \infty$.

Bayes factors for contingency tables

The S-values for 2x2 contingency tables will also be compared to an existing Bayes factor for analyzing 2x2 contingency tables. 'Objective' Bayes factors for 2x2 contingency tables have been proposed Gunel and Dickey [1974], and have recently been implemented by Jamil et al. [2017]. In these GD (Gunel-Dickey) Bayes factors, standard forms of Dirichlet and Gamma *objective* priors on proportions and cell counts are proposed, resulting in analytically solvable posterior distributions

[Gunel and Dickey, 1974]. For a multinomial sampling plan, where n_a and n_b are fixed, this yields:

$$B_{10} = \frac{\binom{n}{n_a}}{\binom{n_0}{n_{a0}}\binom{n_1}{n_{a1}}} \frac{(n+1)}{(n_0+1)(n_1+1)}. \tag{1.4.11}$$

The performance of the GD Bayes factor as defined here and Fisher's exact test will be compared to the performance of the various safe tests proposed in Chapter 4.

1.5 Thesis outline

In Chapter 2, it will be explored whether the GROW S-value for 2x2 contingency tables can be found through finding the JIPr as described in (1.3.6) using numerical optimization. Optimization results will be compared to a theoretical minimum, and the obtained S-values will be tested numerically for indeed achieving the best growth in the worst case, and for *being a safe test*. In Chapters 3 and 4, two different versions of the GROW S-value for 2x2 contingency tables will be developed: the *default (simple)* S-value and the *KL-based S-value*. Both S-values will be compared with respect to the *power* they achieve. In Chapter 5, it will be elaborated on whether the found S-values could be used for *optional stopping*, in addition to optional continuation (the difference between the two will be elaborated on there). In Chapter 6, the performance of both S-values on real-world examples will be evaluated, and it will be discussed if there are cases when one of the two versions is preferable. In Chapter 7 it will be investigated whether an extension of the S-value for 2x2 contingency tables to the Cochran-Mantel-Haenszel test can be constructed.

Chapter 2

Finding the GROW S-value for 2x2 tables

To determine the GROW S-value for 2x2 contingency tables, it will first be investigated whether the JIPr between P_{W_1} and P_{W_0} as described by (1.3.6) can be found through numerical optimization. To convert (1.3.6) into a problem numerically solvable, first, it will be defined how Θ_1 and Θ_0 will be discretised. Next, the BFGS optimisation method will be introduced. In the third subsection, an example of a relatively simple $\mathcal{H}_1$ will be considered, for which the JIPr will be determined analytically. The JIPr found through optimization will be compared to this analytical solution to ensure the BFGS algorithm can really provide a solution close to the JIPr. In the last section, a more intricate example with composite hypotheses will be considered. As the JIPr is not known analytically here, it will be checked *numerically* whether the candidate S-value found through BFGS is a real S-value, and whether the minimized KL-divergence indeed equals its expected growth rate in the worst case.

2.1 Finding the GROW S-value for a *subset* of the alternative hypothesis

As mentioned in Section 1.3, we will require $\mathcal{H}_1$ and $\mathcal{H}_0$ to be non-overlapping. However, in the 2x2 scenario, they do (almost) overlap, as the difference between θ_a and θ_b in the alternative hypothesis can take on values arbitrarily close to 0. One can observe that if we would let $\mathcal{H}_1$ and $\mathcal{H}_0$ contain very similar distributions, the Kullback-Leibler divergence of the JIPr would take on values close to zero. Then, the GROW S-value would tend to the constant 1 independent of the data in the worst case under the alternative hypothesis, and we would not gain any capital during testing. To prevent this, we define a *restricted* alternative hypothesis depending on some divergence measure δ to determine the JIPr for: $\Theta_1(\delta) \subset \Theta_1$ [GHK].

For the 2x2 contingency table tests, two divergence measures are proposed to be used for δ. First, a *default* S-value is defined as the S-value based on $\Theta_1(\delta)$, with [GHK]:

$$\Theta_1(\delta) = \{\theta : |\theta_a - \theta_b| \geq \delta, \theta \in \Theta_1\}, \tag{2.1.1}$$

with $(\theta_a, \theta_b) \in [0,1]^2$. In other words, the default S-value includes all distributions in the restricted alternative hypothesis where there is a certain minimal absolute difference between the group means. This is a natural choice for the restricted alternative hypothesis, as researchers in practice often have to prespecify which difference between proportions in groups they consider *substantively relevant.*

Secondly, a *Kullback-Leibler (KL)*-based S-value is proposed, where δ represents the minimal *Reverse Information Projection (RIPr)* KL-divergence between each distribution $P_{\theta_1} \in \mathcal{H}_1$ on

$\{P_{W_0} : W_0 \in W(\Theta_0)\}$. Thus, for the KL-based divergence measure δ we have [GHK]:

$$\Theta_1(\delta) = \left\{ \theta_1 : \min_{W \in W(\Theta_0)} D(P_{\theta_1} || P_W) \geq \delta \right\}. \tag{2.1.2}$$

An intuitive explanation for choosing the KL-based divergence measure is now presented. If we were really sure we only wanted to test a singleton $\mathcal{H}_1$, the GROW S-value would be found through RIPr to find the corresponding $P_{W_0^*}$. The expected capital growth would then reduce to the KL divergence between P_{θ_1} and $P_{W_0^*}$. We could now compose $\Theta_1(\delta)$ as the set of all $\theta_1 \in \Theta_1$ that would achieve a certain expected capital growth, would we be certain the data were generated under that P_{θ_1}. In other words, we set a boundary on the minimum expected growth we want the distributions to achieve, would we believe them to generate the data; we do not include distributions in our set that would yield a small probability of rejecting the null hypothesis if we believed they were true.

Not the entire set of the parameters in $\Theta_1(\delta)$ needs to be used for finding the JIPr, as we know that the distributions that will yield the worst case capital growth are the ones closest or most similar to Θ_0. Hence, for finding the JIPr, we only consider the distributions where $|\theta_a - \theta_b| = \delta$ for the default S-value, and we only consider the distributions where $D(P_{\theta_1} || P_{W_0^*}) = \delta$ for the KL-based S-value. These subsets of $\Theta_1(\delta)$ will be denoted as $\bar{\Theta}_1(\delta)$.

To now convert (1.3.6) to a problem numerically solvable, $\bar{\Theta}_1(\delta)$ and Θ_0 need to be discretized. According to Carathéodory's theorem, any point x in a convex set $R \subset \mathbb{R}^k$ is in the convex hull of a subset $R' \subset R$, with $|R'| \leq k+1$ [Carathéodory, 1907]. Since the probability distributions on $\mathcal{Y}^n$ for $\mathcal{H}_1$ and $\mathcal{H}_0$ are completely determined by the distributions on the sufficient statistics n_{a1} and n_{b1}, and these can only take on a finite number of values, the probability distributions are determined by k components, for some finite k. It follows that instead of using the Bayes marginals over the entire sets $\bar{\Theta}_1(\delta)$ and Θ_0, we can restrict our search to a k-dimensional vector of parameters for $\mathcal{H}_1$ and $\mathcal{H}_0$ [GHK]. Hence, the marginal distributions simplify to:

$$p_{W_j}(y^n) = \sum_{i=1}^{k} w_j(\theta_{j.i}) p_{\theta_i}(y^n), \tag{2.1.3}$$

with W_j a distribution given by a probability vector $\vec{w}_j = (w_j(\theta_{j.1}), w_j(\theta_{j.1}), \ldots, w_j(\theta_{j.k}))$. It is unknown *which* support points should exactly be included in these grids: grid precision and location of grid points will be experimented with in Chapter 4. For now, we assume that grid points $(\theta_{0.1}, \ldots, \theta_{0.k})$ and $(\theta_{1.1}, \ldots, \theta_{1.k})$ have been given, and write $\vec{w}_{j.i} = w_j(\theta_{j.i})$, so $\vec{w}_j = (\vec{w}_{j.1}, \ldots, \vec{w}_{j.k})$.

2.2 BFGS for finding the JIPr

To find the JIPr, constrained versions of the gradient descent method 'BFGS' were implemented. BFGS stands for Broyden, Fletcher, Goldfarb and Shanno, the names of the authors who all independently developed and published about the method in 1970 [Papakonstantinou, 2011]. BFGS is a quasi-Newton algorithm for optimisation. Newton's algorithm for optimization is a form of gradient descent. Would we aim to minimize a function f with respect to the parameter vector $\vec{w} = (\vec{w}_0, \vec{w}_1)$, an update at step t of the algorithm for the current parameter vector $\vec{w}^{(t)}$ would be [Bonnans et al., 2006]:

$$\vec{w}^{(t+1)} = \vec{w}^{(t)} - \alpha \mathbf{H}(\vec{w}^{(t)})^{-1} f'(\vec{w}^{(t)}), \tag{2.2.1}$$

with α a certain step size, and $\mathbf{H}(\vec{w}^{(t)})^{-1}$ the inverse of the Hessian; i.e. at each step $\vec{w}^{(t)}t$ is updated in the opposite direction of the gradient, and this step is adjusted by the Hessian. One of the properties of Newton's algorithm is that it converges fast to the closest extremal point, which could be a problem when dealing with local minima or maxima [Bonnans et al., 2006]. However, as Kullback-Leibler divergence $D(P_{W_1}||P_{W_0})$ is convex in both P_{W_1} and P_{W_0}, finding the JIPr is a double convex minimization problem [Cover and Thomas, 1990] [Van Erven and Harremoës, 2014]. Another difficulty of Newton's algorithm is the Hessian: for $\vec{w} \in [0,1]^n$, the Hessian is an $n \times n$ matrix, hence, computational complexity for determining the inverse of the Hessian at each step grows quadratically with the number of parameters to optimize over.

This is where quasi-Newton methods like BFGS come in; with those, the inverse of the Hessian is not calculated exactly, but an approximation is used. After each update of $\vec{w}_t$, the inverse of the Hessian for the next update step $\mathbf{H}(\vec{w}^{(t+1)})^{-1}$ is approximated with the current Hessian $\mathbf{H}(\vec{w}^{(t)})$ and some correction. This correction satisfies the *quasi Newton equation*: this equation ensures that the estimated Hessian represents the mean change in the derivatives between $\vec{w}^{(t)}$ and $\vec{w}^{(t+1)}$, and thus behaves like the Hessian at least in that interval [Bonnans et al., 2006].

Since the parameters to optimize over for the JIPr are the *prior distributions* on Θ_0 and $\bar{\Theta}_1(\delta)$, they are bounded by the following constraints:

$$\vec{w}_0 \geq 0, \tag{2.2.2}$$

$$\vec{w}_1 \geq 0, \tag{2.2.3}$$

$$\sum_{i=1}^{k} \vec{w}_{0.i} = 1, \tag{2.2.4}$$

$$\sum_{i=1}^{k} \vec{w}_{1.i} = 1. \tag{2.2.5}$$

Hence, *constrained* variants of BFGS were implemented for finding the JIPr. Note that the equality sign in (2.2.4) should be switched to an inequality sign when one is not sure the regularity conditions discussed in Section 1.3 are met. Experiments in this thesis were performed with an inequality sign in the constraint for the weights in on the parameters in the null hypothesis, but resulting priors were proper in all cases, i.e. all weights in the null hypothesis consistently summed to 1 (for example see Figures 2.1 and 4.2). Two variants of constrained optimization were compared: an implementation of sequential least-squares quadratic programming, `slsqp`, from the `nloptr` package [Johnson, 2019], and the `constrOptim` function from the standard `stats` package.

A first experiment was conducted with both methods to investigate how close both methods came to the KL-divergence minimum, and to compare computation times. A default $\Theta_1(\delta)$ with $\delta = 0.6$ was defined, and the grid precision for the alternative and null hypothesis was varied. Results are summarized in Table 2.1. It can be observed that `slsqp` from the `nloptr` package yielded lower computations times than `ConstrOptim`, even tough the `slsqp` algorithm iterates 1000 times with the default settings, and `ConstrOptim` 100 times. Further, the obtained minima were lower for `slsqp`, indicating that the results found with `slsqp` are closer to the GROW S-value. Therefore, it was decided to continue with the `slsqp` algorithm for the rest of the experiments and

Table 2.1: Computation time and obtained KL divergence minima for sequential quadratic programming and constrained optimisation algorithms. The default settings for each of the algorithms were used: the constrained optimisation algorithm uses 100 iterations, and the quadratic programming algorithm 1000 iterations with the default settings. The JIPr was searched for with the default $\bar{\Theta}_1(\delta)$ with $\delta = 0.6$, and for $n_a = n_b = 10$.

Size of discretized versions of $\bar{\Theta}_1$ and Θ_0	time (s)		minimum	
	`ConstrOptim`	`slsqp`	`ConstrOptim`	`slsqp`
5	0.2429	0.0219	4.8441	4.8363
10	1.0383	0.0210	3.2117	3.2105
15	2.4189	0.0292	3.2014	3.1678
20	2.7802	0.4284	3.2001	3.1757
25	3.1426	0.3816	3.2005	3.1660
30	3.8401	0.7012	3.2416	3.1699
35	4.5202	0.4516	3.2584	3.1657
40	5.4509	0.8427	3.2630	3.1680
45	6.1220	1.2686	3.2863	3.1656
50	6.9703	3.2390	3.3063	3.1671

in the implementation of the package.

It was noted that `slsqp` in some cases yielded small deviations from the equality constraint $w_1 \geq 0$ and $w_0 \geq 0$; the algorithm could sometimes return weights with a very small negative value, close to machine precision. This is illustrated in Table 2.2. Therefore, all weights smaller than 0 were set to a very small positive value (1e-16), and weights were standardised to sum to 1 again. Since S^* then has been adapted, it is not guaranteed these weights provide the GROW S-value anymore. Therefore, after correction the set of weights were offered to the `ConstrOptim` as start parameters, to check if these weights indeed provided the GROW S-value, and a lower minimum was indeed not found. This check was implemented in the final software as well.

Table 2.2: Weights for $\bar{\Theta}_1(0.6)$ after finding the JIPR with sequential quadratic programming, the size of the grid representing $\bar{\Theta}_1(0.6)$ was equal to 10. Note the very small negative weights assigned to the fifth, sixth and tenth parameter pairs.

θ_a	θ_b	$w_{1,\theta}$
0.00	0.60	2.06e-17
0.10	0.70	3.25e-2
0.20	0.80	4.36e-1
0.30	0.90	3.15e-2
0.40	1.00	-2.83e-16
0.60	0.00	-2.75e-16
0.70	0.10	3.15e-2
0.80	0.20	4.36e-1
0.90	0.30	3.15e-2
1.00	0.40	-1.10e-16

2.3 Comparing BFGS results to a theoretical minimum

To investigate whether the parameters achieving the minimum in (1.3.6) could indeed be found with BFGS, a situation where the location of the minimum can essentially be determined analytically was considered.

To do this, we can use a second order Taylor approximation of the expected value of $\log(p_{\theta_0}(y^n))$ around θ_1 to create an asymptotic expression for the KL divergence between two distributions P_{θ_1} and P_{θ_0} that are both members of an exponential family, as is the case here [Grünwald, 2007]:

$$D(P_{\theta_1}||P_{\theta_0}) = \frac{1}{2}(\theta_1 - \theta_0)^T I(\theta_1)(\theta_1 - \theta_0) + O(||\theta_1 - \theta_0||^3). \tag{2.3.1}$$

It can be observed that the Euclidean distance between θ_1 and θ_0 (scaled by the Fisher information of θ_1) provides an approximation of the KL divergence $D(P_{\theta_1}||P_{\theta_0})$. If we now consider a singleton $\mathcal{H}_1$, i.e. we have a fixed value for θ_1, this suggests that minimizing the Euclidian distance between θ_1 and θ_0 with respect to $\theta_0 \in [0,1]$ could minimize $D(P_{\theta_1}||P_{\theta_0})$. We will now first proceed to showing that the S-value $S(y^n) = \frac{p_{\theta_1}(y^n)}{p_{\theta_0}(y^n)}$ with θ_0 the parameter minimizing the Euclidean distance *is an S-value* for $\mathcal{H}_1 = \{P_{\theta_1}\}$; i.e. its expected value under all $P_{\theta_0} : \theta_0 \in [0,1]$ is smaller than or equal to 1. Thereafter, we will show that this S-value also *provides the RIPr* for this singleton $\mathcal{H}_1$ and $\mathcal{H}_0 = \{P_{\theta_0} : \theta_0 \in [0,1]\}$. Since we consider a singleton $\mathcal{H}_1$, the RIPr is identical to *the GROW S-value.*

The Euclidean distance between θ_1 and θ_0 equals:

$$\begin{aligned} ||\theta_1 - \theta_0||^2 &= (\theta_a - \theta_0)^2 + (\theta_b - \theta_0)^2 \\ &= 2\theta_0^2 - 2\theta_0(\theta_a + \theta_b) + \theta_a^2 + \theta_b^2, \end{aligned} \tag{2.3.2}$$

and differentiating with respect to θ_0 yields:

$$\frac{d}{d\theta_0}||\theta_1 - \theta_0||^2 = 4\theta_0 - 2(\theta_a + \theta_b). \tag{2.3.3}$$

It follows that the Euclidean distance between θ_1 and θ_0 is minimized for $\theta_0 = 0.5(\theta_a + \theta_b)$. We define the candidate S-value constructed with these distributions as S^*:

$$S^*(y^n) = \frac{p_{\theta_1}(y^n)}{p_{\theta_0^*}(y^n)} \quad \text{with} \quad \theta_0^* = 0.5(\theta_a + \theta_b). \tag{2.3.4}$$

The observed number of ones in a given sample y^n will now be denoted as $n_1(y^n)$. Under the other distributions $P_{\theta_0} : \theta_0 \in [0,1]$, the expected value of S^* then equals:

$$\mathbb{E}_{P_{\theta_0}}[S^*(Y^n)] = \sum_{y^n} p_{\theta_0}(y^n)\frac{p_{\theta_1}(y^n)}{p_{\theta_0^*}(y^n)} = \sum_{y^n} \theta_0^{n_1(y^n)}(1-\theta_0)^{n-n_1(y^n)}\frac{p_{\theta_1}(y^n)}{p_{\theta_0^*}(y^n)}. \tag{2.3.5}$$

One can observe that the expected value of S^* is equal to one if $\theta_0 = \theta_0^*$. We will now show that this is the *maximal* expectation obtained for any $\theta_0 \in \Theta_0$: we will show that the expectation of S^* has an extremal point when evaluated at $\theta_0 = \theta_0^*$, i.e. the derivative at that point is 0, and that this is indeed a maximum. The derivative of the expected value with respect to θ_0 equals:

$$\begin{aligned}\frac{\partial}{\partial\theta_0}\mathbb{E}_{p_{\theta_0}}\left[S^*(Y^n)\right] &= \sum_{y^n}(n_1(y^n)(1-\theta_0) - (n-n_1(y^n))\theta_0)\ \theta_0^{n_1(y^n)-1}(1-\theta_0)^{n-n_1(y^n)-1}\frac{p_{\theta_1}(y^n)}{p_{\theta_0^*}(y^n)} \\ &= \sum_{y^n}(n_1(y^n) - n\theta_0)\ \theta_0^{n_1(y^n)-1}(1-\theta_0)^{n-n_1(y^n)-1}\frac{p_{\theta_1}(y^n)}{p_{\theta_0^*}(y^n)}. \end{aligned} \tag{2.3.6}$$

Evaluating this derivative at $\theta_0 = \theta_0^*$ yields:

$$\begin{aligned}\frac{\partial}{\partial\theta_0}\mathbb{E}_{p_{\theta_0}}[S^*]_{|\theta_0=\theta_0^*} &= \sum_{y^n}(n_1(y^n) - n\theta_0^*)\ \theta_0^{n_1(y^n)-1}(1-\theta_0)^{n-n_1(y^n)-1}\frac{p_{\theta_1}(y^n)}{\theta_0^{*n_1(y^n)}(1-\theta_0^*)^{n-n_1(y^n)}} \\ &= \sum_{y^n}(n_1(y^n) - n\theta_0^*)\frac{p_{\theta_1}(y^n)}{\theta_0^*(1-\theta_0^*)} \\ &= \sum_{y^n} n_1(y^n)\ \frac{p_{\theta_1}(y^n)}{\theta_0^*(1-\theta_0^*)} - \sum_{y^n} n\ \frac{p_{\theta_1}(y^n)}{(1-\theta_0^*)} \\ &= \frac{\sum_{y^n}[n_1(y^n)p_{\theta_1}(y^n)]}{\theta_0^*(1-\theta_0^*)} - \frac{n}{(1-\theta_0^*)} \\ &= \frac{\mathbb{E}_{P_{\theta_1}}[N_1]}{\theta_0^*(1-\theta_0^*)} - \frac{n}{(1-\theta_0^*)} \\ &= \frac{n_a\theta_a + n_b\theta_b}{\theta_0^*(1-\theta_0^*)} - \frac{n}{(1-\theta_0^*)} \quad \text{and, since } n_a = n_b: \\ &= \frac{\frac{1}{2}(\theta_a+\theta_b)n}{\theta_0^*(1-\theta_0^*)} - \frac{n}{(1-\theta_0^*)} \\ &= \frac{\theta_0^* n}{\theta_0^*(1-\theta_0^*)} - \frac{n}{(1-\theta_0^*)} = 0. \end{aligned} \tag{2.3.7}$$

In the same way we can evaluate the second derivative to illustrate that a (global) maximum is achieved at θ_0^*. It follows that

$$\forall\theta_0 \in \Theta_0: \quad \mathbb{E}_{P_{\theta_0}}\left[S^*(Y^n)\right] \leq 1, \tag{2.3.8}$$

and we can conclude that S^* is an S-value.

Lemma 1.3 of GHK shows that when there exists a probability distribution P_{θ_0} that provides the RIPr of P_{W_1} on P_{W_0} and thus results in a *safe* test, this distribution is the *unique* one resulting in the RIPr: there does not exist any other marginal distribution providing the RIPr for that P_{W_1}. Therefore, we must conclude that that S^* is the only S-value and hence the GROW S-value.

In conclusion, we now have derived that for a singleton $\mathcal{H}_1 = \{\theta_1 = (\theta_a, \theta_b)\}$, all mass in W_0^* should be put on $\theta_0 = 0.5(\theta_a + \theta_b)$ in the JIPr. To test the BFGS algorithm, the JIPr was searched for with $n_a = n_b = 10$, a singleton $\mathcal{H}_1$ with $\theta_1 = (0.3, 0.7)$ and 7 equally spaced grid points on $[0, 1]$ representing Θ_0. Results are summarized in Table 2.3, it can be observed that indeed all mass was put on $\theta_0 = 0.5 = 0.5(\theta_a + \theta_b)$. Results were similar when different starting parameters were used for the BFGS algorithm. Hence, the BFGS algorithm appears to succeed in finding the JIPr for the 2x2 tables safe test in this simple instance.

Table 2.3: Weights obtained for the RIPr (JIPr with a singleton $\mathcal{H}_1$, $\{\theta_a = 0.3, \theta_b = 0.7\}$) for $n_a = n_b = 10$ with different starting parameters: uniform, random, and with all mass put on a 'wrong' distribution.

	Uniform		Random		One point	
θ_0	start	result	start	result	start	result
0.00	0.14	0.00	0.12	0.00	1.00	0.00
0.17	0.14	0.00	0.17	0.00	0.00	0.00
0.33	0.14	0.00	0.05	0.00	0.00	0.00
0.50	0.14	1.00	0.17	1.00	0.00	1.00
0.67	0.14	0.00	0.20	0.00	0.00	0.00
0.83	0.14	0.00	0.12	0.00	0.00	0.00
1.00	0.14	0.00	0.17	0.00	0.00	0.00

2.4 Extending BFGS to composite hypotheses

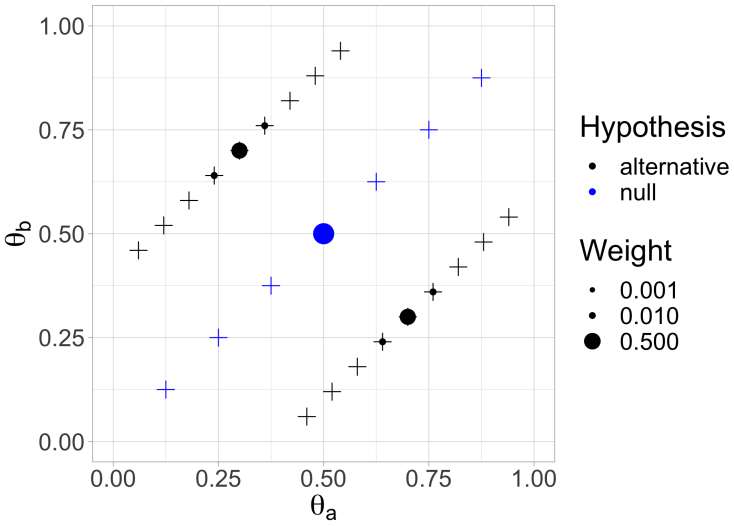
The performance of BFGS for finding the JIPr will now be assessed in a situation extendable to hypothesis testing: a default $\Theta_1(\delta)$ was defined with $\delta = 0.4$, and represented by an equally spaced grid consisting of 18 points (see Figure 2.1a). Θ_0 was represented by a grid of 7 points on $[0, 1]$, and JIPr was performed for $n_a = n_b = 10$. Results are summarized in Figure 2.1a: almost all mass in W_1^* was put on the two points on the line $\theta_b = 1-\theta_a$. In W_0^*, almost all mass was put on $\theta_0 = 0.5$.

It was checked numerically whether the found W_1^* and W_0^* indeed provided an S-value through investigating the expected value of S under various $P_{\theta_0} \in \mathcal{H}_0$. As illustrated in Figure 2.1b, the expected value was less than or equal to 1 for all investigated distributions from the null hypothesis. Hence, we conclude that the S-value found through BFGS indeed provides a *safe* test.

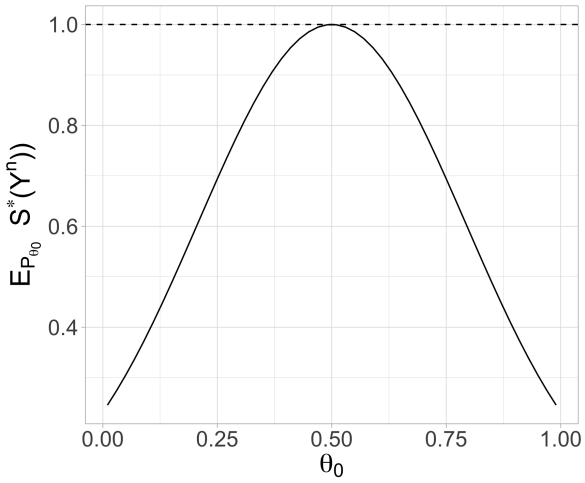
To next investigate if the found S-value was GROW, it was investigated whether the expected growth of S^* under a selection of P_{θ_1}, $\theta_1 \in \bar{\Theta}_1(\delta)$ was not lower than the expected growth under $P_{W_1^*}$, i.e. if the *worst-case* property of GROW S-values as described in (1.3.7) was satisfied. This was investigated for a grid of $\theta_1 \in \bar{\Theta}_1(\delta)$. In Figure 2.1c it is illustrated that for this S^* found with BFGS, the expected capital growth under these P_{θ_1} is higher than or approximately equal to the expected capital growth under $P_{W_1^*}$; this can be interpreted as a confirmation that the found S-value is indeed GROW. From this experiment, we can conclude that would we use this GROW S-value for testing, the expected capital growth would be at least 1.03 under all distributions in $\Theta_1(\delta)$.

Conclusion

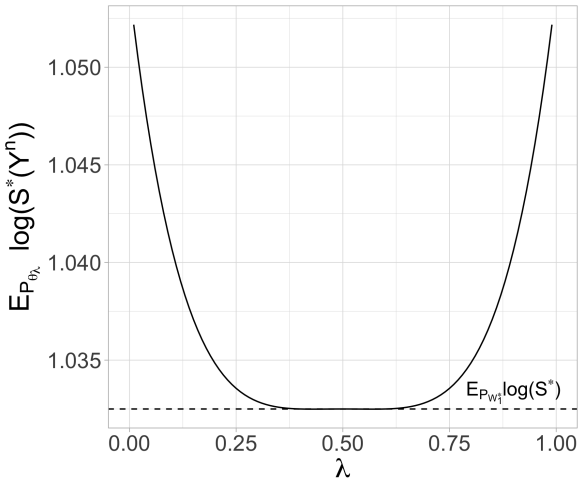
In this chapter, it has been illustrated that the GROW S-value can be found with BFGS. In a simple case, the JIPr obtained with BFGS matched the JIPr theoretically determined. For composite hypotheses, it was numerically determined that the found S-value was safe and GROW. Throughout the next chapters, $\Theta_1(\delta)$ defined with the two different divergence measures and the patterns of the resulting GROW S-values will be investigated.



(a) $\bar{\Theta}_1(\delta)$, Θ_0 and obtained weights



(b) Expected value of S under the null hypothesis



(c) Expected capital growth under various distributions in $\bar{\Theta}_1(\delta)$

Figure 2.1: JIPr found with BFGS for the case where $n_a = n_b = 10$, $\delta = 0.4$. (a): the obtained weights. Crosses indicate positions of grid points in Θ_1 and Θ_0, dots and their size represent the weight from W_0^* and W_1^* for each point. (b): the expected value of S was calculated under P_{θ_0} for a grid of θ_0 values on $[0, 1]$. (c): several θ_a and θ_b combinations that could generate the data under the restricted alternative hypothesis were composed as $\theta_a = 0\lambda + 0.6(1-\lambda)$, and $\theta_b = 0.4\lambda + 1(1-\lambda)$. The dotted line indicates the expected capital growth when data are generated under $P_{W_1^*}$

Chapter 3

The *simple* S-value

In this chapter, *default* S-values are studied: S-values resulting from defining $\Theta_1(\delta)$ with the absolute difference between the group means as a divergence measure. We already saw examples of the JIPr found through BFGS for this $\Theta_1(\delta)$ in the previous section. Given the result in Figure 2.1 it appears that the algorithm converges to a prior where almost all mass is put on the two points on the line $\theta_b = 1 - \theta_a$ with $|\theta_a - \theta_b| = \delta$ in the restricted alternative hypothesis, and on $\theta_0 = 0.5$ in the null. This suggests that if we would put not just almost, but *all* mass on those points, we would still get an S-value. If this were true, one could skip numerically determining the JIPr for finding the GROW S-value for the default case and always use this *simple S-value*, with a huge decrease in computational complexity.

In this chapter, we will first show that a simple S-value is equivalent to the GROW S-value in the *one-sided case*. Thereafter, it will be proven that the simple S-value extended to the two-sided case indeed provides a *safe* test, i.e. it is an S-value. However, it will then numerically illustrated that the simple S-value is not GROW for the entire $\Theta_1(\delta)$ in two-sided testing. Nonetheless, it will also numerically be shown that the subset of $\Theta_1(\delta)$ the simple S-value is not GROW for is very small, and that the simple S-value *becomes* GROW for large n. In the third subsection, the adjustments needed to be made to the simple S-value for the case where $n_a \neq n_b$ are discussed. In the last subsection, design choices for the simple S-value are discussed: it will be investigated how sample size, the distance measure δ and expected capital growth or power relate.

3.1 The simple S-value for one-sided testing when $n_a = n_b$

Sometimes, scenarios arise where a researcher is interested in testing an asymmetric or *one-sided* alternative hypothesis. This is relevant when one can hypothesize the nature of the effect of a predictor or grouping variable before conducting the experiment and hypothesis test. For example, when considering testing the effects of a new drug compared to the effects of a placebo in the clinical setting, one would be interested in testing whether the occurrence of disease symptoms is decreased in the group of patients receiving the drug, and not in testing whether the occurrence of symptoms is increased. Or, in penalized regression, one could be interested in knowing which mutations in tumour suppressor genes in cancer tissue have a positive effect on response to treatment.

In the setting of significance testing, extending the alternative hypothesis to two sides when one is really only interested in testing one side would lead to unnecessary loss of power. When a test is two-sided, the set of extreme test statistics is twice as big as for one-sided testing, and the corresponding p-values are higher for the same observations (and S-values are lower), requiring more extreme results before the significance level is reached. Hence, when one is very sure of a

one-sided nature of ones hypothesis before inspecting one's data, it is preferable to use a one-sided test.

For the one-sided case where one hypothesizes that *the proportion in group b is bigger than in group a*, the null hypothesis as defined in Section 1.4 is tested against the alternative hypothesis:

$$\mathcal{H}_1' = \{P_{\theta_1} = P_{\theta_a,\theta_b} : \theta_1 \in \Theta_1, \theta_b > \theta_a\}. \tag{3.1.1}$$

As with the two sided test, for designing the S-value, we restrict ourselves to a subset of the alternative hypothesis: $\Theta_1'(\delta) = \{\theta_1 : \theta_b - \theta_a \geq \delta, \theta_1 \in \Theta_1\}$. In Appendix, Figure A.1a, the JIPr found with BFGS for the one-sided scenario with $\delta = 0.6$ is illustrated. In the restricted alternative hypothesis, one support point on on the line $\theta_b = 1 - \theta_a$ was found, and in the null hypothesis, all mass was put on $\theta_0^* = 0.5$. We propose the corresponding simple S-value to be defined as follows [1]:

$$S(y^n) = \frac{p_{\theta_1^*}(y^n)}{p_{\theta_0^*}(y^n)} \quad ; \quad \theta_1^* = \left(\frac{1}{2} - \frac{1}{2}\delta, \frac{1}{2} + \frac{1}{2}\delta\right). \tag{3.1.2}$$

One can observe that this definition coincides with the S-value presented in Section 2.3; hence we are already certain the one-sided simple S-value provides a safe test.

It will now numerically illustrated that when testing a one-sided alternative hypothesis with $n_a = n_b$, the simple S-value is the *GROW* S-value, e.g. expected growth rate of the S-value under $P_{\theta_1^*}$ is smaller than or equal to the expected growth rate under all other distributions in the restricted alternative hypothesis. The expected capital growth of the simple S-value under different distributions from $\Theta(\delta)$ for $\delta = 0.6$ is depicted in Appendix Figure A.1b: it can be observed that the expected capital growth is now equal to the worst-case capital growth for all $P_{\theta_1} \in \mathcal{H}_1$, and that the GROW S-value has been achieved. Note that this is a numerical proof only; we have not given a formal proof here.

3.2 The simple S-value for two-sided testing when $n_a = n_b$

The simple S-value for one-sided testing presented above will now be extended to the two-sided case. Let us define Θ_1 as the default $\Theta_1(\delta)$ (see (2.1.1)). For the case where $n_a = n_b$, the hypothesized $P_{W_1^*}$ corresponding to *the simple S-value is then defined as*:

$$p_{W_1^*}(y^n) := \frac{1}{2}\left(p_{\theta_{1.1}}(y^n) + p_{\theta_{1.2}}(y^n)\right), \tag{3.2.1}$$

with:

$$\theta_{1.1} := \left(\frac{1}{2} - \frac{1}{2}\delta, \frac{1}{2} + \frac{1}{2}\delta\right), \tag{3.2.2}$$

and:

$$\theta_{1.2} := \left(\frac{1}{2} + \frac{1}{2}\delta, \frac{1}{2} - \frac{1}{2}\delta\right). \tag{3.2.3}$$

[1] If one would want to test the hypothesis that the proportion in group a is bigger than in group b, one would have to swap θ_a and θ_b in this definition.

The proposed $p_{W_0^*}$ corresponds to p_{θ_0} with $\theta_0^* = 0.5$, i.e $w_0^*(\theta_0^*) = 1$. It is defined as Bernoulli(0.5) and consequently assigns the same probability to every $y^n \in \mathcal{Y}^n$:

$$p_{\theta_0^*}(y^n) = 0.5^{n_1}(1-0.5)^{n-n_1} = 0.5^n. \tag{3.2.4}$$

We will now prove that this proposed simple S-value $\frac{p_{W_1^*}}{p_{\theta_0^*}}$ *is* an S-value. Let us define:

$$S_1(y^n) = \frac{p_{\theta_{1.1}}(y^n)}{p_{\theta_0^*}(y^n)}, \quad \text{and} \quad S_2(y^n) = \frac{p_{\theta_{1.2}}(y^n)}{p_{\theta_0^*}(y^n)}. \tag{3.2.5}$$

In Section 2.3 it was illustrated that, for $n_a = n_b$, likelihood ratios between $p_{\theta_1} : \theta_1 = (\theta_a, \theta_b)$ and a $p_{\theta_0^*}$, where $\theta_0 = \frac{1}{2}(\theta_a + \theta_b)$, are S-values, i.e. their expected value is smaller than or equal to 1 for all $P_{\theta_0} \in \mathcal{H}_0$. For S_1 and S_2 this yields:

$$\frac{1}{2}(\theta_a + \theta_b) = \frac{1}{2}\left(\frac{1}{2} - \frac{1}{2}\delta + \frac{1}{2} + \frac{1}{2}\delta\right) = \frac{1}{2} = \theta_0^*, \tag{3.2.6}$$

and we can conclude that S_1 and S_2 are S-values. We can observe that any convex combination of S_1 and S_2 also yields an S-value:

$$S_{\text{new}}(y^n) = w_1 S_1(y^n) + (1-w_1)S_2(y^n) = \frac{w_1 p_{\theta_{1.1}}(y^n) + (1-w_1)p_{\theta_{1.2}}(y^n)}{0.5^n}, \tag{3.2.7}$$

$$\begin{aligned} \mathbb{E}_{P_0}\left[S_{\text{new}}(Y^n)\right] &= w_1 \mathbb{E}_{P_0}\left[S_1(Y^n)\right] + (1-w_1)\mathbb{E}_{P_0}\left[S_2(Y^n)\right] \\ &\leq w_1 + (1-w_1) \\ &= 1. \end{aligned} \tag{3.2.8}$$

It can be observed that the simple S-value is such a convex combination of S_1 and S_2, with $w_1 = \frac{1}{2}$; we can conclude that the simple S-value is an S-value.

The simple S-value is not GROW for the entire $\Theta_1(\delta)$

To show that the simple S-value cannot be the GROW S-value, we now mix $p_{W_1^*}$ with another marginal distribution $p_{W_1'}$, creating a new marginal distribution to construct the S-value with:

$$p_{W_{1\alpha}} \coloneqq (1-\alpha)p_{W_1^*} + \alpha p_{W_1'} \quad \text{with } 0 \leq \alpha \leq 1.$$

If $p_{W_1^*}$ was the alternative marginal distribution achieving the GROW S-value, the KL divergence $D(P_{W_{1\alpha}} || P_{\theta_0^*})$ would be minimized at $\alpha = 0$, thus, the derivative of the KL divergence with respect to α would be nonnegative. Evaluating this derivative at 0 yields:

$$\frac{\partial}{\partial\alpha}\left[D(P_{W_{1\alpha}}||P_{\theta_0^*})\right]\Big|_{\alpha=0} =$$

$$\frac{\partial}{\partial\alpha}\left[\sum_{y^n} p_{W_{1\alpha}}(y^n)\ \left[\log(p_{W_{1\alpha}}(y^n)) - \log(p_{\theta_0^*}(y^n))\right]\right]\Big|_{\alpha=0} =$$

$$\sum_{y^n}\left[\frac{\partial}{\partial\alpha}\left[p_{W_{1\alpha}}(y^n)\right]\right]\left[\log(p_{W_{1\alpha}}(y^n)) - \log(p_{\theta_0^*}(y^n))\right]$$

$$+ p_{W_{1\alpha}}(y^n)\left[\frac{\partial}{\partial\alpha}\left[\log(p_{W_{1\alpha}}(y^n)) - \log(p_{\theta_0^*}(y^n))\right]\right]\Big|_{\alpha=0} =$$

$$\sum_{y^n}\left[p_{W_1'}(y^n) - p_{W_1^*}(y^n)\right]\left[\log(p_{W_1^*}(y^n)) - \log(p_{\theta_0^*}(y^n))\right] + p_{W_1'}(y^n) - p_{W_1^*}(y^n) =$$

$$\sum_{y^n} p_{W_1'}(y^n)\left[\log(p_{W_1^*}(y^n)) - \log(p_{\theta_0^*}(y^n))\right]$$

$$- \sum_{y^n} p_{W_1^*}(y^n)\left[\log(p_{W_1^*}(y^n)) - \log(p_{\theta_0^*}(y^n))\right] + \sum_{y^n} p_{W_1'}(y^n) - \sum_{y^n} p_{W_1^*}(y^n) =$$

$$E_{P_{W_1'}}\left[\log\left[\frac{p_{W_1^*}(Y^n)}{p_{\theta_0^*}(Y^n)}\right]\right] - D(P_{W_1^*}||P_{\theta_0^*}). \tag{3.2.9}$$

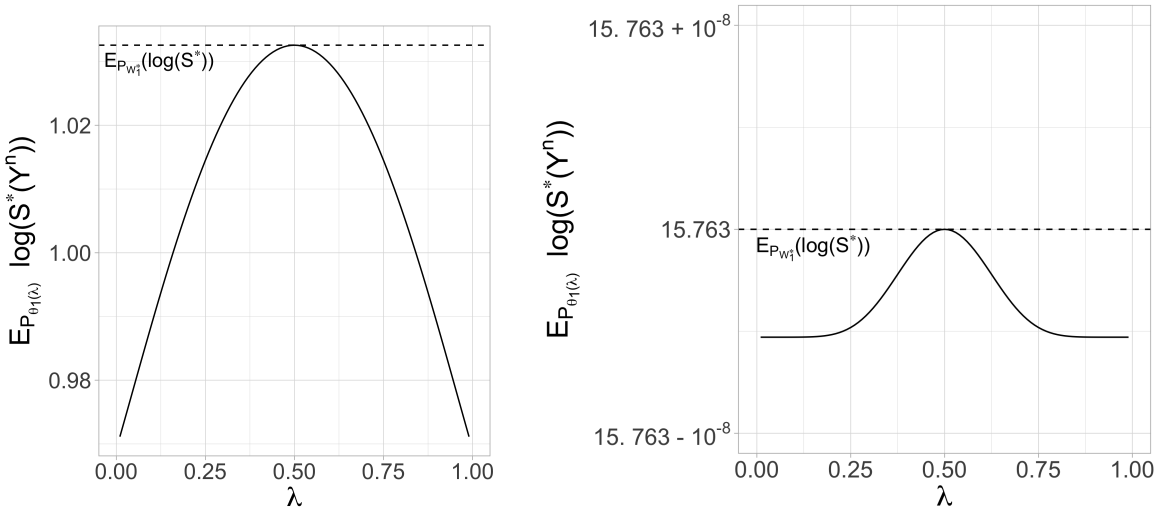
It follows that this derivative can only be 0 or positive when the expected capital growth under $P_{W_1'}$ is bigger than or equal to the expected capital growth under $P_{W_1^*}$; this coincides with the property of GROW S-values described in (1.3.7). However, as we show below, evaluating (3.2.9) for some n_a, n_b and δ and for several choices of $p_{W_1'} = p_{\theta_1}$ with $\theta_1 \in \Theta(\delta)$ indeed reveals negative derivatives at $\alpha = 0$. Hence, we can conclude that increasing α and creating a mixture of $P_{W_1^*}$ and another distribution to create a new marginal can *decrease* $D(P_{W_1}||P_{W_0})$ further, and that the KL-divergence minimum has not been reached with the simple S-value.

This is illustrated in Figure 3.1, where the expected value of $\log(S)$ under various distributions $P_\theta : \theta \in \Theta_1$ was calculated with the simple S-value. In (a), it can be observed that the expected capital growth for all these distributions is lower than or equal to the expected capital growth when data are generated by the marginal distribution $P_{W_1^*}$. In (b), this experiment was extended to $n_a = n_b = 100$; it can be observed that the expected growth in the data-generating distributions is still lower than the expected growth under the marginal, but the differences are almost negligible.

The simple S-value *is* GROW for an 'angled' restricted alternative

The 'simple' S-value is not the GROW S-value for all distributions $P_{\theta_1} : \theta_1 \in \Theta_1(\delta)$, but we will now show that when one is willing to allow for a slight *angle* in the shape of the restricted alternative hypothesis to find the GROW S-value for, $\Theta_1(\delta)$, the simple S-value coincides with the GROW S-value.

The proposed form of an angled restricted alternative hypothesis is illustrated in Figure 3.2. If one defines δ as the minimum relevant substantive difference between θ_a and θ_b one is interested in, a restricted alternative hypothesis bounded by two straight lines through the points $(0, \delta), (1-\delta, 1)$ and $(\delta, 0), (1, 1-\delta)$ is obtained. To create an angled restricted alternative hypothesis, one could instead create four lines based on a $\delta' > \delta$ as in (a). This creates an angle between $\Theta_1(\delta)$ and its



(a) $n = 20$ (b) $n = 200$

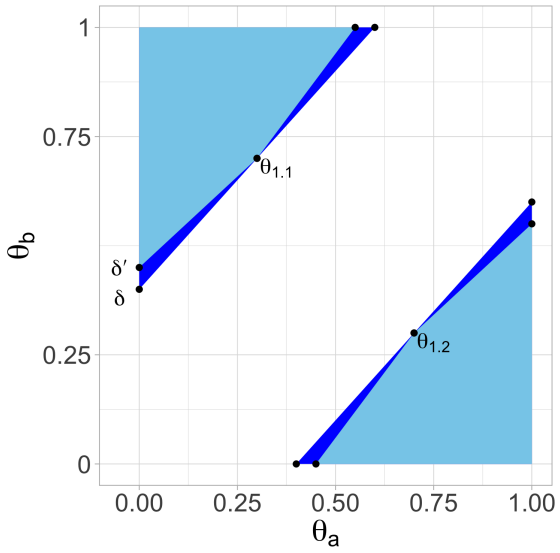
Figure 3.1: Exploring the expected value of $\log(S^*)$ under different distributions from the restricted alternative hypothesis. Various $\theta_1 = (\theta_a$ and $\theta_b)$ were constructed as: $\theta_a = 0\lambda + 0.6(1-\lambda)$, and $\theta_b = 0.4\lambda + 1(1-\lambda)$ for generating data under various P_{θ_1}.

new shape $\Theta_1'(\delta)$. If we take this angle to be large enough, the simple S-value *does* provide the worst case expected capital growth for all P_θ with $\theta \in \Theta_1'(\delta)$ (b), as we verify numerically below.

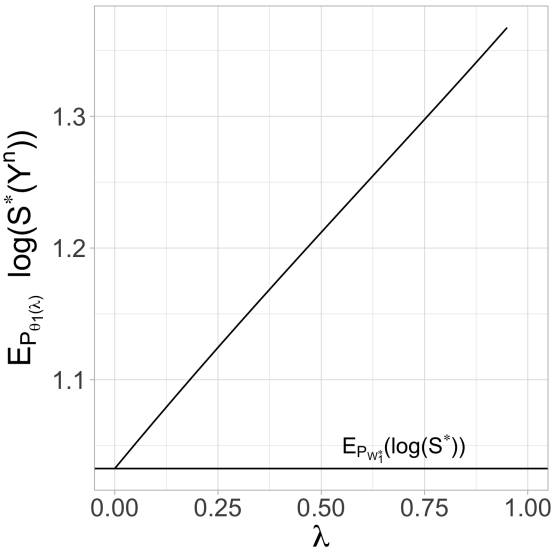
Table 3.1: $\min \delta' - \delta$ for achieving worst case capital growth with three point S-value and an angled $\Theta_1'(\delta)$

δ \ n	10	20	30	40	50	60	70	80	90	100
0.1	0.1589	0.1009	0.0746	0.0589	0.0483	0.0406	0.0348	0.0301	0.0263	0.0232
0.2	0.0968	0.0502	0.0310	0.0206	0.0142	0.0100	0.0072	0.0052	0.0038	0.0028
0.3	0.0552	0.0215	0.0100	0.0049	0.0025	0.0013	0.0007	0.0004	0.0003	0.0002
0.4	0.0284	0.0073	0.0022	0.0007	0.0003	0.0001	0.0001	0.0001	0.0001	0.0001
0.5	0.0126	0.0018	0.0003	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
0.6	0.0044	0.0003	0.0001	0.0001	0.0001	0.0001	0.0001	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
0.7	0.0011	0.0001	0.0001	0.0001	0.0001	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
0.8	0.0002	0.0001	0.0001	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
0.9	0.0001	0.0001	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$

The minimal δ' resulting in a large enough angle for the simple S-value to satisfy the criterion in (1.3.7) (and to be GROW) was determined for several combinations of δ and n. One can observe in Table 3.1 that the angle needed to add to $\Theta_1(\delta)$ for the simple S-value to be GROW vanished as $n \to \infty$: *the subset of* $\Theta_1(\delta)$ *for which the simple S-value is suboptimal vanishes as* n *increases.* The angle is also very small for small values of n. Moreover there appeared to be no difference in obtained power between the simple S-value and the BFGS GROW S-value, as illustrated in Figure 3.3. This is why we propose that, in the case of a restricted alternative hypothesis based on a minimal substantive difference between the proportions, the simple S-value can be used without losing a relevant proportion of expected capital growth.



(a) $\Theta_1'(\delta)$, $\delta' = 0.45$. Support points from the simple S-value are indicated as $\theta_{1.1}$ and $\theta_{1.2}$.



(b) Expected capital growth under distributions from the *angled* restricted alternative hypothesis

Figure 3.2: The angled $\Theta_1'(\delta)$ and expected capital growth with the simple S-value under this angled alternative for $n = 20$, $\delta = 0.4$ and $\delta' = 0.45$. In (b), the various $\theta_1 = (\theta_a$ and $\theta_b)$ were constructed as mixtures of $\theta_{1.1}$ and $(0, \delta')$: $\theta_a = 0\lambda + 0.3(1-\lambda)$, and $\theta_b = 0.45\lambda + 0.7(1-\lambda)$ for generating data under various P_{θ_1} from the *angled* restricted alternative hypothesis.

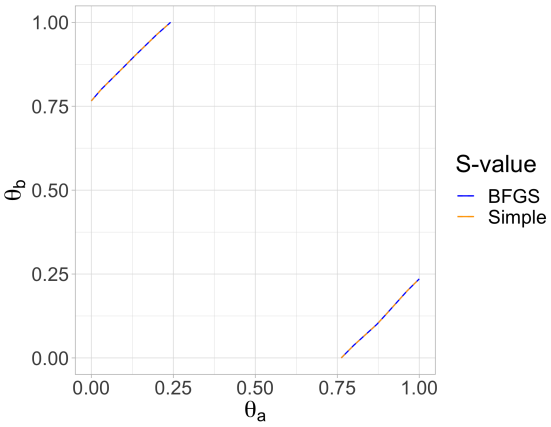


Figure 3.3: Parameters from distributions for which power $1 - \beta = 0.8$ is obtained, with either the simple S-value or the S-value obtained through BFGS. A Monte-Carlo study with 1000 repetitions for each combination of θ_a and θ_b was conducted. $n = 20, n_a = n_b$.

3.3 The simple S-value when $n_a \neq n_b$

We have seen how a simple S-value can be designed for the 2x2 contingency table test in the case of *equal group sizes*. In this section, we will explored whether an analogue of this simple S-value can be designed for the case where one aims to perform the 2x2 test for unequal group sizes.

Numerical experiments performed for finding the JIPr for the two-sided test in cases where

$n_a \neq n_b$ revealed a different pattern in the distribution of the support points in $\Theta_1(\delta)$ than for the case were $n_a = n_b$; examples are depicted in Figure 3.4. For the one sided-test, support points were distributed differently in both $\Theta_1(\delta)$ *and* Θ_0 (experiments not shown). This led to the conjecture that an extension of an analogue of the one-sided simple S-value to the two-sided case as was performed above for the equal group sizes could not be performed for the unequal group sizes.

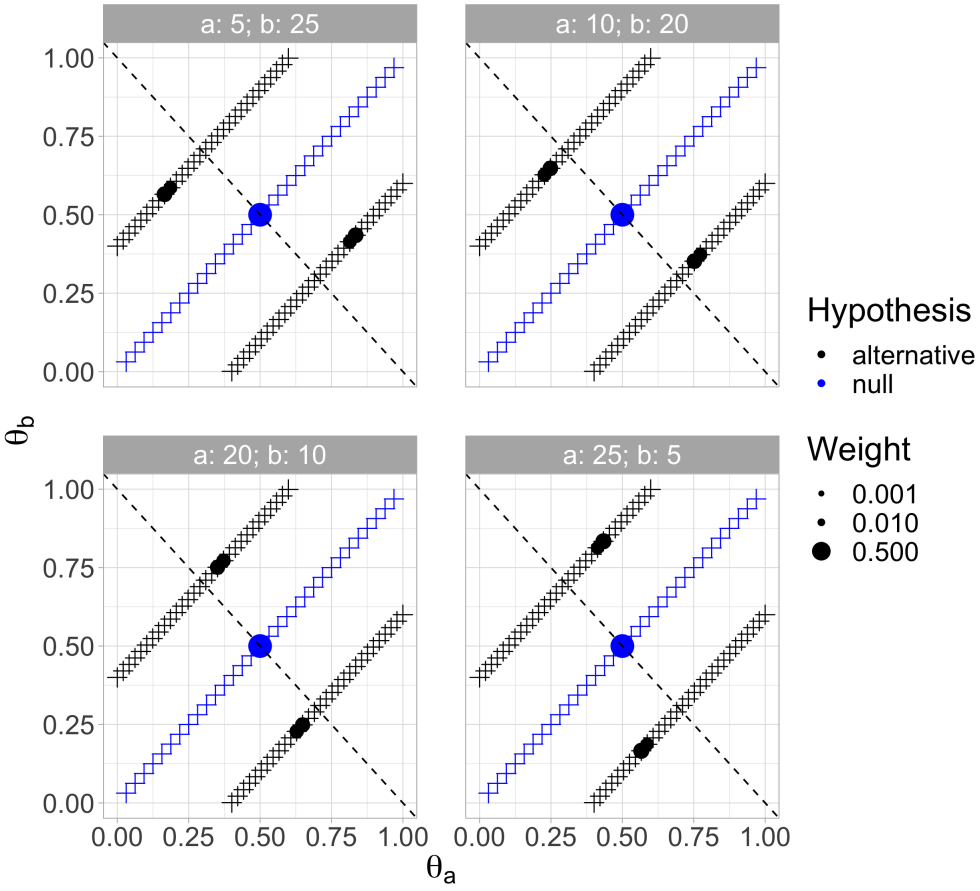


Figure 3.4: JIPr found through BFGS for various cases where $n_a \neq n_b$, with $\delta = 0.4$ and $n = 30$. Crosses indicate grid points included in $\Theta_1(\delta)$ and Θ_0; dots indicate found support points and the size of the dots indicates the weight assigned to the support points.

In Figure 3.4 it can be observed that in the JIPr for the two-sided test for unequal group sizes, the mass in the null hypothesis appeared to be put almost entirely on the point 0.5, as with the simple S-value for $n_a = n_b$. The mass in the restricted alternative hypothesis was still put almost on two points, but on different points than for the case where $n_a = n_b$. In Section 3.2 we defined the support points $\theta_{1.1}$ and $\theta_{1.2}$ for the simple S-value for equal group sizes. We now propose that $\theta_{1.1}$ and $\theta_{1.2}$ can be adjusted for each n_a and n_b combination with some value ϕ to achieve generalization of the simple S-value:

$$\theta_{1.1,\phi} = \theta_{1.1} + \phi \quad ; \quad \theta_{1.2,\phi} = \theta_{1.2} - \phi \; ; \quad \phi \in [-(0.5 - 0.5\delta), 0.5 - 0.5\delta], \tag{3.3.1}$$

which can be used for constructing the *generalized simple S-value*:

$$S^*_\phi(y^n) = \frac{p_{W^*_{1,\phi}}}{p_{W^*_0}} = \frac{\frac{1}{2}\left(p_{\theta_{1.1},\phi}(y^n) + p_{\theta_{1.2},\phi}(y^n)\right)}{0.5^n}, \tag{3.3.2}$$

with $\phi = 0$ when $n_a = n_b$.

This definition is motivated by the results from the numerical experiment depicted in Figure 3.4: for $n_a \neq n_b$, the support points appear to be translated along the lines $\theta_b = \theta_a + \delta$ and $\theta_b = \theta_a - \delta$; $\theta_{1.1}$ appears to be shifted downward for $n_a < n_b$, and $\theta_{1.2}$ upward. The opposite pattern was observed for $n_a > n_b$. The size of the shifts appears to be different for each combination of n_a and n_b for fixed n (see Figure 3.4).

For each combination of n, n_a, n_b and δ, the optimal value of ϕ can be obtained by minimizing:

$$\mathbb{E}_{P_{W^*_{1,\phi}}}[\log S^*_\phi(Y^n)] = \sum_{y^n} p_{W^*_{1,\phi}}(y^n) \log \left[\frac{p_{W^*_{1,\phi}}(y^n)}{0.5^n}\right]. \tag{3.3.3}$$

(3.3.3) was optimized with respect to ϕ for various n, n_a, n_b and δ (results not shown). As expected the estimated value of ϕ was equal to 0 for $n_a - n_b = 0$. For more extreme differences between n_a and n_b, larger adjustments were needed for minimizing expected capital growth. This relationship appeared to be linear, except for larger values of δ. It was numerically shown that the candidate generalized simple S-value is indeed an S-value, as illustrated in Appendix Figure A.2. As with the simple safe test for $n_a = n_b$, the adjusted simple S-value for $n_a \neq n_b$ indeed achieved worst-case expected capital growth for (slightly angled) $\Theta'(\delta)$, as illustrated in Appendix Table A.1. Hence, again, we propose that the generalized simple S-value can be used when $n_a \neq n_b$, as the expected growth is only suboptimal for a very small subset of the parameters, as with the simple S-value.

Overview of proofs seen so far
So far, simple S-values for *equal* group sizes n_a and n_b have been presented, both for the one-sided and the two-sided case. For these cases, formal proofs that the simple S-value *is* an S-value have been given. For the one-sided case, a numerical proof was presented that the simple S-value is GROW. For the two-sided case, a numerical proof that the simple S-value is GROW for an *angled* subset of the restricted alternative hypothesis was given. For the case where $n_a \neq n_b$, the generalized simple S-value was defined; we merely have a numerical proof that this is an S-value. For this S-value, a numerical proof that it is GROW for an angled subset of the restricted alternative hypothesis was presented as well.

3.4 Choices for the restricted alternative hypothesis

It has now been determined how the simple S-value can be defined for a given divergence measure δ, which determines the restricted alternative hypothesis $\Theta_1(\delta)$. However, it is not yet clear *how* δ *should be picked when designing a safe test.* In this subsection, four scenarios where one wants to achieve a certain expected capital growth or *power* with the simple S-value are described, and choices for δ and the sample size n in these cases are discussed. These four scenarios were implemented in the `safe2x2` package as well.

We will first consider the case where one does not know what significance level α will be used before designing an experiment. In this situation, we do presume that one expects to collect a

certain substantial amount of evidence against $\mathcal{H}_0$: we assume that the desired expected capital growth $L = \mathbb{E}_{P_{\theta_1}}[\log(S(Y^n))]$ is known in advance, even tough n or θ_1 may not be known. The design of the S-value then depends on two parameters: δ, determining the set of parameters in the restricted alternative hypothesis, and n, the sample size.

Scenario 1

In the first scenario, there exists a fixed sample size n, for example when a researcher is limited by available resources. The expected capital growth L then depends on $\Theta_1(\delta)$ used for designing the simple S-value. In Figure 3.5a this relation is illustrated; it can be observed that expected capital growth increases with δ. This logically follows from the properties of Kullback-Leibler divergence being a measure of dissimilarity between distributions; if $\Theta_1(\delta)$ is very dissimilar to Θ_0, the minimal KL-divergence between P_{W_1} and P_{W_0} will increase. The option to find the minimal δ^* such that a certain minimal capital growth under the restricted alternative hypothesis is achieved for a fixed sample size was added to the `safe2x2` package, where the following algorithm was implemented:

---

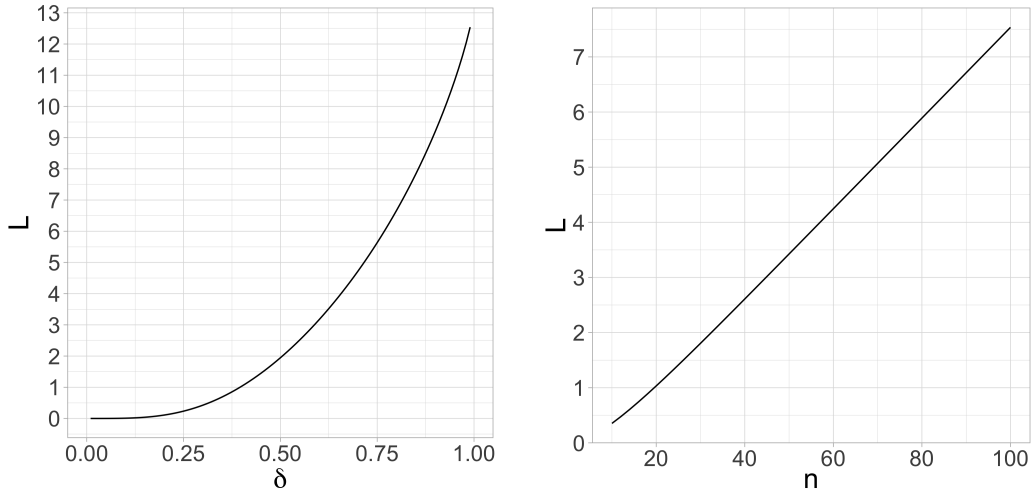
for $\delta \in \{0.01, \ldots, 0.99\}$ **do**
  Construct $S^* = \frac{p_{W_1^*}(y^n)}{p_{W_0^*}(y^n)}$ with $\bar{\Theta}_1(\delta)$ (directly by using the simple S-value or through finding the JIPr)
  if $\mathbb{E}_{P_{W_1^*}}[\log(S^*(Y^n))] \geq L$ **then**
    return δ, S^*
  end if
end for
return message: desired capital growth not reached

Scenario 2

In the second scenario the researcher knows the subset of the alternative hypothesis $\Theta_1(\delta)$ a certain expected capital growth L should be achieved for, and wants to determine the minimal sample size n to reach this expected capital growth. This scenario occurs when one is very certain about the substantively relevant subset $\bar{\Theta}_1(\delta) \subset \Theta_1$, for example when introducing a new drug to the market is only considered useful when a certain proportion of patients experiences alleviation of symptoms when using it.

Expected capital growth increases linearly with the sample size for a fixed δ, as is illustrated in Figure 3.5b for the case where $n_a = n_b$. The option to find the smallest sample size, such that a certain capital growth is reached was added to the `safe2x2` package as well. The function is illustrated in the algorithm below. Note that some $n^{\max}$ has to be specified to prevent the algorithm from running undesirably long, and that n will be increased with some a_{iter} and b_{iter} during each iteration, depending on the ratio between group size a and b one aims to achieve.

$n = 0$
while $n < n^{max}$ **do**
  $n = n + a_{\text{iter}} + b_{\text{iter}}$
  Construct $S^* = \frac{p_{W_1^*}(y^n)}{p_{W_0^*}(y^n)}$ (directly by using the simple S-value or through finding the JIPr)
  if $\mathbb{E}_{P_{W_1^*}}[\log(S^*(Y^n))] \geq L$ **then**
    return n, S^*
  end if
end while
return message: desired capital growth not reached below maximum n



(a) L and δ at $n = 20$ (b) L and n at $\delta = 0.4$

Figure 3.5: Relation between expected capital growth L and δ (a) or sample size n (b) for the simple S-value for $n_a = n_b$.

Scenario 3: finding the simple S-value that optimizes power for a fixed sample size n

As in the first scenario, in the third scenario we will consider again the case where one has collected a sample with a certain size n, and now wants to determine the δ to construct the S-value with.This δ is chosen to maximize the power obtained under all distributions in the researcher's restricted alternative hypothesis of interest. *Note that we have two separate δ here: we will use δ^* as the distance measure to construct the S-value $S^*_{\delta^*}$ with, and we will use δ° to define the restricted alternative hypothesis one is interested in.* Thus, we aim to find δ^* such that in the worst case, for data generated under some $P_\theta : \theta \in \Theta_1(\delta^\circ)$ the best power is achieved. I.e. for some fixed δ°, one searches for the δ^* that achieves (adapted from GHK):

$$1 - \beta^* := \max_{\delta^*} \min_{\theta \in \Theta_1(\delta^\circ)} P_\theta \left(S^*_{\delta^*} \geq \frac{1}{\alpha} \right), \tag{3.4.1}$$

is searched for. Since the minimum in the equation above is always achieved at the boundary of $\Theta_1(\delta^\circ)$, the distributions with the smallest difference between θ_a and θ_b, we only need to investigate the distributions that lie on this boundary, defined as $\bar{\Theta}_1(\delta^\circ)$ (see Section 2.1). Hence, it would be equivalent to replace $\Theta_1(\delta^\circ)$ by $\bar{\Theta}_1(\delta^\circ)$ in the equation above.

As an illustration, results from experiments for $n = 20$ and $n = 100$ are presented in Tables 3.2 and 3.3. Per θ in $\bar{\Theta}_1(\delta^\circ)$, 1000 Monte Carlo simulations were performed to determine the probability of rejecting the null hypothesis when data were generated under P_θ with S-values designed with a grid of δ^* values.The δ^* achieving the best power for each δ° are depicted in Tables 7 and 8. As noted by GHK, optimal choices of δ^* for each n appear to exist: all found δ^* are similar for each n. In fact, there appears to be a *region of uniformly most powerful* δ^* for each n. This suggests that an equivalent of the 'uniformly most powerful test' for classical hypothesis tests exists for safe testing; this will be elaborated on in the discussion, Section 8.2. For example for $n = 20$, δ^* values $0.50, 0.55$ and 0.60 all result in the same (optimal) worst-case power. The value $\delta^* = 0.45$ appears to lie outside the 'uniformly most powerful region' and yields lower worst-case power. For $n = 100$, the 'most powerful region' appears to have lower values and ranges from 0.21 to 0.30.

Table 3.2: Finding the δ^* that provides the highest power in the worst case for $\Theta(\delta^\circ)$ at $\alpha = 0.05$ for $n = 20$

δ°	δ^*	$1-\beta^*$	$\min(1-\beta)$ for $\delta^* = 0.55$	$\min(1-\beta)$ for $\delta^* = 0.60$	$\min(1-\beta)$ for $\delta^* = 0.45$
0.30	0.50	0.04	0.04	0.04	0.01
0.40	0.50	0.15	0.15	0.15	0.04
0.50	0.50	0.37	0.37	0.37	0.16
0.60	0.50	0.60	0.60	0.60	0.37
0.70	0.50	0.81	0.81	0.81	0.62
0.80	0.50	0.94	0.94	0.94	0.85
0.90	0.50	0.99	0.99	0.99	0.98

Table 3.3: Finding the δ^* that provides the highest power in the worst case for $\Theta(\delta^\circ)$ at $\alpha = 0.05$ for $n = 100$

δ°	δ^*	$1-\beta^*$	$\min(1-\beta)$ for $\delta^* = 0.25$	$\min(1-\beta)$ for $\delta^* = 0.30$	$\min(1-\beta)$ for $\delta^* = 0.20$
0.20	0.21	0.10	0.10	0.10	0.06
0.30	0.21	0.61	0.61	0.61	0.52
0.40	0.21	0.91	0.91	0.91	0.87
0.50	0.21	0.99	0.99	0.99	0.99
0.60	0.21	1.00	1.00	1.00	1.00

All in all, these results suggest that for scenario 3, for each sample size n, a standard δ^* can be provided as the optimal δ to construct the simple S-value with: a table with these values will be included in the `safe2x2` package for a few standard cases ($n_a = n_b$, $\alpha = 0.05$). A function to determine δ^* for some n and α will also be included for the other cases, where the following algorithm will be implemented:

---

for $\delta^\circ \in \{0.1, \ldots, 0.9\}$ **do**
  for $\theta_1 \in \bar{\Theta}_1(\delta^\circ)$ **do**
    Compute n_{a1} and n_{b1} M times from n_a Bernoulli(θ_a) and n_b Bernoulli(θ_b) trials
    for $\delta \in \{0.1, \ldots, 0.9\}$ **do**
      Construct $S^*_\delta = \frac{p_{W_1^*}(y^n)}{p_{W_0^*}(y^n)}$
      Estimate $1 - \beta_{\delta,\theta} = \frac{1}{M}\sum_{i=1}^{M} S^*(n_{a1}, n_{b1}) \geq \frac{1}{\alpha}$
    end for
  end for
  δ_i^* is the δ that achieves:

$$\max_{\delta} \min_{\theta \in \bar{\Theta}_1(\delta^\circ)} 1 - \beta_{\delta,\theta}$$

end for
return mode(δ^*)

Scenario 4: finding the minimal sample size to achieve a certain power for a certain difference of interest δ°

In the fourth and last scenario considered, we want to determine the minimal sample size (and its corresponding δ^*) needed to achieve some minimal power while testing with a certain significance level, and we are interested again in data generated under $\bar{\Theta}_1(\delta^\circ)$. Thus, it is aimed to find the smallest sample size $n_{\min}$ with a δ^* such that:

$$\min_{\theta \in \bar{\Theta}_1(\delta^\circ)} P_\theta \left(S_{\delta^*}(y^{n_{\min}}) \geq \alpha \right) \geq 1 - \beta. \tag{3.4.2}$$

To find the combination of $n_{\min}$ and δ^* for each δ°, samples were drawn for each $\theta \in \bar{\Theta}_1(\delta^\circ)$ to estimate the power of each combination of n and S_δ, and n was increased until the desired power of 0.80 was reached. This is a computational intensive algorithm, as illustrated in the pseudocode below:

$n = 0$
while $n < n^{max}$ **do**
  $n = n + a_{\text{iter}} + b_{\text{iter}}$
  for $\delta^* \in \{0.01, \ldots, 0.99\}$ **do**
    Construct $S^* = \frac{p_{W_1^*}(y^n)}{p_{W_0^*}(y^n)}$ with $\bar{\Theta}_1 \delta^*$
    for $\theta_1 \in \bar{\Theta}_1(\delta^\circ)$ **do**
      Compute n_{a1} and n_{b1} M times from n_a Bernoulli(θ_a) and n_b Bernoulli(θ_b) trials
      Estimate $1 - \beta_\theta = \frac{1}{M}\sum_{i=1}^{M} S^*(n_{a1}, n_{b1}) \geq \frac{1}{\alpha}$
    end for
    if $\min_\theta[1 - \beta_\theta] \geq$ desired power **then**
      return δ^*, n, S^*
    end if
  end for
end while
return message: desired power not reached below maximum n

Results from an experiment for a grid of δ° with $\alpha = 0.05$ and $n_a = n_b$ are illustrated in Figure 3.6. It can be observed that $n_{\min}$ is decreasing in δ°: to detect a smaller difference between the groups, a larger sample size is needed. For values of $\delta^\circ < 0.3$, $n_{\min}$ exceeded 200. As also observed in the two experiments in scenario 3, the δ^* yielding optimal power appears to be decreasing in $n_{\min}$: with a smaller sample size, larger values of δ^* appear to yield optimal power.

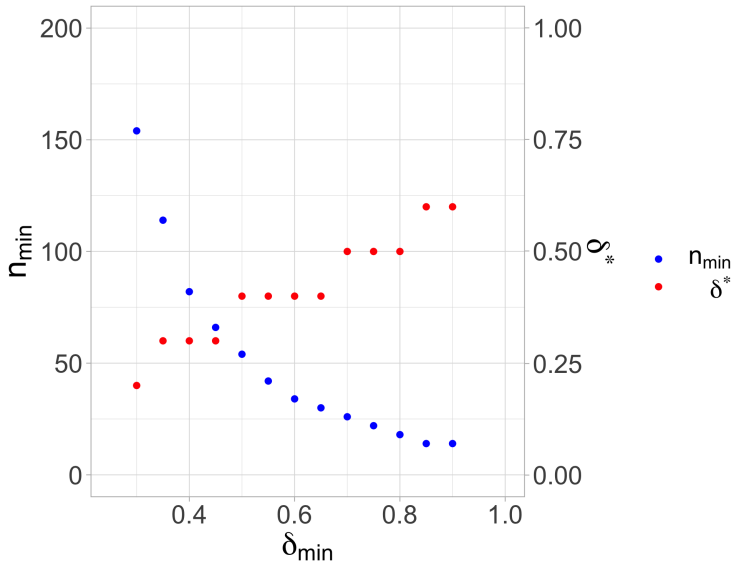


Figure 3.6: minimal n achieving $\min 1 - \beta \geq 0.80$ at δ^* for data generated under $\Theta_1(\delta_{\min})$

Conclusion

In this section, GROW S-values designed with the *default* divergence measure were investigated. From the derivations and numerical illustrations presented, it was concluded that for the default case the *simple* S-value can be used, which is defined for each δ in the case of equal group sizes. Further, the simple S-value can be adjusted with parameter ϕ found through optimization in the case of unequal group sizes. These results imply that the numerical BFGS approach for finding the JIPr in the default case can be omitted. Lastly, it was investigated how the *value* of the divergence measure δ can be chosen. In four scenarios, it was investigated how δ and sample size n relate to expected capital growth and power of the safe test.

Chapter 4

The KL-divergence based S-value

As discussed in Chapter 2, a second divergence measure δ to define $\Theta_1(\delta)$ for finding the JIPr and designing the S-value is proposed to be the minimal Kullback-Leibler divergence between singleton distributions in $\{P_\theta : \theta \in \Theta_1\}$ and P_{W_0}. In the next section it will be discussed how $\Theta_1(\delta)$ can be found, how the weight in the JIPr is distributed for various sample sizes and how this KL-based S-value behaves in terms of *power* compared to the simple S-value, Fisher's exact test and the Gunel-Dickey Bayes factor.

4.1 Reverse information projection for determining the KL-based restricted alternative hypothesis

In Chapter 2, the KL-based restricted alternative hypothesis was defined as:

$$\Theta_1(\delta) = \left\{ \theta_1 : \min_{W \in W(\Theta_0)} D(P_{\theta_1} || P_W) \geq \delta \right\}. \tag{4.1.1}$$

As was seen throughout Chapters 2 and 3, RIPr support points differ for different combinations of n_a and n_b, and achieved expected capital growth $D(P_{\theta_1}||P_{W_0*})$ increases with n. Hence, the KL-based $\Theta_1(\delta)$ is different for each specific combination of n_a and n_b. To determine $\Theta_1(\delta)$ for some n_a and n_b one would need to find the distribution $P_{W_0^*}$ minimizing the expected capital growth for each $P_{\theta_1} : \theta_1 \in \Theta_1$ through reverse information projection, and record the achieved infimum of the expected capital growth. This is computationally intensive.

A first simplification is that, in the case of $n_a = n_b$, for any θ_1, the RIPr is achieved for a prior W_0^* that puts all mass on a single θ_0^*, as we showed in Section 2.3:

$$\Theta_1(\delta) = \left\{ \theta_1 : \min_{\theta_0' \in [0,1])} D(P_{\theta_1} || P_{\theta_0'}) \geq \delta \right\}. \tag{4.1.2}$$

To further decrease the required computational steps for determining the KL-based alternative, we can use the property that KL-divergence behaves *linear* in sample size. This is because data are independently and identically distributed under all $P \in \mathcal{H}_0 \cup \mathcal{H}_1$. For now, for simplicity, we will assume that $n_a = n_b = \frac{n}{2}$. Then, we have for $n' = mn$:

$$D(P_{\theta_1}(Y^{n'})||P_{\theta_0}(Y^{n'})) = mD(P_{\theta_1}(Y^n)||P_{\theta_0}(Y^n)) \tag{4.1.3}$$

We thus can thus define the standardized KL-divergence:

$$D_{\text{stn}}(P_{\theta_1}||P_{\theta_0}) := \frac{D(P_{\theta_1}(Y^n)||P_{\theta_0}(Y^n))}{n}, \tag{4.1.4}$$

and can use this standardized definition to *define* $\Theta_1(\delta)$ *for every* $n \in \{2, 4, 6, \dots\}$ *if we have determined the RIPrs and resulting expected capital growth for* Θ_1 *for one particular* n', since can now rewrite $\Theta_1(\delta)$ as:

$$\Theta_1(\delta) = \left\{\theta_1 : \min_{\theta_0' \in [0,1]} D(P_{\theta_1}(Y^n)||P_{\theta_0'}(Y^n)) \geq \delta\right\} = \left\{\theta_1 : \min_{\theta_0' \in [0,1]} D_{\text{stn}}(P_{\theta_1}||P_{\theta_0'}) \geq \frac{\delta}{n}\right\} \tag{4.1.5}$$

If we would now want to determine $\Theta_1(\delta)$ for n, but we only know the RIPr KL-divergence for n', we could determine $D_{\text{stn}}(P_{\theta_1}||P_{\theta_0})$ for every P_{θ_1} with (4.1.4), and use this to determine $D(P_{\theta_1}(Y^n)||P_{\theta_0}(Y^n))$, as we have that:

$$D(P_{\theta_1}(Y^n)||P_{\theta_0}(Y^n)) = nD_{\text{stn}}(P_{\theta_1}||P_{\theta_0})) = n\frac{D(P_{\theta_1}(Y^{n'})||P_{\theta_0}(Y^{n'}))}{n'} \tag{4.1.6}$$

Note that this can be generalized for any constant ratio $\frac{n_a}{n_b}$ in n and n'.

As an example, $D(P_{\theta_1}(Y^n)||P_{\theta_0}(Y^n))$ was determined for each $\theta_1 \in \Theta_1$ (with Θ_1 represented by an equally-spaced grid) and the corresponding RIPr for $n = 2$, $n_a = n_b = 1$. The resulting grid of minimized KL-divergence for each θ_1 was used to construct $\bar{\Theta}(\delta)$ for various δ for $n = 20$ and $n = 100$: these sets are depicted in Figure 4.1. To construct for example $\bar{\Theta}(\delta)$ with $\delta = \log(20)$ for $n = 20$, all points in Θ_1 achieving $D(P_{\theta_1}(Y^{n=2})||P_{\theta_0}(Y^{n=2})) = 2\frac{\log(20)}{20}$ were selected.

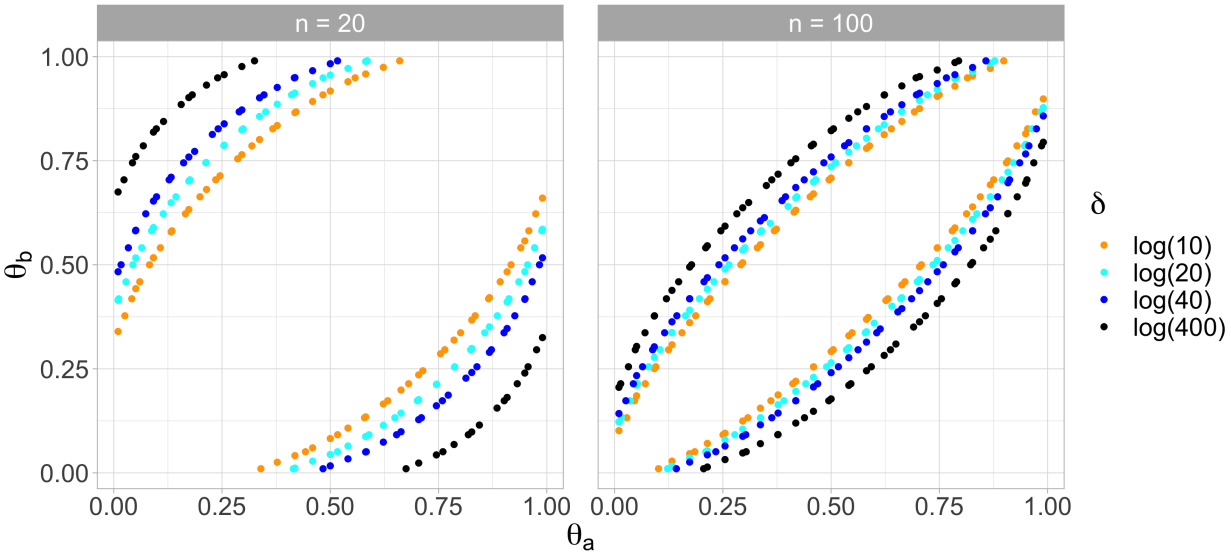
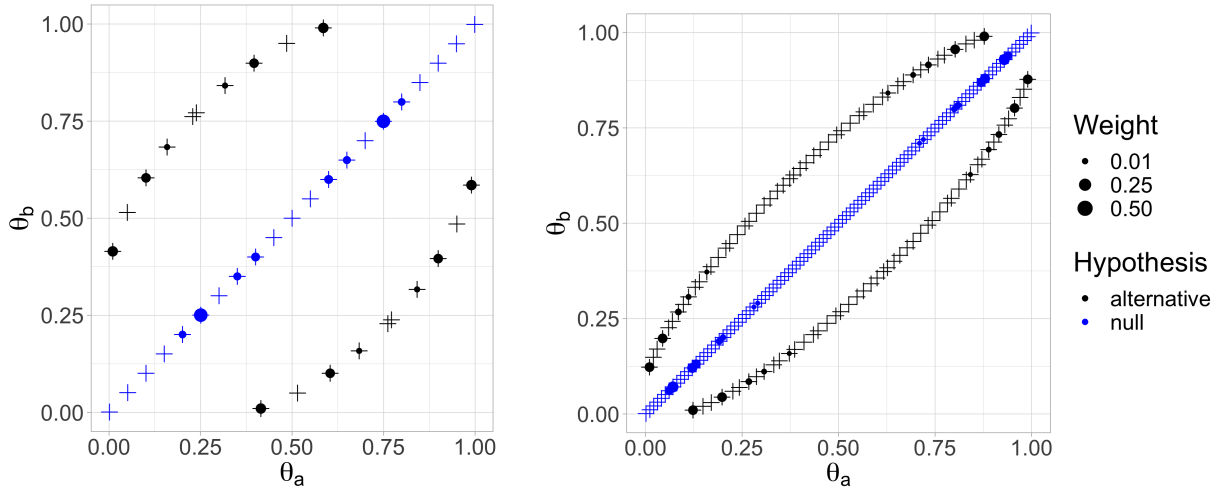


Figure 4.1: Grids representing the KL-based $\Theta_1(\delta)$ for various values of δ

4.2 JIPr for the KL-based S-value

Optimization experiments were carried out to investigate the support points in $\Theta_1(\delta)$ and Θ_0 in the JIPr. Patterns obtained after finding the JIPr with the KL-based $\Theta_1(\delta)$, $\delta = \log(20)$ are illustrated in Figure 4.2 for $n = 20$ and $n = 100$. Unlike with the default S-value, there appear to be many support points for both n and it is not evident how an equivalent of the 'simple S-value' could be constructed in this case. Support points in $P_{W_1^*}$ and $P_{W_0^*}$ appear to be evenly spread out for $n = 20$; for $n = 100$ they appear to exist mainly towards the edges of the parameter space.



(a) $n = 20$, $\delta = \log(20)$ (b) $n = 100$, $\delta = \log(20)$

Figure 4.2: JIPr for KL-based alternative hypotheses with $\delta = \log(20)$. Crosses indicate positions of grid points in Θ_1 and Θ_0, dots and their size represent the weight from W_0^* and W_1^* for each grid point.

Next, it was investigated if a minimum grid precision for $\Theta_1(\delta)$ and Θ_0 could be determined: exceeding the minimum grid precision is unwanted, as this increases computation time. We assume that as long as increasing the number of grid points in $\Theta_1(\delta)$ and Θ_0 decreases the found minimized KL-divergence $D(P_{W_1^*}||P_{W_0^*})$, the JIPr between P_{W_1} and P_{W_0} has not been found. Results from two numerical experiments for $n = 20$ and $n = 100$ are summarized in Table 4.1a and b. The first experiments were conducted with equally spaced grids, as illustrated in Figure 4.2. Increasing grid precision beyond 40 did not appear to decrease the found minimum further for $n = 20$. For $n = 100$, results appeared to be inconsistent; the found minimum *increased* for increasing grid precision from 50 tot 75, and from 100 to 125.

Because of the observation that for large n, most support points for the obtained prior in the JIPr appeared towards the edges of the parameter space, it was hypothesized that small changes in the placing of the grid points in those regions might therefore influence the achieved KL divergence minimum. Spreading grid precision uniformly might therefore not lead to the best results for large n. Increasing grid precision unequally, i.e. finer towards the ends of the parameter space, and coarser in the middle, might save computation time required for reaching the same minimal KL divergence. Obtained KL divergence minima for this type of grid for various precisions are illustrated in Table 4.1. It can be observed that the results now have improved for $n = 100$, but not for $n = 20$.

Hence, it was concluded that it is not clear what the optimal minimum grid precision and location of the grid points are for finding the JIPr in the KL-based case for each n. As computation times increase quickly with growing grid precision, as can be observed in Table 4.1, it is infeasible to include too many grid points in the JIPr algorithm. To this end, while designing the `safe2x2` package, the user is given the opportunity to choose her or his own grid precision for performing the JIPr, and to pass her or his own grid points for the restricted alternative hypothesis. Lastly, after the presumed JIPr has been obtained, the user can choose to perform a check to see if the found S-value indeed provides the worst-case growth under the restricted alternative hypothesis

Table 4.1: Time to find the JIPr and the resulting minimized KL-divergence between $P_{W_1^*}$ and $P_{W_0^*}$ for the KL-divergence based S-value for various grid precisions on $\Theta_1(\delta)$ and Θ_0 with $\delta = \log(20)$.

(a) $n = 20$

	Equally spaced grid		Unequally spaced grid	
precision	time (s)	$D(P_{W_1^*}\|\|P_{W_0^*})$	time (s)	$D(P_{W_1^*}\|\|P_{W_0^*})$
10	1.9	2.26897	1.1	2.26897
20	2.1	2.24152	1.0	2.24216
30	3.5	2.24151	1.9	2.24109
40	5.4	2.24108	3.3	2.24134
50	7.4	2.24108	7.0	2.24136

(b) $n = 100$

	Equally spaced grid		Unequally spaced grid	
precision	time (s)	$D(P_{W_1^*}\|\|P_{W_0^*})$	time (s)	$D(P_{W_1^*}\|\|P_{W_0^*})$
25	3.1	2.29104	2.3	2.28031
50	12.1	2.26570	8.9	2.26998
75	36.3	2.26683	33.9	2.26544
100	65.2	2.26547	154.5	2.26440
125	242.1	2.26821	248.8	2.26114

used for designing the S-value, as in (1.3.7).

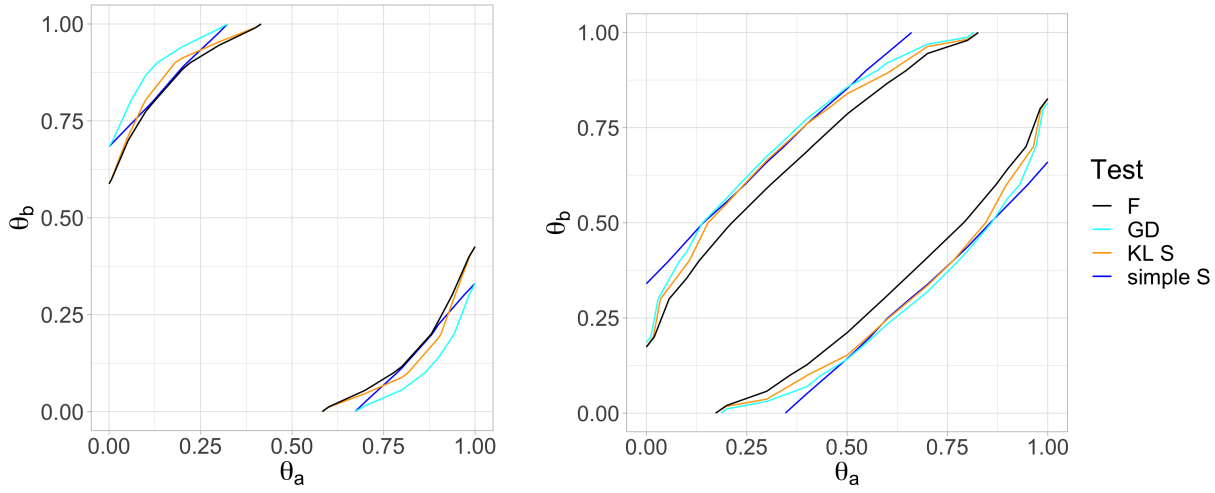
4.3 Comparing the power of the two proposed S-values

Two different S-values have so far been proposed: in the previous chapter, we constructed the 'simple S-value', that (almost) provides optimal growth in the worst case for a restricted alternative hypothesis composed of all distributions with a minimum absolute difference between θ_a and θ_b. In this chapter, the KL-based S-value was introduced. We will now investigate how these S-values behave in terms of *power*.

As explored in Chapter 3, for the simple S-value, there appears to exist a (range of) δ^* yielding the most powerful S-value for each n, independent of the restriction $\Theta_1(\delta^\circ)$ that is used for the alternative. For $n = 20$, one of these δ^* was 0.5, and for $n = 100$, δ^* was found to equal 0.21. In a similar experiment, the δ^* yielding the most powerful S-value in the KL-based case was determined for these sample sizes: the optimal δ^* was $\log(10)$ for $n = 20$, and $\log(16)$ for $n = 100$ (see Appendix Tables A.2 and A.3), still independent of δ°. The grid precision was set to $n + 1$ for both cases.

The performance of these 'most powerful' S-values for $n = 20$ and $n = 100$ were compared to Fisher's exact test and the Gunel-Dickey Bayes factor with respect to power. Data were simulated for a grid of distributions with parameters in Θ_1 and the power for each test for each distribution was estimated. Results are summarized in Figure 4.3; the distributions with the smallest difference between θ_a and θ_b that yielded a power of at least 0.80 are indicated by the lines for each test method. It can be observed that for both sample sizes, the simple S-value and the KL-based S-value performed similarly for most data-generating distributions, except for distributions where θ_a or θ_b had values close to 0 or 1; there, the KL-based S-value provided higher power. The S-values performed similarly or even slightly better (for $n = 20$) compared to the Gunel-Dickey Bayes factor.

For $n = 20$, both S-values appeared to yield a slightly higher power than Fisher's exact test for parameters in the middle of the parameter space. For $n = 100$, Fisher's exact test yielded a higher power overall.



(a) $n = 20$ (b) $n = 100$

Figure 4.3: Power obtained with different tests for 2x2 contingency tables: the lines indicate distributions where a power of 0.80 was achieved. The simple S-value was designed with $\delta^* = 0.5$ for $n = 20$ and $\delta^* = 0.21$ for $n = 100$. The KL-based S-value was designed with $\delta^* = \log(10)$ for $n = 20$ and $\delta^* = \log(16)$ for $n = 100$.

Conclusion

In this section, it was described how the restricted alternative hypothesis $\Theta_1(\delta)$ can be constructed for the KL-based divergence measure through reverse information projection. Examples of the resulting KL-based GROW-S-value found through optimization were presented: a 'simple' equivalent does not appear to exist here. However, experiments aimed at investigating the power of the default and KL-based S-values revealed that the KL-based S-value might perform better in cases where data are generated by distributions towards the edges of the parameter space. This will be illustrated further in Chapter 6. Further, for large n, it appeared that Fisher's exact test yielded a higher power overall compared to both S-values. This implies that to achieve a similar power with the safe test, a larger sample should be collected than when testing with Fisher's exact test. This will be elaborated on in Chapter 5.

Chapter 5

Optional stopping with the simple S-value

In the previous chapter we have observed that with Fisher's exact test, less samples are needed to reach a certain power than with the (simple) S-value for 2x2 contingency tables. However, as mentioned in the introduction of this thesis, the simple S-value allows for *optional stopping* during an experiment. In this section, first, the interpretation of the definition of (robustness under) optional stopping that will be used in this thesis is clarified. It will thereafter be proved that the simple S-value can be used with optional stopping. Lastly, the expected number of samples collected with optional stopping will be compared to the number of samples needed to collect with Fisher's exact test.

5.1 Robustness under the frequentist interpretation of optional stopping

Researchers have a tendency to collect data until a point is proven; this is called *optional stopping* [Hendriksen et al., 2018]. For example, imagine a biomedical researcher in a lab conducting an experiment to prove the association between a mutation in a gene and the presence of a certain feature in tissue samples. The researcher has performed a power analysis before the start of her project, but after searching for the estimated amount of samples in lab archives and conducting the experiments, she sees that the number of samples with the feature is higher in the group with the mutation, but the hypothesis test result is not significant. The researcher now decides to revisit the archives and add a few more samples to the experiment, leading to a significant result when calculating this p-value as if this new sample size had been fixed in advance. The researcher has now used optional stopping: she has performed multiple hypothesis tests on her data during the collection of samples, and has used these tests to decide when she was going to stop collecting data.

Such a practice is abundant throughout all fields of science[1], but it violates the assumptions of frequentist hypothesis testing: as with most classic forms of frequentist hypothesis testing the *entire research protocol should be determined upfront to ascertain a type-I error guarantee* [John et al., 2012]. Recently, a lot of studies on tests and their robustness under optional stopping have been published, but many of these studies use different definitions of robustness under optional stopping [Hendriksen et al., 2018]. For example, the value taken by a Bayes factor, when equipped with priors that do not depend on the stopping time, does not depend on the stopping rule that was actually used. This has led to claims that Bayesian methods are valid under optional stopping [Hendriksen et al., 2018]. However, independence of the stopping time does not imply a type-I error guarantee of the test [Hendriksen et al., 2018]: tests that do retain a type-I error guarantee

[1] An exception are clinical trials, where a preregistration of the entire research protocol is mandatory

under optional stopping are called *robust* under the frequentist interpretation of optional stopping.

5.2 The simple S-value is robust under optional stopping

It will now be shown that the simple S-value is robust under the frequentist interpretation of optional stopping. With optional stopping, the sample space is $\mathcal{Y}^T$, with T the maximal number of samples that will be collected. For this thesis, we will assume T is a finite number. Testing can be then performed on any sequence $y^n = \{y_1, y_2, \ldots, y_n\}$ with $n \leq T$. For an S-value to be robust under frequentist optional stopping, we require that the probability of rejecting when data are generated under the null hypothesis for at least one of these intermediate sequences x^n is bounded by α, i.e. the type-I error is bounded by α, no matter when we stop:

$$\forall\, P_{\theta_0} \in \mathcal{H}_0 :\; P_{\theta_0}\left(\exists n \in \{1, \ldots, T\} : S_n \text{ well-defined and } S_n(Y^n) \geq \frac{1}{\alpha}\right) \leq \alpha, \tag{5.2.1}$$

with $S_n(y^n)$ the S-value corresponding to the safe test for sample size n. With the simple S-value proposed in the previous chapters for $n_a = n_b$ we have $S_2 = S_4 = \cdots = S_T$; i.e. the priors used for constructing the S-value W_1^* and W_0^* are the same for every sample size n. S_n is only well-defined for *even* n: we have to assume that *group sizes are equal for every* n, i.e. we have $n_a = n_b$ at each time point, and data are collected such that $x^T = \{a, b, a, b, \ldots, a, b\}$. n can thus only take on even values and samples are added in groups of 2. From the definition of the simple S-value, which is a convex combination of two one-sided simple S-values denoted as S_a and S_b, it follows that:

$$\begin{aligned} S_{n+2}(y^{n+2}) &= \frac{1}{2}(S_{a,n+2}(y^{n+2}) + S_{b,n+2}(y^{n+2})) \\ &= \frac{1}{2}(S_{a,n}(y^n)S_{a,2}(y^2) + S_{b,2}(y^2)S_{b,n}(y^n)) \end{aligned} \tag{5.2.2}$$

This process, $S_2, S_4, S_6, \ldots$ is then a *nonnegative supermartingale* under each $P \in \mathcal{H}_0$ [Shafer et al., 2011], since:

$$\begin{aligned} \mathbb{E}_{Y_{n+1},Y_{n+2}\sim P}[S_{n+2}|Y^n] &= \frac{1}{2}\mathbb{E}_{Y_1',Y_2'\sim P}[S_{a,2}(Y_1',Y_2')S_{a,n}(Y^n) + S_{b,2}(Y_1',Y_2')S_{b,n}(Y^n)|Y^n] \\ &= \frac{1}{2}\mathbb{E}_{Y_1',Y_2'\sim P}[S_{a,2}(Y_1',Y_2')S_{a,n}|Y^n] + \frac{1}{2}\mathbb{E}_{Y_1',Y_2'\sim P}[S_{b,2}(Y_1',Y_2')S_{b,n}|Y^n] \\ &= \frac{1}{2}\mathbb{E}_{Y_1',Y_2'\sim P}[S_{a,2}(Y_1',Y_2')]S_{a,n} + \frac{1}{2}\mathbb{E}_{Y_1',Y_2'\sim P}[S_{b,2}(Y_1',Y_2')]S_{b,n} \\ &\leq \frac{1}{2}(S_{a,n} + S_{b,n}) \\ &= S_n. \end{aligned} \tag{5.2.3}$$

and it is well-known that (5.2.1) holds for nonnegative supermartingales [Shafer et al., 2011]. Note that the generalized S-value is also a supermartingale if we use the same adjustment ϕ for every sample size; in the numerical experiments described in Section 3 it was observed that ϕ was almost similar for every n with the same ratio $\frac{n_a}{n_b}$. As elaborated on in Chapter 8, in *future work* it will be aimed to determine the relation between ϕ and $\frac{n_a}{n_b}$ analytically.

For the KL-based S-value found through JIPr, we do not necessarily have that $S_n = S_{n+2}$; a different W_1^* and W_0^* can be obtained for every n, and the (GROW) S-value for a certain sample size n would not necessarily be a (GROW) S-value for another sample size. This implies that a different $P_{W_1^*}$ and $P_{W_0^*}$ would be used for each sample size, and it is not guaranteed that (5.2.2) holds. Hence, the KL-based S-value is not necessarily a supermartingale, and not robust under optional stopping.

5.3 Optional stopping with the simple S-value can decrease the expected sample size actually collected

In the previous chapter, it was illustrated that with Fisher's exact test, a smaller sample size was needed than with the simple S-value before a certain power on a specific $\bar{\Theta}_1(\delta^\circ)$ was reached: a possible disadvantage of the simple S-value. We will now proceed with an example illustrating that, since the simple S-value allows for optional stopping, the expected number of samples needed to reach significance can in fact be lower than the number of samples needed to collect when using Fisher's exact test.

Let us for now define the set of distributions we are interested in as $\Theta_1(\delta^\circ)$, with $\delta^\circ = 0.4$. For clarifying this example, we zoom in on one arbitrarily picked specific distribution P_{θ_1}, with $\theta_1 = (0.3, 0.7)$. For this specific distribution, the minimal sample size for each test to reach a minimum power $1-\beta$ at a certain significance level α can be determined as :

$$n_{\min} = \arg\min_n P_{\theta_1}(\text{pval}(y^n) \leq \alpha) \geq 1-\beta, \tag{5.3.1}$$

where pval in de above equation is replaced by $\frac{1}{S}$ when a safe test is used.

First, $n_{\min}$ for reaching a power of 0.80 with Fisher's exact test was determined. 10,000 samples were generated under P_{θ_1} for each $n \in \{10, 12, \ldots, 100\}$, and Fisher's exact test was performed on each sample to estimate $P_{\theta_1}(T(y^n) \leq \alpha)$ for each n, with $n_a = n_b = \frac{n}{2}$. This simulation experiment revealed that 58 samples (29 for groups a and b) need to be collected to reach a minimum power of 0.80.

For the simple S-value, simulation experiments were also carried out to determine the smallest sample size such that a power of 0.8 was reached for data generated under P_{θ_1}. For each n, different values of δ were explored to determine the δ^* yielding the highest power for that n, as described in Section 3.4. With 10,000 Monte Carlo repetitions for each combination of δ and n, the desired power was reached at $n = 74$, with $\delta^* = 0.3$. This is a bigger sample size than the sample size needed to reach the desired power with Fisher's exact test.

Since the simple S-value allows for optional stopping, the collected number of samples might be smaller then $T = n_{\max}$ after an experiment in hindsight. The expected number of samples to be collected, $\mathbb{E}[\tau]$, can be determined with a simulation study as follows:

$$\mathbb{E}_{P_{\theta_1}}[\tau] \approx \frac{1}{M}\sum_{i=1}^{M} \arg\min_n \left[S(Y_{(i)}^n) \geq \frac{1}{\alpha} \right], \tag{5.3.2}$$

i.e. M samples $Y_{(1)}^T, Y_{(2)}^T, \ldots, Y_{(M)}^T$ were generated from P_{θ_1}. Each partial sequence $Y^2, Y^4, \ldots, Y^T$ with $x^T = (a, b, a, b, \ldots, a, b)$ was tested for significance with the simple S-value with $\delta^* = 0.3$.

For each simulation $i \in \{1, \ldots, M\}$, the smallest n such that Y^n yielded a significant S-value was recorded. If no n smaller than $T = 74$ existed for a sample $Y^T_{(i)}$ such that $S(Y^n_{(i)}) \geq \frac{1}{\alpha}$, T was recorded as the number of samples collected.

The retrieved expected stopping time after this experiment (with $M = 1000$) was 48; the expected sample size needed to collect is 10 samples smaller than the sample size one would use for an experiment to analyze with Fisher's exact test when one aims to reach the same power for P_{θ_1}.

The above described experiment was repeated for several possible values for δ°; for each $\bar{\Theta}_1(\delta^\circ)$, the minimal sample size to reach a power of 0.80 *in the worst case* $\theta \in \bar{\Theta}_1(\delta^\circ)$ was determined for Fisher's exact test and the safe test with the simple S-value. From each $\bar{\Theta}_1(\delta^\circ)$ the expected stopping time for the simple S-value *in the worst case* was then calculated as in (5.3.2). Results are summarized in Figure 5.1: it can be observed that for all δ°, the sample size at the expected stopping time (S (os)) is smaller than or approximately equal to the sample size for Fisher's exact test.

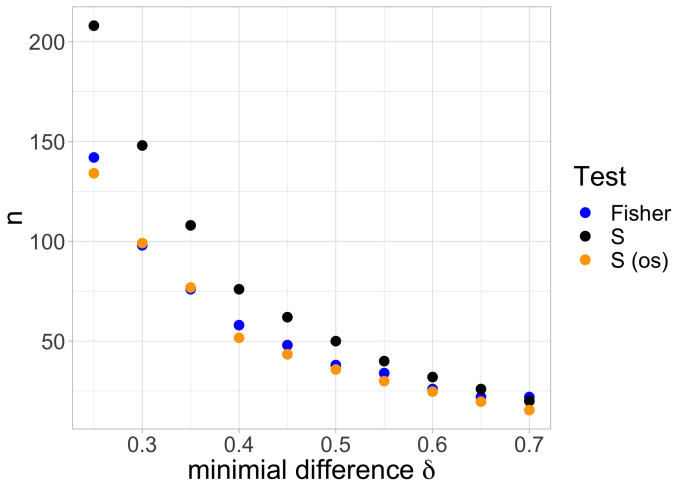


Figure 5.1: Number of samples needed to collect to yield a power of 0.80 for testing at significance level 0.05 for all $P_\theta : \theta \in \bar{\Theta}_1(\delta)$ for Fisher's exact test (Fisher) and the simple S-value (S), and the actual expected number of samples collected after optional stopping (S(os)).

Conclusion

All in all, this section highlighted an important advantage of the simple S-value. As the simple S-value is defined *independently* of sample size, it is a supermartingale and provides robustness under the frequentist interpretation of optional stopping. Allowing for optional stopping decreases the actual sample size one needs to collect: it was illustrated that for most alternative hypotheses $\bar{\Theta}_1(\delta^\circ)$, the expected sample size for the simple S-value was lower than or similar to the sample size needed to collect with Fisher's exact test for achieving a certain minimum power.

Chapter 6

Real-world examples

In this section, the use of the safe test for 2x2 contingency tables is illustrated on real-world examples. The use of the S-values as Bayes factors and as conservative p-values is explored, and the process of choosing $\Theta_1(\delta)$ in the presence or absence of a known α is illustrated. The simple and KL-based S-values will be compared in each example.

6.1 Using S-values as Bayes factors: the doll data

In the original article presenting the implementation of the Gunel-Dickey Bayes factors [Jamil et al., 2017], the 'doll' dataset is used as an example to illustrate the performance of the Bayes factor for the comparison of two proportions in a 2x2 contingency table (see Table 6.1). The doll study concerns an experiment illustrating the existence of racial preference in school children in the American state Nebraska: 89 school children with an Afro-American ethnic background and 71 with a Caucasian background were observed to determine their preference for playing with white or black dolls [Hraba and Grant, 1970]. The null hypothesis that the proportion of black dolls preferred is the same in both groups of children is tested against the alternative hypothesis that the proportions differ between the groups.

Table 6.1: Doll data: comparison of preferences of school children with an Afro-American (AA) or Caucasian (C) ethnic background for doll colours

	Child ethnicity	
Doll	AA	C
Black	62	11
White	27	60
Total	89	71

In Jamil et al. [2017] the Gunel-Dickey Bayes factor based on these research data is calculated to investigate how much the data should "shift our beliefs" about the null and alternative hypotheses. We will now design an S-value *with the aim to use it as a Bayes factor* and compare the resulting safe Bayes factor with the Gunel-Dickey Bayes factor.

As described in Sections 3 and 4, in the absence of a significance level α to test against, the design of the S-value depends on three variables: the sample size n, the desired expected capital growth L and the choice of the restricted alternative hypothesis $\Theta_1(\delta)$ of interest. In this example, n is fixed, with $n_a = 89$ and $n_b = 71$. In Jamil et al. [2017] it is claimed that "strong evidence"

against the null hypothesis is present when a Bayes factor exceeds the value 10. Thus, in our safe test, we want to associate $\mathcal{H}_1$ with a δ such that if $\delta^\circ = \theta_a - \theta_b$ in the real data-generating distribution is bigger than or equal to δ, we can expect to gain strong evidence in favour of the alternative hypothesis. We will pick the δ to design our S-value with as the smallest δ that achieves a capital growth $L = \log(10)$ (see Section 3.4).

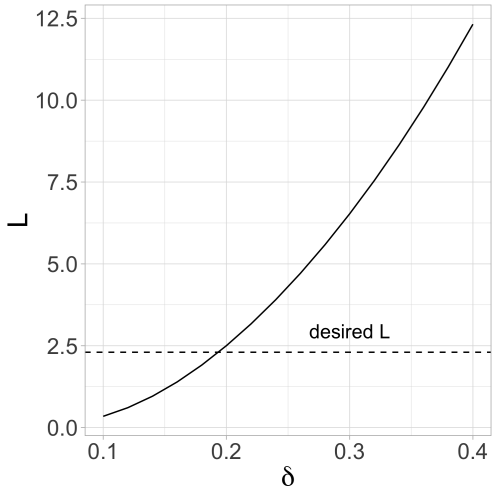
For both the simple (default) S-value and the KL-based S-value, $\Theta_1(\delta)$ was determined as it would be in the `FindDeltaForLDefault` and `FindDeltaForLKL` functions in the `safe2x2` package. For the simple S-values, the grid [0.001, 0.900] of possible δ was explored; for each δ, since $n_a \neq n_b$, the ϕ to translate the simple S-value with was found through optimization, and the expected capital growth was determined. The smallest δ at which the expected capital growth of $\log(10)$ was achieved was 0.194, as can be observed in Figure 6.1; the corresponding $\Theta_1(\delta)$ and $\Theta_0 = 0.5$ are depicted in Figure 6.1b.

For the KL-based S-value, a grid of possible values of δ was explored as well. For all δ in $\{\log(10), \log(15), \log(20), \log(50), \log(100), \log(200), \log(400), \log(600)\}$ the JIPr was determined with $\Theta_1(\delta)$ and Θ_0, with a precision of 50 for both sets to increase the speed of the exploration process. The obtained expected capital growth for each δ is depicted in Figure 6.1c; it can be observed that the desired capital growth rate of $\log(10)$ is achieved for $\delta = \log(50)$. The JIPr for this δ was searched for with a higher precision $(n + 1)$ to determine the GROW S-value; the distribution of the weights in the final S-value (or here: our safe Bayes factor) can be observed in Figure 6.1d.

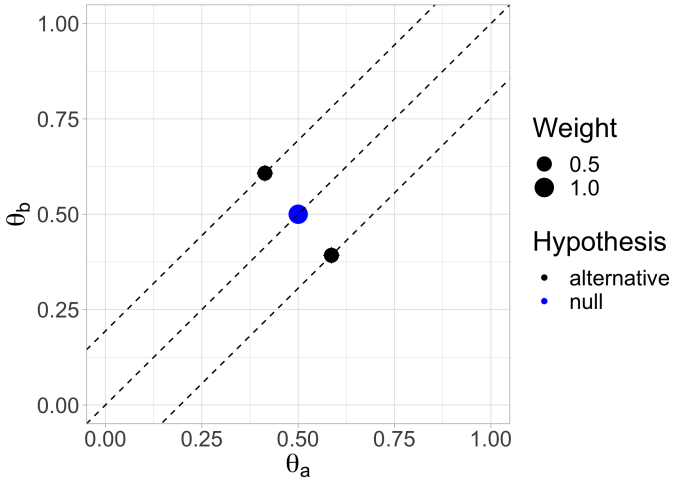
Now that the S-values for the tests we want to perform have been determined, we are allowed to inspect the data. The observed ratio for preferring to play with black dolls is $\frac{62}{89} = 0.70$ in the afro-american group, and $\frac{11}{71} = 0.15$ in the Caucasian group (as can be observed in Table 6.1), with a corresponding odds ratio of $\frac{62*60}{27*11} = 12.5$. Jamil et al. [2017] obtained a Gunel-Dickey Bayes factor of $e^{23.03}$ on these data. Our simple S_{δ^*} revealed a Bayes factor of $e^{13.10} \approx 490,000$. The KL-based S_{δ^*} yielded a Bayes factor of $e^{14.03} \approx 1,250,000$. All Bayes factors yield 'extreme' evidence in preference of the alternative hypothesis (according to the Bayes factor interpretation as proposed by Jamil et al. [2017]: $B_{10} > 100$ indicates extreme evidence). Although the safe Bayes factors yield lower values, we would still conclude that we would have to significantly 'shift our beliefs' from the null to the alternative hypothesis from all tests.

6.2 Using S-values as conservative p-values, part I: canine therapy data

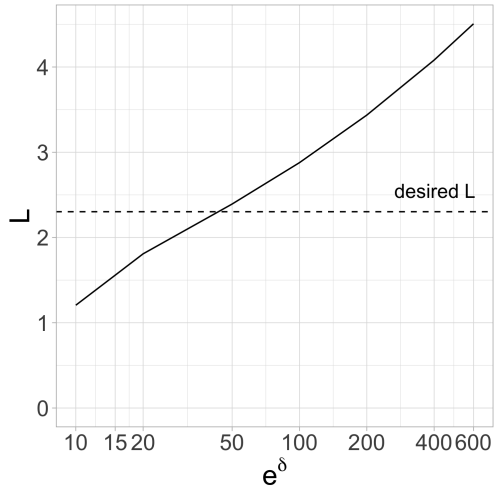
Next, the performance of the simple and KL-based S-values as conservative p-values was compared to the outcome of Fisher's exact test. As a first example, a recent study about the impact of canine therapy for anxious patients at an emergency department was used [Kline et al., 2019]. According to Kline et al. [2019], patients at the emergency department often experience high anxiety, as objectively determined by several anxiety scales. However, most clinicians do not generally address this anxiety in patients, as they are presumably wary of labelling patients as anxious (as such a diagnosis might discount the primary reason the patient searches emergency care for) and they are also reluctant with prescribing anxiolytics (medication aimed at relieving anxiety) because of their side effects. Addressing anxiety at the emergency department in another way might improve patients' experiences and emergency department visit outcomes. As therapy dogs have proven to reduce anxiety in other patient groups, the authors hypothesized that they might decrease anxiety



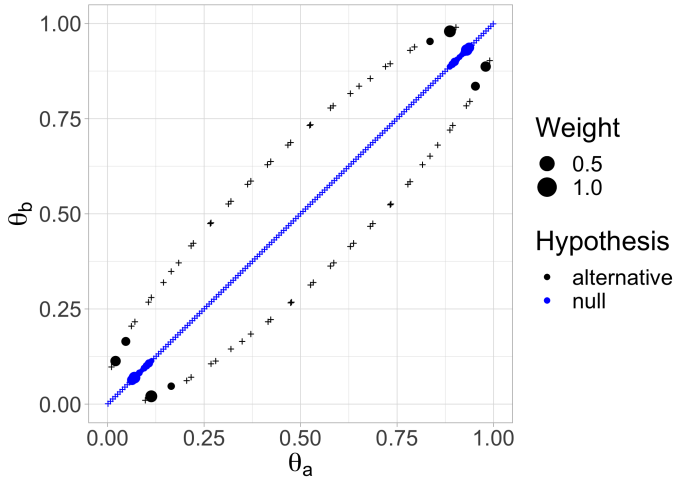
(a) Obtained expected capital growth for (generalized) simple S-values



(b) Simple S-value at $\delta = 0.194$



(c) Obtained expected capital growth with KL-based S-values constructed with various δ



(d) JIPr with KL-based δ, $\delta = \log(50)$

Figure 6.1: Determining δ for the simple S-value and the KL-based S-value for the doll experiment; $n_a = 89$, $n_b = 71$. Figures (a) and (c) depict the expected capital growth $L = \mathbb{E}_{P_{W_1^*}}[\log(S_\delta^*(Y^n))]$ for various δ used to construct the S-value with. Figures (b) and (d) depict the support points and weights in the final S-values: crosses indicate positions of grid points in Θ_1 and Θ_0, dots and their size represent the weight from W_0^* and W_1^* for each point.

in anxious emergency department visitors as well.

One of the hypotheses in Kline et al. [2019] was the hypothesis that therapy dogs sessions could change the proportion of patients needing anxiolytics in the emergency department. To investigate this, a randomised controlled trial was performed at their emergency department in Indiana, USA: 40 anxious patients at the emergency department were given a 15-minute session with a trained therapy dog, and 40 patients were merely observed as controls. The number of patients needing anxiolytic medication was recorded in each group, and Fisher's exact test was performed to test

the hypothesis that the proportions of patients needing medication were the same in the canine therapy group and the control group versus the alternative hypothesis, that these proportions differed. P-values below 0.05 were considered significant by the authors [Kline et al., 2019].

We now proceed to designing the simple S-value and the KL-based S-value for this test. As the sample size n and the significance level for testing α are known, we could look for the δ^* that maximises the expected *power* in the worst case, as described in Section 3.4.

Table 6.2: Finding the δ^* that provides the highest power in the worst case for $\Theta(\delta^\circ)$ for the simple S-value for $n_a = n_b = 40$; $\alpha = 0.05$.

δ°	δ^*	$\min(1-\beta)$
0.20	0.21	0.04
0.30	0.21	0.40
0.40	0.21	0.79
0.50	0.21	0.96
0.60	0.21	0.99
0.70	0.21	1.00

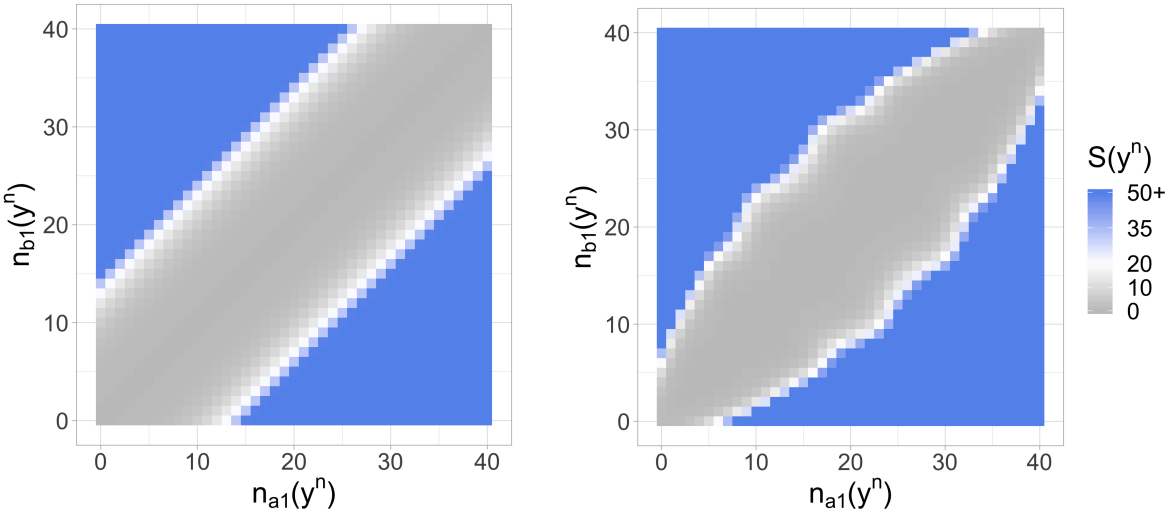
For the design of the simple S-value, δ^* and its corresponding worst-case power were found for a grid of data-generating distributions $\bar{\Theta}(\delta^\circ))$ with $\delta^\circ \in [0.2, 0.7]$. For each distribution in each δ°, 1000 samples were generated and S-values for various S_{δ^*} were calculated. Results for the obtained best worst-case power are illustrated in Table 6.2; it can be observed that $\delta^* = 0.21$ is the most prevalent value. A plot of the obtained S-values for a grid of possible combinations of outcomes n_{a1} and n_{b1} for the simple S-value with this δ^* is depicted in Figure 6.2a.

For the design of the KL-based S-value, data were again generated under $\Theta_1(\delta^\circ)$ for various (δ°) and the δ^* yielding the best power in the worst case for each (δ°) was recorded. Results are summarised in Table 6.3; it can be observed that overall $\delta^* = \log(20)$ yielded the best worst-case power for all but one δ°. A plot of the obtained test values for a grid of possible combinations of outcomes n_{a1} and n_{b1} with this KL-based S-value is depicted in Figure 6.2b.

Table 6.3: Finding the δ^* that provides the highest power in the worst case for $\Theta(\delta^\circ)$ for the KL-based S-value for $n_a = n_b = 40$; $\alpha = 0.05$.

δ°	δ^*	$\min(1-\beta)$
$\log(16)$	$\log(100)$	0.303
$\log(100)$	$\log(20)$	0.576
$\log(200)$	$\log(20)$	0.669
$\log(400)$	$\log(20)$	0.716
$\log(1000)$	$\log(20)$	0.809
$\log(2000)$	$\log(20)$	0.863

After the design of the S-values, the observed data were inspected: in the group receiving canine therapy, 1 out of 40 patients required anxiolytics, whereas in the control group, 7 patients received anxiolytics [Kline et al., 2019]. With the simple S-value, the obtained value was 1.06. With the KL-based S-value, it was 7.31. This difference is interesting; the KL-based S-value yields a much



(a) Obtained value per observed sequence with the simple S-value, $\delta^* = 0.21$

(b) Obtained value per observed sequence the KL-based S-value, $\delta^* = \log(20)$

Figure 6.2: Obtained S-values for each possible outcome y^n for the simple S-value and the KL-based S-value for the canine experiment; $n_a = 40$, $n_b = 40$, $\alpha = 0.05$. The white area indicates outcomes that result in an S-value of $\alpha^{-1} = 20$; the blue area indicates observations that will lead to rejection of the null hypothesis.

higher S-value than the simple S-value. This could possibly be explained through the difference in performance of the S-values for 'extreme outcomes' with either low values of n_{a1} or n_{b1}, as illustrated in Figure 6.2. The KL-based S-value yields higher values than the simple S-value for such outcomes. Fisher's exact test yielded a p-value of 0.056. The S-values as conservative p-values were higher ($\frac{1}{1.06} = 0.94$ and $\frac{1}{7.31} = 0.14$, respectively) than the p-value from Fisher's exact test. Based on all three tests, the null hypothesis that canine therapy does not have an influence on the proportion of patients requiring anxiolytics can not be rejected.

6.3 Using S-values as conservative p-values, part II: ECMO data

As a last experiment on real data, the performance of the simple and KL-based S-values as conservative p-values on a study where $n_a \neq n_b$ was explored. The selected study was regarded as particularly interesting for assessing the performance of the safe tests in comparison to Fisher's exact test, as multiple hypotheses were tested in the study that yielded both borderline significant p-values, and very low p-values.

The study concerned a retrospective study in patients who received ECMO treatment: a treatment that for a short term can replace the function of the heart and lungs, in case there is severe failure of one of these organs [Fletcher-Sandersjöö et al., 2017]. With ECMO treatment, a patient's blood is redirected through a machine that substitutes the functions of the heart and lungs. However, this machine can not entirely mimic the environment the blood is naturally present in. Normally, the blood vessels constantly send signals to the blood to confirm that they are intact, and that the blood does not need to clot (an extensive overview can, for example, be found in Turner [2018]). However, the artificial blood vessels in the ECMO machine can not send such

signals; therefore, strong anticoagulant (anti-clotting) medication has to be administered to the patient to prevent massive blood clotting in the ECMO machine. A frequent adverse effect of strong anticoagulants is the development of severe haemorrhaging (internal bruising), of which intracranial haemorrhaging (ICH) is one of the most devastating complications, with a high mortality rate [Fletcher-Sandersjöö et al., 2017].

Hence, the authors aimed to identify characteristics associated with survival after 30 days in patients who developed ICH while on ECMO treatment. To achieve this, they collected data from 17 patients who survived ICH during ECMO, and data from 48 deceased patients, and they collected information on several characteristics of specific kinds of damage in the brain.

As with the canine therapy data example, first, the 'most powerful' δ^* for this combination of $n_a = 17$ and $n_b = 48$ was determined for the simple- and the KL-based S-values. For the simple S-value, δ^* was 0.33, and for the KL-based S-value, δ^* was $\log(16)$. Safe tests were conducted for a selection of the hypothesis tests performed in Fletcher-Sandersjöö et al. [2017] and conservative p-values were calculated; results are summarized in Table 6.4. One can observe that the S-values are more conservative than the p-values obtained through Fisher's exact test: with Fisher's exact test, four out of five null hypotheses were rejected, but with the safe tests, only three null hypotheses were rejected. As with the canine data example, most of the collected data were observations with extremely low values in group a; as we saw above, the KL-based S-value yields higher S-values (and lower p-values) in such cases, which could explain the higher S-values and lower conservative p-values resulting from the KL-based test in this example.

Table 6.4: Observed occurrence of brain damage characteristics in patients that survived after an ICH during ECMO ($n_a = 17$) and deceased patients ($n_b = 48$) and (strict and conservative) p-values resulting from Fisher's exact test, the simple safe test and the KL-based safe test.

Characteristic	n_{a1}	n_{b1}	Fisher	Simple S-value	KL-based S-value
Intraparenchymal haemorrhage	9	39	0.050	0.277	0.242
Intraventricular heamorrhage	1	24	0.001	0.016	0.013
Hydrocephalus	1	30	<0.001	0.002	0.002
Absent basal cisterns	1	35	<0.001	<0.001	<0.001
Ischemic stroke	4	21	0.161	>1	0.84

Conclusion

Throughout Chapters 3 and 4 we observed an important advantage of the simple S-value over the KL-based S-value: for finding the KL-based S-value, numerical optimization over all parameters in the weights vectors W_0 and W_1 has to be carried out, whereas the (generalized) simple S-value only has to be optimized for ϕ, the adjustment needed in the case of unequal group sizes. The experiments in the current section however illustrate that cases exist where use of the KL-based S-value should be preferred: when one expects that the differences between the groups exist at the *edges* of the parameter space, i.e. one of the group means is expected to be close to 0 or 1, the KL-based S-value yields higher values than the simple S-value. These results are in agreement with the power experiment conducted in Chapter 4, Figure 4.3.

Chapter 7

A safe version of the Cochran-Mantel-Haenszel test

In the preceding sections of this thesis, safe versions of the 2x2 contingency table test were developed. It will now be investigated if a safe version of *stratified* versions of the 2x2 contingency table test can be developed as well. First, the motivation for designing such a test will be elaborated on. Next, the null- and alternative hypotheses and corresponding likelihoods for two stratified versions of the 2x2 contingency table test will be defined: for the classical CMH test, and for a modified version of that test, the Breslow-Day test. In the third subsection results from numerical experiments for finding the JIPr between the distributions in the null and restricted alternative hypothesis for the defined likelihoods are depicted, and again *a simple S-value is designed for equal subgroup sizes*. Lastly, power of the obtained safe versions of the CMH test and the Breslow-Day test is compared to the power of their classical frequentist counterparts.

7.1 Motivation for a *stratified* version of the 2x2 contingency table test

During the midst of the twentieth century, a pressing problem in statistics was the analysis of case-control studies [Breslow, 1996]. With a case-control study design, one aims to compare a group of subjects with a certain symptom or disease with a group of control subjects with regard to the presence of certain characteristics (e.g. interventions, habits, physical or genetic features) [Hofman and Grobbee, 2003]. In this way, characteristics associated with the disease can be discovered, and this might provide insight into its pathophysiology. It is evident that a 2x2 contingency table test would be the hypothesis test of choice for this situation, where we are comparing two groups with regard to the presence or absence of a characteristic. However, case-control studies are often designed retrospectively; subjects with and without the disease are searched for in archives, and information on the features one is interested in is extracted from the archives as well, or through questionnaires. Compared to a prospective design, where patients and control subjects are included in the study in a controlled environment, the retrospective design has some clear disadvantages, as stated by Mantel and Haenszel in their original paper presenting the CMH test:

> "*In addition to the failings shared with the forward study, the retrospective study is further exposed to misleading associations arising from the circumstances under which test and control subjects are obtained.*" - Mantel and Haenszel [1959], p. 723.

In other words, since a retrospective study is not conducted under a controlled environment, circumstances that could have influenced the observations in the groups might be present, influencing

the hypothesis test's result. Therefore, standard 2x2 contingency table tests' assumptions might not be satisfied, and test outcomes could be misleading. This is called *Simpson's paradox*: associations found in a 2x2 contingency table test could disappear or even have the opposite direction after splitting the table into subtables through conditioning on a third factor [Agresti, 2002].

Mantel and Haenszel in 1959 proposed a solution for this challenge in the form of what is most commonly referred to as the Cochran-Mantel-Haenszel test [Mantel and Haenszel, 1959], as the CMH-test is a modified version of the Chi-squared test, building on earlier work of Cochran [Breslow, 1996, Cochran, 1954]. When one is aware of a factor that might have confounded the association between the variables of interest, one can stratify the data with it. Instead of comparing the total (marginal) observed cell counts with expected cell counts, the total (marginal) table is split up into partial tables conditional on the possibly confounding factor, and the observed cell counts are compared to the expected cell counts in each of the subtables.

7.2 Defining hypotheses and likelihoods for the safe CMH-test

Let us define the obtained subtables after conditioning as table 1, table 2, ..., table k when we are stratifying over k categories. As with the 2x2 test, we observe a sequence of outcomes y^n, with $y_j \in \{0, 1\}$, and have fixed group assignments for each observation, x^n, with $x_j \in \{a, b\}$. We now also have fixed observations of a third feature we are conditioning on, z^n, with $z_j \in \{1, \ldots, k\}$. Each observation y_j with group x_j will be placed in subtable i if $z_j = i$. The total number of observations in one subtable is then denoted as $n_{.i}$. The total counts in each group within each subtable are denoted as $n_{a.i}$ and $n_{b.i}$. This notation for the stratified data is summarized in Table 7.1 below.

Table 7.1: Partial table i after stratification

	0	1	total
a	$n_{a0.i}$	$n_{a1.i}$	$n_{a.i}$
b	$n_{b0.i}$	$n_{b1.i}$	$n_{b.i}$
	$n_{0.i}$	$n_{1.i}$	$n_{.i}$

With the stratified/ conditional versions of the 2x2 contingency table test, one aims to test the null hypothesis *that the odds ratios O_i in each partial table are equal to one* [Agresti, 2002], which implies testing that $\frac{\theta_{a.i}(1-\theta_{b.i})}{(1-\theta_{a.i})\theta_{b.i}} = 1$. Clearly, this is equivalent to testing the null hypothesis that the data within each subtable are distributed as a single Bernoulli distribution with mean $\theta_{0.i}$. To see where the name odds ratio test comes from, note that the odds ratios are defined as the empirical analogue of the above:

$$O_i = \frac{n_{a1.i}n_{b0.i}}{n_{a0.i}n_{b1.i}} = \frac{\frac{n_{a1.i}}{n_{a.i}}}{\frac{(n_{a.i}-n_{a1.i})}{n_{a.i}}} \frac{\frac{(n_{b.i}-n_{b1.i})}{n_{b.i}}}{\frac{n_{b1.i}}{n_{b.i}}} = \frac{\hat{\theta}_{a.i}(1-\hat{\theta}_{b.i})}{(1-\hat{\theta}_{a.i})\hat{\theta}_{b.i}}. \tag{7.2.1}$$

From a simple derivation (see Appendix (A.3.1)) it follows that $O_i = 1$ if (and only if) $\hat{\theta}_{a.i} = \hat{\theta}_{b.i}$. Under the null hypothesis, the probability of observing a sequence $y^n \in Y^n$ then becomes:

$$p_{\theta_0}(y^n|z^n) = \prod_{i=0}^{k} \theta_{0.i}^{n_{1.i}}(1-\theta_{0.i})^{n_{0.i}}, \tag{7.2.2}$$

and θ_0 now is a vector of the means in each subtable: $\theta_0 = (\theta_{0.1}, \theta_{0.2}, \ldots, \theta_{0.k})$, with $\theta_{0.i} \in [0, 1]$.

Under the alternative hypothesis, the group means for group a and b are different within each subtable. Let us denote the group mean in group a in subtable i as $\theta_{a.i}$, and the group mean in group b as $\theta_{b.i}$. The likelihood under the alternative hypothesis of observing a sequence $y^n \in Y^n$ then becomes:

$$p_{\theta_1}(y^n|x^n, z^n) = \prod_{i=1}^{k} \theta_{a.i}^{n_{a1.i}}(1-\theta_{a.i})^{n_{a0.i}}\theta_{b.i}^{n_{b1.i}}(1-\theta_{b.i})^{n_{b0.i}}. \tag{7.2.3}$$

Hence, θ_1 is again a vector of parameters: $\theta_1 = (\theta_{a.1}, \theta_{b.1}, \ldots, \theta_{a.k}, \theta_{b.k})$, with $\theta_{.i} \in [0, 1]$.

With regard to the choice of the alternative hypothesis, two different versions of this "stratified" version of the 2x2 contingency table test will now be considered. When Mantel and Haenszel first introduced the CMH-test, they aimed to test the alternative hypothesis that the difference in proportions between the groups is present *and consistent* after conditioning: the proportion should be consistently bigger or smaller in group a compared to group b within each subtable [Mantel and Haenszel, 1959]. This then yields the following composition of the alternative hypothesis:

$$\mathcal{H}_{1,\text{CMH}} = \{P_{\theta_1} : \forall i \in \{1, \ldots, k\} : \theta_{a.i} - \theta_{b.i} > 0\} \cup \{P_{\theta_1} : \forall i \in \{1, \ldots, k\} : \theta_{a.i} - \theta_{b.i} < 0\}. \tag{7.2.4}$$

When there would be *opposite associations* in the subtables after conditioning, these would be missed with the CMH test. When one would also want to allow for testing for contradictory associations after conditioning, one should therefore use a different test: the test for homogeneity of the odds, or the *Breslow-Day* test [Breslow, 1996]. With this test, one then can test the null hypothesis that the odds ratios are equal to one to the alternative that they are not equal to one, yielding the following composition of the alternative hypothesis:

$$\mathcal{H}_{1,\text{B}} = \{P_{\theta_1} : \forall i \in \{1, \ldots, k\} : \theta_{a.i} \neq \theta_{b.i}\}. \tag{7.2.5}$$

As with the 2x2 contingency tables, priors W_1 and W_0 can be put on Θ_0 and Θ_1 for each of these alternative hypotheses, and the GROW S-value can be found through minimizing the KL-divergence over these priors. The resulting composition of the GROW S-value for both alternative hypotheses constructed with the *default* distance measure will be explored below.

7.3 Finding the JIPr for the default safe CMH-test

Let us consider the stratified extension of the *default* S-value for 2x2 contingency tables. Analogous to the 2x2 contingency table test, we define $\bar{\Theta}_1(\delta)$ as:

$$\bar{\Theta}_1(\delta) = \{\theta_1 : \forall i \in \{1, \ldots, k\} : |\theta_{a.i} - \theta_{b.i}| = \delta\}. \tag{7.3.1}$$

$\bar{\Theta}_1(\delta)$ and Θ_0 now have to be discretized to enable numerical approximation of the JIPr, so they were again approximated with grids. To do so for $\bar{\Theta}_1(\delta)$, first, a two-dimensional grid of points representing the parameters for one of the subtables is constructed, $\Theta_i(\delta)$, analogous to how $\bar{\Theta}_1(\delta)$ would be constructed with the safe 2x2 contingency table test. An example of such a grid is depicted in Figure 2.1. Next, the k-fold cross-product of this grid, $\Theta_i(\delta)^k$, is constructed to form $\bar{\Theta}_1(\delta)$; for the CMH test, combinations where the sign of $\theta_{a.i} - \theta_{b.i}$ is not equal for all parameter pairs have to be removed. For Θ_0, the grid can be constructed as all combinations of size k of an

equally spaced grid on $[0,1]$: $\Theta_0 = [0,1]^k$.

As an example, grids were determined and the JIPr was found for the CMH and Breslow-Day tests for the case where $n_{a.1} = n_{b.1} = n_{a.2} = n_{b.2} = 10$, $\delta = 0.4$ and $k = 2$. The grid precision in $\Theta_i(\delta)$ was set to 6 and $|\Theta_0|$ was 25 in both experiments. The JIPr was again found with `slsqp`. $\Theta_1(\delta)$ and the found W_1^* for both cases are summarized in Table 7.2.

Results from this experiment gave rise to the hypothesis that an extension of the *simple S-value* can be designed for the CMH and Breslow-Day tests as well. In W_0^*, all mass was put on $\theta_0 = (0.5, 0.5)$. Further, it was observed that in W_1^*, mass was put uniformly on *all possible combinations of the support points of the simple S-value for 2x2 contingency tables.* I.e. in the Breslow day test, all mass was put uniformly on the four combinations of the support points $\theta_{1.1}$ and $\theta_{1.2}$ as defined in (3.2.1).

Computational complexity

Increasing precision in the grid in a single subtable for the CMH- and Breslow-Day test leads to a big increase in total computational complexity. For the Breslow-Day test, when the number of grid points in the restricted alternative hypothesis is equal to g for one subtable, this results in a size g^k for the entire $\Theta_1(\delta)$. With the CMH test, the size of the restricted alternative hypothesis set becomes $\frac{g^k}{2^{k-1}}$. Thus, the number of weights to optimize over for finding the JIPr increases exponentially with the number of subtables k.

While finding the JIPr through numerical optimization, the likelihood under the alternative hypothesis has to be calculated for each possible combination of θ_1, $n_{a1.i}$ and $n_{b1.i}$ to calculate the expected value of $\log(S(Y^n))$. As k increases, the number of likelihoods to calculate and store thus increases exponentially, leading to an exponential increase in required multiplication actions at each iteration of the optimization process. For example, for $n_{a.1} = n_{b.1} = n_{a.2} = n_{b.2} = 20$, there exist $(20+1)^4 = 194,481$ possible combinations of $n_{a.1}, n_{b.1}, n_{a.2}$ and $n_{b.2}$. Would we want to use a grid with precision 50 for the alternative hypothesis, we would need to store a matrix with $\approx 10 \times 10^6$ float values. With double precision, this object would require $8 \times 10 \times 10^6$ bytes $= 80$ MB to store. Extending this setup to $k = 3$ would result in a matrix of 34 GB and would require 4×10^9 multiplications per iteration: it is infeasible to perform such computations on most personal computers. As a first solution for this, in the next section, the simple S-value will be extended to the CMH and Breslow-Day tests, which also enables omitting finding the JIPr for large sample sizes.

In Section 4, numerical experiments revealed that finding the JIPr with `slsqp` for a grid of Θ_1 and Θ_0 with precisions > 100 can lead to computation times of up to 5 minutes. Hence, it appears unwanted to include too many grid points while determining the JIPr. As we also have seen in Section 4, there appear to be at least 12 support points in the JIPr for 2x2 tables in the KL-based restricted alternative hypothesis, and these support points have different locations for different n. Hence, to at least cover all these support points for the parameters in all subtables for the CMH-test and Breslow-Day test, the JIPr would have to be determined for a very large grid of parameters. It was therefore decided that designing the KL-based S-value for the CMH and Breslow-Day tests was for now infeasible.

Table 7.2: Grids used for discretizing $\Theta_1(\delta)$ and found weights for W_1^* for the JIPr for the Breslow-Day and CMH version of the stratified 2x2 contingency table test. Note that the grid representing the restricted alternative hypothesis for the CMH test contains all grid points from the Breslow-Day test, except for the grid points where the difference between the two group means does not have the same sign in each subtable. $k = 2$, $n_{a.1} = n_{b.1} = n_{a.2} = n_{b.2} = 10$ and $\delta = 0.4$. Highlighted rows indicate found support points in $\Theta_1(\delta)$.

Breslow-Day					CMH				
$\theta_{a.1}$	$\theta_{b.1}$	$\theta_{a.2}$	$\theta_{b.2}$	w_1	$\theta_{a.1}$	$\theta_{b.1}$	$\theta_{a.2}$	$\theta_{b.2}$	w_1
0.3	0.7	0.3	0.7	0.249	0.3	0.7	0.3	0.7	0.500
0.3	0.7	0.7	0.3	0.249					
0.3	0.7	0.1	0.5	0.000	0.3	0.7	0.1	0.5	0.000
0.3	0.7	0.5	0.1	0.000					
0.3	0.7	0.5	0.9	0.000	0.3	0.7	0.5	0.9	0.000
0.3	0.7	0.9	0.5	0.000					
0.7	0.3	0.3	0.7	0.249					
0.7	0.3	0.7	0.3	0.249	0.7	0.3	0.7	0.3	0.500
0.7	0.3	0.1	0.5	0.000					
0.7	0.3	0.5	0.1	0.000	0.7	0.3	0.5	0.1	0.000
0.7	0.3	0.5	0.9	0.000					
0.7	0.3	0.9	0.5	0.000	0.7	0.3	0.9	0.5	0.000
0.1	0.5	0.3	0.7	0.000	0.1	0.5	0.3	0.7	0.000
0.1	0.5	0.7	0.3	0.000					
0.1	0.5	0.1	0.5	0.000	0.1	0.5	0.1	0.5	0.000
0.1	0.5	0.5	0.1	0.000					
0.1	0.5	0.5	0.9	0.000	0.1	0.5	0.5	0.9	0.000
0.1	0.5	0.9	0.5	0.000					
0.5	0.1	0.3	0.7	0.000					
0.5	0.1	0.7	0.3	0.000	0.5	0.1	0.7	0.3	0.000
0.5	0.1	0.1	0.5	0.000					
0.5	0.1	0.5	0.1	0.000	0.5	0.1	0.5	0.1	0.000
0.5	0.1	0.5	0.9	0.000					
0.5	0.1	0.9	0.5	0.000	0.5	0.1	0.9	0.5	0.000
0.5	0.9	0.3	0.7	0.000	0.5	0.9	0.3	0.7	0.000
0.5	0.9	0.7	0.3	0.000					
0.5	0.9	0.1	0.5	0.000	0.5	0.9	0.1	0.5	0.000
0.5	0.9	0.5	0.1	0.000					
0.5	0.9	0.5	0.9	0.000	0.5	0.9	0.5	0.9	0.000
0.5	0.9	0.9	0.5	0.000					
0.9	0.5	0.3	0.7	0.000					
0.9	0.5	0.7	0.3	0.000	0.9	0.5	0.7	0.3	0.000
0.9	0.5	0.1	0.5	0.000					
0.9	0.5	0.5	0.1	0.000	0.9	0.5	0.5	0.1	0.000
0.9	0.5	0.5	0.9	0.000					
0.9	0.5	0.9	0.5	0.000	0.9	0.5	0.9	0.5	0.000

7.4 The simple S-value for 2x2 contingency tables can be extended to the safe CMH-test

Analogous to Chapter 3, it will now be illustrated that the simple S-value can be extended to the safe versions of the CMH-test and the Breslow-Day test. The support point in the null hypothesis with the simple S-value in this case is defined as:

$$\theta^*_{0,\text{simple}} = \{0.5\}^k. \tag{7.4.1}$$

For the *simple* Breslow-Day test, the support points in the alternative are defined as the k-fold cartesian product of the support points in the restricted alternative hypothesis for the simple S-value for the 2x2 contingency tables:

$$\Theta^*_{1,\text{Breslow}}(\delta) = \{\theta_{1.1}, \theta_{1.2}\}^k, \tag{7.4.2}$$

with $\theta_{1.1}$ and $\theta_{1.2}$ as defined in (3.2.1), and with $W_{1,\text{Breslow}}$ a probability distribution with uniform mass on the 2^k parameters in $\Theta^*_{1,\text{Breslow}}(\delta)$.

For the *simple* CMH-test, after removal of the points where the sign of $\theta_{a.i} - \theta_{b.i}$ is not equal for all i, the alternative hypothesis has two support points:

$$\Theta^*_{1,\text{CMH}}(\delta) = \left\{(\theta_{1.1})^k, (\theta_{1.2})^k\right\}, \tag{7.4.3}$$

with $W_{1,\text{CMH}}$ the probability distribution with uniform mass on these 2 points. In Appendix Figure A.3, it is numerically illustrated that the simple CMH test is a safe test.

The simple S-value is not the GROW S-value for the entire restricted alternative hypothesis $\Theta_1(\delta)$. However, as with the 2x2 contingency tables case, an *angled* restricted alternative hypothesis for each subtable can be defined, $\Theta'_i(\delta) \subset \Theta_i(\delta)$ (for an example, see Figure 3.2). The restricted alternative hypothesis of interest $\Theta'_1(\delta)$ can then be defined of the k-fold cross-product of $\Theta'_i(\delta)$. For this subset of the restricted alternative hypothesis, it can be shown that the simple S-value *is* GROW for a certain minimal adjustment δ', i.e. the expected capital growth under each $P_{\theta_1} : \theta_1 \in \Theta'_1(\delta)$ is greater than or equal to the expected capital growth under $P_{W^*_1}$.

Table 7.3: CMH: minimal adjustment δ' of the restricted alternative hypothesis for the simple S-value to be GROW for various combinations of δ and n. $k = 2$, $n_{a.1} = n_{b.1} = n_{a.2} = n_{b.2} = \frac{n}{4}$

n \ δ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
20	0.20000	0.10000	0.04000	0.01000	0.00400	0.00040	0.00010	0.00010	0.00010
28	0.10000	0.04000	0.02000	0.00400	0.00100	0.00010	0.00010	0.00010	0.00010
36	0.10000	0.04000	0.01000	0.00200	0.00020	0.00010	0.00010	0.00010	0.00010
44	0.10000	0.02000	0.01000	0.00100	0.00010	0.00010	0.00010	0.00010	0.00010
52	0.10000	0.02000	0.00400	0.00040	0.00010	0.00010	0.00010	0.00010	$< 10^{-4}$
60	0.10000	0.02000	0.00200	0.00020	0.00010	0.00010	0.00010	0.00010	$< 10^{-4}$

In Tables 7.3 and 7.4, the minimal adjustment δ' needed to make to the restricted alternative hypothesis for the simple S-value to be GROW for various combinations of the sample size n and distance measure δ are presented for the safe CMH and Breslow-Day tests, respectively. The

Table 7.4: Breslow-Day: minimal adjustment δ' of the restricted alternative hypothesis for the simple S-value to be GROW for various combinations of δ and n. $k = 2$, $n_{a.1} = n_{b.1} = n_{a.2} = n_{b.2} = \frac{n}{4}$

n \ δ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
20	0.20000	0.20000	0.10000	0.04000	0.02000	0.01000	0.00200	0.00020	0.00010
28	0.20000	0.10000	0.04000	0.02000	0.01000	0.00200	0.00040	0.00010	0.00010
36	0.20000	0.10000	0.04000	0.01000	0.00400	0.00100	0.00010	0.00010	0.00010
44	0.10000	0.10000	0.02000	0.01000	0.00200	0.00020	0.00010	0.00010	0.00010
52	0.10000	0.04000	0.02000	0.00400	0.00100	0.00010	0.00010	0.00010	0.00010
60	0.10000	0.04000	0.02000	0.00400	0.00040	0.00010	0.00010	0.00010	0.00010

experiments summarized in these Tables were conducted for $k = 2$, and equal sizes of subtables and subgroups. It can be observed that $\delta' \to 0$ as $n \to \infty$ and $\delta \to 1$: hence, we can conclude that *for equal subtable and subgroup size, the subset of the restricted alternative hypothesis the simple S-value is not GROW for becomes negligibly small quickly.* In the Appendix Tables A.4 and A.5, it is illustrated that these results extend to the case where $k = 3$, and to the case where $n_{.1} \neq n_{.2}$, but $n_{a.i} = n_{b.i}$.

For the case where $\exists i \in \{1, \ldots, k\}$ s.t. $n_{a.i} \neq n_{b.i}$, the obtained W_1^* and W_0^* after finding the JIPr did not mimic the pattern of the simple S-value. An example is presented in Appendix Table A.6. Hence, it is proposed that for designing the safe test for the CMH and Breslow-Day test in this case, the JIPr should be approximated through optimization. As with the 2x2 contingency table test, the found minimal KL-divergence between P_{W_1} and P_{W_0} did not decrease further after the use of a certain minimal grid precision; this is illustrated in Table 7.5. It can be observed that the found minimum was the lowest after using a grid with 72 points for representing the restricted alternative hypothesis for the case where $n = 30$. Also, it can be noted that with the grid with 8 points the found minimum was very high; this can be explained because the points around the line $\theta_{b.i} = 1 - \theta_{a.i}$ were not included in this set. As can be observed in Tables 7.2 and A.6, points including these parameter values receive a lot of weight with the JIPr.

Table 7.5: Obtained KL-divergence and computation time for finding the JIPr for the CMH test with $n_{a.i} = 10$ and $n_{b.i} = 5$, $k = 2$ and $\delta = 0.4$, with different grid sizes for discretizing the restricted alternative hypothesis.

$\lvert\Theta_i(\delta)\rvert$	$\lvert\Theta_1(\delta)\rvert$	KL divergence	time (s)
2	2	1.7910	0.2
4	8	3.8728	2.0
6	18	1.7433	1.3
8	32	1.6139	1.3
10	50	1.6049	2.3
12	72	1.5602	6.2
14	98	1.5710	3.8
16	128	1.5628	11.5
18	162	1.5614	21.4
20	200	1.5622	35.3

7.5 Power of the safe CMH test

To compare the power of the safe CMH and Breslow-Day tests to their frequentist counterparts, the simple $S^*_{\delta^*}$ providing the best power in the worst-case scenario for different alternative hypotheses of interest $\bar{\Theta}_1(\delta^\circ)$ was determined, as described in Section 3.4. For the case where $k = 2$ and the size of each subgroup is 10, the found δ^* for the CMH and Breslow-Day test was 0.31, as illustrated in Appendix Table A.7 for the CMH test.

To enable visualizing the obtained power of the tests in two dimensions, 'observed' values of $n_{a1.2}$ and $n_{b1.2}$ were set fixed for each experiment. In Figure 7.1, results of experiments with both the simple S-value and the S-value found through optimization for three different fixed values of $n_{a1.2}$ and $n_{b1.2}$ are summarized. It can be observed that the power of the safe tests is lower than the power obtained with the classical version of the CMH test. The behaviour of the safe tests is similar to the behaviour of the frequentist CMH test; when the observed odds ratio in subtable 2 is fixed in favour of group a through setting $n_{a1.2} = 8$ and $n_{b1.2} = 2$, only distributions with a high mean for group a and a low mean for group b in subtable 1 lead to a high power for both tests (a), and vice versa (c). When the observed odds ratio in subtable 2 is set to 1, the resulting power line is symmetric and only very extreme distributions for subtable one lead to a high power (b). Further, perhaps the most important observation from this experiment: the simple S-value did not perform worse than the S-value found through optimization, leading to the conclusion that *the simple S-value for the CMH-test can be used without loss of power compared to the S-value found through optimization.*

This experiment was repeated with the simple S-value for $n = 100$, with a found best worst-case δ^* of 0.21. Results are summarized in Appendix Figure A.4. Observed patterns were similar, but for smaller differences between the group means a minimum power of 0.80 was now achieved.

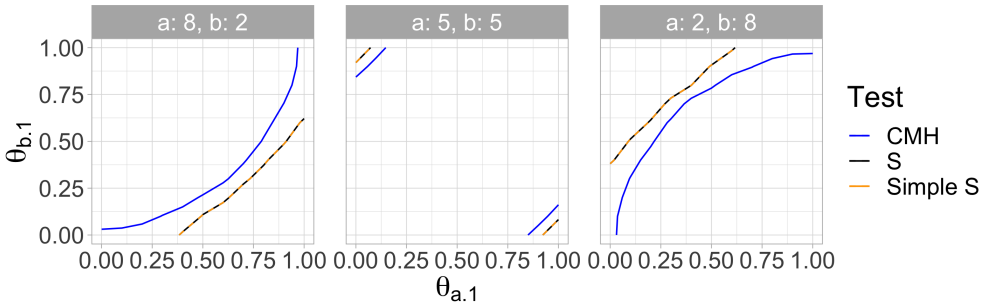


Figure 7.1: $(1 - \beta) = 0.8$ with $\delta^* = 0.31$ for the GROW S-value, the simple S-value and the CMH test for various fixed values of $n_{a1.2}$ and $n_{b1.2}$. $n = 40$, $\alpha = 0.05$. All subgroup sizes were equal to 10.

This experiment was conducted as well for the Breslow-Day version of the stratified 2x2 contingency table test. To perform the Breslow-Day test, the `BreslowDayTest` function from the `DescTools` package was used [Signorell, 2019]. Results should be interpreted with caution, as the classical version of the Breslow-Day test is only certainly valid when most of the expected cell counts are bigger than 5 [Breslow and Day, 1980]. Results for two different fixed values of $n_{a1.2}$ and $n_{b1.2}$ are depicted in Figure 7.2. It can be observed that results are now symmetric for an odds ratio fixed in favour of group a in subtable 2, or for an odds ratio equal to 1, in contrast

to the CMH test; this is as expected, as with the Breslow-Day test, only the *homogeneity* of the odds-ratios accross subtables is tested. Again, the simple S-value and the S-value found through optimization perform identically; however, the power of both safe tests is substantively lower than the power of the classical Breslow-Day test.

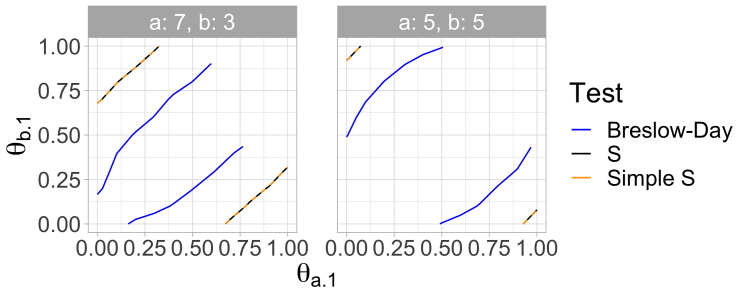


Figure 7.2: $(1 - \beta) = 0.8$ with $\delta^* = 0.31$ for the GROW S-value, the simple S-value and the Breslow-Day test for different fixed values of $n_{a1.2}$ and $n_{b1.2}$. $n = 40$, $\alpha = 0.05$, and all subgroup sizes were equal to 10.

Conclusion

In summary, default S-values for two stratified versions of the 2x2 contingency table test were developed: for the CMH test and for the Breslow-Day test. It was concluded that, like with the S-value for 2x2 contingency tables, a *simple* S-value could be defined in the case of equal subgroups within the subtables. For unequal subgroups, the JIPr still has to be found through optimization with BFGS. The designed S-values were compared to their classical frequentist counterparts with respect to power: again, the power of the safe tests was slightly lower. However, the fact that the simple S-value for the CMH and Breslow-Day tests could be used for optional stopping, and the S-values in general could be used for optional continuation do provide important advantages compared to the classical versions.

Chapter 8

Conclusion and future work

In this thesis two first versions of safe equivalents for the 2x2 contingency table test were presented. For the first version, the GROW S-value was found for a restricted alternative hypothesis where a minimal *absolute* difference between θ_a and θ_b existed. It was discovered that numerical optimization for finding the JIPr between P_{W_1} and P_{W_0} for defining the GROW S-value can be omitted. One could in this case use the proposed *simple* S-value if one is willing to accept that this simple S-value is not GROW for a very small subset of the restricted alternative hypothesis. It was numerically illustrated that this subset disappears as the sample size n grows: for large n, the simple S-value *is* GROW. For the case where $n_a = n_b$, formal proofs that the simple S-value *is* an S-value were given; for the generalized simple S-value, for $n_a \neq n_b$, a numerical proof was given. Further, numerical experiments aimed at investigating the power of the simple S-value revealed no difference in the power of the simple test and its equivalent found through numerical optimization. The simple S-value was also extended to stratified versions of the 2x2 contingency table test: the CMH test and the Breslow-Day test. For these extensions, it was also numerically illustrated that the simple S-test provides the (GROW) S-value for (a subset of) the restricted alternative hypothesis.

Further, the simple S-value provides a type-I error guarantee under the frequentist interpretation of *optional stopping*. As the construction of the (generalized) simple S-value does not depend on the sample size but only on the ratio $\frac{n_a}{n_b}$, this S-value can be used for testing any initial segment $x_1, x_2, \ldots, x_\tau$ of $x_1, \ldots, x_T$ with $\tau \leq T$ the simple S-value is well-defined for. Hence, the simple S-value is a nonnegative supermartingale, and it was concluded that the simple S-value could be used for optional stopping. Numerical experiments revealed that when using the simple S-value for optional stopping, the expected number of samples needed for gaining a certain power could even be lower than with Fisher's exact test.

For the second version of the safe 2x2 test, expected worst-case growth under the alternative hypothesis was minimized for the subset of distributions of the alternative hypothesis that all achieve a certain expected capital growth (i.e. minimized KL divergence with respect to $\mathcal{H}_0$), would we believe the data came from that distribution. A 'simple' (or analytical) version of this S-value could not be obtained for this divergence measure; for every sample size and value of the divergence measure δ, the JIPr has to be found through optimization for constructing the KL-based S-value, leading to increased computation times compared to the simple S-value. In addition, this version could not be used for optional stopping. However, using the KL-based S-value could be preferable in some cases: it was shown that it yields higher power than the simple S-value when data are generated from distributions at the edges of the parameter space ($[0, 1]^2$), i.e. when either θ_a or θ_b are close to 0 or 1. In some real-world examples with observed data with such estimated group means, the KL-based S-value mimicked results from Fisher's exact test more closely than

the simple S-value.

To conclude this thesis, some open problems resulting from the work presented in this manuscript are described. These problems are currently being worked on; results will probably be published in the form of a paper during the coming months.

8.1 Reparameterization of the safe test for contingency tables

In the safe tests for 2x2 contingency tables in this thesis, the likelihood under the alternative hypothesis was parameterized with two parameters: θ_a, representing the probability of observing $y = 1$ for an observation in group a, and θ_b. Prior distributions on (θ_a, θ_b) providing the GROW S-value were searched for numerically in this thesis. However, the likelihood under the alternative hypothesis could also be reparameterized in such a way that only one parameter represents the difference between θ_a and θ_b. For example, we could define a new parameter, a *grand mean* θ_g, as:

$$\theta_g = \frac{\theta_a + \theta_b}{2}. \tag{8.1.1}$$

θ_a and θ_b can then be expressed as a function of this grand mean and a parameter representing the difference between the group means, ξ:

$$\theta_a = \theta_g + \frac{\xi}{2} \quad ; \quad \theta_b = \theta_g - \frac{\xi}{2}. \tag{8.1.2}$$

Would we then express this grand mean and the adjustment ξ on the *log-odds* scale, the domain of ξ would then range from $-\infty$ to ∞, possibly enabling the use of an equivalent of the *Bayesian T-test* for testing the null hypothesis $\xi = 0$ and thus $\theta_a = \theta_b$ [GHK]. An uninformative prior could then be put on θ_g; it would have to be experimented with which priors on θ_g and ξ give rise to a safe test that is GROW. Possibly, the generalized simple S-value could be expressed in this reparameterization as well, allowing for an analytical determination of the adjustment parameter ϕ.

8.2 Uniformly most powerful safe tests

As seen in Chapter 3 of this thesis, the choice of the distance measure δ to design the S-value with can heavily influence the behaviour of the safe test with respect to expected capital growth, and the obtained power under a certain sample size. It would be desirable to standardize this choice of δ in some way.

Almost a century ago, Neyman and Pearson first proposed the idea of a *uniformly most powerful test* (UMPT) [Neyman and Pearson, 1933]. These UMPTs can be defined as "statistical hypothesis tests that provide the greatest power among all tests of a given size" [Johnson, 2013]. Johnson [2013] already proposed to use this idea for developing a new form of objective Bayesian tests, where the alternative hypothesis is defined such that the probability that the Bayes factor exceeds a specific threshold is maximized. We now propose that this idea could be extended to safe testing as well.

As was already observed in Chapters 3 and 4, certain values of δ for designing the S-value appeared to yield optimal power for all distributions in Θ_1 in numerical experiments with a fixed

sample size. In future work, it will be attempted to proof that an UMP δ exists and to analytically find it. To do so, we need to prove that:

$$\forall \theta_1, \theta_1' \in \Theta_1 : \arg\max_{\delta \in [0,1]} P_{\theta_1}\left(S^*_\delta(Y^n) \geq \frac{1}{\alpha}\right) = \arg\max_{\delta' \in [0,1]} P_{\theta_1'}\left(S^*_{\delta'}(Y^n) \geq \frac{1}{\alpha}\right), \tag{8.2.1}$$

and we aim to find an analytic expression for the arg max. To this end, the S-value in its *reparameterized* form as described in the previous paragraph might be used.

8.3 When to *stop* experiments

In the introduction of this thesis, it was emphasized that one reason why S-values are particularly useful is that they provide a type-I error guarantee under *optional continuation*. Hence, when one personally believes that $\mathcal{H}_1$ is true, but can not reject $\mathcal{H}_0$ based on a first experiment, one could start a new experiment, and multiply the new resulting S-value with the previous one. But, when this resulting 'combined' S-value is again not significant, should experiments be stopped? Or should one continue carrying out new experiments and tests until the null hypothesis is rejected? *When collected data contain a lot of evidence in favour of $\mathcal{H}_0$, it would be more efficient to stop designing new experiments and thus 'accept' $\mathcal{H}_0$ (and 'reject' $\mathcal{H}_1$) "for ever"*. Throughout this subsection, it will be hypothesized how such *evidence in favour of $\mathcal{H}_0$* could be recognized within the safe testing framework.

Sir Ronald Fisher, in the midst of developing the paradigm-shifting theory of p-value testing, already empathised that *the null hypothesis can not be proved through p-value testing*:

> "*[...] it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation.*" - Fisher [1935a]

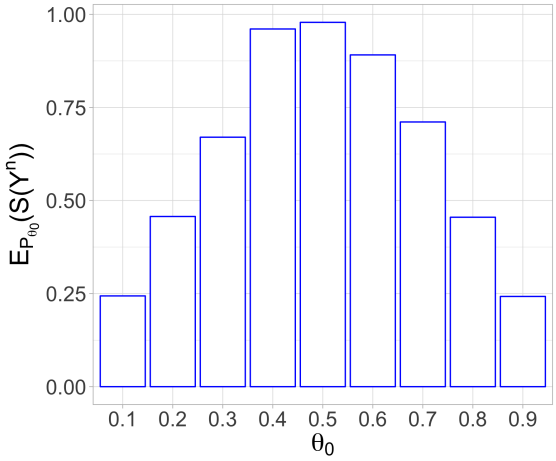
This is because p-value tests have been designed to give insight in the compatibility of the data with the *null* hypothesis. In the Neyman-Pearsonian view of testing (and the safe testing framework), they enable a type-I error guarantee: we can define the chance of rejecting $\mathcal{H}_0$ while it is in fact true. However, no guarantees for the opposite case are given: when a p-value based hypothesis test is not significant, one can not know whether this happened because of futility of the sample, or because $\mathcal{H}_0$ is true.

S-values are designed to have an expected value of at most 1 under the null hypothesis, but this property does not give us any information about the actual *values* that are generated under the distributions from the null hypothesis. Their expected value could take on any value in $[0, 1]$, and nothing is yet known about the *variance* of the S-values under the null hypothesis. Would we know for a specific safe test that all S-values calculated with data that are generated under the null are very close to 0, we could at least decide to *continue testing if we would observe S-values exceeding 1, as this would then be a very unlikely observation under the null.*

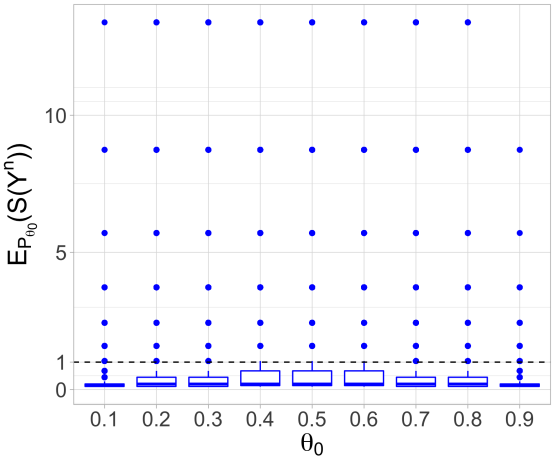
However, numerical experiments point towards infeasibility of this approach. An example of estimated expected values and variance of the simple S-value for 2x2 contingency tables under various Bernoulli distributions is summarized in Figure 8.1. It can be observed that the expected value of the S-value under $\mathcal{H}_0$ is close to 1 for some of the distributions, and under all distributions, S-values close to or bigger than 1 were observed in a simulation experiment. Hence, for 2x2 contingency table tests, the S-values appear to take on values in a broad range under the null

Figure 8.1: Estimated expected value of S under various distributions from $\mathcal{H}_0$ (a) and a boxplot illustrating the spread of of the observed S-values ≤ 20 (b). The simple S-value was used here with $\delta = 0.21$; $n = 100$. $10,000$ samples were generated per distribution P_{θ_0}: $\theta_0 \in \{0.1, 0.2, \ldots, 0.9\}$.

(a) The expected value of the S-value under $\mathcal{H}_0$



(b) Boxplot highlighting the spread of the S-value under $\mathcal{H}_0$



hypothesis, and a clear-cut distinction between data generated from the null or alternative outside of the rejection range does not appear to exist.

Another approach could be to always design a *pair* of safe tests: S_{10}, representing the evidence in favour of $\mathcal{H}_1$ against $\mathcal{H}_0$, and S_{01}, representing the evidence in favour of $\mathcal{H}_0$ against $\mathcal{H}_1$. Note that, unlike with Bayes factors, we can not use S_{10}^{-1} as a representation of the evidence for $\mathcal{H}_0$, as it is not guaranteed that this is a safe test, and this test is not necessarily GROW, as it has been optimized for worst case expected capital growth under $\mathcal{H}_1$. The GROW S_{01} would be defined as:

$$S_{01}^*(y^n) = \frac{P_{W_0^*}(y^n)}{P_{W_1^*}(y^n)} \quad \text{with} \quad (W_0^*, W_1^*) = \arg\min_{W_0} \min_{W_1} D(P_{W_0} || P_{W_1}). \tag{8.3.1}$$

One could then decide to stop planning subsequent experiments when S_{01} is sufficiently high (for example $\geq \frac{1}{\alpha_0}$): one could then decide to 'reject' the alternative hypothesis $\mathcal{H}_1$. The probability of rejecting $\mathcal{H}_1$ while it is in fact true would then become the second 'type I error', representing the probability of aborting experiments while in fact a futile sample has been collected.

Bibliography

A. Agresti. *An Introduction to Categorical Data Analysis*. Wiley, 2002.

V. Amrhein, S. Greenland, and B. McShane. Scientists rise up against statistical significance. *Nature*, 567:305–307, 2019.

J.O. Berger. Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, 18: 1–32, 2003.

J.O. Berger, J.M. Bernardo, and D. Sun. Overall objective priors. *Bayesian Analysis*, 10(1): 189–221, 2015.

J. Frédéric Bonnans, J. Charles Gilbert, C. Lemaréchal, and C.A. Sagastizábal. *Numerical optimization: theoretical and practical aspects*. Springer Berlin Heidelberg, 2006.

N.E. Breslow. Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association*, 91(433):14–28, 1996.

N.E. Breslow and N.E. Day. Statistical methods in cancer research. Volume I - the analysis of case-control studies. *IARC Sci. Publ.*, 32:5 – 338, 1980.

C. Carathéodory. Über den variabilitätsbereich der koeffizienten von potenzreihen, die gegebene werte nicht annehmen. *Mathematische Annalen*, 64(1):95–115, 1907.

Iain Chalmers and Paul Glasziou. Avoidable waste in the production and reporting of research evidence. *The Lancet*, 374(9683):86–89, 2009.

W.G. Cochran. Some methods for strengthening the common χ^2 tests. *Biometrics*, 10(4):417–451, 1954.

T.M. Cover and J.A. Thomas. *Elements of information theory. 2nd ed.* Wiley, 1990.

W.E. Deming. *Out of the Crisis*. MIT Press, 1982.

R.A. Fisher. *The design of experiments*. Oliver & Boyd, Oxford, England, 1935a.

R.A. Fisher. The logic of inductive inference. *Journal of the Royal Statistical Society Series A*, 98: 39–54, 1935b.

R.A. Fisher. Presidential address. *Sankhya: The Indian Journal of Statistics*, pages 14–17, 1938.

A. Fletcher-Sandersjöö, E.P. Thelin, J. Bartek, Jr., A. Elmi-Terander, M. Broman, and B. Bellander. Management of intracranial hemorrhage in adult patients on extracorporeal membrane oxygenation (ecmo): an observational cohort study. *PLOS ONE*, 12(12):1–17, 2017.

P.D. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

P.D. Grünwald, R. de Heide, and W.M. Koolen. Safe testing. 2019. eprint arXiv:1906.07801.

E. Gunel and J. Dickey. Bayes factors for independence in contingency tables. *Biometrika*, 61(3): 545–557, 1974.

A. Hendriksen, R. de Heide, and P. Grünwald. Optional stopping with Bayes Factors: a categorization and extension of folklore results, with an application to invariant situations. 2018. eprint arXiv:1807.09077v1.

A. Hofman and D.E. Grobbee. *Klinische epidemiologie. Tweede druk.* Reed Business, 2003.

J. Hraba and G. Grant. Black is beautiful: a reexamination of racial preference and identification. *Journal of Personality and Social Psychology*, 16(3):398–402, 1970.

T. Jamil, A. Ly, R.D. Morey, J. Love, M. Marsman, and E.J. Wagenmakers. Default gunel and dickey bayes factors for contingency tables. *Behaviour research methods*, 49:638–652, 2017.

L.K. John, G. Loewenstein, and D. Prelec. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, 23:524–32, 2012.

S.G. Johnson. *The NLopt nonlinear-optimization package*, 2019. URL `http://ab-initio.mit.edu/nlopt`.

V.E. Johnson. Uniformly most powerful Bayesian tests. *The Annals of Statistics*, 41(1):1716 – 1741, 2013.

R.E. Kass and A.E. Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90 (430):773–795, 1995.

J.A. Kline, M.A. Fisher, K.L. Pettit, C.T. Linville, and A.M. Beck. Controlled clinical trial of canine therapy versus usual care to reduce patient anxiety in the emergency department. *PLOS ONE*, 14(1):e0209232, 2019.

N. Mantel and W. Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4):719–48, 1959.

D.N. McCloskey and S.T. Ziliak. The standard error of regressions. *Journal of Economic Literature*, 34:97–114, 1996.

J. Neyman and E.S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289 – 337, 1933.

J.M. Papakonstantinou. *Historical Development of the BFGS Secant Method and Its Characterization Properties*. BiblioBazaar, 2011.

S.M. Ross. *Introduction to probability models*. Elsevier, 2010.

P. Schober, S.M. Bossers, P. Dong, C. Boer, and L.A. Schwarte. What do anesthesiologists know about p values, confidence intervals, and correlations: a pilot survey. *Anesthesiology Research and Practice*, 4201289, 2017.

G. Shafer, A. Shen, N. Vereshchagin, and V. Vovk. Test martingales, Bayes factors and p-values. *Statistical Science*, pages 84–101, 2011.

A. Signorell. *DescTools: Tools for Descriptive Statistics*, 2019. URL `https://cran.r-project.org/package=DescTools`. R package version 0.99.28.

J.A.C. Sterne and G.D. Smith. Sifting the evidence—what's wrong with significance tests? *BMJ*, 322:226–231, 2001.

J. ter Schure and P.D. Grünwald. Accumulation bias in meta-analysis: the need to consider time in error control. *arXiv e-prints*, May 2019.

R.J. Turner. *Endothelial pathology in preeclampsia.* PhD thesis, Leiden University, 2018. Introduction, pages 9 - 41.

S.L. van der Pas. *Much ado about the p-value.* Bachelor thesis, Leiden University, 2010.

T. van Erven and P. Harremoës. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

J. Venn and F. Galton. *Cambridge Anthropometry.* Journal of the Anthropological Institute of Great Britain and Ireland. Harrison and Sons, Printers, 1888.

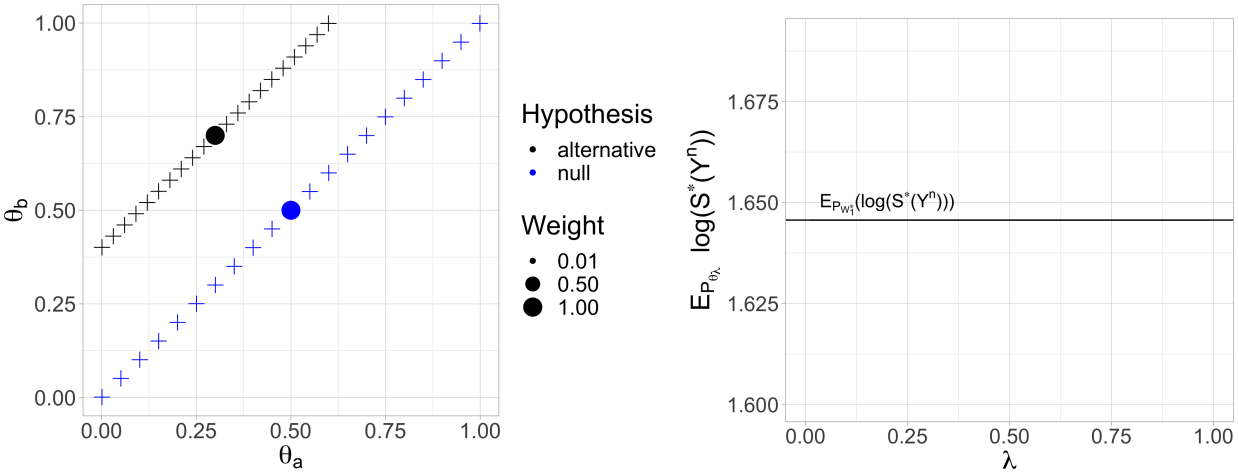
C. Woolston. Psychology journal bans p values. *Nature*, 519:9–9, 2015. doi: 10.1038/519009f.

H.R. Wulff, B. Andersen, P. Brandenhoff, and F. Guttler. What do doctors know about statistics? *Statistics in Medicine*, 6:3–10, 1987.

Appendix A

Figures, Tables and derivations

A.1 Chapter 3



(a) JIPr found with BFGS, $n = 20$, $\delta = 0.6$

(b) Expected capital growth for the simple S-value

Figure A.1: The JIPr found through BFGS for the one-sided version of the 2x2 tables test, and expected capital growth under the simple S-value. (a): crosses indicate positions of grid points in Θ_1 and Θ_0, dots and their size represent the weight from W_0^* and W_1^* for each point. (b): expected capital growth under various distributions from the restricted alternative hypothesis. Various $\theta_1 = (\theta_a$ and $\theta_b)$ were constructed as: $\theta_a = 0\lambda + 0.6(1-\lambda)$, and $\theta_b = 0.4\lambda + 1(1-\lambda)$ for generating data under various P_{θ_1}. Note that the expected capital growth is similar for all distributions, and equal to the expected capital growth under $P_{W_1^*}$.

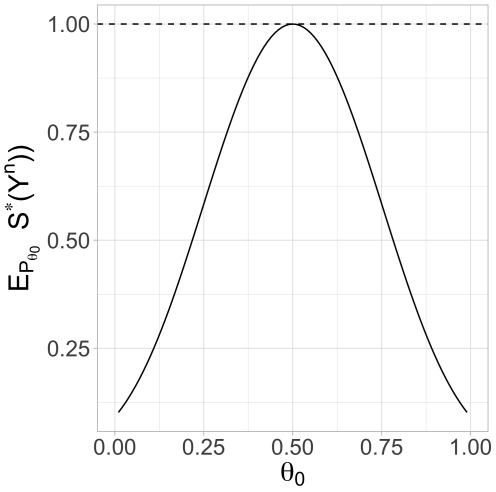


Figure A.2: Expected value of the generalized simple S under the null hypothesis for $n_a = 10$, $n_b = 20$, $\delta = 0.4$.

Table A.1: $\min \delta' - \delta$ for achieving worst case capital growth with the generalized simple S-value and an angled $\Theta_1'(\delta)$. $\frac{n_a}{n_b} = 0.5$ for all experiments.

δ \ n	30	48	66	84	102	120
0.1	0.09038	0.06121	0.04553	0.03556	0.02863	0.02353
0.2	0.04350	0.02282	0.01320	0.00797	0.00493	0.00311
0.3	0.01756	0.00602	0.00224	0.00087	0.00035	0.00015
0.4	0.00534	0.00096	0.00019	0.00004	0.00001	0.00001
0.5	0.00105	0.00008	0.00001	0.00001	0.00001	0.00001
0.6	0.00012	0.00002	0.00002	0.00002	0.00002	0.00002
0.7	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001
0.8	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001
0.9	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001

A.2 Chapter 4

Table A.2: KL-based divergence measure: δ^* achieving the maximum power in the worst case for various δ°. $n_a = n_b = 10$.

δ°	δ^*	$\min(1-\beta)$
log(10)	log(10)	0.184
log(16)	log(10)	0.314
log(20)	log(10)	0.344
log(50)	log(10)	0.527
log(100)	log(10)	0.632
log(200)	log(10)	0.734
log(400)	log(10)	0.800

Table A.3: KL-based divergence measure: δ^* achieving the maximum power in the worst case for various δ°. $n_a = n_b = 50$.

δ°	δ^*	$\min(1-\beta)$
log(16)	log(16)	0.339
log(40)	log(16)	0.483
log(100)	log(16)	0.613
log(200)	log(16)	0.704
log(400)	log(16)	0.762
log(1000)	log(16)	0.841

A.3 Chapter 7

The odds ratio in a partial table is equal to one only if the estimated group means are equal:

$$\begin{aligned} \frac{\hat{\theta}_{a.i}(1-\hat{\theta}_{b.i})}{(1-\hat{\theta}_{a.i})\hat{\theta}_{b.i}} &= 1 \\ \hat{\theta}_{a.i}(1-\hat{\theta}_{b.i}) &= (1-\hat{\theta}_{a.i})\hat{\theta}_{b.i} \\ \hat{\theta}_{a.i} - \hat{\theta}_{a.i}\hat{\theta}_{b.i} &= \hat{\theta}_{b.i} - \hat{\theta}_{a.i}\hat{\theta}_{b.i} \\ \hat{\theta}_{a.i} &= \hat{\theta}_{b.i}. \end{aligned} \tag{A.3.1}$$

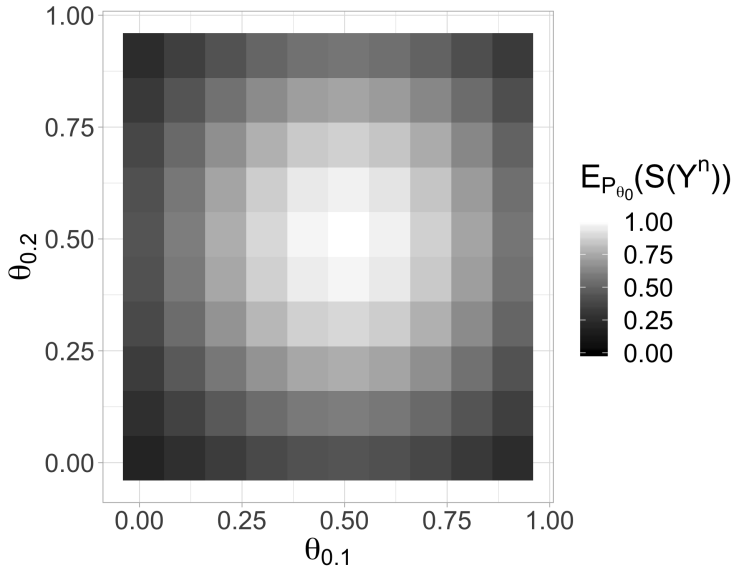


Figure A.3: Expected value of the simple CMH S-value under the null hypothesis for subgroup sizes equal to 5, $\delta = 0.4$, $k = 2$. Note that the expected value does not exceed 1.

Table A.4: CMH with $k = 3$: minimal adjustment δ' of the restricted alternative hypothesis for the simple S-value to be GROW. $n_{a.i} = n_{b.i} = \frac{n}{6}$

n \ δ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
24	0.10000	0.10000	0.02000	0.01000	0.00100	0.00010	0.00010	0.00010	0.00010
30	0.10000	0.04000	0.02000	0.00400	0.00040	0.00010	0.00010	0.00010	0.00010
36	0.10000	0.04000	0.01000	0.00200	0.00020	0.00010	0.00010	0.00010	0.00010
42	0.10000	0.04000	0.01000	0.00100	0.00010	0.00010	0.00010	0.00010	0.00000
48	0.10000	0.02000	0.00400	0.00040	0.00010	0.00010	0.00010	0.00010	0.00000

Table A.5: Minimal adjustment δ' of the angle of the restricted alternative hypothesis for the simple S-value to be GROW for the CMH test with $n_{a.i} = n_{b.i}$, but $n_{.1} = n_{.2} + 4$.

n \ δ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
16	0.20000	0.10000	0.04000	0.02000	0.01000	0.00100	0.00020	0.00010	0.00010
24	0.10000	0.10000	0.02000	0.01000	0.00200	0.00020	0.00010	0.00010	0.00010
32	0.10000	0.04000	0.01000	0.00400	0.00040	0.00010	0.00010	0.00010	0.00010
40	0.10000	0.04000	0.01000	0.00100	0.00010	0.00010	0.00010	0.00010	0.00010
48	0.10000	0.02000	0.00400	0.00040	0.00010	0.00010	0.00010	0.00010	0.00010
56	0.10000	0.02000	0.00400	0.00020	0.00010	0.00010	0.00010	0.00010	0.00010

Table A.6: Found weights for W_1^* for the JIPr for the CMH version of the stratified 2x2 contingency table test, with $k = 2$, $n_{a.1} = 10$, $n_{b.1} = 5$, $n_{a.2} = 11$, $n_{b.2} = 8$ and $\delta = 0.4$. Highlighted rows indicate found support points in $\Theta_1(\delta)$.

$\theta_{a.1}$	$\theta_{b.1}$	$\theta_{a.2}$	$\theta_{b.2}$	w_1
0.0	0.4	0.0	0.4	0.0000
0.0	0.4	0.3	0.7	0.0000
0.0	0.4	0.6	1.0	0.0000
0.3	0.7	0.0	0.4	0.0000
0.3	0.7	0.3	0.7	0.4641
0.3	0.7	0.6	1.0	0.0018
0.6	1.0	0.0	0.4	0.0000
0.6	1.0	0.3	0.7	0.0338
0.6	1.0	0.6	1.0	0.0002
0.4	0.0	0.4	0.0	0.0002
0.4	0.0	0.7	0.3	0.0338
0.4	0.0	1.0	0.6	0.0000
0.7	0.3	0.4	0.0	0.0018
0.7	0.3	0.7	0.3	0.4641
0.7	0.3	1.0	0.6	0.0000
1.0	0.6	0.4	0.0	0.0000
1.0	0.6	0.7	0.3	0.0000
1.0	0.6	1.0	0.6	0.0000

Table A.7: δ^* values achieving the best worst-case growth for finding the overall best δ^* with respect to power for the simple safe CMH test. $\alpha = 0.05$, $k = 2$ and $n_{a.i} = n_{b.i} = 10$. δ^* values on $\{0.2, 0.21, \ldots, 0.6\}$ were tried, with 1000 repetitions for each $\theta_1 \in \bar{\Theta}(\delta^\circ)$

δ°	δ^*	$\min(1-\beta)$
0.20	0.31	0.01
0.30	0.31	0.11
0.40	0.31	0.41
0.50	0.31	0.70
0.60	0.31	0.91
0.70	0.31	0.99
0.80	0.31	1.00

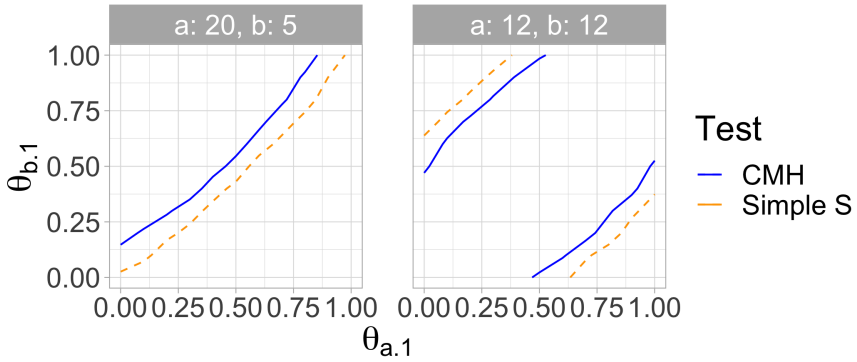


Figure A.4: $(1-\beta) = 0.8$ with $\delta^* = 0.21$ for the simple S-value and the CMH test for various fixed values of $n_{a1.2}$ and $n_{b1.2}$. $n = 100$, $\alpha = 0.05$

Appendix B

R package: preliminary manual

In this manual of the preliminary version of the `safe2x2` package, all functions included in the package are described and examples of their use can be found. The functions in this package can be used by downloading the `safe2x2` package and placing the folder in RStudio's active working directory. Make sure the `nloptr` library is installed. Different functions can then be called by loading them into the working environment through `source('safe2x2/... .R')`, as illustrated throughout the examples in this manual. A set of *default* tables is included as well, with pre-calculated results from the presented algorithms. A list of available tables and instructions to load them are included at the end of this manual.

Currently, the final version of this package is still under development: a similar package for the safe version of the t-test is being developed at the CWI, and to ensure the content and functionality of the two packages align, the final functions in the `safe2x2` package may still undergo some slight moderations.

B.0 General functions

`jipr`

Description
Perform optimization to find the joint information projection (JIPr) between the set of all Bayes marginals on the null hypothesis and on a subset of the alternative hypothesis defined by some divergence measure for the 2x2 contingency table test. The Bayes marginals providing the JIPr are used to compose the GROW S-value, which can subsequently be used to perform a safe test.

Usage
```
jipr(na, nb, delta, KLbased, H1prec, H0prec, providingOwnH1, ownH1, checkGROW)
```

Arguments

na	sample size of group a
nb	sample size of group b
delta	minimal divergence between the null and alternative hypothesis. For the KL divergence measure, the *standardized* KL divergence should be provided. Default FALSE.
KLbased	indication whether the default or KL-based S-value should be retrieved. Default FALSE.
H1prec	grid precision for discretizing the alternative hypothesis. Default 20.
H0prec	grid precision for discretizing the null hypothesis. Default 20.
providingOwnH1	boolean indicating whether the user is providing their own grid representing the alternative hypothesis. Default FALSE.
ownH1	alternative hypothesis grid provided by the user. Default NA.
checkGROW	after the JIPr has been found, it will numerically be checked if the obtained minimum indeed is the worst-case expected capital growth for a grid of distributions from the alternative hypothesis. Default FALSE.

Details

Through optimization with the BFGS algorithm, the JIPr between the set of all Bayes marginals on grids representing the null and alternative hypothesis is found. This process was described in detail in Section 2 of this thesis. The definition of the standardized KL divergence is discussed in Section 4. It is recommended to increase the grid precision in both hypotheses to at least $n+1$. If a lower precision is used to decrease computation time, it is recommended to set the 'checkGROW' argument to TRUE to ensure enough grid points were included to provide a guarantee of the worst-case expected capital growth.

Value

`$na`	sample size of group a
`$nb`	sample size of group b
`$H0set`	the parameters in the null hypothesis grid
`$H1set`	the parameters in the alternative hypothesis grid
`$w0`	the JIPr weights for the null hypothesis
`$w1`	the JIPr weights for the alternative hypothesis
`$message`	If `checkGROW` was set to `T`, the result of the check
`$value`	The KL divergence of the obtained JIPr

Example

```
source('safe2x2/jipr.R')
jipr(na = 10,
     nb = 10,
     KLbased=T,
     H0prec = 21,
     H1prec= 21,
     delta=log(20)/20,
     checkGROW = TRUE)
```

jiprCMH

Description
Provides the safe test null and alternative hypothesis grids and corresponding weights for the CMH or Breslow-Day test. In case of equal group sizes within each subtable, the simple S-value is returned. Otherwise, the GROW S-value is found through optimization.

Usage
```
jiprCMH(group.sizes, k, delta, Breslow.Day, H0.1d.precision, H1.1d.precision)
```

Arguments

group.sizes	vector of group sizes, provided as `c(na.1, nb.1, na.2, nb.2, ...)`
k	number of subtables
delta	minimal divergence between the null and alternative hypothesis
Breslow.Day	indication whether the Breslow-Day test should be used. Default FALSE.
H0.1d.precision	grid precision for discretizing the null hypothesis for one subtable. Default 3.
H1.1d.precision	grid precision for discretizing the alternative hypothesis for one subtable. Default 6.

Details
Grids representing the null- and alternative hypothesis are constructed as described in Chapter 7 of this thesis. Next, the Bayes marginals providing the JIPr are found as described in Chapter 2 of this thesis, or the simple S-value is returned.

Value

`$group.sizes`	vector of group sizes
`$H0set`	the parameters in the null hypothesis grid
`$H1set`	the parameters in the alternative hypothesis grid
`$w0`	the JIPr weights for the null hypothesis
`$w1`	the JIPr weights for the alternative hypothesis
`$min`	The KL divergence of the obtained JIPr

Example

```
source('safe2x2/jiprCMH.R')
jiprCMH(c(20,20,20,20),
        k=2,
        delta = 0.3)
```

S_test

Description
Perform a safe test for 2x2 contingency tables with a previously designed S-value.

Usage
```
S_test(na1, nb1, res, alpha)
```

Arguments

na1	number of ones observed in group a
nb1	number of ones observed in group b
res	result object from `jipr` or one of the other functions in the `safe2x2` package
alpha	significance level to test at. Default 0.05.

Details
Calculates the S-value. As a side effects, prints if the null hypothesis should be rejected at a certain significance level. Throws an exception if `na1` or `nb1` exceeds the group size.

Value
Returns the S-value.

Example

```
source('safe2x2/jipr.R')
source('safe2x2/S_test.R')
res1 <- jipr(na = 10,
     nb = 10,
     KLbased=T,
     H0prec = 21,
     H1prec= 21,
     delta=log(20)/20)
S_test(na1 = 1, nb1 = 8, res = res1, alpha = 0.1)
```

STestCMH

Description
Perform a safe test for the CMH or Breslow-Day test with a previously designed S-value.

Usage
```
S_test(observed.counts, res, alpha)
```

Arguments

observed.counts	vector of observed counts in each subgroup: `c(na1.1, nb1.1, na1.2, ... )`
res	result object from `jiprCMH` or one of the other functions in the `safe2x2` package
alpha	significance level to test at. Default 0.05.

Details
Calculates the S-value. As a side effect, prints if the null hypothesis should be rejected at a certain significance level.

Value
Returns the S-value.

Example

```
source('safe2x2/jiprCMH.R')
source('safe2x2/STestCMH.R')
res1 <- jiprCMH(c(5,10,5,10),
                k=2,
                delta = 0.3)
STestCMH(c(1, 9, 2, 8), res = res1, alpha = 0.05)
```

B.1 Scenario 1: determining the minimal δ to achieve a certain minimal expected capital growth

FindDeltaForLDefault

Description
For the simple S-value, the smallest δ^* such that the S-value designed with $\Theta_1(\delta^*)$ minimally achieves expected growth L for a given sample size n is found.

Usage
```
FindDeltaForLDefault(na, nb, L, precision, deltas)
```

Arguments

na	sample size of group a
nb	sample size of group b
L	minimal expected capital growth
precision	precision of found delta (≥ 0.0001). Default 0.01.
deltas	vector of delta values to try. Default `seq(0.01, 0.99, 0.1)`.

Details
For a given n and L, it is aimed to find the smallest δ^* that achieves:

$$\min_{P_{\theta_1}:\theta_1 \in \Theta_1(\delta^*)} \mathbb{E}_{P_{\theta_1}} [\log(S_{\delta^*}(Y^n))] \geq L,$$

as described in Section 3 of this thesis. The function calls itself recursively until the specified precision is reached. As a side effect, each δ^* being tried is printed.

Value

`$na`	sample size of group a
`$nb`	sample size of group b
`$H0set`	the parameters in the null hypothesis grid
`$H1set`	the parameters in the alternative hypothesis grid
`$w0`	the weights for the null hypothesis
`$w1`	the weights for the alternative hypothesis
`$delta.star`	the smallest delta achieving the minimal expected growth
`$phi`	the rotation adjustment of the simple s value for $n_a \neq n_b$
`$min.growth`	the expected capital growth of the S-value this `$n.star` and `$delta`

Example

```
source('safe2x2/FindDeltaForLDefault.R')
FindDeltaForLDefault(na = 100, nb = 100, L = log(20), precision = 0.001)
```

FindDeltaForLKL

Description
For the KL-based S-value, the smallest δ^* such that the S-value designed with $\Theta_1(\delta^*)$ minimally achieves expected growth L for a given sample size n is found.

Usage
```
FindDeltaForLKL(na, nb, L, H0prec, H1prec, own.delta.grid, precision.delta.grid)
```

Arguments

na	sample size of group a
nb	sample size of group b
L	minimal expected capital growth
H0prec	the desired precision of the grid for finding the JIPr for the null hypothesis. Default NA.
H1prec	the desired precision of the grid for finding the JIPr for the null hypothesis. Default NA.
own.delta.grid	optionally, provide own (standardized) grid of divergence measures to try, sorted in ascending order. Default NA.
precision.delta.grid	optionally, choose the precision of the grid of deltas used. Default 50.

Details
See `FindDeltaForLDefault`. As the precision for the null and alternative hypothesis grids, the minimum of $2(n+1)$ and 100 is used to limit the algorithm's runtime. For large sample sizes, it is recommended to find the GROW S-value with `jipr.R` with the finally found δ^* from this algorithm with higher precision. As a side effect, prints the delta currently being tried.

Value

`$na`	sample size of group a
`$nb`	sample size of group b
`$H0set`	the parameters in the null hypothesis grid
`$H1set`	the parameters in the alternative hypothesis grid
`$w0`	the weights for the null hypothesis
`$w1`	the weights for the alternative hypothesis
`$delta.star`	the smallest (not standardized) delta achieving the minimal expected growth
`$min.growth`	the expected capital growth of the S-value

Example
```
source('safe2x2/FindDeltaForLKL.R')
FindDeltaForLKL(na = 10, nb = 10, L = log(20),
                own.delta.grid = log(c(10, 20, 50, 100, 200))/20)
```

FindDeltaForLCMH

Description
For the simple safe CMH test, the smallest δ^* such that the S-value designed with $\Theta_1(\delta^*)$ minimally

achieves expected growth L for a given sample size n is found.

Usage

```
FindDeltaForLCMH(group.sizes, k, L, precision, deltas, Breslow.Day)
```

Arguments

group.sizes	list of the group sizes, ordered as `na1, nb1, na2, nb2, ...`
k	the number of subtables
L	minimal expected capital growth
precision	precision of found delta (≥ 0.0001). Default 0.01.
deltas	vector of delta values to try. Default `seq(0.01, 0.99, 0.1)`.
Breslow.Day	logical indicating whether the Breslow-Day test should be used. Default False.

Details

See `FindDeltaForLDefault`.

Value

`$group.sizes`	list of the group sizes, ordered as `na1, nb1, na2, nb2, ...`
`$H0set`	the parameters in the null hypothesis grid
`$H1set`	the parameters in the alternative hypothesis grid
`$w0`	the weights for the null hypothesis
`$w1`	the weights for the alternative hypothesis
`$delta.star`	the smallest delta achieving the minimal expected growth
`$min.growth`	the expected capital growth of the S-value

Example

```
source('safe2x2/FindDeltaForLCMH.R')
FindDeltaForLCMH(group.sizes = c(10,10,5,5), k = 2,
                      L = log(20))
```

B.2 Scenario 2: determining the minimal n to achieve a certain minimal expected capital growth

FindNForLDefault

Description
The smallest sample size n that achieves the desired minimal expected capital growth L when designing the generalized simple S-value with $\Theta_1(\delta^*)$ in the default case is found. The minimal expected growth rate L is given, and the minimal 'relevant' substantive difference δ^* between the group means is given.

Usage
```
FindNForLDefault(L, delta.min, a.iter, b.iter, nmin, nmax)
```

Arguments

L	desired expected capital growth
delta.min	divergence measure identifying subset of the alternative hypothesis to find the S-value for
nmin	n to start searching at. Default 10.
nmax	n to stop searching at. Default 100.
a.iter	step size to take in group a while searching for the minimal n. Default 1.
b.iter	step size to take in group b while searching for the minimal n. Default 1.

Details
As discussed in Section 3 of this thesis, for a given δ^*, we seek the smallest sample size n that achieves:

$$\min_{P_{\theta_1}:\theta_1\in\Theta_1(\delta^*)} \mathbb{E}_{P_{\theta_1}}\left[\log(S_{\delta^*}(Y^n))\right] \geq L.$$

If one plans on collecting unequal group sizes, this can be accounted for by changing the values passed through `a.iter` and `b.iter`. As a side effect, each samle size n being tried is printed.

Value

`$n.star`	the minimal n achieving the desired expected growth at delta min
`$na`	group size in group a at `$n.star`
`$nb`	group size in group b at `$n.star`
`$H0set`	the parameters in the null hypothesis grid
`$H1set`	the parameters in the alternative hypothesis grid
`$w0`	the weights for the null hypothesis
`$w1`	the weights for the alternative hypothesis
`$delta.min`	the initially passed delta min
`$phi`	the rotation adjustment of the simple s value for the case of unequal group sizes

Example
```
source('safe2x2/FindNForLDefault.R')
FindNForLDefault(delta.min = 0.6,L = log(20), a.iter = 2, b.iter = 1)
```

FindNForLKL

Description
The smallest sample size n that achieves the desired minimal expected capital growth L when designing the S-value with $\Theta_1(\delta^*)$ is found. The minimum expected growth rate L is known, and the minimal KL-divergence measure δ^* is given.

Usage `FindNForLKL(delta.standard, L, a.iter, b.iter, nmax, nmin, H0prec, H1prec)`

Arguments

L	desired expected capital growth
delta.standard	the *standardized* KL-divergence identifying subset of the alternative hypothesis to find the GROW S-value for.
nmin	n to start searching at. Default 10.
nmax	n to stop searching at. Default 100.
a.iter	step size to take in group a while searching for the minimal n. Default 1.
b.iter	step size to take in group b while searching for the minimal n. Default 1.
H0prec	the desired precision of the grid for finding the JIPr for the null hypothesis. Default NA.
H1prec	the desired precision of the grid for finding the JIPr for the null hypothesis. Default NA.

Details
See `FindNForLDefault`. The standardized KL-divergence is defined in Section 4 of this thesis. As the RIPr grid on Θ_1 is only determined for one quadrant, the precision in the alternative hypothesis will only take on multiples of 4. Other precisions can be passed and will be rounded up to the closest multiple of 4. When no precision is indicated, during each iteration, a precision of $n+1$ will be used. To improve computation time, it is recommended to run this algorithm with fairly low precision, and to run `jipr` with the finally found n again with higher precision. As a side effect, the n being tried and achieved expected growth for each n are printed.

Value

`$n.star`	the minimal n achieving the desired expected growth at delta min
`$na`	group size in group a at `$n.star`
`$nb`	group size in group b at `$n.star`
`$H0set`	the grid of parameters in the null hypothesis for composing the S-value
`$H1set`	the grid of parameters in the alternative hypothesis for composing the S-value
`$w0`	the found weights for the JIPr on the null hypothesis
`$w1`	the found weights for the JIPr on the alternative hypothesis
`$delta`	delta.standard multiplied by the found `$n.star`
`$min.growth`	the expected capital growth of the S-value this `$n.star` and `$delta`

Example

```
source('safe2x2/FindNForLKL.R')
FindNForLKL(delta.standard = log(20)/20, L = log(20), a.iter = 1, b.iter = 1)
```

FindNForLCMH

Description
The smallest sample size n that achieves the desired expected capital growth L when designing the S-value with $\Theta_1(\delta^*)$ is found. The minimum expected growth rate L is known, and the minimal 'relevant' substantive difference δ^* between the group means in each subtable is known.

Usage `FindNForLCMH(delta.standard, L, k, nmax, group.iters, group.sizes.start, Breslow.Day)`
Arguments

L	desired expected capital growth
delta.min	divergence measure identifying subset of the alternative hypothesis to find the S-value for
k	number of subtables
nmax	n to stop searching at
group.iters	vector of increments for the group sizes during each iteration, provided as `c(na.1, nb.1, na.2, nb.2, ...)`
group.sizes.start	vector of group sizes at the start, provided as `c(na.1, nb.1, na.2, nb.2, ...)`
Breslow.Day	indicates whether the Breslow-Day version should be implemented; Default FALSE.

Details
See `FindNForLDefault`.

Value

`$n.star`	the minimal n achieving the desired expected growth at delta min
`$group.sizes`	list of groups and their sizes at`$n.star`
`$H0set`	
`$H1set`	
`$w0`	
`$w1`	
`$delta.min`	the initially passed delta min

Example

```
source('safe2x2/FindNForLCMH.R')
FindNForLCMH(delta.min = 0.5,
             L = log(10),
             group.iters = c(1,2,1,2),
             group.sizes.start = c(4,8,4,8),
             k = 2, Breslow.Day = T)
```

B.3 Scenario 3: determining the optimal δ^* for designing the S-value with a certain sample size n and significance level α

`FindDeltaUmpDefault`

Description

A numerical approximation of the δ^* leading to the best power in the worst case for *most or all* $\{\bar{\Theta}(\delta^\circ) : \delta^\circ \in [0,1]\}$ for a certain n_a and n_b is found for the simple S-value. Worst-case power is evaluated for a grid of $\delta^\circ \in [0,1]$, for a grid of candidate $\delta^* \in [0,1]$.

Usage

```
FindDeltaUmpDefault(na, nb, M, alpha, desired.precision,
        delta.mins, delta.stars, number.for.seed)
```

Arguments

na	sample size in group a
nb	sample size in group b
M	number of monte carlo samples per distribution. Default 100.
alpha	significance level. Default 0.05.
desired.precision	desired precision of the found delta.star. Default 0.1.
delta.mins	grid of deltas identifying data-generating distributions. Default `delta.mins = seq(0, 1, 0.1)`.
delta.stars	grid of possible delta stars to try. Default `seq(0.1, 0.8, by = 0.1)`.
number.for.seed	number to set seed with: if none given, set based on current time Default NA.

Details

M monte carlo simulations are carried out per distribution in a grid representing each alternative hypothesis subset defined with `delta.mins`. A specific seed for these simulations can be set through `number.for.seed`. As a side effect, prints progress while iterating over different grids to achieve the desired precision. Also plots for each δ° the achieved power in the worst case.

Value

`$na`	sample size of group a
`$nb`	sample size of group b
`$H0set`	the parameters in the null hypothesis grid
`$H1set`	the parameters in the alternative hypothesis grid
`$w0`	the weights for the null hypothesis
`$w1`	the weights for the alternative hypothesis
`$delta.star`	the found 'most powerful' delta.

Example

```
source('safe2x2/FindDeltaUmpDefault.R')
FindDeltaUmpDefault(na = 20, nb  =20, desired.precision = 0.01)
```

FindDeltaUmpKL

Description

The δ^* leading to the best power in the worst case for *most or all* $\{\bar{\Theta}(\delta^\circ) : \delta^\circ \in [0,1]\}$ for a certain n is found for the KL-based divergence measure. Worst-case power is evaluated for a grid of δ°, for a grid of candidate δ^*.

Usage

```
FindDeltaUmpKL(na, nb, M, alpha, delta.mins, delta.stars,
         H0prec, H1pred, number.for.seed)
```

Arguments

na	sample size in group a
nb	sample size in group b
alpha	significance level to test at. Default 0.05.
M	number of monte carlo samples per distribution. Default 100.
delta.mins	grid of deltas identifying data-generating distributions. Default `log(c(10, 20, 50, 100, 200, 400, 600))`.
delta.stars	grid of possible delta stars to try. Default `log(c(10, 20, 50, 100))`.
H0prec	the desired precision of the grid for finding the JIPr for the null hypothesis. Default NA.
H1prec	the desired precision of the grid for finding the JIPr for the null hypothesis. Default NA.
number.for.seed	number to set seed with: if none given, set based on current time. Default NA.

Details

See `FindDeltaUmpDefault`. The grid precision in the null and alternative hypothesis will be taken as $n+1$, if not provided. It is recommended to run the algorithm with a low grid precision, and to use `jipr.R` to design the GROW S-value with higher precision with the found δ^*.

Value

`$na`	sample size of group a
`$nb`	sample size of group b
`$H0set`	the parameters in the null hypothesis grid
`$H1set`	the parameters in the alternative hypothesis grid
`$w0`	the weights for the null hypothesis
`$w1`	the weights for the alternative hypothesis
`$delta.star`	the found 'most powerful' delta.

Example

```
source('safe2x2/FindDeltaUmpKL.R')
FindDeltaUmpKL(na = 10, nb = 10)
```

FindDeltaUmpCMH

Description

The δ^* leading to the best power in the worst case for *most or all* $\{\bar{\Theta}(\delta^\circ) : \delta^\circ \in [0,1]\}$ for a certain n is found for the default distance measure for the simple CMH test. Worst-case power is evaluated for a grid of $\delta^\circ \in [0,1]$, for a grid of candidate $\delta^* \in [0,1]$

Usage

```
FindDeltaUmpCMH(group.sizes, k, M, alpha, desired.precision,  delta.mins, delta.stars,
        number.for.seed, Breslow.Day)
```

Arguments

group.sizes	list of the group sizes, ordered as `na1, nb1, na2, nb2, ...`
k	number of subtables
M	number of monte carlo samples per distribution. Default 100.
alpha	significance level. Default 0.05.
desired.precision	desired precision of the found delta.star. Default 0.1.
delta.mins	grid of deltas identifying data-generating distributions. Default `delta.mins = seq(0, 1, 0.1)`.
delta.stars	grid of possible delta stars to try. Default `seq(0.1, 0.8, by = 0.1)`.
number.for.seed	number to set seed with: if none given, set based on current time Default NA.
Breslow.Day	indicates whether the Breslow-Day version should be implemented; Default FALSE.

Details

See `FindDeltaUmpDefault`. The grid representing the restricted alternative hypothesis to simulate the power for can become large with k; to decrease simulation time, when the grid exceeds size 100, a random sample is taken from the grid points.

Value

`$delta.star`	the found 'most powerful' delta.
`$group.sizes`	list of groups and their sizes
`$H0set`	the parameters in the null hypothesis grid
`$H1set`	the parameters in the alternative hypothesis grid
`$w0`	the weights for the null hypothesis
`$w1`	the weights for the alternative hypothesis

Example

```
source('safe2x2/FindDeltaUmpCMH.R')
FindDeltaUmpCMH(group.sizes = c(10, 10, 10, 10), k = 2)
```

B.4 Scenario 4: determining the minimal sample size for achieving a certain power for a subset of the alternative hypothesis $\Theta(\delta^\circ)$ while testing at significance level α

`FindNForPowerDefault`

Description
The desired type I-error α, the type II-error β and the subset of distributions $\{P_{\theta_1} : \theta_1 \in \Theta_1(\delta^\circ)\}$ for which the desired power $1-\beta$ should be achieved with the S-value are given. δ° is measured as the default divergence measure in this case. The smallest n that achieves this, and the corresponding δ^* and simple S-value are returned.

Usage

```
FindNForPowerDefault(delta.min, alpha, beta, M, nmin, nmax, a.iter, b.iter,
.         deltas,desired.precision.sample.size, number.for.seed)
```

Arguments

delta.min	divergence measure to indicate the subset of the alternative hypothesis to achieve the desired power for
alpha	significance level for testingDefault 0.05.
beta	desired type II error probability. Default 0.20.
M	number of simulations per distribution to approximate the power. Default 100.
nmin	n to start searching at. Default 20.
nmax	n to stop searching at. Default 200.
a.iter	step size taken in group a while searching for the minimal sample size. Default 5.
b.iter	step size taken in group b while searching for the minimal sample size. Default 5.
deltas	grid of possible delta stars to try. Default `seq(0.1, 0.8, 0.1)`.
desired.precision.sample.size	final desired precision to determine the sample size with. Default 1.
number.for.seed	number to set seed with: if none given, set based on current time Default NA.

Details The algorithm implemented is discussed in detail in Section 3 of this thesis. It is aimed to find the smallest n and smallest δ^* that achieve:

$$\min_{\theta \in \bar{\Theta}_1(\delta^\circ)} P_\theta \left(S_{\delta^*}(y^{n_{\min}}) \geq \alpha \right) \geq 1 - \beta.$$

As a side effect, the n being tried, and worst case power and the corresponding data generating distribution are printed.

Value

`$n.star`	the minimal n achieving the desired power
`$d.star`	the minimal δ^* achieving the desired power at n.star
`$na`	group size a at n.star
`$nb`	group size b at n.star
`$H0set`	the parameters in the null hypothesis grid
`$H1set`	the parameters in the alternative hypothesis grid
`$w0`	the weights for the null hypothesis
`$w1`	the weights for the alternative hypothesis

Example

```
source('safe2x2/FindNForPowerDefault.R')
FindNForPowerDefault(delta.min = 0.5)
```

FindNForPowerKL

Description

The desired type I-error α, the type II-error β and the subset of distributions $\{P_{\theta_1} : \theta_1 \in \Theta_1(\delta^\circ)\}$ for which the desired power $1-\beta$ should be achieved with the safe test are known. δ° is measured as the KL-based divergence measure in this case. The smallest n that achieves this in combination with the smallest δ^* and S-value result object to use for further testing are returned.

Usage

```
FindNForPowerKL(delta.min, alpha, beta, M, nmin, nmax, a.iter, b.iter,  delta.stars,
         desired.precision.sample.size, number.for.seed)
```

Arguments

delta.min	standard divergence measure to indicate the subset of the alternative hypothesis to achieve the desired power for
alpha	significance level for testing. Default 0.05.
beta	desired type II error probability. Default 0.20.
M	number of simulations per distribution to approximate the power. Default 100.
nmin	n to start searching at. Default 20.
nmax	n to stop searching at. Default 20.
a.iter	step size taken in group a while searching for the minimal sample size. Default 5.
b.iter	step size taken in group b while searching for the minimal sample size Default 5.
delta.stars	grid of possible delta stars to try. Default `log(c(10, 20, 50, 100))`.
desired.precision.sample.size	final desired precision to determine the sample size with. Default 1.
number.for.seed	number to set seed with: if none given, set based on current time. Default NA.

Details

See `FindNForPowerDefault`. Note that the maximum grid precision of the null and alternative

hypothesis is set to 100 in this algorithm. For high sample sizes, it is recommended to use `jipr.R` with the finally obtained n.star and delta.star with increased precision. As a side effect, prints the sample sizes being tried.

Value

`$n.star`	the minimal n achieving the desired power
`$delta.star`	the minimal δ^* achieving the desired power at n.star
`$na`	group size a at n.star
`$nb`	group size b at n.star
`$H0set`	the parameters in the null hypothesis grid
`$H1set`	the parameters in the alternative hypothesis grid
`$w0`	the weights for the null hypothesis
`$w1`	the weights for the alternative hypothesis

Example

```
source('safe2x2/FindNForPowerKL.R')
FindNForPowerKL(delta.min = log(50)/20)
```

FindNForPowerCMH

Description
The desired type I-error α, the type II-error β and the subset of distributions $\{P_{\theta_1} : \theta_1 \in \Theta_1(\delta^\circ)\}$ for which the desired power $1-\beta$ should be achieved with the safe test are given. δ° is measured as the default divergence measure in this case. The smallest n that achieves this, and the corresponding δ^* and simple S-value are returned.

Usage

```
FindNForPowerCMH(delta.min, group.sizes.start, group.iters, k, alpha, beta,
          M, deltas, Breslow.Day, number.for.seed)
```

Arguments

delta.min	divergence measure to indicate the subset of the alternative hypothesis to achieve the desired power for
group.iters	vector of increments for the group sizes during each iteration, provided as `c(na.1, nb.1, na.2, nb.2, ...)`
group.sizes.start	vector of group sizes at the start, provided as `c(na.1, nb.1, na.2, nb.2, ...)`
k	number of subtables
alpha	significance level for testing. Default 0.05.
beta	desired type II error probability. Default 0.20.
M	number of simulations per distribution to approximate the power. Default 100.
nmax	n to stop searching at. Default 100.
deltas	grid of possible delta stars to try. Default `seq(0.1, 0.8, 0.1)`.
Breslow.Day	indicates whether the Breslow-Day version should be implemented; Default FALSE.
number.for.seed	number to set seed with: if none given, set based on current time. Default NA.

Details

See `FindNForPowerDefault`. As a side effect, prints each n being tried.

Value

`$n.star`	the minimal n achieving the desired power
`$d.star`	the minimal δ^* achieving the desired power at n.star
`$group.sizes`	list of groups and their sizes at`$n.star`
`$H0set`	the parameters in the null hypothesis grid
`$H1set`	the parameters in the alternative hypothesis grid
`$w0`	the weights for the null hypothesis
`$w1`	the weights for the alternative hypothesis

Example

```
source('safe2x2/FindNForPowerCMH.R')
FindNForPowerCMH(delta.min = 0.5,
                 group.iters = c(1,1,1,1),
                 group.sizes.start = c(5,5,5,5),
                 k = 2)
```

B.5 List of default tables

Outcomes of several of the algorithms provided in this package were previously calculated and recorded for certain 'standard' cases. Results are presented in several data frames attached to this package; these are described in the table below.

Table name	Description
`DFdeltaForLKLnaIsnb.Rdata`	For the KL-based divergence measure, results for Situation 1 for several combinations of n and L are provided, for $n_a = n_b$.
`DFnForLKLnaIsnb.Rdata`	For the KL-based divergence measure, results for Situation 2 for several combinations of δ and L are provided, for $n_a = n_b$.
`DFDeltaUmpFornaIsnb.Rdata`	For the default divergence measure, for several n the 'most powerful' delta is provided, for $n_a = n_b$.
`DFDeltaUmpKLFornaIsnb.Rdata`	For the KL-based divergence measure, for several n the 'most powerful' delta is provided, for $n_a = n_b$.
`DFNForPowernaIsNbDefault.Rdata`	For the default divergence measure, results for Situation 4 for various $\delta°$, with $n_a = n_b$, $\alpha = 0.05$ and $\beta = 0.20$ are provided.
`DFNForPowernaIsNbKL.Rdata`	For the KL-based divergence measure, results for Situation 4 for various $\delta°$, with $n_a = n_b$, $\alpha = 0.05$ and $\beta = 0.20$ are provided.

These data frames can be loaded and viewed as follows:

```
get(load('safe2x2/data/*data_frame_name*'))
```