Safe Hypothesis Tests for the 2×2 Contingency Table

Reuben Adams (4939034)

Supervisors: Peter Grünwald, Joris Bierkens

A thesis submitted to the department of Statistics, TU Delft, in partial fulfillment of the requirements for the degree of MSc Applied Mathematics

Abstract

Safe hypothesis tests are tests that are robust under accumulation bias, namely when there are dependencies between the results of previous studies and the decision whether to conduct further studies. We construct two types of safe test for the 2×2 contingency table, the conditional and unconditional safe tests. In general safe tests are given by an information projection that may be difficult to compute. The conditional tests we construct however are given either in explicit form or implicitly via a defining equation. The same can be said of many of the unconditional tests we construct, for which we prove a number of theoretical results enabling their quick calculation when not given explicitly. The method we develop to accomplish this may perhaps be used to identify optimal safe tests in many other scenarios.

Acknowledgements

I would like to thank Peter Grünwald for introducing me to the fascinating and and highly relevant topic of safe hypothesis testing. His enthusiasm and unfailing positivity is highly infectious and I enjoyed our virtual meetings tremendously. Encountering open questions so early in the thesis made for an exciting journey and I hope I have managed to contribute something of use. I am very grateful to have had Peter Grünwald as a supervisor.

I am also immensely thankful to Joris Bierkens who extremely kindly offered to co-supervise at short notice. I am very grateful for his insightful suggestions.

Contents

Notation

$\mathrm{KL}(P\|\|Q)$	The Kullback–Leibler divergence between the distributions P and Q
$\mathrm{kl}(p\|\|q)$	The Kullback–Leibler divergence between two Bernoulli distributions with parameters p and q respectively
Θ	The parameter set
Θ_0, Θ_1	The null and alternative parameter sets respectively
$\mathcal{W}(A)$	The set of all distributions on a given set A
$\mathcal{E}(\Theta_0)$	The set of all S-values defined with respect to the null parameter set Θ_0

For 2×2 contingency tables in particular, we also use the following notation.

θ_a, θ_b	The probability of observing a 1 in group a or b respectively under the alternative hypothesis
θ_a^L, θ_a^U	The lowest and highest values of θ_a permitted under the prior knowledge
θ_b^L, θ_b^U	The lowest and highest values of θ_b permitted under the prior knowledge
PKR	$[\theta_a^L, \theta_a^U] \times [\theta_b^L, \theta_b^U]$, the prior knowledge rectangle
$\theta_0(p)$	(p, p), where $p \in [0, 1]$
Θ_0	$\{\theta_0(p) : p \in [0, 1]\}$
Θ_0'	$\Theta_0 \cap$ PKR, the restricted null parameter set for the given PKR
I^0_{PKR}	$[\max\{\theta_a^L, \theta_b^L\}, \min\{\theta_a^U, \theta_b^U\}]$, namely the $p \in [0, 1]$ such that $\theta_0(p) \in \Theta_0'$
Θ_1	$\{(\theta_a, \theta_b) \in [0, 1]^2 : \theta_b \geq \theta_a\}$
Θ_1'	$\Theta_1 \cap$ PKR, the restricted alternative parameter set for the given PKR
$\delta = \delta(\theta_a, \theta_b)$	$\theta_b - \theta_a$, the risk difference for the given (θ_a, θ_b)
$\lambda = \lambda(\theta_a, \theta_b)$	θ_b / θ_a, the relative risk for a given (θ_a, θ_b) (also called the risk ratio)
$\psi = \psi(\theta_a, \theta_b)$	$\frac{\theta_b}{1-\theta_b} \frac{1-\theta_a}{\theta_a}$, the odds ratio for a given (θ_a, θ_b) (for conditional S-values, we will occasionally specify values of ψ without reference to a specific (θ_a, θ_b))
$\Theta_1(\delta)$	$\{(\theta_a, \theta_b) \in \Theta_1 : \delta(\theta_a, \theta_b) \geq \delta\}$ for the given risk difference threshold δ
$\Theta_1(\lambda)$	$\{(\theta_a, \theta_b) \in \Theta_1 : \lambda(\theta_a, \theta_b) \geq \lambda\}$ for the given relative risk threshold λ
$\Theta_1(\psi)$	$\{(\theta_a, \theta_b) \in \Theta_1 : \psi(\theta_a, \theta_b) \geq \psi\}$ for the given odds ratio threshold ψ
$\Theta_1(\epsilon)$	Used to denote either $\Theta_1(\delta)$, $\Theta_1(\lambda)$ or $\Theta_1(\psi)$ when it is not specified which is the parameter of interest
$\Theta_1(\epsilon)'$	$\Theta_1(\epsilon) \cap$ PKR for the given PKR

Chapter 1

Introduction

1.1 The need for a new type of statistical hypothesis test

Suppose we wish to know whether a novel drug is effective in treating a particular condition. A clinical trial with 100 patients may be conducted in which 50 are randomly selected to receive the drug while the remaining 50 receive the placebo, and the number of patients in each group making a recovery within one year being recorded. Suppose, at the conclusion of the trial, 38 of the patients who received the drug recovered, while 32 of the patients who received the placebo recovered. Clearly this is evidence in favour of the hypothesis that the drug is more effective than the placebo, but is it sufficient evidence to be conclusive? Or perhaps the result is only suggestive and further trials should be conducted before prescribing the drug.

Statistical hypothesis tests are methods that can be used to quantify evidence in order to more objectively determine the degree of certainty one should place in either of two opposing hypotheses, referred to as the null and alternative hypotheses. In the case of the clinical trial, the null hypothesis is that the drug is only as effective as the placebo, while the alternative hypothesis is that it is more effective than the placebo. While hypothesis tests come in a number of forms, they broadly fit into two methodologies; Frequentist and Bayesian.

Frequentists generally quantify evidence by asking the following question: "If the null hypothesis is indeed true and we repeat the trial in identical manner, what is the probability that we will get evidence *at least as extreme* as we obtained the first time round?" This probability is called a p-value. If the calculated p-value is very small, this means that the evidence we in fact observed was quite unlikely to occur under the null hypothesis, meaning the null hypothesis is likely to be false. A commonly used threshold is $p = 0.05$, where the null hypothesis is deemed to be false (or is 'rejected') if and only if the p-value is less than 0.05.

Bayesians, on the other hand, first attempt to quantify their prior credence in each of the hypotheses before seeing any data. Their prior belief in the odds of the alternative versus the null hypothesis is referred to as the 'prior odds'. The evidence is then used to update the 'prior odds' to 'posterior odds', namely the odds of the alternative versus the null hypothesis after observing the evidence. This is done by using Bayes' rule, which, informally, allows one to calculate the probability of each hypothesis *after* seeing the evidence, provided one has access to the probabilities that each hypothesis would *produce* the observed evidence were it true.

There is a long history of disagreement among statisticians about whether the Frequentist or the Bayesian approach is correct. While the methods are not wholly incompatible, they are underpinned by fundamentally different philosophies. Since both are frequently used, this is problematic for the meta-analyst wishing to combine the results of many different trials that may have been analysed using both Frequentist and Bayesian methods.

Safe hypothesis testing is a new method of statistical inference developed by Grünwald et al. [5] that, while developed from a Frequentist framework, permits the construction of hypothesis tests with both Frequentist and Bayesian interpretations. Thus, statisticians of both schools can use these tests in accordance with their own private philosophies, but in a way that is still straightforward to translate and reinterpret.

A further, significant, advantage of safe hypothesis tests is that they are unaffected by accumulation bias. To explain this, suppose a meta-analyst working for a pharmaceutical company has an incentive to show that a particular drug is effective. If the evidence so far does not look sufficiently convincing, the meta-analyst may decide to wait for more results before 'cashing out' and conducting the meta-analysis. Since trials inherently involve an element of randomness, this choice in timing means there are more chances for the cumulative evidence to look convincing. In this way, the meta-analyst can increase the probability that the evidence will look convincing *at the time of the analysis*. This is clearly a form of bias and is one aspect of 'accumulation bias'. Safe tests, by their construction, are impossible to 'game' in this way; they are safe under optional continuation.

One of the most commonly encountered situations in which a statistical test is required is that of determining whether there is a connection between two binary variables. In a clinical trial for example, we want to know whether there is a connection between whether the patient receives the drug or the placebo and whether the patient recovers or not. Tests in these scenarios are referred to as 2×2 contingency table tests. In this thesis we construct safe versions of such tests. For an outline of the thesis and our contributions, see section 1.5.

1.2 The importance of full knowledge of the experimental procedure

For a statistical experiment, the p-value is defined as the probability of obtaining data at least as extreme as the data actually observed. From this definition it is clear that in order to calculate the p-value, the exact details of the experimental procedure that produced the results must be known. To see how things can go wrong if this knowledge cannot be obtained in full, suppose we are told that a coin was tossed $n = 100$ times came up heads $m = 59$ times and that we want to use this data to test whether the coin is biased in favour of heads. Let q be the true probability of the coin coming up heads. Denote the null hypothesis that the coin is fair by $\mathcal{H}_0 : q = 1/2$ and the alternative hypothesis that the coin is biased in favour of heads by $\mathcal{H}_1 : q > 1/2$, and suppose we pick significance level $\alpha = 0.05$. Let M be the number of heads, so that under the null hypothesis $M \sim \text{Bin}(100, 0.5)$. We then have the p-value $P = P_0(M \geq 59) = 0.04431 < \alpha$. May we therefore conclude that the coin is biased in favour of heads? In fact we cannot. To see why, note that in our analysis we assumed that the total number of tosses $n = 100$ was fixed before the experiment began, but we were not given this information, we were only told that the coin *happened* to be tossed 100 times. Thus the assumption that $M \sim \text{Bin}(n, 1/2)$ where $n = 100$ was assumed to be fixed in advance is not justified.

Indeed, suppose it is now revealed that it was decided in advance that the coin would first be tossed 50 times, at which point the coin would tossed a further 50 times if and only if the above (incorrect) analysis would fail to reject the null hypothesis. Perhaps after the first 50 tosses the coin in fact only came up heads 28 times, at which point we would not have concluded the coin was biased in favour of heads, since $P(M_{50} \geq 28) = 0.2399438 > \alpha$. Thus, since the decision was made to continue for another 50 tosses, we were given an extra chance to conduct our flawed analysis and possibly conclude that the coin was biased in favour of heads. Calculating a valid p-value, would require incorporating knowledge of the decision made at the halfway point.

Note that it follows directly from the definition that if P is a p-value then $P_0(P \leq \alpha) \leq \alpha$, where P_0 represents the distribution of the data under the null hypothesis (for continuous random variables this will be an equality, whereas for discrete random variables equality is not always attained). Let N be the number of times the coin is tossed, so that $N \in \{50, 100\}$, and let P_{50} and P_{100} be the invalid p-values calculated as above by falsely assuming N is fixed from the beginning of the experiment as 50 or 100 respectively. Therefore the invalid p-value P calculated above is equal to P_N, where N is revealed in the course of the experiment. We can now see that P_N is not a valid p-value by showing that $P_0(P_N \leq \alpha) > \alpha$. First, let M_n be the number of heads after n tosses. Then we have

- $P_0(M_{50} \geq 31) = 0.05946$ and $P_0(M_{50} \geq 32) = 0.03245$. Therefore $P_{50} = P_0(M_{50} \geq m_{50}) \leq \alpha$ iff $m_{50} \geq 32$.
- $P_0(M_{100} \geq 58) = 0.06660$ and $P_0(M_{100} \geq 59) = 0.04431$. Therefore $P_{100} = P_0(M_{100} \geq m_{100}) \leq \alpha$ iff $m_{100} \geq 59$.

It then follows that

$$\begin{aligned} P_0(P_N \leq \alpha) &= P_0(P_N \leq \alpha, N = 50) + P_0(P_N \leq \alpha, N = 100) && (1.1)\\ &= P_0(P_{50} \leq \alpha) + P_0(P_{50} > \alpha, P_{100} \leq \alpha) && (1.2)\\ &= P_0(M_{50} \geq 32) + P_0(M_{50} < 32, M_{100} \geq 59) && (1.3)\\ &= 0.03245 + \sum_{k=9}^{31} P_0(M_{50} = k, M_{100} - M_{50} = 59 - k) && (1.4)\\ &= 0.03245 + 0.03025 && (1.5)\\ &= 0.0627, && (1.6) \end{aligned}$$

which is indeed greater than α. In other words, if the coin is in fact unbiased and the experiment was conducted repeatedly, by this faulty analysis we would conclude that the data was at most 5% likely *more* than 5% of the time. Clearly the analysis underestimates the true p-value and so rejects the null hypothesis too readily.

1.3 Meta-Analysis and Accumulation Bias

Large sample sizes are crucial for statistical tests in order to maximize power (the probability of correctly rejecting the null hypothesis). Therefore one advantage of conducting a meta-analysis is the increase in power that comes from pooling the results of many studies. In general, the hope is that by combining the data of many studies one can be more confident of the inferences drawn.

Suppose, after a systematic review of the literature, a series of N studies is found, each testing the same null hypothesis and providing a p-value. Let P_n be the p-value of the n'th study, namely the probability, under the null hypothesis, that the results obtained by repeating that experiment would be at least as extreme as those actually observed. One might then suppose that the probability under the null hypothesis that *all* the studies, were they to be repeated, would produce results at least as extreme as those observed would then be $P \coloneqq \prod_{n=1}^{N} P_n$. However, this is not the case, since it would be making the same mistake as in the coin tossing example: the number of studies is not fixed in advance. In the coin tossing example, it is possible to overcome this difficulty by incorporating the decision process behind the total number of tosses. However this approach cannot possibly work in the case of meta-analyses since the full experimental protocol leading to the full series of studies has so many complicating factors that cannot be quantified before the first study commences. For example:

- A highly powered study concluding with a significant finding may be deemed conclusive enough that further studies are not performed, thus terminating the series.
- The statistician performing the meta-analysis may make the decision of when to perform the meta-analysis based on the results of the studies known at each point in time. This is like the coin tossing example above.
- An unexpected cut or influx of funding to research institutes may increase or decrease the number of studies performed.
- If the first study produces significant results, this may lead to a flurry of replications from researchers also wishing to obtain significant results.
- Conversely, if the first studies produce results that are not significant, this may be sufficient to dissuade researchers from pursuing the same lines, thus terminating the series.

In summary, the problem is that while the results of any two studies may be independent, the *existence* of later studies may be dependent on the results of previous studies. Second, the timing of the meta-analysis may also depend on the results of the studies hitherto available. These two dependencies have been collectively termed *Accumulation Bias* [11]. Moreover, the dependencies may be complex and impossible to quantify, meaning it is impossible to incorporate them into the meta-analysis since this would involve knowing the probability of every possible sequence of events in advance.

As in the coin tossing example, one solution would be to stipulate in advance the number of studies that will be conducted before the meta-analysis. This would eliminate dependencies between the existence of studies and dependencies between the results of the studies and the timing of the meta-analysis. However, this would lead to research waste and hamper the progress of science. For example, if the first few studies produce highly significant results (they have p-values far below the significance threshold), then, since further studies are unlikely to overturn these results, resources may be better spent on follow up research or entirely different lines of research. Conversely, if the first few studies produce results that give very little evidence against the null hypothesis, it may be best to abandon the line of research. In medical contexts it may be especially important to preserve the option to terminate the experiment. Further, it is impossible to rule out the possibility that future studies will become impossible due to unforeseen circumstances such as a cut in funding.

To avoid causing research waste, one might suggest a more nuanced policy whereby a series of studies is permitted to be terminated or extended, provided all the ways in which this may happen are pre-registered. In other words, we make require the pre-registration of a protocol fully specifying the decisions that will be made on whether to terminate or continue the series of studies based on the results so far accumulated. Such a policy however is highly unrealistic since the registration process would be extremely involved, requiring agreements between all researchers with any intention to pursue a line of research even before seeing the first study. Further, the issue remains that series can be terminated by outside events that are impossible to quantify.

A much more practical approach is to find a method of statistical analysis that is robust to post-hoc decisions on whether to continue or terminate a series. This is exactly the problem that analysis via S-values solves.

1.4 How S-values solve the problem of Accumulation Bias

We require a test statistic which gives a type I error guarantee that is robust to accumulation bias. This may seem like a lot to ask of a test statistic given the number of possible strategies that could be used to decide when to terminate the series and conduct a meta-analysis. In fact, we are asking that the type I error guarantee hold even for adversarial decisions on when to terminate.

Suppose we have an almost surely positive test statistic S such that the expectation of S under the null hypothesis is at most one. We will refer to such test statistics as S-values. Note that such a test statistic is unlikely to be large under the null hypothesis. Thus, if S is large, this indicates that the null hypothesis may not be correct. In fact, $1/S$ is a conservative p-value by Markov's inequality

$$P_0(1/S \leq \alpha) = P_0(S \geq 1/\alpha) \leq \frac{\mathbf{E}_0[S]}{1/\alpha} \leq \frac{1}{1/\alpha} = \alpha. \tag{1.7}$$

We can reframe this as a bet that pays out $\$S$ per dollar invested, where under the null hypothesis the bet is at most fair, namely $\mathbf{E}_0[S] \leq 1$. If we win a large sum after investing \$1 in this bet, we may suspect that the null hypothesis is not correct. Suppose now that $S_1, S_2, \ldots$ is a (possibly infinite) sequence of S-values. If we consider these as successive bets, where at each stage we either invest all our accumulated capital into the next bet or decide to cash out, our final winnings is equal to $S^K = S_1 S_2 \cdots S_K$, where K is the number of bets we made.

Suppose the S-values $S_1, S_2, \ldots$ are independent. Then $S^n := S_1 S_2 \cdots S_n$ is a super-martingale under the null hypothesis, since by definition each S-value has expectation at most one under the null. It follows from martingale theory that whatever strategy one employs for deciding when to 'cash-out', one cannot in expectation make a profit from bets that are at best fair, namely bets for which the winnings W satisfies $\mathbf{E}_0[W] \leq 1$. Somewhat more precisely, suppose each S-value S_i is a function of data Z_i, which is a random variable on sample space $\mathcal{Z}_i$. Further, suppose that for each k, the decision on whether to continue to bet $k+1$ can be based not just on the S-values $S_1, S_2, \ldots S_k$, but on the data $Z_1, Z_2, \ldots, Z_k$ on which those S-values are calculated. Formally, a decision rule can be modelled as a function

$$f : \bigcup_{k=1}^{\infty} \left(\prod_{i=1}^{k} \mathcal{Z}_i \right) \to \{\textsc{stop}, \textsc{continue}\}. \tag{1.8}$$

If we now define the stopping time K by

$$K := \min\{k : f(Z_1, \ldots, Z_k) = \text{STOP}\}, \tag{1.9}$$

we see that our final winnings is given by $S^K := S_1 S_2 \cdots S_k$. Now the random variable K is a stopping time in the sense of random process theory and this implies[1], since S^n is a super-martingale, that $\mathbf{E}_0[S^K] \leq 1$. Thus S^K is also an S-value and any choice of decision rule cannot destroy the type I error guarantee. In particular, suppose we use the greedy decision rule that continues betting until $S^n \geq 1/\alpha$. We then see, again by Markov's inequality, that

$$P_0\left(\exists k : \prod_{i=1}^{k} S_i \geq 1/\alpha\right) = P_0(S^K \geq 1/\alpha) \leq 1/\alpha, \tag{1.10}$$

implying that it is likely we will never cash out at all. While such an adversarial strategy would constitute p-hacking were we working with p-values, we see that S-values are robust to this kind of 'gaming'. It is thus clear that whatever decision rule is followed to decide whether to conduct further studies or to move to meta-analysis, multiplying all S-values together gives a conservative p-value once the reciprocal is taken. Thus S-values can provide a type I error guarantee even in the presence of accumulation bias. Finally, we note that while we assumed $S_1, S_2, \ldots$ were independent, the result that S^K is also an S-value for any stopping time K in fact generalizes beyond far beyond that. For details, see [5, section 2].

Can we always find a non-trivial test statistic with expectation at most one under the null hypothesis? If we can find several such test statistics, how should we choose between them? In the following section, we outline the theory of these so-called S-values in more detail and provide answers to these questions.

1.5 Thesis outline

Chapter two outlines some of the general theory developed by Grünwald et al. [5] on safe hypothesis testing that will form the foundation of this thesis. Section 2.1 formally defines safe tests in terms of S-values and gives the first examples of S-values, namely Bayes factors for simple null hypotheses. Section 2.2 then introduces the GROW criterion that will be used to select S-values that 'grow' as quickly as possible under the alternative hypothesis. Section 2.3 presents the main result found in [5] which characterizes the GROW S-value in terms of information projections. This gives the GROW S-value as the solution to a convex optimization problem that can then be approximated numerically. In section 2.4 we apply the result to a large class of statistical models, namely exponential families, and prove a corollary of a theorem found in [5]. We will use this corollary in chapter four when constructing conditional S-values. Finally, in section 2.5, we explore the connection between GROW S-values and the concept of a uniformly most powerful Bayesian test (UMPBT), developed by Johnson [7]. The connection is discussed in [5] and in this chapter we give a rigorous proof.

The short chapter three introduces 2×2 contingency tables. After providing some brief context in section 3.1, section 3.2 fixes the notation that will be used in the rest of the thesis. Section 3.3 introduces Fisher's exact test, which is commonly used when evaluating 2×2 tables and will be used as a benchmark by which to measure the power of the S-values we construct in subsequent chapters. In section 3.4 we clarify the difference between 'conditional' and 'unconditional' S-values; the two types of S-values that we will construct in this thesis. Finally, section 3.5 discusses the relation of this thesis to previous work.

Chapter four is where we begin constructing safe tests for 2×2 contingency tables, starting with conditional tests. The GROW S-value can be found in explicit form by using a corollary proved in chapter two, since the underlying distribution forms an exponential family in the conditional setting. The uniformly most powerful GROW S-values (UMPG S-values) are defined and can also be found using theory from chapter two, but only in implicit form. We then evaluate the growth and power of

[1] We have omitted many details here. For a rigorous measure-theoretic proof, see [5, section 2].

these S-values. Although they are developed in the conditional setting, we show that these tests can be used in a valid way also in the unconditional setting. Their evaluation in this setting is postponed to chapter seven, so that they can be compared with the S-values constructed in the following chapters.

Chapter five explores the unconditional setting, which in many cases is closer to reality (indeed, for the drug trial example, the unconditional case refers to the scenario where the total number of patients who recover, referred to as N_1, is not known in advance). Similar to the conditional S-values constructed in chapter four, we will construct two types, namely the *unconditional GROW* S-values and, where they exist, the *unconditional UMPG* S-values. We discuss a number of ways in which the parameter sets can be restricted to ensure that they are positively separated (which is necessary to prevent the GROW S-value being degenerate). We do this by setting threshold values for the risk difference, relative risk or the odds ratio; all parameters that are frequently used in clinical research when analysing 2×2 tables. We also allow for the possibility that the practitioner has prior knowledge, showing how this can be incorporated into the test. Chapter five focuses on situations in which the shortcuts of chapter six cannot be used and the information projection must be approximated directly. While this can be very slow, we provide a simplified expression for the gradient of the objective function that can speed up calculations by around an order of magnitude.

In chapter six we provide a number of theoretical results that bypass the slow information projection approximation used in chapter five and significantly speed up the calculation of the GROW S-values. We show that in most cases of interest the GROW S-values can be calculated very quickly without an information projection. Indeed, in two cases we provide explicit formulas for the GROW S-values. The method developed in this chapter to identify GROW S-values may perhaps be used in other testing scenarios. Although the chapter is heavily algebraic, the results can frequently be visualized geometrically. This is by far the longest chapter and contains a significant proportion of the original work in this thesis. For a more detailed summary of the results, see the introduction to the chapter.

Chapter seven collates the unconditional S-values developed in chapters three to six and evaluates them in terms of growth and power. Unfortunately there is not a clear winner overall, although it can be seen that in some cases certain S-values are to be preferred.

Chapter 2

S-values and safe hypothesis testing

In this chapter we outline some of the general theory developed by Grünwald et al. [5] on safe hypothesis testing that will form the foundation of this thesis. Section 2.1 formally defines safe tests in terms of S-values and gives the first examples of S-values, namely Bayes factors for simple null hypotheses. Section 2.2 then introduces the GROW criterion that will be used to select S-values that 'grow' as quickly as possible under the alternative hypothesis. Section 2.3 presents the main result found in [5] which characterizes the GROW S-value in terms of information projections. This gives the GROW S-value as the solution to a convex optimization problem that can then be approximated numerically (see chapter 5). In section 2.4 we apply the result to a large class of statistical models, namely exponential families and prove a corollary of the relevant theorem found in [5]. Finally, in section 2.5, we explore the connection between GROW S-values and the concept of a uniformly most powerful Bayesian test (UMPBT), developed by Johnson [7]. This connection is discussed in [5] and here we give a rigorous proof.

2.1 Definitions and first examples

Suppose we have the model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, indexed by $\Theta = \Theta_0 \cup \Theta_1$, where the null and alternative hypotheses are $\mathcal{H}_0 : \theta \in \Theta_0$ and $\mathcal{H}_1 : \theta \in \Theta_1$ respectively. Denoting our data by the random variable $Z \sim P_\theta$ for some unknown $\theta \in \Theta$, we assume that each P_θ has an associated density or probability mass function p_θ.[1] At this point we make no assumption on the structure of Θ; it is simply an indexing set, so the model may be parametric or non-parametric. The following definitions are made.

Definition 1. An S-value is any non-negative statistic $S = S(Z)$ such that $\mathbf{E}_\theta[S] \leq 1$ for all $\theta \in \Theta_0$.

Definition 2. For any S-value S and significance level α, we define the associated safe test $T_\alpha(S)$ as the test that rejects $\mathcal{H}_0$ iff $S \geq 1/\alpha$. We see by (1.7) that $T_\alpha(S)$ has type I error bound α.

Since S-values are defined with respect to Θ_0, we are justified in denoting the set of S-values for a given null parameter set Θ_0 by $\mathcal{E}(\Theta_0)$. As a first example, suppose the null hypothesis is simple, namely $\Theta_0 = \{\theta_0\}$ is a singleton. Then for any $\theta_1 \in \Theta_1$, the test statistic

$$T_{\theta_1}(Z) := \frac{p_{\theta_1}(Z)}{p_{\theta_0}(Z)} \tag{2.1}$$

is an S-value. This can be seen by directly calculating the expectation of $T_{\theta_1}(Z)$ when $Z \sim P_{\theta_0}$ as follows

$$\mathbf{E}_{Z\sim P_{\theta_0}}\left[\frac{p_{\theta_1}(Z)}{p_{\theta_0}(Z)}\right] = \int p_{\theta_0}(z)\frac{p_{\theta_1}(z)}{p_{\theta_0}(z)}\mathrm{d}z \tag{2.2}$$

$$= \int p_{\theta_1}(z)\mathrm{d}z \tag{2.3}$$

$$= 1. \tag{2.4}$$

[1] For the purposes of this thesis, Z will always be a discrete random variable and p_θ will be a probability mass function. If Z is continuous however, it is assumed that all the distributions P_θ have densities p_θ with respect to a common measure, say μ.

Now, for any set A, let $\mathcal{W}(A)$ be the set of probability distributions on A. For any $W_1 \in \mathcal{W}(\Theta_1)$ (which we will also call a *prior* as discussed below), let P_{W_1} denote the marginal distribution of Z when $\theta \sim W_1$. If W_1 has a density w_1, then P_{W_1} has density

$$p_{W_1}(z) \coloneqq \int_{\Theta_1} w_1(\theta) p_\theta(z) \mathrm{d}\theta. \tag{2.5}$$

More generally, if W_1 does not necessarily have a density (for example if it is a point mass), we can define

$$p_{W_1}(z) \coloneqq \mathbf{E}_{\theta \sim W_1}[p_\theta(z)]. \tag{2.6}$$

The Bayesian approach to testing $\mathcal{H}_1 : \theta \in \Theta_1$ against $\mathcal{H}_0 : \theta \in \Theta_0$ is to take prior probabilities π_1 and π_0 as estimates of $P(\theta \in \Theta_1)$ and $P(\theta \in \Theta_0)$, respectively. Also, priors $W_1 \in \mathcal{W}(\Theta_1)$ and $W_0 \in \mathcal{W}(\Theta_0)$ are used to reflect prior beliefs on the likelihood of different values of θ. Once data Z is observed, Bayes rule can be used to obtain, for $i \in \{0, 1\}$,

$$P(\theta \in \Theta_i | Z) = \frac{p(Z|\theta \in \Theta_i) P(\theta \in \Theta_i)}{p(Z)} \tag{2.7}$$

$$= \frac{p_{W_i}(Z)\pi_i}{p(Z)}. \tag{2.8}$$

The densities p_{W_1} and p_{W_0} are referred to as the Bayes marginal probability distributions. The posterior odds can then be calculated as

$$\frac{P(\theta \in \Theta_1 | Z)}{P(\theta \in \Theta_0 | Z)} = \frac{p_{W_1}(Z)}{p_{W_0}(Z)} \frac{\pi_1}{\pi_0}, \tag{2.9}$$

where the likelihood ratio

$$\mathrm{BF}_{10}(Z) \coloneqq \frac{p_{W_1}(Z)}{p_{W_0}(Z)}. \tag{2.10}$$

is referred to as the Bayes factor. Thus the posterior odds is equal to the prior odds π_1/π_0 times the Bayes factor.

If a definitive decision is required, a threshold $\gamma > 1$ may be chosen to determine when to reject the null hypothesis. More precisely, it may be decided that $\mathcal{H}_0$ will be rejected in favour of $\mathcal{H}_1$ iff $\mathrm{BF}_{10} \geq \gamma$. We will refer to such a test as a *Bayesian hypothesis test* and we will return to them in section 2.5 when discussing uniformly most powerful Bayesian tests. This is reminiscent of the safe test $T_\alpha(S)$, for any S-value S, which rejects $\mathcal{H}_0$ iff $S \geq 1/\alpha$. Indeed, if $\Theta_0 = \{\theta_0\}$ is a singleton then BF_{10} is an S-value. This can be see by direct calculation, in exactly the same way as above

$$\mathbf{E}_{Z \sim P_{\theta_0}}[\mathrm{BF}_{10}(Z)] = \mathbf{E}_{Z \sim P_{\theta_0}}\left[\frac{p_{W_1}(Z)}{p_{\theta_0}(Z)}\right] \tag{2.11}$$

$$= \int p_{\theta_0}(z) \frac{p_{W_1}(z)}{p_{\theta_0}(z)} \mathrm{d}z \tag{2.12}$$

$$= \int p_{W_1}(z) \mathrm{d}z \tag{2.13}$$

$$= 1. \tag{2.14}$$

Note that BF_{10} is an S-value for singleton Θ_0 even if $W_1 \in \mathcal{W}(\Theta_1)$ is an *arbitrary* probability distribution, not necessarily in alignment with our prior beliefs about the likelihood of different values of θ. We may therefore wonder, for general Θ_0, whether BF_{10} is an S-value for *arbitrary* probability distributions W_1 and W_0. We will see in the following sections that while this is not true in general, there do always exist probability distributions W_1 and W_0 for which the corresponding Bayes factor BF_{10} is an S-value, even for non-trivial Θ_0. In fact—in a sense to be defined later—the 'best' S-value is always a Bayes factor.

We should note that in this thesis we will refer to elements of $\mathcal{W}(\Theta_0)$ and $\mathcal{W}(\Theta_1)$ as 'priors', and test statistics $p_{W_1}(Z)/p_{W_0}(Z)$ as 'Bayes factors' *whether or not* the priors W_1 and W_0 align with our prior beliefs on the likelihood of different values of θ. For the purposes of this thesis, a prior is *any* distribution on (a subset of) the null or alternative parameter set.

2.2 Optimal S-values: growth rate and the GROW criterion

Since large values of S are unlikely under the null hypothesis, a good S value should be likely to be large under the alternative hypothesis, which can then be interpreted as evidence against the null. Ordinarily, for example in the Neyman-Pearson setting, this would be formalized by saying we want a statistical test with large power, where power is defined as the smallest probability of correctly rejecting the null hypothesis, namely the power is equal to

$$1 - \beta = \inf_{\theta \in \Theta_1'} P_\theta(T_\alpha(S) = \text{REJECT}) = \inf_{\theta \in \Theta_1'} P_\theta(S \geq 1/\alpha), \tag{2.15}$$

where β is the type II error. This is the choice made in the Neyman-Pearson paradigm; once a type I error threshold α has been fixed, the most powerful test satisfying the type I error constraint is chosen. With such a worst-case methodology, it is usually necessary to take a subset $\Theta_1' \subseteq \Theta_1$ to ensure Θ_0 and Θ_1 are strictly separated.

Instead of working with power however, [5] define the *growth rate* of an S-value at any particular $\theta \in \Theta_1$ as

$$\text{GR}_\theta(S) \coloneqq \mathbf{E}_{Z \sim P_\theta}[\log S], \tag{2.16}$$

and the *worst case growth rate* as

$$\text{GR}(S) \coloneqq \inf_{\theta \in \Theta_1} \mathbf{E}_{Z \sim P_\theta}[\log S]. \tag{2.17}$$

[5] then formulate the *GROW criterion* (Growth Rate Optimal in the Worst case), which states that the S-value that maximizes the worst case growth rate should be chosen. This S-value is then denoted S^* and is referred to as the GROW S-value. Thus S^*, if it exists, achieves

$$\sup_{S \in \mathcal{E}(\Theta_0)} \inf_{\theta \in \Theta_1} \mathbf{E}_{Z \sim P_\theta}[\log S]. \tag{2.18}$$

While growth rate is analogous to power, choosing an S-value with larger growth rate does not necessarily mean it will have larger power (and vice versa). Just as in the Neyman-Pearson paradigm, it may be necessary to restrict Θ_1 to a subset Θ_1'. This is to ensure that the parameter sets are sufficiently separated so that the worst case growth rate is not degenerate. For any given $\Theta_1' \subseteq \Theta_1$, the corresponding GROW S-value is denoted $S^*_{\Theta_1'}$.

A number of explanations are given in [5] as to why the logarithm of S is taken in the definition of growth rate, rather than defining the growth rate at θ by $\mathbf{E}_{Z \sim P_\theta}[f(S)]$ for some other function f, perhaps the identity. First, if f is the identity, it ends up being the case that the GROW S-value is frequently zero with positive probability. This is undesirable from the viewpoint of optional continuation (which is indeed the purpose of S-values), since if any S-value in a sequence of S-values is zero, the product will remain zero after that point, meaning the null hypothesis will never be rejected however strong any subsequent evidence against it is. A similar problem arises when f is any other polynomial. However, the problem does not arise with the logarithm.

A second justification is that, if the alternative hypothesis is true, we would like the running product of a series of S-values to grow as quickly as possible. This is because a large product is interpreted as greater evidence against the null. It is shown in [5, section 3.1] that the logarithm is the natural choice since it minimizes the average time at which the running product doubles. For a more detailed discussion and other reasons for choosing the logarithm, see [5, section 3.1].

It is not immediately apparent how one would go about finding the GROW S-value. However, Grünwald et al. [5] provide a remarkable theorem (Theorem 1), which states that the GROW S-value is in fact a Bayes factor—as in 2.10—where the priors W_1, W_0 are such that the KL divergence between the marginals P_{W_1}, P_{W_0} is minimized. These special priors are denoted by W_1^* and W_0^*, and in general are unlikely to coincide with our prior beliefs on the likelihood of different values of θ. Furthermore, they prove that the GROW S-value is 'essentially unique', meaning any other S-value satisfying the GROW criterion is almost surely equal to S^*, regardless of the true value of the parameter θ. We discuss this theorem in the following section.

2.3 Characterizing the GROW S-value using information projection

Let P and Q be any two distributions defined on the same probability space $\mathcal{Z}$, such that P is absolutely continuous[2] with respect to Q. We define the Kullback–Leibler divergence from Q to P by

$$\mathrm{KL}(P||Q) := \int_{\mathcal{Z}} \log\left(\frac{dP}{dQ}\right) dP, \tag{2.19}$$

where dP/dQ is the Radon-Nikodym derivative of P with respect to Q.

Now let $\mathcal{P}$ and $\mathcal{Q}$ be two arbitrary sets of distributions, where every distribution in both sets is defined on the same probability space $\mathcal{Z}$. For a given $Q \in \mathcal{Q}$, we may define

$$d(\mathcal{P}, Q) := \inf_{P \in \mathcal{P}} \mathrm{KL}(P||Q). \tag{2.20}$$

If there exists a unique $P \in \mathcal{P}$ achieving this infimum, it is denoted by P^* and is referred to as the *information projection* (IP) of Q onto $\mathcal{P}$, namely

$$P^* := \arg\min_{P \in \mathcal{P}} \mathrm{KL}(P||Q). \tag{2.21}$$

Likewise, for any $P \in \mathcal{P}$, we may define the *reverse information projection* (RIP) of P onto $\mathcal{Q}$ by

$$Q^* := \arg\min_{Q \in \mathcal{Q}} \mathrm{KL}(P||Q), \tag{2.22}$$

provided Q^* exists and is unique. Finally, we may define the *joint information projection* (JIP) of $\mathcal{P}$ and $\mathcal{Q}$ onto each other by

$$(P^*, Q^*) := \arg\min_{(P,Q) \in \mathcal{P} \times \mathcal{Q}} \mathrm{KL}(P||Q), \tag{2.23}$$

provided P^* and Q^* exist and are unique.

Now take $\mathcal{P} = \{P_{W_1} : W_1 \in \mathcal{W}(\Theta_1)\}$ and $\mathcal{Q} = \{P_{W_0} : W_0 \in \mathcal{W}(\Theta_0)\}$ and suppose the JIP $(P^*, Q^*) \in \mathcal{P} \times \mathcal{Q}$ exists and is achieved uniquely by the priors W_1^* and W_0^*, namely $P^* = P_{W_1^*}$ and $Q^* = P_{W_0^*}$. Then the following theorem states that the GROW S-value S^* is given by the Bayes factor generated by these priors, namely,

$$S^* = \frac{p_{W_1^*}(Z)}{p_{W_0^*}(Z)}. \tag{2.24}$$

Further, the worst case growth rate of S^* is attained at W_1^*, where it is equal to $\mathrm{KL}(P_{W_1^*}||P_{W_0^*})$, which can be thought of as the minimum 'distance' between distributions in $\mathcal{P}$ and $\mathcal{Q}$. Thus, informally, the greater the separation between $\mathcal{P}$ and $\mathcal{Q}$, the more S^* will grow (at least in the worst case) and so the easier it will be to correctly reject the null hypothesis. Finally, it states that the GROW S-value S^* is 'essentially unique' where this is taken to mean that if $\tilde{S}$ is any other S-value satisfying the GROW criterion, then $P_\theta(S^* = \tilde{S}) = 1$ for all $\theta \in \Theta_0 \cup \Theta_1$. Since this means their growth rates and powers at any θ are equal, it is irrelevant from an inference perspective which we choose to use. We now give the theorem.

Theorem 1 (Grünwald, [5]). *Let $\Theta_1' \subseteq \Theta_1$ and suppose that for all $\theta_0 \in \Theta_0$ and $W_1 \in \mathcal{W}(\Theta_1')$ we have that P_{θ_0} is absolutely continuous relative to P_{W_1}. If $\inf_{(W_1,W_0) \in \mathcal{W}(\Theta_1') \times \mathcal{W}(\Theta_0)} \mathrm{KL}(P_{W_1}||P_{W_0}) < \infty$ and the infimum is achieved uniquely by (W_1^*, W_0^*), then the GROW S-value $S^*_{\Theta_1'}$ exists, is essentially unique, and is given by*

$$S^*_{\Theta_1'}(Z) = \frac{p_{W_1^*}(Z)}{p_{W_0^*}(Z)}. \tag{2.25}$$

Further,

$$\inf_{W_1 \in \mathcal{W}(\Theta_1')} \mathbf{E}_{P_{W_1}}\left[\log S^*_{\Theta_1'}\right] = \mathbf{E}_{P_{W_1^*}}\left[\log S^*_{\Theta_1'}\right] = \mathrm{KL}(P_{W_1^*}||P_{W_0^*}), \tag{2.26}$$

meaning $S^*_{\Theta_1'}$ achieves its worst case growth rate at W_1^*.

[2]Meaning $Q(A) = 0 \implies P(A) = 0$ for any set A in the underlying σ-algebra. While we do not explicitly refer to σ-algebras in this thesis, it is always assumed that we take the Borel σ-algebra, which is the σ-algebra generated by the open sets.

This characterization of $S^*_{\Theta'_1}$ clarifies the importance of ensuring the parameter sets are sufficiently separated. For if the infimum is achieved and equals zero, we have

$$\mathrm{KL}(P_{W_1^*}||P_{W_0^*}) = 0 \implies P_{W_1^*} = P_{W_0^*} \implies S_{\Theta'_1} \equiv 1, \tag{2.27}$$

meaning the test $T_\alpha(S_{\Theta'_1})$ is useless as it never leads to a rejection of the null hypothesis.

Although the above theorem does not provide a closed formula for the GROW S-value, the characterization reformulates the search as a convex optimization problem since the KL divergence is jointly-convex. More precisely, the map $\kappa : \mathcal{P} \times \mathcal{Q} \to [0, \infty]$ given by

$$(P, Q) \mapsto \mathrm{KL}(P||Q) \tag{2.28}$$

is jointly convex in the sense that for any two pairs of distributions $(P_1, Q_1), (P_2, Q_2) \in \mathcal{P} \times \mathcal{Q}$ and any $\alpha \in [0, 1]$, we have (see [13])

$$\begin{aligned}
\kappa(\alpha(P_1, Q_1) + (1-\alpha)(P_2, Q_2)) &= \kappa(\alpha P_1 + (1-\alpha)P_2, \alpha Q_1 + (1-\alpha)Q_2) && (2.29)\\
&\coloneqq \mathrm{KL}(\alpha P_1 + (1-\alpha)P_2||\alpha Q_1 + (1-\alpha)Q_2) && (2.30)\\
&\leq \alpha \mathrm{KL}(P_1||Q_1) + (1-\alpha)\mathrm{KL}(P_2||Q_2) && (2.31)\\
&= \alpha\kappa(P_1, Q_1) + (1-\alpha)\kappa(P_2, Q_2). && (2.32)
\end{aligned}$$

Further, the map $m : \mathcal{W}(\Theta'_1) \times \mathcal{W}(\Theta_0) \to \mathcal{P} \times \mathcal{Q}$ given by

$$(W_1, W_0) \mapsto (P_{W_1}, P_{W_0}) \tag{2.33}$$

is linear as follows. Let $W_i, W'_i \in \mathcal{W}(\Theta_i)$ for $i \in \{0, 1\}$ and $\alpha \in [0, 1]$. Then

$$\begin{aligned}
m(\alpha(W_1, W_0) + (1-\alpha)(W'_1, W'_0)) &= m((\alpha W_1 + (1-\alpha)W'_1, \alpha W_0 + (1-\alpha)W'_0)) && (2.34)\\
&= (P_{\alpha W_1 + (1-\alpha)W'_1}, P_{\alpha W_0 + (1-\alpha)W'_0}), && (2.35)
\end{aligned}$$

where, for $i \in \{0, 1\}$, the distribution $P_{\alpha W_i + (1-\alpha)W'_i}$ is given by

$$\begin{aligned}
P_{\alpha W_i + (1-\alpha)W'_i}(z) &= \mathbf{E}_{\theta \sim \alpha W_i + (1-\alpha)W'_i}[P_\theta(z)] && (2.36)\\
&= \alpha \mathbf{E}_{\theta \sim W_i}[P_\theta(z)] + (1-\alpha)\mathbf{E}_{\theta \sim (1-\alpha)W'_i}[P_\theta(z)] && (2.37)\\
&= \alpha P_{W_i}(z) + (1-\alpha)P_{W'_i}(z). && (2.38)
\end{aligned}$$

Substituting this into (2.38), we have

$$\begin{aligned}
m(\alpha(W_1, W_0) + (1-\alpha)(W'_1, W'_0)) &= (\alpha P_{W_1} + (1-\alpha)P_{W'_1}, \alpha P_{W_0} + (1-\alpha)P_{W'_0}) && (2.39)\\
&= \alpha(P_{W_1}, P_{W_0}) + (1-\alpha)(P_{W'_1}, P_{W'_0}). && (2.40)
\end{aligned}$$

Finally, since the composition of a linear map with a convex map is itself convex, we see that the map $\kappa \circ m : \mathcal{W}(\Theta'_1) \times \mathcal{W}(\Theta_0) \to [0, \infty]$, which is then given by

$$(W_1, W_0) \mapsto \mathrm{KL}(P_{W_1}||P_{W_0}), \tag{2.41}$$

is convex. Since the parameter sets may be of arbitrary size, this may still be an infinite-dimensional convex optimization problem. Notwithstanding, we will see later how discretizing the parameter sets does not destroy convexity and leads to a convex optimization problem reasonable enough to permit numerical approximation.

2.4 Exponential families

Let $\Theta \subseteq \mathbb{R}$ be a one-dimensional parameter set and suppose we have a set of distributions on $\mathbb{R}^n$ parametrized by Θ, say $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. We call such a set a *one-parameter exponential family* if

there exist functions h, η, T and A such that each P_θ has a probability mass function or a density p_θ (relative to some common background measure λ say) that can be expressed in the following form

$$p_\theta(x) = h(x) \exp\big(\eta(\theta)T(x) - A(\theta)\big). \tag{2.42}$$

In this case, η is referred to as the *canonical parameter* and the set $E \coloneqq \{\eta(\theta) : \theta \in \Theta\}$ as the *canonical parameter space*.

The same exponential family may be parametrized in different ways. An important example is the so-called *mean-value parametrization*. Suppose we have a one-parameter exponential family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where θ is in fact the canonical parameter. We may define the *mean-value parameter* μ by

$$\mu(\theta) \coloneqq \mathbf{E}_{X \sim P_\theta}[X]. \tag{2.43}$$

It can be shown that $\mu(\theta)$ is strictly increasing in θ and so the map (2.43) is one-to-one [6, Section 18.3]. This means the family $\mathcal{P}$ can be re-parametrized in terms of μ as follows

$$\mathcal{P} = \{Q_\mu : \mu \in M\}, \quad \text{where} \quad M \coloneqq \{\mu(\theta) : \theta \in \Theta\} \quad \text{and} \quad Q_{\mu(\theta)} \coloneqq P_\theta. \tag{2.44}$$

The mean-value parametrization will be useful in the later construction of some S-values (see chapter 4, section 3).

Grünwald et al. [5, Section 4.1, proof in Appendix D] provide the following theorem, which states that for exponential families with mean-value parameter θ and alternative parameter set $\Theta_1 = [\theta_1, \infty) \cap \Theta$, the GROW S-value is achieved by the prior W_1^* that puts all its mass on θ_1.

Theorem 2. *Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}$ be a one-parameter exponential family for sample space $\mathcal{Z}$, given in its mean-value parametrization. Suppose $0 \in \Theta$, let $\Theta_0 = \{0\}$ and take $\Theta_1(\underline{\delta}) = [\underline{\delta}, \infty) \cap \Theta$ for some threshold value $\underline{\delta}$ contained in the interior of Θ. Then the GROW S-value $S_{\Theta_1(\underline{\delta})}$, referred to as the $\underline{\delta}$-GROW S-value and denoted by $S^*_{\underline{\delta}}$, is given by*

$$S^*_{\underline{\delta}}(Z) \coloneqq S^*_{\Theta_1(\underline{\delta})}(Z) = \frac{p_{\underline{\delta}}(Z)}{p_0(Z)}, \tag{2.45}$$

where, for each θ, p_θ is the density or probability mass function of the distribution P_θ.

Suppose we have a one-parameter exponential family as in Theorem 2, but we wish to test $\Theta_0 = \{\theta_0\}$ against $\Theta_1(\underline{\delta})$, for some $\theta_0 \neq 0$ in the interior of Θ such that $\underline{\delta} > \theta_0$. Theorem 2 can be readily extended to this scenario using simple properties of the KL-divergence and exponential families. We have the following proposition, which will be used later when we look at Fisher's noncentral hypergeometric distribution. Intuitively, this is a simple corollary of the previous result, since shifting all the distributions and random variables does not change the KL–divergences or the fact that we are dealing with an exponential family. However, a precise proof requires some work.

Proposition 3. *Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}$ be a one-parameter exponential family for sample space $\mathcal{Z}$, given in its mean-value parametrization. Suppose $\Theta_0 = \{\theta_0\}$ and $\Theta_1(\underline{\delta}) = [\underline{\delta}, \infty) \cap \Theta$, where $\underline{\delta} > \theta_0$ and both θ_0 and $\underline{\delta}$ lie in the interior of Θ. Then the GROW S-value $S_{\Theta_1(\underline{\delta})}$, referred to as the $\underline{\delta}$-GROW S-value and denoted by $S^*_{\underline{\delta}}$, is given by*

$$S^*_{\underline{\delta}}(Z) \coloneqq S^*_{\Theta_1(\underline{\delta})}(Z) = \frac{p_{\underline{\delta}}(Z)}{p_{\theta_0}(Z)}, \tag{2.46}$$

where, for each θ, p_θ is the density or probability mass function of the distribution P_θ.

Proof. Let $\mathcal{Q} = \{Q_\varphi : \varphi \in \Phi\}$, where $\Phi \coloneqq \{\theta - \theta_0 : \theta \in \Theta\}$ and Q_φ is the distribution of $Z - \theta_0$ for $Z \sim P_{\varphi + \theta_0}$. Now, since $\mathcal{P}$ is an exponential family, the density of P_θ can be written as

$$p_\theta(z) = h(z) \exp\big(\eta(\theta)T(z) - A(\theta)\big), \tag{2.47}$$

for some h, η, T and A. Since, for any φ, Q_φ is simply a translation of $P_{\varphi+\theta_0}$, then the density of Q_φ can be written as

$$q_\varphi(y) = p_{\varphi+\theta_0}(y + \theta_0) \tag{2.48}$$

and so

$$q_\varphi(y) = h(y+\theta_0)\exp\big(\eta(\varphi+\theta_0)T(y+\theta_0) - A(\varphi+\theta_0)\big) \tag{2.49}$$
$$= \tilde{h}(y)\exp\big(\tilde{\eta}(\varphi)\tilde{T}(y) - \tilde{A}(\varphi)\big), \tag{2.50}$$

where $\tilde{h}(h) \coloneqq h(y+\theta_0)$, $\tilde{\eta}(\varphi) \coloneqq \eta(\varphi+\theta_0)$, $\tilde{T}(y) \coloneqq T(y+\theta_0)$ and $\tilde{A}(\varphi) \coloneqq A(\varphi+\theta_0)$. Therefore $\mathcal{Q}$ is also an exponential family. Further, since

$$\mathbf{E}_{Y\sim Q_\varphi}[Y] = \mathbf{E}_{Y\sim Q_\varphi}[Y+\theta_0] - \theta_0 \tag{2.51}$$
$$= \mathbf{E}_{Z\sim P_{\varphi+\theta_0}}[Z] - \theta_0 \tag{2.52}$$
$$= \varphi + \theta_0 - \theta_0 \tag{2.53}$$
$$= \varphi, \tag{2.54}$$

we see that $\mathcal{Q} = \{Q_\varphi : \varphi \in \Phi\}$ is in fact in the mean-value parametrization. Lastly, since $\theta_0 \in \Theta_0$, by the definition of Φ we have $0 \in \Phi$.

We can therefore apply Theorem 2 to $\mathcal{Q}$, with $\Phi_0 = \{0\}$ and $\Phi_1(\underline{\delta}) \coloneqq [\underline{\delta} - \theta_0, \infty) \cap \Phi$, to obtain

$$S^*_{\Phi_1(\underline{\delta})}(Y) = \frac{q_{\underline{\delta}-\theta_0}(Y)}{q_0(Y)}. \tag{2.55}$$

Recall that GROW S-values are constructed using the JIP, so that the previous line implies

$$\dot{W}_{\underline{\delta}-\theta_0} = \operatorname*{arg\,min}_{W_1\in\mathcal{W}(\Phi_1(\underline{\delta}))} \mathrm{KL}(Q_{W_1}||Q_0), \tag{2.56}$$

where $\dot{W}_{\underline{\delta}-\theta_0}$ represents a point mass on $\underline{\delta}-\theta_0$.

By definition of Q_φ, we have that $Y \sim Q_\varphi$ implies $Y+\theta_0 \sim P_{\varphi+\theta_0}$. We now see that this holds more generally. Namely, for any $W_1 \in \mathcal{W}(\Phi_1(\underline{\delta}))$, let $f(W_1) \in \mathcal{W}(\Theta_1(\underline{\delta}))$ be defined such that

$$\varphi \sim W_1 \implies \varphi + \theta_0 \sim f(W_1). \tag{2.57}$$

We then have

$$q_{W_1}(y) \coloneqq \mathbf{E}_{\varphi\sim W_1}[q_\varphi(y)] \tag{2.58}$$
$$= \mathbf{E}_{\varphi\sim W_1}[p_{\varphi+\theta_0}(y+\theta_0)] \qquad \text{by (2.48)} \tag{2.59}$$
$$= \mathbf{E}_{\theta\sim f(W_1)}[p_\theta(y+\theta_0)] \tag{2.60}$$
$$= p_{f(W_1)}(y+\theta_0), \tag{2.61}$$

and so $Y \sim Q_{W_1}$ implies $Y+\theta_0 \sim P_{f(W_1)}$.

Note that, since $\Phi_1(\underline{\delta})$ is a translation of $\Theta_1(\underline{\delta})$, $f : \mathcal{W}(\Phi_1(\underline{\delta})) \to \mathcal{W}(\Theta_1(\underline{\delta}))$ is a bijection. Now, it is a fact that the KL-divergence is invariant under transformations of the variables that are differentiable and invertible, which includes translations [10]. Thus, for any $W_1 \in \mathcal{W}(\Phi_1(\underline{\delta}))$, we have

$$\mathrm{KL}(Q_{W_1}||Q_0) = \mathrm{KL}(P_{f(W_1)}||P_{\theta_0}). \tag{2.62}$$

Combining these two facts, (2.56) becomes

$$\dot{W}_{\underline{\delta}-\theta_0} = \operatorname*{arg\,min}_{W_1\in\mathcal{W}(\Phi_1(\underline{\delta}))} \mathrm{KL}(Q_{W_1}||Q_0) \tag{2.63}$$
$$= \operatorname*{arg\,min}_{W_1\in\mathcal{W}(\Phi_1(\underline{\delta}))} \mathrm{KL}(P_{f(W_1)}||P_{\theta_0}) \tag{2.64}$$

and so

$$\dot{W}_{\underline{\delta}} = f(\dot{W}_{\underline{\delta}-\theta_0}) = \operatorname*{arg\,min}_{W_1\in\mathcal{W}(\Theta_1(\underline{\delta}))} \mathrm{KL}(P_{W_1}||P_{\theta_0}). \tag{2.65}$$

Since Θ_0 is a singleton, we therefore see that the RIP is given by $P_{\dot{W}_{\underline{\delta}}} = P_{\underline{\delta}}$ and P_{θ_0}, so that, by Theorem 1, the GROW S-value $S_{\Theta_1(\underline{\delta})}$ is given by

$$S_{\Theta_1(\underline{\delta})}(Z) = \frac{p_{\underline{\delta}}(Z)}{p_{\theta_0}(Z)}. \tag{2.66}$$

□

2.5 UMPGTs and connections to UMPBTs

Suppose, for a given experiment, that there is a natural way of constructing a restriction $\Theta_1(\underline{\delta})$ of Θ_1 based on some parameter δ with a threshold value $\underline{\delta} > 0$ (where $\underline{\delta}$ is perhaps deemed a minimum clinically relevant effect size). In such cases, following the terminology used in [5], we will refer to the resulting GROW S-value $S^*_{\underline{\delta}} \coloneqq S^*_{\Theta_1(\underline{\delta})}$ as the *$\underline{\delta}$-GROW S-value*. For example, if we are testing whether the mean of a normal distribution is equal to zero or not, we may choose $\Theta_0 = \{0\}$ and $\Theta_1(\underline{\mu}) = \{\mu \in \mathbb{R} : |\mu| \geq \underline{\mu}\}$ based on a given threshold $\underline{\mu} > 0$, and refer to the resulting S-value $S^*_{\underline{\mu}} \coloneqq S^*_{\Theta_1(\underline{\mu})}$ as the $\underline{\mu}$-GROW S-value.

For a given significance level α, we can inspect the power of $S^*_{\underline{\delta}}$ across the *whole* of Θ_1 for each each $\underline{\delta} > 0$. Then, if there exists a $\underline{\delta} > 0$ such that

$$\forall \theta \in \Theta_1 \;\; \forall \underline{\delta}' \geq 0 \quad P_\theta(S^*_{\underline{\delta}} \geq 1/\alpha) \geq P_\theta(S^*_{\underline{\delta}'} \geq 1/\alpha), \tag{2.67}$$

this threshold value $\underline{\delta}$ is denoted by δ^* and referred to as a *uniformly most powerful threshold* (UMP threshold). Likewise, we call $S^*_{\delta^*}$ a *uniformly most powerful GROW S-value* (UMPG S-value) and the associated test $T_\alpha(S^*_{\mu^*})$ a *uniformly most powerful GROW test of significance level* α, or UMPGT(α) for short. Note that in some cases this terminology is somewhat misleading, since there may exist some other subset $\Theta_1' \subseteq \Theta_1$ not of the form $\Theta_1(\underline{\delta})$ such that the test $T_\alpha(S^*_{\Theta_1'})$ based on the GROW S-value $S^*_{\Theta_1'}$ is more powerful than $T_\alpha(S^*_{\delta^*})$. However, in the case of one-parameter exponential families, we will see that $T_\alpha(S^*_{\delta^*})$ is always more powerful than $T_\alpha(S^*_{\Theta_1'})$ for *any* GROW S-value $S^*_{\Theta_1'}$, meaning in that case the terminology 'UMPG S-value' is fully justified. This is shown in detail in Theorem 5.

Recall that in this thesis we will be concerned solely with one-sided tests since, as noted in [5], finding GROW S-values for two-sided tests produces extra difficulties. Now, as we saw in Theorem 1, GROW S-values are Bayes factors for very particular priors on the null and alternative parameter sets. Thus a test based on a GROW S-value is an instance of a Bayesian hypothesis test, in which the null hypothesis test is rejected iff a Bayes factor exceeds a given threshold. To find a UMPGT(α), we go via the concept of a uniformly most powerful Bayesian test (UMPBT), first introduced by Johnson [7]. The idea is that for a fixed parameter space Θ and null hypothesis $\mathcal{H}_0 : \theta \sim \pi_0$, there may exist an alternative hypothesis $\mathcal{H}_1 : \theta \sim \pi_1$ that maximises the probability that the associated Bayes factor

$$\mathrm{BF}_{10}(\mathbf{X}) \coloneqq \frac{p_{\pi_1}(\mathbf{X})}{p_{\pi_0}(\mathbf{X})} \tag{2.68}$$

exceeds a certain threshold γ *uniformly* across Θ. More precisely, Johnson gives the following definition.

Definition 3. A uniformly most powerful Bayesian test for evidence threshold $\gamma > 0$ in favor of the alternative hypothesis $\mathcal{H}_1$ against a fixed null hypothesis $\mathcal{H}_0$, denoted by UMPBT(γ), is a Bayesian hypothesis test with Bayes factor $\mathrm{BF}_{10}(\mathbf{X})$, where $\mathrm{BF}_{10}(\mathbf{X})$ satisfies the following inequality for any alternative hypotheses $\mathcal{H}_2 : \theta \sim \pi_2(\theta)$:

$$\forall \theta_t \in \Theta \;\; P_{\theta_t}(\mathrm{BF}_{10}(\mathbf{X}) > \gamma) \geq P_{\theta_t}(\mathrm{BF}_{20}(\mathbf{X}) > \gamma). \tag{2.69}$$

Johnson then states and proves the following theorem on the existence of a UMPBT(γ) in the case of exponential families with a one-sided test against a point null hypothesis. It shows that a UMPBT(γ) can be constructed by using a prior π_1 with very restricted support, usually a single point.

Theorem 4. *(Johnson's theorem) Assume that $x_1, \ldots, x_n$ are i.i.d. from an exponential family with a density (or p.m.f. in the case of discrete data) with canonical form*

$$p_\theta(x) = h(x) \exp\big(\eta(\theta)T(x) - A(\theta)\big), \tag{2.70}$$

where η is monotonic. Consider a one-sided test of a fixed point null hypothesis $\mathcal{H}_0 : \theta = \theta_0$ against an arbitrary alternative hypothesis. Let γ denote the evidence threshold for a UMPBT(γ). Define $g_{\gamma,n}(\theta, \theta_0)$ according to

$$g_{\gamma,n}(\theta, \theta_0) = \frac{\log \gamma + n(A(\theta) - A(\theta_0))}{\eta(\theta) - \eta(\theta_0)} \tag{2.71}$$

In addition, define u to be 1 or -1 according to whether η is monotonically increasing or decreasing, respectively, and define v to be either 1 or -1 according to whether the alternative hypothesis requires θ to be greater than or less than θ_0, respectively. Then a UMPBT(γ) can be obtained by restricting the support of π_1 to values of θ that belong to the set

$$\arg\min_{\theta} uvg_{\gamma,n}(\theta, \theta_0). \tag{2.72}$$

Thus finding a UMPBT(γ) reduces to minimizing $uvg_{\gamma,n}(\theta, \theta_0)$.

The above theorem is used in [5] to show the existence of a UMPGT(α). The key idea is that since GROW tests (with significance level α) are themselves Bayesian tests (with threshold $1/\alpha$), if a UMPBT($1/\alpha$) turns out to be a GROW test, it is automatically a UMPGT(α).[3] The authors then state that Johnson's theorem implies a UMPBT($1/\alpha$) can be found by solving $\mathrm{KL}(P_\theta||P_{\theta_0}) = (-\log\alpha)/n$ for θ. We now state a slightly modified version of their theorem and provide a proof.

Theorem 5. *Consider the setting of Johnson's theorem with parameter set Θ and hypotheses $\mathcal{H}_0 : \theta = \theta_0$ and $\mathcal{H}_1 : \theta \in \Theta_1$ where $\Theta_1 := [\theta_0, \infty) \cap \Theta$. Assume that A and η are differentiable, $\eta'(\theta) > 0$ for all θ and that, in the case of continuous random variables, we can take the derivative of $\int h(x) \exp\left(\eta(\theta)T(x) - A(\theta)\right) dx$ with respect to θ through the integral. For $\theta \geq \theta_0$ define*

$$d_n(\theta) := \mathrm{KL}(P_\theta^n||P_{\theta_0}^n), \tag{2.73}$$

where P_θ^n denotes the n-fold product measure of P_θ. Suppose d_n is continuous and strictly increasing in θ. Let $L := \lim_{\theta\to\infty} d_n(\theta)$. Then for all $\alpha \in (e^{-L}, 1)$ there exists a UMPGT(α), namely the test $T_\alpha(S^*_{\theta^*})$ based on the θ^*-GROW S-value $S^*_{\theta^*}$, where θ^* is the unique solution to $d_n(\theta^*) = -\log\alpha$. Further, the terminology UMPGT(α) is fully justified in this setting since $T_\alpha(S^*_{\theta^*})$ is more powerful than $T_\alpha(S^*_{\Theta_1'})$ based on GROW S-value $S^*_{\Theta_1'}$ for any alternative parameter set $\Theta_1' \subseteq \Theta_1$.

In order to prove the above theorem, we first need the following lemma.

Lemma 6. *Fix $\gamma > 0$ and suppose the conditions of the above theorem hold. Then $\frac{\partial}{\partial\theta} g_{\gamma,n}(\theta, \theta_0) = 0$ if and only if $d_n(\theta) = \log\gamma$.*

Proof. First, the Kullback-Leibler divergence can be simplified as follows.

$$\begin{aligned}
d_1(\theta) &= \int h(x) \exp\left(\eta(\theta)T(x) - A(\theta)\right) \log \frac{h(x)\exp\left(\eta(\theta)T(x) - A(\theta)\right)}{h(x)\exp\left(\eta(\theta_0)T(x) - A(\theta_0)\right)} \mathrm{d}x && (2.74)\\
&= \int h(x) \exp\left(\eta(\theta)T(x) - A(\theta)\right) \left((\eta(\theta) - \eta(\theta_0))T(x) + A(\theta_0) - A(\theta)\right) \mathrm{d}x && (2.75)\\
&= (\eta(\theta) - \eta(\theta_0)) \int T(x)h(x) \exp\left(\eta(\theta)T(x) - A(\theta)\right) \mathrm{d}x && (2.76)\\
&\quad + (A(\theta_0) - A(\theta)) \int h(x) \exp\left(\eta(\theta)T(x) - A(\theta)\right) \mathrm{d}x && (2.77)\\
&= (\eta(\theta) - \eta(\theta_0))\mathbf{E}_\theta[T(X)] + A(\theta_0) - A(\theta). && (2.78)
\end{aligned}$$

Then, since the KL-divergence is additive for independent distributions, we have

$$d_n(\theta) = n\Big[(\eta(\theta) - \eta(\theta_0))\mathbf{E}_\theta[T(X)] + A(\theta_0) - A(\theta)\Big] \tag{2.79}$$

Next, since distributions integrate to one, we have that for all θ

$$\int h(x) \exp\left(\eta(\theta)T(x) - A(\theta)\right) \mathrm{d}x = 1. \tag{2.80}$$

[3] The UMPGT(α) will in fact be based on a GROW S-value S^* for which the optimal prior W_1^* puts all its mass on the value of θ closest to θ_0. This makes it straightforward to implement.

Hence, by taking a derivative with respect to θ through the integral, we have

$$\int h(x)(\eta'(\theta)T(x) - A'(\theta)) \exp\left(\eta(\theta)T(x) - A(\theta)\right) \mathrm{d}x = 0 \tag{2.81}$$

$$\implies \eta'(\theta) \int T(x)h(x) \exp\left(\eta(\theta)T(x) - A(\theta)\right) \mathrm{d}x = A'(\theta) \int h(x) \exp\left(\eta(\theta)T(x) - A(\theta)\right) \mathrm{d}x \tag{2.82}$$

$$\implies \eta'(\theta)\mathbf{E}_\theta[T(X)] = A'(\theta). \tag{2.83}$$

Using the quotient rule for differentiation and then substituting the above line for $A'(\theta)$ gives

$$\frac{\partial}{\partial\theta} g_{\gamma,n}(\theta, \theta_0) = \frac{(\eta(\theta) - \eta(\theta_0))nA'(\theta) - \big[\log\gamma + n(A(\theta) - A(\theta_0))\big]\eta'(\theta)}{(\eta(\theta) - \eta(\theta_0))^2} \tag{2.84}$$

$$= \frac{(\eta(\theta) - \eta(\theta_0))n\eta'(\theta)\mathbf{E}_\theta[T(X)] - \big[\log\gamma + n(A(\theta) - A(\theta_0))\big]\eta'(\theta)}{(\eta(\theta) - \eta(\theta_0))^2} \tag{2.85}$$

$$= \eta'(\theta)\frac{n\big[(\eta(\theta) - \eta(\theta_0))\mathbf{E}_\theta[T(X)] + A(\theta_0) - A(\theta)\big] - \log\gamma}{(\eta(\theta) - \eta(\theta_0))^2} \tag{2.86}$$

$$= \eta'(\theta)\frac{d_n(\theta) - \log\gamma}{(\eta(\theta) - \eta(\theta_0))^2}. \tag{2.87}$$

The result now follows since by the assumption that $\eta'(\theta) > 0$ for all θ. □

We can now prove Theorem 5.

Proof (of Theorem 5). Since $d_n(\theta_0) = \mathrm{KL}(P^n_{\theta_0} || P^n_{\theta_0}) = 0$, and d_n is continuous and strictly increasing, d_n takes on every value in $[0, L)$ exactly once. Thus, since $e^{-L} < \alpha < 1$ and so $0 < -\log\alpha < L$, there is then a unique solution θ^* to $d_n(\theta^*) = -\log\alpha$. Defining $\gamma = 1/\alpha$, by Lemma 6 we then have

$$\frac{\partial}{\partial\theta} g_{\gamma,n}(\theta^*, \theta_0) = 0, \tag{2.88}$$

so that, by Johnson's Theorem (Theorem 4), a UMPBT(γ) can be obtained for $\gamma = 1/\alpha$ by choosing a prior π_1 with support a subset of $\arg\min_\theta g_{\gamma,n}(\theta, \theta_0)$. We now show that $\arg\min_\theta g_{\gamma,n}(\theta, \theta_0)$ is a single point, and in fact

$$\theta^* = \arg\min_\theta g_{\gamma,n}(\theta, \theta_0). \tag{2.89}$$

Since the null hypothesis is simple, this implies that the UMPBT(γ) uses point mass priors π_1 and π_0 on θ^* and θ_0 respectively. Thus it uses the Bayes factor

$$\mathrm{BF}_{10}(\mathbf{X}) = \frac{p_{\pi_1}(\mathbf{X})}{p_{\pi_0}(\mathbf{X})} = \frac{p_{\theta^*}(\mathbf{X})}{p_{\theta_0}(\mathbf{X})} \tag{2.90}$$

and rejects the null hypothesis iff $p_{\theta^*}(\mathbf{X})/p_{\theta_0}(\mathbf{X}) \geq \gamma$.

To show (2.89), note that since A and η are assumed to be differentiable, $g_{\gamma,n}$ is partially differentiable with respect to θ except at $\theta = \theta_0$. Therefore its minimum is attained either at the unique stationary point θ^*, the endpoint θ_0 or at infinity (note $\theta^* = \theta_0$ is impossible since it would imply $d(\theta^*) = d(\theta_0) = 0$ whereas in fact $d(\theta^*) = -\log\alpha > 0$).

We first exclude the possibility that the minimum is attained at θ_0 by showing that $g_{\gamma,n}(\theta, \theta_0) \to \infty$ as $\theta \to \theta_0$. Since A and η are differentiable, they are continuous. Thus for θ close enough to θ_0, we have $|A(\theta) - A(\theta_0)| \leq 1/(2n\log\gamma)$ and so

$$g_{\gamma,n}(\theta, \theta_0) \coloneqq \frac{\log\gamma + n(A(\theta) - A(\theta_0))}{\eta(\theta) - \eta(\theta_0)} \tag{2.91}$$

$$\geq \frac{1/2\log\gamma}{\eta(\theta) - \eta(\theta_0)}, \tag{2.92}$$

which is positive (as η is strictly increasing) and tends to ∞ as $\theta \to \theta_0$.

Second, we exclude the possibility that the minimum is attained at infinity by showing that $g_{\gamma,n}$ is strictly increasing for $\theta \geq \theta^*$. This follows immediately by noting that

$$\frac{\partial}{\partial\theta} g_{\gamma,n}(\theta, \theta_0) = \eta'(\theta) \frac{d_n(\theta) - d_n(\theta^*)}{(\eta(\theta) - \eta(\theta_0))^2} > 0 \quad \text{for } \theta > \theta^*, \tag{2.93}$$

since $\eta'(\theta) > 0$ for all θ and d_n is strictly increasing. Thus (2.89) holds and the UMPBT(γ) rejects $\mathcal{H}_0$ iff $p_{\theta^*}(\mathbf{X})/p_{\theta_0}(\mathbf{X}) \geq \gamma = 1/\alpha$.

Now, since we are dealing with a one-parameter exponential family, we can apply Proposition 3 to see that for any threshold value $\underline{\theta} > \theta_0$, the $\underline{\theta}$-GROW S-value $S^*_{\underline{\theta}} := S^*_{\Theta_1(\underline{\theta})}$ is given by

$$S^*_{\underline{\theta}}(\mathbf{X}) = \frac{p_{\underline{\theta}}(\mathbf{X})}{p_{\theta_0}(\mathbf{X})}. \tag{2.94}$$

In particular, we find that the θ^*-GROW test $T_\alpha(S^*_{\theta^*})$ based on the θ^*-GROW S-value $S^*_{\theta^*}$ rejects $\mathcal{H}_0$ iff $p_{\theta^*}(\mathbf{X})/p_{\theta_0}(\mathbf{X}) \geq 1/\alpha$ and therefore coincides with the UMPBT(γ) given in (2.90). Finally, recall Theorem 1, in particular (2.25), which shows that the GROW S-value $S^*_{\Theta_1'}$ based on *any* $\Theta_1' \subseteq \Theta_1$ is given by a Bayes factor, so that all GROW tests $T_\alpha(S^*_{\Theta_1'})$ are instances of Bayesian tests. Thus, since the uniformly most powerful Bayesian test UMPBT(γ) is in fact equal to the θ^*-GROW test $T_\alpha(S^*_{\theta^*})$, we see that $T_\alpha(S^*_{\theta^*})$ is the most powerful of *all* the GROW tests, so that the terminology 'UMP GROW test for significance level α is fully justified in this case. □

The remainder of this thesis is devoted to finding GROW S-values and UMPGT(α)'s (when they exist) in the specific case of 2×2 contingency tables.

Chapter 3

2×2 Contingency tables

In this short chapter we introduce 2×2 contingency tables. After providing some brief context in section 3.1, section 3.2 fixes the notation that will be used in the rest of the thesis. Section 3.3 introduces Fisher's exact test, which is commonly used when evaluating 2×2 tables and will be used as a benchmark by which to measure the power of the S-values we construct in subsequent chapters. In section 3.4 we clarify the difference between 'conditional' and 'unconditional' S-values. Finally, section 3.5 discusses the relation of this thesis to previous work.

3.1 Context

Suppose a number of categorical random variables are measured across a sample of individuals. The resulting frequencies can then be represented in a contingency table, which may be used to investigate any dependencies that may exist between the variables. For the purposes of this thesis, we will be interested only in 2×2 contingency tables, which are the result of measuring two variables, where each variable can take only two values. Although simple, this scenario is very common, for example in clinical trials.

Suppose one wants to determine whether a particular treatment improves recovery or increases the chance of survival. After choosing a (perhaps arbitrary) cut-off point, recovery or survival may be modelled as a binary random variable. If the subset of the patients who receive the treatment (as opposed to the placebo) are selected randomly, the resulting data can be used in a causal investigation of the efficacy of the treatment.

We now outline the mathematical details. Although what follows is clearly not specific to clinical trials, for clarity we will continue to use words such as 'patient' and 'recover'.

3.2 Mathematical setup

Suppose we have a sample of n patients, n_a of which are randomly selected to receive the placebo, while the remaining n_b receive the treatment. Let θ_a be the probability that an individual in the placebo group will recover and θ_b be the probability that an individual in the treatment group will recover. Let N_{a1} and N_{b1} be the numbers of patients that recover in the placebo and treatment groups respectively. Likewise, let N_{a0} and N_{b0} be the number of patients in the two groups that do not recover. According to our model, we have

$$N_{a1} \sim \text{Bin}(n_a, \theta_a) \quad \text{and} \quad N_{b1} \sim \text{Bin}(n_b, \theta_b), \tag{3.1}$$

where N_{a1} and N_{b1} are independent. Throughout this thesis we will abbreviate $Z = (N_{a1}, N_{b1})$.

Suppose after the trial has been completed, we have data $(n_{a1}, n_{a0}, n_{b1}, n_{b0})$, summarized in the table below. We want to infer—for a one-sided test—whether $\theta_b = \theta_a$ (the treatment does not improve the chance of recovery) or $\theta_b > \theta_a$ (the treatment improves the chance of recovery). We take parameter sets $\Theta_0 = \{(\theta_0, \theta_0) : \theta_0 \in [0,1]\}$ and $\Theta_1 = \{(\theta_a, \theta_b) \in [0,1]^2 : \theta_b > \theta_a\}$. The null and alternative hypotheses are then $\mathcal{H}_i = \{P_\theta : \theta \in \Theta_i\}$, for $i = 0, 1$ respectively.

Group \ Recovery	0 (No)	1 (Yes)	Total
a (Placebo)	n_{a0}	n_{a1}	n_a
b (Treatment)	n_{b0}	n_{b1}	n_b
Total	n_0	n_1	n

Since the practitioner selects n_a and n_b, these are assumed to be known. Note that if we know $N_1 = n_1$ then there is only one degree of freedom remaining. Indeed, given the value of just one of N_{a1}, N_{a0}, N_{b1} or N_{b0}, we can infer the other values using the marginal sums. Therefore, given the marginal sums, the data can be summarized simply by N_{b1} (picked arbitrarily). In chapter four we will consider the case in which the value of N_1 is known in advance. While this is unrealistic in the context we have just given, there are situations in which this is possible. Chapter five then considers the case when N_1 is treated as a random variable whose value is not known in advance.

3.3 Fisher's exact test

Many different statistical tests have been proposed for analyzing 2×2 contingency tables. Examples include Pearson's chi-squared test [8], Fisher's exact test [4], Boschloo's test [2] and Barnard's test [1]. Since Fisher's exact test bears some resemblance to the 'conditional S-values' we construct in the next chapter, we will discuss it here.

If the null hypothesis were true—where θ_a and θ_b both equal p say—patients receiving the treatment would face the same chance of recovery as those receiving the placebo. We would therefore expect the proportion of patients that recover to be approximately the same in both groups. For any fixed n_1, we would expect $N_{a1} \approx n_a n_1/n$ and $N_{b1} \approx n_b n_1/n$. Therefore if N_{b1} is 'large', we should interpret this as evidence against the null. To formalize this, we need to be precise about what we mean by 'large'. A key insight is that under the null hypothesis N_1 is a sufficient statistic for p. This implies that, under the null hypothesis, the distribution of $(N_{a1}, N_{b1})|N_1 = n_1$—and hence the distribution of $N_{b1}|N_1 = n_1$—does not depend on the true parameter p.

Thus, Fisher's exact test finds the distribution of N_{b1} when conditioned on the actually observed value $N_1 = n_1$ and, for a one-sided test with significance level α and observed values n_{b1} and n_1, rejects the null hypothesis if and only if

$$P(N_{b1} \geq n_{b1} | N_1 = n_1) \leq \alpha. \tag{3.2}$$

Now, for a given value of n_1, the conditional distribution of N_{b1} under the null hypothesis is the hypergeometric distribution. More precisely,

$$N_{b1} \mid N_1 = n_1 \sim \text{hypg}(n, n_b, n_1) \quad \text{where} \quad P(N_{b1} = n_{b1} \mid N_1 = n_1) = \frac{\binom{n_a}{n_1 - n_{b1}}\binom{n_b}{n_{b1}}}{\binom{n}{n_1}}, \tag{3.3}$$

Therefore Fisher's exact test rejects the null if and only if

$$\sum_{n'_{b1}=n_{b1}}^{\min\{n_b,n_1\}} \frac{\binom{n_a}{n_1 - n_{b1'}}\binom{n_b}{n'_{b1}}}{\binom{n}{n_1}} \leq \alpha. \tag{3.4}$$

To see that the conditional distribution of N_{b1} given $N_1 = n_1$ takes the above form, let P_{θ_0} be the distribution of $Z = (N_{a1}, N_{b1})$ given any true parameter $\theta_0 = (p, p) \in \Theta_0$. Then $N_{a1} \sim \text{Bin}(n_a, p)$ and

$N_{b1} \sim \text{Bin}(n_b, p)$ are independent and, for any fixed values n_{b1} and n_1, we have

$$
\begin{aligned}
P_{\theta_0}(N_{b1} = n_{b1} \mid N_1 = n_1) &:= P_{\theta_0}(N_{b1} = n_{b1} \mid N_1 = n_1) && (3.5)\\
&= \frac{P_{\theta_0}(N_{b1} = n_{b1}, N_1 = n_1)}{P_{\theta_0}(N_1 = n_1)} && (3.6)\\
&= \frac{P_{\theta_0}(N_{a1} = n_1 - n_{b1}, N_{b1} = n_{b1})}{P_{\theta_0}(N_1 = n_1)} && (3.7)\\
&= \frac{P_{\theta_0}(N_{a1} = n_1 - n_{b1}) P_{\theta_0}(N_{b1} = n_{b1})}{P_{\theta_0}(N_1 = n_1)} && (3.8)\\
&= \frac{\binom{n_a}{n_1 - n_{b1}} p^{n_1 - n_{b1}} (1-p)^{n_a - (n_1 - n_{b1})} \binom{n_b}{n_{b1}} p^{n_{b1}} (1-p)^{n_b - n_{b1}}}{\binom{n}{n_1} p^{n_1} (1-p)^{n - n_1}} && (3.9)\\
&= \frac{\binom{n_a}{n_1 - n_{b1}} \binom{n_b}{n_{b1}}}{\binom{n}{n_1}}, && (3.10)
\end{aligned}
$$

which is independent of θ_0 and is in fact the p.m.f. of the hypergeometric distribution with parameters n, n_b, n_1.

Fisher's exact test is exact in the sense that since no asymptotic approximation is been made (as is common in other tests), and instead the true distribution of N_{b1} is used, the test's true significance level is equal to α.[1] While this is based on the assumption that N_1 really is fixed in advance—which of course in the context of clinical trials it is not, since we do not know in advance the total number of patients that will recover—the exactness in fact carries over to the case when N_1 is not known in advance by the law of total probability. More precisely, for any $\theta_0 \in \Theta_0$, we have

$$
\begin{aligned}
P_{\theta_0}(\text{FET rejects } \mathcal{H}_0) &= \sum_{n_1=0}^{n} P_{\theta_0}(\text{FET rejects } \mathcal{H}_0 \mid N_1 = n_1) P_{\theta_0}(N_1 = n_1) && (3.11)\\
&\approx \sum_{n_1=0}^{n} \alpha P_{\theta_0}(N_1 = n_1) && (3.12)\\
&= \alpha. && (3.13)
\end{aligned}
$$

Nevertheless, there is some controversy around conditioning on N_1 [3]. There are three cases; N_1 really is known in advance; N_1 is not known in advance, but we analyse the data as though it were; and N_1 is not known in advance and we do not treat is as though it were. In the following section, we briefly outline how we will construct S-values in these three cases.

3.4 Conditional and unconditional S-values

First, if N_1 truly is known in advance, then the data can be summarized by N_{b1}. We will see in the next chapter that N_{b1} then follows Fisher's noncentral hypgergeometric distribution, which is parametrized by the odds ratio ψ, defined as

$$
\psi = \frac{\theta_b}{1 - \theta_b} \frac{1 - \theta_a}{\theta_a} \tag{3.14}
$$

This is a one-dimensional exponential family, so we can use the relevant results from [5] to find the GROW and UMPG S-values, which will then be referred to as *conditional S-values*.

Second, suppose N_1 is not known in advance, but we analyze the data as though the discovered value of N_1 were in fact known from the beginning. More precisely, we can calculate the conditional S-value using the value of N_1 discovered in the course of the experiment. This is similar to the methodology of Fisher's exact test and in fact this statistical test shares the property of Fisher's exact test that the type I error guarantee is preserved even after taking expectations over N_1. When using the conditional S-value in this way, we still refer to it as the conditional S-value, but we will evaluate its growth and

[1] Strictly speaking, since we are dealing with discrete random variables, the true significance may be slightly less than α.

power in the *unconditional* setting, where N_1 is unknown. The details of these first two choices can be found in the following chapter.

Finally, suppose N_1 is not known in advance and we do not treat it as though it were. In this case we no longer have an exponential family and so we resort to calculating (or approximating) the JIP. Recall that this requires restricting the parameter sets so that they are positively separated. At this point a number of choices can be made. For example, if we are interested in the risk difference $\delta = \delta(\theta_a, \theta_b) \coloneqq \theta_b - \theta_a$, we may pick a threshold value of this parameter that is deemed 'substantial'. We may therefore choose

$$\Theta_0 = \{(\theta_a, \theta_b) \in [0,1]^2 : \theta_a = \theta_b\} \quad \text{and} \quad \Theta_1(\delta) = \{(\theta_a, \theta_b) \in [0,1]^2 : \theta_b > \theta_a + \delta\}. \tag{3.15}$$

We can then calculate the JIP to get the GROW S-value S^*_δ. Thus for each threshold value of δ, we have an associated GROW S-value S^*_δ. Out of these, we can then choose the uniformly most powerful one (if it exists), $S^*_{\delta_{\text{UMP}}}$. These S-values are referred to as *unconditional S-values*. The full details of the different choices to be made and the resulting S-values can be found in chapter 5.

3.5 Relation to previous work

This thesis is of course heavily based on the methodology and results developed by Grünwald et al. in [5]. It also builds on the work of Turner [12], who conducted an empirical investigation of some of the unconditional GROW and UMPG S-values we explore here, bolstered by some theoretical work in specific cases. We extend the work of Turner by investigating other choices for restricting Θ_1. We prove a number of theoretical results that give closed formulas for the GROW S-values in some cases and in other cases give methods for finding the GROW S-values that are dramatically quicker than approximating the JIP directly. While in some cases it is still necessary to approximate the JIP directly (and this is computationally intensive), we have found a simplification of the calculation in the specific case of 2×2 contingency tables that speeds up the calculation by around an order of magnitude. Our calculation of a conditional UMPG S-value in the subsequent chapter is heavily reliant on a theorem of Johnson [7], extended by Grünwald et al. [5].

Chapter 4

Safe tests for 2×2 contingency tables when N_1 is known

In this chapter we construct safe tests for 2×2 contingency tables in the case where N_1 is known in advance. We shall call these tests *conditional* safe tests. In fact, as we will see in Proposition 7, it is also valid to use these tests in the *unconditional* setting, namely when N_1 is not known in advance, simply by substituting the discovered value n_1. This is valid in the sense that the type I error guarantee is preserved.

In section 4.1 we show that the distribution of N_{b1} when conditioned on N_1 is equal to Fisher's noncentral hypergeometric distribution (fnchypg), of which the hypergeometric distribution is a special case. The parameters of this distribution are n, n_b, n_1 and ψ, where ψ is the odds ratio. Since the fnchypg distributions form an exponential family, the methods discussed in chapter 2 can be used to produce GROW and UMPG S-values, which we will call the *conditional GROW* and *conditional UMPG* S-values. We provide a closed formula for the conditional GROW S-value in section 4.2 and an implicit expression for the UMPG S-value (when it exists) in section 4.3. The growth and power of these conditional S-values in the conditional setting (when N_1 is known in advance) are evaluated in section 4.4, while the evaluation in the unconditional setting (when N_1 is not known in advance) is postponed to chapter 7.

4.1 Fisher's noncentral hypergeometric distribution

Recall that under the null hypothesis $\theta = (\theta_a, \theta_b) \in \Theta_0$, the distribution of N_{b1} conditioned on N_1 follows the hypergeometric distribution, namely

$$N_{b1} \mid N_1 = n_1 \sim \text{hypg}(n, n_b, n_1) \quad \text{where} \quad P(N_{b1} = n_{b1} \mid N_1 = n_1) = \frac{\binom{n_a}{n_1 - n_{b1}}\binom{n_b}{n_{b1}}}{\binom{n}{n_1}}, \tag{4.1}$$

We now show that this is in fact a special case, in that for an arbitrary $\theta = (\theta_a, \theta_b) \in [0,1]^2$, the distribution of N_{b1} conditioned on $N_1 = n_1$ is equal to Fisher's noncentral hypergeometric distribution, which is a generalization of the hypergeometric distribution which only depends on θ via the one-dimensional parameter $\psi(\theta_a, \theta_b)$, referred to as the odds ratio. We are thus justified in writing the conditional distribution of N_{b1} as P_ψ. More precisely, we have

$$N_{b1} \mid N_1 = n_1 \sim \text{fnchypg}(n, n_b, n_1, \psi), \quad \text{where} \quad \psi = \psi(\theta_a, \theta_b) = \frac{\theta_b}{1-\theta_b}\frac{1-\theta_a}{\theta_a} \tag{4.2}$$

and[1]

$$P_\psi(N_{b1} = n_{b1}) \coloneqq P_{\theta_a,\theta_b}(N_{b1} = n_{b1} \mid N_1 = n_1) = \frac{\binom{n_a}{n_1 - n_{b1}}\binom{n_b}{n_{b1}}\psi^{n_{b1}}}{\sum_{n'_{b1}=\max\{0, n_1 - n_a\}}^{\min\{n_b, n_1\}} \binom{n_a}{n_1 - n'_{b1}}\binom{n_b}{n'_{b1}}\psi^{n'_{b1}}}. \tag{4.3}$$

[1] Although we are using P for both the conditional distribution of N_{b1} given N_1 and a particular ψ, and the unconditional distribution of $Z = (N_{a1}, N_{b1})$ for a particular θ, no confusion should arise so long as one of the symbols ψ or θ is present. If values are substituted, it should still be clear since ψ is a scalar while θ is a tuple.

Recall that N_{a1} and N_{b1} are independent and have distributions $N_{a1} \sim \text{Bin}(n_a, \theta_a)$ and $N_{b1} \sim \text{Bin}(n_b, \theta_b)$. If we define

$$\psi_a = \frac{\theta_a}{1-\theta_a} \quad \text{and} \quad \psi_b = \frac{\theta_b}{1-\theta_b}, \tag{4.4}$$

so that $\psi = \psi_b/\psi_a$, we can derive (4.3) as follows

$$P_{\theta_a,\theta_b}(N_{b1} = n_{b1} \mid N_1 = n_1) = \frac{P_{\theta_a,\theta_b}(N_{b1} = n_{b1}, N_1 = n_1)}{P_{\theta_a,\theta_b}(N_1 = n_1)} \tag{4.5}$$

$$= \frac{P_{\theta_a,\theta_b}(N_{a1} = n_1 - n_{b1}, N_{b1} = n_{b1})}{\sum_{n'_{b1}=\max\{0,n_1-n_a\}}^{\min\{n_b,n_1\}} P_{\theta_a,\theta_b}(N_{a1} = n_1 - n'_{b1}, N_{b1} = n'_{b1})} \tag{4.6}$$

$$= \frac{P_{\theta_a,\theta_b}(N_{a1} = n_1 - n_{b1}) P_{\theta_a,\theta_b}(N_{b1} = n_{b1})}{\sum_{n'_{b1}=\max\{0,n_1-n_a\}}^{\min\{n_b,n_1\}} P_{\theta_a,\theta_b}(N_{a1} = n_1 - n'_{b1}) P_{\theta_a,\theta_b}(N_{b1} = n'_{b1})} \tag{4.7}$$

$$= \frac{\binom{n_a}{n_1-n_{b1}} \theta_a^{n_1-n_{b1}} (1-\theta_a)^{n_{a0}} \binom{n_b}{n_{b1}} \theta_b^{n_{b1}} (1-\theta_b)^{n_{b0}}}{\sum_{n'_{b1}=\max\{0,n_1-n_a\}}^{\min\{n_b,n_1\}} \binom{n_a}{n_1-n'_{b1}} \theta_a^{n_1-n'_{b1}} (1-\theta_a)^{n'_{a0}} \binom{n_b}{n'_{b1}} \theta_b^{n'_{b1}} (1-\theta_b)^{n'_{b0}}} \tag{4.8}$$

$$= \frac{\binom{n_a}{n_1-n_{b1}} \psi_a^{n_1-n_{b1}} (1-\theta_a)^{n_a} \binom{n_b}{n_{b1}} \psi_b^{n_{b1}} (1-\theta_b)^{n_b}}{\sum_{n'_{b1}=\max\{0,n_1-n_a\}}^{\min\{n_b,n_1\}} \binom{n_a}{n_1-n'_{b1}} \psi_a^{n_1-n'_{b1}} (1-\theta_a)^{n_a} \binom{n_b}{n'_{b1}} \psi_b^{n'_{b1}} (1-\theta_b)^{n_b}} \tag{4.9}$$

$$= \frac{\binom{n_a}{n_1-n_{b1}} \binom{n_b}{n_{b1}} \psi_a^{n_1-n_{b1}} \psi_b^{n_{b1}}}{\sum_{n'_{b1}=\max\{0,n_1-n_a\}}^{\min\{n_b,n_1\}} \binom{n_a}{n_1-n'_{b1}} \binom{n_b}{n'_{b1}} \psi_a^{n_1-n'_{b1}} \psi_b^{n'_{b1}}} \tag{4.10}$$

$$= \frac{\binom{n_a}{n_1-n_{b1}} \binom{n_b}{n_{b1}} \psi_a^{n_1-n_{b1}+n_{b1}} \psi^{n_b 1}}{\sum_{n'_{b1}=\max\{0,n_1-n_a\}}^{\min\{n_b,n_1\}} \binom{n_a}{n_1-n'_{b1}} \binom{n_b}{n'_{b1}} \psi_a^{n_1-n'_{b1}+n'_{b1}} \psi^{n'_b 1}} \tag{4.11}$$

$$= \frac{\binom{n_a}{n_1-n_{b1}} \binom{n_b}{n_{b1}} \psi^{n_{b1}}}{\sum_{n'_{b1}=\max\{0,n_1-n_a\}}^{\min\{n_b,n_1\}} \binom{n_a}{n_1-n'_{b1}} \binom{n_b}{n'_{b1}} \psi^{n'_{b1}}}. \tag{4.12}$$

The sum introduced in the denominator is simply a sum over all possible values of n'_{b1}. This sum has a more complicated form than might first be supposed. This is because the lowest and highest values n'_{b1} can take depend not just on the size of group b but also on the size of group a. More precisely, the lowest value n'_{b1} can take is when as many 1's as possible are in group a. If $n_a \geq n_1$, then we can fit all the 1's in group a. However, if $n_a < n_1$, then we can only fit at most n_a, leaving $n_1 - n_a$ in group b. Thus the smallest value n'_{b1} can take is $\max\{0, n_1 - n_a\}$. Similar reasoning gives $\min\{n_b, n_1\}$ as the upper limit of the sum.

In the rest of the thesis, we will let K denote the set of possible values of n_{b1} for a given n_1, namely

$$K \coloneqq \{k \in \mathbb{N} : \max\{0, n_1 - n_a\} \leq k \leq \min\{n_b, n_1\}\}, \tag{4.13}$$

where $\mathbb{N}$ is the nonnegative integers (which includes 0). Further, for $k \in K$, we will use the abbreviation

$$c_k \coloneqq \binom{n_a}{n_1 - k}\binom{n_b}{k}. \tag{4.14}$$

We can then write (4.12) more concisely as

$$P_\psi(N_{b1} = n_{b1} \mid N_1 = n_1) = \frac{c_{n_{b1}} \psi^{n_{b1}}}{\sum_{k=\max\{0,n_1-n_a\}}^{\min\{n_b,n_1\}} c_k \psi^k}. \tag{4.15}$$

As a sanity check, we can confirm that the denominator is never zero by checking that $K \neq \emptyset$ and

all the binomial coefficients are valid. We have

$$\begin{aligned} K \neq \emptyset &\iff \min\{n_b, n_1\} \geq \max\{0, n_1 - n_a\} && (4.16)\\ &\iff n_b, n_1 \geq \max\{0, n_1 - n_a\} && (4.17)\\ &\iff n_b, n_1 \geq 0 \quad \text{and} \quad n_b, n_1 \geq n_1 - n_a && (4.18)\\ &\iff n_b, n_1 \geq 0 \quad \text{and} \quad n_a + n_b \geq n_1 \quad \text{and} \quad n_1 \geq n_1 - n_a && (4.19)\\ &\iff n_b, n_1 \geq 0 \quad \text{and} \quad n \geq n_1 \quad \text{and} \quad n_a \geq 0 && (4.20)\\ &\iff n_a, n_b \geq 0 \quad \text{and} \quad 0 \leq n_1 \leq n. && (4.21) \end{aligned}$$

Clearly then for all sensible values of (n, n_1, n_b) the sum contains at least one term. In order to do statistical inference however, K must contain at least *two* values, otherwise N_{b1} is constant and nothing can be deduced from knowing its value. Now $|K| \geq 2$ iff all the inequalities above are strict, namely $n_a, n_b > 0$ and $0 < n_1 < n$. While we will always assume $n_a, n_b > 0$, it may occur that $n_1 = 0$ or $n_1 = n$. We will make it clear in this thesis when we are excluding such cases. Finally, we see that

$$\begin{aligned} \max\{0, n_1 - n_a\} \leq k \leq \min\{n_b, n_1\} &\implies 0, n_1 - n_a \leq k \quad \text{and} \quad k \leq n_b, n_1 && (4.22)\\ &\implies 0 \leq k \leq n_b \quad \text{and} \quad n_1 - n_a \leq k \leq n_1 && (4.23)\\ &\implies 0 \leq k \leq n_b \quad \text{and} \quad 0 \leq n_1 - k \leq n_a, && (4.24) \end{aligned}$$

and so all the binomial coefficients are valid for sensible values of (n, n_1, n_b).

Since the conditional distribution of N_{b1} given N_1 depends on θ only via the one-dimensional parameter ψ, we may reparametrize the model in terms of ψ. Thus the null parameter set becomes

$$\Psi_0 \coloneqq \{\psi(\theta_0) : \theta_0 \in \Theta_0\} = \left\{ \frac{p}{1-p} \frac{1-p}{p} : p \in [0,1] \right\} = \{1\}, \tag{4.25}$$

which is a singleton. Further, since $x \mapsto \frac{x}{1-x}$ is a strictly increasing function on $[0, 1)$, we have

$$\theta_a > \theta_b \iff \psi_a > \psi_b \iff \psi > 1. \tag{4.26}$$

Therefore, if we are performing a one sided test with parameter set $\Theta_1 = \{(\theta_a, \theta_b) \in [0,1]^2 : \theta_b > \theta_a\}$, we have

$$\Psi_1 \coloneqq \{\psi(\theta_1) : \theta_1 \in \Theta_1\} = (1, \infty). \tag{4.27}$$

With this parametrization, Ψ_0 is a singleton and so, as we saw above, for any prior $W_1 \in \mathcal{W}(\Psi_1)$ the likelihood ratio

$$S_{W_1}(N_{b1}) \coloneqq \frac{P_{W_1}(N_{b1})}{P_1(N_{b1})} \tag{4.28}$$

is an S-value, where the marginal P_{W_1} is defined by

$$P_{W_1}(N_{b1}) \coloneqq \mathbf{E}_{\psi \sim W_1}[P_\psi(N_{b1})]. \tag{4.29}$$

In particular, for a prior $W_1 = \delta\{\psi\}$ that is a point mass on some ψ, we have

$$\begin{aligned} S_\psi(N_{b1}) &\coloneqq S_{\delta\{\psi\}}(N_{b1}) && (4.30)\\ &= \frac{P_\psi(N_{b1})}{P_1(N_{b1})} && (4.31)\\ &= \frac{\binom{n_a}{n_1 - n_{b1}}\binom{n_b}{n_{b1}}\psi^{n_{b1}}}{\sum_{n'_{b1}=\max\{0,n_1-n_a\}}^{\min\{n_b,n_1\}} \binom{n_a}{n_1-n'_{b1}}\binom{n_b}{n'_{b1}}\psi^{n'_{b1}}} \left(\frac{\binom{n_a}{n_1-n_{b1}}\binom{n_b}{n_{b1}}}{\binom{n}{n_1}} \right)^{-1} && (4.32)\\ &= \frac{\binom{n}{n_1}\psi^{n_{b1}}}{\sum_{n'_{b1}=\max\{0,n_1-n_a\}}^{\min\{n_b,n_1\}} \binom{n_a}{n_1-n'_{b1}}\binom{n_b}{n'_{b1}}\psi^{n'_{b1}}}. && (4.33) \end{aligned}$$

We will now see that the GROW and UMPG S-values are all of this form.

Now it is straightforward to show that any test statistic that is an S-value in the conditional case is also an S-value in the unconditional case. This follows simply by taking expectations over N_1. The following proposition spells this out in more detail and in a more general setting.

Proposition 7. *Let $Z = (X, Y)$ be a random variable, where $Z \sim P_\theta$ for some $\theta \in \Theta_0 \cup \Theta_1$. For any value y and any $\theta_0 \in \Theta_0$, let $P^X_{\theta_0}(\cdot \mid y)$ denote the conditional distribution of X given $Y = y$. Suppose $S(X|Y)$ is a conditional S-value, in the sense that for any value of y known in advance, $S(X|y)$ is an S-value. More precisely, for all values of y we have*

$$\forall \theta_0 \in \Theta_0 \quad \mathbf{E}_{X \sim P^X_{\theta_0}(\cdot|y)}[S(X|y)] \leq 1. \tag{4.34}$$

Then $S(X|Y)$ is also an unconditional S-value, in the sense that

$$\forall \theta_0 \in \Theta_0 \quad E_{Z \sim P_{\theta_0}}[S(X|Y)] \leq 1. \tag{4.35}$$

Proof. This follows simply by taking the expectation over Y. For any $\theta_0 \in \Theta_0$, let $P^Y_{\theta_0}$ be the marginal distribution of Y when $Z \sim P_{\theta_0}$. Then, by the law of total expectation, we have

$$E_{Z \sim P_{\theta_0}}[S(X|Y)] = \mathbf{E}_{Y \sim P^Y_{\theta_0}}\left[\mathbf{E}_{X \sim P^X_{\theta_0}(\cdot|Y)}[S(X|Y)]\right] \tag{4.36}$$

$$\leq \mathbf{E}_{Y \sim P^Y_{\theta_0}}[1] \tag{4.37}$$

$$= 1. \tag{4.38}$$

□

We therefore see that $S_{W_1}(N_{b1})$ is both a conditional and an unconditional S-value for any prior W_1. We will now find which W_1 gives the S-value with the highest growth rate in the conditional case, namely the *conditional GROW S-value*, using the results for exponential families found in [5].

4.2 The conditional GROW S-values

To avoid the GROW S-value being degenerate, we must pick $\Psi_1' \subseteq \Psi_1$ strictly separated from Ψ_0. As in section 2.5, we can define $\Psi(\underline{\psi}) := [\underline{\psi}, \infty)$ for each threshold value $\underline{\psi} > 1$ and refer to the resulting GROW S-value $S^*_{\underline{\psi}} := S^*_{\Psi(\underline{\psi})}$ as the *$\underline{\psi}$-GROW S-value*. Further, for significance level $\alpha \in (0, 1)$, we will call the safe test $T_\alpha(S^*_{\underline{\psi}})$ the *$\underline{\psi}$-GROW test for significance level α*. We now derive a closed formula for the $\underline{\psi}$-GROW S-values. We first need the following lemma.

Lemma 8. *Let $\mathcal{P} = \{P_\eta : \mu \in E\}$ be an exponential family given in the canonical parametrization and let $\mathcal{P} = \{Q_\mu : \mu \in M\}$ be the same family in the mean-value parametrization. Then*

$$\mu(\eta) := \mathbf{E}_{X \sim P_\eta}[X] \tag{4.39}$$

is a strictly increasing function of η.

Proof. See [6, Section 18.3]. □

We can now state and prove the following theorem, which says that the optimal prior W_1^* is simply a point mass on the threshold value $\underline{\psi}$.

Theorem 9. *Let n, n_b and n_1 be fixed, and suppose $N_{b1} \sim$ fnchypg(n, n_b, n_1, ψ) for some $\psi \geq 1$. If we have null parameter set $\Psi_0 = \{1\}$ and, for some fixed threshold value $\underline{\psi} > 1$, alternative parameter set $\Psi_1 = [\underline{\psi}, \infty)$, then the $\underline{\psi}$-GROW S-value $S^*_{\underline{\psi}}$ is given simply by*

$$S^*_{\underline{\psi}}(N_{b1}) = \frac{P_{\underline{\psi}}(N_{b1})}{P_1(N_{b1})}. \tag{4.40}$$

Proof. Let P_ψ denote the probability mass function of a fnchypg(n, n_b, n_1, ψ) distribution. Then, for any n_{b1}, we can write $P_\psi(N_{b1} = n_{b1})$ as

$$P_\psi(N_{b1} = n_{b1}) = \frac{\binom{n_a}{n_1 - n_{b1}}\binom{n_b}{n_{b1}}\psi^{n_{b1}}}{\sum_{n'_{b1}=\max\{0, n_1 - n_a\}}^{\min\{n_b, n_1\}} \binom{n_a}{n_1 - n'_{b1}}\binom{n_b}{n'_{b1}}\psi^{n'_{b1}}} = h(n_{b1})\exp\left[\eta(\psi)T(n_{b1}) - A(\psi)\right], \tag{4.41}$$

where

- $h(n_{b1}) = \binom{n_a}{n_1 - n_{b1}}\binom{n_b}{n_{b1}}$
- $\eta(\psi) = \log \psi$
- $T(n_{b1}) = n_{b1}$
- $A(\psi) = \log\left(\sum_{n'_{b1}=\max\{0,n_1-n_a\}}^{\min\{n_b,n_1\}} \binom{n_a}{n_1-n'_{b1}}\binom{n_b}{n'_{b1}}\psi^{n'_{b1}}\right)$.

We therefore see that $\mathcal{P} \coloneqq \{P_\psi : \psi \geq 1\}$ forms an exponential family. We define the canonical and mean-value parametrizations of $\mathcal{P}$ as follows

- The canonical parametrization is $\mathcal{P} = \{P^{\text{can}}_\eta : \eta \in E\}$, where $E \coloneqq \{\eta(\psi) : \psi \in [1, \infty)\}$, $\eta(\psi) \coloneqq \log \psi$ and $P^{\text{can}}_{\eta(\psi)} \coloneqq P_\psi$.
- The mean-value parametrization is $\mathcal{P} = \{P^{\text{mean}}_\mu : \mu \in M\}$, where $M \coloneqq \{\mu(\eta) : \eta \in E\}$, $\mu(\eta) \coloneqq \mathbf{E}_{N_{b1}\sim P^{\text{can}}_\eta}[N_{b1}]$ and $P^{\text{mean}}_{\mu(\eta)} \coloneqq P^{\text{can}}_\eta$.

Since η is strictly increasing in ψ and, by Lemma 8, μ is strictly increasing in η, we see that μ is strictly increasing in ψ. Thus the null and alternative parameter sets $\Psi_0 = \{1\}$ and $\Psi_1 = [\underline{\psi}, \infty)$ induce null and alternative parameter sets for the mean $M_0 \coloneqq \{\mu(\eta(1))\}$ and $M_1(\underline{\psi}) \coloneqq [\mu(\eta(\underline{\psi})), a)$, where $a > \mu\eta(\underline{\psi})$ and $\mu(\eta(\underline{\psi})) > \mu(\eta(1))$. We can now apply Proposition 3, which states that the GROW S-value $S^*_{M_1(\underline{\psi})}$ is given by

$$S^*_{M_1(\underline{\psi})}(N_{b1}) = \frac{P^{\text{mean}}_{\mu(\eta(\underline{\psi}))}(N_{b1})}{P^{\text{mean}}_{\mu(\eta(1))}(N_{b1})} = \frac{P^{\text{can}}_{\eta(\underline{\psi})}(N_{b1})}{P^{\text{can}}_{\eta(1)}(N_{b1})} = \frac{P_{\underline{\psi}}(N_{b1})}{P_1(N_{b1})}. \tag{4.42}$$

Finally, since we have simply reparametrized, we know that $S^*_{\Psi_1(\underline{\psi})} = S^*_{M_1}(\underline{\psi})$ and the result follows. □

Note that by Proposition 7 we know that $S^*_{\underline{\psi}}$ is also an *unconditional* S-value, though it is not necessarily the unconditional GROW S-value.

4.3 The conditional UMPG S-value

As discussed in section 2.5, instead of deciding a fixed threshold value $\underline{\psi} > 1$ in advance and using the $\underline{\psi}$-GROW S-value $S^*_{\underline{\psi}}$, we can choose the threshold value $\underline{\psi}$ such that the $\underline{\psi}$-GROW test $T_\alpha(S^*_{\underline{\psi}})$ has the greatest power of all the $\underline{\psi}$-GROW tests, *uniformly* over $\Psi_1 = [1, \infty)$. If such a $\underline{\psi}$ exists, it called the UMP threshold and is denoted by ψ^*. The resulting S-value $S^*_{\psi^*}$ is called a uniformly most powerful GROW S-value for significance level α, abbreviated to the UMPG S-value for significance level α. Further, the test safe $T_\alpha(S^*_{\psi^*})$ is called a uniformly most powerful GROW test for significance level α, abbreviated to UMPGT(α). Formally, we are asking whether, for given n, n_b, n_1 and $\alpha \in (0, 1)$, there exists ψ^* such that for all $\psi_1, \psi_2 > 1$ we have

$$P_{\psi_1}(S^*_{\psi^*} \geq 1/\alpha) \geq P_{\psi_1}(S^*_{\psi_2} \geq 1/\alpha). \tag{4.43}$$

Since $\mathcal{P} = \{P_\psi : \psi \geq 1\}$ forms an exponential family, we can apply Theorem 5 to get Proposition 10 below. This states that not only does a UMPGT(α) exist (provided α is not too small), it is in fact more powerful than *any* safe GROW test. Thus the terminology UMPGT(α) is fully justified. Recall the abbreviations $K \coloneqq \{k : \max\{0, n_1 - n_a\} \leq k \leq \min\{n_b, n_1\}\}$ and

$$c_k \coloneqq \binom{n_a}{n_1 - k}\binom{n_b}{k}. \tag{4.44}$$

Proposition 10. *Fix n, n_b and n_1 and take $\Psi_0 = \{1\}$. Let $k_{\max} \coloneqq \max(K) = \min\{n_b, n_1\}$, and suppose $|K| \geq 2$. Then for all $\alpha \in (c_{k_{\max}}/\binom{n}{n_1}, 1)$ there exists a UMPGT(α), namely $T_\alpha(S^*_{\psi^*})$, where ψ^* is the unique solution to $d(\psi^*) = -\log \alpha$, where*

$$d(\psi) \coloneqq \text{KL}(P_\psi || P_1). \tag{4.45}$$

In order to prove Proposition 10, we require the following important fact for exponential families in the mean-value parameterization.

Lemma 11. *Let $\mathcal{P} = \{P_\mu : \mu \in M\}$ be an exponential family given in the mean-value parametrization. Then, for any $\mu' \in M$, we have that*

$$d(\mu) \coloneqq \mathrm{KL}(P_\mu || P_{\mu'}) \tag{4.46}$$

is a strictly convex function of μ.

Proof. See [6, Section 18.4]. □

We can now prove Proposition 10.

Proof. First note that if $|K| = 1$ then N_{b1} can only take one value, meaning all S-values are identically equal to one. In this uninteresting case all the associated tests are UMP since they all have the same power. For the remainder of the proof assume $|K| > 1$.

Recall that $\mathcal{P} \coloneqq \{P_\psi : \psi \geq 1\}$ forms an exponential family:

$$P_\psi(N_{b1} = n_{b1}) = h(n_{b1}) \exp\left[\eta(\psi) T(n_{b1}) - A(\psi)\right], \tag{4.47}$$

where

- $h(n_{b1}) = \binom{n_a}{n_1 - n_{b1}} \binom{n_b}{n_{b1}}$
- $\eta(\psi) = \log \psi$
- $T(n_{b1}) = n_{b1}$
- $A(\psi) = \log\left(\sum_{n'_{b1}=\max\{0, n_1 - n_a\}}^{\min\{n_b, n_1\}} \binom{n_a}{n_1 - n'_{b1}} \binom{n_b}{n'_{b1}} \psi^{n'_{b1}}\right)$.

To prove the proposition, it therefore suffices to verify the conditions of Theorem 5.

Clearly, η and A are differentiable and η is strictly increasing. We do not need to worry about swapping an integral and a derivative since N_{b1} is a discrete random variable. Since for all $k \in K$, $\psi \mapsto P_\psi(k)$ is continuous, $d(\psi)$ is also continuous. To see that $d(\psi)$ is strictly increasing, we use Lemmas 11 and 8, which make use of both the canonical and mean-value parametrizations (of which our parametrization in terms of ψ is neither). We therefore use the same notation for these parametrizations as in the proof of Theorem 9. We then have

$$\begin{aligned} d(\psi) &\coloneqq \mathrm{KL}(P_\psi || P_1) && (4.48)\\ &= \mathrm{KL}(P^{\mathrm{can}}_{\eta(\psi)} || P^{\mathrm{can}}_{\eta(1)}) && (4.49)\\ &= \mathrm{KL}(P^{\mathrm{mean}}_{\mu(\eta(\psi))} || P^{\mathrm{mean}}_{\mu(\eta(1))}). && (4.50) \end{aligned}$$

Now $\eta(\psi) = \log \psi$ is clearly strictly increasing in ψ. Further, by Lemma 8, we see that $\mu(\eta)$ is strictly increasing in η. Thus $\mu(\eta(\psi))$ is strictly increasing in ψ. Now since $d(1) = 0$ and the KL-divergence is always nonnegative, Lemma 11 implies that $d(\psi)$ is strictly increasing for $\psi \geq 1$.

It remains to show that $\lim_{\psi \to \infty} d(\psi) = \log\left(\binom{n}{n_1} / c_{k_{\max}}\right)$. For $k \in K$, let

$$f_k(\psi) \coloneqq \sum_{l \in K} c_l \psi^{l-k}. \tag{4.51}$$

For $k < k_{\max}$ there exists $l > k$ and so $f_k(\psi) \to \infty$ as $\psi \to \infty$. If $k = k_{\max}$, then all terms in the sum tend to zero except the last, which equals $c_{k_{\max}}$. So $f_{k_{\max}}(\psi) \to c_{k_{\max}}$ as $\psi \to \infty$. Using these facts,

we have

$$d(\psi) = \sum_{k\in K} \frac{c_k\psi^k}{\sum_{l\in K} c_l\psi^l}\left[\log\frac{c_k\psi^k}{\sum_{m\in K} c_m\psi^m} - \log\frac{c_k}{\binom{n}{n_1}}\right] \tag{4.52}$$

$$= \sum_{k\in K} \frac{c_k}{\sum_{l\in K} c_l\psi^{l-k}}\left[\log\frac{1}{\sum_{m\in K} c_m\psi^{m-k}} + \log\binom{n}{n_1}\right] \tag{4.53}$$

$$= \sum_{k\in K} \frac{c_k}{f_k(\psi)}\left[\log\frac{1}{f_k(\psi)} + \log\binom{n}{n_1}\right] \tag{4.54}$$

$$= \sum_{k<k_{\max}} \frac{c_k}{f_k(\psi)}\left[\log\frac{1}{f_k(\psi)} + \log\binom{n}{n_1}\right] \quad + \quad \frac{c_{k_{\max}}}{f_{k_{\max}}(\psi)}\left[\log\frac{1}{f_{k_{\max}}(\psi)} + \log\binom{n}{n_1}\right] \tag{4.55}$$

$$\to \frac{c_{k_{\max}}}{c_{k_{\max}}}\left[\log\frac{1}{c_{k_{\max}}} + \log\binom{n}{n_1}\right] \quad \text{as } \psi\to\infty \tag{4.56}$$

$$= \log\left(\frac{\binom{n}{n_1}}{c_{k_{\max}}}\right). \tag{4.57}$$

□

4.4 Growth and power of conditional S-values

We now inspect the (worst case) growth rate and power of the conditional GROW and conditional UMPG S-values in the case where $n_a = n_b = 50$. We took $n_1 \in \{5, 10, 20, 50\}$, noting that, by symmetry, $n_1' = n - n_1$ will give the same results in reverse. For each of these values of n_1, a UMPG S-value exists and is denoted by $S^*_{\psi^*(n_1)}$, where $\psi^*(n_1)$ is the UMP threshold given by Proposition 10 as the solution to

$$\mathrm{KL}(P_{\psi,n_1}||P_{1,n_1}) = -\log\alpha, \tag{4.58}$$

where P_{ψ,n_1} denotes the fnchypg(n, n_b, n_1, ψ) distribution. We start at $n_1 = 5$ since (4.58) only has a solution for $n_1 \geq 5$.

For a sequence of twenty evenly spaced odds ratio thresholds from $\underline{\psi} = 1$ to $\underline{\psi} = 5$, we calculated the (worst case) growth rate and power of the conditional GROW S-value $S^*_{\Psi_1(\underline{\psi})}$ and the conditional UMPG S-value $S^*_{\psi^*(n_1)}$ on $\Psi_1(\underline{\psi})$. The results are found in figure 4.1 to 4.4, where we write simply ψ for the threshold value rather than $\underline{\psi}$. It is worth stressing that, as the threshold value $\underline{\psi}$ increases, the GROW S-value changes while the UMPG S-value does not. Nevertheless, when $\underline{\psi} = \psi^*$, the two S-values coincide. Therefore the growth and power curves meet when $\underline{\psi} = \psi^*$. This is not always apparent in the figures if ψ^* is very large.

Note that the GROW S-value always has growth rate at least as high as the UMPG S-value, and the UMPG S-value always has power at least as high as the GROW S-value. Indeed, this follows from the definition of these S-values. Interestingly however, there are cases where their power appears equal while the GROW S-value has larger growth rate (for example when $n_1 = 50$ and $\underline{\psi} > 3$). In these cases, it seems that the GROW S-value is clearly the better choice. We also see that Fisher's exact test always has power at least as large as either S-value (for $n_1 = 5$ the power of Fisher's exact test seems to be identical to the power of the conditional UMPG S-value). Indeed, it should perhaps be expected that since S-values satisfy the extra requirement of having a type I error guarantee robust to optional continuation, that they will necessarily perform worse in some other respect. The fact that the S-values have lower power than Fisher's exact test is by no means a definitive argument against using them.

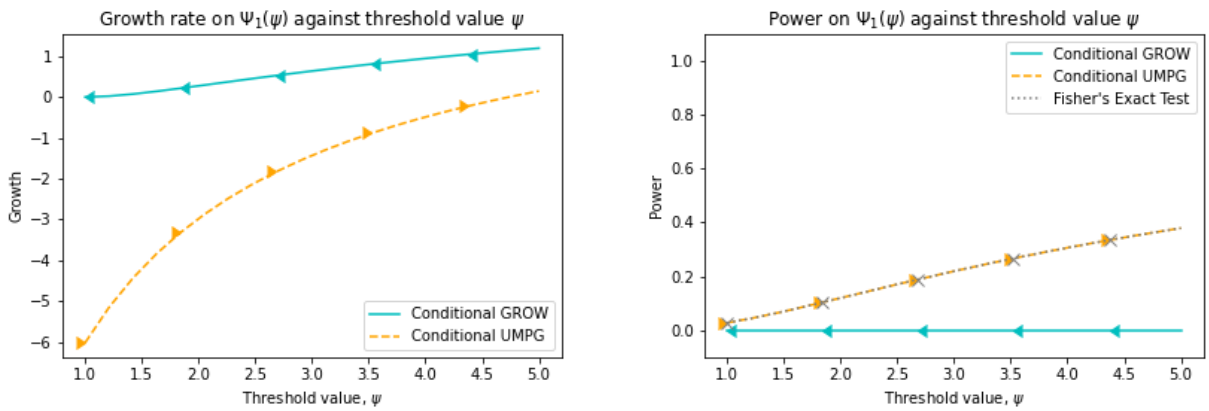

Figure 4.1: $n_1 = 5$. There exists a UMP ψ, namely $\psi^* = 44.14$.

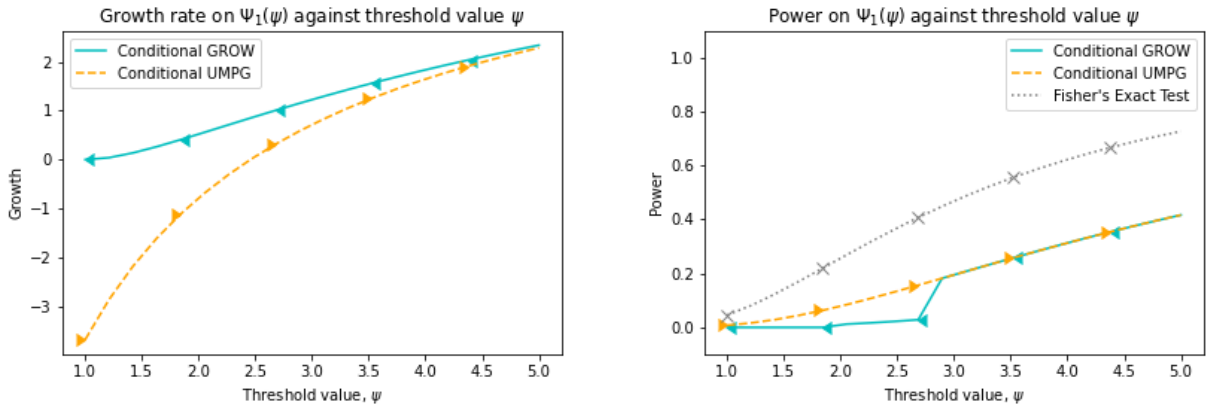

Figure 4.2: $n_1 = 10$. There exists a UMP ψ, namely $\psi^* = 6.65$.

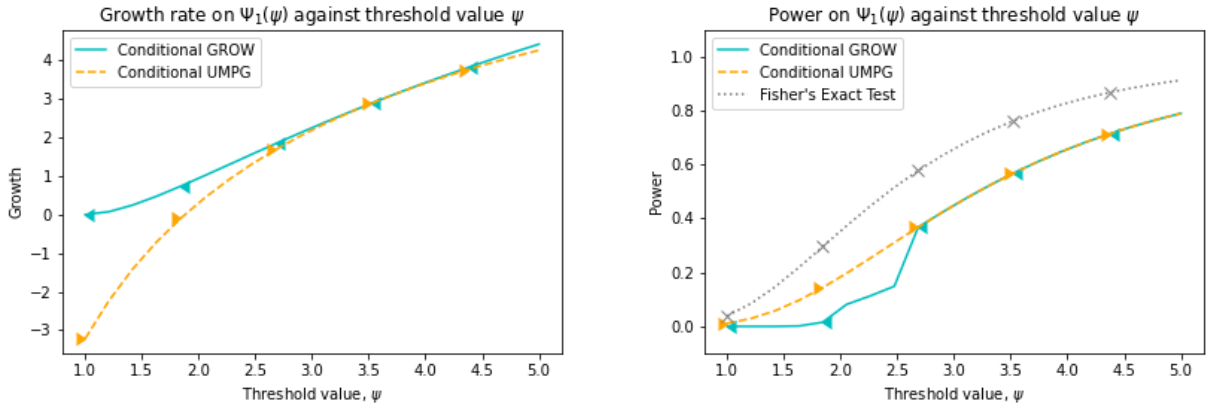

Figure 4.3: $n_1 = 20$. There exists a UMP ψ, namely $\psi^* = 3.62$.

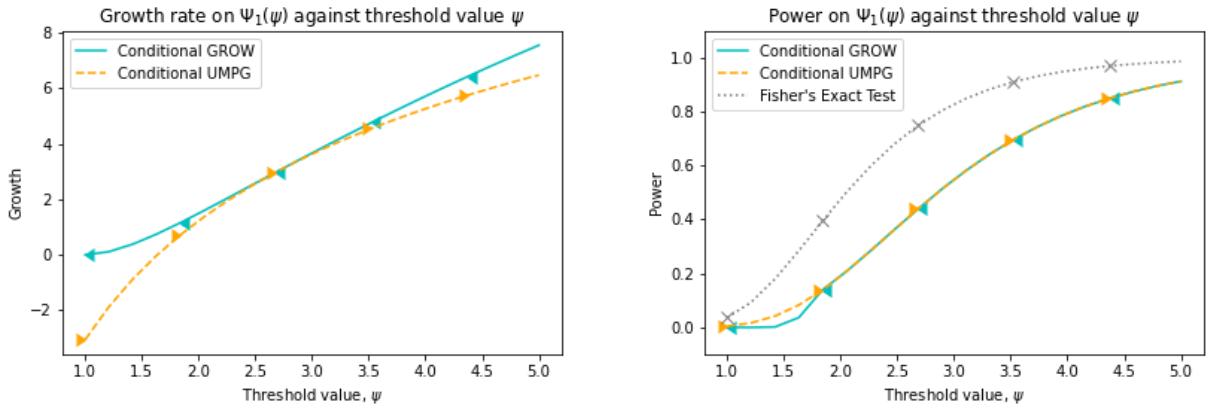

Figure 4.4: $n_1 = 50$. There exists a UMP ψ, namely $\psi^* = 2.69$.

Chapter 5

Safe tests for 2×2 contingency tables when N_1 is unknown

In the previous chapter it was assumed that N_1 was known in advance. Here we drop that assumption and construct safe tests for 2×2 contingency tables where N_1 truly is a random variable whose value is not known before the experiment begins. This is, in most cases, much closer to reality. For example, when running a clinical trial, the total number of patients who will recover is not known in advance. Since these tests do not condition on the value of N_1, we refer to them as *unconditional* safe tests. The S-values on which the unconditional safe tests are based will be referred to as the *unconditional* S-values. As before, we will construct two types, namely the *unconditional GROW* S-values and, where they exist, the *unconditional UMPG* S-values.

Given restricted parameter sets, the GROW S-value can be estimated by numerically approximating the JIP between the sets of Bayesian marginals. Practitioners analysing 2×2 contingency tables commonly use the risk difference, relative risk or the odds ratio between θ_a and θ_b to quantify the effect size. In section 5.1 we describe how threshold values of any of these parameters can be used to construct a subset of the alternative parameter set. In the case of the risk difference, this suffices to strictly separate the parameter sets, but for the relative risk or the odds ratio further restriction must be made. We permit further restriction in the form of a prior knowledge rectangle, representing the range of values of θ_a and θ_b that the practitioner deems reasonable. In section 5.2 we recap the conditions of the main result of [5] and refer to Appendix A for a discussion of their justification in our case. We also specify what we mean by a UMPG S-value in the unconditional setting.

Section 5.3 outlines the process used to discretize the parameter sets so that the JIP—and therefore the GROW S-value—can be approximated numerically. While numerically approximating the JIP is rather computationally intensive, we provide a simplified formula for the gradient of the objective function in Proposition 12 that speeds up the calculation by around an order of magnitude. Later, in chapter six, we will prove a number of results that bypass the JIP approximation altogether by either reducing the problem to a much simpler computation or even finding an explicit expression for the GROW S-value.

5.1 Parameter of interest and prior knowledge

The risk difference δ, relative risk λ and odds ratio ψ for any $(\theta_a, \theta_b) \in [0,1]^2$ are defined by

$$\delta = \delta(\theta_a, \theta_b) := \theta_b - \theta_a \tag{5.1}$$

$$\lambda = \lambda(\theta_a, \theta_b) := \theta_b / \theta_a \tag{5.2}$$

$$\psi = \psi(\theta_a, \theta_b) := \frac{\theta_b}{1-\theta_b} \frac{1-\theta_a}{\theta_a}. \tag{5.3}$$

When discussing facts that hold whichever parameter is chosen, we use $\epsilon = \epsilon(\theta_a, \theta_b)$ to denote an arbitrary choice. Recall that we are only considering one-sided tests of $\theta_b > \theta_a$ versus $\theta_b = \theta_a$. This corresponds to the one-sided test of $\delta > 0$, $\lambda > 1$ or $\psi > 1$ versus $\delta = 0$, $\lambda = 1$ or $\psi = 1$. Thus, given

any threshold value $\underline{\delta} > 0, \underline{\lambda} > 1$ or $\underline{\psi} > 1$ deemed a *minimum clinically relevant effect size*, we define the following restrictions[1] of Θ_1

$$\Theta_1(\underline{\delta}) \coloneqq \{(\theta_a, \theta_b) \in [0,1]^2 : \theta_b \geq \theta_a + \underline{\delta}\} \tag{5.4}$$

$$\Theta_1(\underline{\lambda}) \coloneqq \{(\theta_a, \theta_b) \in [0,1]^2 : \theta_b/\theta_a \geq \underline{\lambda}\} \tag{5.5}$$

$$\Theta_1(\underline{\psi}) \coloneqq \left\{(\theta_a, \theta_b) \in [0,1]^2 : \frac{\theta_b}{1-\theta_b}\frac{1-\theta_a}{\theta_a} \geq \underline{\psi}\right\}. \tag{5.6}$$

Thus, in general, we have

$$\Theta_1(\underline{\epsilon}) \coloneqq \{(\theta_a, \theta_b) \in [0,1]^2 : \epsilon(\theta_a, \theta_b) \geq \underline{\epsilon}\}. \tag{5.7}$$

As mentioned above, $\Theta_1(\underline{\delta})$ is already strictly separated from Θ_0 for any $\underline{\delta} > 0$. However, for the other parameter choices, we have

$$\Theta_1(\underline{\lambda}) \cap \Theta_0 = \{(0,0)\} \quad \text{and} \quad \Theta_1(\underline{\psi}) \cap \Theta_0 = \{(0,0),(1,1)\}, \tag{5.8}$$

for any $\underline{\lambda} > 1$ and $\underline{\psi} > 1$. In these cases the GROW S-value is degenerate unless we make further restrictions. We now discuss prior knowledge that may be incorporated to achieve this, when it exists. We will allow prior knowledge also in the case of the risk difference.

Suppose the practitioner already has knowledge on which values of θ_a and θ_b are reasonable. For example, if the probability of recovery without treatment *or* placebo is known to be 0.1 from data drawn from the general population, it may be reasonable to assume that receiving a placebo cannot decrease the probability of recovery, namely, the practitioner has the prior knowledge that $\theta_a \geq 0.1$. More generally, the practitioner may have prior knowledge on the range of values that θ_a and θ_b can reasonably take. We will refer to this as the *prior knowledge rectangle* (PKR). More precisely, suppose the practitioner knows that $\theta_a \in [\theta_a^L, \theta_a^U]$ and $\theta_b \in [\theta_b^L, \theta_b^U]$ for some $\theta_a^L, \theta_a^U, \theta_b^L, \theta_b^U \in [0,1]$, where $\theta_a^L \leq \theta_a^U$ and $\theta_b^L \leq \theta_b^U$. Then the prior knowledge rectangle is defined by

$$\text{PKR} \coloneqq [\theta_a^L, \theta_a^U] \times [\theta_b^L, \theta_b^U]. \tag{5.9}$$

We allow the possibility that $\theta_a^L = \theta_a^U$ (or $\theta_b^L = \theta_b^U$). This may reflect, for example, knowledge of previous large studies using the same placebo for the same condition that determined the probability of recovery for the placebo group to a high accuracy. We now define

$$\Theta_1(\underline{\epsilon})' \coloneqq \Theta_1(\underline{\epsilon}) \cap \text{PKR} \quad \text{and} \quad \Theta_0' \coloneqq \Theta_0 \cap \text{PKR}. \tag{5.10}$$

Further, it may occasionally be necessary to refer to

$$\Theta_1' \coloneqq \Theta_1 \cap \text{PKR}, \tag{5.11}$$

the alternative parameter set restricted by the prior knowledge rectangle but not by any threshold value.

5.2 The unconditional GROW and UMPG S-values

Given a threshold value $\underline{\epsilon}$ and prior knowledge rectangle PKR such that $\Theta_1(\underline{\epsilon})'$ and Θ_0' are strictly separated, we wish to find the the GROW S-value $S^*_{\Theta_1(\underline{\epsilon})'}$. It is these such S-values that we will refer to as the *unconditional GROW S-values*. Recall Theorem 1, which states that these S-values exist and are given by

$$S^*_{\Theta_1(\underline{\epsilon})'}(Z) = \frac{p_{W_1^*}(Z)}{p_{W_0^*}(Z)}, \quad \text{where} \quad (W_1^*, W_0^*) \coloneqq \inf_{(W_1,W_0)\in\mathcal{W}(\Theta_1')\times\mathcal{W}(\Theta_0)} \text{KL}(P_{W_1}||P_{W_0}), \tag{5.12}$$

provided the following conditions hold

[1] This notation is ambiguous once numerical values for the threshold is substituted. In such cases we will make it clear which parameter is being used.

1. For all $\theta_0 \in \Theta_0$ and $W_1 \in \mathcal{W}(\Theta_1')$ we have that P_{θ_0} is absolutely continuous relative to P_{W_1}.
2. The infimum $\inf_{(W_1,W_0)\in\mathcal{W}(\Theta_1')\times\mathcal{W}(\Theta_0)} \mathrm{KL}(P_{W_1}||P_{W_0})$ is finite.
3. The infimum is achieved by some (W_1^*, W_0^*).
4. The infimum is achieved uniquely.

For an exploration of whether these conditions indeed hold in the cases we consider, see appendix A, the conclusion of which is essentially that in every case, each condition either holds or its failure to hold is not problematic, meaning we can always apply the theorem.

As in the conditional case, instead of fixing a threshold value $\underline{\epsilon}$, we may instead choose the value of ϵ that uniformly maximizes the power of $S^*_{\Theta_1(\epsilon)'}$ over Θ_1'. More precisely, for a given PKR, let E be the values of ϵ such that $\Theta_1(\epsilon)' \neq \emptyset$ (for example, if we have the risk difference δ and PKR $= [0.5, 1] \times [0, 1]$, then $E = (0, 0.5]$). Then, given unconditional GROW S-values $\{S^*_{\Theta_1(\epsilon)'} : \epsilon \in E\}$, if there exists $\epsilon^* \in E$ such that, for a given significance level $\alpha \in (0, 1)$, we have

$$\forall \theta \in \Theta_1',\ \forall \epsilon \in E \quad P_\theta\big(S^*_{\Theta_1(\epsilon^*)'} \geq 1/\alpha\big) \geq P_\theta\big(S^*_{\Theta_1(\epsilon)'} \geq 1/\alpha\big), \tag{5.13}$$

we refer to ϵ^* as the *uniformly most powerful (UMP) threshold*, the GROW S-value $S^*_{\Theta_1(\epsilon^*)'}$ as the *unconditional uniformly most powerful GROW (UMPG) S-value* and the test $T_\alpha(S^*_{\Theta_1(\epsilon^*)'})$ as an *unconditional uniformly most powerful GROW test for significance level* α, abbreviated to UMPGT(α).

5.3 Estimating GROW S-values by numerically approximating the JIP

Chapter six discusses cases in which there are shortcuts to finding the unconditional GROW S-values. However, there are still cases where it is necessary to estimate the unconditional GROW S-values by directly approximating the JIP numerically. Although, as seen in Chapter 2, this is a convex optimization problem, it can still be very computationally intensive. We now discuss how the parameter sets can be discretized to make the problem amenable to numerical methods, and how the gradient of the objective function can be found in closed form, which speeds up the calculation by around an order of magnitude.

Suppose, for given threshold value $\underline{\epsilon}$ and PKR, we discretize the null and alternative parameter sets by

$$\dot{\Theta}_0' = \{\theta_{0,1}, \ldots, \theta_{0,K_0}\} \subseteq \Theta_0' \quad \text{and} \quad \dot{\Theta}_1(\underline{\epsilon})' = \{\theta_{1,1}, \ldots, \theta_{1,K_1}\} \subseteq \Theta_1(\underline{\epsilon})' \tag{5.14}$$

respectively, where $K_0, K_1 \in \mathbb{N}$. A natural choice is to construct an evenly spaced $n \times n$ grid of points

$$\dot{\Theta}_n \coloneqq \{(i/n, j/n) : i, j \in [n]\}, \tag{5.15}$$

where $[n] \coloneqq \{0, 1, \ldots, n\}$, and then take

$$\dot{\Theta}_0' = \Theta_0' \cap \dot{\Theta}_n \quad \text{and} \quad \dot{\Theta}_1(\underline{\epsilon})' = \Theta_1(\underline{\epsilon})' \cap \dot{\Theta}_n. \tag{5.16}$$

This is the discretization we used in our experiments.

For any $K \in \mathbb{N}$, let $\triangle^K$ denote the set of distributions on K points, namely

$$\triangle^K \coloneqq \left\{ \mathbf{w} \in \mathbb{R}^K : \sum_{k=1}^K w_k = 1 \text{ and } w_k \geq 0 \text{ for } k = 1, \ldots, K \right\}. \tag{5.17}$$

Then, for any $\mathbf{w} = (\mathbf{w}_0, \mathbf{w}_1) \in \triangle^{K_0} \times \triangle^{K_1}$, let $W_0(\mathbf{w}) \in \mathcal{W}(\dot{\Theta}_0')$ be the prior that puts puts mass $w_{0,k}$ on $\theta_{0,k}$ for each $k = 1, \ldots, K_0$, and let $W_1(\mathbf{w}) \in \mathcal{W}(\dot{\Theta}_1(\underline{\epsilon})')$ be the prior that puts puts mass $w_{1,k}$ on $\theta_{1,k}$ for each $k = 1, \ldots, K_0$. The corresponding marginal distributions on $Z = (N_{a1}, N_{b1})$ are then given by

$$P_{W_i(\mathbf{w})} \coloneqq \sum_{k=1}^{K_i} w_{i,k} P_{\theta_{i,k}} \quad \text{for } i \in \{0, 1\}. \tag{5.18}$$

Finally, let $f : \triangle^{K_0} \times \triangle^{K_1} \to [0, \infty)$ be defined by

$$f(\mathbf{w}) \coloneqq \mathrm{KL}(P_{W_1(\mathbf{w})} || P_{W_0(\mathbf{w})}). \tag{5.19}$$

We therefore have

$$\begin{aligned} \inf_{(W_0,W_1) \in \mathcal{W}(\dot{\Theta}_0') \times \mathcal{W}(\dot{\Theta}_1(\underline{\epsilon})')} \mathrm{KL}(P_{W_1} || P_{W_0}) &= \inf_{\mathbf{w} \in \triangle^{K_0} \times \triangle^{K_1}} \mathrm{KL}(P_{W_1(\mathbf{w})} || P_{W_0(\mathbf{w})}) && (5.20) \\ &= \inf_{\mathbf{w} \in \triangle^{K_0} \times \triangle^{K_1}} f(\mathbf{w}) && (5.21) \\ &= \min_{\mathbf{w} \in \triangle^{K_0} \times \triangle^{K_1}} f(\mathbf{w}), && (5.22) \end{aligned}$$

where the last line follows since f is a continuous function on a compact set and so attains its infimum at some $\mathbf{w}^*$.

Let $\tilde{W}_i^* \coloneqq W_i(\mathbf{w}^*)$ for $i \in \{0, 1\}$. Then we can define the likelihood ratio

$$\tilde{S}^*(Z) \coloneqq \frac{P_{\tilde{W}_1^*}(Z)}{P_{\tilde{W}_i^*}(Z)}, \tag{5.23}$$

which, by Theorem 1, is then the GROW S-value for the discretized parameter sets. Suppose the conditions of Theorem 1 are satisfied for the original (non-discretized) parameter sets and the GROW S-value for these original parameter sets is given by

$$S^* \coloneqq \frac{P_{W_1^*}(Z)}{P_{W_0^*}(Z)}, \quad \text{where} \quad (W_1^*, W_0^*) \coloneqq \operatorname*{arg\,min}_{(W_1,W_0) \in \mathcal{W}(\Theta_0') \times \mathcal{W}(\Theta_1(\underline{\epsilon})')} \mathrm{KL}(P_{W_1} || P_{W_0}). \tag{5.24}$$

We will take $\tilde{S}^*$ as our approximation of S^*. By Theorem 1, we know that the worst case growth rates for $\tilde{S}^*$ and S^* are given by

$$\mathrm{GR}(\tilde{S}^*) = \mathrm{KL}(P_{\tilde{W}_1^*} || P_{\tilde{W}_0^*}) \quad \text{and} \quad \mathrm{GR}(S^*) = \mathrm{KL}(P_{W_1^*} || P_{W_0^*}). \tag{5.25}$$

Now, if our discretizations are sufficiently fine, we expect that

$$\begin{aligned} \mathrm{KL}(P_{W_1^*} || P_{W_0^*}) &= \min_{(W_0,W_1) \in \mathcal{W}(\Theta_0') \times \mathcal{W}(\Theta_1(\underline{\epsilon})')} \mathrm{KL}(P_{W_1} || P_{W_0}) && (5.26) \\ &\approx \min_{(W_0,W_1) \in \mathcal{W}(\dot{\Theta}_0') \times \mathcal{W}(\dot{\Theta}_1(\underline{\epsilon})')} \mathrm{KL}(P_{W_1} || P_{W_0}) && (5.27) \\ &= \mathrm{KL}(P_{\tilde{W}_1^*} || P_{\tilde{W}_0^*}), && (5.28) \end{aligned}$$

meaning the worst case growth rates of $\tilde{S}^*$ and S^* are approximately equal. Thus $\tilde{S}^*$ has approximately optimal worst case growth rate and so in this crucial sense it is a good approximation to S^*.

In summary, the optimization problem that we will solve numerically is to minimize

$$f(\mathbf{w}) \coloneqq \mathrm{KL}(P_{W_1(\mathbf{w})} || P_{W_0(\mathbf{w})}) \tag{5.29}$$

subject to

$$\begin{aligned} \sum_{k=1}^{K_i} w_{i,k} = 1 \quad \text{for } i \in \{0, 1\}, && (5.30) \\ w_{0,k} \in [0, 1] \quad \text{for } k = 1, \ldots, K_0, && (5.31) \\ w_{1,k} \in [0, 1] \quad \text{for } k = 1, \ldots, K_1. && (5.32) \end{aligned}$$

Recall (chapter 2) that this is a convex optimization problem. Further, for any z, the gradient of $p_{W_i(\mathbf{w})}(z)$ with respect to $\mathbf{w}$ is well-defined for $i \in \{0, 1\}$. Finally, since Z takes on finitely many values, we see that the gradient of $f(\mathbf{w})$ with respect to $\mathbf{w}$ is also well-defined. Since optimization packages can often take the gradient as an argument in order to speed up calculation, the following proposition will be useful.

Proposition 12. *The gradient of $f(\mathbf{w})$ is well-defined and is given by*

$$\frac{\partial f}{\partial w_{0,k}}(\mathbf{w}) = -\sum_z \frac{P_{W_1(\mathbf{w})}(z) P_{\theta_{0,k}}(z)}{P_{W_0(\mathbf{w})}(z)} \tag{5.33}$$

for $k = 1, 2, \ldots, K_0$, and

$$\frac{\partial f}{\partial w_{1,k}}(\mathbf{w}) = 1 + \sum_z P_{\theta_{1,k}}(z) \log \frac{P_{W_1(\mathbf{w})}(z)}{P_{W_0(\mathbf{w})}(z)} \tag{5.34}$$

for $k = 1, 2, \ldots K_1$.

Proof. Using the definition of the marginals $P_{W_0(\mathbf{w})}$ and $P_{W_1(\mathbf{w})}$, we can write $f(\mathbf{w})$ as follows

$$\begin{aligned}
f(\mathbf{w}) &= D(P_{W_1(\mathbf{w})} || P_{W_0(\mathbf{w})}) && (5.35)\\
&= \mathbf{E}_{Z \sim P_{W_1(\mathbf{w})}} \left[\log \frac{P_{W_1(\mathbf{w})}(Z)}{P_{W_0(\mathbf{w})}(Z)} \right] && (5.36)\\
&= \sum_{r=1}^{K_1} w_{1,r} \mathbf{E}_{Z \sim P_{\theta_{1,r}}} \left[\log \frac{P_{W_1(\mathbf{w})}(Z)}{P_{W_0(\mathbf{w})}(Z)} \right] && (5.37)\\
&= \sum_{r=1}^{K_1} w_{1,r} \mathbf{E}_{Z \sim P_{\theta_{1,r}}} \left[\log P_{W_1(\mathbf{w})}(Z) - \log P_{W_0(\mathbf{w})}(Z) \right] && (5.38)\\
&= \sum_{r=1}^{K_1} w_{1,r} \mathbf{E}_{Z \sim P_{\theta_{1,r}}} \left[\log \sum_{s=1}^{K_1} w_{1,s} P_{\theta_{1,s}}(Z) - \log \sum_{t=1}^{K_0} w_{0,t} P_{\theta_{0,t}}(Z) \right]. && (5.39)
\end{aligned}$$

Now since $Z = (N_{a1}, N_{b1})$ is a discrete random variable, the expectations are finite sums and so we are permitted to pass derivatives inside the expectations. First, let $k \in \{1, \ldots, K_0\}$. Then by passing the derivative through the sum and expectations, we have

$$\begin{aligned}
\frac{\partial f}{\partial w_{0,k}}(\mathbf{w}) &= \sum_{r=1}^{K_1} w_{1,r} \mathbf{E}_{Z \sim P_{\theta_{1,r}}} \left[\frac{-P_{\theta_{0,k}}(Z)}{\sum_{t=1}^{K_0} w_{0,t} P_{\theta_{0,t}}(Z)} \right] && (5.40)\\
&= \sum_{r=1}^{K_1} w_{1,r} \mathbf{E}_{Z \sim P_{\theta_{1,r}}} \left[\frac{-P_{\theta_{0,k}}(Z)}{P_{W_0(\mathbf{w})}(Z)} \right] && (5.41)\\
&= \mathbf{E}_{Z \sim P_{W_1(\mathbf{w})}} \left[\frac{-P_{\theta_{0,k}}(Z)}{P_{W_0(\mathbf{w})}(Z)} \right] && (5.42)\\
&= -\sum_z \frac{P_{W_1(\mathbf{w})}(z) P_{\theta_{0,k}}(z)}{P_{W_0(\mathbf{w})}(z)}, && (5.43)
\end{aligned}$$

which can be fairly efficiently computed by first computing the three distributions $P_{W_1(\mathbf{w})}$, $P_{\theta_{0,k}}$ and $P_{W_1(\mathbf{w})}$, and then combining them as in the given sum. Next, let $k \in \{1, \ldots K_1\}$. Similar to above, except now using the product rule, we have

$$\frac{\partial f}{\partial w_{1,k}}(\mathbf{w}) = \sum_{r=1}^{K_1} w_{1,r} \frac{\partial}{\partial w_{1,k}} \mathbf{E}_{Z\sim P_{\theta_{1,r}}} \left[\log \sum_{s=1}^{K_1} w_{1,s} P_{\theta_{1,s}}(Z) - \log \sum_{t=1}^{K_0} w_{0,t} P_{\theta_{0,t}}(Z) \right] \tag{5.44}$$

$$+ \sum_{r=1}^{K_1} \frac{\partial w_{1,r}}{\partial w_{1,k}} \mathbf{E}_{Z\sim P_{\theta_{1,r}}} \left[\log \sum_{s=1}^{K_1} w_{1,s} P_{\theta_{1,s}}(Z) - \log \sum_{t=1}^{K_0} w_{0,t} P_{\theta_{0,t}}(Z) \right] \tag{5.45}$$

$$= \sum_{r=1}^{K_1} w_{1,r} \mathbf{E}_{Z\sim P_{\theta_{1,r}}} \left[\frac{P_{\theta_{1,k}}(Z)}{\sum_{s=1}^{K_1} w_{1,s} P_{\theta_{1,s}}(Z)} \right] \tag{5.46}$$

$$+ \mathbf{E}_{Z\sim P_{\theta_{1,k}}} \left[\log \sum_{s=1}^{K_1} w_{1,s} P_{\theta_{1,s}}(Z) - \log \sum_{t=1}^{K_0} w_{0,t} P_{\theta_{0,t}}(Z) \right] \tag{5.47}$$

$$= \sum_{r=1}^{K_1} w_{1,r} \mathbf{E}_{Z\sim P_{\theta_{1,r}}} \left[\frac{P_{\theta_{1,k}}(Z)}{P_{W_1(\mathbf{w})}(Z)} \right] + \mathbf{E}_{Z\sim P_{\theta_{1,k}}} \left[\log P_{W_1(\mathbf{w})}(Z) - \log P_{W_0(\mathbf{w})}(Z) \right] \tag{5.48}$$

$$= \mathbf{E}_{Z\sim P_{W_1(\mathbf{w})}} \left[\frac{P_{\theta_{1,k}}(Z)}{P_{W_1(\mathbf{w})}(Z)} \right] + \mathbf{E}_{Z\sim P_{\theta_{1,k}}} \left[\log \frac{P_{W_1(\mathbf{w})}(Z)}{P_{W_0(\mathbf{w})}(Z)} \right] \tag{5.49}$$

$$= \sum_z \frac{P_{W_1(\mathbf{w})}(z) P_{\theta_{1,k}}(z)}{P_{W_1(\mathbf{w})}(z)} + \sum_z P_{\theta_{1,k}}(z) \log \frac{P_{W_1(\mathbf{w})}(z)}{P_{W_0(\mathbf{w})}(z)} \tag{5.50}$$

$$= 1 + \sum_z P_{\theta_{1,k}}(z) \log \frac{P_{W_1(\mathbf{w})}(z)}{P_{W_0(\mathbf{w})}(z)}. \tag{5.51}$$

Again, this can be fairly efficiently computed by first computing the three distributions $P_{\theta_{1,k}}$, $P_{W_0(\mathbf{w})}$ and $P_{W_1(\mathbf{w})}$, and then combining them as in the given sum. □

We can now see the effect of passing the gradient to the minimization procedure. We used the `minimize` function from the `optimize` package of the SciPy library with the `SLSQP` method, for which the gradient vector is optional. If the gradient vector is not passed to the optimization method, the gradient is estimated using the method of finite differences with two points. For each coordinate, this would require two evaluations of f, each evaluation requiring a single sum over z. On the other hand, using the above expression for the true gradient vector, each coordinate requires a single sum over z. We may initially suppose that this halves the overall computation time of the minimization problem. However, as figure 5.1 shows, the computation time can be much more than halved by using the true gradient vector, especially for discretizations with many elements. This is likely because the true gradient allows faster convergence to the minimum than an estimated gradient.

We now present the results of numerically approximating the solution to the convex optimization problem (5.29) to (5.32). Note that chapter six provides quicker methods or even closed formulas for the GROW S-value in many cases. However, the method used to derive such results depends on the parameter sets being convex. Since this does not apply to $\Theta_1(\underline{\psi})'$, in this case it is necessary to approximate the JIP directly.

We considered ten values of $\underline{\psi}$ evenly spaced between 1.5 and 4.5. To ensure the parameter sets were positively separated, we used prior knowledge rectangle PKR $= [0.2, 0.8]^2$. To reduce computation time, for each threshold value $\underline{\psi}$ we assumed that the optimal prior W_1^* places all its mass on the boundary of $\Theta_1(\underline{\psi})'$. We therefore discretized $\Theta_1(\underline{\psi})'$ by taking K evenly spaced points along its boundary. We also discretized Θ_0' by using K evenly spaced points. The objective function (5.29) was then minimized using the `minimize` function from the `optimize` package of the SciPy library with the `SLSQP` method.

We initialized the weights to the discrete uniform distribution and passed the `minimize` function the simplified gradient from Proposition 12. The initial and optimized weights can be seen in figure 5.2, which shows the unit square (representing all possible values of $(\theta_a, \theta_b) \in [0,1]^2$), the prior knowledge rectangle and the boundary of $\Theta_1(\underline{\psi})$, namely the curve $\psi(\theta_a, \theta_b) = \underline{\psi}$. It can be seen that the optimial priors put nearly all the mass on just two points each. While W_0^* seems to put its mass on *four* points, these are two pairs of neighbouring discretization points. Indeed, experiments with larger numbers of

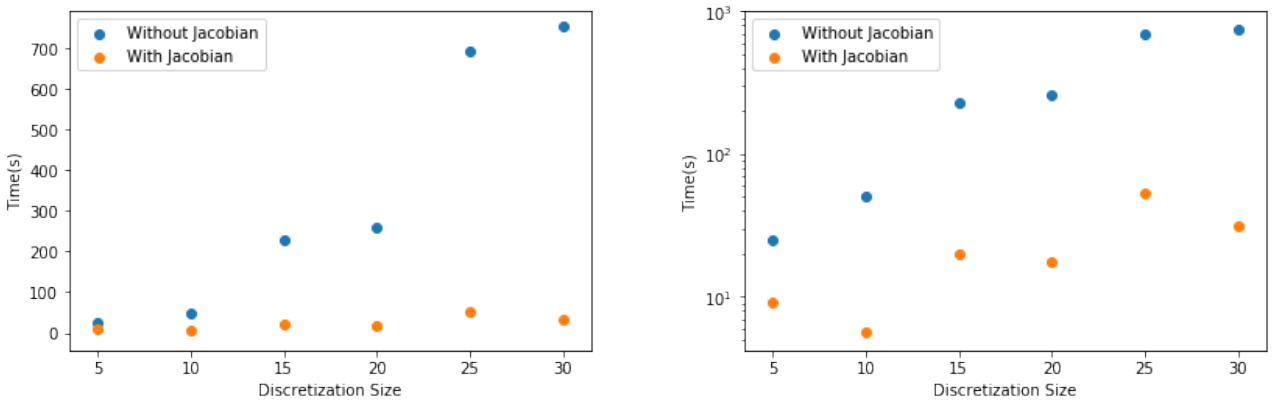

Figure 5.1: Comparison of computation times with and without the gradient. In both cases, $n_a = n_b = 10$, $\Theta_0 = \{(p, p) : p \in [0, 1]\}$ and $\Theta_1 = \{(\theta_a, \theta_a + \delta) : \theta_a \in [0, 1 - \delta]\}$ for some $\delta > 0$. A discretization of size K means both Θ_0 and Θ_1 were discretized by K evenly spaced points. The log plot on the right demonstrates that the time taken to compute the GROW S-value is approximately an order of magnitude lower when using the true gradient.

discretization points indicate that this may simply be an artefact of using a discretization, and that the truly optimal prior W_1^* is indeed a mixture of just two point masses.

Using the approximately optimal priors, each of the GROW S-values $S^*_{\Theta_1(\underline{\psi})'}$ was approximated, along with its (worst case) growth and power over $\Theta_1(\underline{\psi})'$. The results can be seen in figure 5.3. Finally, it was checked whether a UMP threshold ψ^* exists, which it did not. Therefore, for this choice of prior knowledge at least, no unconditional UMPG S-value exists.

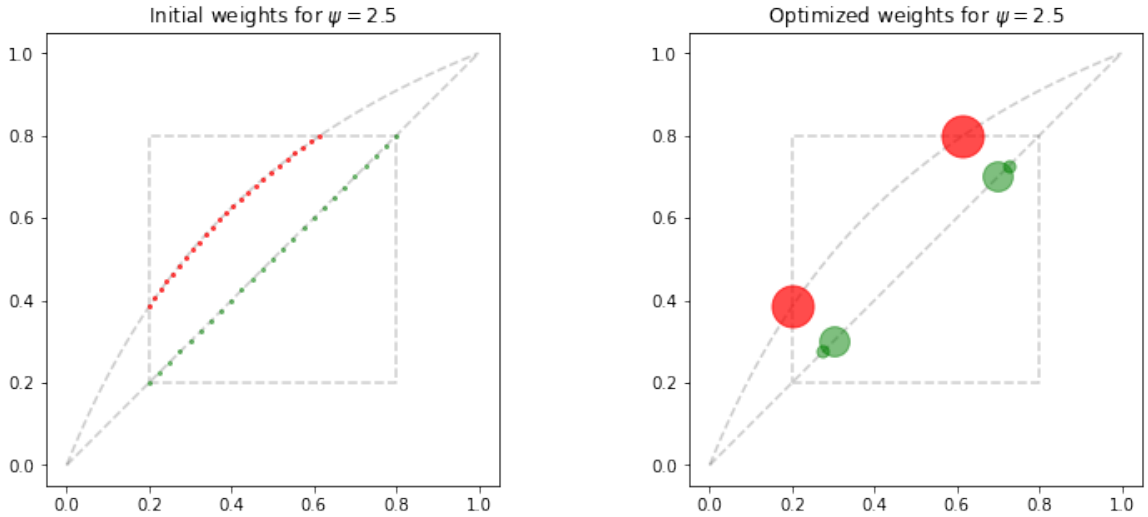

Figure 5.2: The initial and optimized weights for $n_a = n_b = 50$ and the odds ratio. Θ_0' and the lower-right border of $\Theta_1(\underline{\psi})'$ were discretized into 25 evenly spaced points, initialized to the uniform distribution. After optimizing, it can be see than the JIP approximation concentrates its mass on a small number of points. The two pairs of green circles become closer for finer discretizations, indicating that the fact their are four rather than two is an artefact of having to discretize the parameter sets.

5.4 Using conditional S-values in the unconditional setting

Recall the conditional setting, where it is assumed that $N_1 = n_1$ is known in advance. The data can then be summarized by N_{b1}, which has Fisher's noncentral hypergeometric distribution for some odds

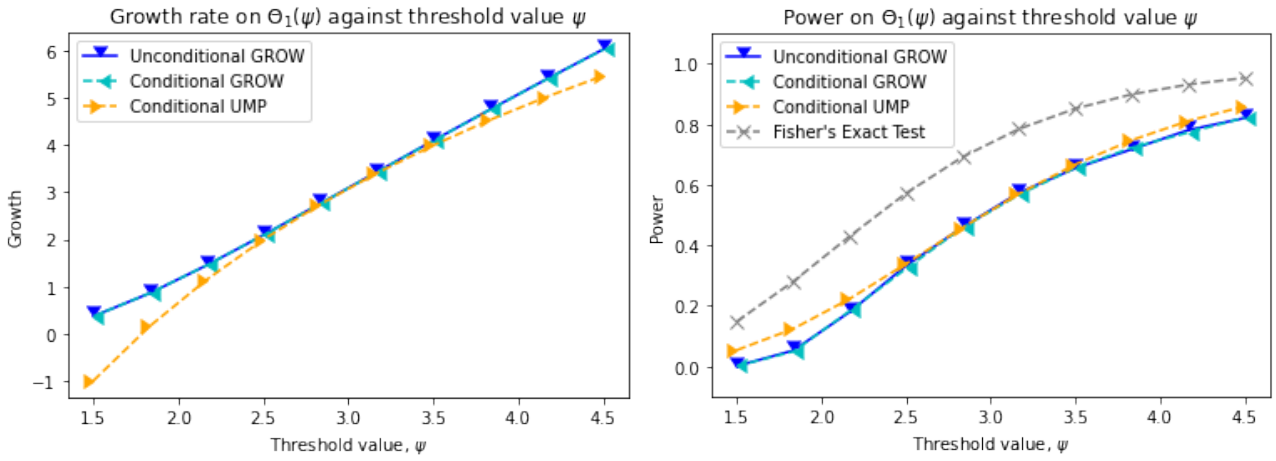

Figure 5.3: The growth and power of the unconditionl GROW, conditional GROW and conditional UMP S-values with Fisher's exact test for comparison. Here $n_a = n_b = 50$ and 10 evenly spaced values between 1.5 and 4.5 have been used for the threshold $\underline{\psi}$.

ratio ψ, namely

$$N_{b1} \sim \text{fnchypg}(n, n_b, n_1, \psi). \tag{5.52}$$

Further, given parameter sets $\Psi_0 = \{1\}$ and $\Psi_1(\underline{\psi}) = [\underline{\psi}, \infty)$ for some threshold value $\underline{\psi}$, we saw that the GROW S-value $S^*_{\underline{\psi}} := S^*_{\Psi(\underline{\psi})}$ is given by

$$S^*_{\underline{\psi}}(N_{b1}) = \frac{P_{\underline{\psi}}(N_{b1})}{P_1(N_{b1})}, \tag{5.53}$$

since the optimal prior $W_1^* \in \mathcal{W}(\Psi_1(\underline{\psi}))$ is always a point mass on $\underline{\psi}$, independently of n_1.

Now, as we saw in Proposition 7, conditional S-values are also unconditional S-values. Thus for any $\psi > 1$, the conditional S-value

$$S_{\psi}(N_{b1}) := \frac{P_{\psi}(N_{b1})}{P_1(N_{b1})} \tag{5.54}$$

(which is indeed a conditional S-value since the null parameter set is a singleton) can be legitimately used for statistical inference even when N_1 is not in fact known in advance and is instead discovered in the course of the experiment. In fact, for any function $W : \{0, 1, \ldots, n_a + n_b\} \to \mathcal{W}(\Psi_1)$, we can define the random variable $S_W(N_{b1}|N_1)$ by

$$S_W(N_{b1}|N_1) := \frac{P_{W(N_1)}(N_{b1})}{P_1(N_{b1})}. \tag{5.55}$$

Then, since $S_W(N_{b1}|n_1)$ is a conditional S-value for each n_1 (again since the null parameter set is a singleton), Proposition 7 gives that $S_W(N_{b1}|N_1)$ is an unconditional S-value.

Suppose we are in the unconditional setting, with 'unconditional' parameter sets Θ_0' and $\Theta_1(\underline{\epsilon})'$ for some threshold value $\underline{\epsilon}$ and prior knowledge rectangle PKR. Then we can define the 'conditional' parameter sets Ψ_0 and Ψ_1 to be the sets of values of the odds ratio consistent with Θ_0' and $\Theta_1(\underline{\epsilon})'$ respectively, namely

$$\Psi_0 := \{\psi(\theta_a, \theta_b) : (\theta_a, \theta_b) \in \Theta_0'\} = \{1\} \quad \text{and} \tag{5.56}$$

$$\Psi_1 := \{\psi(\theta_a, \theta_b) : (\theta_a, \theta_b) \in \Theta_1(\underline{\epsilon})'\} = [\underline{\psi}, \psi_{\max}]. \tag{5.57}$$

Thus the 'unconditional' parameter sets define a threshold value $\underline{\psi}$ for the odds ratio[2] that is independent of N_1. We consider two possibilities for S-values defined as in (5.55). First, we can take

[2]It is worth emphasizing that the parameter ϵ may already be the odds ratio ψ. In this case we have that the induced threshold is equal to the original threshold, so that there is no ambiguity in denoting both by $\underline{\psi}$.

$W(n_1) = \underline{\psi}$ for all n_1, so that $S_W(N_{b1}|N_1)$ is simply the conditional GROW S-value but in the unconditional setting. Second, for any $\psi \geq 1$ and $0 < n_1 < n$, let P_{ψ,n_1} denote the fnchypg(n, n_b, n_1, ψ) distribution. Then, for a given significance level α, we can take $W(n_1) = \psi^*(n_1)$ where $\psi^*(n_1)$ is defined as the unique solution to $d(\psi) = -\log \alpha$, where $d(\psi) \coloneqq \mathrm{KL}(P_{\psi,n_1}||P_{1,n_1})$, as in Proposition 10. Then $S_W(N_{b1}|N_1)$ is simply the conditional UMPG S-value but in the unconditional setting. For an evaluation of these S-values in terms of growth and power, see chapter seven.

Chapter 6

The DOT S-value and bypassing the JIP approximation

Preliminary results of the numerical convex optimization used to find W_1^* and W_0^* seem to show that both of the priors put all their mass on single points provided $\Theta_1(\epsilon)'$ is convex. If this is indeed the case, then it may be possible to find closed formulas for the point masses and therefore for S^*. Thus, the goal of this chapter is to find when the priors W_1^* and W_0^* defining the GROW S-value are in fact point masses.

Recall that for a given Θ_0' and $\Theta_1(\epsilon)'$ the GROW S-value is given by

$$S^*(Z) = \frac{P_{W_1^*}(Z)}{P_{W_0^*}(Z)}, \quad \text{where} \quad (W_1^*, W_0^*) \coloneqq \underset{(W_1,W_0)\in\mathcal{W}(\Theta_1(\epsilon)')\times\mathcal{W}(\Theta_0')}{\arg\min} \mathrm{KL}(P_{W_1}||P_{W_0}). \tag{6.1}$$

If W_1^* and W_0^* happen to be point masses, they must of course be masses on the two points θ_1 and θ_0 minimizing $\mathrm{KL}(P_{\theta_1}||P_{\theta_0})$. Defining

$$\dot{S}^*(Z) \coloneqq \frac{P_{\theta_1^*}(Z)}{P_{\theta_0^*}(Z)}, \quad \text{where} \quad (\theta_1^*, \theta_0^*) \coloneqq \underset{(\theta_1,\theta_0)\in\Theta_1(\epsilon)'\times\Theta_0'}{\arg\min} \mathrm{KL}(P_{\theta_1}||P_{\theta_0}), \tag{6.2}$$

we therefore want to know whether $W_i^* = \delta_{\theta_i^*}$ for $i \in \{0, 1\}$, since this would give $S^* = \dot{S}^*$. We will refer to $\dot{S}^*$ as the *unconditional DOT* S-value[1] even if we do not know whether $S^* = \dot{S}^*$ (this will be justified in Theorem 15 where we see that the DOT S-value is indeed an S-value). At first glance it seems that this alone would greatly reduce the difficulty of computing S^*. However, it is not quite so straightforward; while $\mathrm{KL}(P_{W_1}||P_{W_0})$ is convex in (W_1, W_0) and so can be minimized by simple gradient descent (albeit over a large number of discretization points), $\mathrm{KL}(P_{\theta_1}||P_{\theta_0})$ might not be convex in (θ_1, θ_0) if we are restricting to point masses. However, as we demonstrate in this section, there are cases in which $S^* = \dot{S}^*$ where there exist closed formulas for θ_1^* and θ_0^*. Computing S^* in theses cases is therefore very straightforward. Moreover, in such cases S^* has a cleaner interpretation.

In section 6.1, Theorem 13, we determine, for a fixed θ_1 which θ_0 minimizes $\mathrm{KL}(P_{\theta_1}||P_{\theta_0})$. This is denoted θ_0^*, or $\theta_0^*(\theta_1)$ for clarity. Theorem 15 in section 6.2 then shows that $p_{\theta_1}(Z)/p_{\theta_0^*}(Z)$ is an S-value. Sections 6.3 and 6.4 attempt to generalize this by replacing θ_1 with an arbitrary prior W_1. While we are able to find an expression for the associated θ_0^*, it is seen that the likelihood ratio $p_{W_1}(Z)/p_{\theta_0^*}(Z)$ is not necessarily an S-value. In section 6.5, Theorem 19, we prove that if the parameter sets are convex and $\dot{S}^*$ achieves its minimum growth rate at θ_1^* defined in 6.2, then $\dot{S}^*$ is in fact the GROW S-value S^*. Section 6.6 then explores the different cases of parameter and prior knowledge rectangle, making significant use of Theorem 19 to show that indeed $\dot{S}^* = S^*$ in many cases. The method we use relies on Lemma 20, which shows that the optimal θ_1^* always lies on the boundary of $\Theta_1(\epsilon)'$. This, combined with the aforementioned Theorem 19, makes it substantially easier to find S^* when $S^* = \dot{S}^*$, since the problem is reduced to a one-dimensional minimization rather than a very high-dimensional constrained minimization as when finding the JIP. While it may be tempting to believe that we always have $S^* = \dot{S}^*$, in section 6.7 we conclude the chapter by showing providing a counterexample to this conjecture.

[1] Informally, 'DOT' can be thought of as abbreviating 'deltas on thetas'.

6.1 Finding the closest P_{θ_0} to a fixed P_{θ_1}

For fixed n_a, n_b, PKR and $\theta_1 \in [0,1]^2$, there is a simple geometric way to find

$$\theta_0^* \coloneqq \underset{\theta_0 \in \Theta_0'}{\arg\min}\, \text{KL}(P_{\theta_1} || P_{\theta_0})^2. \tag{6.3}$$

First draw a straight line with gradient $-n_a/n_b$ through θ_1 and label the point where it intersects Θ_0 by $\theta_0^\times = (p^\times, p^\times)$. Then let $\theta_0^\perp$ be the point in Θ_0' that is closest in Euclidean distance to $\theta_0^\times$, namely $\theta_0^\perp = (p^\perp, p^\perp)$ where

$$p^\perp \coloneqq \underset{p \in I_{\text{PKR}}}{\arg\min}\, |p - p^\times|. \tag{6.4}$$

The following theorem then states that $\theta_0^* = \theta_0^\perp$. Since the prior knowledge is always a rectangle, this implies that θ_0^* is to the lower right of θ_1.

Theorem 13. *Fix n_a, n_b, PKR and $\theta_1 \coloneqq (\theta_a, \theta_b) \in [0,1]^2$ and let $\theta_0^* = (p^*, p^*)$ and $\theta_0^\perp = (p^\perp, p^\perp)$ be defined as above. Then*

$$\theta_0^* = \theta_0^\perp. \tag{6.5}$$

Proof. First note that

$$p^\times = \frac{n_a\theta_a + n_b\theta_b}{n}. \tag{6.6}$$

This is easily verified by calculating the gradient between $\theta_1 = (\theta_a, \theta_b)$ and $\theta_0^\times = (p^\times, p^\times)$:

$$\frac{p^\times - \theta_b}{p^\times - \theta_a} = \frac{n_a\theta_a + n_b\theta_b - n\theta_b}{n_a\theta_a + n_b\theta_b - n\theta_a} \tag{6.7}$$

$$= \frac{n_a(\theta_a - \theta_b)}{n_b(\theta_b - \theta_a)} \tag{6.8}$$

$$= -\frac{n_a}{n_b}. \tag{6.9}$$

Now, for all $p \in [0,1]$, let $\theta_0^{(p)} \coloneqq (p,p)$ and define

$$f(p) \coloneqq \text{KL}(P_{\theta_1} || P_{\theta_0^{(p)}}). \tag{6.10}$$

Recall that our prior knowledge is of the form $\text{PKR} = [\theta_a^L, \theta_a^U] \times [\theta_b^L, \theta_b^U]$. Therefore $\Theta_0' = \{p \in [0,1] : \theta_0^{(p)} \in I_{\text{PKR}}\}$, where $I_{\text{PKR}} \coloneqq [\max\{\theta_a^L, \theta_b^L\}, \min\{\theta_a^U, \theta_b^U\}]$. Note that by definition of $\theta_0^* = (p^*, p^*)$, $f(\cdot)$ takes its minimum value over I_{PKR} at p^*. Since projecting $\theta_0^\times$ onto Θ_0' is equivalent to projecting $p^\times$ onto the interval I_{PKR}, we have one of the following cases:

1. $p^\times \in I_{\text{PKR}}$, which implies $p^\perp = p^\times$,
2. $p^\times < \min I_{\text{PKR}}$, which implies $p^\perp = \min I_{\text{PKR}}$, or
3. $p^\times > \max I_{\text{PKR}}$, which implies $p^\perp = \max I_{\text{PKR}}$.

Our goal is to show that $f(\cdot)$ has a unique minimum over I_{PKR} at $p^\perp$, so that $p^* = p^\perp$ and hence $\theta_0^* = \theta_0^\perp$. In all three cases, the result will follow if we can show that the global minimum of $f(\cdot)$ over $[0,1]$ is attained at $p^\times$ and that $f(\cdot)$ is strictly increasing away from $p^\times$. This can be seen by inspecting $f'(\cdot)$ directly as follows. First, by independence,

$$f(p) \coloneqq \text{KL}(P_{\theta_1} || P_{\theta_0^{(p)}}) \tag{6.11}$$

$$= n_a \text{kl}(\theta_a || p) + n_b \text{kl}(\theta_b || p) \tag{6.12}$$

$$= n_a \left(\theta_a \log \frac{\theta_a}{p} + (1-\theta_a) \log \frac{1-\theta_a}{1-p} \right) + n_b \left(\theta_b \log \frac{\theta_b}{p} + (1-\theta_b) \log \frac{1-\theta_b}{1-p} \right). \tag{6.13}$$

[2] We are abusing notation by calling this point θ_0^* since it is dependent upon θ_1. However, the notation $\theta_0^*(\theta_1)$ would be quite cumbersome for many parts of this section. We therefore omit the dependence on θ_1 and trust that it is clear from the context whether $\theta_0^* = \theta_0^*(\theta_1)$ for some θ_1, or whether θ_0^* is defined by (6.2).

Next, taking the derivative with respect to p, we obtain

$$f'(p) = \frac{\mathrm{d}}{\mathrm{d}p}\Big[- n_a\theta_a \log p - n_a(1-\theta_a)\log(1-p) - n_b\theta_b \log p - n_b(1-\theta_b)\log(1-p))\Big] \tag{6.14}$$

$$= -\frac{n_a\theta_a + n_b\theta_b}{p} + \frac{n_a(1-\theta_a) + n_b(1-\theta_b)}{1-p} \tag{6.15}$$

$$= \frac{pn - (n_a\theta_a + n_b\theta_b)}{p(1-p)} \tag{6.16}$$

$$= \frac{n(p-p^{\times})}{p(1-p)}. \tag{6.17}$$

Therefore $f'(p) = 0 \iff p = p^{\times}$ and so $f(\cdot)$ has a unique stationary point at $p^{\times}$. Moreover, since the sign of $f'(p)$ is equal to the sign of $p - p^{\times}$, we see that $f(\cdot)$ is strictly increasing away from $p^{\times}$. □

Corollary 14. *For any $\theta_1 = (\theta_a, \theta_b) \in \Theta_1$, its associated $\theta_0^* = (p^*, p^*)$ lies to the lower right of θ_1, namely*

$$\theta_a \leq p^* \quad \textit{and} \quad \theta_b \geq p^*. \tag{6.18}$$

Proof. With some thought this should be clear geometrically. Nevertheless, we give the following algebraic proof.

Note $\theta_1 \in \Theta_1$ implies $\theta_b \geq \theta_a$. If no projection is required to obtain θ_0^*, then

$$p^* = \frac{n_a\theta_a + n_b\theta_b}{n}, \tag{6.19}$$

which is a mixture of θ_a and θ_b and so clearly $\theta_a \leq p^* \leq \theta_b$. If projection is required, we either have

$$p^{\times} < \min(I_{\mathrm{PKR}}) \quad \text{or} \quad p^{\times} > \max(I_{\mathrm{PKR}}). \tag{6.20}$$

In the first case, we have $p^* = \min(I_{\mathrm{PKR}})$. Suppose, for a contradiction, that (p^*, p^*) does not lie to the lower right of θ_1, namely

$$\theta_a > p^* \quad \text{or} \quad \theta_b < p^*. \tag{6.21}$$

Recalling $I_{\mathrm{PKR}} = [\max\{\theta_a^L, \theta_b^L\}, \min\{\theta_a^U, \theta_b^U\}]$, this implies

$$\theta_a > \max\{\theta_a^L, \theta_b^L\} \quad \text{or} \quad \theta_b < \max\{\theta_a^L, \theta_b^L\}. \tag{6.22}$$

In the first case, since $\theta_b \geq \theta_a$, we have $\theta_a, \theta_b > \max\{\theta_a^L, \theta_b^L\}$. Now $p^{\times}$ is a mixture of θ_a and θ_b and so this implies $p^{\times} > \max\{\theta_a^L, \theta_b^L\} = \min(I_{\mathrm{PKR}})$, which is a contradiction. In the second case we either have $\theta_b < \theta_b^L$, which is clearly a contradiction, or $\theta_b < \theta_a^L$, which implies $\theta_a < \theta_a^L$ (since $\theta_b \geq \theta_a$), which is also a contradiction. Since every option ends in a contradiction, we therefore have that θ_0^* does in fact lie to the lower right of θ_1. By symmetry, this also holds in the second case of (6.20). □

6.2 Showing that the DOT S-value is indeed an S-value

For a fixed $\theta_1 \in [0,1]^2$ and its associated $\theta_0^* \coloneqq \arg\min_{\theta_0 \in \Theta_0'} \mathrm{KL}(P_{\theta_1} || P_{\theta_0})$, the natural question to ask now is whether $P_{\theta_1}(Z)/P_{\theta_0^*}(Z)$ is an S-value. The following theorem shows that the answer is yes, regardless of the parameter choice or prior knowledge.

Theorem 15. *For any n_a, n_b and any $\theta_1 = (\theta_a, \theta_b) \in (0,1)^2$, let $\theta_0^* = (p^*, p^*)$ be defined by*

$$\theta_0^* \coloneqq \operatorname*{arg\,min}_{\theta_0 \in \Theta_0'} \mathrm{KL}(P_{\theta_1} || P_{\theta_0}). \tag{6.23}$$

Then

$$S_{\theta_1, \theta_0^*}(Z) \coloneqq \frac{P_{\theta_1}(Z)}{P_{\theta_0^*}(Z)} \tag{6.24}$$

is an S-value.

Proof. Define $R(p) \coloneqq \mathbf{E}_{Z\sim P_{\theta_0^{(p)}}}[S_{\theta_1,\theta_0^*}]$ for $p \in [0,1]$. We need to show that $R(p) \leq 1$ for all $p \in I_{\text{PKR}}$. In fact, we will show that $R(p) \leq 1$ for all $p \in [0,1]$. We do this by showing that $R(\cdot)$ achieves its maximum value at p^* which suffices since

$$R(p^*) = \mathbf{E}_{Z\sim P_{\theta_0^*}}\left[\frac{P_{\theta_1}(Z)}{P_{\theta_0^*}(Z)}\right] = \sum_z \frac{P_{\theta_0^*}(z)P_{\theta_1}(z)}{P_{\theta_0^*}(z)} = \sum_z P_{\theta_1}(z) = 1. \tag{6.25}$$

We first expand the definition of $R(p)$. By independence,

$$\begin{aligned} R(p) &\coloneqq \mathbf{E}_{Z\sim P_{\theta_0^{(p)}}}\left[\frac{P_{\theta_1}(Z)}{P_{\theta_0^*}(Z)}\right] && (6.26)\\ &= \mathbf{E}_{N_{a1}\sim\text{Bin}(n_a,p),N_{b1}\sim\text{Bin}(n_b,p)}\left[\frac{P_{\theta_1}(N_{a1},N_{b1})}{P_{\theta_0^*}(N_{a1},N_{b1})}\right] && (6.27)\\ &= \mathbf{E}_{N_{a1}\sim\text{Bin}(n_a,p)}\left[\frac{P_{\theta_1}(N_{a1})}{P_{\theta_0^*}(N_{a1})}\right]\mathbf{E}_{N_{b1}\sim\text{Bin}(n_b,p)}\left[\frac{P_{\theta_1}(N_{b1})}{P_{\theta_0^*}(N_{b1})}\right]. && (6.28)\end{aligned}$$

Looking at just the first of these terms for the moment, we have

$$\begin{aligned} \mathbf{E}_{N_{a1}\sim\text{Bin}(n_a,p)}\left[\frac{P_{\theta_1}(N_{a1})}{P_{\theta_0^*}(N_{a1})}\right] &= \mathbf{E}_{N_{a1}\sim\text{Bin}(n_a,p)}\left[\frac{\theta_a^{N_{a1}}(1-\theta_a)^{n_a-N_{a1}}}{(p^*)^{N_{a1}}(1-p^*)^{n_a-N_{a1}}}\right] && (6.29)\\ &= \left(\frac{1-\theta_a}{1-p^*}\right)^{n_a}\mathbf{E}_{N_{a1}\sim\text{Bin}(n_a,p)}\left[\left(\frac{\theta_a(1-p^*)}{(1-\theta_a)p^*}\right)^{N_{a1}}\right] && (6.30)\\ &= \left(\frac{1-\theta_a}{1-p^*}\right)^{n_a}\left[1-p+p\frac{\theta_a(1-p^*)}{(1-\theta_a)p^*}\right]^{n_a} \quad \text{using the PGF} && (6.31)\\ &= \left(\frac{1-\theta_a}{1-p^*}\right)^{n_a}\left[1+p\frac{\theta_a-p^*}{(1-\theta_a)p^*}\right]^{n_a}. && (6.32)\end{aligned}$$

Note that it is permitted to divide by $1-\theta_a$ and $1-\theta_b$ since $\theta_1 \in (0,1)^2$. Further, we can divide by $1-p^*$ since $\theta_a, \theta_b < 1$ implies $p^\times < 1$ and so $p^* < 1$. Combining the two terms again, we have

$$\begin{aligned} R(p) &= \left(\frac{1-\theta_a}{1-p^*}\right)^{n_a}\left(\frac{1-\theta_b}{1-p^*}\right)^{n_b}\left[1+p\frac{\theta_a-p^*}{(1-\theta_a)p^*}\right]^{n_a}\left[1+p\frac{\theta_b-p^*}{(1-\theta_b)p^*}\right]^{n_b} && (6.33)\\ &= \gamma[1+c_a p]^{n_a}[1+c_b p]^{n_b}, && (6.34)\end{aligned}$$

where

$$\gamma \coloneqq \left(\frac{1-\theta_a}{1-p^*}\right)^{n_a}\left(\frac{1-\theta_b}{1-p^*}\right)^{n_b} \quad \text{and} \quad c_i \coloneqq \frac{\theta_i-p^*}{(1-\theta_i)p^*} \text{ for } i \in \{a,b\}. \tag{6.35}$$

Since $\theta_1 \in (0,1)^2$, we know that $\gamma > 0$ and

$$c_i = \frac{\theta_i-p^*}{(1-\theta_i)p^*} > \theta_i - p^* \geq 0 - 1 = -1. \tag{6.36}$$

Therefore $1 + c_i p > 0$ for $i \in \{a,b\}$ and we are justified writing the logarithm of $R(\cdot)$ as

$$r(p) \coloneqq \log R(p) = \log\gamma + n_a\log(1+c_a p) + n_b\log(1+c_b p). \tag{6.37}$$

Taking the derivative, we see that

$$r'(p) = \frac{n_a c_a}{1+c_a p} + \frac{n_b c_b}{1+c_b p} \tag{6.38}$$

and so

$$\begin{aligned} r'(p) = 0 &\iff \frac{n_a c_a}{1+c_a p} + \frac{n_b c_b}{1+c_b p} = 0 && (6.39)\\ &\iff n_a c_a(1+c_b p) + n_b c_b(1+c_a p) = 0 && (6.40)\\ &\iff p = p_0 \coloneqq -\frac{n_a c_a + n_b c_b}{n c_a c_b}, && (6.41)\end{aligned}$$

so that $r(\cdot)$ has a single stationary point at p_0, provided $p_0 \in [0,1]$. Inspecting the second derivative, we have

$$r''(p) = -\frac{n_a c_a^2}{(1+c_a p)^2} - \frac{n_b c_b^2}{(1+c_b p)^2}, \tag{6.42}$$

which is nonnegative for all $p \in [0,1]$. Therefore $r(\cdot)$ is (weakly) concave on $[0,1]$ and thus on $I_{\mathrm{PKR}} \subseteq [0,1]$.

Going back to (6.38), we see that

$$\begin{aligned}
r'(p^*) &= \frac{n_a}{1/c_a + p^*} + \frac{n_b}{1/c_b + p^*} && (6.43)\\
&= \frac{n_a}{\frac{(1-\theta_a)p^*}{\theta_a - p^*} + p^*} + \frac{n_b}{\frac{(1-\theta_b)p^*}{\theta_b - p^*} + p^*} && (6.44)\\
&= \frac{n_a(\theta_a - p^*)}{(1-\theta_a)p^* + p^*(\theta_a - p^*)} + \frac{n_b(\theta_b - p^*)}{(1-\theta_b)p^* + p^*(\theta_b - p^*)} && (6.45)\\
&= \frac{n_a(\theta_a - p^*)}{p^*(1-p^*)} + \frac{n_b(\theta_b - p^*)}{p^*(1-p^*)} && (6.46)\\
&= \frac{n_a\theta_a + n_b\theta_b - np^*}{p^*(1-p^*)} && (6.47)\\
&= \frac{n(p^\times - p^*)}{p^*(1-p^*)}. && (6.48)
\end{aligned}$$

Thus $r'(p^*)$ has the same sign as $p^\times - p^*$. We now conclude the proof by considering the three possible cases.

1. If $p^\times \in I_{\mathrm{PKR}}$, then $p^* = p^\times$. This implies that $r'(p^*) = 0$ and so $p^* = p_0$, the unique stationary point of $r(\cdot)$. Since $r(\cdot)$ is (weakly) concave, this implies that $r(\cdot)$ achieves its maximum over $[0,1]$ at p^*. Since the map $x \mapsto \log x$ is increasing, we therefore see that $R(\cdot)$ also achieves its maximum over $[0,1]$ at p^*.

2. If $p^\times < \min I_{\mathrm{PKR}}$ then $p^* = \min I_{\mathrm{PKR}}$. Therefore $p^\times < p^*$ and we see from (6.48) that $r'(\min I_{\mathrm{PKR}}) = r'(p^*) < 0$. Since $r(\cdot)$ is (weakly) concave, this implies that $r(\cdot)$ is non-increasing on I_{PKR}. Since the map $x \mapsto \log x$ is increasing, we therefore see that $R(\cdot)$ is also non-increasing on I_{PKR}. Thus $R(\cdot)$ attains its maximum value at $\min I_{\mathrm{PKR}} = p^*$.

3. If $p^\times > \max I_{\mathrm{PKR}}$ then $p^* = \max I_{\mathrm{PKR}}$ and the rest follows similarly to case 2.

Thus in all three cases $R(\cdot)$ achieves its maximum over $[0,1]$ at p^*, where $R(p^*) = 1$. Therefore $R(p) \le 1$ for all $p \in I_{\mathrm{PKR}} \subseteq [0,1]$ and $S_{\theta_1,\theta_0^*} \coloneqq \frac{P_{\theta_1}}{P_{\theta_0^*}}$ is an S-value. □

Since θ_1 may be chosen arbitrarily, the above theorem holds for $\theta_1 = \theta_1^*$ as defined in (6.2), proving that $\dot{S}^*(Z) = P_{\theta_1^*}(Z)/P_{\theta_0^*}(Z)$ is an S-value. We state this as a corollary.

Corollary 16. *For any n_a, n_b, let*

$$\dot{S}^*(Z) \coloneqq \frac{P_{\theta_1^*}(Z)}{P_{\theta_0^*}(Z)}, \quad \textit{where} \quad (\theta_1^*, \theta_0^*) \coloneqq \operatorname*{arg\,min}_{(\theta_1,\theta_0)\in\Theta_1'\times\Theta_0'} \mathrm{KL}(P_{\theta_1}||P_{\theta_0}). \tag{6.49}$$

Then $\dot{S}^*$ is an S-value.

6.3 Finding the closest P_{θ_0} to a fixed P_{W_1}

Suppose we have a fixed prior $W_1 \in \mathcal{W}(\Theta_1)$ on the alternative parameter set and we form the test statistic

$$T_{W_1}(Z) \coloneqq \frac{P_{W_1}(Z)}{P_{\theta_0^*}(Z)} \quad \text{where} \quad \theta_0^* \coloneqq \operatorname*{arg\,min}_{\theta_0\in\Theta_0'} \mathrm{KL}(P_{W_1}||P_{\theta_0}). \tag{6.50}$$

Is T_{W_1} then an S-value? In general, the answer is no. For example, suppose we are interested in the risk difference with threshold $\delta = 0.05$, we have no prior knowledge (namely PKR $= [0,1]^2$) and $n_a = n_b$. If we take the prior $W_1 \in \mathcal{W}(\Theta_1(\delta))$ that is uniform across the boundary of $\Theta_1(\delta)$, namely

$$\theta_a \sim U[0, 0.95] \quad \text{and} \quad \theta_b = \theta_a + 0.05, \tag{6.51}$$

then we will show that T_{W_1} is not an S-value. We begin with the following lemma which gives a closed form expression for θ_0^*.

Lemma 17. *Fix n_a, n_b and $W_1 \in \mathcal{W}(\Theta_1)$. Suppose we have no prior knowledge, namely,* PKR $= [0,1]^2$. *Then $\theta_0^* \coloneqq \arg\min_{\theta_0 \in \Theta_0'} \mathrm{KL}(P_{W_1} || P_{\theta_0})$ is given by $\theta_0^* = (p^*, p^*)$, where*

$$p^* = \frac{n_a \mathbf{E}_{W_1}[\theta_a] + n_b \mathbf{E}_{W_1}[\theta_b]}{n}. \tag{6.52}$$

Proof. For all $p \in [0,1]$, define $f(p) \coloneqq \mathrm{KL}(P_{W_1} || P_{\theta_0^{(p)}})$. Then

$$f(p) = \mathbf{E}_{Z \sim P_{W_1}}\left[\log \frac{P_{W_1}(Z)}{P_{\theta_0^{(p)}}(Z)}\right] \tag{6.53}$$

$$= \mathbf{E}_{Z \sim P_{W_1}}\left[\log P_{W_1}(Z)\right] - \mathbf{E}_{Z \sim P_{W_1}}\left[\log P_{\theta_0^{(p)}}(Z)\right]. \tag{6.54}$$

Inspecting the second term and using independence, we see that

$$\mathbf{E}_{Z \sim P_{W_1}}\left[\log P_{\theta_0^{(p)}}(Z)\right] = \mathbf{E}_{Z \sim P_{W_1}}\left[\log\left(P_{\theta_0^{(p)}}(N_{a1}) P_{\theta_0^{(p)}}(N_{b1})\right)\right] \tag{6.55}$$

$$= \mathbf{E}_{Z \sim P_{W_1}}\left[\log\left(\binom{n_a}{N_{a1}} p^{N_{a1}}(1-p)^{n_a - N_{a1}} \binom{n_b}{N_{b1}} p^{N_{b1}}(1-p)^{n_b - N_{b1}}\right)\right] \tag{6.56}$$

$$= \log\left(\binom{n_a}{N_{a1}}\binom{n_b}{N_{b1}}\right) + n\log(1-p) + \log\left(\frac{p}{1-p}\right)\mathbf{E}_{Z \sim P_{W_1}}\left[N_{a1} + N_{b1}\right]. \tag{6.57}$$

Since the first terms in (6.54) and (6.57) are independent of p, they disappear when we take the derivative of f with respect to p. Thus

$$f'(p) = \frac{\mathrm{d}}{\mathrm{d}p}\left[-n\log(1-p) - \log\left(\frac{p}{1-p}\right)\mathbf{E}_{Z \sim P_{W_1}}\left[N_{a1} + N_{b1}\right]\right] \tag{6.58}$$

$$= \frac{n}{1-p} - \left(\frac{1}{p} + \frac{1}{1-p}\right)\mathbf{E}_{Z \sim P_{W_1}}\left[N_{a1} + N_{b1}\right] \tag{6.59}$$

$$= \frac{n}{1-p} - \frac{1}{p(1-p)}\mathbf{E}_{Z \sim P_{W_1}}\left[N_{a1} + N_{b1}\right]. \tag{6.60}$$

Inspecting the expectation, we have

$$\mathbf{E}_{Z \sim P_{W_1}}\left[N_{a1} + N_{b1}\right] = \mathbf{E}_{(\theta_a, \theta_b) \sim W_1}\left[\mathbf{E}_{Z \sim P_\theta}\left[N_{a1} + N_{b1}\right]\right] \tag{6.61}$$

$$= \mathbf{E}_{(\theta_a, \theta_b) \sim W_1}\left[n_a\theta_a + n_b\theta_b\right] \tag{6.62}$$

$$= n_a \mathbf{E}_{(\theta_a, \theta_b) \sim W_1}\left[\theta_a\right] + n_b \mathbf{E}_{(\theta_a, \theta_b) \sim W_1}\left[\theta_b\right]. \tag{6.63}$$

Setting the derivative equal to zero, we see that

$$f'(p) = 0 \iff \frac{n}{1-p} - \frac{1}{p(1-p)}\left(n_a \mathbf{E}_{(\theta_a, \theta_b) \sim W_1}\left[\theta_a\right] + n_b \mathbf{E}_{(\theta_a, \theta_b) \sim W_1}\left[\theta_b\right]\right) = 0 \tag{6.64}$$

$$\iff p = p_0 \coloneqq \frac{n_a \mathbf{E}_{(\theta_a, \theta_b) \sim W_1}\left[\theta_a\right] + n_b \mathbf{E}_{(\theta_a, \theta_b) \sim W_1}\left[\theta_b\right]}{n}. \tag{6.65}$$

Note that $\mathbf{E}_{W_1}[\theta_x] \in [0,1]$ for $x \in \{a, b\}$, so p_0, being a mixture of $\mathbf{E}_{W_1}[\theta_a]$ and $\mathbf{E}_{W_1}[\theta_b]$, is also in $[0,1]$. Substituting p_0 back into the expression for $f'(p)$, we have

$$f'(p) = \frac{n}{1-p} - \frac{np_0}{p(1-p)} \tag{6.66}$$

and so the second derivative is

$$f''(p) = \frac{n}{(1-p)^2} + \frac{np_0}{p^2(1-p)^2}(1-2p) \tag{6.67}$$

$$= \frac{n}{(1-p)^2}\left[1 + \frac{p_0(1-2p)}{p^2}\right]. \tag{6.68}$$

Therefore

$$f''(p_0) \geq 0 \iff 1 + \frac{1-2p_0}{p_0} \geq 0 \tag{6.69}$$

$$\iff 1 - 2p_0 \geq -p_0 \tag{6.70}$$

$$\iff p_0 \leq 1. \tag{6.71}$$

Since $p_0 \in [0,1]$, we conclude that p_0 is the unique global minimum of $f(\cdot)$ on $[0,1]$. □

6.4 For fixed $W_1 \in \mathcal{W}(\Theta_1)$, the closest P_{θ_0} does not necessarily give an S-value

We can now use the closed form expression for θ_0^* given by Lemma 17 to show that T_{W_1} is not necessarily an S-value for arbitrary W_1.

Lemma 18. *Let $W_1 \in \mathcal{W}(\Theta_1)$ where $\Theta_1 \subseteq [0,1]^2$ is arbitrary. Define the test statistic T_{W_1} by*

$$T_{W_1}(Z) \coloneqq \frac{P_{W_1}(Z)}{P_{\theta_0^*}(Z)} \quad \text{where} \quad \theta_0^* \coloneqq \operatorname*{arg\,min}_{\theta_0 \in \Theta_0'} \mathrm{KL}(P_{W_1} || P_{\theta_0}). \tag{6.72}$$

Then T_{W_1} is not in general an S-value.

Proof. For a fixed $W_1 \in \mathcal{W}(\Theta_1)$, let $p = p^*$ be defined as in Lemma 17, so that

$$T_{W_1}(Z) = \frac{P_{W_1}(Z)}{P_{\theta_0^{(p)}}(Z)}. \tag{6.73}$$

For all $q \in [0,1]$, let $g(q)$ be the expectation of T_{W_1} when $Z \sim P_{\theta_0^{(q)}}$. Then

$$g(q) \coloneqq \mathbf{E}_{Z \sim P_{\theta_0^{(q)}}}\left[\frac{P_{W_1}(Z)}{P_{\theta_0^{(p)}}(Z)}\right] \tag{6.74}$$

$$= \mathbf{E}_{Z \sim P_{\theta_0^{(q)}}}\left[\frac{\mathbf{E}_{\theta_1 \sim W_1}\left[P_{\theta_1}(Z)\right]}{P_{\theta_0^{(p)}}(Z)}\right] \tag{6.75}$$

$$= \mathbf{E}_{Z \sim P_{\theta_0^{(q)}}}\left[\mathbf{E}_{\theta_1 \sim W_1}\left[\frac{P_{\theta_1}(Z)}{P_{\theta_0^{(p)}}(Z)}\right]\right] \tag{6.76}$$

$$= \mathbf{E}_{\theta_1 \sim W_1}\left[\mathbf{E}_{Z \sim P_{\theta_0^{(q)}}}\left[\frac{P_{\theta_1}(Z)}{P_{\theta_0^{(p)}}(Z)}\right]\right] \quad (\textit{Fubini}) \tag{6.77}$$

$$= \mathbf{E}_{(\theta_a,\theta_b) \sim W_1}\left[\mathbf{E}_{(N_{a1},N_{b1}) \sim P_{\theta_0^{(q)}}}\left[\frac{\binom{n_a}{N_{a1}}\theta_a^{N_{a1}}(1-\theta_a)^{n_a-N_{a1}}\binom{n_b}{N_{b1}}\theta_b^{N_{b1}}(1-\theta_b)^{n_b-N_{b1}}}{\binom{n_a}{N_{a1}}p^{N_{a1}}(1-p)^{n_a-N_{a1}}\binom{n_b}{N_{b1}}p^{N_{b1}}(1-p)^{n_b-N_{b1}}}\right]\right] \tag{6.78}$$

$$= \mathbf{E}_{(\theta_a,\theta_b) \sim W_1}\Bigg[\frac{(1-\theta_a)^{n_a}(1-\theta_b)^{n_b}}{(1-p)^n} \times \mathbf{E}_{N_{a1} \sim \mathrm{Bin}(n_a,q)}\left[\left(\frac{\theta_a(1-p)}{p(1-\theta_a)}\right)^{N_{a1}}\right]\mathbf{E}_{N_{b1} \sim \mathrm{Bin}(n_b,q)}\left[\left(\frac{\theta_b(1-p)}{p(1-\theta_b)}\right)^{N_{b1}}\right]\Bigg], \tag{6.79}$$

where we have used independence in the last two lines. The use of Fubini is justified since the integrand is nonnegative. Inspecting one of the inner expectations of the final line and using the formula for the probability generating function of a binomial random variable, we obtain

$$\mathbf{E}_{N_{a1}\sim\mathrm{Bin}(n_a,q)}\left[\left(\frac{\theta_a(1-p)}{p(1-\theta_a)}\right)^{N_{a1}}\right] = \left[1+q\left(\frac{\theta_a(1-p)}{p(1-\theta_a)}-1\right)\right]^{n_a} \tag{6.80}$$

$$= \left[1+q\frac{\theta_a-p}{p(1-\theta_a)}\right]^{n_a}. \tag{6.81}$$

$$\tag{6.82}$$

Substituting this back into (6.79) we get,

$$g(q) = \mathbf{E}_{(\theta_a,\theta_b)\sim W_1}\left[\frac{(1-\theta_a)^{n_a}(1-\theta_b)^{n_b}}{(1-p)^n}\left[1+q\frac{\theta_a-p}{p(1-\theta_a)}\right]^{n_a}\left[1+q\frac{\theta_b-p}{p(1-\theta_b)}\right]^{n_b}\right] \tag{6.83}$$

$$= \frac{1}{p^n(1-p)^n}\mathbf{E}_{(\theta_a,\theta_b)\sim W_1}\Big[\left[p(1-\theta_a)+q(\theta_a-p)\right]^{n_a}\left[p(1-\theta_b)+q(\theta_b-p)\right]^{n_b}\Big] \tag{6.84}$$

$$= \frac{1}{p^n(1-p)^n}\mathbf{E}_{(\theta_a,\theta_b)\sim W_1}\Big[\left[p(1-q)+\theta_a(q-p)\right]^{n_a}\left[p(1-q)+\theta_b(q-p)\right]^{n_b}\Big]. \tag{6.85}$$

To prove the lemma, we need to find an example where this is larger than 1. First, suppose we have $n_a = n_b = 1$. Then

$$p = \frac{\mathbf{E}[\theta_a]+\mathbf{E}[\theta_b]}{2}. \tag{6.86}$$

This allows us to simplify $g(q)$ as follows

$$g(q) = \frac{1}{p^2(1-p)^2}\mathbf{E}_{(\theta_a,\theta_b)\sim W_1}\Big[\left[p(1-q)+\theta_a(q-p)\right]\left[p(1-q)+\theta_b(q-p)\right]\Big] \tag{6.87}$$

$$= \frac{1}{p^2(1-p)^2}\mathbf{E}_{(\theta_a,\theta_b)\sim W_1}\Big[p^2(1-q)^2+p(1-q)(q-p)[\theta_a+\theta_b]+(q-p)^2\theta_a\theta_b\Big] \tag{6.88}$$

$$= \frac{1}{p^2(1-p)^2}\Big[p^2(1-q)^2+p(1-q)(q-p)\big(\mathbf{E}[\theta_a]+\mathbf{E}[\theta_b]\big)+(q-p)^2\mathbf{E}[\theta_a\theta_b]\Big] \tag{6.89}$$

$$= \frac{1}{p^2(1-p)^2}\Big[p^2(1-q)^2+2p^2(1-q)(q-p)+(q-p)^2\mathbf{E}[\theta_a\theta_b]\Big] \quad (\textit{using } (6.86)) \tag{6.90}$$

$$= \frac{1}{(1-p)^2}\left[(1-q)^2+2(1-q)(q-p)+(q-p)^2\frac{\mathbf{E}[\theta_a\theta_b]}{p^2}\right] \tag{6.91}$$

$$= \frac{1}{(1-p)^2}\left[(1-q)^2+2(1-q)(q-p)+(q-p)^2\right]+\frac{(q-p)^2}{(1-p)^2}\left[\frac{\mathbf{E}[\theta_a\theta_b]}{p^2}-1\right] \tag{6.92}$$

$$= \frac{\big((1-q)+(q-p)\big)^2}{(1-p)^2}+\frac{(q-p)^2}{(1-p)^2}\left[\frac{\mathbf{E}[\theta_a\theta_b]}{p^2}-1\right] \tag{6.93}$$

$$= 1+\frac{(q-p)^2}{(1-p)^2}\left[\frac{\mathbf{E}[\theta_a\theta_b]}{p^2}-1\right]. \tag{6.94}$$

$$\tag{6.95}$$

Thus $P_{W_1}(Z)/P_{\theta_0^{(p)}}(Z)$ is an S-value iff $g(q)\le 1$ for all $q\in[0,1]$. For any particular q, we have

$$g(q)\le 1 \iff \frac{(q-p)^2}{(1-p)^2}\left[\frac{\mathbf{E}[\theta_a\theta_b]}{p^2}-1\right]\le 0 \tag{6.96}$$

$$\iff q=p \quad \text{or} \quad \mathbf{E}[\theta_a\theta_b]\le p^2. \tag{6.97}$$

Recalling the definition of p, we have

$$\mathbf{E}[\theta_a\theta_b]\le p^2 \iff \mathbf{E}[\theta_a\theta_b]\le\left(\frac{\mathbf{E}[\theta_a]+\mathbf{E}[\theta_b]}{2}\right)^2 \tag{6.98}$$

$$\iff 4\mathbf{E}[\theta_a\theta_b]\le\mathbf{E}[\theta_a]^2+2\mathbf{E}[\theta_a]\mathbf{E}[\theta_b]+\mathbf{E}[\theta_b]^2. \tag{6.99}$$

Now suppose $\Theta_1 \subseteq \{(\theta_a, \theta_b) \in [0,1]^2 : \theta_b \geq \theta_a\}$ so that $\mathbf{E}[\theta_a] \leq \mathbf{E}[\theta_b]$. Using the fact that the function $x \mapsto x^2$ is convex, we also have $\mathbf{E}[\theta_a]^2 \leq \mathbf{E}[\theta_a^2]$. Thus (6.99) implies

$$\mathbf{E}[\theta_a \theta_b] \leq p^2 \implies 4\mathbf{E}[\theta_a^2] \leq \mathbf{E}[\theta_a^2] + 3\mathbf{E}[\theta_b]^2 \tag{6.100}$$

$$\implies \mathbf{E}[\theta_a^2] \leq \mathbf{E}[\theta_b]^2. \tag{6.101}$$

We now provide an example for which (6.101) does not hold. As outlined previously, we consider the risk difference with arbitrary threshold $\delta > 0$ in the case of no prior knowledge (PKR $= [0,1]^2$). Now let $W_1 \in \mathcal{W}(\Theta_1(\delta))$ be uniform across the boundary of $\Theta_1(\delta)$, namely

$$\theta_a \sim U[0, 1-\delta] \quad \text{and} \quad \theta_b = \theta_a + \delta. \tag{6.102}$$

Recall that for a uniformly distributed random variable $X \sim \mathrm{U}[\alpha, \beta]$, the first second moments are given by

$$\mathbf{E}[X] = \frac{\alpha + \beta}{2} \quad \text{and} \quad \mathbf{E}[X^2] = \frac{\alpha^2 + \alpha\beta + \beta^2}{3}. \tag{6.103}$$

Therefore

$$\mathbf{E}[\theta_a^2] = \frac{(1-\delta)^2}{3} \quad \text{and} \quad \mathbf{E}[\theta_b]^2 = \frac{(1+\delta)^2}{4} \tag{6.104}$$

If we now set let $\delta = 0.05$, we have

$$\mathbf{E}[\theta_a^2] = \frac{0.95^2}{3} \approx 0.30 \quad \text{and} \quad \mathbf{E}[\theta_b]^2 = \frac{1.05^2}{4} \approx 0.28 \tag{6.105}$$

and so $\mathbf{E}[\theta_a^2] > \mathbf{E}[\theta_b]^2$. Working backwards, we now see that (6.101) fails and so (6.97) fails for any $q \neq p$. Therefore $g(q) > 1$ for all $q \neq p$ and T_{W_1} is not an S-value. □

6.5 A sufficient condition for the DOT S-value to be GROW

Suppose Θ_1 is convex and define the following

$$(\theta_1^*, \theta_0^*) \coloneqq \underset{(\theta_1,\theta_0)\in\Theta_1\times\Theta_0}{\arg\min} \mathrm{KL}(P_{\theta_1} || P_{\theta_0}) \quad \text{and} \tag{6.106}$$

$$(W_1^*, W_0^*) \coloneqq \underset{(W_1,W_0)\in\mathcal{W}(\Theta_1)\times\mathcal{W}(\Theta_0)}{\arg\min} \mathrm{KL}(P_{W_1} || P_{W_0}). \tag{6.107}$$

Our aim now is to show that W_1^* and W_0^* are simply point masses on θ_1^* and θ_0^* respectively, so that $P_{\theta_1^*}(Z)/P_{\theta_0^*}(Z)$ is in fact the GROW S-value. This will make finding the GROW S-value substantially less computationally intensive. The following theorem gives a sufficient condition for this to be true. It will be extensively used in the next section to find closed form or implicit expressions for GROW S-values that dramatically speed up their computation. Note that this theorem is not specific to the setting of 2×2 contingency tables, so that the methods developed in the next section may prove to be useful in other statistical settings.

Theorem 19. *Suppose the conditions of Theorem 1 hold and that in addition Θ_1 and Θ_0 are convex. Define the DOT S-value by*

$$\dot{S}^*(Z) \coloneqq \frac{P_{\theta_1^*}(Z)}{P_{\theta_0^*}(Z)}, \quad \textit{where} \quad (\theta_1^*, \theta_0^*) \coloneqq \underset{(\theta_1,\theta_0)\in\Theta_1\times\Theta_0}{\arg\min} \mathrm{KL}(P_{\theta_1} || P_{\theta_0}). \tag{6.108}$$

For each $\theta_1 \in \Theta_1$, let $g(\theta) \coloneqq \mathbf{E}_{Z\sim P_{\theta_1}}[\log \dot{S}^*]$ be the growth rate of $\dot{S}^*$ at θ_1. If $g(\cdot)$ achieves its minimum at θ_1^* then $P_\theta(S^* = \dot{S}^*) = 1$ for all $\theta \in \Theta_0 \cup \Theta_1$, so that it is irrelevant from an inference perspective which we use.

Proof. Let $\theta_1^*, \theta_0^*, W_1^*$ and W_0^* be defined as above. For each $\alpha \in [0,1]$, let $W_{i,\alpha} \coloneqq (1-\alpha)\delta_{\theta_i^*} + \alpha W_i^*$ for $i \in \{0,1\}$, where δ_θ denotes the point mass on θ. Now the function

$$f(\alpha) \coloneqq \mathrm{KL}(P_{W_{1,\alpha}} || P_{W_{0,\alpha}}) \tag{6.109}$$

is convex since the KL divergence is jointly convex. Further, for any $\alpha \in [0,1]$, we have

$$f(\alpha) = \mathrm{KL}(P_{W_{1,\alpha}} || P_{W_{0,\alpha}}) \geq \min_{(W_1,W_0) \in \mathcal{W}(\Theta_0) \times \mathcal{W}(\Theta_1)} \mathrm{KL}(P_{W_1} || P_{W_0}) = \mathrm{KL}(P_{W_1^*} || P_{W_0^*}) = f(1), \tag{6.110}$$

so f achieves its minimum value at 1. We will now show that $f'(0) \geq 0$. To see that this proves the result, note that the continuity and convexity of f then implies $f(0) \leq f(1) = \min_{\alpha \in [0,1]} f(\alpha)$ and so $f(0) = f(1)$. Thus

$$\mathrm{KL}(P_{\theta_1^*} || P_{\theta_0^*}) = \mathrm{KL}(P_{W_1^*} || P_{W_0^*}), \tag{6.111}$$

which, since we are assuming the conditions of Theorem 1 hold (in particular that the infimum over priors is obtained uniquely), implies that W_1^* and W_0^* are point masses on θ_1^* and θ_0^* respectively, and so clearly S^* and $\dot{S}^*$ are almost surely equal. We will check that $f'(0) \geq 0$ by calculating and simplifying the derivative. First note that for $i \in \{0,1\}$ we have

$$P_{W_{i,\alpha}} = P_{(1-\alpha)\delta_{\theta_i^*} + \alpha W_i^*} = (1-\alpha) P_{\theta_i^*} + \alpha P_{W_i^*}. \tag{6.112}$$

Using this, we have

$$\begin{aligned}
f(\alpha) &\coloneqq \mathrm{KL}(P_{W_{1,\alpha}} || P_{W_{0,\alpha}}) && (6.113)\\
&= D\big((1-\alpha)P_{\theta_1^*} + \alpha P_{W_1^*} || (1-\alpha)P_{\theta_0^*} + \alpha P_{W_0^*}\big) && (6.114)\\
&= \mathbf{E}_{Z \sim (1-\alpha)P_{\theta_1^*}(Z) + \alpha P_{W_1^*}(Z)} \left[\log \frac{(1-\alpha)P_{\theta_1^*}(Z) + \alpha P_{W_1^*}(Z)}{(1-\alpha)P_{\theta_0^*}(Z) + \alpha P_{W_0^*}(Z)} \right] && (6.115)\\
&= (1-\alpha)\mathbf{E}_{Z \sim P_{\theta_1^*}(Z)} \left[\log \frac{(1-\alpha)P_{\theta_1^*}(Z) + \alpha P_{W_1^*}(Z)}{(1-\alpha)P_{\theta_0^*}(Z) + \alpha P_{W_0^*}(Z)} \right] && (6.116)\\
&\quad + \alpha \mathbf{E}_{Z \sim P_{W_1^*}(Z)} \left[\log \frac{(1-\alpha)P_{\theta_1^*}(Z) + \alpha P_{W_1^*}(Z)}{(1-\alpha)P_{\theta_0^*}(Z) + \alpha P_{W_0^*}(Z)} \right]. && (6.117)
\end{aligned}$$

Assuming we can pass the derivative through the expectation, we obtain

$$\begin{aligned}
f'(0) &= -\mathbf{E}_{Z \sim P_{\theta_1^*}} \left[\log \frac{P_{\theta_1^*}(Z)}{P_{\theta_0^*}(Z)} \right] + \mathbf{E}_{Z \sim P_{\theta_1^*}} \left[\frac{-P_{\theta_1^*}(Z) + P_{W_1^*}(Z)}{P_{\theta_1^*}(Z)} \right] && (6.118)\\
&\quad - \mathbf{E}_{Z \sim P_{\theta_1^*}} \left[\frac{-P_{\theta_0^*}(Z) + P_{W_0^*}(Z)}{P_{\theta_0^*}(Z)} \right] + \mathbf{E}_{Z \sim P_{W_1^*}} \left[\log \frac{P_{\theta_1^*}(Z)}{P_{\theta_0^*}(Z)} \right] && (6.119)\\
&= -\mathbf{E}_{Z \sim P_{\theta_1^*}} \left[\log \frac{P_{\theta_1^*}(Z)}{P_{\theta_0^*}(Z)} \right] + 1 - \mathbf{E}_{Z \sim P_{\theta_1^*}} \left[\frac{P_{W_0^*}(Z)}{P_{\theta_0^*}(Z)} \right] + \mathbf{E}_{Z \sim P_{W_1^*}} \left[\log \frac{P_{\theta_1^*}(Z)}{P_{\theta_0^*}(Z)} \right] && (6.120)\\
&\geq -\mathbf{E}_{Z \sim P_{\theta_1^*}} \left[\log \frac{P_{\theta_1^*}(Z)}{P_{\theta_0^*}(Z)} \right] + \mathbf{E}_{Z \sim P_{W_1^*}} \left[\log \frac{P_{\theta_1^*}(Z)}{P_{\theta_0^*}(Z)} \right]. && (6.121)
\end{aligned}$$

The last line is obtained by switching $P_{\theta_1^*}$ and $P_{W_0^*}$ and then recalling that $\frac{P_{\theta_1^*}}{P_{\theta_0^*}}$ is an S-value:

$$\mathbf{E}_{Z \sim P_{\theta_1^*}} \left[\frac{P_{W_0^*}(Z)}{P_{\theta_0^*}(Z)} \right] = \mathbf{E}_{Z \sim P_{W_0^*}} \left[\frac{P_{\theta_1^*}(Z)}{P_{\theta_0^*}(Z)} \right] \leq 1. \tag{6.122}$$

It remains to show that

$$\mathbf{E}_{Z \sim P_{W_1^*}} \left[\log \frac{P_{\theta_1^*}(Z)}{P_{\theta_0^*}(Z)} \right] \geq \mathbf{E}_{Z \sim P_{\theta_1^*}} \left[\log \frac{P_{\theta_1^*}(Z)}{P_{\theta_0^*}(Z)} \right], \tag{6.123}$$

but this follows directly from the assumption that $\dot{S}^*$ achieves its minimum growth rate at θ_1^* as follows

$$\mathbf{E}_{Z\sim P_{W_1^*}}\left[\log\frac{P_{\theta_1^*}(Z)}{P_{\theta_0^*}(Z)}\right] = \mathbf{E}_{\theta\sim W_1^*}\left[\mathbf{E}_{Z\sim P_\theta}\left[\log\frac{P_{\theta_1^*}(Z)}{P_{\theta_0^*}(Z)}\right]\right] \tag{6.124}$$

$$\geq \mathbf{E}_{\theta\sim W_1^*}\left[\mathbf{E}_{Z\sim P_{\theta_1^*}}\left[\log\frac{P_{\theta_1^*}(Z)}{P_{\theta_0^*}(Z)}\right]\right] \tag{6.125}$$

$$= \mathbf{E}_{Z\sim P_{\theta_1^*}}\left[\log\frac{P_{\theta_1^*}(Z)}{P_{\theta_0^*}(Z)}\right]. \tag{6.126}$$

□

While we have not been able to prove that the reverse of the theorem is true, namely that if $P_\theta(S^* = \dot{S}^*) = 1$ for all $\theta \in \Theta_0 \cup \Theta_1$ then $g(\cdot)$ achieves its minimum at θ_1^*, the following argument suggests that it is indeed true. Suppose, for a contradiction, that $g(\cdot)$ does not achieve its minimum at θ_1^*, namely there exists $\theta \in \Theta_1$ such that $g(\theta) < g(\theta_1^*)$. We then have

$$\mathbf{E}_{Z\sim P_\theta}[\log S^*] = \mathbf{E}_{Z\sim P_\theta}[\log \dot{S}^*] = g(\theta) < g(\theta_1^*) = \mathbf{E}_{Z\sim P_{\theta_1^*}}[\log \dot{S}^*] = \mathbf{E}_{Z\sim P_{\theta_1^*}}[\log S^*], \tag{6.127}$$

and thus

$$\mathbf{E}_{Z\sim P_\theta}[\log S^*] < \mathbf{E}_{Z\sim P_{\theta_1^*}}[\log S^*]. \tag{6.128}$$

Now Theorem 1 states that S^* achieves its minimum growth rate at W_1^* and so

$$\mathbf{E}_{Z\sim P_{W_1^*}}[\log S^*] \leq \mathbf{E}_{Z\sim P_\theta}[\log S^*] < \mathbf{E}_{Z\sim P_{\theta_1^*}}[\log S^*], \tag{6.129}$$

which implies that $W_1^* \neq \delta_{\theta_1^*}$. It thus seems unlikely that the priors W_0^* and $\delta_{\theta_0^*}$ could be such that the difference between W_1^* and $\delta_{\theta_1^*}$ is always cancelled out, for every $\theta \in \Theta_0 \cup \Theta_1$.

6.6 Finding the DOT S-value for 2×2 tables and checking whether it is GROW

The following lemma states that θ_1^* always lies on the boundary of $\Theta_1(\epsilon)'$. This will be useful in finding (θ_1^*, θ_0^*) in the different cases since it reduces the search to a one-dimensional space, as we will explain below.

Lemma 20. *For any parameter of interest with threshold ϵ, and any prior knowledge rectangle* PKR*, we have*

$$\theta_1^* \in \mathrm{BD}(\Theta_1(\epsilon)'), \quad \textit{where} \quad (\theta_1^*, \theta_0^*) \coloneqq \underset{(\theta_1,\theta_0)\in\Theta_1(\epsilon)'\times\Theta_0'}{\arg\min} \mathrm{KL}(P_{\theta_1}||P_{\theta_0}). \tag{6.130}$$

Proof. For any fixed $\theta_1 = (\theta_a, \theta_b) \in \Theta_1(\epsilon)'$, recall that its associated $\theta_0^* \coloneqq \arg\min_{\theta_0\in\Theta_0'} \mathrm{KL}(P_{\theta_1}||P_{\theta_0})$ lies to the lower right of θ_1 (this is the case even if there is a projection involved). Let $\tilde{\theta}_1 \in \Theta_1$ be any point inside the rectangle spanned by θ_1 and θ_0^*. In other words, let $\tilde{\theta}_1$ satisfy

$$p \leq \tilde{\theta}_a \leq \theta_a \quad \text{and} \quad p \leq \tilde{\theta}_b \leq \theta_b. \tag{6.131}$$

Then, by independence and the strict monotonicity of the KL divergence between two Bernoulli distributions, we have

$$\mathrm{KL}(P_{\theta_1}||P_{\theta_0^{(p)}}) = n_a\mathrm{KL}(\mathrm{Bern}(\theta_a)||\mathrm{Bern}(p)) + n_b\mathrm{KL}(\mathrm{Bern}(\theta_b)||\mathrm{Bern}(p)) \tag{6.132}$$

$$\geq n_a\mathrm{KL}(\mathrm{Bern}(\tilde{\theta}_a)||\mathrm{Bern}(p)) + n_b\mathrm{KL}(\mathrm{Bern}(\tilde{\theta}_b)||\mathrm{Bern}(p)) \tag{6.133}$$

$$= \mathrm{KL}(P_{\tilde{\theta}}||P_{(p,p)}). \tag{6.134}$$

Thus for $(\theta_1^*, \theta_0^*) \coloneqq \arg\min_{(\theta_1,\theta_0)\in\Theta_1(\epsilon)'\times\Theta_0'} \mathrm{KL}(P_{\theta_1}||P_{\theta_0})$, it must be the case that all of the points to the lower right of θ_1^* lie outside $\Theta_1(\epsilon)'$. For all the permutations of parameter of interest and prior knowledge that we have considered, this implies θ_1^* lies on the boundary of $\Theta_1(\epsilon)'$. □

For each combination of parameter of interest and prior knowledge, the above lemma suggests the following method for finding (θ_1^*, θ_0^*) as follows. First, parametrize the boundary of $\Theta_1(\epsilon)'$ by

$$\mathrm{BD}(\Theta_1(\epsilon)') = \{(q, \varphi(q)) : q \in [0,1]\} \cap \mathrm{PKR}, \quad \text{for some } \varphi. \tag{6.135}$$

For each q, define

$$f(q) \coloneqq \mathrm{KL}(P_{\theta_1(q)} || P_{\theta_0^*(q)}), \tag{6.136}$$

where $\theta_1(q) \coloneqq (q, \varphi(q))$ and $\theta_0^*(q) \coloneqq \arg\min_{\theta_0 \in \Theta_0'} \mathrm{KL}(P_{\theta_1(q)} || P_{\theta_0})$. Defining

$$q^* \coloneqq \underset{q:\theta_1(q) \in \Theta_1(\epsilon)'}{\arg\min} f(q), \tag{6.137}$$

we then have

$$(\theta_1^*, \theta_0^*) = (\theta_1(q^*), \theta_0(q^*)). \tag{6.138}$$

Thus finding (θ_1^*, θ_0^*) reduces to a one-dimensional optimization problem. We use this method in a number of cases below.

We now prove the following lemma, which implies that $\dot{S}^*$ achieves its minimum growth rate at some θ_G lying on the boundary of $\Theta_1(\epsilon)'$, provided no projection is required to obtain θ_0^*. This gets us part of the way to satisfying the requirement of Theorem 19 since, as we have just seen (Lemma 20), θ_1^* lies on the boundary of $\Theta_1(\epsilon)'$ regardless of the parameter of interest, its threshold value, or the prior knowledge. We will then consider each case separately to see whether θ_G in fact equals θ_1^* and so, by Theorem 19, $\dot{S}^* = S^*$.

Lemma 21. *Let $\theta = (\theta_a, \theta_b) \in [0,1]^2$ be such that $\theta_b > \theta_a$ and define the test statistic*

$$T_\theta(Z) = \frac{P_\theta(Z)}{P_{\theta_0^*}(Z)}, \quad \textit{where} \quad \theta_0^* \coloneqq \underset{\theta_0 \in \Theta_0'}{\arg\min}\, \mathrm{KL}(P_\theta || P_{\theta_0}). \tag{6.139}$$

For any $\theta' = (\theta_a', \theta_b') \in [0,1]^2$, let $g(\theta')$ be the growth rate of T_θ at θ'. Then

$$g(\theta') = c_a \theta_a' + c_b \theta_b' + c, \tag{6.140}$$

for constants c_a, c_b and c. Further, $c_a \leq 0$ and $c_b \geq 0$, meaning $g(\cdot)$ is decreasing in θ_a' and increasing in θ_b'. For all combinations of parameter of interest and prior knowledge, this implies that the growth rate of T_θ is minimized on the boundary of $\Theta_1(\epsilon)'$.

Proof. Let $\theta_0^* = (p^*, p^*)$. Using independence we have

$$
\begin{aligned}
g(\theta') &:= \mathbf{E}_{Z\sim P_{\theta'}}[\log T_\theta] && (6.141)\\
&= \mathbf{E}_{Z\sim P'_\theta}\left[\log \frac{P_\theta(Z)}{P_{\theta_0^*}(Z)}\right] && (6.142)\\
&= \mathbf{E}_{Z\sim P_{\theta'}}\left[\log \frac{\mathrm{Bin}(N_{a1}; n_a, \theta_a)\mathrm{Bin}(N_{b1}; n_b, \theta_b)}{\mathrm{Bin}(N_{a1}; n_a, p^*)\mathrm{Bin}(N_{b1}; n_b, p^*)}\right] && (6.143)\\
&= \mathbf{E}_{N_{a1}\sim\mathrm{Bin}(n_a,\theta'_a)}\left[\log \frac{\mathrm{Bin}(N_{a1}; n_a, \theta_a)}{\mathrm{Bin}(N_{a1}; n_a, p^*)}\right] + \mathbf{E}_{N_{b1}\sim\mathrm{Bin}(n_b,\theta'_b)}\left[\log \frac{\mathrm{Bin}(N_{b1}; n_b, \theta_b)}{\mathrm{Bin}(N_{b1}; n_b, p^*)}\right] && (6.144)\\
&= \sum_{x\in\{a,b\}} \mathbf{E}_{N_{x1}\sim\mathrm{Bin}(n_x,\theta'_x)}\left[\log \frac{\binom{n_x}{N_{x1}}\theta_x^{N_{x1}}(1-\theta_x)^{n_x-N_{x1}}}{\binom{n_x}{N_{x1}}(p^*)^{N_{x1}}(1-p^*)^{n_x-N_{x1}}}\right] && (6.145)\\
&= \sum_{x\in\{a,b\}} \mathbf{E}_{N_{x1}\sim\mathrm{Bin}(n_x,\theta'_x)}\left[\log \left(\frac{\theta_x}{p^*}\right)^{N_{x1}}\left(\frac{1-\theta_x}{1-p^*}\right)^{n_x-N_{x1}}\right] && (6.146)\\
&= \sum_{x\in\{a,b\}} \log\left(\frac{\theta_x}{p^*}\right)\mathbf{E}_{N_{x1}\sim\mathrm{Bin}(n_x,\theta'_x)}[N_{x1}] + \log\left(\frac{1-\theta_x}{1-p^*}\right)\mathbf{E}_{N_{x1}\sim\mathrm{Bin}(n_x,\theta'_x)}[n_x - N_{x1}] && (6.147)\\
&= \sum_{x\in\{a,b\}} n_x\theta'_x \log\left(\frac{\theta_x}{p^*}\right) + n_x(1-\theta'_x)\log\left(\frac{1-\theta_x}{1-p^*}\right) && (6.148)\\
&= \sum_{x\in\{a,b\}} n_x\left[\theta'_x \log\left(\frac{\theta_x}{1-\theta_x}\frac{1-p^*}{p^*}\right) + \log\left(\frac{1-\theta_x}{1-p^*}\right)\right]. && (6.149)\\
&= c_a\theta'_a + c_b\theta'_b + c, && (6.150)
\end{aligned}
$$

where

$$
c_x = n_x \log\left(\frac{\theta_x}{1-\theta_x}\frac{1-p^*}{p^*}\right) \quad \text{for } x \in \{a,b\} \quad \text{and} \quad c = \sum_{x\in\{a,b\}} n_x \log\left(\frac{1-\theta_x}{1-p^*}\right). \tag{6.151}
$$

Thus $g(\cdot)$ is linear in θ'_a and θ'_b and so to minimize it we simply need to find the sign of the coefficients c_a and c_b. Recall Corollary 14, which states that $\theta_0^* = (p^*, p^*)$ lies to the lower right of θ_1, namely

$$
\theta_a \le p^* \quad \text{and} \quad \theta_b \ge p^*. \tag{6.152}
$$

Since the function $x \mapsto \frac{x}{1-x}$ is increasing on $[0,1]$, we have

$$
\theta_a \le p^* \implies \frac{\theta_a}{1-\theta_a} \le \frac{p^*}{1-p^*} \implies \log\left(\frac{\theta_a}{1-\theta_a}\frac{1-p^*}{p^*}\right) \le 0. \tag{6.153}
$$

Likewise,

$$
\theta_b \ge p^* \implies \frac{\theta_b}{1-\theta_b} \ge \frac{p^*}{1-p^*} \implies \log\left(\frac{\theta_b}{1-\theta_b}\frac{1-p^*}{p^*}\right) \ge 0. \tag{6.154}
$$

This shows that $g(\cdot)$ is decreasing in θ'_a and increasing in θ'_b. For all combinations of parameter of interest and prior knowledge, this implies that the growth rate of T_θ is minimized on the boundary of $\Theta_1(\epsilon)'$. □

Finding θ_1^* and θ_0^* in the case of the risk difference

In the case of the risk difference, we are only able to find explicit formulas for (θ_1^*, θ_0^*) when $n_a = n_b$. For a fixed $\delta \in (0,1)$, we parametrize the boundary of $\Theta_1(\delta)$ as follows. Take $\varphi(q) = q + \delta$, so that (6.135) becomes

$$
\mathrm{BD}(\Theta_1(\delta)) = \{\theta_1(q) : q \in [0, 1-\delta]\} \cap \mathrm{PKR} \tag{6.155}
$$

where $\theta_1(q) \coloneqq (q, q+\delta)$. Suppose for the moment that we have no prior knowledge. Then there is no need for projection to find $\theta_0^*(q)$ for each $\theta_1(q)$. Therefore $\theta_0^*(q) \coloneqq \arg\min_{\theta_0 \in \Theta_0'} \mathrm{KL}(P_{\theta_1(q)} || P_{\theta_0})$ is given by $(p^*(q), p^*(q))$, where $p^*(q)$ is given by (6.6), namely

$$p^*(q) = \frac{n_a q + n_b(q+\delta)}{n} = q + r\delta, \tag{6.156}$$

where $r \coloneqq n_b/n$. As in (6.136), for each q we define

$$\begin{aligned} f(q) &\coloneqq \mathrm{KL}(P_{\theta_1(q)} || P_{\theta_0^*(q)}) && (6.157)\\ &= \mathrm{KL}(P_{(q,q+\delta)} || P_{(q+r\delta, q+r\delta)}) && (6.158)\\ &= n_a \mathrm{kl}(q || q+r\delta) + n_b \mathrm{kl}(q+\delta || q+r\delta), && (6.159) \end{aligned}$$

where the last line follows by independence.

The following lemma shows that $f(\cdot)$ is a strictly convex function of q (since linear combinations of strictly convex functions are themselves strictly convex) and so has a unique minimum occurring either at an endpoint of $[0, 1-\delta]$ or at a unique stationary point if one exists. This minimum can be found in closed form in the case when $n_a = n_b$, where we have $q^* = \frac{1-\delta}{2}$, so that

$$\theta_1^* = \left(\frac{1-\delta}{2}, \frac{1+\delta}{2}\right) \quad \text{and} \quad \theta_0^* = \left(\frac{1}{2}, \frac{1}{2}\right). \tag{6.160}$$

This is stated and proved below as Corollary 24. When $n_a \neq n_b$, the following lemma at least shows that $f(\cdot)$ is straightforward to minimize computationally. Since the proof of the lemma is somewhat tedious and uninsightful, we relegate it to Appendix B.

Lemma 22. *Let $\gamma \in [-1, 1] \setminus \{0\}$ and define $f(q) = \mathrm{kl}(q+\gamma || q)$ for $\max\{0, -\gamma\} \leq q \leq \min\{1, 1-\gamma\}$. Then f is strictly convex.*

Theorem 23. *Fix n_a, n_b (not necessarily equal) and risk difference threshold $\delta \in (0,1)$ and let $r \coloneqq n_b/n$. Then the equation*

$$n_a \log \frac{q(1-(q+r\delta))}{(q+r\delta)(1-q)} + n_b \log \frac{(q+\delta)(1-(q+r\delta))}{(q+r\delta)(1-(q+\delta))} = 0 \tag{6.161}$$

has a unique solution in $[0, 1-\delta]$, which we denote q^*. Define

$$(\theta_1^*, \theta_0^*) \coloneqq \underset{(\theta_1, \theta_0) \in \Theta_1(\delta)' \times \Theta_0'}{\arg\min} \mathrm{KL}(P_{\theta_1} || P_{\theta_0}). \tag{6.162}$$

Then $\theta_1^* = (q^*, q^*+\delta)$ and $\theta_0^* = (q^*+r\delta, q^*+r\delta)$, provided these lie in PKR*. Further, in this case the GROW S-value S^* is given by*

$$S^* = \dot{S}^* \coloneqq \frac{P_{\theta_1^*}}{P_{\theta_0^*}}. \tag{6.163}$$

Proof. By Lemma 20 we know that θ_1^* lies on the boundary of $\Theta_1(\delta)'$. Let us first suppose that we have no prior knowledge. Parametrize the boundary of $\Theta_1(\delta)$ as in (6.155), namely

$$\mathrm{BD}(\Theta_1(\delta)) = \{(q, q+\delta) : q \in [0, 1-\delta]\}. \tag{6.164}$$

As in (6.156), for each $\theta_1(q) \coloneqq (q, q+\delta)$, the corresponding $\theta_0^*(q) = (p^*(q), p^*(q))$ is given by $p^*(q) = q + r\delta$. Finally, for $q \in [0, 1-\delta]$, we define

$$f(q) \coloneqq \mathrm{KL}(P_{\theta_1(q)} || P_{\theta_0^*(q)}). \tag{6.165}$$

Expanding the KL divergence, we have

$$\begin{aligned} f(q) &= \mathrm{KL}(P_{(q,q+\delta)} || P_{(q+r\delta, q+r\delta)}) && (6.166)\\ &= n_a \mathrm{kl}(q || q+r\delta) + n_b \mathrm{kl}(q+\delta || q+r\delta) && (6.167)\\ &= n_a \left[q \log \frac{q}{q+r\delta} + (1-q) \log \frac{1-q}{1-(q+r\delta)} \right] && (6.168)\\ &\quad + n_b \left[(q+\delta) \log \frac{q+\delta}{q+r\delta} + (1-(q+\delta)) \log \frac{1-(q+\delta)}{1-(q+r\delta)} \right]. && (6.169) \end{aligned}$$

Differentiating with respect to q, we obtain

$$f'(q) = n_a \left[\log \frac{q}{q+r\delta} + 1 - \frac{q}{q+r\delta} - \log \frac{1-q}{1-(q+r\delta)} - 1 + \frac{1-q}{1-(q+r\delta)} \right] \tag{6.170}$$

$$+ n_b \left[\log \frac{q+\delta}{q+r\delta} + 1 - \frac{q+\delta}{q+r\delta} - \log \frac{1-(q+\delta)}{1-(q+r\delta)} - 1 + \frac{1-(q+\delta)}{1-(q+r\delta)} \right] \tag{6.171}$$

$$= n_a \left[\log \frac{q(1-(q+r\delta))}{(q+r\delta)(1-q)} + \frac{-q(1-(q+r\delta)) + (1-q)(q+r\delta)}{(q+r\delta)(1-(q+r\delta))} \right] \tag{6.172}$$

$$+ n_b \left[\log \frac{(q+\delta)(1-(q+r\delta))}{(q+r\delta)(1-(q+\delta))} + \frac{-(q+\delta)(1-(q+r\delta)) + (1-(q+\delta))(q+r\delta)}{(q+r\delta)(1-(q+r\delta))} \right] \tag{6.173}$$

$$= n_a \left[\log \frac{q(1-(q+r\delta))}{(q+r\delta)(1-q)} + \frac{r\delta}{(q+r\delta)(1-(q+r\delta))} \right] \tag{6.174}$$

$$+ n_b \left[\log \frac{(q+\delta)(1-(q+r\delta))}{(q+r\delta)(1-(q+\delta))} - \frac{(1-r)\delta}{(q+r\delta)(1-(q+r\delta))} \right] \tag{6.175}$$

$$= n_a \log \frac{q(1-(q+r\delta))}{(q+r\delta)(1-q)} + n_b \log \frac{(q+\delta)(1-(q+r\delta))}{(q+r\delta)(1-(q+\delta))} + \frac{(n_a r - n_b(1-r))\delta}{(q+r\delta)(1-(q+r\delta))} \tag{6.176}$$

$$= n_a \log \frac{q(1-(q+r\delta))}{(q+r\delta)(1-q)} + n_b \log \frac{(q+\delta)(1-(q+r\delta))}{(q+r\delta)(1-(q+\delta))}, \tag{6.177}$$

where the last line follows since

$$n_a r - n_b(1-r) = \frac{n_a n_b}{n} - n_b + \frac{n_b^2}{n} = \frac{n_b}{n}(n_a - n + n_b) = 0. \tag{6.178}$$

Since $\delta \in (0,1)$ and $r < 1$, using continuity we have

$$\lim_{q \downarrow 0} f'(q) = -\infty \quad \text{and} \quad \lim_{q \uparrow 1-\delta} f'(q) = \infty. \tag{6.179}$$

Therefore, since Lemma 22 implies $f(\cdot)$ is convex, we know that there exists a unique $q^* \in [0, 1-\delta]$ for which $f'(q^*) = 0$. This proves the first part of the theorem, namely that (6.161) has a unique solution in $[0, 1-\delta]$.

If a convex function has a stationary point, then this is in fact the global minimum. Thus $q^* = \arg\min_{q \in [0,1-\delta]} f(q)$ and so

$$\mathrm{KL}(P_{\theta_1(q^*)} || P_{\theta_0(q^*)}) = \min_{(\theta_1,\theta_0) \in \Theta_1(\delta) \times \Theta_0} \mathrm{KL}(P_{\theta_1} || P_{\theta_0}). \tag{6.180}$$

Since restricting the parameter sets cannot decrease the minimum, we have

$$\min_{(\theta_1,\theta_0) \in \Theta_1(\delta) \times \Theta_0} \mathrm{KL}(P_{\theta_1} || P_{\theta_0}) \leq \min_{(\theta_1,\theta_0) \in \Theta_1(\delta)' \times \Theta_0'} \mathrm{KL}(P_{\theta_1} || P_{\theta_0}). \tag{6.181}$$

Now, by assumption $\theta_1(q^*), \theta_0(q^*) \in \mathrm{PKR}$ and so

$$\min_{(\theta_1,\theta_0) \in \Theta_1(\delta)' \times \Theta_0'} \mathrm{KL}(P_{\theta_1} || P_{\theta_0}) \leq \mathrm{KL}(P_{\theta_1(q^*)} || P_{\theta_0(q^*)}). \tag{6.182}$$

These three lines together imply that (6.181) is in fact an equality, which proves the second part of the theorem, namely $\theta_1^* = \theta_1(q^*) = (q^*, q^* + \delta)$ and $\theta_0^* = \theta_0(q^*) = (q^* + r\delta, q^* + r\delta)$.

We now show that $\dot{S}^*$ achieves its minimum growth rate at θ_1^*. By Theorem 19, this will complete the proof. By Lemma 21, we know that the minimum growth rate of $\dot{S}^*$ is attained at the boundary

of $\Theta_1(\delta)'$. Let $g(q)$ equal the growth rate of $\dot{S}^*$ at $\theta_1(q) = (q, q+\delta)$, for $q \in [0, 1-\delta]$. We have

$$
\begin{aligned}
g(q) &\coloneqq \mathbf{E}_{Z\sim P_{\theta_1(q)}}\left[\log \frac{P_{\theta_1^*}}{P_{\theta_0^*}}\right] && (6.183)\\
&= \mathbf{E}_{Z\sim P_{\theta_1(q)}}\left[\log \frac{P_{\theta_1(q)}}{P_{\theta_0^*}} - \log \frac{P_{\theta_1(q)}}{P_{\theta_1^*}}\right] && (6.184)\\
&= \mathrm{KL}(P_{\theta_1(q)}||P_{\theta_0^*}) - \mathrm{KL}(P_{\theta_1(q)}||P_{\theta_1^*}) && (6.185)\\
&= \mathrm{KL}(P_{(q,q+\delta)}||P_{(p^*,p^*)}) - \mathrm{KL}(P_{(q,q+\delta)}||P_{(q^*,q^*+\delta)}) && (6.186)\\
&= n_a \mathrm{kl}(q||p^*) + n_b \mathrm{kl}(q+\delta||p^*) - n_a \mathrm{kl}(q||q^*) - n_b \mathrm{kl}(q+\delta||q^*+\delta) && (6.187)\\
&= n_a \left[q \log \frac{q}{p^*} + (1-q)\log \frac{1-q}{1-p^*} - q\log\frac{q}{q^*} - (1-q)\log\frac{1-q}{1-q^*}\right] && (6.188)\\
&+ n_b \Bigg[(q+\delta)\log\frac{q+\delta}{p^*} + (1-(q+\delta))\log\frac{1-(q+\delta)}{1-p^*} && (6.189)\\
&\qquad - (q+\delta)\log\frac{q+\delta}{q^*+\delta} - (1-(q+\delta))\log\frac{1-(q+\delta)}{1-(q^*+\delta)}\Bigg] && (6.190)\\
&= n_a \left[q\log\frac{q^*}{p^*} + (1-q)\log\frac{1-q^*}{1-p^*}\right] && (6.191)\\
&\qquad + n_b \left[(q+\delta)\log\frac{q^*+\delta}{p^*} + (1-(q+\delta))\log\frac{1-(q^*+\delta)}{1-p^*}\right]. && (6.192)
\end{aligned}
$$

Differentiating with respect to q, we obtain

$$
\begin{aligned}
g'(q) &= n_a \left[\log\frac{q^*}{p^*} - \log\frac{1-q^*}{1-p^*}\right] + n_b \left[\log\frac{q^*+\delta}{p^*} - \log\frac{1-(q^*+\delta)}{1-p^*}\right] && (6.193)\\
&= n_a \log\frac{q^*(1-p^*)}{p^*(1-q^*)} + n_b \log\frac{(q^*+\delta)(1-p^*)}{p^*(1-(q^*+\delta))}, && (6.194)
\end{aligned}
$$

which is independent of q. If we now substitute the expression for p^*, namely $p^* = q^* + r\delta$, we have

$$
g'(q) = n_a \log\frac{q^*(1-(q^*+r\delta))}{(q^*+r\delta)(1-q^*)} + n_b \log\frac{(q^*+\delta)(1-(q^*+r\delta))}{(q^*+r\delta)(1-(q^*+\delta))}. \tag{6.195}
$$

This is equal to zero by the definition of q^*. Therefore $g(\cdot)$ is constant and so trivially attains its minimum at q^*, regardless of the prior knowledge. □

In the case of $n_a = n_b$, (6.161) has a closed form solution, namely $q^* = (1-\delta)/2$. We state this as a corollary.

Corollary 24. *Fix $n_a = n_b$ and risk difference threshold $\delta \in (0,1)$. Then the GROW S-value S^* is given by*

$$
S^*(Z) = \frac{P_{\theta_1^*}(Z)}{P_{\theta_0^*}(Z)}, \quad \textit{where} \quad \theta_1^* = \left(\frac{1-\delta}{2}, \frac{1+\delta}{2}\right) \quad \textit{and} \quad \theta_0^* = \left(\frac{1}{2}, \frac{1}{2}\right), \tag{6.196}
$$

provided θ_1^* and θ_0^* lie in PKR*. Further, the worst case growth of this S-value is*

$$
GR(S^*) = n\mathrm{kl}\left(\frac{1+\delta}{2}\middle|\middle|\frac{1}{2}\right). \tag{6.197}
$$

Proof. By Theorem 23, it suffices to show that $q = (1-\delta)/2$ solves (6.161) in the case of $r \coloneqq n_b/n = 1/2$. Substituting $n_a = n_b$ and $r = 1/2$, the equation becomes

$$
\frac{n}{2}\log\frac{q(q+\delta)(1-(q+\delta/2))^2}{(1-q)(1-(q+\delta))(q+\delta/2)^2} = 0 \tag{6.198}
$$

If we now substitute $q = (1-\delta)/2$ into the left hand side, we obtain

$$
\frac{n}{2}\log\frac{\frac{1-\delta}{2}(\frac{1-\delta}{2}+\delta)(1-(\frac{1-\delta}{2}+\delta/2))^2}{(1-\frac{1-\delta}{2})(1-(\frac{1-\delta}{2}+\delta))(\frac{1-\delta}{2}+\delta/2)^2} = \frac{n}{2}\log\frac{\frac{1-\delta}{2}\frac{1+\delta}{2}(\frac{1}{2})^2}{\frac{1+\delta}{2}\frac{1-\delta}{2}(\frac{1}{2})^2} = 0. \tag{6.199}
$$

The growth of S^* is minimized at θ_1^*. Thus

$$\begin{aligned}
\mathrm{GR}(S^*) &= \mathbf{E}_{Z\sim P_{\theta_1^*}}\left[\log\frac{P_{\theta_1^*}}{P_{\theta_0^*}}\right] && (6.200)\\
&= \mathrm{KL}(P_{\theta_1^*}||P_{\theta_0^*}) && (6.201)\\
&= n_a\mathrm{kl}\left(\frac{1-\delta}{2}\middle\|\frac{1}{2}\right) + n_b\mathrm{kl}\left(\frac{1+\delta}{2}\middle\|\frac{1}{2}\right) && (6.202)\\
&= \frac{n}{2}\mathrm{kl}\left(1-\frac{1+\delta}{2}\middle\|\frac{1}{2}\right) + \frac{n}{2}\mathrm{kl}\left(\frac{1+\delta}{2}\middle\|\frac{1}{2}\right) && (6.203)\\
&= n\mathrm{kl}\left(\frac{1+\delta}{2}\middle\|\frac{1}{2}\right). && (6.204)
\end{aligned}$$

The final line follows by the identity $\mathrm{kl}(p||q) = \mathrm{kl}(1-p||q)$, which is easily seen by writing out both sides. □

The proof of Corollary 24 relies on Theorem 23 which in turn depends on the fiddly Lemma 22. However, there exists a more direct proof without reference to this lemma, which we now present as an alternative.

Proof. (*Alternative proof of Corollary 24*) Recalling (6.177) from Theorem 23 (the derivation of which is independent of Lemma 22), we have

$$f'(q) = \frac{n}{2}\log\frac{q(q+\delta)(1-(q+\delta/2))^2}{(1-q)(1-(q+\delta))(q+\delta/2)^2}. \tag{6.205}$$

Thus

$$\begin{aligned}
f'(q) > 0 &\iff \frac{q(q+\delta)(1-(q+\delta/2))^2}{(1-q)(1-(q+\delta))(q+\delta/2)^2} > 1 && (6.206)\\
&\iff q(q+\delta)(1-(q+\delta/2))^2 - (1-q)(1-(q+\delta))(q+\delta/2)^2 > 0 && (6.207)\\
&\iff \frac{\delta^3}{4} + \frac{\delta^2}{2}q - \frac{\delta^2}{4} > 0 && (6.208)\\
&\iff \delta + 2q - 1 > 0 && (6.209)\\
&\iff q > \frac{1-\delta}{2}. && (6.210)
\end{aligned}$$

Clearly, by symmetry, we also have

$$f'(q) < 0 \iff q > \frac{1-\delta}{2} \quad \text{and} \quad f'(q) = 0 \iff q = \frac{1-\delta}{2}. \tag{6.211}$$

Thus $f(\cdot)$ is strictly increasing away from $\frac{1-\delta}{2}$ and so has a unique minimum at $q^* \coloneqq \frac{1-\delta}{2}$. □

Finding θ_1^* and θ_0^* in the case of the relative risk

In the case of the relative risk, $\Theta_1(\lambda)$ and Θ_0 get arbitrarily close near $(0,0)$, and so S^* is degenerate unless we have prior knowledge that positively separates $\Theta_1(\lambda)$ and Θ_0. When dealing with the relative risk we therefore always assume that either $\theta_a^L > 0$ or $\theta_b^L > 0$.

For a fixed $\lambda > 1$, we parametrize the boundary of $\Theta_1(\lambda)$ as follows. Take $\varphi(q) = \lambda q$, so that (6.135) becomes

$$\mathrm{BD}(\Theta_1(\lambda)) = \{\theta_1(q) : q \in [0, 1/\lambda]\} \cap \mathrm{PKR}, \tag{6.212}$$

where $\theta_1(q) \coloneqq (q, \lambda q)$. If, for a given $\theta_1(q)$, no projection is required, then

$$\theta_0^*(q) \coloneqq \operatorname*{arg\,min}_{\theta_0\in\Theta_0'} \mathrm{KL}(P_{\theta_1(q)}||P_{\theta_0}) \tag{6.213}$$

is given by $(p^*(q), p^*(q))$, where $p^*(q)$ is given by (6.6), namely

$$p^*(q) = \frac{n_a q + n_b \lambda q}{n} = rq. \tag{6.214}$$

where $r := \frac{n_a + \lambda n_b}{n}$. As in (6.136), for each q we define

$$f(q) := \mathrm{KL}(P_{\theta_1(q)} || P_{\theta_0^*(q)}). \tag{6.215}$$

If, for a given q, no projection is required, then this can be written as

$$\begin{aligned} f(q) &:= \mathrm{KL}(P_{\theta_1(q)} || P_{\theta_0^*(q)}) && (6.216)\\ &= \mathrm{KL}(P_{(q,\lambda q)} || P_{(rq,rq)}) && (6.217)\\ &= n_a \mathrm{kl}(q||rq) + n_b \mathrm{kl}(\lambda q||rq), && (6.218) \end{aligned}$$

where the last line follows by independence.

The following lemma shows that $f(\cdot)$ is strictly increasing in q, since both of its terms are. This means that if, for $q_{\min} := \min\{q : (q, \lambda q) \in \Theta_1(\lambda)'\}$, no projection is required, we have $q^* = q_{\min}$ and $p^* = p^*(q^*) = rq^*$, so that

$$\theta_1^* = (q_{\min}, \lambda q_{\min}) \quad \text{and} \quad \theta_0^* = (rq_{\min}, rq_{\min}). \tag{6.219}$$

We restate this as Theorem 26, where we give a condition for no projection being required for $q_{\min}$ and prove the result more formally. First we state the following necessary lemma. Its proof is uninsightful and so can be found in Appendix B.

Lemma 25. *Let $\gamma > 0$ and define $f(q) := \mathrm{kl}(q||\gamma q)$. Then f is strictly increasing in q.*

Theorem 26. *Fix arbitrary n_a and n_b (not necessarily equal) and relative risk threshold $\lambda > 1$ and let $r := (n_a + \lambda n_b)/n$. Suppose either $\theta_a^L > 0$ or $\theta_b^L > 0$, so that $\Theta_1(\lambda)'$ and Θ_0' are positively separated, and define*

$$(\theta_1^*, \theta_0^*) := \operatorname*{arg\,min}_{(\theta_1,\theta_0) \in \Theta_1(\delta)' \times \Theta_0'} \mathrm{KL}(P_{\theta_1} || P_{\theta_0}). \tag{6.220}$$

Assume $\theta_b^U \geq \lambda \theta_a^L$ (otherwise $\Theta_1(\lambda)'$ is empty). If $\theta_b^L \leq r\theta_a^L \leq \theta_a^U$, then the GROW S-value S^* is given by

$$S^*(Z) = \dot{S}^*(Z) := \frac{P_{\theta_1^*}(Z)}{P_{\theta_0^*}(Z)} \quad \textit{where} \quad \theta_1^* = (\theta_a^L, \lambda \theta_a^L) \quad \textit{and} \quad \theta_0^* = (r\theta_a^L, r\theta_a^L). \tag{6.221}$$

Proof. As before, for q such that $\theta_1(q) \in \Theta_1(\lambda)'$, define

$$\theta_0^\times(q) := \operatorname*{arg\,min}_{\theta_0 \in \Theta_0} \mathrm{KL}(P_{\theta_1(q)} || P_{\theta_0}) \quad \text{and} \quad \theta_0^*(q) := \operatorname*{arg\,min}_{\theta_0 \in \Theta_0'} \mathrm{KL}(P_{\theta_1(q)} || P_{\theta_0}), \tag{6.222}$$

and recall that $\theta_0^\times = (p^\times(q), p^\times(q))$ is given by (6.6), namely

$$p^\times(q) = \frac{n_a q + n_b \lambda q}{n} = rq. \tag{6.223}$$

Further, define

$$f(q) := \mathrm{KL}(P_{\theta_1(q)} || P_{\theta_0^*(q)}) \quad \text{and} \quad f^\times(q) := \mathrm{KL}(P_{\theta_1(q)} || P_{\theta_0^\times(q)}). \tag{6.224}$$

By Lemma 25, $f^\times(\cdot)$ is strictly increasing in q and so attains its minimum at $q_{\min} := \min\{q : \theta_1(q) \in \Theta_1(\lambda)'\}$. Note that

$$(q, \lambda q) \in \mathrm{PKR} := [\theta_a^L, \theta_a^U] \times [\theta_b^L, \theta_b^U] \iff (q, q) \in [\theta_a^L, \theta_a^U] \times [\theta_b^L/\lambda, \theta_b^U/\lambda] \tag{6.225}$$

and so $q_{\min} = \max\{\theta_a^L, \theta_b^L/\lambda\}$. Moreover, since $r := (n_a + \lambda n_b)/n < (\lambda n_a + \lambda n_b)/n = \lambda$, the assumption that $\theta_b^L \leq r\theta_a^L$ implies $\theta_b^L < \lambda \theta_a^L$ and so $q_{\min} = \theta_a^L$. We now show that no projection is required for $(\theta_a^L, \lambda \theta_a^L)$, namely $p^\times(\theta_a^L) \in I_{\mathrm{PKR}} := [\max\{\theta_a^L, \theta_b^L\}, \min\{\theta_a^U, \theta_b^U\}]$. First, note

$$p^\times(\theta_a^L) = \frac{n_a \theta_a^L + n_b \lambda \theta_a^L}{n} = r\theta_a^L. \tag{6.226}$$

By assumption, $\theta_b^L \le r\theta_a^L \le \theta_a^U$. Now $r \coloneqq (n_a + \lambda n_b)/n > (n_a + n_b)/n = 1$ implies $r\theta_a^L \ge \theta_a^L$. Further, using again that $r < \lambda$, and the assumption that $\theta_b^U \ge \lambda\theta_a^L$, we have $r\theta_a^L < \lambda\theta_a^L \le \theta_b^U$. In summary,

$$\theta_a^L, \theta_b^L \le r\theta_a^L \quad \text{and} \quad r\theta_a^L \le \theta_a^U, \theta_b^U \tag{6.227}$$

and so $r\theta_a^L \in I_{\text{PKR}}$. This means no projection is required, meaning $p^*(\theta_a^L) = p^\times(\theta_a^L)$ and so $\theta_0^*(\theta_a^L) = \theta_0^\times(\theta_a^L)$. By the definitions of $f(\cdot)$ and $f^\times(\cdot)$, this implies $f(\theta_a^L) = f^\times(\theta_a^L)$.

Since $\Theta_0' \subseteq \Theta_0$, we know that for any q such that $(q, \lambda q) \in \Theta_1(\lambda)'$, we have $f(q) \ge f^\times(q)$ (this again follows from the definitions of $f(\cdot)$ and $f^\times(\cdot)$). Thus

$$f(q) \ge f^\times(q) \ge f^\times(\theta_a^L) = f(\theta_a^L) \tag{6.228}$$

and so $f(\cdot)$ also attains its minimum at θ_a^L. Taken together, this implies

$$\theta_1^* = (\theta_a^L, \lambda\theta_a^L) \quad \text{and} \quad \theta_0^* = (r\theta_a^L, r\theta_a^L). \tag{6.229}$$

To complete the proof, it suffices (by Theorem 19) to show that $\dot{S}^*$ achieves its minimum growth rate at $\theta_1^* = (\theta_a^L, \lambda\theta_a^L)$. Let $g(q)$ be the growth rate of $\dot{S}^*$ at $\theta_1(q)$. Recall (6.149), which states that for any $\theta \in \Theta_1(\epsilon)'$, its associated θ_0^* and an arbitrary $\theta' \in \Theta_1(\epsilon)'$, the growth rate of $T_\theta \coloneqq P_\theta / P_{\theta_0^*}$ at θ' is given by

$$g(\theta') \coloneqq \mathbf{E}_{Z\sim P_{\theta'}}\left[\log \frac{P_{\theta_1}}{P_{\theta_0^*}}\right] = \sum_{x\in\{a,b\}} n_x \left[\theta_x' \log\left(\frac{\theta_x}{1-\theta_x}\frac{1-p^*}{p^*}\right) + \log\left(\frac{1-\theta_x}{1-p^*}\right)\right]. \tag{6.230}$$

Thus, substituting $\theta = \theta_1^* = (\theta_a^L, \lambda\theta_a^L)$, $\theta_0^* = (r\theta_a^L, r\theta_a^L)$ and $\theta' = \theta_1(q) = (q, \lambda q)$, we have

$$\begin{aligned}
g(q) &\coloneqq \mathbf{E}_{Z\sim P_{\theta_1(q)}}\left[\frac{P_{\theta_1^*}(Z)}{P_{\theta_0^*}(Z)}\right] && (6.231)\\
&= \sum_{x\in\{a,b\}} n_x \left[\theta_1(q)_x \log\left(\frac{\theta_{1,x}^*}{1-\theta_{1,x}^*}\frac{1-p^*}{p^*}\right) + \log\left(\frac{1-\theta_{1,x}^*}{1-p^*}\right)\right] && (6.232)\\
&= n_a\left[q\log\left(\frac{\theta_a^L}{1-\theta_a^L}\frac{1-r\theta_a^L}{r\theta_a^L}\right) + k_a\right] + n_b\left[\lambda q\log\left(\frac{\lambda\theta_a^L}{1-\lambda\theta_a^L}\frac{1-r\theta_a^L}{r\theta_a^L}\right) + k_b\right] && (6.233)\\
&= n_a q\log\left(\frac{\theta_a^L}{1-\theta_a^L}\right) + \lambda n_b q\log\left(\frac{\lambda\theta_a^L}{1-\lambda\theta_a^L}\right) + (n_a+\lambda n_b)q\log\left(\frac{1-r\theta_a^L}{r\theta_a^L}\right) + k && (6.234)\\
&= q\left[n_a\log\left(\frac{\theta_a^L}{1-\theta_a^L}\right) + \lambda n_b\log\left(\frac{\lambda\theta_a^L}{1-\lambda\theta_a^L}\right) + (n_a+\lambda n_b)\log\left(\frac{n-(n_a+\lambda n_b)\theta_a^L}{(n_a+\lambda n_b)\theta_a^L}\right)\right] + k && (6.235)\\
&= q\Bigg[-n_a\log\big(1-\theta_a^L\big) + \lambda n_b\log\lambda - \lambda n_b\log\big(1-\lambda\theta_a^L\big) && (6.236)\\
&\qquad + (n_a+\lambda n_b)\log\left(\frac{n-(n_a+\lambda n_b)\theta_a^L}{n_a+\lambda n_b}\right)\Bigg] + k, && (6.237)
\end{aligned}$$

where k_a, k_b and k are constants that are independent of q and in (6.235) we substituted the value of r. Thus, as expected, $g(q)$ is linear in q. Let $c(\theta_a^L, \lambda)$ denote the coefficient of q in the final line. We now show that for all values of n_a, n_b, θ_a^L and λ, we have $c(\theta_a^L, \lambda) > 0$. This completes the proof since it shows that $g(\cdot)$ is strictly increasing and therefore obtains its minimum value at the minimum value of q, namely θ_a^L.

First, for fixed n_a, n_b, we have that

$$f(\lambda) \coloneqq c(0,\lambda) = \lambda n_b\log\lambda + (n_a+\lambda n_b)\log\left(\frac{n}{n_a+\lambda n_b}\right) \tag{6.238}$$

is positive for all $\lambda > 1$. This is because $f(1) = 0$ and $f'(\lambda) > 0$ for all $\lambda > 1$ as follows

$$f'(\lambda) = n_b + n_b \log \lambda + n_b \log \left(\frac{n}{n_a + \lambda n_b} \right) - n_b \tag{6.239}$$

$$= n_b \log \left(\frac{\lambda n}{n_a + \lambda n_b} \right) \tag{6.240}$$

$$= n_b \log \left(\frac{\lambda n_a + \lambda n_b}{n_a + \lambda n_b} \right) > 0. \tag{6.241}$$

Further, for any $\lambda > 1$, the partial derivative of $c(\theta_a^L, \lambda)$ with respect to θ_a^L is positive. This follows by direct calculation

$$\frac{\partial c}{\partial \theta_a^L} = \frac{n_a}{1 - \theta_a^L} + \frac{\lambda^2 n_b}{1 - \lambda \theta_a^L} - \frac{(n_a + \lambda n_b)^2}{n - (n_a + \lambda n_b)\theta_a^L} \tag{6.242}$$

$$= \frac{n_a(1 - \lambda\theta_a^L) + \lambda^2 n_b(1 - \theta_a^L)}{(1 - \theta_a^L)(1 - \lambda\theta_a^L)} - \frac{(n_a + \lambda n_b)^2}{n - (n_a + \lambda n_b)\theta_a^L} \tag{6.243}$$

$$= \frac{n_a + \lambda^2 n_b - \lambda(n_a + \lambda n_b)\theta_a^L}{(1 - \theta_a^L)(1 - \lambda\theta_a^L)} - \frac{(n_a + \lambda n_b)^2}{n - (n_a + \lambda n_b)\theta_a^L} \tag{6.244}$$

$$= \frac{\left[n_a + \lambda^2 n_b - \lambda(n_a + \lambda n_b)\theta_a^L\right]\left[n - (n_a + \lambda n_b)\theta_a^L\right] - (n_a + \lambda n_b)^2(1 - \theta_a^L)(1 - \lambda\theta_a^L)}{(1 - \theta_a^L)(1 - \lambda\theta_a^L)\left[n - (n_a + \lambda n_b)\theta_a^L\right]} \tag{6.245}$$

$$= \frac{(n_a + \lambda n_b)\left[- \lambda n - (n_a + \lambda^2 n_b) + (\lambda + 1)(n_a + \lambda n_b)\right]\theta_a^L + n(n_a + \lambda^2 n_b) - (n_a + \lambda n_b)^2}{(1 - \theta_a^L)(1 - \lambda\theta_a^L)\left[n - (n_a + \lambda n_b)\theta_a^L\right]} \tag{6.246}$$

$$= \frac{n(n_a + \lambda^2 n_b) - (n_a + \lambda n_b)^2}{(1 - \theta_a^L)(1 - \lambda\theta_a^L)\left[n - (n_a + \lambda n_b)\theta_a^L\right]} \tag{6.247}$$

$$= \frac{n_a n_b + \lambda^2 n_a n_b - 2\lambda n_a n_b}{(1 - \theta_a^L)(1 - \lambda\theta_a^L)\left[n - (n_a + \lambda n_b)\theta_a^L\right]} \tag{6.248}$$

$$= \frac{n_a n_b (1 - \lambda)^2}{(1 - \theta_a^L)(1 - \lambda\theta_a^L)\left[n - (n_a + \lambda n_b)\theta_a^L\right]}. \tag{6.249}$$

Line (6.247) follows by using $n = n_a + n_b$ and expanding the coefficient of the θ_a^L in the numerator to see that the coefficient equals zero. Now, since we assumed $\theta_a^L \in (0, 1/\lambda)$ where $\lambda > 1$, we know that $1 - \theta_a^L > 0$ and $1 - \lambda\theta_a^L > 0$. Further, since $\lambda\theta_a^L < 1$, we have $n - (n_a + \lambda n_b)\theta_a^L > n - \lambda n \theta_a^L = n(1 - \lambda\theta_a^L) > 0$. Thus the final expression is positive and the proof is complete. □

6.7 The DOT S-value is not always GROW

At this point we might conjecture that the optimal priors are always point masses, provided Θ_0' and $\Theta_1(\epsilon)'$ are convex. However, this turns out to be false, as the following example shows. The example is rather simple. As we will see later in Theorem 26, it comes from violating the condition that $\theta_b^L \leq r\theta_a^L \leq \theta_a^U$ by taking $\theta_b^L > 0$ and $\theta_a^L = 0$. We then show that the $\dot{S}^*$ does not achieve its minimum growth rate at θ_1^*, which, by Theorem 19, shows that $\dot{S}^* \neq S^*$.

Lemma 27. *Suppose we have* $\text{PKR} = [0, 1] \times [\theta_b^L, 1]$ *and that our parameter of interest is the relative risk. For any values of* n_a *and* n_b*, there exists a threshold value* λ *such that* $\dot{S}^* \neq S^*$*. Likewise, for any values of* n_a *and* λ*, there exists a value of* n_b *such that* $\dot{S}^* \neq S^*$*.*

Proof. We show that in the two cases either λ or n_b respectively can be chosen such that $\dot{S}^*$ does not achieve its minimum growth rate at θ_1^*. For the moment, let n_a, n_b and λ be fixed, with choices to be made at the end. As before, we know that θ_1^* lies on the boundary of $\Theta_1(\lambda)'$. Parameterize the boundary of $\Theta_1(\lambda)'$ by

$$\Theta_1(\lambda)' = \{\theta_1(q) : q \in [\theta_b^L/\lambda, 1/\lambda]\}, \tag{6.250}$$

where $\theta_1(q) \coloneqq (q, \lambda q)$ and for each q let

$$\theta_0^*(q) \coloneqq \arg\min_{\theta_0 \in \Theta_0'} \mathrm{KL}(P_{\theta_1(q)} || P_{\theta_0}). \tag{6.251}$$

If we define

$$f(q) \coloneqq \mathrm{KL}(P_{\theta_1(q)} || P_{\theta_0^*(q)}), \tag{6.252}$$

then

$$(\theta_1^*, \theta_0^*) = (\theta_1(q^*), \theta_0^*(q^*)), \quad \text{where} \quad q^* \coloneqq \arg\min_{q \in [\theta_b^L/\lambda, 1/\lambda]} f(q). \tag{6.253}$$

Due to the shape of PKR, there exists a value of q, say q_0, such that if $q > q_0$ no projection is required while for $q \leq q_0$ we have $\theta_0^*(q) = (\theta_b^L, \theta_b^L)$. Thus for $q > q_0$ we have $f^{\times}(q) = f(q)$. We have seen in Lemma 25 that $f^{\times}(\cdot)$ is strictly increasing in q. This implies that $q^* \leq q_0$ and so $\theta_0^* = (\theta_b^L, \theta_b^L)$. Therefore it suffices to find the minimum of $f(\cdot)$ on the interval $[\theta_b^L/\lambda, q_0]$, where

$$f(q) = \mathrm{KL}(P_{\theta_1(q)} || P_{(\theta_b^L, \theta_b^L)}). \tag{6.254}$$

We now show that $f'(\theta_b^L/\lambda) < 0$, so that $q^* > \theta_b^L/\lambda$. For $q \leq q_0$, we have

$$\begin{aligned} f(q) =& \mathrm{KL}(P_{(q,\lambda q)} || P_{(\theta_b^L, \theta_b^L)}) && (6.255)\\ =& \, n_a \mathrm{kl}(q || \theta_b^L) + n_b \mathrm{kl}(\lambda q || \theta_b^L) && (6.256)\\ =& \, n_a \left[q \log \frac{q}{\theta_b^L} + (1-q) \log \frac{1-q}{1-\theta_b^L} \right] + n_b \left[\lambda q \log \frac{\lambda q}{\theta_b^L} + (1-\lambda q) \log \frac{1-\lambda q}{1-\theta_b^L} \right]. && (6.257) \end{aligned}$$

Differentiating,

$$\begin{aligned} f'(q) =& \, n_a \left[\log \frac{q}{\theta_b^L} + 1 - \log \frac{1-q}{1-\theta_b^L} - 1 \right] + n_b \left[\lambda \log \frac{\lambda q}{\theta_b^L} + \lambda - \lambda \log \frac{1-\lambda q}{1-\theta_b^L} - \lambda \right] && (6.258)\\ =& \, n_a \log \frac{q(1-\theta_b^L)}{\theta_b^L (1-q)} + \lambda n_b \log \frac{\lambda q (1-\theta_b^L)}{\theta_b^L (1-\lambda q)}. && (6.259) \end{aligned}$$

If we now set $q = \theta_b^L/\lambda$, we obtain

$$\begin{aligned} f'\left(\frac{\theta_b^L}{\lambda}\right) =& \, n_a \log \frac{1-\theta_b^L}{\lambda - \theta_b^L} + \lambda n_b \log \frac{\theta_b^L (1-\theta_b^L)}{\theta_b^L (1-\theta_b^L)} && (6.260)\\ =& \, n_a \log \frac{1-\theta_b^L}{\lambda - \theta_b^L}. && (6.261) \end{aligned}$$

Since $\lambda > 1$, this shows that $f'(\theta_b^L/\lambda) < 0$ and so $q^* > \theta_b^L/\lambda$.

Recall Lemma 21, which states that the growth rate of $\dot{S}^*$ is minimized somewhere on the boundary of $\Theta_1(\lambda)'$. Let $g(q)$ be the growth rate of $\dot{S}^*$ at $\theta_1(q)$. Recalling (6.149), we have

$$\begin{aligned} g(q) \coloneqq& \, \mathbf{E}_{Z \sim P_{\theta_1(q)}} \left[\log \frac{P_{\theta_1^*}}{P_{\theta_0^*}} \right] && (6.262)\\ =& \, \mathbf{E}_{Z \sim P_{(q,\lambda q)}} \left[\log \frac{P_{(q^*, \lambda q^*)}}{P_{(\theta_b^L, \theta_b^L)}} \right] && (6.263)\\ =& \, n_a q \log \frac{q^*}{1-q^*} \frac{1-\theta_b^L}{\theta_b^L} + n_b \lambda q \log \frac{\lambda q^*}{1-\lambda q^*} \frac{1-\theta_b^L}{\theta_b^L} + c. && (6.264) \end{aligned}$$

where c is a constant that does not depend on q. Differentiating,

$$g'(q) = \left[n_a \log \frac{q^*}{1-q^*} \frac{1-\theta_b^L}{\theta_b^L} + \lambda n_b \log \frac{\lambda q^*}{1-\lambda q^*} \frac{1-\theta_b^L}{\theta_b^L} \right], \tag{6.265}$$

which is a constant. Now $q^* > \theta_b^L/\lambda$ implies $\lambda q^* > \theta_b^L$. Recalling $x \mapsto x/(1-x)$ is a strictly increasing positive function on $[0,1)$, we have

$$q^* > \theta_b^L/\lambda \implies \lambda q^* > \theta_b^L \tag{6.266}$$

$$\implies \frac{\lambda q^*}{1-\lambda q^*} > \frac{1-\theta_b^L}{\theta_b^L} \tag{6.267}$$

$$\implies \frac{\lambda q^*}{1-\lambda q^*}\frac{1-\theta_b^L}{\theta_b^L} > 1 \tag{6.268}$$

$$\implies \log \frac{\lambda q^*}{1-\lambda q^*}\frac{1-\theta_b^L}{\theta_b^L} > 0. \tag{6.269}$$

Further, this final term is increasing in λ. Thus, regardless of the value of

$$n_a \log \frac{q^*}{1-q^*}\frac{1-\theta_b^L}{\theta_b^L}, \tag{6.270}$$

$g'(q)$ is a positive constant if n_b or λ is chosen large enough. This would then imply that the growth rate of $\dot{S}^*$ is minimized by taking q as small as possible, namely at the point $\theta_G = (\theta_b^L/\lambda, \theta_b^L/\lambda)$. Since we have just seen that $q^* > \theta_b^L/\lambda$, we see that $\theta_G \neq \theta_1^*$. Thus, by Theorem 19, $\dot{S}^* \neq S^*$. □

Chapter 7

Results

The plots in this chapter show the growth rate and power of a number of S-values in the unconditional setting (namely where we take expectations over N_1 rather than assuming its value is fixed). We focus on the more computationally feasible S-values, namely the conditional S-values (constructed in the unconditional setting) of section 5.4 and the DOT S-values of chapter six. Since using a threshold value $\underline{\psi}$ of the odds ratio to restrict the alternative parameter set Θ_1 leads to the non-convex $\Theta_1(\underline{\psi})$, the results of the previous chapter do not apply. We therefore consider only the risk difference and relative risk in this chapter.

7.1 Process for generating the S-values

First, recall the process for generating the unconditional DOT S-values:

1. Choose a parameter ϵ (in this chapter δ or λ).
2. Specify the prior knowledge rectangle PKR.
3. For each threshold value $\underline{\epsilon}$, construct the null and alternative parameter sets Θ_0' and $\Theta_1(\underline{\epsilon})'$.
4. For each threshold $\underline{\epsilon}$, calculate the DOT S-value $\dot{S}^*_{\Theta_1(\underline{\epsilon})'}$ using the methods of the previous chapter.

For each $\underline{\epsilon}$, we then calculate the worst case growth rate and power of $\dot{S}^*_{\Theta_1(\underline{\epsilon})'}$ over $\Theta_1(\underline{\epsilon})'$. Finally, we see whether there exists a threshold value ϵ^* such that the power of $\dot{S}^*_{\Theta_1(\epsilon^*)'}$ exceeds that of any other $\dot{S}^*_{\Theta_1(\underline{\epsilon})'}$ *uniformly* over $\Theta_1(\epsilon_0)'$, where for the risk difference $\delta_0 = 0$ and for the relative risk $\lambda_0 = 1$. If such an S-value exists, it will be referred to as the *uniformly most powerful DOT S-value* or simply the UMP DOT S-value.

Second, recall the process for generating the conditional S-values in the unconditional setting. The first three steps are the same as in the unconditional case. Thereafter:

1. Construct the induced parameter sets

$$\Psi_0 \coloneqq \{\psi(\theta_a, \theta_b) : (\theta_a, \theta_b) \in \Theta_0'\} = \{1\} \quad \text{and} \tag{7.1}$$

$$\Psi_1(\underline{\psi}) \coloneqq \{\psi(\theta_a, \theta_b) : (\theta_a, \theta_b) \in \Theta_1(\underline{\epsilon})'\} = [\underline{\psi}, \psi_{\max}]. \tag{7.2}$$

2. For each induced threshold $\underline{\psi}$, the GROW S-value is given by

$$S^*_{\underline{\psi}}(N_{b1}) \coloneqq S^*_{\Psi_1(\underline{\psi})}(N_{b1}) = P_{\underline{\psi},N_1}(N_{b1})/P_{1,N_1}(N_{b1}), \tag{7.3}$$

   where P_{ψ,n_1} denotes the fnchypg(n, n_b, n_1) distribution.

3. The UMPG S-value is given by

$$S^*_{\psi^*(N_1)}(N_{b1}) = P_{\psi(N_1),N_1}(N_{b1})/P_{1,N_1}(N_{b1}), \tag{7.4}$$

   where for each n_1, $\psi^*(n_1)$ is the unique solution to $\mathrm{KL}(P_{\psi,n_1}||P_{1,n_1}) = -\log\alpha$.[1]

[1] For small values of n_1 this equation may not have a solution. In these cases we used an arbitrary value of ψ.

As before, for each threshold value $\underline{\epsilon}$, the worst case growth and power of these S-values is calculated over $\Theta_1(\underline{\epsilon})'$.

In summary, for each threshold $\underline{\epsilon}$ with induced threshold $\underline{\psi}$, we calculate the worst case growth and power over $\Theta_1(\underline{\epsilon})'$ of

1. The DOT S-value $\dot{S}^*_{\Theta_1(\underline{\epsilon})'}$
2. The UMP DOT S-value $\dot{S}^*_{\Theta_1(\epsilon^*)'}$
3. The conditional GROW S-value $S^*_{\underline{\psi}}$
4. The UMPG S-value S^*.

For the risk difference, we repeated this process with prior knowledge of each of the following forms:

1. no prior knowledge, namely PKR $= [0,1]^2$
2. prior knowledge $\theta_a \geq 0.1$, namely PKR $= [0.1,1] \times [0,1]$
3. prior knowledge $\theta_a \leq 0.9$, namely PKR $= [0,0.9] \times [0,1]$
4. prior knowledge $\theta_b \geq 0.1$, namely PKR $= [0,1] \times [0.1,1]$
5. prior knowledge $\theta_b \leq 0.9$, namely PKR $= [0,1] \times [0,0.9]$
6. prior knowledge $\theta_a = 0.3$, namely PKR $= [0.3,0.3] \times [0,1]$
7. prior knowledge $\theta_b = 0.3$, namely PKR $= [0,1] \times [0.3,0.3]$.

For the relative risk some of these choices do not suffice to strictly separate the parameter sets. In this case we just use 2, 4, 6 and 7

7.2 Growth and power plots

We now present all the results in figures 7.1 to 7.11, which were conducted using $n_a = n_b = 50$ throughout. Note that in some cases there is no line for the UMP DOT S-value. This is because in those cases no such UMP S-value exists. In the caption for each figure, we state whether it is known that $\dot{S}^* = S^*$ by theoretical results from chapter six. As before, for all the power plots we have included Fisher's exact test as a benchmark.

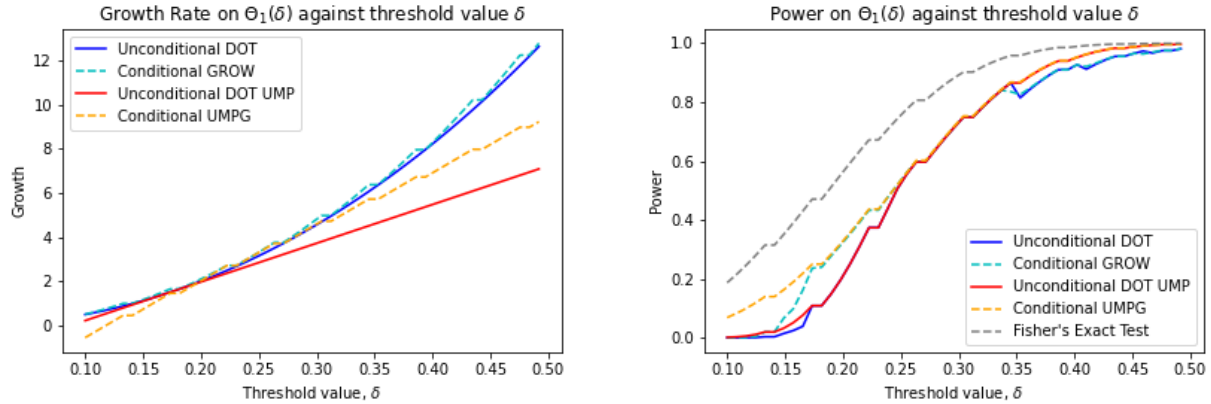

Figure 7.1: Risk difference with no prior knowledge. In this case, for every threshold value considered, it is known theoretically that $\dot{S}^* = S^*$.

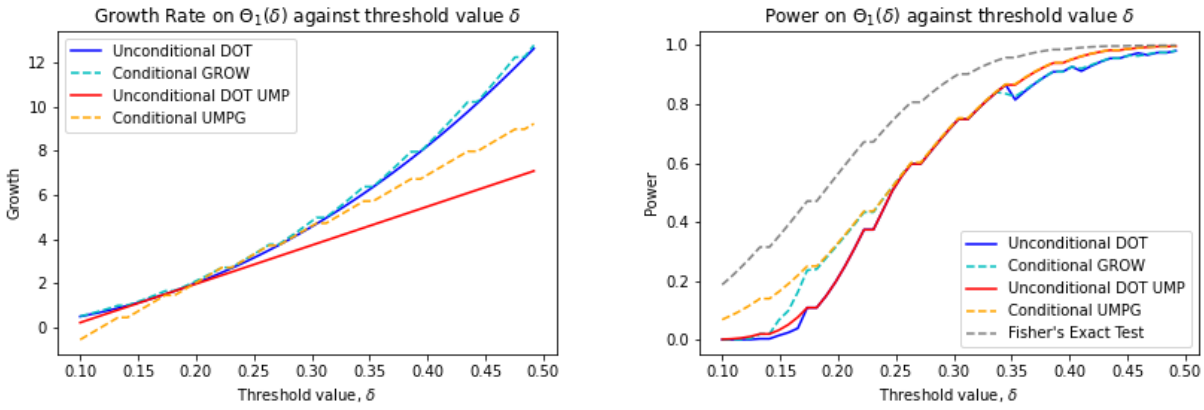

Figure 7.2: Risk difference with prior knowledge $\theta_a \geq 0.1$. In this case, for every threshold value considered, it is known theoretically that $\dot{S}^* = S^*$.

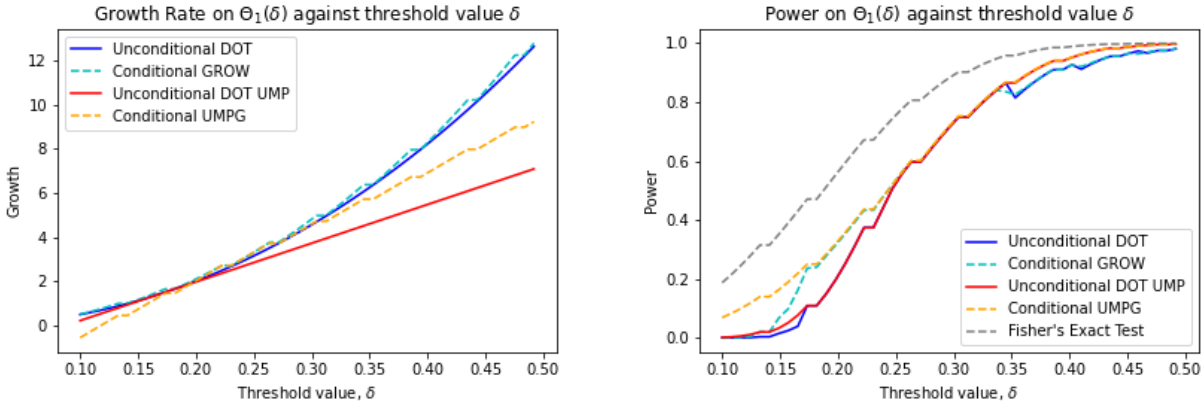

Figure 7.3: Risk difference with prior knowledge $\theta_a \leq 0.9$. In this case, for every threshold value considered, it is known theoretically that $\dot{S}^* = S^*$.

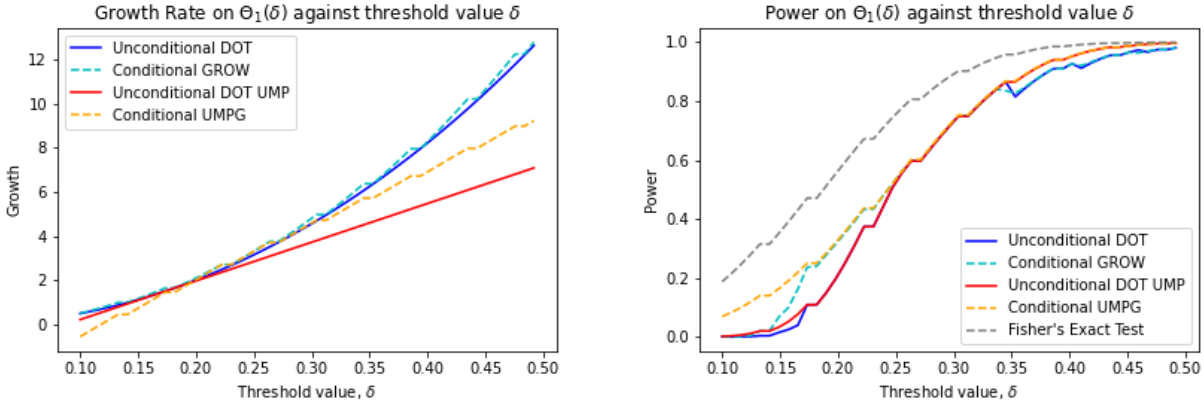

Figure 7.4: Risk difference with prior knowledge $\theta_b \geq 0.1$. In this case, for every threshold value considered, it is known theoretically that $\dot{S}^* = S^*$.

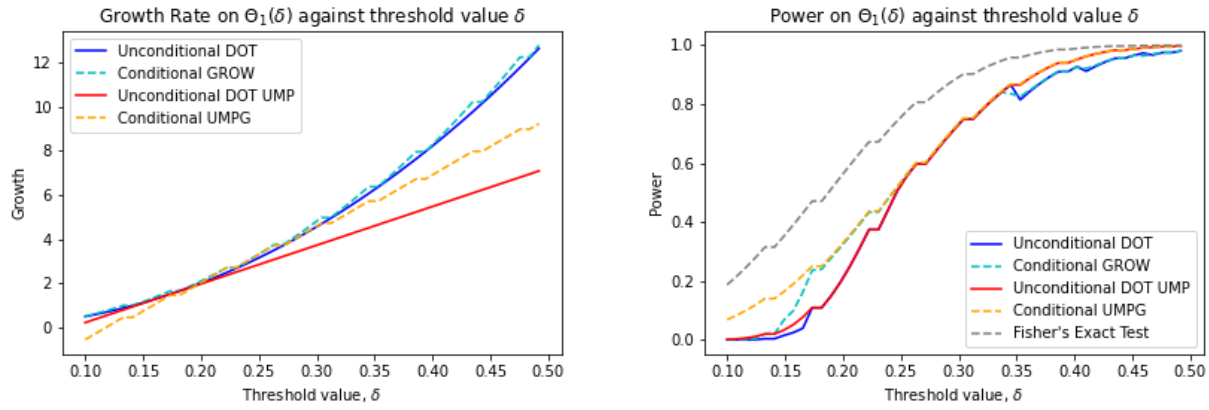

Figure 7.5: Risk difference with prior knowledge $\theta_b \leq 0.9$. In this case, for every threshold value considered, it is known theoretically that $\dot{S}^* = S^*$.

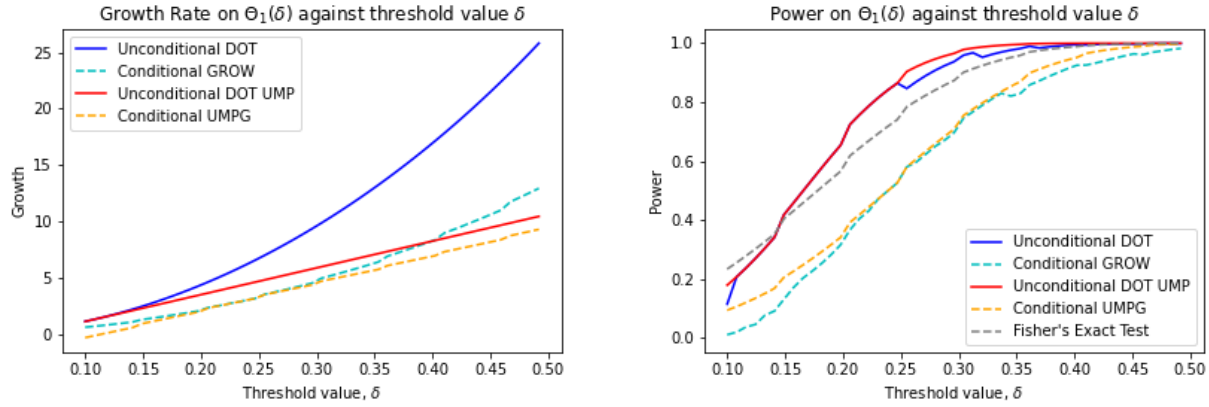

Figure 7.6: Risk difference with prior knowledge $\theta_a = 0.3$. In this case, for every threshold value considered, it is known theoretically that $\dot{S}^* = S^*$.

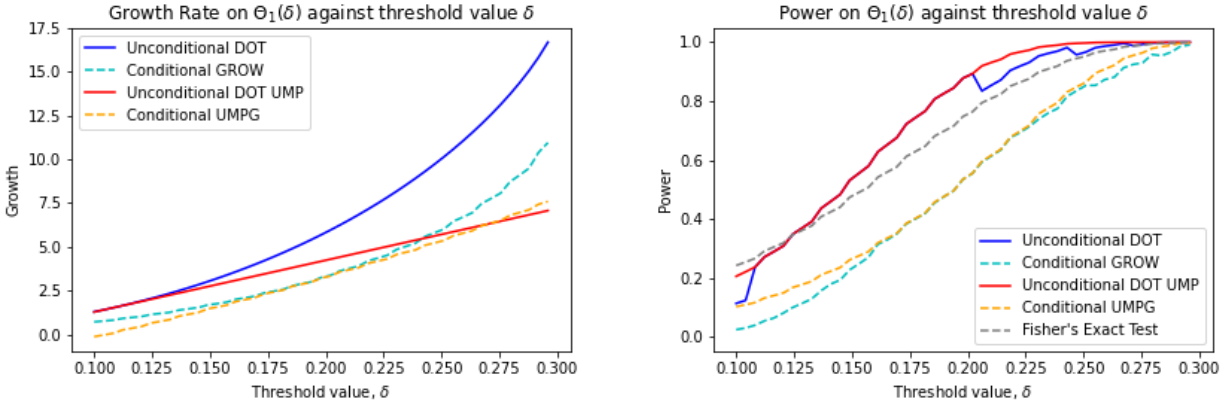

Figure 7.7: Risk difference with prior knowledge $\theta_b = 0.3$. In this case, for every threshold value considered, it is known theoretically that $\dot{S}^* = S^*$.

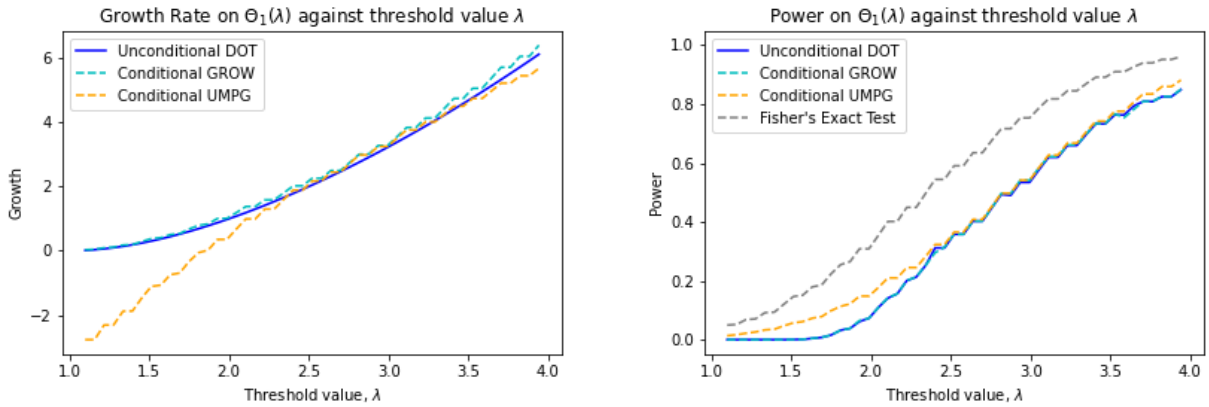

Figure 7.8: Relative risk with prior knowledge $\theta_a \geq 0.1$. In this case, for every threshold value considered, it is known theoretically that $\dot{S}^* = S^*$.

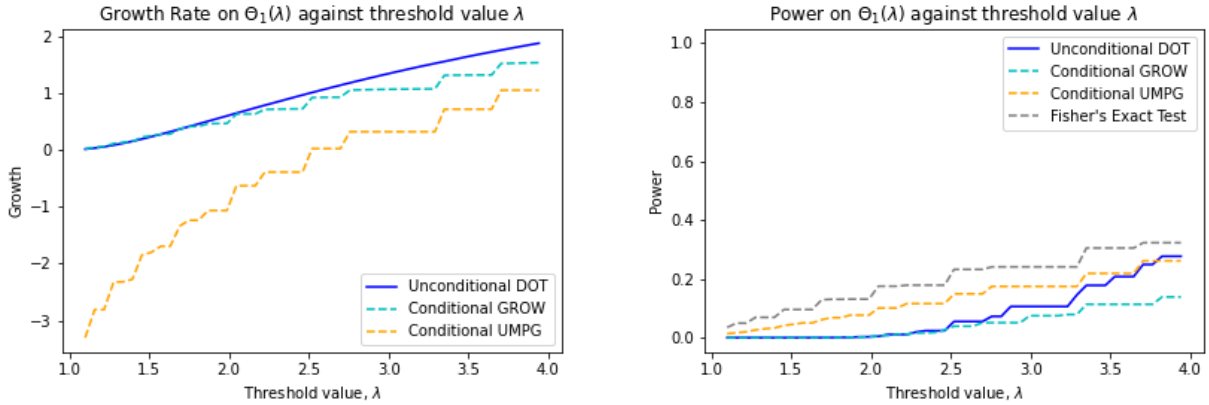

Figure 7.9: Relative risk with prior knowledge $\theta_b \geq 0.1$. In this case, for at least one threshold value considered it has been numerically demonstrated that $\dot{S}^* \neq S^*$, but this is not the case for all threshold values considered.

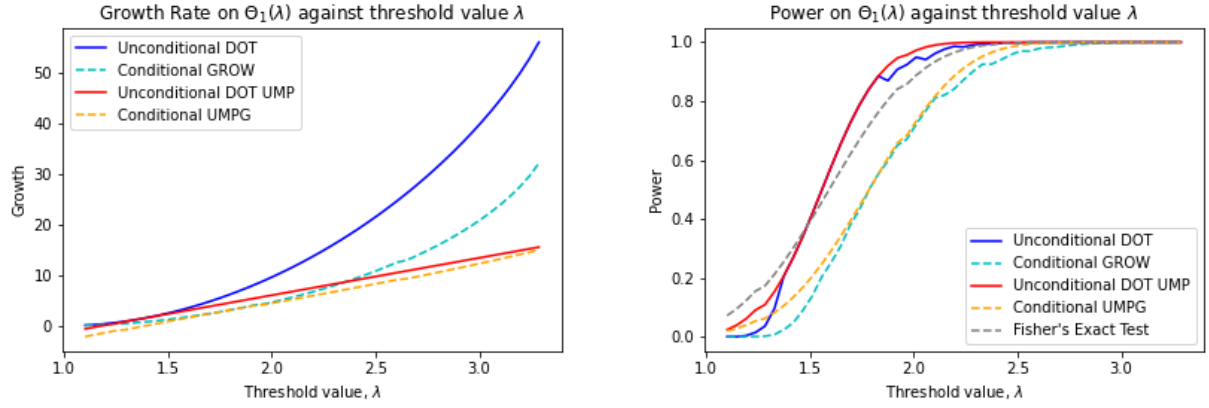

Figure 7.10: Relative risk with prior knowledge $\theta_a = 0.3$. In this case, for every threshold value considered, it is known theoretically that $\dot{S}^* = S^*$.

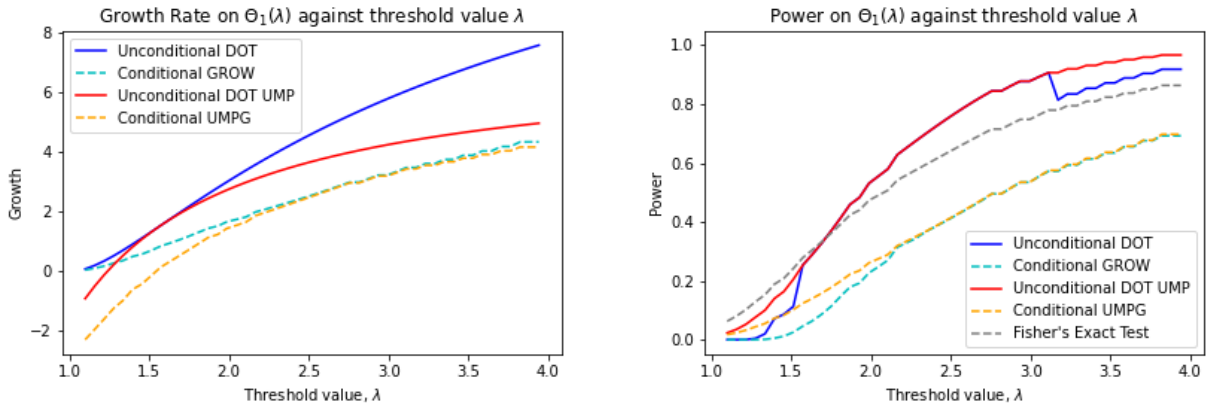

Figure 7.11: Relative risk with prior knowledge $\theta_b = 0.3$. In this case, for every threshold value considered, it is known theoretically that $\dot{S}^* = S^*$.

7.3 Analysis

In most cases we see that Fisher's exact test is more powerful than all of our S-values. However, for degenerate prior knowledge rectangles, namely when one of θ_a or θ_b is known precisely, it appears that the unconditional DOT and unconditional UMP DOT S-values are more powerful, at least for large threshold values. As before, the fact that Fisher's exact test is usually more powerful is perhaps to be expected, since S-values have to satisfy the additional requirement of being safe under optional continuation.

While there is no clear 'winner' for growth rate and power simultaneously among the four S-values, some qualitative judgements can be made. For example, it is noteworthy that between the unconditional DOT and the unconditional UMP DOT S-values, the first is likely to be preferable in many cases. The reason is that while by definition the DOT S-value has growth at least as good and power at most as good as the UMP DOT S-value, it frequently exceeds the growth rate of the UMP DOT S-value by a large margin, while falling short of the UMP DOT S-value in terms of power by only a small margin. The same appears true in the conditional case, namely that the conditional GROW S-value is slightly preferable to the conditional UMPG S-value in many cases.

A significant disadvantage of the conditional UMPG S-value is that its growth rate is occasionally *negative*, for small values of the threshold. This means that there are some cases in which the UMPG S-value may actually *shrink* under the alternative hypothesis. While the unconditional DOT UMP also suffers this disadvantage in a small region of figure 7.11, it is the UMPG S-value that is most affected. Overall, it appears that there should be a slight preference for the unconditional DOT S-value, but this is not clear cut in every case, meaning the practitioner should first check the specifics of their trial before deciding which S-value to use.

7.4 Similarity between growth of unconditional DOT and conditional GROW S-values

Curiously, the growth of the unconditional DOT and conditional GROW S-values is very similar if not identical in a number of cases. We explore here why that might be.

First, as in [5], for any $\rho > 0$, we can define

$$\Theta_1(\rho) := \left\{ \theta_1 \in \Theta_1 : \inf_{\theta_0 \in \Theta_0} \mathrm{KL}(P_{\theta_1} || P_{\theta_0}) \geq \rho \right\}. \tag{7.5}$$

The boundary $\mathrm{BD}(\Theta_1(\rho))$ of $\Theta_1(\rho)$, is then referred to in [5] as the 'lemon' due to its bulging shape around the main diagonal. Intuitively, to find the (θ_1^*, θ_0^*) for a given Θ_0' and $\Theta_1(\underline{\epsilon})'$ minimizing $\mathrm{KL}(P_{\theta_1} || P_{\theta_0}$, one can imagine increasing the value of ρ until the 'lemon' intersects first intersects $\Theta_1(\underline{\epsilon})'$. The point of intersection is then θ_1^*. Now $\Theta_1(\underline{\epsilon})'$ is a polygon for the risk difference and relative

risk. Further, the shape of the curves $\psi(\theta_a, \theta_b) = \psi_0$ are quite similar to the lemon and they share the same symmetry about the line $\theta_b = 1 - \theta_a$. It is not hard to imagine therefore that increasing ψ_0 to find the point where $\psi(\theta_a, \theta_b) = \psi_0$ first intersects $\Theta_1(\underline{\epsilon})'$ might also give θ_1^*. In the case of the risk difference with no prior knowledge, the symmetry alone suffices to see this. Letting $\theta_1^* = (\theta_{1,a}^*, \theta_{1,b}^*)$, this implies that the induced threshold for the odds ratio is given by $\underline{\psi} = \psi(\theta_{1,a}^*, \theta_{1,b}^*)$. Suppose we are indeed in this scenario and let $\theta_0^* = (p^*, p^*)$. Then the conditional GROW S-value is given by

$$S_{\underline{\psi}}^*(N_{b1}|N_1) = S_{\psi(\theta_{1,a}^*, \theta_{1,b}^*)}^*(N_{b1}|N_1) \tag{7.6}$$

$$:= \frac{P_{\psi(\theta_{1,a}^*, \theta_{1,b}^*), N_1}(N_{b1})}{P_{1,N_1}(N_{b1})} \tag{7.7}$$

$$= \frac{P_{(\theta_{1,a}^*, \theta_{1,b}^*)}(N_{b1}|N_1)}{P_{(p^*,p^*)}(N_{b1}|N_1)} \quad (\text{since } \psi(p^*, p^*) = 1) \tag{7.8}$$

$$= \frac{P_{(\theta_{1,a}^*, \theta_{1,b}^*)}(N_{a1}, N_{b1}) P_{(\theta_{1,a}^*, \theta_{1,b}^*)}(N_1)}{P_{(p^*,p^*)}(N_{a1}, N_{b1}) P_{(p^*,p^*)}(N_1)}. \tag{7.9}$$

Further, the unconditional DOT S-value is given by

$$\dot{S}^*(Z) = \frac{P_{\theta_1^*}(Z)}{P_{\theta_0^*}(Z)} = \frac{P_{(\theta_{1,a}^*, \theta_{1,b}^*)}(N_{a1}, N_{b1})}{P_{(p^*,p^*)}(N_{a1}, N_{b1})}. \tag{7.10}$$

Therefore, taking the ratio, we have

$$\frac{\dot{S}^*(Z)}{S_{\underline{\psi}}^*(Z)} = \frac{P_{(\theta_{1,a}^*, \theta_{1,b}^*)}(N_1)}{P_{(p^*,p^*)}(N_1)} = \frac{P_{\theta_1^*}(N_1)}{P_{\theta_0^*}(N_1)}. \tag{7.11}$$

Using this fact, we can inspect the difference in growth rate between the two S-values as follows. For any $\theta_1 = (\theta_a, \theta_b) \in \Theta_1(\underline{\epsilon})'$, we have

$$\mathbf{E}_{Z \sim P_{\theta_1}}[\log \dot{S}^*] - \mathbf{E}_{Z \sim P_{\theta_1}}[\log S_{\underline{\psi}}^*] = \mathbf{E}_{Z \sim P_{\theta_1}}\left[\log \frac{\dot{S}^*}{S_{\underline{\psi}}^*}\right] \tag{7.12}$$

$$= \mathbf{E}_{Z \sim P_{\theta_1}}\left[\log \frac{P_{\theta_1^*}(N_1)}{P_{\theta_0^*}(N_1)}\right]. \tag{7.13}$$

Now if $\dot{S}^*$ is in fact the GROW S-value then it achieves its minimum growth rate at θ_1^*. Further, for any N_1, $S_{\underline{\psi}}^*$ achieves its minimum conditional growth rate at $\psi^* = \psi(\theta_1^*)$. Suppose this implies $S_{\underline{\psi}}^*$ achieves its minimum *unconditional* growth rate at θ_1^*.

We now take a particular example, namely that of the risk difference with threshold $\delta = 0.2$ and $n_a = n_b = 40$. Recall that we then have

$$\theta_1^* = \left(\frac{1-\delta}{2}, \frac{1+\delta}{2}\right) = (0.4, 0.6) \quad \text{and} \quad \theta_0^* = (0.5, 0.5). \tag{7.14}$$

Plotting $\log\left(\dot{S}^*/S_{\underline{\psi}}^*\right)$ against n_1, we have the figure 7.12. Note that the difference is approximately zero in a neighbourhood of $n_1 = 40$, which is the expected value of N_1. Since this is where the majority of the mass of the distribution of N_1 lies, we see that 7.13 is indeed small, meaning the difference in growth between the two S-values is small (provided all the assumptions we made along the way are satisfied).

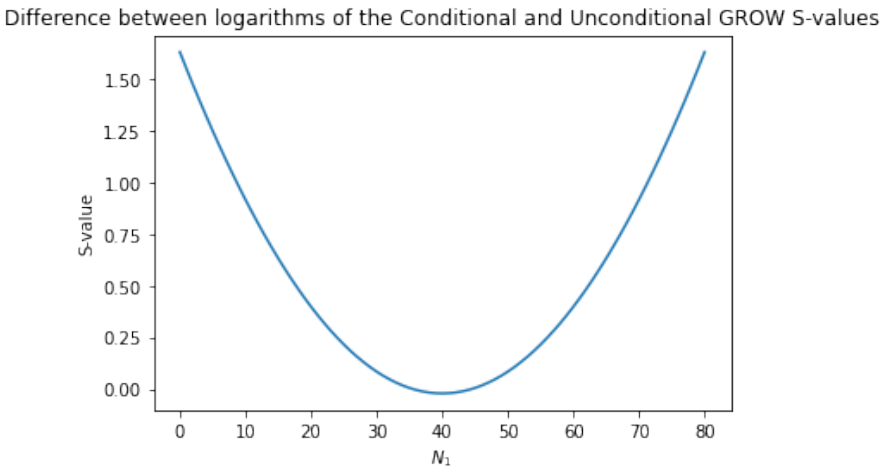

Figure 7.12: The logarithm of the ratio $\dot{S}^*(n_1)/S^*_{\underline{\psi}}(n_1)$ plotted against n_1.

Chapter 8

Conclusion

The bulk of the contribution of this thesis is the series of results either giving closed expressions for the S-values we considered, or providing calculation methods that are dramatically faster than approximating the JIP directly. While the evaluation of our safe tests did not provide a definite answer to the question of which should be used for statistical inference, the methods used to construct them and the results developed along the way contribute to the general understanding of safe tests for 2×2 tables. Some of our results and methods may find wider application in the study of safe tests more generally; perhaps most promisingly Theorem 19, which can be used to check whether it suffices to look at point mass priors. If future research validates the conjecture that this theorem can be reversed, it could also be used in other cases to falsify the claim that point mass priors suffice, showing that the JIP approximation may have to be used.

It was seen in chapter six that while the unconditional GROW S-value for 2×2 tables is frequently given by point priors, this is not always the case, even when the parameter sets are convex. Nevertheless, the cases in which point priors did not suffice were also the cases in which projection was required to find the optimal θ_0^*. This suggests the conjecture that, provided the parameter sets are convex *and* no projection is required to obtain θ_0^*, namely $\theta_0^* = (p^*, p^*)$ where

$$p^* = p^\times = \frac{n_a \theta_a + n_b \theta_b}{n}, \tag{8.1}$$

then point priors suffice. This would be consistent with the results found in this thesis and may be a direction for future work.

Appendix A

Conditions of Theorem 1

Recall the conditions of Theorem 1, which is the main result of the paper [5]:

1. For all $\theta_0 \in \Theta_0$ and $W_1 \in \mathcal{W}(\Theta_1')$ we have that P_{θ_0} is absolutely continuous relative to P_{W_1}.
2. The infimum $\inf_{(W_1,W_0)\in\mathcal{W}(\Theta_1')\times\mathcal{W}(\Theta_0)} \mathrm{KL}(P_{W_1}||P_{W_0})$ is finite.
3. The infimum is achieved by some (W_1^*, W_0^*).
4. The infimum is achieved uniquely.

We will now explore when these conditions hold for our different choices of $\Theta_1(\underline{\epsilon})'$ and Θ_0'.

The first condition holds if Θ_1' does not overlap with the boundary of the unit square. To see this, note that for all $\theta = (\theta_a, \theta_b) \in (0,1)^2$, $N_{a1} \in [n_a]$ and $N_{b1} \in [n_b]$, we have $P_\theta(N_{a1}, N_{b1}) \neq 0$. Thus for any $W_1 \in \mathcal{W}((0,1)^2)$, we also have $P_{W_1}(N_{a1}, N_{b1}) \neq 0$, and so the absolutely continuous condition trivially holds.

However, the first condition may not hold if Θ_1' does overlap with the boundary of the unit square. For example, suppose there exists $\theta_0 = (p,p) \in \Theta_0'$ for some $p \in (0,1)$ and some $\theta_1 = (\theta_a, \theta_b) \in \Theta_1'$ such that $\theta_a = 0$ ($\theta_b = 0$ will lead to the same problem by swapping the values of N_{a1} and N_{b1}). Let W_1 be a point mass on θ_1, and take $N_{a1} = 1$ and $N_{b1} = k$. Then $P_{W_1}(N_{a1}, N_{b1}) = 0$, while $P_{\theta_0}(N_{a1}, N_{b1}) \neq 0$. Thus P_{θ_0} is not absolutely continuous relative to P_{W_1}.

In many of our cases $\Theta_1(\underline{\epsilon})'$ *does* contain elements of the boundary of the unit square. This can be remedied by taking the intersection with $[c, 1-c]^2$, for any $c > 0$. However, note that all the numerical experiments of chapter five and theoretical results of chapter six demonstrate that the supports of W_1^* and W_0^* are strictly separated from the boundary of the unit square for every choice of parameter, threshold and prior knowledge rectangle for which W_1^* and W_0^* are not degenerate. Hence their supports already lie in $[c, 1-c]^2$, for some positive c. Therefore the parameter sets can be safely intersected with this smaller square without increasing the infimum or altering the priors which achieve it. This then ensures the first condition holds without altering the GROW S-value. Thus we can safely ignore condition one.

For the second condition to hold, there need only exist $\theta_0 = (p,p) \in \Theta_0'$ and $\theta_1 = (\theta_a, \theta_b) \in \Theta_1'$ such that $p, \theta_a, \theta_b \in (0,1)$. By independence, we have

$$\mathrm{KL}(P_{\theta_1}||P_{\theta_0}) = n_a \mathrm{kl}(\theta_a||p) + n_b \mathrm{kl}(\theta_b||p) \tag{A.1}$$

$$= n_a \left[\theta_a \log \frac{\theta_a}{p} + (1-\theta_a) \log \frac{1-\theta_a}{1-p}\right] + n_b \left[\theta_b \log \frac{\theta_b}{p} + (1-\theta_b) \log \frac{1-\theta_b}{1-p}\right], \tag{A.2}$$

which is finite since none of the values θ_a θ_b or p is equal to 0 or 1. Such nontrivial p, θ_a and θ_b exist in all the cases we consider.

We now prove Theorem 30, which states that the third condition holds provided $\Theta_1(\underline{\epsilon})'$ and Θ_0' are compact. This suffices, since in all the cases we consider $\Theta_1(\underline{\epsilon})'$ and Θ_0' are indeed compact. The result relies on Prokhorov's Theorem and a theorem due to Posner [9, Theorem 1], for which the following definitions are required.

Definition 4. A metric space (X, d) is *complete* if every Cauchy sequence in X has a limit in X.

Definition 5. A metric space (X, d) is *separable* if X contains a countable dense subset.

Definition 6. A topological space (X, τ) is *sequentially compact* if every sequence of points in X has a convergent subsequence converging to a point in X.

Definition 7. The *Borel σ-algebra* on a metric space (X, d) is the σ-algebra generated by the open sets defined by d.

Definition 8. Let (X, d) be a metric space with Borel σ-algebra $\mathcal{B}$. Let $\mathcal{P}$ denote the collection of all probability measures defined on $\mathcal{B}$. A subset $\mathcal{Q} \subseteq \mathcal{P}$ is *tight* if for any $\epsilon > 0$ there exists a compact $K_\epsilon \subset X$ such that, for all measures $\mu \in \mathcal{Q}$,

$$\mu(K_\epsilon) > 1 - \epsilon. \tag{A.3}$$

Definition 9. Let (X, d) be a metric space with Borel σ-algebra $\mathcal{B}$. A sequence of probability measures $P_1, P_2, \dots$ on $\mathcal{B}$ is said to converge weakly to the probability measure P if

$$\mathbf{E}_{X \sim P_n}[f(X)] \to \mathbf{E}_{X \sim P}[f(X)] \tag{A.4}$$

for every bounded continuous real-valued function f on X. This is written $P_n \Rightarrow P$.

Theorem 28 (Prokhorov's Theorem)**.** *Let (X, d) be a separable metric space with Borel σ-algebra $\mathcal{B}$ and let $\mathcal{P}$ denote the collection of all probability measures defined on $\mathcal{B}$. A subset $\mathcal{Q} \subseteq \mathcal{P}$ is tight if and only if the closure of $\mathcal{Q}$ is sequentially compact in the space $\mathcal{P}$ equipped with the topology of weak convergence.*

Theorem 29 (Posner, [9])**.** *Let (X, d) be a complete separable metric space and let $P_n \Rightarrow P$ and $Q_n \Rightarrow Q$. Then*

$$\mathrm{KL}(P||Q) \leq \liminf_{n \to \infty} \mathrm{KL}(P_n||Q_n). \tag{A.5}$$

In other words, $\mathrm{KL}(P||Q)$ *is jointly lower semi-continuous in P and Q.*

With these definitions and theorems, we can now state and prove the following theorem, which states that condition three holds provided the parameter sets are compact.

Theorem 30. *Let $\Theta_1', \Theta_0' \subseteq [0, 1]^2$ be compact. Let n_a and n_b be positive integers and, for any $\theta = (\theta_a, \theta_b) \in [0, 1]^2$ let P_θ denote the distribution of $Z = (N_{a1}, N_{b1})$, where N_{a1} and N_{b1} are independent with distributions*

$$N_{a1} \sim Bin(n_a, \theta_a) \quad \text{and} \quad N_{b1} \sim Bin(n_b, \theta_b). \tag{A.6}$$

Then there exist W_1^* and W_0^* such that

$$\mathrm{KL}(P_{W_1^*}||P_{W_0^*}) = \inf_{(W_1, W_0) \in \mathcal{W}(\Theta_1') \times \mathcal{W}(\Theta_0)} \mathrm{KL}(P_{W_1}||P_{W_0}). \tag{A.7}$$

Proof. By the definition of the infimum, there exists a sequence $(W_1^{(n)}, W_0^{(n)})$ for $n = 1, 2, \dots$ such that

$$\mathrm{KL}(P_{W_1^{(n)}}||P_{W_1^{(n)}}) \to \inf_{(W_1, W_0) \in \mathcal{W}(\Theta_1') \times \mathcal{W}(\Theta_0)} \mathrm{KL}(P_{W_1}||P_{W_0}) \tag{A.8}$$

as $n \to \infty$. Let $\mathcal{Q}_i = \{W_i^{(n)} : n = 1, 2, \dots\}$ for $i \in \{0, 1\}$. Then, since every $W_i^{(n)}$ is a probability distribution on the compact set $[0, 1]^2$, we see that $\mathcal{Q}_1$ and $\mathcal{Q}_0$ are tight (we simply take $K_\epsilon = [0, 1]^2$ for each $\epsilon > 0$). Thus, by Prokhorov's Theorem (Theorem 28), their closures are sequentially compact, meaning they have subsequences weakly converging to some $\tilde{W}_1$ and $\tilde{W}_0$ respectively. By taking the subsequences sequentially, we can thus find a subsequence $(\tilde{W}_1^{(n)}, \tilde{W}_0^{(n)})$ of $(W_1^{(n)}, W_0^{(n)})$ such that

$$\tilde{W}_1^{(n)} \Rightarrow \tilde{W}_1 \quad \text{and} \quad \tilde{W}_0^{(n)} \Rightarrow \tilde{W}_0 \tag{A.9}$$

as $n \to \infty$. Now, let $f : \mathcal{Z} \to [-B, B]$ be an arbitrary bounded continuous real-valued function. We can then define $g : [0,1]^2 \to \mathbb{R}$ by

$$g(\theta) \coloneqq \mathbf{E}_{Z \sim P_\theta}[f(Z)] \tag{A.10}$$

for any $\theta = (\theta_a, \theta_b) \in [0,1]^2$. Then for any θ

$$|g(\theta)| \leq \mathbf{E}_{Z \sim P_\theta}[|f(Z)|] \leq \mathbf{E}_{Z \sim P_\theta}[B] \leq B, \tag{A.11}$$

so g is bounded. Further, since $P_\theta(z)$ is continuous in θ for every $z \in \mathcal{Z}$, we see that g is continuous. Thus, for $i \in \{0, 1\}$, by the definition of weak convergence we have

$$\begin{aligned}
\mathbf{E}_{Z \sim P_{\tilde{W}_i^{(n)}}}[f(Z)] &= \mathbf{E}_{\theta \sim \tilde{W}_i^{(n)}}\left[\mathbf{E}_{Z \sim P_\theta}[f(Z)]\right] && \text{(A.12)}\\
&= \mathbf{E}_{\theta \sim \tilde{W}_i^{(n)}}[g(\theta)] && \text{(A.13)}\\
&\Rightarrow \mathbf{E}_{\theta \sim \tilde{W}_i}[g(\theta)] && \text{(A.14)}\\
&= \mathbf{E}_{\theta \sim \tilde{W}_i}\left[\mathbf{E}_{Z \sim P_\theta}[f(Z)]\right] && \text{(A.15)}\\
&= \mathbf{E}_{Z \sim P_{\tilde{W}_i}}[f(Z)]. && \text{(A.16)}
\end{aligned}$$

Since f was an arbitrary bounded continuous real-valued function, we see that $P_{\tilde{W}_1^{(n)}} \Rightarrow P_{\tilde{W}_1}$ and $P_{\tilde{W}_0^{(n)}} \Rightarrow P_{\tilde{W}_0}$. Finally, by the joint lower semi-continuity of the KL-divergence (Theorem 29), we see that

$$\begin{aligned}
\mathrm{KL}(P_{\tilde{W}_1} || P_{\tilde{W}_0}) &\leq \liminf_{n \to \infty} \mathrm{KL}(P_{\tilde{W}_1^{(n)}} || P_{\tilde{W}_0^{(n)}}) && \text{(A.17)}\\
&= \lim_{n \to \infty} \mathrm{KL}(P_{\tilde{W}_1^{(n)}} || P_{\tilde{W}_0^{(n)}}) && \text{(A.18)}\\
&= \inf_{(W_1, W_0) \in \mathcal{W}(\Theta_1') \times \mathcal{W}(\Theta_0)} \mathrm{KL}(P_{W_1} || P_{W_0}) && \text{(A.19)}
\end{aligned}$$

and so

$$\mathrm{KL}(P_{\tilde{W}_1} || P_{\tilde{W}_0}) = \inf_{(W_1, W_0) \in \mathcal{W}(\Theta_1') \times \mathcal{W}(\Theta_0)} \mathrm{KL}(P_{W_1} || P_{W_0}). \tag{A.20}$$

□

Appendix B

Proofs

We restate and give the proof of Lemma 22.

Lemma. *Let* $\gamma \in [-1, 1] \setminus \{0\}$ *and define* $f(q) = \mathrm{kl}(q + \gamma || q)$ *for* $\max\{0, -\gamma\} \leq q \leq \min\{1, 1 - \gamma\}$. *Then* f *is strictly convex.*

Proof. We show that f is always either continuous or infinite at the endpoints and has positive second derivative on the open interval.

First, suppose $\gamma > 0$, so that $q \in [0, 1 - \gamma]$. Then

$$f(0) = \mathrm{kl}(\gamma||0) = \gamma \log \frac{\gamma}{0} + (1 - \gamma) \log \frac{1 - \gamma}{1} = \infty, \tag{B.1}$$

$$f(1 - \gamma) = \mathrm{kl}(1||1 - \gamma) = 1 \log \frac{1}{1 - \gamma} + 0 \log \frac{0}{\gamma} = -\log(1 - \gamma) \quad \text{and} \tag{B.2}$$

$$\lim_{q \uparrow 1-\gamma} f(q) = \lim_{q \uparrow 1-\gamma} \left\{ (q + \gamma) \log \frac{q + \gamma}{q} + (1 - (q + \gamma)) \log \frac{1 - (q + \gamma)}{1 - q} \right\} \tag{B.3}$$

$$= -\log(1 - \gamma), \tag{B.4}$$

using continuity and the fact that $x \log x \to 0$ as $x \downarrow 0$. Therefore f is infinite at the left endpoint and continuous at the right endpoint.

Likewise, if $\gamma < 0$, so that $q \in [-\gamma, 1]$, we have

$$f(-\gamma) = \mathrm{kl}(0|| - \gamma) = 0 \log \frac{0}{-\gamma} + 1 \log \frac{1}{1 + \gamma} = -\log(1 + \gamma), \tag{B.5}$$

$$f(1) = \mathrm{kl}(1 + \gamma||1) = (1 + \gamma) \log \frac{1 + \gamma}{1} - \gamma \log \frac{-\gamma}{0} = \infty \quad \text{and} \tag{B.6}$$

$$\lim_{q \downarrow -\gamma} f(q) = \lim_{q \downarrow -\gamma} \left\{ (q + \gamma) \log \frac{q + \gamma}{q} + (1 - (q + \gamma)) \log \frac{1 - (q + \gamma)}{1 - q} \right\} \tag{B.7}$$

$$= -\log(1 - \gamma), \tag{B.8}$$

again using continuity and the fact that $x \log x \to 0$ as $x \downarrow 0$. Therefore f is infinite at the right endpoint and continuous at the left endpoint.

We now calculate the second derivative of f in the open interval. Expanding the KL divergence, we have

$$f(q) = (q + \gamma) \log \frac{q + \gamma}{q} + (1 - (q + \gamma)) \log \frac{1 - (q + \gamma)}{1 - q}, \tag{B.9}$$

and, taking the derivative, we get

$$f'(q) = \log\frac{q+\gamma}{q} + (q+\gamma)\left(\frac{1}{q+\gamma} - \frac{1}{q}\right) - \log\frac{1-(q+\gamma)}{1-q} \tag{B.10}$$

$$+ (1-(q+\gamma))\left(-\frac{1}{1-(q+\gamma)} + \frac{1}{1-q}\right) \tag{B.11}$$

$$= \log\frac{q+\gamma}{q} - \frac{\gamma}{q} - \log\frac{1-(q+\gamma)}{1-q} - \frac{\gamma}{1-q}. \tag{B.12}$$

The second derivative is then

$$f''(q) = \frac{1}{q+\gamma} - \frac{1}{q} + \frac{\gamma}{q^2} + \frac{1}{1-(q+\gamma)} - \frac{1}{1-q} - \frac{\gamma}{(1-q)^2} \tag{B.13}$$

$$= \frac{\gamma^2(2\gamma q - \gamma - 3q(1-q) + 1)}{q^2(1-q)^2(q+\gamma)(1-(q+\gamma))}. \tag{B.14}$$

Since we are now working in the open interval, we have $q+\gamma > 0$, $q > 0$, $1-q > 0$ and $1-(q+\gamma) > 0$. Therefore

$$f''(q) > 0 \iff 2\gamma q - \gamma - 3q(1-q) + 1 > 0 \tag{B.15}$$

$$\iff 3q^2 + (2\gamma - 3)q + 1 - \gamma > 0. \tag{B.16}$$

The discriminant of this quadratic is $(2\gamma-3)^2 - 12(1-\gamma) = 4\gamma^2 - 3$. Thus if $\gamma \in (-\sqrt{3}/2, \sqrt{3}/2)$ the quadratic is always positive and f is strictly convex. Alternatively, suppose $\gamma < -\sqrt{3}/2$, so that $q \in [-\gamma, 1]$. Let q_-, q_+ be the lower and upper solutions of the quadratic respectively. We have

$$q_+ \coloneqq \frac{3 - 2\gamma + \sqrt{4\gamma^2 - 3}}{6} < -\gamma \iff \sqrt{4\gamma^2 - 3} < -4\gamma - 3 \tag{B.17}$$

$$\iff 4\gamma^2 - 3 < 16\gamma^2 + 24\gamma + 9 \quad (\text{since } 4\gamma^2 - 3 > 0) \tag{B.18}$$

$$\iff \gamma^2 + 2\gamma + 1 > 0 \tag{B.19}$$

$$\iff (\gamma+1)^2 > 0. \tag{B.20}$$

Since the last line is trivially true, we see that $q_+ < -\gamma$ and so $f''(q) > 0$ for $q \in [-\gamma, 1]$.

Likewise, if $\gamma > \sqrt{3}/2$, so that $q \in [0, 1-\gamma]$, we have

$$q_- \coloneqq \frac{3 - 2\gamma - \sqrt{4\gamma^2 - 3}}{6} > 1 - \gamma \iff \sqrt{4\gamma^2 - 3} < 4\gamma - 3 \tag{B.21}$$

$$\iff 4\gamma^2 - 3 < 16\gamma^2 - 24\gamma + 9 \quad (\text{since } 4\gamma^2 - 3 > 0) \tag{B.22}$$

$$\iff \gamma^2 + 2\gamma + 1 > 0 \tag{B.23}$$

$$\iff (\gamma+1)^2 > 0. \tag{B.24}$$

Again, since the last line is trivially true, we see that $q_- > 1-\gamma$ and so $f''(q) > 0$ for $q \in [0, 1-\gamma]$. □

We now restate and give the proof of Lemma 25

Lemma. *Let $\gamma > 0$ and define $f(q) \coloneqq \mathrm{kl}(q||\gamma q)$. Then f is strictly increasing in q.*

Proof. By direct calculation, we show that $f'(0) > 0$ and $f''(q) > 0$ for all q, which, since f is continuously differentiable, implies that $f'(q) > 0$ for all q.

$$f(q) = q\log\frac{q}{\gamma q} + (1-q)\log\frac{1-q}{1-\gamma q} \tag{B.25}$$

$$= -q\log\gamma + (1-q)\log\frac{1-q}{1-\gamma q}. \tag{B.26}$$

$$\tag{B.27}$$

Differentiating, we get

$$f'(q) = -\log\gamma + (1-q)\left[-\frac{1}{1-q} + \frac{\gamma}{1-\gamma q}\right] - \log\frac{1-q}{1-\gamma q} \tag{B.28}$$

$$= -\log\gamma - 1 + \frac{\gamma(1-q)}{1-\gamma q} - \log\frac{1-q}{1-\gamma q} \tag{B.29}$$

$$= -\log\gamma + \frac{\gamma-1}{1-\gamma q} - \log\frac{1-q}{1-\gamma q}. \tag{B.30}$$

This gives $f'(0) = -\log\gamma + \gamma - 1$ which is positive by the well-known inequality $\log x < x - 1$. Differentiating again, we have

$$f''(q) = \frac{\gamma(\gamma-1)}{(1-\gamma q)^2} + \frac{1}{1-q} - \frac{\gamma}{1-\gamma q} \tag{B.31}$$

$$= \frac{\gamma(\gamma-1)}{(1-\gamma q)^2} + \frac{1-\gamma}{(1-q)(1-\gamma q)} \tag{B.32}$$

$$= \frac{\gamma-1}{(1-\gamma q)^2(1-q)}[\gamma(1-q) - (1-\gamma q)] \tag{B.33}$$

$$= \frac{(\gamma-1)^2}{(1-\gamma q)^2(1-q)}, \tag{B.34}$$

which is positive since $q \in I_{\mathrm{PKR}} \cap [0, 1/\lambda]$, so $q \leq 1/\lambda < 1$ since $\lambda > 1$. Thus $f'(0) > 0$ and $f''(q) > 0$ for all q, meaning $f'(q) > 0$ for all q and so $f(\cdot)$ is strictly increasing. □

Bibliography

[1] G. A. Barnard. A new test for 2× 2 tables. *Natur*, 156(3954):177, 1945.

[2] R. Boschloo. Raised conditional level of significance for the 2x2-table when testing the equality of two probabilities. *Statistica Neerlandica*, 24:1 – 9, 04 2008.

[3] L. Choi, J. D. Blume, and W. D. Dupont. Elucidating the foundations of statistical inference with 2 x 2 tables. *PloS one*, 10(4):e0121263, 2015.

[4] R. A. Fisher. On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

[5] P. Grünwald, R. de Heide, and W. Koolen. Safe testing. *arXiv preprint arXiv:1906.07801*, 2019.

[6] P. D. Grünwald and A. Grunwald. *The minimum description length principle*. MIT press, 2007.

[7] V. E. Johnson. Uniformly most powerful bayesian tests. *Annals of statistics*, 41(4):1716, 2013.

[8] K. Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

[9] E. Posner. Random coding strategies for minimum entropy. *IEEE Transactions on Information Theory*, 21(4):388–391, 1975.

[10] Y. Qiao and N. Minematsu. A study on invariance of f-divergence and its application to speech recognition. *IEEE Transactions on Signal Processing*, 58(7):3884–3890, 2010.

[11] J. ter Schure and P. Grünwald. Accumulation bias in meta-analysis: the need to consider time in error control. *F1000Research*, 8, 2019.

[12] R. Turner. Safe tests for 2 x 2 contingency tables and the cochran-mantel-haenszel test. Master's thesis, University of Leiden, 2019.

[13] T. Van Erven and P. Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.