S.L. van der Pas

Much ado about the *p*-value

Fisherian hypothesis testing versus an alternative test, with an application to highly-cited clinical research.

Bachelor thesis, June 16, 2010

Thesis advisor: prof.dr. P.D. Grünwald

Mathematisch Instituut, Universiteit Leiden

Contents

Introduction

πάντες ἄνθρωποι τοῦ εἰδέναι ὀρέγονται φύσει.
All men by nature desire to know.
— Aristoteles, *Metaphysica* I.980a.

How can we acquire knowledge? That is a fundamental question, with no easy answer. This thesis is about the statistics we use to gain knowledge from empirical data. More specifically, it is a study of some of the current statistical methods that are used when one tries to decide whether a hypothesis is correct, or which hypothesis from a set of hypotheses fits reality best. It tries to assess whether the current statistical methods are well equipped to handle the responsibility of deciding which theory we will accept as the one that, for all intents and purposes, is true.

The first chapter of this thesis touches on the debate about how knowledge *should* be gained: two popular methods, one ascribed to R. Fisher and the other to J. Neyman and E. Pearson, are explained briefly. Even though the two methods have come to be combined in what is often called Null Hypothesis Significance Testing (NHST), the ideas of their founders clashed and a glimpse of this quarrel can be seen in the Section that compares the two approaches.

The first chapter is introductory; the core part of this thesis is in Chapters 2 and 3. Chapter 2 is about Fisherian, p-value based hypothesis testing and is primarily focused on the problems associated with it. It starts out with discussing what kind of knowledge the users of this method believe they are obtaining and how that compares to what they are actually learning from the data. It will probably not come as a surprise to those who have heard the jokes about psychologists and doctors being notoriously bad at statistics that the perception of the users does not match reality. Next, problems inherent to the method are considered: it depends on data that were never observed and on such sometimes uncontrollable factors as whether funding will be revoked or participants in studies will drop out. Furthermore, evidence against the null hypothesis seems to be exaggerated: in Lindley's famous paradox, a Bayesian analysis of a random sample will be shown to lead to an entirely different conclusion than a classical frequentist analysis. It is also proven that sampling to a foregone conclusion is possible using NHST. Despite all these criticisms, p-values are still very popular. In order to understand why their use has not been abandoned, the chapter concludes with some arguments in defense of using p-values.

In Chapter 3, a somewhat different statistical test is considered, based on a likelihood ratio. In the first section, a property of the test regarding error rates is proven and it is compared to a standard Neyman-Pearson test. In the next section, the test is compared to Fisherian hypothesis testing and is shown to fare better on all points raised in Chapter 2. This thesis then takes a practical turn by discussing some real clinical studies, taking an article by J.P.A. Ioannidis as a starting point. In this article, some disquieting claims about the correctness of highly cited medical research articles were made. This might be partly due to the use of unfit statistical methods. To illustrate this, two calibrations are performed on 15 of the articles studied by Ioannidis. The first of these calibrations is based on the alternative test introduced in this chapter, the second one is based on Bayesian arguments similar to those in chapter two. Many results that were considered 'significant', indicated by small p-values, turn out not to be significant anymore when calibrated.

The conclusion of the work presented in this thesis is therefore that there is much ado about the p-value, and for good reasons.

1 Overview of frequentist hypothesis testing

1.1 Introduction

What is a 'frequentist'? A frequentist conceives of probability as limits of relative frequencies. If a frequentist says that the probability of getting heads when flipping a certain coin is $\frac{1}{2}$, it is meant that if the coin were flipped very often, the relative frequency of heads to total flips would get arbitrarily close to $\frac{1}{2}$ [1, p.196]. The tests discussed in the next two sections are based on this view of probability. There is another view, called Bayesian. That point of view will be explained in Section 2.5.1.

The focus of the next chapter of this thesis will be the controversy that has arisen over the use of p-values, which are a feature of Fisherian hypothesis testing. Therefore, a short explanation of this type of hypothesis testing will be given. Because the type I errors used in the Neyman-Pearson paradigm will play a prominent part in Chapter 3, a short introduction to Neyman-Pearson hypothesis testing will be useful as well. Both of these paradigms are frequentist in nature.

1.2 Fisherian hypothesis testing

A 'hypothesis test' is a bit of a misnomer in a Fisherian framework, where the term 'significance test' is to be preferred. However, because of the widespread use of 'hypothesis test', this term will be used in this thesis as well. The p-value is central to this test. The p-value was first introduced by Karl Pearson (not the same person as Egon Pearson from the Neyman-Pearson test), but popularized by R.A. Fisher [2]. Fisher played a major role in the fields of biometry and genetics, but is most well-known for being the 'father of modern statistics'. As a practicing scientist, Fisher was interested in creating an objective, quantitative method to aid the process of inductive inference [3]. In Fisher's model, the researcher proposes a null hypothesis that a sample is taken from a hypothetical population that is infinite and has a known sampling distribution. After taking the sample, the p-value can be calculated. To define a p-value, we first need to define a *sample space* and a *test statistic.*

Definition 1.1 (sample space) The *sample space* $\mathcal{X}$ is the set of all outcomes of an event that may potentially be observed. The set of all possible samples of length n is denoted $\mathcal{X}^n$.

Definition 1.2 (test statistic) A *test statistic* is a function $T : \mathcal{X} \to \mathbb{R}$.

We also need some notation: $P(A|H_0)$ will denote the probability of the event A, under the assumption that the null hypothesis H_0 is true. Using this notation, we can define the p-value.

Definition 1.3 (p-value) Let T be some test statistic. After observing data x_0, then
$p = P(T(X) \geq T(x_0)|H_0)$.

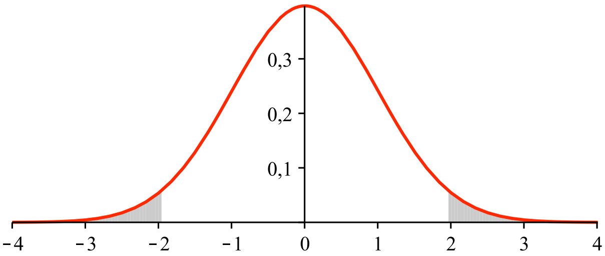


Figure 1: For a standard normal distribution with $T(X) = |X|$, the p-value after observing $x_0 = 1.96$ is equal to the shaded area (graph made in Maple 13 for Mac).

The statistic T is usually chosen such that large values of T cast doubt on the null hypothesis H_0. Informally, the p-value is the probability of the observed result or a more extreme result, assuming the null hypothesis is true. This is illustrated for the standard normal distribution in Figure 1. Throughout this thesis, all p-values will be two-sided, as in the example. Fisher considered p-values from single experiments to provide inductive evidence against H_0, with smaller p-values indicating greater evidence. The rationale behind this test is Fisher's famous disjunction: if a small p-value is found, then either a rare event has occured or else the null hypothesis is false.

It is thus only possible to reject the null hypothesis, not to prove it is true. This Popperian viewpoint is expressed by Fisher in the quote:

> *"Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis."*
> — R.A. Fisher (1966).[1]

Fisher's p-value was not part of a formal inferential method. According to Fisher, the p-value was to be used as a measure of evidence, to be used to reflect on the credibility of the null hypothesis, in light of the data. The p-value in itself was not enough, but was to be combined with other sources of information about the hypothesis that was being studied. The outcome of a Fisherian hypothesis test is therefore an inductive inference: an inference about the population based on the samples.

1.3 Neyman-Pearson hypothesis testing

The two mathematicians Jerzy Neyman and Egon Pearson developed a different method of testing hypotheses, based on a different philosophy. Whereas a Fisherian hypothesis test only requires one hypothesis, for a Neyman-Pearson hypothesis test two hypotheses need to be specified: a null hypothesis H_0 and an alternative hypothesis H_A. The reason for this, as explained by Pearson, is:

> *"The rational human mind did not discard a hypothesis until it could conceive at least one plausible alternative hypothesis".*
> — E.S Pearson (1990.)[2]

Consequently, we will compare two hypotheses. When deciding between two hypotheses, two types of error can be made:

Definition 1.4 (Type I error) A *type I error* occurs when H_0 is rejected while H_0 is true. The probability of this event is usually denoted by α.

Definition 1.5 (Type II error) A *type II error* occurs when H_0 is accepted while H_0 is false. The probability of this event is usually denoted by β.

	accept H_0	reject H_0
H_0 true	✓	type I error (α)
H_A true	type II error (β)	✓

Table 1: Type I and type II errors.

The *power* of a test is then the probability of rejecting a false null hypothesis, which equals $1 - \beta$. When designing a test, first the type I error probability α is specified. The best test is then the one that minimizes the type II error β within the bound set by α. That this 'most powerful test' has the form of a likelihood ratio test is proven in the famous Neyman-Pearson lemma, which is discussed in Section 3.1.3. There is a preference for choosing α small, usually equal to 0.05, whereas β can be larger. The α and β error rates then define a 'critical' region for the test statistic. After an experiment, one should only report whether the result falls in the critical region, not where it fell. If the test statistic

[1] Fisher, R.A. (1966^8), *The design of experiments*, Oliver & Boyd (Edinburg), p.16, cited by [2, p.298].

[2] Pearson, E.S. (1990), *'Student'. A statistical biography of William Sealy Gosset*, Clarendon Press (Oxford), p.82, cited by [2, p.299].

falls in that region, H_0 is rejected in favor of H_A, else H_0 is accepted. The outcome of the test is therefore not an inference, but a behavior: acceptance or rejection.

An important feature of the Neyman-Pearson test is that it is based on the assumption of repeated random sampling from a defined population. Then, α can be interpreted as the long-run relative frequency of type I errors and β as the long-run relative frequency of type II errors [2]. The test does not include a measure of evidence. An important quote of Neyman and Pearson regarding the goal of their test is:

> *"We are inclined to think that as far as a particular hypothesis is concerned, no test based upon a theory of probability can by itself provide any valuable evidence of the truth or falsehood of a hypothesis.*
>
> *But we may look at the purpose of tests from another viewpoint. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not often be wrong."*
> — J. Neyman and E. Pearson (1933).[3]

Therefore, their test cannot measure evidence in an individual experiment, but limits the number of mistakes made over many different experiments. Whether we *believe* a hypothesis we 'accept' is not the issue, it is only necessary to *act* as though it were true. Goodman compares this to a system of justice that is not concerned with whether an individual defendant is guilty or innocent, but tries to limit the overall number of incorrect verdicts, either unjustly convicting someone, or unjustly acquitting someone [4, p.998]. The preference for a low type I error probability can then be interpreted as a preference for limiting the number of persons unjustly convicted over limiting the number of guilty persons that go unpunished.

1.4 Differences between the two approaches

Nowadays, the two approaches are often confused and used in a hybrid form: an experiment is designed to control the two types of error (typically $\alpha = 0.05$ and $\beta < 0.20$). After the data have been observed, the p-value is calculated. If it is less than α, the results are declared 'significant'. This fusion is quite remarkable, as there are many differences that make the two types of tests incompatible. Most of these differences are apparent from the discussion in the previous two sections. Some aspects that spring to mind are that the outcome of a Fisherian test is an inference, while the outcome of a Neyman-Pearson test is a behaviour, or that a type I error rate α is decided in advance, while the p-value can only be calculated after the data have been collected. The distinction between p and α is made very clear in Table 2, reproduced in abbreviated form from Hubbard [2, p.319].

Table 2: Contrasting p's and α's

p-value	α level
Fisherian significance level	Neyman-Pearson significance level
Significance test	Hypothesis test
Evidence against H_0	Type I error - erroneous rejection of H_0
Inductive inference - guidelines for interpreting strength of evidence in data	Inductive behavior - guidelines for making decisions based on data
Data-based random variable	Pre-assigned fixed value
Property of data	Property of test
Short-run - applies to any single experiment/study	Long-run - applies only to ongoing, identical repetitions of original experiment/study, not to any given study

[3] Neyman, J., Pearson, E. (1933), On the problem of the most efficient tests of statistical hypotheses, *Philosophical Transactions of the Royal Society London A*, **231**, 289-337, p.290-291, cited in [3, p.487].

It is somewhat ironic that both methods have come to be combined, because their 'founding fathers' did not see eye to eye at all and spoke derisively about each other. Some quotes from Fisher about the Neyman-Pearson test are very clear about this:[4]

> *"The differences between these two situations seem to the author many and wide, and I do not think it would have been possible had the authors of this reinterpretation had any real familiarity with work in the natural sciences, or consciousness of those features of an observational record which permit of an improved scientific understanding."*
>
> *"The concept that the scientific worker can regard himself as an inert item in a vast cooperative concern working according to accepted rules is encouraged by directing attention away from his duty to form correct scientific conclusions, to summarize them and to communicate them to his scientific colleagues, and by stressing his supposed duty mechanically to make a succession of automatic "decisions". "*
>
> *"The idea that this responsibility can be delegated to a giant computer programmed with Decision Functions belongs to a phantasy of circles rather remote from scientific research. The view has, however, really been advanced (Neyman, 1938) that Inductive Reasoning does not exist, but only "Inductive Behaviour"! "*
> — R. Fisher (1973).[5]

Even though Fisher and Neyman and Pearson obviously considered their tests to be incompatible, the union of both seems to be irreversible. How did this come to be? Some factors seem to be that Neyman and Pearson used, for convenience, Fisher's 5% and 1% significance levels to define their type I error rates and that the terminology is ambiguous. Nowadays, many textbooks and even the American Psychological Association's *Publication Manual* present the hybrid form as one unified theory of statistical inference. It should therefore not come as a surprise that Hubbard found that out of 1645 articles using statistical tests from 12 psychology journals, at least 1474 used a hybrid form of both methods. Perhaps somewhat more surprisingly, he also found that many critics of null hypothesis significance testing used p's and α's interchangeably [2]. Until awareness of the historical debate that has preceded the combination of the Fisherian and the Neyman-Pearson paradigms raises, this confusion will continue to exist and continue to hinder correct interpretation of results.

The confusion of the two paradigms is problematic, but it turns out that it is but one of the many problems associated with hypothesis testing. In the next chapter, the main criticisms on the use of p-values will be discussed. These concern both the incorrect interpretation of what p-values are and properties that raise doubts whether p-values are fit to be used as measures of evidence. This is very disconcerting, as p-values are a very popular tool used by medical researchers to decide whether differences between groups of patients (for example, a placebo group and a treatment group) are significant. The decisions based on this type of research affect the everyday life of many people. It is therefore important that these decisions are made by means of sound methods. If we cannot depend on p-values to aid us in a desirable manner while making these decisions, we should consider alternatives. Therefore, after reviewing the undesirable properties of p-values in Chapter 2, in Chapter 3 an alternative test will be considered and compared to a p-value test. A similar test is then applied to medical research articles. The particular articles discussed have been selected from a set of highly cited medical research articles considered in an article that shows that approximately 30% of them has later been contradicted or shown to have claimed too strong results. The application of the alternative test will show that the use of p-values is very common, but can give a misleading impression of the probability that the conclusions drawn based on them are correct.

[4] Many more can be found in Hubbard [2], 299-306.

[5] R. Fisher, (1973^3) *Statistical methods and scientific inference*, Macmillan (New York), p.79-80, p.104-105, cited in [3, p.488].

2 Problems with p-values

2.1 Introduction

> *"I argue herein that NHST has not only failed to support the advance of psychology as a science but also has seriously impeded it. (...) What's wrong with NHST? Well, among many other things, it does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!"*
> — J. Cohen (1994), [5, p.997].

> *"This paper started life as an attempt to defend p-values, primarily by pointing out to theoreticians that there are more things in the clinical trials industry than are dreamed of in their lecture courses and examination papers. I have, however, been led inexorably to the opposite conclusion, that the current use of p-values as the 'main means' of assessing and reporting the results of clinical trials is indefensible."*
> — P.R. Freeman (1993), [6, p.1443].

> *"The overall conclusion is that P values can be highly misleading measures of the evidence provided by the data against he null hypothesis."*
> — J.O. Berger and T. Sellke (1987), [7, p.112].

For decades, articles criticizing the use of Null Hypothesis Significance Testing (NHST) have been published. This has led to the banning or serious discouragement of the use of p-values in favor of confidence intervals by some journals, most prominently by the *American Journal of Public Health* (*AJPH*) and *Epidemiology*, both under the influence of editor Kenneth Rothman.[6] There was no official ban on p-values at *AJPH*, but Rothmans revise-and-submit letters spoke volumes, for example [8, p.120]:

> *"All references to statistical hypothesis testing and statistical significance should be removed from the paper. I ask that you delete p values as well as comments about statistical significance. If you do not agree with my standards (concerning the inappropriateness of significance tests), you should feel free to argue the point, or simply ignore what you may consider to be my misguided view, by publishing elsewhere. "*

However, this is an exception and most journals continue to use p-values as measures of statistical evidence. In this section, the most serious criticisms on p-values are reviewed. The starting point for this section has been Wagenmakers [9], but some additional literature providing more details on the various problems has been used and is cited at the appropriate places. In an effort to prevent that NHST will be convicted without a fair trial, the most common answers of NHST advocates to these criticisms are included in Section 2.7.

[6] Note, however, that even though the use of confidence intervals is also advocated in highly cited articles such as Gardner, M.J., Altman, D.G., (1986), Confidence intervals rather than P values: estimation rather than hypothesis testing, *British Medical Journal*, **292**, 746-750, confidence intervals have many of the same problems as p-values do. They, too, are often misinterpreted and affected by optional stopping. See for example Mayo, D.G. (2008), How to discount double-counting when it counts: some clarifications, *British Journal for the Philosophy of Science*, **59**, 857-879 (especially page 866-7), Lai, T.L., Su, Z., Chuang, C.S. (2006), Bias correction and confidence intervals following sequential tests, *IMS Lecture Notes - Monograph series. Recent developments in nonparametric inference and probability*, **50**, 44-57, Coad, D.S., Woodroofe, M.B. (1996), Corrected confidence intervals after sequential testing with applications to survival analysis, *Biometrika*, **83** (4), 763-777, Belia, S., Fidler, F., Williams, J., Cumming, G. (2005), Researchers misunderstand confidence intervals and standard error bars, *Psychological Methods*, **10**, 389-396, [8] and [6, p.1450]. That optional stopping will be a problem when used in the form 'if θ_0 is not in the confidence interval, then H_0 can be rejected' seems to be intuitively clear from the duality of confidence intervals and hypothesis tests (see [10, p.337]): the confidence interval consists precisely of all those values of θ_0 for which the null hypothesis $H_0 : \theta = \theta_0$ is accepted.

2.2 Misinterpretation

There are many misconceptions about how p-values should be interpreted. One misinterpretation, that a p-value is the same as a type I error rate α, has already been touched upon in Section 1.4. There is, however, an arguably more harmful misinterpretation, which will be discussed in this section.

The most problematic misinterpretation is caused by the fact that the p-value is a conditional value. It is the probability of the observed or more extreme data, given H_0, which can be represented by $P(x|H_0)$. However, many researchers confuse this with the probability that H_0 is true, given the data, which can be represented by $P(H_0|x)$. This is a wrong interpretation, because the p-value is calculated on the assumption that the null hypothesis true. Therefore, it cannot also be a measure of the probability that the null hypothesis is true [4]. That these two conditional probabilities are not the same is shown very strikingly by Lindley's paradox, which will be discussed in Section 2.5.2. Unfortunately, this misconception is very widespread [2, 6]. For example, the following multiple choice question was presented to samples of doctors, dentist and medical students as part of a short test of statistical knowledge:

> A controlled trial of a new treatment led to the conclusion that it is significantly better than placebo ($p < 0.05$). Which of the following statements do you prefer?
>
> 1. It has been proved that treatment is better than placebo.
> 2. If the treatment is not effective, there is less than a 5 per cent chance of obtaining such results.
> 3. The observed effect of the treatment is so large that there is less than a 5 per cent chance that the treatment is no better than placebo.
> 4. I do not really know what a p-value is and do not want to guess.

There were 397 responders. The proportions of those responders who chose the four options were 15%, 19%, 52% and 15%, respectively. Of the responders who had recently had a short course on statistical methods, the proportion choosing option 2 (which, for the record, is the correct answer) increased at the expense of option 4, but the proportion choosing option 3 remained about the same. The ignorance may even be greater, because the test was administered by mail and the response rate was only about 50% [6, p.1445].

Cohen thinks the cause of this misinterpretation is *"a misapplication of deductive syllogistic reasoning."* [5, p.998]. This is most easily explained by means of an example. Start out with:

> If the null hypothesis is correct, then this datum can not occur.
> It has, however, occured.
> Therefore, the null hypothesis is false.

For example:

> Hypothesis: all swans are white.
> While visiting Australia, I saw a black swan.
> Therefore, my hypothesis is false.

This reasoning is correct. Now tweak the language so that the reasoning becomes probabilistic:

> If the null hypothesis is correct, then these data are highly unlikely.
> These data have occured.
> Therefore, the null hypothesis is unlikely.

This reasoning is incorrect. This can be seen by comparing it to this syllogism:

> If a woman is Dutch, then she is probably not the queen of the Netherlands.
> This woman is the queen of the Netherlands.
> Therefore, she is probably not Dutch.

Although everyone would agree with the first statement, few will agree with the conclusion. In this example, it is easy to see that the logic must be wrong, because we know from experience that the conclusion is not correct. But when the syllogism is about more abstract matters, this mistake is easily made and unfortunately, is made very often, as was illustrated by the previous example of the statistical knowledge of medical professionals.

2.3 Dependence on data that were never observed

The phenomenon that data that have not been observed have an influence on p-values is shown in this section by means of two examples.

Assume that we want to test two hypotheses, H_0 and H_0'. We do not want to test them against each other, but we are interested in each hypothesis separately. Suppose that X has the distribution given in Table 3 [11, p.108].

Table 3: Two different sampling distributions

x	0	1	2	3	4
$H_0(x)$	.75	.14	.04	.037	.033
$H_0'(x)$	.70	.25	.04	.005	.005

We use the test statistic $T(x) = x$. Suppose that $x = 2$ is observed. The corresponding p-values for both hypotheses are:

$$p_0 = P(x \geq 2|H_0) = 0.11 \qquad p_0' = P(x \geq 2|H_0') = 0.05.$$

Therefore, the observation $x = 2$ would provide 'significant evidence against H_0' at the 5% level', but would not even provide 'significant evidence against H_0 at the 10% level'. Note that under both hypotheses the observation $x = 2$ is equally likely. If the hypotheses would be considered against each other, this observation would not single out one of the hypotheses as more likely than the other. Therefore, it seems strange to give a different weight of evidence to the observation when each hypothesis is considered in isolation. As Sir Harold Jeffreys famously wrote:

> *"What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred."*
> — H. Jeffreys, (1961).[7]

Another famous example of this phenomenon was given by J.W. Pratt, which is summarized below:

> An engineer measures the plate voltage of a random sample of electron tubes with a very accurate volt-meter. The measurements are all between 75 and 99 and look normally distributed. After performing a normal analysis on these results, the statistician visits the engineer and notices that his volt meter reads only as far as 100, so the population appears to be censored, which would call for a new analysis. However, the engineer tells the statistician that he had another very accurate volt meter which read as far as 1000 volts, which he would have used if any voltage had been over 100. Therefore, the population is effectively uncensored. The next day, the engineer calls the statistician to tell him that the high-range volt meter was not working on the day of the experiment. Thus, a new analysis is required after all.

This seems a strange practice, because the sample would have provided the same data, whether the high-range volt meter worked that day or not. However, a traditional statistician may prefer a new analysis because the data are censored, even though none of the actual observations were censored. The reason for this is that when replicate data sets are generated according to the null hypothesis, the shape of the sampling distribution is changed when the data are censored, as none of the observations can exceed 100. Therefore, even if the observed data were not censored, censoring does have an effect on the p-value [9, p.782-783].

[7] Jeffreys, H. (1961^3), *Theory of probability*, Clarendon Press (Oxford), p.385, cited by [11, p.108].

What these two examples illustrate is that unobserved data (i.e. $x = 3$ and $x = 4$ in the first example and voltages over 100 volts in the second example) can influence a p-value, which seems counterintuitive.

2.4 Dependence on possibly unknown subjective intentions

p-Values depend on subjective intentions of the researcher, as can be seen from the following example, in which p-values will be shown to depend on the sampling plan and other factors that might be unknown.

Suppose that two researchers want to test whether someone can distinguish between Coca-Cola and Fanta that has been colored black. The null hypothesis is that the subject cannot distinguish the two soft drinks from each other. Both researchers decide on a different sampling plan.

Researcher 1

Researcher 1 prepares twelve drinks for the experiment. After each cup, the subject is asked which drink he has just had. As the test statistic T, the number of correct guesses is used. In this situation, the binomial model with $n = 12$ can be applied, given by:

$$P(T(x) = k|\theta) = \binom{12}{k}\theta^k(1-\theta)^{12-k},$$

with θ reflecting the probability that the persons identifies the drink correctly. The null hypothesis can now be modelled by using $\theta = \frac{1}{2}$.

Suppose the data x are: CCCCCWWCCCCW, with C a correct guess and W a wrong guess. The two-sided p-value is then:

$$p = P\left(T(x) \geq 9|\theta = \tfrac{1}{2}\right) + P\left(T(x) \leq 3|\theta = \tfrac{1}{2}\right) \approx 0.146.$$

Thus, researcher 1 cannot reject the null hypothesis.

Researcher 2

Reseacher 2 did not decide in advance how many drinks she would offer the subject but keeps giving him drinks until he guesses wrong for the third time. The test statistic T is now the number of drinks the researcher offers the subject until the third wrong guess. In this situation, the negative binomial model should be applied, given by:

$$P(T(x) = n|\theta) = \binom{n-1}{k-1}\theta^{n-k}(1-\theta)^k,$$

with $1-\theta$ reflecting the probability that the subject identifies the drink incorrectly and k the number of wrong guesses. The null hypothesis can again be modelled by using $\theta = \frac{1}{2}$.

Suppose researcher 2 gets the same data x as researcher 1: CCCCCWWCCCCW. The p-value is then:

$$p = P\left(T(x) \geq 12|\theta = \tfrac{1}{2}\right) = \sum_{n=12}^{\infty}\binom{n-1}{2}\left(\frac{1}{2}\right)^n = 1 - \sum_{n=1}^{11}\binom{n-1}{2}\left(\frac{1}{2}\right)^n \approx 0.033.$$

Hence, researcher 2 *does* obtain a significant result, with the exact same data as researcher 1!

Discussion

From this example, we see that the same data can yield different p-values, depending on the intention with which the experiment was carried out. In this case, it is intuitively clear why the same data do not yield the same p-values, because the sampling distribution is different for each experiment. This dependence on the sampling plan *is* problematic however, because few researchers are completely aware of all of their own intentions. Consider for example a researcher whose experiment involves 20 subjects [9]. A standard null hypothesis test yields $p = 0.045$, which leads to the rejection of the null hypothesis. Before the researcher publishes his findings, a statistician asks him: "What would you have done if the effect had not been significant after 20 subjects?" His answer may be that he does not know, that he would have tested 20 more subjects and then stopped, that it depends on the p-value he obtained or on whether he had time to test more subjects or on whether he would get more funding. In all these circumstances, the p-value has to either be adjusted or is not defined at all. The only answer that would not have affected the p-value would have been: "I would not have tested any more subjects."

And this is not the only question a researcher has to ask himself beforehand. He should also consider what he would do if participants dropped out, if there were anomalous responses, if the data turn out to be distributed according to a different distribution than expected, etc. It is impossible to specify all these things beforehand and therefore impossible to calculate the correct p-value. Many people feel that a statistical test should only depend on the data itself, not on the intentions of the researcher who carried out the experiment.

2.5 Exaggeration of the evidence against the null hypothesis

This section contains two examples that show that a small p-value does not necessarily imply that the probability that the null hypothesis is correct, is low. In fact, it can be quite the opposite: even though the p-value is arbitrarily small, the probability that the null hypothesis is true can be more than 95%. This is Lindley's renowned paradox in a nutshell. It will be proven to be true in Section 2.5.2. A more general example in Section 2.5.3 will also show that the evidence against the null hypothesis can be much less serious than a p-value may lead one to think when applying the wrong syllogistic reasoning discussed in Section 2.2

2.5.1 Bayes' theorem

In order to understand the proofs in the following two sections, some familiarity with Bayes' theorem is required and therefore, this theorem will now be stated and proven.

Theorem 2.1 (Bayes' theorem) If for two events A and B it holds that $P(A) \neq 0, P(B) \neq 0$, then:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Proof of theorem 2.1
Using the definition of conditional probability, we can write:
$P(A|B) = \dfrac{P(A|B)}{P(B)}$.
$P(B|A) = \dfrac{P(A|B)}{P(A)}$.
The second identity implies $P(A|B) = P(B|A)P(A)$. Substituting this in the first identity proves the theorem.

□

To calculate $P(B)$, the law of total probability is often used: $P(B) = \sum_i P(B|A_i)P(A_i)$. This can be extended for continuous variables x and λ [1, p.198]:

$$w(\lambda|x) = \frac{v(x|\lambda)w(\lambda)}{v(x)} = \frac{v(x|\lambda)w(\lambda)}{\int v(x|\lambda)w(\lambda)d\lambda}.$$

All the factors in this expression have their own commonly used names:

$w(\lambda|x)$ is the *posterior density* of λ, given x.
$w(\lambda)$ is the *prior density* of λ.
$v(x)$ is the *marginal density* of x.

The theorem itself is a mathematical truth and therefore not controversial at all. However, its application sometimes is. The reason for this is the prior $w(\lambda)$. The prior represents your beliefs about the value of λ. For example, before you are going to test whether you have a fever using a thermometer, you may believe that values between 36 °C and 40 °C are quite likely and therefore, you would put most mass on these values in your prior distribution of your temperature. This subjective element of Bayes' theorem is what earns it most of its criticism. Sometimes, everyone agrees what the prior probability of some hypothesis is, for example in HIV testing (see Section 3.2.4). But in most cases, there is no agreement on what the shape of the prior should be. For example, what is the prior probability that a new treatment is better than a placebo? The owner of the pharmaceutical company that produces the medicine will probably have a different opinion on that than a homeopathic healer. However, if priors are not too vague and variable, they frequently have a negligible effect on the conclusions obtained from Bayes' theorem and two people with widely divergent prior opinions but reasonably open minds will be forced into close agreement about future observations by a sufficient amount of data [1]. An alternative solution is to perform some sort of sensitivity analysis using different types of prior [12] or to derive 'objective' lower bounds (see Section 2.5.3).

When a prior probability can be derived by frequentist means, frequentists apply Bayes' theorem too. What is different about Bayesian statistics? Bayesian statistics is an approach to statistics in which *all* inferences are based on Bayes' theorem. An advantage of the Bayesian approach is that it allows to express a degree of belief about any unknown but potentially observable quantity, for example the probability that the Netherlands will host the Olympic games in 2020. For a frequentist, this might be difficult to interpret as part of a long-run series of experiments. Bayes' theorem also allows us to calculate the probability of the null hypothesis given the data, which is in most cases impossible from a frequentist perspective. Even though the p-value is often thought of as precisely this probability, Lindley's paradox will show that this interpretation can be very much mistaken. A frequentist may counter by saying that he does not believe Bayesian statistics to be correct, thereby solving the paradox. Nevertheless, even as a frequentist, it would be good to know that the result of Bayesian statistics is approximately the same as the result of frequentist statistics in those cases where Bayesian statistics make sense even to a frequentist. However, Lindley's paradox shows that this is not the case, which should make a frequentist feel somewhat uncomfortable.

2.5.2 Lindley's paradox

Lindley's paradox is puzzling indeed, at least for those who confuse the p-value with the probability that the null hypothesis is true. The opening sentence of Lindley's article summarizes the paradox adequately [13, p.187]:

> An example is produced to show that, if H is a simple hypothesis and x the result of an experiment, the following two phenomena can occur simultaneously:
>
> (i) a significance test for H reveals that x is significant at, say, the 5% level
>
> (ii) the posterior probability of H, given x, is, for quite small prior probabilities of H, as high as 95%.

This is contrary to the interpretation of many people of the p-value (see Section 2.2).

Now, for the paradox: consider a random sample $x^n = x_1, \ldots, x_n$ from a normal distribution with unknown mean θ and known variance σ^2. The null hypothesis H_0 is: $\theta = \theta_0$. Let the prior probability that θ equals θ_0 be c. Suppose that the remainder of the prior probability is distributed uniformly over some interval I containing θ_0. By $\overline{x}$, we will denote the arithmetic mean of the observations and we will assume that it is well within the interval I. After noting that $\overline{x}$ is a minimal sufficient statistic for the mean of the normal distribution, we can now calculate the posterior probability that the null hypothesis is true, given the data:

$$
\begin{aligned}
P(H_0|\overline{x}) &= \frac{P(\overline{x}|H_0)P(H_0)}{P(\overline{x}|H_0)P(H_0) + P(\overline{x}|\theta \neq \theta_0)P(\theta \neq \theta_0)} \\
&= \frac{c \cdot \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} e^{-\frac{n}{2\sigma^2}(\overline{x}-\theta_0)^2}}{c \cdot \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} e^{-\frac{n}{2\sigma^2}(\overline{x}-\theta_0)^2} + (1-c)\int_{\theta \in I} \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} e^{-\frac{n}{2\sigma^2}(\overline{x}-\theta)^2} \cdot \frac{1}{|I|} d\theta} \\
&= \frac{c e^{-\frac{n}{2\sigma^2}(\overline{x}-\theta_0)^2}}{c e^{-\frac{n}{2\sigma^2}(\overline{x}-\theta_0)^2} + (1-c)\int_{\theta \in I} e^{-\frac{n}{2\sigma^2}(\overline{x}-\theta)^2} \cdot \frac{1}{|I|} d\theta}
\end{aligned}
$$

Now substitute $\overline{x} = \theta_0 + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$, where $z_{\alpha/2}$ is the value such that this will produce a sample mean that will yield $p = \alpha$:

$$
= \frac{c e^{-\frac{1}{2} z_{\alpha/2}^2}}{c e^{-\frac{1}{2} z_{\alpha/2}^2} + \frac{1-c}{|I|} \int_{\theta \in I} e^{-\frac{n}{2\sigma^2}(\overline{x}-\theta)^2} d\theta}
$$

Now use: $\int_{\theta \in I} e^{-\frac{n}{2\sigma^2}(\overline{x}-\theta)^2} d\theta \leq \int e^{-\frac{n}{2\sigma^2}(\overline{x}-\theta)^2} d\theta = \frac{\sqrt{2\pi}\sigma}{\sqrt{n}} \int \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} e^{-\frac{n}{2\sigma^2}(\overline{x}-\theta)^2} d\theta = \frac{\sqrt{2\pi}\sigma}{\sqrt{n}}$ to get:

$$
\geq \frac{c e^{-\frac{1}{2} z_{\alpha/2}^2}}{c e^{-\frac{1}{2} z_{\alpha/2}^2} + \frac{1-c}{|I|} \frac{\sqrt{2\pi}\sigma}{\sqrt{n}}}
$$

The paradox is apparent now. Because $\frac{1-c}{|I|} \frac{\sqrt{2\pi}\sigma}{\sqrt{n}}$ goes to zero as n goes to infinity, $P(H_0|\overline{x})$ goes to one as n goes to infinity. Thus, indeed a sample size n, dependent on c and α, can be produced such that if a significance test is significant at the $\alpha\%$ level, the posterior probability of the null hypothesis is 95%. Hence, a standard frequentist analysis will lead to an entirely different conclusion than a Bayesian analysis: the former will reject H_0 while the latter will see no reason to believe that H_0 is not true based on this sample. A plot of this lower bound for $P(H_0|\overline{x})$ for $c = \frac{1}{2}$, $\sigma = 1$ and $|I| = 1$ for various p-values can be found in Figure 2.

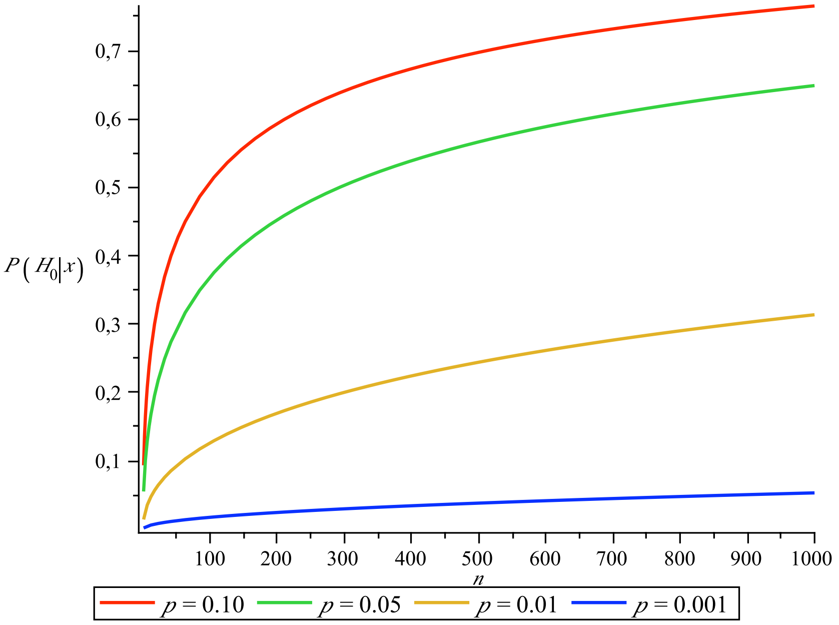


Figure 2: Lower bound on $P(H_0|\overline{x})$ for $c = \frac{1}{2}$, $\sigma = 1$ and $|I| = 1$ for various values of $z_{\alpha/2}$ (graph made in Maple 13 for Mac).

Why is there such a great discrepancy between the result of a classical analysis and a Bayesian analysis? Lindley himself noted that this is not an artefact of this particular prior: the phenomenon would persist with almost any prior that has a concentration on the null value and no concentrations elsewhere. Is this type of prior reasonable? Lindley thinks it is, because singling out the hypothesis $\theta = \theta_0$ is itself evidence that the value θ_0 is in some way special and is therefore likely to be true. Lindley gives several examples of this, one of which is about telepathy experiments where, if no telepathic powers are present, the experiment has a success ratio of $\theta = \frac{1}{5}$. This value is therefore fundamentally different from any value of θ that is not equal to $\frac{1}{5}$.

Now assume that the prior probability exists and has a concentration on the null value. Some more insight can be provided by rewriting the posterior probability as

$$P(H_0|\overline{x}) \geq \frac{cf_n}{cf_n + \frac{1-c}{|I|}}, \text{ where } f_n = \sqrt{\frac{n}{2\pi\sigma^2}} e^{-\frac{1}{2}z^2_{\alpha/2}}.$$

Naturally, $f_n \to \infty$ if $n \to \infty$. This therefore behaves quite differently than the p-value, which is the probability of the observed outcome and more extreme ones. In Lindley's words:

> *"In fact, the paradox arises because the significance level argument is based on the area under a curve and the Bayesian argument is based on the ordinate of the curve."*
> — D.V. Lindley (1957), [13, p.190].

There is more literature on the nature and validity of Lindley's paradox. Because they utilize advanced mathematical theories and do not come to a definite conclusion, they fall outside of the scope of this thesis.[8]

2.5.3 Irreconcilability of p-values and evidence for point null hypotheses

Lindley's paradox raises concerns mostly for large sample sizes. Berger and Sellke showed that p-values can give a very misleading impression as to the validity of the null hypothesis for *any* sample size and *any* prior on the alternative hypothesis [7].

Consider a random sample $x^n = x_1, \ldots, x_n$ having density $f(x^n|\theta)$. The null hypothesis H_0 is: $\theta = \theta_0$, the alternative hypothesis H_1 is: $\theta \neq \theta_0$. Let π_0 denote the prior probability of H_0 and $\pi_1 = 1 - \pi_0$ the prior probability of H_1. Suppose that the mass on H_1 is spread out according to the density $g(\theta)$. Then we can apply Bayes' theorem to get:

$$\begin{aligned} P(H_0|x^n) &= \frac{\pi_0 f(x^n|\theta_0)}{\pi_0 f(x^n|\theta_0) + (1-\pi_0)\int f(x^n|\theta)g(\theta)d\theta} \\ &= \left[1 + \frac{1-\pi_0}{\pi_0} \cdot \frac{\int f(x^n|\theta)g(\theta)d\theta}{f(x^n|\theta)}\right]^{-1}. \end{aligned}$$

The posterior odds ratio of H_0 to H_1 is:

$$\begin{aligned} \frac{P(H_0|x^n)}{P(H_1|x^n)} &= \frac{P(H_0|x^n)}{1 - P(H_0|x^n)} \\ &= \frac{\pi_0 f(x^n|\theta_0)}{\pi_0 f(x^n|\theta_0) + (1-\pi_0)\int f(x^n|\theta)g(\theta)d\theta} \cdot \frac{\pi_0 f(x^n|\theta_0) + (1-\pi_0)\int f(x^n|\theta)g(\theta)d\theta}{(1-\pi_0)\int f(x^n|\theta)g(\theta)d\theta} \\ &= \frac{\pi_0}{1-\pi_0} \cdot \frac{f(x^n|\theta_0)}{\int f(x^n|\theta)g(\theta)d\theta}. \end{aligned}$$

The second fraction is also known as the *Bayes factor* and will be denoted by B_g, where the g corresponds to the prior $g(\theta)$ on H_1:

$$B_g(x^n) = \frac{f(x^n|\theta_0)}{\int f(x^n|\theta)g(\theta)d\theta}.$$

[8] See for example: Shafer, G. (1982), Lindley's paradox, *Journal of the American Statistical Association*, **77** (378), 325-334, who thinks that "The Bayesian analysis seems to interpret the diffuse prior as a representation of strong prior evidence, and this may be questionable". He shows this using the theory of belief functions. See also Tsao, C.A. (2006), A note on Lindley's paradox, *Sociedad de Estadística e Investigación Operativa Test*, **15** (1), 125-139. Tsao questions the point null approximation assumption and cites additional literature discussing the paradox.

The Bayes factor is used very frequently, as it does not involve the prior probabilities of the hypotheses. It is often interpreted as the odds of the hypotheses implied by the data alone. This is of course not entirely correct, as the prior on H_1, g, is still involved. However, lower bounds on the Bayes factor over all possible priors can be considered to be 'objective'.

The misleading impression p-values can give about the validity of the null hypothesis will now be shown by means of an example.

Suppose that x^n is a random sample from a normal distribution with unknown mean θ and known variance σ^2. Let $g(\theta)$ be a normal distribution with mean θ_0 and variance σ^2. As in Lindley's paradox, we will use the sufficient statistic $\overline{x}$. Then:

$$\begin{aligned}\int f(\overline{x}|\theta)g(\theta)d\theta &= \int \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} e^{-\frac{n}{2\sigma^2}(\overline{x}-\theta)^2} \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{1}{2\sigma^2}(\theta-\theta_0)^2} d\theta \\ &= \frac{\sqrt{n}}{2\pi\sigma^2} \int e^{-\frac{1}{2\sigma^2}(n(\overline{x}-\theta)^2+(\theta-\theta_0)^2)} d\theta\end{aligned}$$

Because $n(\overline{x}-\theta)^2 + (\theta-\theta_0)^2 = (n+1)\left(\theta - \frac{\overline{x}n+\theta_0}{n+1}\right)^2 + \frac{1}{1+\frac{1}{n}}(\overline{x}-\theta_0)^2$, this is equal to:

$$\begin{aligned}&= \frac{1}{\sqrt{2\pi}\sigma\sqrt{1+\frac{1}{n}}} e^{-\frac{1}{2\sigma^2(1+\frac{1}{n})}(\overline{x}-\theta_0)^2} \cdot \int \frac{\sqrt{n+1}}{\sqrt{2\pi}\sigma} e^{-\frac{n+1}{2\sigma^2}\left(\theta-\frac{\overline{x}n+\theta_0}{n+1}\right)^2} d\theta \\ &= \frac{1}{\sqrt{2\pi}\sigma\sqrt{1+\frac{1}{n}}} e^{-\frac{1}{2\sigma^2(1+\frac{1}{n})}(\overline{x}-\theta_0)^2}.\end{aligned}$$

Thus, the Bayes factor is equal to:

$$B_g(\overline{x}) = \frac{\frac{\sqrt{n}}{\sqrt{2\pi}\sigma} e^{-\frac{n}{2\sigma^2}(\overline{x}-\theta_0)^2}}{\frac{1}{\sqrt{2\pi}\sigma\sqrt{1+\frac{1}{n}}} e^{-\frac{1}{2\sigma^2(1+\frac{1}{n})}(\overline{x}-\theta_0)^2}}$$

Now substitute $z = \frac{\sqrt{n}|\overline{x}-\theta_0|}{\sigma}$:

$$= \sqrt{1+n} \cdot \frac{e^{-\frac{1}{2}z^2}}{e^{-\frac{z^2}{2(n+1)}}} = \sqrt{1+n} \cdot e^{-\frac{n}{2(n+1)}z^2}.$$

The posterior probability of H_0 is therefore:

$$\begin{aligned}P(H_0|\overline{x}) &= \left[1 + \frac{1-\pi_0}{\pi_0} \cdot \frac{1}{B_g(\overline{x})}\right]^{-1} \\ &= \left[1 + \frac{1-\pi_0}{\pi_0} \cdot \frac{1}{\sqrt{1+n}} \cdot e^{\frac{n}{2(n+1)}z^2}\right]^{-1}\end{aligned}$$

Lindley's paradox is also apparent from this equation: for fixed z (for example $z = 1.96$, corresponding to $p = 0.05$), $P(H_0|\overline{x})$ will go to one if n goes to infinity. However, what the authors wished to show with this equation is that there is a great discrepancy between the p-value and the probability that H_0 is true, even for small n. This can easily be seen from the plot of the results for $\pi_0 = \frac{1}{2}$, for values of z corresponding to $p = 0.10, p = 0.05, p = 0.01$ and $p = 0.001$ in Figure 3.

This is not an artefact of this particular prior. As will be shown next, we can derive a lower bound on $P(H_0|x)$ for *any* prior. First, some notation. Let G_A be the set of all distributions, $\underline{P}(H_0|x, G_A) = \inf_{g\in G_A} P(H_0|x)$ and $\underline{B}(x, G_A) = \inf_{g\in G_A} B_g(x)$. If a maximum likelihood estimate of θ, denoted by $\hat{\theta}(x)$, exists for the observed x, then this is the parameter most favored by the data. Concentrating the density under the alternative hypothesis on $\hat{\theta}(x)$ will result in the smallest possible Bayesfactor [1, p.228], [7, p.116]. Thus:

$$\underline{B}(x, G_A) = \frac{f(x|\theta_0)}{f(x|\hat{\theta}(x))} \quad \text{and hence,} \quad \underline{P}(H_0|x, G_A) = \left[1 + \frac{1-\pi_0}{\pi_0} \cdot \frac{1}{\underline{B}(x, G_A)}\right]^{-1}.$$

Let us continue with the example. For the normal distribution, the maximum likelihood estimate of the mean is $\hat{\theta} = \overline{x}$. Hence:

$$\underline{B}(x, G_A) = \frac{\frac{\sqrt{n}}{\sqrt{2\pi}\sigma} e^{-\frac{n}{2\sigma^2}(\overline{x}-\theta_0)^2}}{\frac{\sqrt{n}}{\sqrt{2\pi}\sigma} e^{-\frac{n}{2\sigma^2}(\overline{x}-\overline{x})^2}} = e^{-\frac{n}{2\sigma^2}(\overline{x}-\theta_0)^2} = e^{-\frac{1}{2}z^2}.$$

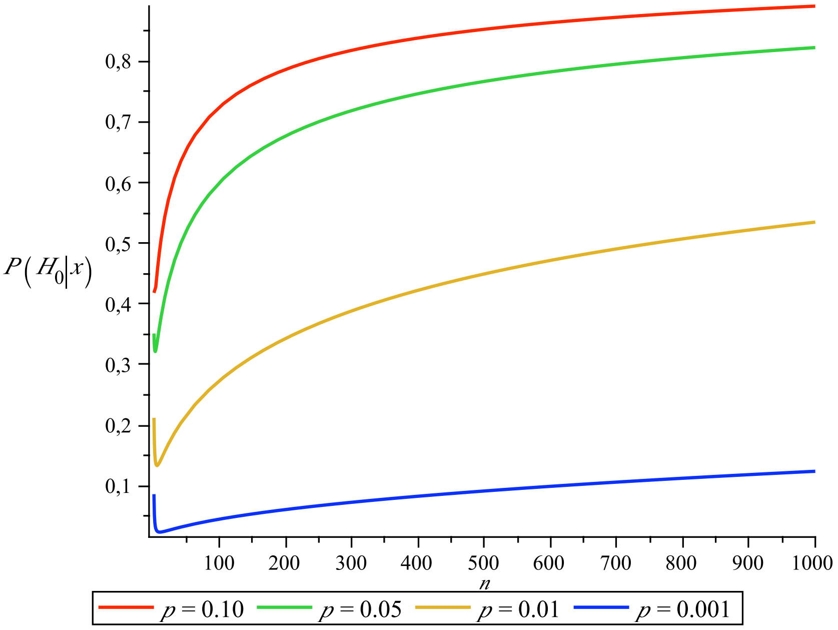


Figure 3: $P(H_0|x)$ for $\pi_0 = \frac{1}{2}$ and fixed z (graph made in Maple 13 for Mac).

And thus:

$$\underline{P}(H_0|x, G_A) = \left[1 + \frac{1-\pi_0}{\pi_0} \cdot e^{\frac{1}{2}z^2}\right]^{-1}$$

Again setting $\pi_0 = \frac{1}{2}$, Table 4 shows the two-sided p-values and corresponding lower bounds on $P(H_0|x, G_A)$ for this example [7, p.116]. The lower bounds on $P(H_0|x, G_A)$ are considerably larger than the corresponding p-values, casting doubt on the premise that small p-values constitute evidence against the null hypothesis.

Table 4: Comparison of p-values and $\underline{P}(H_0|x, G_A)$ for $\pi_0 = \frac{1}{2}$

z	p-value	$\underline{P}(H_0\|x, G_A)$
1.645	0.10	0.205
1.960	0.05	0.128
2.576	0.01	0.035
3.291	0.001	0.0044

2.5.4 Discussion

As we saw, the magnitude of the p-value and the magnitude of the evidence against the null hypothesis can differ greatly. Why is this? The main reason is the conditioning that occurs: $P(H_0|x)$ only depends on the data x that are observed, while to calculate a p-value, one replaces x by the knowledge that x is in $A := \{y : T(y) \geq T(x)\}$ for some test statistic T. There is an important difference between x and A, which can be illustrated with a simple example [7, p.114]. Suppose X is measured by a weighing scale that occasionally "sticks". When the scale sticks, a light is flashed. When the scale sticks at 100, one only knows that the true x was larger than 100. If large X cast doubt on the null hypothesis, then the occurence of a stick at 100 constitutes greater evidence that the null hypothesis is false than a true reading of $x = 100$. Therefore, it is not very surprising that the use of A in frequentist calculations overestimates the evidence against the null hypothesis.

These results also shed some light on the validity of the p *postulate*. The p postulate states that identical p-values should provide identical evidence against the null hypothesis [9, p.787]. Lindley's paradox casts great doubt on the p postulate, as it shows that the amount of evidence for the null hypothesis depends on the sample size. The same phenomenon can be observed in Figure 3. However, studies have found that psychologists were more willing to reject a null hypothesis when the sample size increased, with the p-value held constant [9, p.787].

Even a non-Bayesian analysis suggests that the p postulate is invalid. For this, consider a trial in which patients receive two treatments, A and B. They are then asked which treatment they prefer. The null hypothesis is that there is no preference. Using the number of patients who prefer treatment A as the test statistic T, the probability of k patients out of n preferring treatment A is equal to:

$$P(T(x) = k) = \binom{n}{k}\left(\frac{1}{2}\right)^n .$$

And the two-sided p-value is:

$$p = \sum_{j=0}^{n-k}\binom{n}{j}\left(\frac{1}{2}\right)^n + \sum_{j=k}^{n}\binom{n}{j}\left(\frac{1}{2}\right)^n .$$

Now consider the data in Table 5.[9]

Table 5: Four theoretical studies.

n	numbers preferring A:B	% preferring A	two-sided p-value
20	15:5	75.00	0.04
200	115:85	57.50	0.04
2000	1046:954	52.30	0.04
2000000	1001445:998555	50.07	0.04

Even though the p-value is the same for all studies, a regulatory authority would probability not treat all studies the same. The study with $n = 20$ would probably be considered inconclusive due to a small sample size, while the study with $n = 2000000$ would be considered to provide almost conclusive evidence that there is no difference between the two treatments. These theoretical studies therefore suggest that the interpretation of a p-value depends on sample size, which implies that the p postulate is false.

2.6 Optional stopping

Suppose a researcher is convinced that a new treatment is significantly better than a placebo. In order to convince his colleagues of this, he sets up an experiment to test this hypothesis. He decides to not fix a sample size in advance, but to continue collecting data until he obtains a result that would be significant if the sample size had been fixed in advance. However, unbeknownst to the researcher, the treatment is actually not better than the placebo. Will the researcher succeed in rejecting the null hypothesis, even though it is true? The answer is yes, if he has enough funds and patience, he certainly will.

Suppose that we have data $x^n = x_1, x_2, \ldots, x_n$ that are normally distributed with unknown mean θ and known standard deviation $\sigma = 1$. The null hypothesis is: $\theta = 0$. The test statistic is then $Z_n = \overline{x}\sqrt{n}$. When the null hypothesis is true, Z_n follows a standard normal distribution. If n is fixed in advance, the two-sided p-value is $p = 2(1 - \Phi(|Z_n|))$. In order to obtain a significant result, the researcher must find $|Z_n| > k$ for some k. His stopping rule will then be: "Continue testing until $|Z_n| > k$, then stop." In order to show that this strategy will always be succesful, we need the law of

[9] Data slightly adapted from Freeman, [6, p.1446], because his assumptions were not entirely clear and his results do not completely match the values I calculated myself based on a binomial distribution.

the iterated logarithm:

Theorem 2.2 (Law of the iterated logarithm) If $x_1, x_2, \ldots$ are *iid* with mean equal to zero and variance equal to one, then

$$\limsup_{n \to \infty} \frac{\sum_{i=1}^{n} x_i}{\sqrt{2n \log \log n}} = 1 \text{ almost surely.}$$

This means that for $\lambda < 1$, the inequality:

$$\sum_{i=1}^{n} x_i > \lambda \sqrt{2n \log \log n}$$

holds with probability one for infinitely many n.

The proof will be omitted, but can be found in Feller [14, p.192]. This theorem tells us that with probability one, because $\sum_{i=1}^{n} x_i = \sqrt{n} Z_n$, the inequality

$$Z_n > \lambda \sqrt{2 \log \log n},$$

for $\lambda < 1$, will hold for infintely many n. Therefore, there is a value of n such that $Z_n > k$ and therefore, such that the experiment will stop while yielding a significant result.

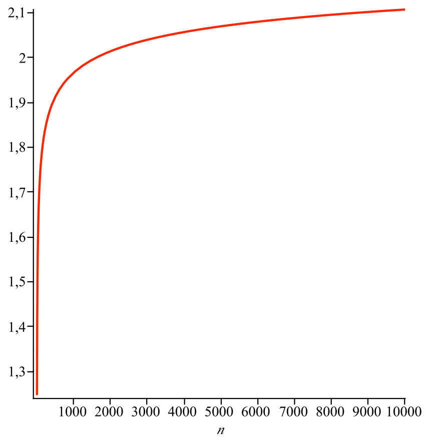


Figure 4: Graph of $\sqrt{2 \log \log n}$ (graph made in Maple 13 for Mac).

This procedure is also known as 'sampling to a foregone conclusion' and generally frowned upon. Is it always cheating? Maybe it was in the previous example, but consider a researcher who designs an experiment on inhibitory control in children with ADHD and decides in advance to test 40 children with ADHD and 40 control children [9, p.785]. After 20 children in each group have been tested, she examines the data and the results demonstrate convincingly what the researcher hoped to demonstrate. However, the researcher cannot stop the experiment now, because then she would be guilty of optional stopping. Therefore, she has to continue spending time and money to complete the experiment. Or alternatively, after testing 20 children in each group, she was forced to stop the experiment because of a lack of funding. Even though she found a significant result, she would not be able to publish her findings, once again because she would be guilty of optional stopping. It seems undesirable that results can become useless because of a lack of money, time or patience.

2.7 Arguments in defense of the use of p-values

Considering all problems with p-values raised in the previous sections, one could wonder why p-values are still used. Schmidt and Hunter surveyed the literature to identify the eight most common arguments [15]. For each of them, they cite examples of the objection found in literature. The objections are:

1. Without significance tests, we cannot know whether a finding is real or just due to chance.
2. Without significance tests, hypothesis testing would be impossible.
3. The problems do not stem from a reliance on significance testing, but on a failure to develop a tradition of replication of findings.
4. Significance testing has an important practical value: it makes interpreting research findings easier by allowing one to eliminate all findings that are not statistically significant from further consideration.
5. Confidence intervals are themselves significance tests.
6. Significance testing ensures objectivity in the interpretation of research data.
7. Not the use, but the misuse of significance testing is the problem.
8. Trying to reform data analysis methods is futile and errors in data interpretation will eventually be corrected as more research is conducted on a given question.

Schmidt and Hunter argue that all these objections are false. For some objections, this is obvious, for some maybe less clearly so. Be that as it may, it is remarkable that none of these objections to abandoning p-value based hypothesis testing contradict the claims made in previous sections, but rather direct the attention to practical problems or point to the misunderstandings of others as the root of the problem.

Nickerson has contributed to the discussion by means of a comprehensive review of the arguments for and against null hypothesis significance testing [16]. He agrees with the critique, but tries to mitigate it by describing circumstances in which the problems are not so bad.

For example, when discussing Lindley's paradox, he produces a table based on the equation

$$P(H_0|D) = \frac{P(D|H_0)P(H_0)}{P(D|H_0)P(H_0) + P(D|H_A)P(H_A)}.$$

Assume $P(H_0) = P(H_A) = \frac{1}{2}$. Then we can calculate $P(H_0|D)$ for various values of $P(D|H_0)$ and $P(D|H_A)$. Nickersons results are in Figure 5.

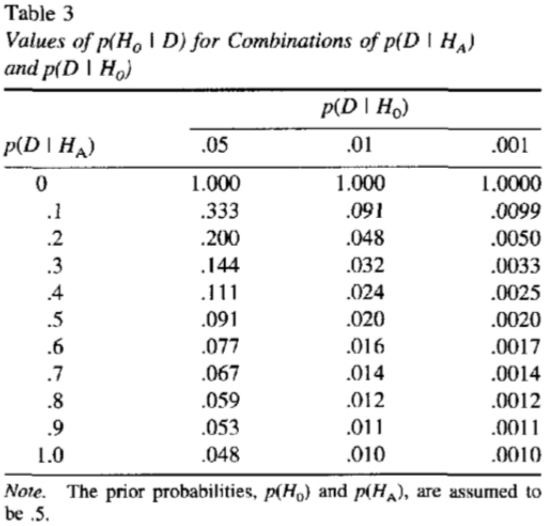


Table 3

Values of $p(H_0 \mid D)$ for Combinations of $p(D \mid H_A)$ and $p(D \mid H_0)$

	$p(D \mid H_0)$		
$p(D \mid H_A)$	.05	.01	.001
0	1.000	1.000	1.0000
.1	.333	.091	.0099
.2	.200	.048	.0050
.3	.144	.032	.0033
.4	.111	.024	.0025
.5	.091	.020	.0020
.6	.077	.016	.0017
.7	.067	.014	.0014
.8	.059	.012	.0012
.9	.053	.011	.0011
1.0	.048	.010	.0010

Note. The prior probabilities, $p(H_0)$ and $p(H_A)$, are assumed to be .5.

Figure 5: Values of $P(H_0|D)$ for various values of $P(D|H_0)$ and $P(D|H_A)$ as calculated by Nickerson.

Nickerson notes that if $P(D|H_A)$ is larger, $P(D|H_0)$ is a better approximation for $P(H_0|D)$ and that if $P(D|H_A) = 0.5$, $P(H_0|D)$ is about twice the value of $P(D|H_0)$. Furthermore, a $P(D|H_0)$ of 0.01 or 0.001 represents fairly strong evidence against the null, even for relatively small values of $P(D|H_A)$. From the data in the table, he concludes:

> *"In short, although $P(D|H_0)$ is not equivalent to $P(H_0|D)$, if one can assume that $P(H_A)$ is at least as large as $P(H_0)$ and that $P(D|H_A)$ is much larger than $P(D|H_0)$, then a small value of p, that is, a small value of $P(D|H_0)$, can be taken as a proxy for a relatively small value of $P(H_0|D)$. There are undoubtedly exceptions, but the above assumptions seem appropriate to apply in many, if not most, cases."*
> — R.S. Nickerson (2000), [16, p.252].

It does seem reasonable to reject H_0 in favor of H_A if $P(D|H_A)$ is much larger than $P(D|H_0)$ (even though this is not always the correct action, see for example Section 3.2.4), or equivalently, if the Bayes factor is smaller than one. However, it is unclear if these are the circumstances under which most research is performed, though Nickerson seems to think so.

Nickerson also comments on Berger and Sellke (see Section 2.5.3) by stating that with all the distributions for D considered by them, $P(H_0|D)$ varies monotonically by p, which means that the smaller the value of p, the stronger the evidence against H_0 and for H_A is. Therefore, it might be reasonable to consider a small value of p as evidence against H_0, but not as as strong evidence as it was considered previously. This would even support a criticized move by former editor Melton of the *Journal of Experimental Psychology*, who refused to publish results with $p < 0.05$ because he considered 0.05 to be too high a cutoff. Although simply lowering the threshold for p-values would not solve the other problems that were discussed, it is interesting to compare Nickersons conclusion with the discussion in Section 3.3.5.

Regarding the dependence of p-values on both the sampling plan and sample size (including optional stopping), Nickerson points out that there are sequential stopping rules for NHST, whereby it is not necessary to decide on a sample size in advance. These stopping rules are, however, controversial [9, p.785-787].

As to the flawed syllogistic logic, as discussed in Section 2.2, Nickerson counters by saying that the logic is approximately correct *"when the truth of the first premise's antecedent is positively related to the truth of its consequent."* So if the consequent of the conditional premise is very likely to be true, independent of the truth or falsity of the antecedent, the form has little force. An example of this is: "If X, then that person is probably not the queen of the Netherlands." This is likely to be true, almost independent of X. In such cases, syllogistic reasoning should not be applied, but in cases like "If the experimental manipulation did not work, then p would probably be greater than 0.05", it can be [16, p.268].

Lastly, to the objection about the arbitrariness of the value 0.05 (not discussed in previous sections, but also sometimes raised by critics), Nickerson replies that freedom to select your own significance level would add an undesirable element of subjectivity to research. Nickerson discusses many more points in his article. Most of them can be summarized as objection 7 of Schmidt and Hunter: misapplication is the main problem, which recalls Seneca's statement *"A sword does not kill anyone, it is merely the tool of a killer."*[10] However, even though a correct and well thought-out application of the hypothesis testing principle would improve matters, it does not solve all problems.

[10] *gladius neminem occidit: occidentis telum est*, Lucius Annaeus Seneca, *Epistulae morales ad Lucilium* 87.30.

3 An alternative hypothesis test

3.1 Introduction

P. Grünwald suggested a new hypothesis test, with an interesting property regarding optional stopping. For this test, we need to define *probabilistic sources*. In this definition, we'll use the following notation: for a given sample space $\mathcal{X}$, define $\mathcal{X}^+ = \bigcup_{n \geq 1} \mathcal{X}^n$, the set of all possible samples of each length. Define $\mathcal{X}^0 = \{x^0\}$, where x^0 is a special sequence called the empty sample. After defining $\mathcal{X}^* = \mathcal{X}^+ \cup \mathcal{X}^0$, we can define [17, p.53]:

Definition 3.1 (probabilistic source) Let $\mathcal{X}$ be a sample space. A *probabilistic source* with outcomes in $\mathcal{X}$ is a function $P : \mathcal{X}^* \to [0, \infty)$ such that for all $n \geq 0$, all $x^n \in \mathcal{X}^n$ it holds that:

1. $\sum_{z \in \mathcal{X}} P(x^n, z) = P(x^n)$.
2. $P(x^0) = 1$.

These two conditions say that the event that data (x^n, z) arrives is identical to the event that data x^n arrives first and data z arrives afterward. They ensure that $\sum_{x^n \in \mathcal{X}^n} P(x^n) = 1$ for all n. For continuous $\mathcal{X}$, we simply replace the sum in condition 1 by an integral. Intuitively, a probabilistic source is a probability distribution represented by a probability mass function defined over arbitrarily long samples.

Example (Bernoulli, probabilistic source)
Let $\mathcal{X} = \{0, 1\}$ and define $P_\theta : \mathcal{X}^* \to [0, \infty)$ by $P_\theta(x^n) = \theta^{n_1}(1-\theta)^{n-n_1}$, with n_1 the number of ones observed. Let x^n be a sequence of outcomes with n_1 ones, then:

$$\sum_{z \in \mathcal{X}} P(x^n, z) = P(x^n, z = 0) + P(x^n, z = 1) = \theta^{n_1}(1-\theta)^{n-n_1+1} + \theta^{n_1+1}(1-\theta)^{n-n_1}$$
$$= (1 - \theta + \theta)\theta^{n_1}(1-\theta)^{n-n_1} = P(x^n).$$

We can define $P(x^0) = 1$ and the consequence of this condition is well-known to hold:

$$\sum_{z \in \mathcal{X}} P(x^0, z) = \sum_{z \in \mathcal{X}} P(z) = 1.$$

Thus, P_θ is a probabilistic source.

3.1.1 Definition of the test

With this knowledge, we can define the alternative hypothesis test. Let $x^n = x_1, \ldots, x_n$ be a sequence of n outcomes from a sample space $\mathcal{X}$. Choose a fixed value $\alpha^* \in (0, 1]$. Then the test T^* for two point hypotheses P (the null hypothesis) and Q, with P and Q probabilistic sources, is:

$$\begin{cases} \text{If } P(x^n) = 0, \text{reject } P \text{ with type I error probability equal to zero.} \\ \text{If } Q(x^n) = 0, \text{accept } P \text{ with type II error probability equal to zero.} \\ \text{If } \dfrac{P(x^n)}{Q(x^n)} < \alpha^*, \text{reject } P \text{ with type I error probability less than } \alpha^*. \end{cases}$$

This test can be interpreted as a standard frequentist hypothesis test with significance level α^*. As will be proven in Section 3.1.2, the type I error is really strictly smaller than α^*. For the Bernoulli model, this test would take the following form:

Example (Bernoulli)
Suppose we want to test the hypothesis $\theta = \theta_0$ against $\theta = \theta_A$ with significance level $\alpha^* = 0.05$. After observing $x^n = x_1, \ldots, x_n$, n outcomes from $\mathcal{X} = \{0, 1\}$, with $n_1 = \sum_{i=1}^n x_i$ and $n_0 = n - n_1$ the test will take the form: reject if

$$\frac{P(x^n|\theta_0)}{P(x^n|\theta_A)} = \frac{\theta_0^{n_1}(1-\theta_0)^{n_0}}{\theta_A^{n_1}(1-\theta_A)^{n_0}} = \left(\frac{\theta_0}{\theta_A}\right)^{n_1} \cdot \left(\frac{1-\theta_0}{1-\theta_A}\right)^{n_0} < 0.05.$$

3.1.2 Derivation

That the type I error probability of the test T^* is less than α^* can be derived by means of Markov's inequality.

Theorem 3.2 (Markov's inequality) Let Y be a random variable with $P(Y \geq 0) = 1$ for which $E(Y)$ exists and let $c > 0$, then: $P(Y \geq c) \leq \frac{E[Y]}{c}$.

Proof of Theorem 3.2
This proof is for the discrete case, but the continuous case is entirely analogous.

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

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

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

□

Corollary 3.3 In Markov's inequality, equality holds if and only if $P(Y = c) + P(Y = 0) = 1$.

Proof of Corollary 3.3
As is apparent from the proof, equality holds if and only if

1. $E(Y) = \sum_{y \geq c} yP(y)$ and
2. $\sum_{y \geq c} yP(y) = \sum_{y \geq c} cP(y)$.

Condition 1 is met if and only if $\sum_{y<c} yP(y) = 0$, which is exactly the case if $P(Y < c) = 0$ or if $P(Y = 0|Y < c) = 1$.
Condition 2 is met if and only if $P(Y > c) = 0$.
Combining these two conditions gives the stated result.

□

We can now prove the claim made in Section 3.1 about the type I error:

Theorem 3.4 Assume for all $x^n \in \mathcal{X}^n$, $Q(x^n) \neq 0, P(x^n) \neq 0$ and choose $\alpha^* \in (0, 1]$. The type I error probability of the test T^* is strictly less than α^*.

Proof of Theorem 3.4
This can be proven by applying Markov's inequality to $Y = \frac{Q(x^n)}{P(x^n)}$. Note that Y cannot take on negative values, because both Q and P can only take on values between zero and one. Then:

$$P\left(\frac{P(x^n)}{Q(x^n)} \leq \alpha^*\right) = P\left(\frac{Q(x^n)}{P(x^n)} \geq \frac{1}{\alpha^*}\right) \overset{(\star)}{<} \frac{E\left[\frac{Q(x^n)}{P(x^n)}\right]}{\frac{1}{\alpha^*}} = \alpha^* \sum_{x^n} P(x^n) \cdot \frac{Q(x^n)}{P(x^n)} = \alpha^* \sum_{x^n} Q(x^n) = \alpha^*.$$

($\star$) The inequality is strict, because as $Q(x^n) \neq 0$, $P(Y = 0) = 0$. Is it possible that $P(Y = \frac{1}{\alpha^*}) = 1$? If $P(Y = \frac{1}{\alpha^*}) = 1$, we can never reject, in which case the test would not be useful. However, this will not happen if P and Q are not defective: if $P \neq Q$, then there must be a x^n such that $P(x^n) > Q(x^n)$. For this x^n it holds that $\frac{Q(x^n)}{P(x^n)} \neq \frac{1}{\alpha^*}$, because $\alpha^* \leq 1$ and $\frac{Q(x^n)}{P(x^n)} < 1$. It also holds that $P(x^n) > 0$ and therefore: $P\left(Y \neq \frac{1}{\alpha^*}\right) > 0$. Therefore, by Corollary 3.3, the inequality is strict.

Because $P\left(\frac{P(x^n)}{Q(x^n)} \leq \alpha^*\right) \geq P\left(\frac{P(x^n)}{Q(x^n)} < \alpha^*\right)$, we can conclude that $P\left(\frac{P(x^n)}{Q(x^n)} < \alpha^*\right) < \alpha^*$. Therefore, the probability under the null hypothesis P that the null hypothesis is rejected is less than α^*. This probability is the probability of making a type I error.

□

3.1.3 Comparison with a Neyman-Pearson test

This test seems to be very similar to the test for simple hypotheses that is most powerful in the Neyman-Pearson paradigm, since this test also takes the form of a likelihood ratio test. This can be seen from the Neyman-Pearson lemma [10]:

Lemma 3.5 (Neyman-Pearson lemma) Suppose that H_0 and H_1 are simple hypotheses and that the test that rejects H_0 whenever the likelihood ratio is less than c has significance level α. Then any other test for which the significance level is less than or equal to α has power less than or equal to that of the likelihood ratio test.

How does the test T^* differ from a Neyman-Pearson test? The difference is how the value at which the test rejects is chosen. For the Neyman-Pearson test, the value c is calculated using the prescribed type I error α and n. For example: suppose we have a random sample $x_1, \ldots, x_n$ from a normal distribution having known variance σ^2 and unknown mean θ. With $H_0 : \theta = \theta_0$, $H_A : \theta = \theta_A$ and $\theta_0 < \theta_A$, we can calculate the likelihood ratio:

$$\Lambda(x^n) = \frac{P(x^n|H_0)}{P(x^n|H_A)} = e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n (x_i-\theta_0)^2 - \sum_{i=1}^n (x_i-\theta_A)^2\right)}.$$

Small values of this likelihood ratio correspond to small values of $\sum_{i=1}^n (x_i - \theta_A)^2 - \sum_{i=1}^n (x_i - \theta_0)^2$. This expression reduces to $2n\overline{x}(\theta_0 - \theta_A) + n(\theta_A^2 - \theta_0^2)$. Because $\theta_0 < \theta_A$, the likelihood ratio is small if $\overline{x}$ is large. By the Neyman-Pearson lemma, the most powerful test rejects for $\overline{x} > x_0$ for some x_0. x_0 is chosen so that $P(\overline{x} > x_0|H_0) = \alpha/2$ and this leads to the well-known test of rejecting H_0 if $\overline{x} > \theta_0 + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$. This can be substituted in the likelihood ratio to find the value c such that if $\frac{P(x^n|H_0)}{P(x^n|H_A)} < c$, the test will reject. As is clear from this discussion, c depends on both α and n. The difference between a Neyman-Pearson test and the test T^* is clear now: if we want the test T^* to be significant at the level α, we simply reject when the likelihood ratio is smaller than α. The value at which the test T^* rejects is therefore independent of n, while this is not the case for the Neyman-Pearson test.

3.2 Comparison with Fisherian hypothesis testing

3.2.1 Interpretation

Is this test susceptible to the same misinterpretation as the p-value test? The value α^* is similar to the type I error of the Neyman-Pearson paradigm: α^* is a bound on the probability that the null hypothesis will be rejected even though it is true. Therefore, it seems likely that this test will be misunderstood in the same way as the type I error rate α from the Neyman-Pearson test is. As discussed in Section 1.4, many researchers follow Neyman and Pearson methodologically, but Fisher philosophically. As p-values and α are already regularly confused, this might also be a problem for this test. However, the problem seems to be mostly that the p-value is interpreted as if it were a type I error rate. Therefore, a more interesting question in this setting is whether type I error rates are misinterpreted by themselves. Nickerson confirms that this happens and lists several common misinterpretations [16].

The first one is very similar to the main misinterpretation of p-values: many people interpret it as the probability that a type I error has been made, after the null hypothesis has been rejected. Thus α is interpreted as the probability that H_0 is true even though it has been rejected, which can be represented as $P(H_0|R)$. It is, however, the probability that H_0 is rejected even though it is true, or $P(R|H_0)$.

Another easily made mistake is illustrated by the quote

> *"In a directional empirical prediction we can say that 1 or 5% of the time (as we choose) we will be wrong in rejecting the null hypothesis on the basis of such data as these."*

This is not correct, as this not only depends on the probability that the null hypothesis is rejected when it is true, but it also depends on how often the null hypothesis is rejected when it is in fact not true. The researcher can only be wrong in rejecting the null hypothesis in the former situation, so α is an upper bound for the probability of a type I error. Only if all null hypotheses are true would the author of the quote be right.

Nickerson describes some more misconceptions, but these seem to be the most prevalent ones. I see no reason to believe that the test T^* will not be plagued by the same misunderstandings.

3.2.2 Dependence on data that were never observed

The test T^* only depends on the observed data x^n, the probability distributions under two specified hypotheses and α^*. The value of α^* is chosen independent of the data. Data that are 'more extreme' than the observed data do not play any role in this test. Therefore, data that were never observed do not influence the outcome of this test. This can be illustrated by means of the two examples from Section 2.3.

First, the testing of the two hypotheses H_0 and H_0', where X has the distribution given in Table 3. In Section 2.3, we were not interested in testing the hypotheses against each other, but in the fact that one hypothesis would be rejected and the other would not after observing data that had the same probability of occuring under both hypotheses. With the test T^* we cannot consider a hypothesis separately, but it is insightful to test the two hypotheses against each other and reflect on the outcome: does it seem to be intuitively correct? To perform the test, we fix a value $\alpha^* \in (0,1]$ and observe $x = 2$. Then:

$$\frac{P_{H_0}(x)}{P_{H_0'}(x)} = \frac{P_{H_0'}(x)}{P_{H_0}(x)} = \frac{0.04}{0.04} = 1 \not< \alpha^*.$$

Therefore, neither H_0 nor H_0' will be rejected in favor of the other on the basis of this test, which seems to be intuitively correct behaviour.

Secondly, the censoring problem. Because the data were not actually censored and because this test does not take into account what could happen if the experiment would be repeated, the censoring does not have any effect on the form of the test: α^* would not be chosen differently and the probabilities of the data under the hypotheses do not change. Repeat tests might give different results, but only the current data are taken into consideration while performing the test T^*. The type I error α^* does not depend on repeat sampling, as is clear from the proof of Theorem 3.4. If the data were actually censored the test would, of course, have to be adjusted, because it is not directly clear what $P(x)$ should be if we know x to be censored.

3.2.3 Dependence on possibly unknown subjective intentions

In what ways could this test depend on subjective intentions? α^* clearly does not depend on subjective intentions, because it is picked in advance. How about the hypotheses P and Q then? Do they, for example, depend on the sampling plan? Let us see what happens when we apply the test to the example from Section 2.4.

For this test, we need to specify an alternative hypothesis. We will test $H_0 : \theta = \theta_0$ against $H_A : \theta = \theta_A$. After choosing α^* and obtaining the data $x =$ CCCCCWWCCCCW, the test will reject for the first researcher if :

$$\frac{P_{H_0}(x)}{P_{H_A}(x)} = \frac{\binom{12}{9}\theta_0^9(1-\theta_0)^3}{\binom{12}{9}\theta_A^9(1-\theta_A)^3} = \left(\frac{\theta_0}{\theta_A}\right)^9 \cdot \left(\frac{1-\theta_0}{1-\theta_A}\right)^3 < \alpha^*.$$

The test will reject for the second researcher if:

$$\frac{P_{H_0}(x)}{P_{H_A}(x)} = \frac{\binom{11}{2}\theta_0^9(1-\theta_0)^3}{\binom{11}{2}\theta_A^9(1-\theta_A)^3} = \left(\frac{\theta_0}{\theta_A}\right)^9 \cdot \left(\frac{1-\theta_0}{1-\theta_A}\right)^3 < \alpha^*.$$

So for the binomial en negative binomial distributions, both researchers would draw the same conclusions from the same data.

Is it just a fluke that this goes smoothly for this particular example? It is not. The reason for this is that the result of using a different sampling plan is that your observations are from a population of a different size, so the number of ways in which one can obtain a certain result changes. However, because this test uses a likelihood *ratio*, the factors resulting from the population size cancel out. The test T^* is therefore independent of the sampling plan.

3.2.4 Exaggeration of the evidence against the null hypothesis

Does rejection at a small value of α^* imply that the null hypothesis has a low probability of being true? As discussed in Section 1.3, in the Neyman-Pearson paradigm we cannot know whether the particular hypothesis we are testing is true or not, but only that if we repeat the test often, we will not make too many mistakes. That we will not make too many mistakes in the long run is true for this test as well, but we know a little more. We can see this using Bayes factors, introduced in Section 2.5.3. By using Bayes' theorem for both $P(H_0|x)$ and $P(H_A|x)$, we can write:

$$\frac{P(H_0|x)}{P(H_A|x)} = \frac{P(x|H_0)}{P(x|H_A)} \cdot \frac{P(H_0)}{P(H_A)}.$$

In this expression, $\frac{P(x|H_0)}{P(x|H_A)}$ is the Bayes factor. We can now restate the test T^* by saying:

> Reject the null hypothesis if the Bayes factor is less than α^*.

So the null hypothesis will be rejected if the observed data are less likely under the null hypothesis than under the alternative hypothesis. The test therefore seems to have an advantage over a p-value test, because a p-value test only takes into account whether the data are improbable under the null hypothesis or not. It does not, however, consider whether the data are even less likely under an alternative hypothesis.

The comparison with the probability of the data under the alternative hypothesis is an improvement over p-values, but examples can still be constructed where the null hypothesis is rejected even though it has a high probability of being true. A simple example of this phenomenon can be found in medical testing.

Example

In 2005, the U.S. Preventive Services Task Force reported that a large study of HIV testing in 752 U.S. laboratories found a sensitivity of 99.7% and a specificity of 98.5% for enzyme immunoassay. In the U.S., the adult HIV/AIDS prevalence rate is estimated to be 0.6%.[11] Suppose that a U.S. adult who is not at an elevated risk for HIV is tested for HIV by means of an enzyme immunoassay. The

[11] Sensitivity, specificity and the prevalence rate were retrieved from http://en.wikipedia.org/wiki/HIV_test#Accuracy_of_HIV_testing and http://en.wikipedia.org/wiki/List_of_countries_by_HIV/AIDS_adult_prevalence_rate on April 19th 2010.

null hypothesis H_0 is that that person is not infected with HIV, the alternative hypothesis H_A is that that person is infected with HIV. Let T^+ denote a positive test result, then we can derive from the data just stated:
$P(T^+|H_0) = 0.015$
$P(T^+|H_A) = 0.997$
$P(H_0) = 0.994$
$P(H_A) = 0.006$
Now suppose that the person has a positive test result. Applying the test T^* with significance level $\alpha^* = 0.05$, we get

$$\frac{P(T^+|H_0)}{P(T^+|H_A)} = \frac{0.015}{0.997} \approx 0.015 < \alpha^*.$$

Thus, the null hypothesis is rejected. However, by applying Bayes' theorem, we get:

$$\begin{aligned} P(H_0|T^+) &= \frac{P(T^+|H_0)P(H_0)}{P(T^+|H_0)P(H_0) + P(T^+|H_A)P(H_A)} \\ &= \frac{0.015 \cdot 0.994}{0.015 \cdot 0.994 + 0.997 \cdot 0.006} \approx 0.71. \end{aligned}$$

So there is an approximate 71% chance that the person is not infected with HIV, but the test T^* would reject the null hypothesis at a quite low significance level. Because the test T^* does not take the prior probability of the null hypothesis into account, the probability that the null hypothesis is true even though it is rejected by the test T^* can still be quite high. Therefore, if the prior probability of H_0 is much larger than the prior probability of H_A, then this test can also exaggerate the evidence against the null hypothesis. However, this will not happen often, as we have already proven that we will only falsely reject H_0 in at most $100\alpha\%$ of applications of the test.

3.2.5 Optional stopping

The test T^* has an interesting and attractive property regarding optional stopping, which we can prove using the next theorem:

Theorem 3.7 Assume for all $x^n \in \mathcal{X}^n$: $P(x^n) \neq 0, Q(x^n) \neq 0$ and choose $\alpha^* \in (0, 1]$. The probability under the null hypothesis P that there exists an $n' \in \{1, \dots, n\}$ such that $\frac{P(x^{n'})}{Q(x^{n'})} < \alpha^*$ is strictly smaller than α^*.

Proof of Theorem 3.7
For every $x^n = x_1, \dots, x_n$, define:

$$i_0 := \min_{1 \leq i \leq n} \left\{ i \,\middle|\, \frac{P(x^i)}{Q(x^i)} < \alpha^* \right\}.$$

If $\left\{ i \,\middle|\, \frac{P(x^i)}{Q(x^i)} < \alpha^* \right\}$ is empty, set $i_0 = n + 1$.
Construct a new probabilistic source Q' in the following manner:

$$Q'(x^i) := \begin{cases} Q(x^i) & 0 \leq i \leq i_0 \\ P(x_{i_0+1}, \dots, x_i | x^{i_0}) Q(x^{i_0}) & i_0 + 1 \leq i \leq n \end{cases}$$

Q' is a probabilistic source, because the two conditions from the definition hold:

1. For all $x^n \in X^n$, one of the following statements must be true:
   (a) $\left\{ i \,\middle|\, \frac{P(x^i)}{Q(x^i)} < \alpha^* \right\} = \emptyset$
   In that case: $Q'(x^n) = Q(x^n)$ for that x^n and because condition 1 holds for Q, it also holds for Q' for that x^n.

(b) $\left\{i \,\middle|\, \frac{P(x^i)}{Q(x^i)} < \alpha^*\right\} \neq \emptyset$
Then, there exists an $i_0 \in \{1, \dots, n\}$ for this x^n such that we can write:

$$\sum_{z \in X} Q'(x^n, z) = \sum_{z \in X} Q(x^{i_0}) P(x_{i_0+1}, \dots, x_n, z | x^{i_0})$$

Because $P(x^{i_0}) \neq 0$, this is equal to:

$$\begin{aligned} &= \sum_{z \in X} Q(x^{i_0}) P(x_{i_0+1}, \dots, x_n, z | x^{i_0}) \cdot \frac{P(x^{i_0})}{P(x^{i_0})} \\ &= \frac{Q(x^{i_0})}{P(x^{i_0})} \sum_{z \in X} P(x^n, z) \end{aligned}$$

And because P is a probabilistic source, this equals:

$$= \frac{Q(x^{i_0})}{P(x^{i_0})} P(x^n) = Q(x^{i_0}) P(x_{i_0+1}, \dots, x_n, z | x^{i_0}) = Q'(x^n).$$

2. By definition of Q' and because Q is a probabilistic source: $Q'(x^0) = Q(x^0) = 1$.

Thus, Q' is a probabilistic source. Now we can apply Markov's inequality to show that:

$$P\left(\frac{P(x^n)}{Q'(x^n)} < \alpha^*\right) < \alpha^*.$$

The next claim, combined with the last inequality, now proves the theorem:

<u>Claim:</u> $\frac{P(x^n)}{Q'(x^n)} < \alpha^*$ if and only if there exists an $i \in \{1, \dots, n\}$ such that $\frac{P(x^i)}{Q(x^i)} < \alpha^*$.

<u>Proof of Claim:</u>
If $\frac{P(x^n)}{Q'(x^n)} < \alpha^*$, there are two possibilities for Q':

1. $Q'(x^i) = Q(x^i)$ for all i. But then: $\frac{P(x^n)}{Q(x^n)} = \frac{P(x^n)}{Q'(x^n)} < \alpha^*$.

2. $Q'(x^i) = Q(x^i)$ for $i \leq i_0$, for some $i_0 \in \{1, \dots, n-1\}$. Then, by definition of Q', $\left\{i \,\middle|\, \frac{P(x^i)}{Q(x^i)} < \alpha^*\right\}$ is not empty.

Conversely, if there exists an $i \in \{1, \dots, n\}$ such that $\frac{P(x^i)}{Q(x^i)} < \alpha^*$, then $\left\{i \,\middle|\, \frac{P(x^i)}{Q(x^i)} < \alpha^*\right\}$ is not empty and therefore there exists an i_0 such that we can write:

$$\frac{P(x^n)}{Q'(x^n)} = \frac{P(x^{i_0}) P(x_{i_0+1}, \dots, x_n | x^{i_0})}{Q(x^{i_0}) P(x_{i_0+1}, \dots, x_n | x^{i_0})} = \frac{P(x^{i_0})}{Q(x^{i_0})} < \alpha^*.$$

□

Using this theorem, we can prove that the test T^* is not susceptible to optional stopping: using a stopping rule that depends on the data does not change the probability of obtaining a significant result. Theorem 3.7 showed that after performing n tests, the probability that the researcher will find an $n' \in \{1, \dots, n\}$ such that his result is significant even though the null hypothesis is true is less than α^*. The next theorem shows that even with unlimited resources, the probability to ever find a significant result is less than α^*.

Theorem 3.8 Assume for all $x^n \in \mathcal{X}^n$: $P(x^n) \neq 0, Q(x^n) \neq 0$ and choose $\alpha^* \in (0,1]$. The probability under the null hypothesis P that there exists an n such that $\frac{P(x^n)}{Q(x^n)} < \alpha^*$ is strictly smaller than α^*.

Proof of Theorem 3.8
Assume that $P\left(\exists n : \frac{P(x^n)}{Q(x^n)} < \alpha^*\right) = \alpha^* + \varepsilon$, $\varepsilon \geq 0$. Then the following holds for all $m \in \mathbb{N}_{>0}$:

$$P\left(\exists n : \frac{P(x^n)}{Q(x^n)} < \alpha^*\right) \leq P\left(\exists n \in \{1, \ldots, m\} : \frac{P(x^n)}{Q(x^n)} < \alpha^*\right) + P\left(\exists n > m : \frac{P(x^n)}{Q(x^n)} < \alpha^*\right).$$

Define $f, g : \mathbb{N}_{>0} \to \mathbb{R}_{\geq 0}$, given by:

$$f(m) = P\left(\exists n \in \{1, \ldots, m\} : \frac{P(x^n)}{Q(x^n)} < \alpha^*\right) \text{ and } g(m) = P\left(\exists n > m : \frac{P(x^n)}{Q(x^n)} < \alpha^*\right)$$

If $m^* > m$, then $f(m^*) \geq f(m)$ and $g(m^*) \leq g(m)$. Hence, f is monotonically increasing and g is monotonically decreasing in m and $f(m) + g(m) \geq \alpha^* + \varepsilon$ is constant for all m.
By Theorem 3.7, $f(m) < \alpha^*$ for all m. So f is monotonically increasing and bounded above by α^*, which implies, by the monotone convergence theorem, that $\lim_{m \to \infty} f(m) := l$ exists and that $l \leq \alpha^*$. Therefore, for all m_1, m_2 with $m_1 < m_2$:

$$P\left(\exists n \in \{m_1 + 1, \ldots, m_2\} : \frac{P(x^n)}{Q(x^n)} < \alpha^*\right) = f(m_2) - f(m_1) \leq l - f(m_1) \to 0$$

if $m_1, m_2 \to \infty$. Thus, $g(m) \downarrow 0$ and we already know that $f(m) < \alpha^*$ for all m. However, these two statements combined contradict that $f(m) + g(m) \geq \alpha^* + \varepsilon$ for all m. Therefore, the initial assumption $P\left(\exists n : \frac{P(x^n)}{Q(x^n)} < \alpha^*\right) \geq \alpha^*$ must be incorrect.

□

Hence, optional stopping will not be an issue with this test: the probability that a researcher will find a sample size at which he can reject is less than α^*.

3.3 Application to highly cited but eventually contradicted research

3.3.1 Introduction

In 2005, J.P.A. Ioannidis published a disconcerting study of highly cited (more than 1000 times) clinical research articles [19]. Out of 49 highly cited original clinical research studies, seven were contradicted by subsequent studies, seven had found effects that were stronger than those of subsequent studies, 20 were replicated, 11 remained unchallenged and four did not claim that the intervention was effective. This begs the question: *why* are so many of our research findings apparently incorrect? Ioannidis considered factors such as non-randomization, the use of surrogate markers, lack of consideration of effect size and publication bias. He did not mention, however, that the discrepancy might partly be explained by faulty statistical methods. When keeping in mind the previous sections, this does not seem to be a very unreasonable suggestion. This section is not an attempt to prove that this is the case, but an attempt to illustrate that it is a possibility. This is done by applying a variant of the test T^* to a selection of the articles studied by Ioannidis. It will be interesting to see whether the results are still considered to be significant by this test and it will provide a nice example of the way statistics are used and reported in medical literature.

3.3.2 'Calibration' of p-values

In this section, we will 'calibrate' p-values in order to obtain a significance level of the same type as α^* in Section 3. We will not be able to use the exact same test T^*, as most tests in literature are of the type: 'H_0 a precise hypothesis, H_1 may be composite' and the test T^* was only designed to handle two point hypotheses. Therefore, we will consider two calibrations. The first one is by Vovk and has some similarities to the test T^*. The second one is by Sellke, Bayarri and Berger and is based on a different principle. Both will be explained and then applied to a subset of the articles studied by Ioannidis.

Calibration by Vovk

By Markov's inequality (Theorem 3.2), we know that if we have some test $R(X)$ with $P(R(X) \geq 0) = 1$ such that $E_{P_0}\left[\frac{1}{R(X)}\right] \leq 1$, then $P_0(R(X) < \alpha) \leq \alpha$, for any $\alpha \in (0,1]$. Here, P_0 denotes the probability under the null hypothesis. Then we can apply the reasoning of Theorem 3.8 to show that optional stopping is not a problem. The only difference will be that the outside inequalities will not be strict anymore, so $P_0(\exists n : R(x^n) < \alpha) \leq \alpha$. Knowing this, the following theorem by Vovk proves to be very useful [20]:

Theorem 3.9 Let $p(X)$ be the p-value obtained from a test of the form 'H_0 a precise hypothesis, H_1 may be composite'. If $f : [0,1] \to \mathbb{R}^+$ is non-decreasing, continuous to the right and $\int_0^1 \frac{1}{f(x)} dx \leq 1$, then $E_{P_0}\left[\frac{1}{f(p(X))}\right] \leq 1$.

This theorem implies that if we find such an f, then we will have: $P_0(f(p(x)) < \alpha) \leq \alpha$ and $P_0(\exists n : f(p(x^n)) < \alpha) \leq \alpha$ for any $\alpha \in (0,1]$.

How to find such an f? Vovk suggested several types. One of them will be used throughout this thesis:

$$f^* : [0,1] \to \mathbb{R}^+ \text{ given by } f^*(p) = \frac{p^{1-\varepsilon}}{\varepsilon}$$

for any $\varepsilon \in (0,1)$. It is clear that f^* is continuous and non-decreasing and

$$\int_0^1 \frac{1}{f^*(p)} dp = \int_0^1 \varepsilon p^{\varepsilon - 1} dp = \varepsilon \left[\frac{p^\varepsilon}{\varepsilon}\right]_0^1 = 1.$$

It seems somewhat strange that it does not matter what test the source of the p-value is. In order to understand this, an analysis of Theorem 3.9 for discretized p-values is enlightening. The proof for continuous p-values can be found in Vovk [20].

'Analysis' of Theorem 3.9

Let $\beta \in (0,1)$ and consider a test that can only output discretized p-values: $p \in \{\beta^k | k \in \mathbb{N}_{\geq 0}\}$. Every test can be transformed to such a - slightly weaker - test and can be approximated very closely by choosing β almost equal to one.

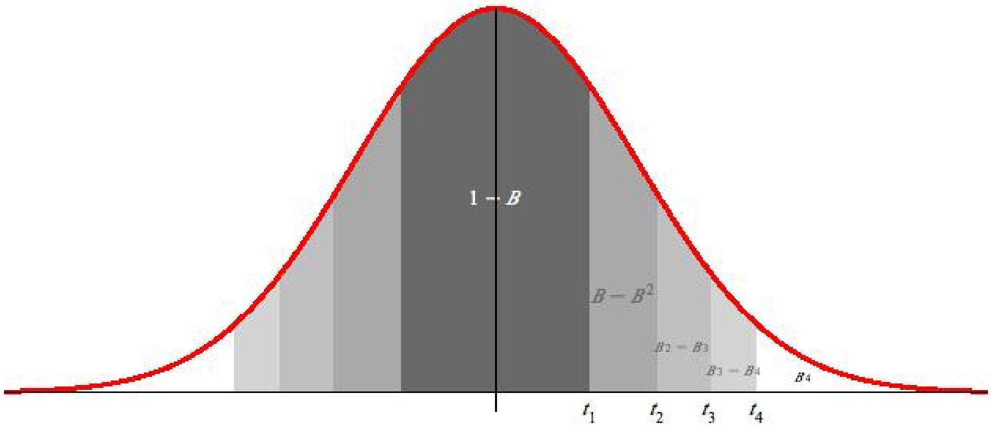


Figure 6: A graph of the standard normal distribution with $T(X) = |X|$, illustrating the discretization (graph made in Maple 13 for Mac).

Let T be the test statistic and define $t_1, t_2, \ldots$ to be the boundary values such that if $t_k \leq T(X) < t_{k+1}$, then $p = \beta^k$ for all $k \in \mathbb{N}_{\geq 0}$. An example of this procedure for the normal distribution can be seen in Figure 6. Then, after setting $P(t_0 \leq T(X) < t_1) = P(T(X) < t_1)$, it holds that $P_0(t_k \leq T(X) < t_{k+1}) = \beta^k(1-\beta)$ for all $k \in \mathbb{N}_{\geq 0}$. Thus:

$$V := E_{P_0}\left[\frac{1}{f(p(X))}\right] = \sum_{x^n} P(x^n) \cdot \frac{1}{f(p(x^n))} = \sum_{k=0}^{\infty} \sum_{x^n : t_k \leq T(x^n) < t_{k+1}} P(x^n) \cdot \frac{1}{f(p(x^n))}$$
$$= \sum_{k=0}^{\infty} \beta^k(1-\beta) \cdot \frac{1}{f(\beta^k)} = (1-\beta)\sum_{k=0}^{\infty} \beta^k \cdot \frac{1}{f(\beta^k)}.$$

Therefore, we should choose an f such that $(1-\beta)\sum_{k=0}^{\infty} \beta^k \cdot \frac{1}{f(\beta^k)} \leq 1$. By defining $\Delta_{\beta_k} = \beta^k - \beta^{k+1}$, we can rewrite this as: $\sum_{k=0}^{\infty} \Delta_{\beta_k} \frac{1}{f(\beta^k)} \leq 1$. Because β^k ranges from approximately zero to one, this is an approximation of the condition $\int_0^1 \frac{1}{f(x)} dx \leq 1$.

Let us check that this holds for f^*:

$$V = (1-\beta)\sum_{k=0}^{\infty} \beta^k \cdot \varepsilon \cdot \beta^{k(\varepsilon-1)} = \varepsilon(1-\beta)\sum_{k=0}^{\infty} (\beta^\varepsilon)^k = \frac{\varepsilon(1-\beta)}{1-\beta^\varepsilon}$$

$V = 1$ if $\varepsilon(1-\beta) = 1 - \beta^\varepsilon$. An exponential function has a maximum of two intersections with a linear function and it is easy to see that $\varepsilon = 0$ and $\varepsilon = 1$ do the job. Now $V \leq 1$ for all $\varepsilon \in (0, 1]$ and hence the requirement $V \leq 1$ is satisified by f^*. Therefore, it is possible to obtain such a calibration that does not depend on the source of the p-value.

Calibration by Sellke, Bayarri and Berger

Another calibration was proposed by Sellke, Bayarri and Berger [21]. This calibration is based on a different principle and partly continues the discussion on lower bounds on $P(H_0|x)$ in Section 2.5.3. This calibration can be used for a precise null hypothesis and an alternative does not need to be specified. The calibration is: if $p < \frac{1}{e} \approx 0.368$, then the type I error when rejecting the null hypothesis is:

$$\alpha(p) = \frac{1}{1 - \frac{1}{ep\log(p)}}.$$

Note that this calibration does not meet Vovk's conditions, since a quick Maple check reveals that $\int_\varepsilon^{1-\varepsilon} \frac{1}{\alpha(x)} dx > 3$ for some small ε. This calibration can be derived by a Bayesian analysis, by using the fact that, under the null hypothesis, the distribution of the p-value is Uniform[0,1]. Thus, we will test:

$$H_0 : p \sim \text{Uniform}[0,1] \text{ vs } H_1 : p \sim f(p|\xi).$$

If large values of the test statistic T cast doubt on H_0, then the density of p under H_1 should be decreasing in p. Therefore, the class of Beta$(\xi, 1)$ densities, given by $f(p|\xi) = \xi p^{\xi-1}, \xi \in (0, 1]$ seem to be a reasonable choice for $f(p|\xi)$. The null hypothesis is then: $p \sim f(p|1)$. Then, as shown in Section 2.5.3, the minimum Bayes factor is equal to:

$$\underline{B}(p) = \frac{f(p|1)}{\sup_\xi \xi p^{\xi-1}} = \frac{1}{\sup_\xi \xi p^{\xi-1}}.$$

By setting $\frac{d}{d\xi}\xi p^{\xi-1} = (1 + \xi\log(p))p^{\xi-1}$ equal to zero, we find $\xi = -\frac{1}{\log(p)}$. Substitution into $\xi p^{\xi-1}$ gives $-\frac{1}{\log(p)} p^{-\frac{1}{\log(p)}-1} = -\frac{1}{p\log(p)}\left(e^{\log(p)}\right)^{-\frac{1}{\log(p)}} = -\frac{1}{ep\log(p)}$. Hence:

$$\underline{B}(p) = -ep\log(p).$$

If we assume $P(H_0) = P(H_1)$, then a lower bound on the posterior probability of H_0 for the Beta$(\xi, 1)$ alternatives is:

$$\underline{P}(H_0|p) = \left[1 + \frac{1-\pi_0}{\pi_0} \cdot \frac{1}{\underline{B}(p)}\right]^{-1} = \frac{1}{1 - \frac{1}{ep\log(p)}}.$$

This calibration can thus be interpreted as a lower bound on $P(H_0|p)$ for a broad class of priors. The authors showed that this lower bound can also be interpreted as a lower bound on the frequentist type I error probability, but those arguments, involving conditioning and ancillary statistics, are beyond the scope of this thesis.

3.3.3 Example analysis of two articles

Because the abstract is expected to hold the most crucial information, including the p-values for the outcomes based on which the conclusions were made, the 'results' sections of the abstracts of the articles considered by Ioannidis were the starting-points for this analysis. A quick look at the results revealed a complication: most authors base their conclusions on multiple p-values. To simplify comparisons between studies, I have tried to discern the primary outcome. In some cases, this was very clearly stated by the authors. Sometimes, it required a closer reading of the article. An example of both types is provided.

An article in which the primary outcome was clearly stated is CIBIS-II investigators and committees (1999), The Cardiac Insufficiency Bisoprolol Study II (CIBIS-II): a randomised trial, *Lancet* **353**, 9-13. They investigated the efficacy of bisoprolol, a β_1 selective adrenoceptor blocker, in decreasing all-cause mortality in chronic heart failure. Their results section in the abstract reads:

> *"**Findings** CIBIS-II was stopped early, after the second interim analysis, because bisoprolol showed a significant mortality benefit. All-cause mortality was significantly lower with bisoprolol than on placebo (156 [11.8%] vs 228 [17.3%] deaths) with a hazard ratio of 0.66 (95% CI 0.54 - 0.81, p<0.0001). There were significantly fewer sudden deaths among patients on bisoprolol than in those on placebo (48 [3.6%] vs 83 [6.3%] deaths), with a hazard ratio of 0.56 (0.39 - 0.80, p=0.0011). Treatment effects were independent of the severity or cause of heart failure."*

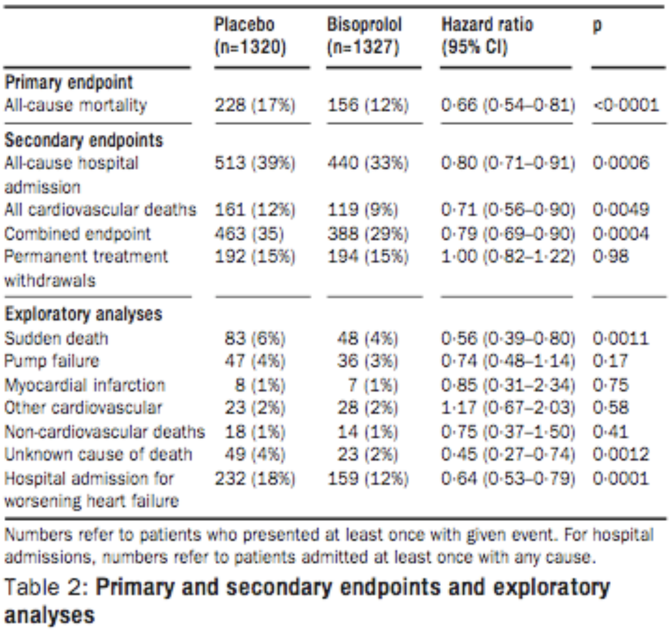
This in itself is very clear. The first null hypothesis is presumably that there is no difference in mortality when taking placebo or bisoprolol. The probability of finding the observed number of deaths in both groups, or more extreme numbers, is smaller than 0.0001. The second null hypothesis is that the probability of sudden death is the same for patients taking placebo and patients taking bisoprolol. The probability of finding the observed number of sudden deaths in both groups or more extreme numbers under this null hypothesis is 0.0011.

The authors of this study chose to state only two p-values in the abstract. The assumption that those must be the p-values for the primary outcomes turned out to be correct: it is stated in the full text of the article and can easily be seen from the table reproduced in Figure 7. All relevant information can be found in that table.

Let us analyze this article according to the two calibrations. First, with Vovk's function $f^*(p) = \frac{p^{1-\varepsilon}}{\varepsilon}$. There is no obvious choice for ε. Instead of analysing these results using a fixed value for ε, I will find the ε^* such that f^* is minimized. In this way, a 'minimum significance level', denoted by α_0 at which the test based on f^* would have rejected the results will be obtained. Values of α_0 smaller than or equal to the arbitrary but popular value of 0.05 will be considered significant. All minimizations have been performed using the Maple command `minimize`. I found the cuttoff for the function f^* to be approximately $p = 0.003202$.

Furthermore, I will apply the calibration by Sellke, Bayarri and Berger: $\alpha_1(p) = \left[1 - [ep\log(p)]^{-1}\right]^{-1}$, again considering values of α_1 smaller than or equal to 0.05 to be significant. This means we can reject for $p < 0.00341$ (value calculated using Maple's `solve` function).

I will only analyze the p-values for the primary and secondary endpoints, omitting 'permanent



	Placebo (n=1320)	Bisoprolol (n=1327)	Hazard ratio (95% CI)	p
Primary endpoint				
All-cause mortality	228 (17%)	156 (12%)	0·66 (0·54–0·81)	<0·0001
Secondary endpoints				
All-cause hospital admission	513 (39%)	440 (33%)	0·80 (0·71–0·91)	0·0006
All cardiovascular deaths	161 (12%)	119 (9%)	0·71 (0·56–0·90)	0·0049
Combined endpoint	463 (35)	388 (29%)	0·79 (0·69–0·90)	0·0004
Permanent treatment withdrawals	192 (15%)	194 (15%)	1·00 (0·82–1·22)	0·98
Exploratory analyses				
Sudden death	83 (6%)	48 (4%)	0·56 (0·39–0·80)	0·0011
Pump failure	47 (4%)	36 (3%)	0·74 (0·48–1·14)	0·17
Myocardial infarction	8 (1%)	7 (1%)	0·85 (0·31–2·34)	0·75
Other cardiovascular	23 (2%)	28 (2%)	1·17 (0·67–2·03)	0·58
Non-cardiovascular deaths	18 (1%)	14 (1%)	0·75 (0·37–1·50)	0·41
Unknown cause of death	49 (4%)	23 (2%)	0·45 (0·27–0·74)	0·0012
Hospital admission for worsening heart failure	232 (18%)	159 (12%)	0·64 (0·53–0·79)	0·0001

Numbers refer to patients who presented at least once with given event. For hospital admissions, numbers refer to patients admitted at least once with any cause.

Table 2: **Primary and secondary endpoints and exploratory analyses**

Figure 7: Table with p-values, reproduced from 'The Cardiac Insufficiency Bisoprolol Study II'.

treatment withdrawals', because this is not considered significant by the authors. If $p < c$ is reported instead of $p = c$, I will analyze it as if $p = c$ was reported. The results are in Table 6:

Table 6: Results for 'The Cardiac Insufficiency Bisoprolol Study II'.

endpoint	p-value	α_0	α_0 significant?	α_1	α_1 significant?
all-cause mortality	*0.0001*	*0.00250*	*yes*	*0.00250*	*yes*
all-cause hospital admission	0.0006	0.0121	yes	0.0120	yes
all cardiovascular deaths	0.0049	0.0708	no	0.0662	no
combined endpoint	0.0004	0.00851	yes	0.00844	yes

Italics indicate the primary outcome. The results for both tests are remarkably similar. All results hold up very well for both calibrations, except for the endpoint 'all cardiovascular deaths': this would not have been considered significant by tests based on the two calibrations. The primary endpoint, however, would quite convincingly be considered significant by both tests. This is in accordance with Ioannidis' findings, as this study is categorized by Ioannidis as 'replicated'.

It is somewhat harder to find the primary endpoint in the results section of Ziegler, E.J., Fisher, C.J., Sprung, C.L., et al. (1991), Treatment of gram-negative bacteremia and septic shock with HA-1A human monoclonal antibody against endotoxin. A randomized, double-blind, placebo-controlled trial, *The New England Journal of Medicine*, **324** (7), 429-436. They evaluated the efficacy and safety of HA-1A, a human monoclonal IgM antibody, in patients with sepsis and a presumed diagnosis of gram-negative infection. Their results section reads:

> *"Results. Of 543 patients with sepsis who were treated, 200 (37 percent) had gram-negative bacteremia as proved by blood culture. For the patients with gram-negative bacteremia followed to death or day 28, there were 45 deaths among the 92 recipients of placebo (49 percent) and 32 deaths among the 105 recipients of HA-1A (30 percent; P = 0.014). For the patients with gram-negative bacteremia and shock at entry, there were 27 deaths among the 47 recipients of placebo (57 percent) and 18 deaths among the 54 recipients of HA-1A (33 percent; P=0.017). Analyses that stratified according to the severity of illness at entry showed improved survival with HA-1A treatment in both severely ill and less severely ill patients. Of the 196 patients with gram-negative bacteremia who were followed to hospital discharge or death, 45 of the 93 given placebo (48 percent) were*

> *discharged alive, as compared with 65 of 103 treated with HA-1A (63 percent; P = 0.038). No benefit of treatment with HA-1A was demonstrated in the 343 patients with sepsis who did not prove to have gram-negative bacteremia. For all 543 patients with sepsis who were treated, the mortality rate was 43 percent among the recipients of placebo and 39 percent among those given HA-1A (P = 0.24). All patients tolerated HA-1A well, and no anti-HA-1A antibodies were detected."*

What is the primary outcome here? The full text of the article states twelve p-values. From the first sentence of the 'Discussion' ("The results of this clinical trial show that adjunctive therapy with HA-1A, a human monoclonal antibody against endotoxin, reduces mortality significantly in patients with sepsis and gram-negative bacteremia.") I gathered that the primary outcome was 'mortality in patients with sepsis and gram-negative bacteremia', with an associated p-value of $p = 0.014$.

What about the other p-values? Some were only indirectly related to the outcomes, such as: "This analysis indicates that shock was an important determinant of survival (P = 0.047)" and "Pretreatment APACHE II scores were highly correlated with death among the patients given placebo in all populations examined (P = 0.0001)". Some of the p-values were deemed to indicate non-significance and will be omitted. Determining which p-values concern primary or secondary outcomes requires a read-through of the full text. The classification in Table 7 is therefore mine. The analysis of the p-values was done in exactly the same way as in the previous example. Both calibrations deem none of the outcomes to be significant. This is once again in accordance with Ioannidis' findings, as he classified this study as 'contradicted'.

Table 7: Results for 'Treatment of gram-negative bacteremia and septic shock with HA-1A human monoclonal antibody against endotoxin'.

endpoint	p-value	α_0	α_0 significant?	α_1	α_1 significant?
mortality (sepsis + bacteremia)	*0.014*	*0.162*	*no*	*0.140*	*no*
mortality (shock + bacteremia)	0.017	0.188	no	0.158	no
mortality (bacteremia)	0.012	0.144	no	0.126	no
resolution of complications	0.024	0.243	no	0.196	no
discharge alive	0.038	0.338	no	0.252	no

3.3.4 Results

Ioannidis distinguished five study categories: contradicted (7), initially stronger effects (7), replicated (20), unchallenged (11) and negative (4). Because the replicated studies are most likely to have found a true effect and the contradicted studies are least likely to have found a true effect, I have analyzed only articles from those categories. I analyzed all 7 contradicted studies and the bottom half of the replicated studies, as ordered in table 2 in [19, p.222]. The analysis was performed in the same way as in Section 3.3.3. Studies 49 en 56, both replicated studies, have been omitted, because they do not report any p-values. The results are summarized in Table 8 and Table 9. Because in every single cases α_0 was significant if and only if α_1 was significant, I have merged their separate columns into one column.

The complete data on which the summary was based can be found in Appendix A (for the contradicted studies) and Appendix B (for the replicated studies). All articles are only referred to by number: this number is the same as the reference number in Ioannidis' article. The complete references can also be found in the appendices.

Table 8: Results for the 7 contradicted studies.

study	primary p-value	α_0	α_1	α_0, α_1 sign.?	#other p-values	p-value range	# α_0, α_1 sign.
13[1]	0.0001	0.00250	0.00250	yes	3	0.0004 - 0.42	1
15	0.014	0.162	0.140	no	4	0.012 - 0.038	0
20[1]	0.003	0.0474	0.0452	yes	6	0.02 - 0.22	0
21[1]	0.001	0.0188	0.0184	yes	0	-	0
22	0.008	0.105	0.095	no	12	0.002 - 0.028	1
42	0.001	0.0188	0.0184	yes	4	0.001 - 0.03	1
51	0.005	0.0720	0.0672	no	3	0.0001 - 0.015	1

Table 9: Results for 8 replicated studies.

study	primary p-value	α_0	α_1	α_0, α_1 sign.?	#other p-values	p-value range	# α_0, α_1 sign.
36	0.001	0.0188	0.0184	yes	2	0.04 - 0 .07	0
37	0.001	0.0188	0.0184	yes	4	0.001 - 0.05	3
38	0.001	0.0188	0.0184	yes	17	0.001 - 0.048	11
41[2]	0.0003	0.00661	0.00657	yes	0	-	0
45	0.001	0.0188	0.0184	yes	6	0.001 - 0.02	4
53	0.001	0.0188	0.0184	yes	3	0.001- 0.002	3
55	0.0001	0.00250	0.00250	yes	3	0.0004 - 0.0049	2
57[3]	0.00001	0.000313	0.000313	yes	1	0.002	1

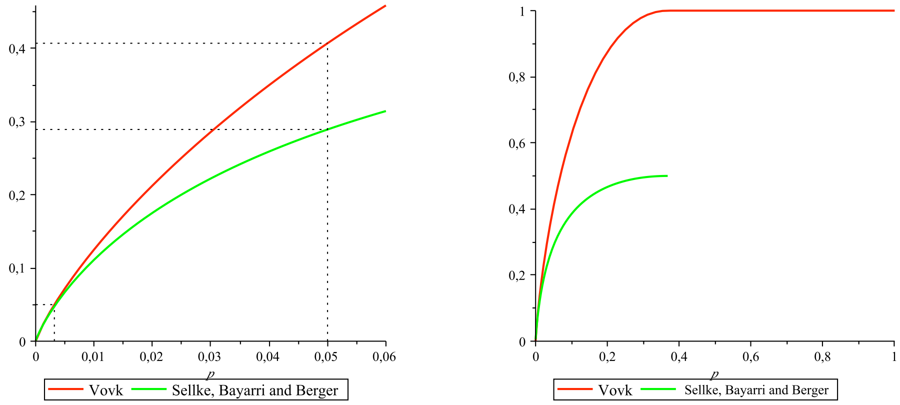
[1] : Many results were reported only in the form of 95% confidence intervals.

[2]: A p-value was only provided for the primary outcome. For the secondary outcomes, 95% confidence intervals were provided.

[3]: Additional results not concerning the primary outcome were reported by means of 95% confidence intervals.

3.3.5 Discussion

First off, both calibrations yielded remarkably similar results. A plot of both calibrations (see Figure 8) reveals that their values are very similar for small values of p, but that Vovks calibration is considerably larger than the one by Sellke, Bayarri and Berger for larger values of p. This can be explained because Vovks calibration provides an upper bound on the type I error rate, while the other calibration provides a lower bound. The calibration by Sellke, Bayari and Berger is only plotted for $0 < p < \frac{1}{e}$, as it is invalid for larger values of p. Because most p-values were smaller than 0.05, both calibrations gave approximately the same results. As noted, their 'cutoff for significance' is almost the same: for Vovks calibration, it is approximately $p = 0.0032$ and for the calibration by Sellke, Bayarri and Berger, it is approximately $p = 0.0034$.



(a) Plot of the two calibrations, zoomed in.

(b) Plot of $\min_\varepsilon f(p)$ (red) and $(1 - (ep\log(p))^{-1})^{-1}$ (green).

Figure 8: Comparison of the two calibrations (graphs made in Maple 13 for Mac).

It is remarkable that for the replicated studies, all p-values associated with the primary outcome remained significant. Some p-values for secondary outcomes lost their significance, but overall 67% (24 out of 36) held up. For the contradicted, four out of seven p-vaues corresponding to the primary outcome did hold up, but many p-values concerning primary outcomes became insignificant under both calibrations: only 13% (4 out of 32) remained significant. However, it is difficult to compare the contradicted and the replicated groups, because they are both small and the way in which p-values were used and reported was not uniform for all articles. Therefore, no conclusions can be made based on Table 8 and Table 9. Nevertheless, they do seem to suggest that the problems may be alleviated somewhat by choosing a smaller 'critical' p-value, for example $p = 0.001$. This would not free us from all the trouble: sampling to a foregone conclusion would still be possible, Lindley's paradox would still be troublesome, the interpretation of a p-value would not change, p-values would still depend on sampling plans and other subjective intentions and the evidence against the null hypothesis would still be exaggerated. However, many of these problems depend on the sample size n, which implies that a much larger sample size would be necessary in order for trouble to start. Choosing a smaller cutoff would therefore not solve the entire conflict, but it might mitigate it.

Conclusion

> *cuiusvis hominis est errare; nullius nisi insipientis perseverare in errore.*
> Anyone is capable of making mistakes, but only a fool persists in his error.
> — Marcus Tullius Cicero, *Orationes Philippicae* 12.2

In the first chapter, both Fisher's and Neyman and Pearson's view on what we can learn about hypotheses from empirical tests were discussed. It turned out that their ideas are incompatible. Fisher thought that we can make an inference about one particular hypothesis, while Neyman and Pearson considered this to be impossible. While their methods have, ironically, been fused into a hybrid form, which follows Neyman and Pearson methodologically, but Fisher philosophically, the fierce debate between their founders is now largely unknown to researchers. This has led to the confusion of the p-value and the type I error probability α.

However, that seems to be a minor problem when compared to the points raised in chapter two, in which p-values were considered. They by themselves are severely misunderstood: researchers are interested in $P(H_0|\text{data})$, but what they are calculating is $P(\text{data}|H_0)$. These two values are often confused and this raises the question whether p-values are useful: if the users do not know what they mean, how can they draw valid conclusions from them?

Putting that aside, there are many additional arguments that makes one question whether we should continue using p-values. Especially the exaggeration of the evidence against the null hypothesis seems to be an undesirable property of p-values. The cause of this discrepancy is that the p-value is a tail-area probability. The logic behind using the observed 'or more extreme' outcomes seems hardly defensible. Furthermore, optional stopping is possible for both researchers with bad intentions and researchers with good intentions. If we only need enough patience and money to prove whatever we want, what is then the value of a 'fact' proven by this method?

However, abandoning the use of p-values is easier said than done. The millions of psychologists and doctors will not be amused if they will have to learn new statistics, especially when they do not understand what is wrong with the methods they have been using all their life. Statistical software would have to be adjusted and new textbooks would need to be written. Besides these huge practical problems, we need an alternative to fill the spot left by the p-values. What test is up for the task? In this thesis, I considered an alternative test based on the likelihood ratio, which compared favorably to the p-value based test. Sampling to a foregone conclusion will likely not succeed using this test and all other points raised against p-values are either not applicable or much less severe. Unfortunately, this alternative test can only handle precise hypotheses, while in practice composite alternative hypotheses are often considered. Thus, this test is not The Answer.

The calibrations did not provide a new test, but they did illustrate that p-values that are considered to be significant are not significant anymore when using different principles. It was very interesting to see how p-values are reported in practice and how little awareness there seems to be that $p < 0.05$ does not imply that one's conclusions are true. The brief survey of the effects of the calibrations on the p-values suggested that matters might improve somewhat by choosing a smaller cutoff for significance, such as $p = 0.001$, but alas, many problems would still remain.

In conclusion, I think we should heed the wise words of Cicero: it seems obvious now that p-values are not fit to be measures of evidence, both because of misuse and more fundamental undesirable properties, and it would not be wise of us to continue using them. While I cannot offer a less problematic test to replace traditional hypothesis tests, I do hope that all the problems associated with p-values will become more well-known, so that conclusions will be drawn more carefully until we have a better method of acquiring knowledge.

References

[1] Edwards, W., Lindman, H., Savage, L.J. (1963), Bayesian statistical inference for psychological research, *Psychological Review*, **70** (3), 193-242.

[2] Hubbard, R. (2004), Alphabet soup: blurring the distinctions between p's and a's in psychological research, *Theory & Psychology*, **14** (3), 295-327.

[3] Goodman, S.N. (1993), p Values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate, *American Journal of Epidemiology*, **137** (5), 485-496.

[4] Goodman, S.N. (1999), Toward evidence-based medical statistcs. 1: The p-value fallacy, *Annals of Internal Medicine*, **130** (12), 995-1004.

[5] Cohen, J. (1994), The earth is round ($p < .05$), *American Psychologist*, **49** (12), 997-1003.

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

[7] Berger, J.O., Sellke, T. (1987), Testing a point null hypothesis: the irreconcilability of p-values and evidence, *Journal of the American Statistical Association*, **82**, 112-122.

[8] Fidler, F., Thomason, N., Cumming, G. et al. (2004), Editors can lead researchers to confidence intervals, but can't make them think. Statistical reform lessons from medicine, *Psychological Science*, **15** (2), 119-126.

[9] Wagenmakers, E.-J. (2007), A practical solution to the pervasive problems of p-values, *Psychonomic Bulletin & Review*, **14** (5), 779-804.

[10] Rice, J.A. (2007^3) *Mathematical statistics and data analysis*, Thomson Brooks/Cole (Belmont).

[11] Berger, J.O., Wolpert, R.L. (1988^2), *The likelihood principle*, Institute of Mathematical Statistics (Hayward).

[12] Spiegelhalter, D.J., Abrams, K.R., Myles, J.P. (2004), *Bayesian approaches to clinical trials and health-care evaluation*, John Wiley & Sons, Ltd (Chichester).

[13] Lindley, D.V., (1957), A statistical paradox, *Biometrika*, **44**, 187-192.

[14] Feller, W. (1957^2) *An introduction to probability theory and its applications*, John Wiley & Sons, Inc (New York).

[15] Schmidt, F.L., Hunter, J.E. (1997), Eight common but false objections to the discontinuation of significance testing in the analysis of research data, in Harlow, L.L., Mulaik, S.A., Steiger, J.H. edd. *What if there were no significance tests?*, 37-64.

[16] Nickerson, R.S. (2000), Null hypothesis significance testing: a review of an old and continuing controversy, *Psychological Methods*, **5** (2), 241-301.

[17] Grünwald, P.D. (2007) *The minimum description length principle*, The MIT Press (Cambridge, Massachusetts).

[18] Barnard, G.A. (1990), Must clinical trials be large? The interpretation of p-values and the combination of test results, *Statistics in Medicine*, **9**, 601-614.

[19] Ioannidis, J.P.A. (2005), Contradicted and initially stronger effects in highly cited clinical research, *Journal of the American Medical Association*, **294** (2), 218-228.

[20] Vovk, V.G. (1993), A logic of probability, with application to the foundations of statistics, *Journal of the Royal Statistical Societ. Series B (Methodological)*, **55** (2), 317-341.

[21] Sellke, T., Bayarri, M.J., Berger, J.O., Calibration of p values for testing precise null hypotheses, *The American Statistician*, **55** (1), 62-71.

A Tables for the contradicted studies

Study 13: Stampfer, M.J, Colditz, G.A., Willett, W.C., et al. (1991), Postmenopausal estrogen therapy and cardiovascular disease. Ten-year follow-up from the Nurses' Health Study, *The New England Journal of Medicine*, **325** (11), 756-762.

Study 15: Ziegler, E.J., Fisher, C.J., Sprung, C.L., et al. (1991), Treatment of gram-negative bacteremia and septic shock with HA-1A human monoclonal antibody against endotoxin. A randomized, double-blind, placebo-controlled trial, *The New England Journal of Medicine*, **324** (7), 429-436.

Study 20: Rimm, E.B., Stampfer, M.J., Ascherio, A., et al. (1993), Vitamin E consumption and the risk of coronary heart disease in men, *The New England Journal of Medicine*, **328** (20), 1450-1456.

Study 21: Stampfer, M.J., Hennekens, C.H., Manson, J.E., et al. (1993), Vitamin E consumption and the risk of coronary disease in women, *The New England Journal of Medicine*, **328** (20), 1444-1449.

Study 22: Rossaint, R., Falke, K.J, Lopez, F., et al. (1993), Inhaled nitric oxid for the adult respiratory distress syndrome, textitThe New England Journal of Medicine, **328** (6), 399-405.

Study 42: The writing group for the PEPI trial (1995), Effects of estrogen or estrogen/progestin regimens on heart disease risk factors in postmenopausal women. The Postmenopausal Estrogen/Progestin Interventions (PEPI) trial, *Journal of the American Medical Association*, **273** (3), 199-206.

Study 51: Stephens, N.G., Parsons, A., Schofield, P.M., et al. (1996), Randomised controlled trial of vitamin E in patients with coronary disease: Cambridge Heart Antioxidant Study (CHAOS), *The Lancet*, **347**, 781-786.

Table 10: Results for study 13.

endpoint	p-value	α_0	α_1	α_0, α_1 significant?
Reduced risk coronary disease	0.0001	0.00250	0.00250	yes
Reduced risk coronary disease former users	0.42	1	0.498	no
Mortality all causes former users	0.0004	0.00851	0.00844	yes
Cardiovascular mortality former users	0.02	0.213	0.175	no

Table 11: Results for study 15.

endpoint	p-value	α_0	α_1	α_0, α_1 significant?
mortality (sepsis + bacteremia)	0.014	0.162	0.140	no
mortality (shock + bacteremia)	0.017	0.188	0.158	no
mortality (bacteremia)	0.012	0.144	0.126	no
resolution of complications	0.024	0.243	0.196	no
discharge alive	0.038	0.338	0.252	no

Table 12: Results for study 20.

endpoint	p-value	α_0	α_1	α_0, α_1 significant?
Coronary disease (vit E)	0.003	0.0474	0.0452	yes
Coronary disease (carotene)	0.02	0.213	0.175	no
Coronary disease (vit E supp)	0.22	0.905	0.475	no
Coronary disease (diet)	0.11	0.660	0.398	no
Overall mortality highest - lowest intake	0.06	0.459	0.315	no
Coronary disease (carotene, former smoker)	0.04	0.350	0.259	no
Coronary disease (carotene, current smoker)	0.02	0.213	0.175	no

Table 13: Results for study 21.

endpoint	p-value	α_0	α_1	α_0, α_1 significant?
Major coronary disease	0.001	0.0188	0.0184	yes

Table 14: Results for study 22.

endpoint	p-value	α_0	α_1	α_0, α_1 significant?
Reduction pulmonary-artery pressure NO	0.008	0.105	0.095	no
Reduction pulmonary-artery pressure pros	0.011	0.135	0.119	no
Cardiac output pros	0.015	0.171	0.146	no
Pulmonary vascular resistance NO	0.008	0.105	0.095	no
Pulmonary vascular resistance pros	0.011	0.135	0.119	no
Systemic vascular resistance pros	0.002	0.0338	0.0327	yes
Decrease intrapulmonary shunting NO	0.028	0.272	0.214	no
Increase arterial oxygenation NO	0.008	0.105	0.095	no
Decrease arterial oxygenation pros	0.005	0.072	0.0672	no
Increase partial pressure oxygen venous NO	0.008	0.105	0.095	no
Increase blood flow lung regions NO	0.011	0.135	0.119	no
Decrease blood flow lung regions NO	0.012	0.144	0.126	no
Decrease $\log_S DQ$ NO	0.011	0.135	0.119	no

Table 15: Results for study 42.

endpoint	p-value	α_0	α_1	α_0, α_1 significant?
Lipoproteins	0.001	0.0188	0.0184	yes
Fibrinogen	0.001	0.0188	0.0184	yes
Glucose	0.01	0.125	0.111	no
Glucose (fasting)	0.03	0.286	0.222	no
Weight gain	0.03	0.286	0.222	no

Table 16: Results for study 51.

endpoint	p-value	α_0	α_1	α_0, α_1 significant?
Cardiovascular death and non-fatal MI	0.005	0.0720	0.0672	no
Non-fatal MI	0.005	0.0720	0.0672	no
Major cardiovascular events	0.015	0.171	0.146	no
Non-fatal MI	0.0001	0.00250	0.00250	yes

B Tables for the replicated studies

Study 36: The EPILOG investigators (1997), Platelet glycoprotein IIb/IIIa receptor blockade and low-dose heparin during percutaneous coronary revascularization, *The New England Journal of Medicine*, **336** (24), 1689-1696.

Study 37: McHutchison, J.G., Gordon, S.C., Schiff, E.R., et al. (1998), Interferon alfa-2b alone or in combination with ribavirin as initial treatment for chronic hepatitis C, *The New England Journal of Medicine*, **339** (21), 1485-1492.

Study 38: The Long-Term Intervention with Pravastatin in Ischaemic Disease (LIPID) study group (1998), Prevention of cardiovascular events and death with pravastatin in patients with coronary heart disease and a broad range of initial cholesterol levels, *The New England Journal of Medicine*, **339** (19), 1349-1357.

Study 41: SHEP cooperativee research group (1991), Prevention of stroke by antihypertensive drug treatment in older persons with isolated systolic hypertension, *Journal of the American Medical Association*, **265** (24), 3255-3264.

Study 45: Downs, J.R., Clearfield, M., Weis, S., et al. (1998), Primary prevention of acute coronary events with lovastatin in men and women with average cholesterol levels: results of AFCAPS/TexCAPS, *Journal of the American Medical Association*, **279** (20), 1615-1622.

Study 53: Poynard, T., Marcellin, P., Lee, S.S., et al. (1998), Randomised trial of interferon α2b plus ribavirine for 48 weeks or for 24 weeks versus interferon α2b plus placebo for 48 weeks for treatment of chronic infection with hepatitis C virus, *The Lancet*, **352**, 1426-1432.

Study 55: CIBIS-II investigators and committees (1999), The Cardiac Insufficiency Bisoprolol Study II (CIBIS-II): a randomised trial, *The Lancet* **353**, 9-13.

Study 57: Fisher, B., Costantino, J.P., Wickerham, D.L., et al. (1998), Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 study, *Journal of the National Cancer Institute*, **90** (18), 1371-1388.

Table 17: Results for study 36.

endpoint	p-value	α_0	α_1	α_0, α_1 significant?
Death, MI or urgent revascularization	0.001	0.0188	0.0184	yes
Death, MI, repeated revascularization low	0.07	0.506	0.336	no
Death, MI, repeated revascularization std	0.04	0.350	0.259	no

Table 18: Results for study 37.

endpoint	p-value	α_0	α_1	α_0, α_1 significant?
Sustained virologic response	0.001	0.0188	0.0184	yes
Increase virologic response 24 vs 48 weeks	0.05	0.407	0.289	no
Rate of response combination 24	0.001	0.0188	0.0184	yes
Rate of response combination 48	0.001	0.0188	0.0184	yes
Histologic improvement	0.001	0.0188	0.0184	yes

Table 19: Results for study 38.

endpoint	p-value	α_0	α_1	α_0, α_1 significant?
Death due to CHD	0.001	0.0188	0.0184	yes
Death due to CVD	0.001	0.0188	0.0184	yes
Overall mortality	0.001	0.0188	0.0184	yes
MI	0.001	0.0188	0.0184	yes
Death due to CHD or nonfatal MI	0.001	0.0188	0.0184	yes
CABG	0.001	0.0188	0.0184	yes
PTCA	0.024	0.243	0.196	no
CABG or PTCA	0.001	0.0188	0.0184	yes
Unstable angina	0.005	0.0720	0.0672	no
Stroke	0.048	0.396	0.284	no
Coronary revascularization	0.001	0.0188	0.0184	yes
Lipid levels	0.001	0.0188	0.0184	yes
Death due to CHD, previous MI	0.004	0.0600	0.0566	no
Overall mortality, previous MI	0.002	0.0338	0.0327	yes
Death due to CHD, previous UA	0.036	0.325	0.245	no
Overall mortality, previous UA	0.004	0.0600	0.0566	no
Less time in hospital	0.001	0.0188	0.0184	yes
Fewer admissions + less time	0.002	0.0338	0.0327	yes

Table 20: Results for study 41.

endpoint	p-value	α_0	α_1	α_0, α_1 significant?
Stroke	0.0003	0.00661	0.00657	yes

Table 21: Results for study 45.

endpoint	p-value	α_0	α_1	α_0, α_1 significant?
First acute major coronary event	0.001	0.0188	0.0184	yes
MI	0.002	0.0338	0.0327	yes
Unstable angina	0.02	0.213	0.175	no
Coronary revascularization	0.001	0.0188	0.0184	yes
Coronary events	0.006	0.0834	0.0770	no
Cardiovascular events	0.003	0.0474	0.0452	yes
Lipid levels	0.001	0.0188	0.0184	yes

Table 22: Results for study 53.

endpoint	p-value	α_0	α_1	α_0, α_1 significant?
Sustained virological response both regimens	0.001	0.0188	0.0184	yes
Sustained response ¡3 factors	0.002	0.0338	0.0327	yes
Sustained normalisation al am 48 wks	0.001	0.0188	0.0184	yes
Histological improvement	0.001	0.0188	0.0184	yes

Table 23: Results for study 55.

endpoint	p-value	α_0	α_1	α_0, α_1 significant?
all-cause mortality	0.0001	0.00250	0.00250	yes
all-cause hospital admission	0.0006	0.0121	0.0120	yes
all cardiovascular deaths	0.0049	0.0708	0.0662	no
combined endpoint	0.0004	0.00851	0.00844	yes

Table 24: Results for study 57.

endpoint	p-value	α_0	α_1	α_0, α_1 significant?
Invasive breast cancer	0.00001	0.000313	0.00313	yes
Noninvasive breast cancer	0.002	0.0338	0.0327	yes