The London School of Economics and Political Science

To p, or not to p?
Quantifying Inferential Decision Errors To Assess Whether Significance Truly Is Significant

James Spencer Abdey

A thesis submitted to the Department of Statistics of the London School of Economics and Political Science for the degree of Doctor of Philosophy, London, September 2009

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

Dedication

In loving memory of Mum

Susan Ivy Abdey

'Sue'

1947-2008

Forever indebted to the sacrifices you made.

Abstract

Empirical testing is centred on p-values. These summary statistics are used to assess the plausibility of a null hypothesis, and therein lies a flaw in their interpretation. Central to this research is accounting for the behaviour of p-values, through density functions, under the alternative hypothesis, H_1. These densities are determined by a combination of the sample size and parametric specification of H_1. Here, several new contributions are presented to reflect p-value behaviour.

By considering the likelihood of both hypotheses in parallel, it is possible to optimise the decision-making process. A framework for simultaneously testing the null and alternative hypotheses is outlined for various testing scenarios. To facilitate efficient empirical conclusions, a new set of critical value tables is presented requiring only the conventional p-value, hence avoiding the need for additional computation in order to apply this joint testing in practice. Simple and composite forms of H_1 are considered.

Recognising the conflict between different schools of thought with respect to hypothesis testing, a unified approach at consolidating the advantages of each is offered. Again, exploiting p-value distributions under various forms of H_1, a revised conditioning statistic for conditional frequentist testing is developed from which original p-value curves and surfaces are produced to further ease decision making.

Finally, attention turns to multiple hypothesis testing. Estimation of multiple testing error rates is discussed and a new estimator for the proportion of true null hypotheses, when simultaneously testing several independent hypotheses, is presented. Under certain conditions it is shown that this estimator is superior to an established estimator.

Acknowledgements

Production of this thesis would not have been possible without the continued help and support of my two principal supervisors — initially Dr Jeremy Penzer, followed by Dr Irini Moustaki. Regular meetings allowed this research to reach its ultimate conclusion. Also, throughout the research period, I am grateful to numerous comments and suggestions received from conference participants following various presentations, both nationally and internationally.

My thanks for financial support go to the Economic and Social Research Council (ESRC) under award PTA-030-2005-00047. Given the ESRC's activities aim to promote research in the social sciences, it is hoped that the methodologies presented here are readily applicable to this domain.

Finally, on a personal level, I would like to dedicate this work to the memory of the one person who has been a unique source of strength and support to me over many years. Mum, I so wish you could have witnessed the completion of this work. Though sadly not possible, I will do my utmost to make you proud in my future endeavours.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

P-values have become a staple measure used in empirical research across a diverse range of disciplines. Theoretical ideas are formulated into hypotheses of interest and subsequently tested to allow researchers to assess the theory's plausibility, and hence application to the real world. Given the prevalence of p-value reporting, it is important to establish the correct interpretation of this summary statistic so that accurate inferences are made. This will therefore justify the widespread usage of p-values among practitioners in response to numerous criticisms of the Fisherian p-value. This tome considers the inferential issues surrounding p-values in detail.

Conditional on collected sample data, the process of statistical testing is to evaluate the feasibility of the null hypothesis, H_0, and, ideally, conclude firmly in favour either of it or the alternative hypothesis, H_1. Three prominent 'schools' of testing exist, propelled by Fisher, Jeffreys and Neyman. Fisher extolled the virtue of the p-value, whose magnitude signals the strength of evidence in favour of H_0. In contrast, Jeffreys' approach favours the use of objective posterior probabilities using a Bayesian framework, whilst Neyman resorted to fixed error probabilities, namely the computation of Type I and Type II errors.

Having such competing doctrines poses a conundrum for practitioners since the schools typically yield different reported results. Which, if any, is uniformly superior? Although Fisherian p-values are widely-utilised, does there exist a way to combine all three approaches? A universal methodology is clearly in demand and its widespread adoption would allow consistency of implementation and an unambiguous interpretation of test results in empirical applications.

Berger, Brown, and Wolpert (1994) present the conditional frequentist approach to testing which provides a basis for the methodological unification of the different schools for simple versus simple hypothesis testing. Here a complementary link with frequentist testing is presented which considers the actual density of the p-values when either hypothesis is true. Such testing procedures are presented for both simple and composite alternative hypotheses, hence are particularly suited for testing non-zero parameters, i.e. when using statistical testing in an exploratory capacity, for example when investigating the presence of a zero or non-zero effect.

Chapter 2 discusses p-value fundamentals incorporating a detailed literature review. Here p-values are introduced as random variables, and their subsequent distributions are discussed. Having examined the properties of the p-value, it is evident that although it is an excellent summary statistic for representing the distance that data fall from a null hypothesis (and being restricted to the common unit interval for *all* test statistics), its namesake of a *probability* value must be taken into account. That is, no p-value can offer a *definitive* conclusion with regards to one or other of the hypotheses, merely that it should be treated as an indicator flagging interesting results which warrant further investigation. Nevertheless, even as an investigative tool the p-value is a most welcome statistic for empirical researchers as a key component in their inferential

arsenals.

Chapter 3 outlines a formal simultaneous hypothesis testing methodology which takes into account the plausibility of H_1 by considering a so-called second-order p-value, p'. Critical value tables in terms of conventional p-values are provided to allow for quick testing of H_1 for a given sample size and effect size. Derivation of the p-value distribution for t-distributed test statistics is presented, reflecting the common use of estimated variances in many testing situations. Attention focuses on the effect size and it is noted that when this takes a negative value (for simple forms of H_1) or a range of negative values (for composite forms of H_1, where the effect size follows a particular probability distribution), the p-value distribution under the alternative departs from its usually perceived shape. The implications of this for statistical inference are discussed.

Chapter 4 revisits the quest for a methodological unification of statistical testing. The conflict between p-values and conditional measures is highlighted and the adoption of the second-order p-value into the conditional frequentist approach is given. This helps to cement the different testing schools by retaining the desirable features of each. By defining a new conditioning statistic, p-value curves and surfaces are constructed to ease the decision-making process for practitioners.

Chapter 5 investigates multiple hypothesis testing. Compound error rates (false discovery and false nondiscovery rates) and their estimation are discussed. This centres around the true proportion of tested hypotheses for which the null is true. A new estimator for this proportion is proposed and its statistical performance relative to an estimator in the literature is investigated. It is shown that under certain conditions this new estimator is preferred. Chapter 6 concludes.

Chapter 2

Background to p-values

2.1 Statistical Testing

Statistical testing is employed across many fields as an elementary inferential procedure for choosing between two specified hypotheses using a set of observed sample data. From the social and behavioural sciences[1] to the biological sciences and beyond, use of this methodology is well-established and the reporting of test results has become standard throughout the empirical literature, having flourished over recent decades.

A designated null hypothesis, H_0, is set *a priori* and held to be true. This is tested against a prior alternative, H_1, which for simplicity is assumed to be one-sided, i.e. directional, at present.[2] Construction of an appropriate test statistic dependent on the integer sample size n, X_n, say, is carried out and its distribution under the null hypothesis (assuming necessary data conditions are satisfied) is obtained.[3]

[1] A recent review of developments in significance testing in Sociology, for example, can be found in Leahey (2005).

[2] For symmetrical test statistic distributions, a two-tailed H_1 poses few problems, discussed later.

[3] The distribution of X_n under H_0 is assumed to be known, either exactly or asymptotically.

At this point, from a pedagogical perspective and for completeness, it is appropriate to distinguish between 'hypothesis testing' and 'significance testing' as subsets of 'statistical testing' as clarified in Huberty (1987) who notes that some authors have failed to make explicit this distinction, for example Carver (1978) and Sawyer and Peter (1983), though not all as in Kempthorne (1976) among others.

Hypothesis testing has its roots in the Neyman-Pearson framework of classical statistics in which a prior (pre-experimental) *fixed* significance level, α, would be set during the test design phase. α therefore corresponds to the size of a test and probability of erroneously rejecting a true null hypothesis, i.e. a Type I error. Typically $\alpha = 0.05$, though other probabilities such as 0.01 or indeed 0.001 may be used when more stringent safeguards are warranted by the researcher to avoid the rejection of a true H_0.[4] Given α and the distribution of X_n under H_0, a critical region, C_α, which spans sufficiently extreme (i.e. improbable) values of the test statistic can be established[5] such that if the realised $x_n \in C_\alpha$, then our belief in H_0 becomes untenable, leading to the oft-cited phrase, 'we reject the null hypothesis in favour of H_1'.[6] Scientific journal articles routinely publish asterisks alongside reported x_n values as appropriate to indicate the level of significance, such as * for statistically significant ($0.01 \leq p < 0.05$) and ** for highly statistically significant ($p < 0.01$).

[4]The precise setting of α should depend on the purpose of the statistical testing. If purely an exploratory study, a liberal α should be set in order to detect potentially interesting results which demand further research. However if the study is the result of previous investigation(s), then a more conservative level should be used. Of course ultimately the choice of α is entirely discretionary. As noted in Alberoni (1962), the choice of α should correspond to the researcher's "threshold for the dismissal of the idea of chance" for that particular hypothesis. Note that the implicit subjectivity attached to α-setting counters the common criticism of Bayesian testing for its use of a subjective prior distribution. This issue is returned to in Chapter 4.

[5]That is, the test statistic distribution is partitioned into two regions: reject and fail to reject.

[6]Another issue is how strictly should a critical region be enforced? If x_n was within or outside of C by an arbitrarily small amount ε, what conclusion should be drawn? Labovitz (1968) discusses.

Significance testing is attributable to R. A. Fisher where the probability of obtaining a draw from X_n at least as extreme as the sample value is used in lieu of C_α, and therefore α. This probability is commonly referred to as the p-value and can be assigned to any observed sample test statistic value. The actual term 'p-value' can be traced back to Deming (1943) as noted in David (1998). The p-value acts as a summary statistic and is often cited in empirical studies to indicate the observed level of significance of tested parameters. Consequently p-values are more informative than a simple 'reject' or 'fail to reject' declaration as they allow the reader to attach his/her own subjective α as appropriate for the study being undertaken, see Kiefer (1977). This is beneficial when a researcher has no justifiable reason to favour any particular statistical significance level. Hence the p-value is a more useful data reduction technique and its prolific and popular adoption as a conventional quasi-sufficient statistic for interpreting hypothesis test results is testament of this.

In practice, applied inferential statistics is a mixture, or hybrid, of the two approaches. Stallings (1985) cautions against treating α and p-values synonymously, despite both camps of statistical tests being foremost probabilistic. An obvious distinction is that the Neyman-Pearson theory is used to obtain a qualitative decision (reject or not reject) while the Fisherian approach yields a significance level. Also, α is a pre-experimental researcher-fixed constant whereas p-values are random variables, sensitive to n under the alternative hypothesis.

2.2 P-values as Random Variables

Suppose X_n is used to test $\mathrm{H}_0 : \theta = a$ against $\mathrm{H}_1 : \theta = b$, $a < b$. Let X_n have the left-continuous distribution function $F_{X_n}(x_n) = \Pr(X_n \leq x_n)$ under H_0, for realised x_n. The p-value statistic, i.e. significance probability, is then a random variable, P, whose realised value, p, is calculated for continuous[7] test statistics as[8]

$$p = \Pr(X_n > x_n) = 1 - F_{X_n}(x_n) = \bar{F}_{X_n}(x_n), \tag{2.1}$$

hence p-values are one-to-one transformations of the random variable X_n, so are themselves random variables.[9] Pearson (1938) refers to this as the 'probability integral transformation' of the sample data. Advantages of the p-value include its simplicity, i.e. a single real number restricted to the unit interval, and also its universal application across test statistics with *any* distribution under H_0 due to the transformation in (2.1). Consequently the unit interval $[0, 1]$ is a common scale for comparison allowing meta-analyses to be performed. Among the numerous examples from social science research are Hedges (1980), Rosenthal (1984), Strube (1985) and Bornstein (1989), though clearly this list is by no means exhaustive.

2.3 P-value Distributions

Given either hypothesis could be true, p-values will have different distributions accordingly. As shown in many studies, under a non-composite null this is a uniform distribution for *any* continuous test statistic. Let p be the realised p-value

[7] For discrete distributions a correction may be required — see Cox (1977).

[8] This yields *exact* p-values, as opposed to *approximate* p-values which are computed by using an approximation of F_{X_n}, for example when F_{X_n} is unknown.

[9] As shown in (2.1), p-values can be viewed in terms of the survival function, $\bar{F}_{X_n}$.

with $F_P(p|\mathrm{H}_0)$ being the corresponding distribution function under H_0, then using (2.1),

$$\begin{aligned} F_P(p|\mathrm{H}_0) &= \Pr(P \leq p|\mathrm{H}_0) \\ &= \Pr(1 - F_{X_n}(x_n) \leq p|\mathrm{H}_0) \\ &= 1 - F_{X_n}(F_{X_n}^{-1}(1-p)) \\ &= 1 - (1-p) \\ &= p, \end{aligned} \tag{2.2}$$

for $p \in [0,1]$. Hence this gives a density function, $f_p(p|\mathrm{H}_0) = 1$, consistent with a uniform density. It should be noted that this density is independent of the test statistic distribution, sample size and effect size. Therefore under H_0 it is impossible to distinguish p-values obtained from small and large samples as well as between tests engineered to have high power and those less powerful.

Denoting the left-continuous distribution function of X_n under the alternative hypothesis by $G_{X_n}(x_n)$, then the p-value distribution function under H_1, $F_P(p|\mathrm{H}_1)$, becomes

$$\begin{aligned} F_P(p|\mathrm{H}_1) &= \Pr(P \leq p|\mathrm{H}_1) \\ &= \Pr(1 - F_{X_n}(x_n) \leq p|\mathrm{H}_1) \\ &= 1 - G_{X_n}(F_{X_n}^{-1}(1-p)), \end{aligned} \tag{2.3}$$

which clearly depends on the test statistic's distribution under both hypotheses. The corresponding density function is merely the likelihood ratio for the hypothesis

test, since by application of the chain rule

$$f_P(p|\mathrm{H}_1) = \frac{\partial}{\partial p} F_P(p|\mathrm{H}_1) = \frac{g_{X_n}(F_{X_n}^{-1}(1-p))}{f_{X_n}(F_{X_n}^{-1}(1-p))}, \tag{2.4}$$

where the densities of X_n under the null and alternative are $f_{X_n}(x_n)$ and $g_{X_n}(x_n)$ respectively. Equation (2.4) provides the density for generic test statistics, while Hung, O'Neill, Bauer, and Köhne (1997) derive this for the case of a standard t test of a sample mean assuming a Gaussian response variable. Assuming simple (i.e. non-composite) forms of H_0 and H_1, they show this to be

$$f_P(p|\mathrm{H}_1) = g_\delta(p) = \frac{\phi(Z_p - \sqrt{n}\delta)}{\phi(Z_p)}, \qquad 0 < p < 1, \tag{2.5}$$

where $\delta = \mu/\sigma$ denotes the effect size, n the sample size, ϕ is the standard Gaussian density and Z_p its $(1-p)^{th}$ percentile.[10] It is clear therefore that for t tests under the alternative, p-value densities are dependent on both the sample size and the effect size. Hung, O'Neill, Bauer, and Köhne (1997) proceed to investigate how the distribution changes as n and δ are varied. Clearly test power increases with n, hence $f_P(p|\mathrm{H}_1)$ becomes steeper (similarly when $|\delta|$ increases) which means $\Pr(P < \alpha|\mathrm{H}_1)$ increases. As a result the expected value and variance of P, $E[P]$ and $\mathrm{Var}(P)$ respectively, decrease under H_1. Note that under the null $\mathrm{H}_0 : \mu = 0$, (2.5) reduces to 1, that is the uniform density. This concept will be revisited in Chapter 4 when a methodological unification of the Fisherian, Neyman-Pearson and Jeffreys' approaches to testing is presented.

An interesting field which utilises p-values is multiple hypothesis testing. In such

[10] Invoking the Central Limit Theorem, the sample mean of data drawn from any distribution is asymptotically Gaussian, with a resulting standard Gaussian test statistic.

instances there is likely to be considerable heterogeneity amongst the p-values when the alternative hypotheses are true, suggesting scope for a suitable approximating methodology. Previous work concerning (2.4) focusing on single testing includes Bahadur (1960) who used a lognormal distribution to approximate the finite sampling distribution. Extensions of this work can be found in Lambert and Hall (1982), who present sufficient conditions for asymptotic lognormality, and Becker (1991) who studies non-null asymptotic distributions of various functions of p-values from one-tailed t tests. Multiple testing will be revisited later in Chapter 5.

2.4 Null Hypothesis Types

The null hypothesis, H_0, is customarily thought of as being either a pointwise null, for example $\mathrm{H}_0 : \mu = \mu_0$, or as a one-sided null, $\mathrm{H}_0 : \mu \leq \mu_0$ or $\mathrm{H}_0 : \mu \geq \mu_0$, with two-sided and one-sided alternatives, $\mathrm{H}_1 : \mu \neq \mu_0$ and $\mathrm{H}_1 : \mu > \mu_0$ or $\mathrm{H}_1 : \mu < \mu_0$, respectively. Although at first sight these appear distinct forms, Schervish (1996) shows that the two types lie at opposite ends of a continuum of null hypotheses bridged by so-called 'interval nulls'.

Such interval hypotheses are an intermediate species where the parameter of interest is reckoned to be in a particular range bounded by lower and upper values, μ_l and μ_u respectively, for example $\mathrm{H}_0 : \mu \in [\mu_l, \mu_u]$. The corresponding alternative will be the complement of this interval with respect to the parameter space Ω, namely $\mathrm{H}_1 : \mu \in \Omega \setminus [\mu_l, \mu_u]$. It is then evident that the pointwise form is the limit of an interval null as μ_l and μ_u converge, while a one-sided null is formed as either μ_l or μ_u tends to minus or plus infinity for lower- and upper-sided nulls respectively.

Schervish (1996) continues to show that under the assumption that all tests

considered are uniformly most powerful and also unbiased, p-values are continuous. For a Gaussian test statistic with known variance σ^2/n, such as $\bar{X}_n \sim N(\mu, \sigma^2/n)$, p-values of the realised $\bar{x}_n$ under H_0 with the various types of null hypothesis are therefore computed as,

$$\begin{aligned}
\mathrm{H}_0 : \mu = \mu_0 \quad &: \quad p = 2 \times \Phi\left(-\frac{\sqrt{n}(|\bar{x}_n - \mu_0|)}{\sigma}\right), \qquad \text{(two-tailed test)} \qquad (2.6)\\
\mathrm{H}_0 : \mu \geq \mu_0 \quad &: \quad p = \Phi\left(\frac{\sqrt{n}(\bar{x}_n - \mu_0)}{\sigma}\right),\\
\mathrm{H}_0 : \mu \leq \mu_0 \quad &: \quad p = \Phi\left(\frac{\sqrt{n}(\mu_0 - \bar{x}_n)}{\sigma}\right),\\
\mathrm{H}_0 : \mu \in [\mu_l, \mu_u] \quad &: \quad p = \begin{cases} \Phi\left(\frac{\sqrt{n}(\bar{x}_n - \mu_l)}{\sigma}\right) + \Phi\left(\frac{\sqrt{n}(\bar{x}_n - \mu_u)}{\sigma}\right) & \text{if } \bar{x}_n < 0.5[\mu_l + \mu_u] \\ \Phi\left(\frac{\sqrt{n}(\mu_l - \bar{x}_n)}{\sigma}\right) + \Phi\left(\frac{\sqrt{n}(\mu_u - \bar{x}_n)}{\sigma}\right) & \text{if } \bar{x}_n \geq 0.5[\mu_l + \mu_u] \end{cases},
\end{aligned}$$

where Φ is the standard Gaussian distribution function. The p-value formulae for interval nulls make use of the result in Lehmann (1986) which states that such a null is rejected if $\sqrt{n}(|\bar{x}_n - 0.5(\mu_l + \mu_u)|)/\sigma > c$, for critical value c such that $\Phi(0.5\sqrt{n}[\mu_l - \mu_u]/\sigma - c) + \Phi(0.5\sqrt{n}[\mu_u - \mu_l]/\sigma - c) = \alpha$. Hence c is a function of both α and $(\mu_u - \mu_l)$, i.e. $C_{\alpha,(\mu_u - \mu_l)}$.

It is then established by Schervish (1996) that pointwise and one-sided null hypotheses are indeed special cases of the interval null. This also extends to the respective p-values, that is p-values of pointwise and one-sided nulls are limits of interval null p-values. This result stems from the continuity of p-values and it is noted that this is not just restricted to Gaussian test statistics, but other distributions including uniform and exponential densities.

The beauty and relevance of this result concern testing the significance of factors such as establishing whether a zero or non-zero effect exists. In clinical trials, asset pricing theories and other deterministic models, we frequently desire to

know whether a hypothesised explanatory variable does have a significant non-zero effect on some response variable. Non-zero coefficients in linear regression models are an example of the modelled effect on a real-world response.

Although any non-zero effect is an effect by definition, if the effect is very small, say $\pm\varepsilon$, then it still may not be worthwhile to view the explanatory factor as significant in a practical sense once a feasibility study or cost-benefit analysis has been carried out to assess the viability of acting on the statistical result of an effect. Potential cases when it would not be appropriate to act on sufficiently small non-zero effect sizes include:

1. The financial cost of manufacturing/supplying a new drug responsible for a marginal improvement in the treatment of a particular disease crowds out the marginal benefit to the patient or society of the new treatment (in terms of increased life expectancy, say).

2. When transaction costs eclipse the expected marginal gain in asset price movements from trading on the basis of an estimated pricing model.

Knowing the direction of an effect is insufficient to conclude that intervention is cost-effective. Size of an effect is what matters. Admittedly, in medical situations, there is an *ethical* obligation to resort to any strictly superior treatment if human longevity can be improved, even if only marginally. Of course pragmatism dictates that the economic marginal costs must influence the ultimate decision.[11]

So when testing for non-zero significance in the explanatory power of a given factor, it would be appropriate to construct an interval null hypothesis such as

[11] In the UK, the National Institute for Health and Clinical Excellence (NICE) is a special health authority charged with assessing whether the health benefits of new treatments offset the financial costs. Formally the quality-adjusted life-year (QALY), which measures an extra year of life after discounting for any disability, pain or other impairment, is used as the decision rule with a cap currently of £30,000 for treatment approval.

$H_0 : \theta \in [-\varepsilon, \varepsilon] \equiv \theta \leq |\varepsilon|$.[12] In practice, *exact* point nulls are rare, instead small interval nulls are more realistic. The limit results referred to above imply that a pointwise null, i.e. $H_0 : \theta = 0$, acts as a good approximation to the interval version as argued by Zellner (1984) and justification of this in both the Classical and Bayesian[13] worlds is offered in Berger and Delampady (1987), hence is suitable to use in clinical trials and related factor significance tests because a single parameter value in H_0 produces a single p-value density, the continuous uniform distribution, thus simplifying matters. Of course, the smaller $|\varepsilon|$ the better the approximation. The distinction between statistical and practical significance has a long history, for example Berkson (1938), Berkson (1942) and Hodges and Lehmann (1954).

2.5 Interpretation of P-values

As already noted above, statistical testing of hypotheses using the Neyman-Pearson theory has its deficiencies due to the arbitrary nature of setting the pre-experimental significance level of the test, α. True, the hypothesis testing methodology has a worthy merit, namely by constructing a test which controls the Type I error through α then the sample size n can be chosen to achieve the desired test power, equivalently reduce the probability of a Type II error, subject to α. Such an approach bypasses the interrelationship between the two types of inferential decision error.

[12] θ represents the parameter under investigation, such as a population mean, μ, or a regression coefficient, β_i, for factor i in a multiple linear regression. Here the interval null limits are equi-distant from zero. This symmetry corresponds to the case where a practical significant non-zero effect is associated with a true θ exceeding zero by some *absolute* arbitrarily small amount ε. If the departure from zero for practical relevance depended on whether θ was positive or negative, then the interval null limits would not be equi-distant. For example, if twice the positive effect is required vis-à-vis the negative effect, then $H_0 : \mu \in [-\varepsilon, 2\varepsilon]$. An economic interpretation would relate to non-constant returns to scale, for example.

[13] The prior density, $\pi(\theta)$, is assumed continuous but with a sharp spike near the value θ used in H_0.

In addition to the *ad hoc* nature of α selection, another disadvantage is that the designated significance level may not actually be fully attainable. This is potentially the case when dealing with discrete parametric distributions under H_0, but also in the non-parametric world. In the latter case, lack of information regarding distributions under the alternative hypothesis precludes consideration of Type II errors.

The p-value is a strictly superior method, since it contains more information about the validity of H_0. For a directional (one-sided) H_1, the p-value of x_n represents the upper or lower (depending on the specification of H_1) tail probability of the test statistic's sampling distribution for observing x_n, or a more extreme result. This was formalised in (2.1).

Extending the Neyman-Pearson methodology, it is noted that $\alpha \in [0, 1]$, hence has a lower and upper bound. A p-value can therefore be interpreted as the minimum lower bound such that H_0 would be rejected. This reflects the "level attained by the sample" terminology in Gibbons and Pratt (1975).

As mentioned, p-values have the advantage of being treated as subjective αs. With this in mind, just as an α-based critical region offers no *definitive* conclusion as to the significance of a test result, p-values should just as equally be viewed as indicative. In fact, empirical research findings are essentially evaluated in terms of their practical (in medical statistics, clinical) rather than statistical significance. Therefore p-values should act as a *partial* aid when forming consequential decisions, hence are ideal for exploratory analyses.

In conjunction with the significance probability, other factors to be taken into account include the size of datasets used, robustness of the sampling procedure implemented as well as its suitability for the application in question. Such issues are clearly statistical in nature, whereas whether there are sufficient resources to

collect good quality data, indeed whether the data are even representative of the target population, are important factors which affect whether inference is reliable, valid and worthwhile. Hence as a guide/indicator p-values are extremely useful, but should not be the sole basis for practical or clinical inference when definitive answers are sought.

2.5.1 Measure of support

Several authors have considered p-values as posterior probabilities. This is understandable since test statistics, hence p-values, are computed once the data have been observed, and posterior probabilities are used to measure the support for hypotheses once data are collected. DeGroot (1973), Casella and Berger (1987) and Berger and Sellke (1987) all explore this view.

By thinking of p-values as posterior probabilities, it is intuitive to view such probabilities as measures of support for the tested hypotheses. The smaller the p-value, the more evidence against H_0, while larger p-values provide greater support. In terms of the actual observed data (since p-values correspond to test statistics, themselves functions of the data), the greater the distance between the sample statistic and its expected value under H_0, the more likely we would be to reject H_0. So for a two-tailed test for population parameter $\theta \in \Theta$, with sample estimator $\hat{\theta}$, as a function of $\hat{\theta}$ the p-value increases as $|\hat{\theta} - \theta|$ decreases, i.e. as the observed statistic comes in closer proximity to its expected value under H_0.

In addition, it would be expected that as H_0 covers a greater proportion of the parameter space, Θ, then a given dataset would provide more compelling support in favour of H_0. If $\Theta \equiv \mathbb{R}$, for a one-sided null $\mathrm{H}_0 : \theta \leq \theta_0$, as θ_0 increases such that the alternative encompassing (θ_0, ∞) is squeezed, the p-value becomes increasing in

θ_0.

Despite the desirable appeal of treating p-values as measures of support of a null hypothesis, Schervish (1996) shows that if significance probabilities were indeed measures of support, then a simple logical condition involving the breadth of the parameter space covered by the null would be satisfied. However the logical condition in question, which involves the coherence of p-values, fails to hold.

Coherence in simultaneous testing is covered in Gabriel (1969). For coherence to hold, if one null is implied by another null, for example $H_0 : \theta \leq \theta_0$ is implied by $H_0' : \theta = \theta_0$[14], then rejection of the implied hypothesis guarantees rejection of the other. Hence the measure of support for the implied hypothesis must be at least as great as the measure of support for the other. Schervish (1996) offers examples which show that p-values are in fact incoherent whereby the p-value of a null hypothesis whose parameter range is a proper subset of the parameter range covered in the other null — using the above, the parameter value in H_0' is a proper subset of that in H_0 — is *greater*. In other words, this suggests we would be more confident that $\theta = \theta_0$ rather than the more liberal $\theta \leq \theta_0$.

Given pointwise null and one-sided null hypotheses are rarely considered simultaneously, such incoherence is likely to be ignored by most researchers. However, since both types of null are special cases of interval nulls, they *can* be compared, hence this lack of coherence is relevant, and so p-values should not be interpreted as measures of support. Consequently, we should restrict the interpretation of p-values to the role of indicator for potentially interesting results which require further investigation. However, the ratio of measures of support with respect to H_0 and H_1 will be considered later.

[14] A proper subset implies the parent set, not vice versa.

2.6 Note on Two-sided Tests

Thus far p-values have been characterised as the probability of a test statistic value at least as extreme as that observed. For one-sided tests, i.e. directional alternative hypotheses, visualisation of this concept is straightforward. However, two-sided tests are appropriate when there is no prior justification for the nature of the directional departure from H_0.

Gibbons and Pratt (1975) offer some ways to accommodate two-sided alternatives when reporting p-values. Briefly, these include simply reporting the one-tailed p-value associated with the actual test statistic outcome plus *some* probability reflecting the other tail. Of course, such an approach leads to a continuum of potential p-values since the other tail probability to add is entirely subjective.

However if the test statistic distribution under H_0 is symmetric, it is natural to report twice the one-tailed p-value as shown in the first equation of (2.6). This corresponds to conventional hypothesis testing in which a test of size α is structured into two equal-sized critical regions, each of area $\alpha/2$. For asymmetric distributions, partition in such a way such that the most unlikely values form the critical region.

The only caveat with a doubling of the one-tailed p-value (in the case of symmetric test statistic distributions) occurs in the case of discrete distributions when the final p-value is not a specific, realisable probability given the null distribution, or indeed the final p-value could exceed one. To circumvent such a problem, construction of the reported p-value should comprise a combination of the one-tailed value plus an attainable probability representing the complementary tail, with the constraint of the sum being no greater than one.

Since statistical testing is based on the notion of the extent of departure from a null hypothesis, this equates to the difference between the observed test statistic value and its expected value, given H_0. In most instances this measure of central tendency is the mean, however other location measures could be considered, specifically the median or mode. Symmetry is a sufficient condition for the mean and median to be equivalent, though not necessarily for the mode, for example continuous uniform densities have a continuum of modes.

Two-sided p-values could therefore be constructed as the one-sided value plus the probability of being equi-distant from the desired location measure in the opposite direction. Formally, for x_n above location parameter l, i.e. in the upper tail, the two-sided p-value would be computed as $p = \Pr(X_n > x_n) + \Pr(X_n < 2l - x_n)$. Similarly for x_n below l, $p = \Pr(X_n < x_n) + \Pr(X_n > 2l - x_n)$. This equates to the doubling of the one-sided p-value when l is the population mean under H_0 for symmetric distributions. For asymmetric distributions such an approach is legitimate when accompanied by an appropriate adjustment for skewness.

2.7 Effect of Sample Size on Interpretation

So far the sample size n has only been considered (besides its role in computing the test statistic, X_n, and test power) with regard to its effect on the p-value density function as highlighted in (2.5). Royall (1986), however, considers the contradictory interpretations by previous authors concerning the impact of n on the strength of evidence against H_0 provided by p-values.

Most practitioners treat the magnitude of p-values as an indication of the falsity of H_0, with smaller values interpreted as providing stronger evidence. As a

consequence of this, the α-postulate of Cornfield (1966) states that equal p-values provide (approximately) equal weights of evidence for or against H_0. Once n is taken into account, this postulate is challenged. It is noted that Lindley and Scott (1984) believe that statistically significant results at 5% carry greater weight against H_0 the smaller the trial size. This contrasts with Peto et al. (1976) who declare that for a given p-value, larger trials afford stronger evidence. In fact both views are accurate when distinguishing between *specific* p-values and merely *significant* p-values.

Let both hypotheses be simple when testing some arbitrary parameter γ, say, such that $H_0 : \gamma = \gamma_0$ and $H_1 : \gamma = \gamma_1$. Also, let the appropriate test statistic be defined for small samples, X_S, and for large samples, X_L, with corresponding densities under H_0 and H_1 being $f_0^S(x_S)$, $f_1^S(x_S)$, $f_0^L(x_L)$ and $f_1^L(x_L)$ for small (S) and large (L) trials respectively. Define $\tilde{x}_S$ under H_0 such that $\Pr_0(X_S \geq \tilde{x}_S) = \alpha$ and similarly define $\tilde{x}_L$ under H_0 such that $\Pr_0(X_L \geq \tilde{x}_L) = \alpha$. Without loss of generality, assume the statistical test is upper-tailed such that $\gamma_0 < \gamma_1$, therefore $\tilde{x}_L < \tilde{x}_S$. This is the case since increasing the sample size will reduce the variance of the sampling distributions under both the null and alternative hypotheses.

We can proceed by generalising Royall (1986) by considering two sequences of trials differing by sample size. Let the (arbitrary) prior probability ratio $\Pr(H_0)/\Pr(H_1) = k$. Ideally, significant results are attributable to true discoveries, rather than false discoveries (i.e. that H_1 is true rather than a Type I error being committed). Let "Sig." denote a significant test statistic, then Bayes' Theorem yields,

$$\frac{\Pr(H_0|\text{Sig.})}{\Pr(H_1|\text{Sig.})} = \frac{\Pr(\text{Sig.}|H_0)\Pr(H_0)/\Pr(\text{Sig.})}{\Pr(\text{Sig.}|H_1)\Pr(H_1)/\Pr(\text{Sig.})} = \frac{\alpha}{\beta}k, \tag{2.7}$$

where β is the test power, namely the ability of the test to reject a false H_0. Larger values of the ratio given in (2.7) represent a greater danger of misleading results.

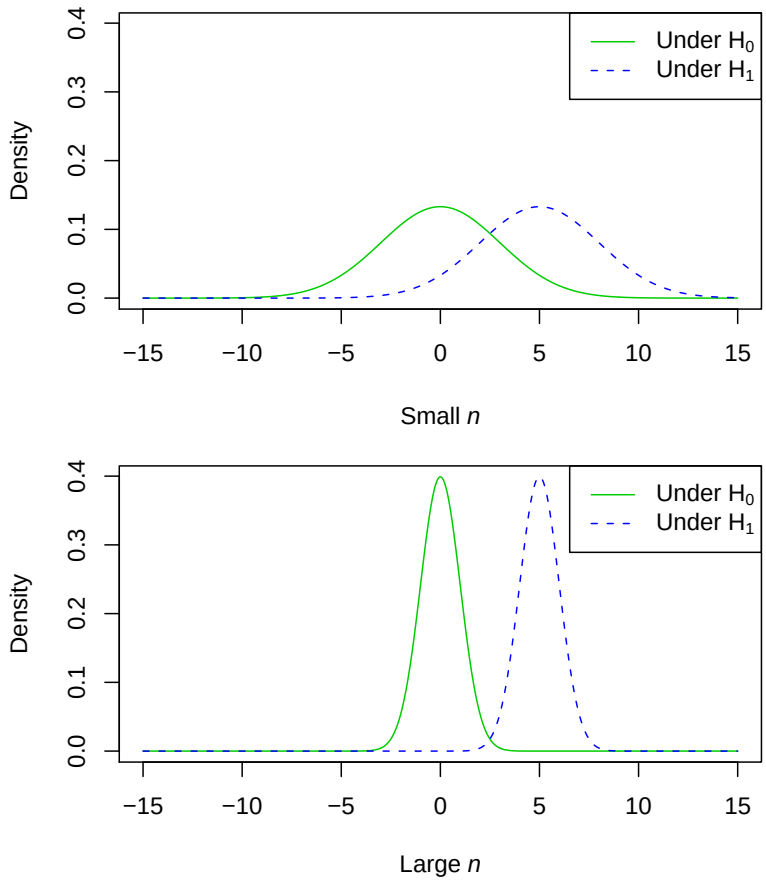


Figure 2.1: Arbitrary test statistic densities under H_0 (mean 0) and H_1 (mean 5) using small and large sample sizes.

Specifically, when $\frac{\alpha}{\beta}k = 1$, then this means that a significant result is as likely to be a false discovery (Type I error) as it is to be a genuine discovery. In multiple testing, this translates into a false discovery rate of 50%. $\frac{\alpha}{\beta}k > (<)1$ means significant results are more (less) likely to be false discoveries.

For sequences of small and large independent trials, k and α are fixed *a priori*, leaving the power, β, to determine the value of the above ratio. Consequently, increasing the sample size will increase β, hence reducing (2.7). This corresponds to fewer misleading significant results (fewer false discoveries), a desirable outcome, hence implies that larger sample sizes provide stronger evidence of the falsity of H_0

which is consistent with the Peto et al. (1976) view. So as expected, larger samples increase inferential accuracy.

Figure 2.1 illustrates sample mean examples for the densities $f_0^S(x_S) = N(0, 9)$, $f_1^S(x_S) = N(5, 9)$, $f_0^L(x_L) = N(0, 1)$ and $f_1^L(x_L) = N(5, 1)$ such that $\gamma_0 = 0$ and $\gamma_1 = 5$. Therefore for $\alpha = 0.05$, $\tilde{x}_S = 4.935$ and $\tilde{x}_L = 1.645$. It follows that under H_1 for small samples, $\beta_S = \Pr_1(X_S \geq \tilde{x}_S) = 1 - \Phi\left(\frac{4.935-5}{3}\right) = 0.5086$. Similarly for large samples, $\beta_L = \Pr_A(X_L \geq \tilde{x}_L) = 1 - \Phi\left(\frac{1.645-5}{1}\right) = 0.9996$.

To be consistent with the Peto et al. (1976) view that large trials yield stronger evidence, let the prior ratio $k = 5$, so in each sequence of independent trials 5/6 are true under H_0, hence 1/6 of trials are true under H_1. Application of (2.7) gives for small samples a ratio of $5 \times 0.05/0.5086 \approx 0.5$, and for large samples a ratio of $5 \times 0.05/0.9996 \approx 0.25$. Therefore, for this parametric specification, approximately 33%[15] of significant results are attributable to false discoveries for small samples; however this reduces to around 25%[16] for large samples. Note that this superiority of large samples is independent of the choice of the prior ratio k and significance level α.

How does this reconcile with the apparently differing views of Peto et al. (1976) who advocate large samples, and Lindley and Scott (1984) who propose small samples? The above likelihood ratio methodology is appropriate for *significant* results, regardless of the specific p-values of individual test statistics. This approach requires rationing the information provided by p-values into an indicator variable transforming p-values greater than α into zeroes, and those less than α into ones.

By considering *specific* p-values only, the view of Lindley and Scott (1984) becomes clear. (2.7) requires adjustment such that instead of clustering significant

[15] $k\alpha/\beta = 1/2$ means 1 in 3 significant results are actually true under H_0, i.e. false discoveries.
[16] $k\alpha/\beta = 1/4$ means 1 in 5 significant results are actually true under H_0, i.e. false discoveries.

results together, only cases when the p-values of x_S and x_L, p_{x_S} and p_{x_L} respectively, are both equal to α are considered. For small and large samples respectively, this gives

$$\frac{\Pr(\mathrm{H}_0|p_{x_S}=\alpha)}{\Pr(\mathrm{H}_1|p_{x_S}=\alpha)} = \frac{\Pr(p_{x_S}=\alpha|\mathrm{H}_0)\Pr(\mathrm{H}_0)/\Pr(p_{x_S}=\alpha)}{\Pr(p_{x_S}=\alpha|\mathrm{H}_1)\Pr(\mathrm{H}_1)/\Pr(p_{x_S}=\alpha)} = \frac{f_0^S(x_S)}{f_1^S(x_S)}k, \tag{2.8}$$

$$\frac{\Pr(\mathrm{H}_0|p_{x_L}=\alpha)}{\Pr(\mathrm{H}_1|p_{x_L}=\alpha)} = \frac{\Pr(p_{x_L}=\alpha|\mathrm{H}_0)\Pr(\mathrm{H}_0)/\Pr(p_{x_L}=\alpha)}{\Pr(p_{x_L}=\alpha|\mathrm{H}_1)\Pr(\mathrm{H}_1)/\Pr(p_{x_L}=\alpha)} = \frac{f_0^L(x_L)}{f_1^L(x_L)}k. \tag{2.9}$$

The relative magnitudes of (2.8) and (2.9) clearly depend on the density ratios evaluated at the test statistic values (x_S and x_L) such that their p-values are equal to α. For all sample sizes, as illustrated in Figure 2.1 with $x_S = 4.935$ and $x_L = 1.645$, the relation

$$\frac{f_0^S(x_S)}{f_1^S(x_S)} < \frac{f_0^L(x_L)}{f_1^L(x_L)} \tag{2.10}$$

holds. Therefore smaller sample sizes lead to a smaller proportion of false discoveries because the relative probability of H_0 is lower (equivalently the relative probability of H_1 is higher), thus supporting Lindley and Scott (1984).

Where does this leave us? Well, the α-postulate that equal p-values provide equal weight against H_0 can only be valid when comparing p-values when the sample sizes used in each test are the same. If not, then the weightier evidence belongs to the larger sample test when p-values are purely categorised as significant or not, otherwise the smaller sample test has greater credence when comparing *exact* p-values.

As noted in Birnbaum (1962), in general the likelihood function offers a better indicator of evidence against H_0 rather than p-value functions. Recall that p-values are calculated under H_0 and so a rare event as indicated by a small p-value does

not suggest *per se* that such an event is common under H_1. Consequently the test power is key to assessing the link between p-values and departures from the null hypothesis.

The effect of sample size on statistical testing is obviously an important issue in meta-analyses. Such consolidation of previous study results is not the only application of these sample size effects. Multiple hypothesis testing, that is simultaneously testing several hypothesis, can lead to false discoveries. However for study comparisons this issue can be circumvented by fixing n across all tests.

2.8 Summary

In this chapter, we have pursued a comprehensive literature review of the popular p-value. Given its prominence in modern statistical testing across the spectrum of empirical research, it is important to understand and properly interpret its true meaning.

It has been shown that p-values are random variables (2.1), and as such they have different distributions corresponding to the null and alternative hypotheses denoted in (2.2) and (2.3) respectively. Since the uniform density of p-values under H_0 is fixed across all continuous test statistics regardless of their distributions, then exploitation of this property can be extremely useful.

The remaining chapters chiefly focus on using the p-value distributions under both hypotheses. Central to this approach is the appreciation that improbable null hypotheses do not guarantee that the alternative hypothesis is more feasible. Hence by accommodating this concept, it is possible to improve the decision-making process. Recognising the fallibility of the real world, inferential mistakes will always

occur. However, by restricting their occurrence and fully understanding the potential extent of the conclusiveness of statistical tests, we can be better informed when interpreting results.

Chapter 3

Purifying p: Extracting Maximum p-value Information

P-values, *in practice*, are treated as a measure of evidence against H_0 with small values, typically those less than a designated significance level conventionally taken to be $\alpha = 0.05$, being deemed statistically significant — that is, based on the data there is sufficient evidence to reject the null in favour of the alternative as the probability of observing test statistic value x_n or a more extreme value, given H_0 is true, is just p. However significance probabilities are purely a measure of how far the observed data fall from the null hypothesis, although the temptation for some is to interpret large p-values as automatic justification for the acceptance of H_0, or outright rejection of H_1. Instead confidence intervals are preferred when faced with insignificant test results, see Gardner and Altman (1986), yet p-values are still popular and widely reported.[1]

However, as has been previously discussed, p-values are incoherent and technically should not be used as measures of support. That said, due to the

[1] As previously argued, this popularity of use is justified when performing exploratory analyses.

widespread use of p-values across many fields of empirical research, it is prudent to extract the maximum value of the information contained in any significance probability. Donahue (1999) pursues this view by considering not only the extent of the data's deviation from the null, but also the likelihood of the data given the alternative hypothesis for a particular example involving simple hypotheses. This concept is especially important since an improbable event under H_0 does not by default mean a likely event under H_1.

It is noted that by convention Type I errors are considered more intolerable than their Type II counterparts,[2] hence most researchers control the former using α and increase n to limit the latter, that is by increasing the test power. Whenever a Type I or Type II error is committed, this is explicable by the data being improbably, though it must be stressed *not impossibly*, extreme by pure chance alone.[3] Of course Type I errors equate to the reported p-values, though computation of the probability of a true Type II error requires accurate specification of the alternative hypothesis. Alternatively a power function can be derived.

The contribution to the academic literature included in this chapter comprises the following. Beginning with simple hypotheses, testing for example a population mean, a simultaneous testing methodology is presented in the classical and Fisherian moulds taking into consideration traditional inferential errors as well as analogous concepts when testing H_1. To achieve this, new terminology is used. Pursuit of the classical approach leads to the advent of critical value tables in terms of conventional p-values which allow the simultaneous testing of H_1 without the need to compute

[2]Choice regarding the precedence of decision errors is likely to be situation-dependent. As such a cost-benefit analysis of these inferential decision errors might be appropriate as discussed in Nagel and Neef (1977).

[3]Remember, events with very small probabilities do occur — just ask lottery jackpot winners. Of course the flip-side of this are the many millions of players who do not win.

so-called 'second-order' p-values which in essence are the p-values associated with the p-value when used as a test statistic for testing H_1.

In order to produce such tables, the p-value density under the alternative hypothesis is required. This has previously been derived for standard Gaussian-distributed test statistics, however here this is extended to Student t-distributed test statistics reflecting the frequent use of estimated variances. Obviously t distributions are asymptotically standard Gaussian in the degrees of freedom, but for small sample studies use of the t distribution is necessary. As a result critical value tables are also produced for various degrees of freedom.

3.1 Second-order p-values

Given two p-values both greater than α, say 0.4 and 0.8, although both 'lend support' to H_0 under classical hypothesis testing, does it mean that the larger value offers stronger evidence in favour of H_0? This is an important issue, since most empirical studies are only concerned about whether the null can be rejected and so ignore test power and the consequences of H_1. To answer this, Donahue (1999) considers not only how far the data fall from the null hypothesis, but also how far the data fall from a *specific* alternative hypothesis. To achieve this, reporting two summary statistics, the original p-value and a 'second-order' p-value, p', defined below, is required. Both act as quasi-*post hoc* risk levels indicating certain inferential decision errors.

Utilising (2.3), for the general case covering all possible test statistic

distributions,

$$\begin{aligned} p' &= \Pr(P > p|\mathrm{H}_1) \\ &= 1 - \Pr(P \leq p|\mathrm{H}_1) \\ &= G_{X_n}(F_{X_n}^{-1}(1-p)), \end{aligned} \tag{3.1}$$

from which an upper tail rejection region (for H_1) can be obtained, for a given significance level, say $\gamma = 0.05$. Consequently, instead of basing inference solely on H_0 as is frequently the case, assessments on the merits of both hypotheses can be presented namely:

i. p provides a summary statistic measuring the deviation of the data from H_0

ii. p' provides a summary statistic measuring the deviation of the data from H_1.

Using this methodology, we are in essence testing both hypotheses simultaneously. Recall from (2.2) that under H_0 all p-values have a uniform distribution over $[0, 1]$. Meanwhile the density under H_1 is given in (2.4) which will be, sometimes, heavily positively skewed as smaller p-values are more likely under the alternative hypothesis. Graphically this is illustrated in Figure 3.1 for the testing of a population mean using two simple hypotheses.

Just as the test level α is entirely subjective, so too is the benchmark level γ used to form decisions about H_1 based on the computed p'-values. Although there is no theoretical necessity to force the condition that $\alpha = \gamma$, it would seem sensible and logical to set these levels the same, say 5%. This would then ensure that the probability of erroneously rejecting the null and alternative hypotheses remains equal.

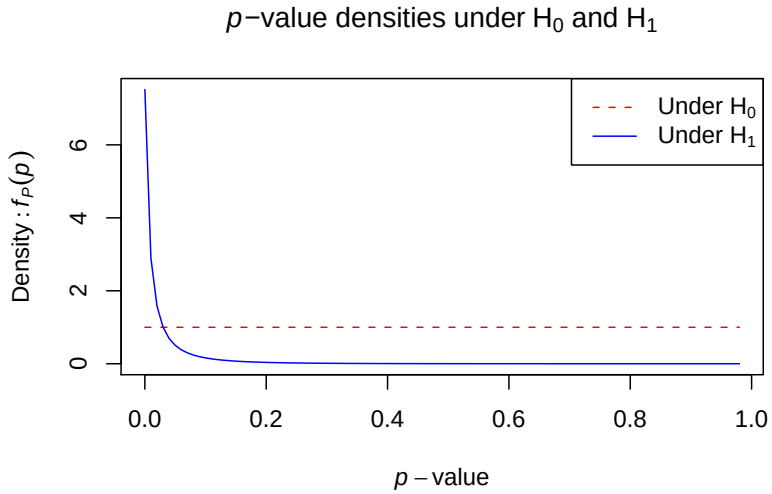


Figure 3.1: P-value densities under both $\mathrm{H}_0 : \mu = 0$ (uniform distribution) and $\mathrm{H}_1 : \mu = 2$ (ratio of two Gaussian densities as per (2.5)) such that $\sqrt{n}\delta = 3.5$.

Table 3.1: Outcome scenarios for the twin hypothesis testing of the null, H_0, and alternative null, H_1, for the exhaustive combinations listed.

	Reject H_0	Not reject H_0	Reject H_1	Not reject H_1
H_0 true	Type I error	Correct decision	Correct decision	Type IIa error
H_1 true	Correct decision	Type II error	Type Ia error	Correct decision

Generalising the ideas presented in Donahue (1999) to create a testing methodology reflecting the traditional classical conventions, I propose to treat H_1 as an *'alternative null'*. The rationale for this stems from the performing of two separate significance tests, both based on the p-value of the test statistic *under the original* H_0. Just as in conventional statistical testing, when testing the alternative null it is still possible to make inferential decision errors, namely rejecting a true alternative null hypothesis and failing to reject a false alternative null hypothesis.

Table 3.1 presents the possible outcomes of this 'twin statistical testing' in terms of outcome errors. The left-hand-side corresponds to the usual possible decisions with respect to the rejection, or otherwise, of H_0, and includes the familiar Type I and Type II errors. The right-hand-side complements this with respect to the rejection, or otherwise, of H_1. As denoted in Table 3.1, inferential decision errors are still possible in this case, and to remain consistent with the traditional error names, I shall denote these as Type Ia and Type IIa errors respectively.[4]

[4]There have been many arguments advancing other types of hypothesis test error. Examples include the Type III error, or error of the third kind, of Mosteller (1948) which represents correctly rejecting H_0, but for the wrong reason. This idea was extended in Kaiser (1960) to cover two-tailed tests — for a two-tailed test, the alternative hypothesis specifies two directions, namely above and below the null value. A Type III error is the correct rejection of H_0, but inferring the incorrect direction of the effect. Such an inferential error could easily occur in practice, because most practitioners rejecting H_0 in favour of a two-sided alternative will automatically conclude the true direction is the one suggested by the sample. For example for a population mean, this might represent a sample mean significantly smaller than the null value (resulting in rejection of H_0) drawn from a population whose mean is greater than the null value.

This implies the need for a revised power definition, as advocated in Leventhal and Huynh (1996), that power should be the conditional probability of rejecting the null hypothesis and correctly identifying the true direction of the difference between the population value of the tested parameter and the null value. Consequently when performing three-choice tests, that is testing three hypotheses, i.e. the null, alternative $<$ null and alternative $>$ null, the power will be reduced when using this revised definition of power. Equivalently the number of observations required will be underestimated during the test design phase if not adjusting the power definition. However, the impact of a Type III error is likely to be small in practice.

Another candidate for 'error[s] of the third kind' is given in Kimball (1957) by providing the right answer to the wrong problem, while Raiffa (1968) suggests this corresponds to answering

It is known that a conventional Type I error (rejection of a true H_0) is set *a priori* to be the test size, α, since p-values are uniformly distributed under the null. With respect to the alternative null, the analogous decision error is the incorrect rejection of the alternative null, logically referred to here as a Type Ia error. Similarly a Type II error is the failure to reject the null when false, hence the alternative null equivalent is classified here as a Type IIa error, that is failure to reject H_1 when false. Equations (3.2) to (3.5) clarify these definitions for simple forms of H_0 and H_1.

$$\begin{aligned}
\Pr(\text{Type I error}) = \Pr(\text{Reject } H_0|H_0) &= \Pr(|X_n| > |c||H_0) && (3.2)\\
&= \Pr(P < \alpha|H_0)\\
&= \alpha \qquad (\text{as } P|H_0 \sim U[0,1]),\\
\Pr(\text{Type Ia error}) = \Pr(\text{Reject } H_1|H_1) &= \Pr(P > d|H_1) && (3.3)\\
&= \Pr(P' < \gamma|H_1)\\
&= \alpha',\\
\Pr(\text{Type II error}) = \Pr(\text{Not reject } H_0|H_1) &= \Pr(|X_n| \leq |c||H_1) && (3.4)\\
&= \Pr(P \geq \alpha|H_1)\\
&= \beta,\\
\Pr(\text{Type IIa error}) = \Pr(\text{Not reject } H_1|H_0) &= \Pr(P \leq d|H_0) && (3.5)\\
&= \Pr(P' \geq \gamma|H_0)\\
&= \beta'.
\end{aligned}$$

completing the wrong problem. Even a Type IV error is considered by Marascuilo and Levin (1970) which is the incorrect interpretation of a correctly rejected hypothesis. Though this has parallels with Mosteller (1948).

Note that the appropriate test statistic for Type I and Type II errors is the usual X_n, i.e. some direct function of the data. Then given α, a critical value, c, is determined. For example, the upper-tailed test of a standard Gaussian test statistic with $\alpha = 0.05$ has $c = 1.645$. A substitute method of course, as discussed previously, is to compare the p-value of the observed sample statistic, x_n, with α. Hence α has a dual role, namely:

i. as the test size (prior probability of a Type I error)

ii. as the critical value for rejection in the p-value world of testing.

However, when using second-order p-values to test the alternative null, the test statistic is no longer the random variable X_n but the random variable P (namely the conventional p-value), derived from X_n as given by (2.1). Given a simple (non-composite) H_1, the p-value density is known to be as per (2.4) in the general case, and by (2.5) for a simple t test (assuming known variance). Using the test statistic distribution functions F_{X_n} and G_{X_n} as appropriate, a critical value, $d \in (0, 1)$, can be computed such that if the observed p-value exceeds d we would reject H_1 (cf. $p > d$ and $|x_n| > |c|$). Again, there exists a parallel methodology, this time involving the p-value of the p-value test statistic, namely the 'second-order' p-value, p', which can be compared to the prior set tolerance, γ. Due to the close relationship between Type I and Type Ia errors, also Type II and Type IIa errors, the probabilities of Type Ia and Type IIa errors are denoted as α' and β' respectively.

3.2 Simultaneous Testing

Small p-values suggest implausible null hypotheses, similarly small p'-values suggest implausible alternative hypotheses. What are the possible implications when

performing these tests simultaneously? This depends on the skewness of $f_P(p|\mathrm{H}_1)$, the p-value density under H_1.[5] Let $\alpha = \gamma$. The critical p-value for the rejection of H_0 is always α because p-values are permanently uniform under the null. However, the critical p-value for rejection of H_1 will depend on the shape of the density $f_P(p|\mathrm{H}_1)$, with d a decreasing function of the (positive) skewness.

So we seek d, such that $\Pr(P > d|\mathrm{H}_1) = \gamma$. There are three scenarios to consider: (i) $d < \alpha$, (ii) $d > \alpha$ and (iii) $d = \alpha$. The inferential conclusions for these different cases are presented in Table 3.2. The critical values referred to as (α, d) only require computation of d values, since α takes the usual value of the desired size of the test. d values will depend on the test statistic distributions as identified in (2.4). Such critical values for Gaussian and t-distributed test statistics are presented later in this chapter.

Note that no mention is given to *accepting* a hypothesis, instead we restrict the terminology only to whether we can reject or fail to reject a hypothesis. This is because rejection of a specific hypothesis, either the null or the alternative, is solely due to the implausibility of the observed test statistic (x_n for testing H_0, p for testing H_1). All we can say is that the data suggest that the rejected hypothesis is implausible. In effect, we are rejecting a particular hypothesis in favour of anything which is *not* the rejected hypothesis. For example, an upper-tailed test with $\mu_0 < \mu_1$, the rejection of H_1 only infers that $\mu \neq \mu_1$. However given our objective is to form some conclusion, a 'net conclusion' is suggested in each viable case.

It follows from Table 3.2 that some awkward (i.e. inconclusive) outcomes can occur. When $d < \alpha$, on occasions we may reject *both* hypotheses. This equates to the implausibility of both H_0 and H_1. If the population parameter value specified

[5]Recall that this is a function of the test statistic densities under both the null and alternative hypotheses, f_{X_n} and g_{X_n} respectively.

Table 3.2: Inferential conclusions for the pair of significance probabilities, (p, p') when testing (H_0, H_1) with critical values (α, d) respectively such that H_0 is rejected when $p < \alpha$ and H_1 is rejected when $p > d$, equivalently when $p' < \gamma$. Three potential scenarios are considered: (i) $d < \alpha$, (ii) $d > \alpha$ and (iii) $d = \alpha$. α does not necessarily equal γ. '-' indicates impossible conditions.

(i) $d < \alpha$	$(d < p \leq 1) \equiv (0 \leq p' < \gamma)$	$(0 \leq p \leq d) \equiv (\gamma \leq p' < 1)$
$0 \leq p < \alpha$	Reject H_0, H_1 Net conclusion: Choose neither	Reject H_0, fail to reject H_1 Net conclusion: Choose H_1
$\alpha \leq p < 1$	Fail to reject H_0, reject H_1 Net conclusion: Choose H_0	- -

(ii) $d > \alpha$	$(d < p \leq 1) \equiv (0 \leq p' < \gamma)$	$(0 \leq p \leq d) \equiv (\gamma \leq p' < 1)$
$0 \leq p < \alpha$	- -	Reject H_0, fail to reject H_1 Net conclusion: Choose H_1
$\alpha \leq p < 1$	Fail to reject H_0, reject H_1 Net conclusion: Choose H_0	Fail to reject H_0, H_1 Net conclusion: Choose both

(iii) $d = \alpha$	$(d < p \leq 1) \equiv (0 \leq p' < \gamma)$	$(0 \leq p \leq d) \equiv (\gamma \leq p' < 1)$
$0 \leq p < \alpha$	- -	Reject H_0, fail to reject H_1 Net conclusion: Choose H_1
$\alpha \leq p < 1$	Fail to reject H_0, Reject H_1 Net conclusion: Choose H_0	- -

in H_1 is considered to be of minimum practical relevance (for example marginal benefit exceeds marginal cost, or marginal return not crowded out by transaction costs) then the simultaneous conditions of $p < \alpha$ and $p' < \gamma$ equate to a statistically significant result ($p < \alpha$), but not a practically relevant one ($p' < \gamma$).

When $d > \alpha$, it is possible for a p-value to be in neither critical region. In such instances it is necessary to carry out a randomisation, see Mood, Graybill, and Boes (1974), to ultimately decide upon one of the hypotheses. Essentially in practice this equates to a follow-up trial independent of the original sample, as repeat sampling is a natural consequence having obtained inconclusive results. However in the event that $d = \alpha$, we have an unambiguous choice of rejection since we are always in one, and only one, critical region for all possible p-values.

3.3 Practical Example

We turn now to the case of a test of a population mean. Let $H_0 : \mu = 0$ and a non-composite, hence directional, alternative hypothesis (which without loss of generality will be assumed to be upper-tailed) be $H_1 : \mu = k$, where $k > 0$ and represents the minimum required population mean value to ensure practical viability (clinical, economic etc.) of, and hence warrant adoption of, the real-world consequence corresponding to H_1 in the event of testing favouring this alternative. This could represent the minimum improvement in survival rates necessary to justify purchasing a new drug, or sufficient financial returns to offset transaction costs.

Let $\bar{Y}_n$ be the sample mean random variable, formed from n independent observations of Y, a Gaussian variable with mean μ and (assumed known) variance, σ^2. The test statistic $Z = \sqrt{n}\bar{Y}_n/\sigma$ has a standard Gaussian distribution

under H_0 with the p-value random variable, $P = 1 - \Phi(Z)$, having its usual interpretation of a Type I error probability if H_0 is rejected when actually true. Under the alternative hypothesis $Z \sim N(\sqrt{n}k/\sigma, 1)$. It is then possible to compute the second-order p-value, p', which is the probability of observing *at least* p, given H_1 is true.[6]

Applying the result of Hung, O'Neill, Bauer, and Köhne (1997), and given in (2.5), the distribution function of the p-value under H_1 is

$$F_P(p|H_1) = \int_0^p \frac{\phi(Z_x - \sqrt{n}\delta)}{\phi(Z_x)}\, dx = 1 - \Phi(Z_p - \sqrt{n}\delta), \tag{3.6}$$

where recall Z_p is the $(1-p)$-th percentile of the standard Gaussian distribution, and $\delta = k/\sigma$. Given the definition of the p'-value in (3.1) for generic test statistic distributions, for the upper-tailed z test,

$$\begin{aligned} p' = \Pr(P > p|H_1) &= 1 - F_P(p|H_1) \\ &= \Phi(Z_p - \sqrt{n}\delta). \end{aligned} \tag{3.7}$$

(3.7) can therefore be used to derive conditions between d and α in terms of the sample size n. If we set $\alpha = \gamma$, so that Pr(Type I error) = Pr(Type Ia error), then we must equate (3.2) and (3.3) such that

$$\begin{aligned} \alpha &= \gamma \\ \Pr(P < \alpha|H_0) &= \Pr(P > d|H_1) \\ &= \Phi(Z_d - \sqrt{n}\delta). \end{aligned} \tag{3.8}$$

[6]This corresponds to p-values (not p'-values) which suggest incompatibility with H_1. Recall that the positively skewed p-value density under H_1 is consistent with a small probability of observing a large p-value.

Table 3.2 ensures for all $z \in Z$ we will always be able to reject one, and only one, of the hypotheses if and only if $d = \alpha$. It then follows from (3.8) that $\Phi(Z_d - \sqrt{n}\delta) = \Phi(Z_\alpha - \sqrt{n}\delta)$ and so $\sqrt{n}\delta = 2Z_\alpha = 2Z_\gamma$ in order for $d = \alpha$.[7] If k is set *a priori*, and, as in this example, σ is known, then the required sample size to obtain $d = \alpha$ becomes

$$n = \left(\frac{2Z_\alpha}{\delta}\right)^2 = \frac{4Z_\alpha^2\sigma^2}{k^2}. \tag{3.9}$$

If the solution of (3.9) does not yield an integer, as is likely, then rounding n up (or otherwise an even larger sample) will deliver $d < \alpha$ (since higher test power increases the skewness of $f_P(p|\mathrm{H}_1)$) potentially resulting in the rejection of both H_0 and H_1 as per Table 3.2 if $d < p < \alpha$, where p is the p-value of realised (here Gaussian) test statistic z. As discussed this should be interpreted as a statistically significant result (due to rejection of H_0), but not practically or scientifically relevant (due to rejection of H_1) given k is the minimum value of μ to be a viable difference.

Similarly, if n is rounded down (or otherwise an even smaller sample) then $d > \alpha$ reflecting lower power. Hence if $\alpha < p < d$ for the p-value of z then we would fail to reject either hypothesis, therefore requiring a subsequent randomisation / further sampling.

---

[7]

$$\begin{aligned}
\Phi(Z_\alpha - \sqrt{n}\delta) &= \alpha \\
\Phi^{-1}(\Phi(Z_\alpha - \sqrt{n}\delta)) &= \Phi^{-1}(\alpha) \\
Z_\alpha - \sqrt{n}\delta &= Z_{1-\alpha} \\
\sqrt{n}\delta &= Z_\alpha - Z_{1-\alpha} \\
\sqrt{n}\delta &= Z_\alpha - (-Z_\alpha) \\
\sqrt{n}\delta &= 2Z_\alpha.
\end{aligned}$$

3.3.1 Empirical illustration

Let $Y \sim N(\mu, 16)$, $\bar{Y}_n \sim N(\mu, 16/n)$, $\mathrm{H}_0 : \mu = 0$ and $\mathrm{H}_1 : \mu = 2$, i.e. simple hypotheses. Given a random sample $y_1, \ldots, y_{49}$, with $\bar{y}_{49} = 0.8$, under H_0, $z = \sqrt{49} \times 0.8/4 = 1.4$ with corresponding p-value, $p = 0.081$, hence significant for $\alpha = 0.1$, not so for $\alpha = 0.05$ when testing H_0. Under H_1, $\delta = \mu/\sigma = 0.5$, therefore the associated p'-value is $p' = \Phi(Z_{0.081} - \sqrt{49}(0.5)) = 0.018$, strongly suggesting rejection of H_1. Figure 3.1 illustrates the p-value densities under both H_0 and H_1 ($\sqrt{n}\delta = 3.5$).

Alternatively, for $\gamma = 0.1$, the critical value d, such that $\Pr(P > d|\mathrm{H}_1) = 0.1$ is 0.013, for $\gamma = 0.05$, $d = 0.032$ and for $\gamma = 0.01$, $d = 0.120$. Hence a p-value of 0.081 leads to rejection of H_1 at the 10% and 5% levels, but not at the 1% level, consistent with the p'-value of 0.018. Note for $d = \alpha$, given μ and σ implies, using (3.9), that $n = 26.28$, 43.296 and 86.59 for $\alpha = 0.1$, 0.05 and 0.01 respectively. Clearly such sample sizes are unobtainable in practice, therefore increase n if you are willing to accept $d < \alpha$, otherwise decrease n to incur $d > \alpha$ allowing the possibility of future randomisations, should $\alpha < p < d$. However *ceteris paribus* higher test power is preferred, hence the larger n the better.

3.4 Additional Issues

The alternative hypothesis so far has been constructed to be simple. Composite forms of H_1 will be considered later, because often when assessing the significance of a potential explanatory variable for example, we are concerned with *any* non-zero effect. In such circumstances it is appropriate to treat the parameter under the alternative as a random variable with some distribution over a specified interval,

typically following a Gaussian or uniform distribution, say.

However to perform statistical testing vis-à-vis H_1, we require the p-value density under the alternative, $f_P(p|H_1)$, hence thus far for simplicity an alternative hypothesis with a specific-valued parameter has been used. This approach is the reason for suggesting that the minimum commercially or scientifically viable parameter value, k, (or other appropriate viability condition) be used in H_1.

3.4.1 Critical values

Given the advent of critical value tables to aid hypothesis testing, there is a clear need and advantage to produce such a table of the critical p-values, d, used in (3.8) to assist in determining the significance of H_1 *without* having to calculate p'-values. Table 3.3 achieves this for selected values of $\sqrt{n}\delta$. For a range of desired γ and $\sqrt{n}\delta$ values, the critical p-values are given, such that if the p-value is above d, we would reject H_1 at that γ level.

Note that this critical value table allows evaluation of the significance of H_1 based solely on the conventional p-value of the test statistic used to assess the plausibility, given the data, of H_0. Also, it is not essential to set $\alpha = \gamma$ (although as argued earlier, it would seem logical to do so) as these critical values are independent of α.

$f_P(p|H_1)$, as previously noted, is in general positively skewed as can be seen for example in Figure 3.1. This skewness increases with n, due to the increased test power when testing H_0. Therefore for larger sample sizes the skewness will be ever more marked resulting in lower critical values, d, through the effect on $\sqrt{n}\delta$. Also as $|k - \mu_0|$ increases (that is, the distance between null and alternative parameter values increases), power also rises which also affects $\sqrt{n}\delta$ since δ is a function of k. From Table 3.3 it can be seen that for $\sqrt{n}\delta > 6$, this critical value tends to, and is

thus indistinguishable, from zero for all γ levels, indicative of permanent rejection of H_1 even for arbitrarily small p-values.

In the limit as $\sqrt{n}\delta \longrightarrow 0$, for example as $\mu|H_1 \longrightarrow 0$, then this represents convergence to H_0. Consequently the p-value density under H_1 converges to the continuous uniform distribution. Therefore the upper bound on d in the limit as $\sqrt{n}\delta \longrightarrow 0$ is just $1-\gamma$, yielding an upper-tail probability of γ for a uniform density over the unit interval, mirroring the p-value critical value α when testing H_0.

However remember rejection of H_1 does not automatically facilitate acceptance of H_0, merely implausibility of the parameter value hypothesised in the alternative. Of course asymptotically thanks to the Central Limit Theorem the sample mean random variable converges to an unbiased degenerate distribution as the variance collapses towards zero. But for modest finite samples, $\sqrt{n}\delta$ will be small permitting useful application of this twin statistical testing methodology. So despite the restrictive nature of simple alternative hypotheses, for modest sample sizes which yield modest values of $\sqrt{n}\delta$, the use of such critical values as a substitute for p' is relevant.

Table 3.3: Critical p-values, d, for various levels of γ and selected values of $\sqrt{n}\delta$ when testing, for *known variance*, non-composite hypotheses featuring population mean parameter θ, i.e. $\mathrm{H}_0 : \theta = 0$ and $\mathrm{H}_1 : \theta = k$, such that $\Pr(P > d|\mathrm{H}_1) = \gamma$, where P is the one-tailed p-value of the test statistic under H_0. Under the null, the test statistic is to have a standard Gaussian distribution and under H_1 the test statistic is Gaussian with unit variance. n is the sample size and $\delta = k/\sigma$ where $k > 0$ is the hypothesised parameter value of θ under H_1.

$\sqrt{n}\delta$ / γ	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00
0.100	0.8489	0.7828	0.7025	0.6109	0.5126	0.4135	0.3197	0.2362	0.1664	0.1115	0.0710	0.0429
0.050	0.9185	0.8739	0.8146	0.7405	0.6535	0.5576	0.4581	0.3612	0.2725	0.1962	0.1345	0.0877
0.025	0.9564	0.9278	0.8869	0.8315	0.7611	0.6772	0.5832	0.4840	0.3859	0.2946	0.2148	0.1492
0.010	0.9811	0.9661	0.9425	0.9076	0.8591	0.7957	0.7178	0.6279	0.5304	0.4311	0.3359	0.2503
0.005	0.9900	0.9810	0.9661	0.9425	0.9076	0.8590	0.7956	0.7176	0.6277	0.5302	0.4309	0.3357

$\sqrt{n}\delta$ / γ	3.25	3.50	3.75	4.00	4.25	4.50	4.75	5.00	5.25	5.50	5.75	6.00
0.100	0.0245	0.0133	0.0068	0.0033	0.0015	0.0006	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000
0.050	0.0542	0.0318	0.0176	0.0093	0.0046	0.0022	0.0010	0.0004	0.0002	0.0001	0.0000	0.0000
0.025	0.0985	0.0618	0.0367	0.0207	0.0110	0.0055	0.0026	0.0012	0.0005	0.0002	0.0001	0.0000
0.010	0.1778	0.1203	0.0773	0.0471	0.0272	0.0149	0.0077	0.0038	0.0017	0.0008	0.0003	0.0001
0.005	0.2501	0.1777	0.1202	0.0772	0.0470	0.0272	0.0148	0.0077	0.0037	0.0017	0.0008	0.0003

3.4.2 Unknown variance

In the above example it was assumed that the variance of the observed data, σ^2, was known. In practice this will probably not be the case, since it is unlikely that we know the variance with certainty, but not the mean. In such instances, σ^2 should be replaced by its well-known unbiased sample estimator,

$$S^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2}{n-1}. \tag{3.10}$$

Of course, if this is the case then the test statistic will only be asymptotically standard Gaussian under H_0, similarly asymptotically Gaussian under H_1. Furthermore, if we seek large sample sizes to obtain consistent variance estimators, then this will have a consequential impact on p'-values due to the effect of higher n on test power.

For small samples, we perform the popular (and uniformly most powerful) t test[8] for $\mathrm{H}_0 : \mu = 0$ versus $\mathrm{H}_1 : \mu = k$. Again without loss of generality, assume $k > 0$. Replacing σ with S, the revised test statistic becomes

$$T = \frac{\sqrt{n}\bar{Y}_n}{S} \sim t_{n-1} \tag{3.11}$$

under H_0. Define $\hat{\delta} = k/S$,[9] therefore $T - \sqrt{n}\hat{\delta}$ also has a t_{n-1} distribution under H_1. As usual, $f_P(p|\mathrm{H}_0) = 1$, and $f_P(p|\mathrm{H}_1)$ is obtainable via (2.4), page 9. Let $T_{p,n-1}$

[8] At this point it should be noted that since the t test is uniformly most powerful for such problems when the variance is unknown, justification for its use relies on power considerations, i.e. among all test procedures satisfying the level constraint, choose the most powerful one. Clearly such an approach treats H_0 and H_1 asymmetrically, since the constraint is formulated under H_0 while the objective is formulated under H_1. Discussion of treating the hypotheses in a more symmetric manner using a Bayesian approach will be discussed in Chapter 4.

[9] Since σ is estimated with S, the quasi-estimated effect size is thus denoted $\hat{\delta}$.

be the $(1-p)$-th percentile of a t distribution with $n-1$ degrees of freedom. It follows that for ν degrees of freedom,

$$\begin{aligned} f_P(p;\nu|\mathrm{H}_1) &= \frac{f_T(t_p-\sqrt{n}\hat{\delta};\nu)}{f_T(t_p;\nu)} \\ &= \frac{\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}}\left(1+\frac{(t_{p,\nu}-\sqrt{n}\hat{\delta})^2}{\nu}\right)^{-\frac{\nu+1}{2}}}{\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}}\left(1+\frac{t_{p,\nu}^2}{\nu}\right)^{-\frac{\nu+1}{2}}} \\ &= \left(\frac{1+\frac{t_{p,\nu}^2}{\nu}}{1+\frac{(t_{p,\nu}-\sqrt{n}\hat{\delta})^2}{\nu}}\right)^{\frac{\nu+1}{2}}, \end{aligned} \tag{3.12}$$

where $f_T(t;\nu)$ is the density function of a t-distribution with ν degrees of freedom. Figure 3.2 illustrates $f_P(p|\mathrm{H}_1)$ for various degrees of freedom for $\sqrt{n}\hat{\delta} = 3.5$. Compare this with Figure 3.1.

In a similar light to the z test, we can proceed to compute p'-values to test H_1. Again it is much more convenient to have access to critical p-values which preclude the need to compute p'-values. Given the plethora of possibilities for the number of degrees of freedom and for $\sqrt{n}\hat{\delta}$, extensive tables are required to provide a complete resource. However Table 3.4 presents a sample of such critical d values for selected $\sqrt{n}\hat{\delta}$ and degrees of freedom values which should prove sufficient in practice using interpolation as an approximation for absent $\sqrt{n}\hat{\delta}$ or ν values. Derivation of these values stems from the distribution function of a Student's t variable,

$$F_T(t;\nu) = \Pr(T \le t;\nu) = \frac{1}{2} + \frac{t\Gamma((\nu+1)/2)_2F_1(\frac{1}{2},(\nu+1)/2;\frac{3}{2};-\frac{t^2}{\nu})}{\sqrt{\pi\nu}\Gamma(\nu/2)}, \tag{3.13}$$

where ${}_2F_1$ is the hypergeometric function. Using (2.3), the t-distribution equivalents

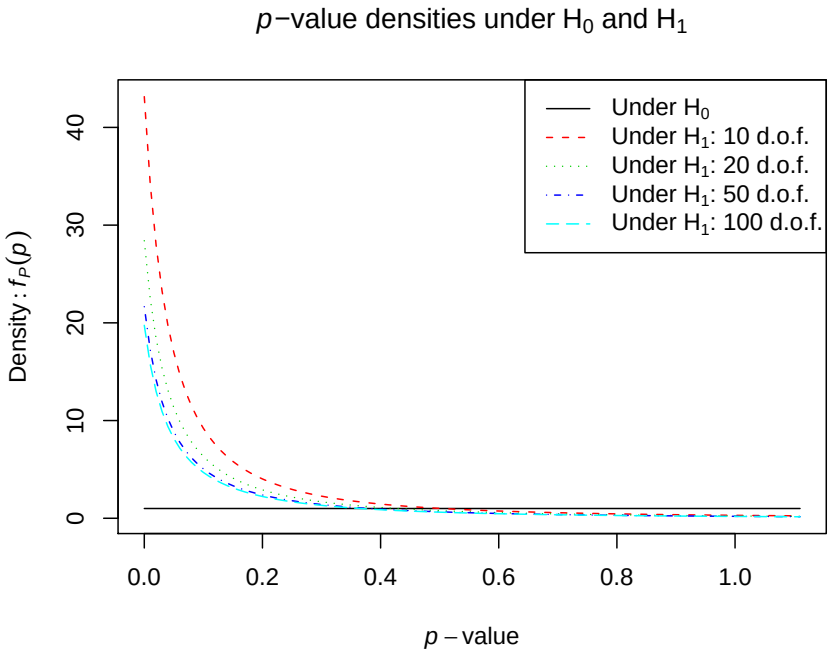


Figure 3.2: P-value densities when variance unknown under both $\mathrm{H}_0 : \mu = 0$ (uniform distribution) and $\mathrm{H}_1 : \mu = 2$ (ratio of two Student's t densities as per (3.12)) for 10, 20, 50 and 100 degrees of freedom such that $\sqrt{n}\hat{\delta} = 3.5$.

of (3.6) and (3.7) become respectively,

$$F_P(p|\mathrm{H}_1) = \int_0^p \frac{f_T(t_x - \sqrt{n}\hat{\delta};\nu)}{f_T(t_x;\nu)}dx = 1 - F_T(t_{p,\nu} - \sqrt{n}\hat{\delta};\nu) \tag{3.14}$$

and

$$p' = \Pr(P > p|\mathrm{H}_1) = F_T(t_{p,\nu} - \sqrt{n}\hat{\delta};\nu). \tag{3.15}$$

Note that as $\sqrt{n}\hat{\delta} \longrightarrow 0$, we again approach the null hypothesis under which p-values are uniformly distributed. Therefore the upper bound on d is again $1-\gamma$ because for a uniform density over the unit interval, the probability of being above $1-\gamma$ is γ.

3.5 Supplementary Hypothesis Test Scenarios

3.5.1 Comparison of means — variances known

When testing the equality of means using two sample means, $\bar{Y}_1$ and $\bar{Y}_2$, with sizes n_1 and n_2 and known variances σ_1^2 and σ_2^2 respectively, the test statistic

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{3.16}$$

has a standard Gaussian distribution under $\mathrm{H}_0 : \mu_1 - \mu_2 = 0$. For the specific alternative $\mathrm{H}_1 : \mu_1 - \mu_2 = k$, then $T - \frac{k}{\sqrt{\omega}}$ also has an $N(0,1)$ distribution, where $\omega = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$. Critical values in Table 3.3 can then be used replacing $\sqrt{n}\delta$ with $\frac{k}{\sqrt{\omega}}$.[10]

[10] $\sqrt{n}\delta$ could be written as $\frac{k}{\sqrt{\omega}}$ with $\omega = \frac{\sigma^2}{n}$ for a single sample case.

3.5.2 Comparison of means — variances unknown

The only difference with the known variance case above, is the use of estimated variances in the standard error term. As per (3.10), let σ_i^2 be estimated by the sample estimator S_i^2, $i = 1, 2$. Acknowledging the Behrens-Fisher problem, allow the revised test statistic under the null of equality to be

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}, \tag{3.17}$$

which is asymptotically Gaussian, but for small samples has a t distribution with $\min(n_1-1, n_2-1)$ degrees of freedom. Therefore as $T - \frac{k}{\sqrt{\hat{w}}}$ has the same distribution, Table 3.4 can be used for critical p-values when testing H_1 as appropriate where $\sqrt{n}\hat{\delta}$ is replaced by $\frac{k}{\sqrt{\hat{\omega}}}$ where $\hat{\omega} = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$.

3.5.3 Comparison of means - variances unknown but assumed equal

When the variances of Y_1 and Y_2 are unknown but assumed equal we use the standard error based on the pooled variance estimator,

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}. \tag{3.18}$$

It follows that for small samples the test statistic

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{3.19}$$

has a t distribution under H_0 with $(n_1 + n_2 - 2)$ degrees of freedom, but note is asymptotically Gaussian. Again, as $T - \frac{k}{\sqrt{\hat{\omega}_p}}$ has the same distribution, where $\hat{\omega}_p = S_p^2(\frac{1}{n_1} + \frac{1}{n_2})$, then Table 3.4 is applicable with $\sqrt{n}\hat{\delta}$ replaced by $\frac{k}{\sqrt{\hat{\omega}_p}}$.

3.5.4 Other situations

This twin-testing methodology is directly applicable to any hypothesis testing scenario provided the test statistic distributions are known either exactly or asymptotically. Clearly, depending on the test statistic distribution, different critical value tables may be required, for example for χ^2- or F-distributed test statistics, but the principles of testing H_1 are unchanged.

3.6 Negative Values for the Effect Size δ

Recall $\delta = \mu/\sigma$. Since $\sigma > 0$, it is possible to obtain negative effect sizes when $\mu < 0$. When testing $H_0 : \mu = 0$, the simple forms of $H_1 : \mu = k$ considered so far have been for positive k. Consequently for $k > 0$, $\sqrt{n}\delta > 0$ and the critical values provided in Table 3.3 are hence applicable. Similarly when variances have to be estimated, Table 3.4 with $\sqrt{n}\hat{\delta} > 0$ should be consulted.

Previously it has been stated that the p-value distribution under the alternative hypothesis is heavily skewed to the right which has meant the probability of observing small p-values is large, and so a small p-value lends support to H_1. This is a fundamental tenet in hypothesis testing whether in the classical or Fisherian mould. However this universal acceptance of right-skewed distributions for $f_P(p|H_1)$ breaks down when negative effect sizes are modelled in H_1. Consequently it will be shown that negative effect sizes induce large values of

$f_P(p|\mathrm{H}_1)$ for large p-values, going against the common perception of the distribution of $f_P(p|\mathrm{H}_1)$ as in Figure 3.1. Thus the cornerstone principle of only small p-values being more common under H_1 is not universally applicable. In such cases when this fails, it follows that the conventional approach to hypothesis testing is wholly inappropriate with significant implications for statistical inference.

To illustrate this phenomenon, Figure 3.3 plots the density and distribution functions respectively for a range of $\sqrt{n}\delta$ values (note that under $\mathrm{H}_0 : \mu = 0$, $\sqrt{n}\delta = 0$) using (2.5) and (3.6). From this figure it is clear that under H_1 the shape of the p-value distribution mirrors the commonly perceived right-skewed distribution for the generic p-value density under the alternative hypothesis, $f_P(p|\mathrm{H}_1)$. When $\delta < 0$, equivalently when $\sqrt{n}\delta < 0$, larger p-values are more likely to arise under H_1 than small p-values as evidenced by the distribution of the probability density. In such cases, significant p-values (i.e. those less than the designated significance level, α) tend to be more likely under H_0 than H_1. This result therefore calls into question the appropriateness of allowing small p-values to signal automatic rejection of H_0 when a negative δ is hypothesised in H_1.

It is possible to conclude from Figure 3.3, page 52, that for simple alternatives with $\delta < 0$, improbable test statistic occurrences under H_0 are even more improbable under H_1. Therefore if a small p-value occurs such that $f_P(p|\mathrm{H}_1) < f_P(p|\mathrm{H}_0)$, then the researcher should conclude in favour of the null hypothesis, even if $p < \alpha$. The next chapter will consider a methodological unification of Bayesian with classical and Fisherian hypothesis testing taking this into account.

Appendix A provides critical value tables for negative effect sizes, analogous to Tables 3.3 and 3.4 for positive effect sizes. The main difference is that a lower-tail test is performed when testing H_1, since sufficiently small p-values warrant rejection

Table 3.4: Critical p-values, d, for various levels of γ and selected values of $\sqrt{n}\hat{\delta}$ when testing, *for unknown variance*, non-composite hypotheses featuring population mean parameter θ, i.e. $\mathrm{H}_0 : \theta = 0$ and $\mathrm{H}_1 : \theta = k$ such that $\Pr(P > d|\mathrm{H}_1) = \gamma$, where P is the one-tailed p-value of the test statistic under H_0. Under the null, the test statistic has a t-distribution with $\nu = n - 1$ degrees of freedom and under H_1 the test statistic achieves the same distribution once $\sqrt{n}\hat{\delta}$ has been subtracted. n is sample size and $\hat{\delta} = k/S$ where $k > 0$ is the hypothesised parameter value of θ under H_1, and $S^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2}{n-1}$. Entries in the table give 10% (top), 5% (middle) and 1% (bottom) significance points respectively.

ν / $\sqrt{n}\hat{\delta}$	10	20	30	50	75	100	200	500
0.25	0.8560	0.8525	0.8513	0.8503	0.8498	0.8496	0.8492	0.8490
	0.9254	0.9221	0.9209	0.9199	0.9195	0.9192	0.9188	0.9186
	0.9846	0.9831	0.9825	0.9819	0.9817	0.9815	0.9813	0.9812
0.50	0.7982	0.7905	0.7880	0.7859	0.7848	0.7843	0.7835	0.7831
	0.8907	0.8825	0.8797	0.8774	0.8762	0.8756	0.8748	0.8742
	0.9765	0.9720	0.9702	0.9686	0.9678	0.9674	0.9668	0.9664
0.75	0.7261	0.7143	0.7103	0.7072	0.7056	0.7048	0.7037	0.7030
	0.8435	0.8293	0.8245	0.8205	0.8185	0.8176	0.8161	0.8152
	0.9641	0.9547	0.9509	0.9477	0.9460	0.9452	0.9439	0.9431
1.00	0.6412	0.6258	0.6208	0.6168	0.6148	0.6138	0.6123	0.6114
	0.7823	0.7615	0.7545	0.7489	0.7461	0.7447	0.7426	0.7413
	0.9459	0.9289	0.9223	0.9166	0.9137	0.9122	0.9099	0.9086
1.25	0.5474	0.5297	0.5239	0.5193	0.5171	0.5159	0.5143	0.5133
	0.7069	0.6799	0.6711	0.6640	0.6605	0.6587	0.6561	0.6546
	0.9195	0.8921	0.8816	0.8729	0.8684	0.8661	0.8626	0.8605
1.50	0.4504	0.4316	0.4255	0.4206	0.4183	0.4171	0.4153	0.4142
	0.6194	0.5878	0.5775	0.5695	0.5655	0.5635	0.5605	0.5588
	0.8825	0.8419	0.8270	0.8146	0.8084	0.8052	0.8005	0.7976
1.75	0.3567	0.3378	0.3317	0.3269	0.3245	0.3233	0.3215	0.3204
	0.5243	0.4900	0.4791	0.4706	0.4664	0.4643	0.4612	0.4594
	0.8327	0.7772	0.7576	0.7417	0.7338	0.7298	0.7238	0.7202
2.00	0.2721	0.2538	0.2479	0.2432	0.2409	0.2397	0.2380	0.2369
	0.4275	0.3930	0.3821	0.3736	0.3694	0.3674	0.3643	0.3625
	0.7687	0.6983	0.6746	0.6558	0.6464	0.6418	0.6348	0.6307

ν / $\sqrt{n}\hat{\delta}$	10	20	30	50	75	100	200	500
2.25	0.2003	0.1831	0.1775	0.1730	0.1708	0.1697	0.1681	0.1671
	0.3355	0.3026	0.2923	0.2842	0.2803	0.2783	0.2754	0.2737
	0.6907	0.6081	0.5814	0.5606	0.5504	0.5454	0.5378	0.5334
2.50	0.1429	0.1270	0.1218	0.1176	0.1156	0.1146	0.1131	0.1121
	0.2537	0.2236	0.2142	0.2069	0.2033	0.2015	0.1989	0.1973
	0.6013	0.5110	0.4831	0.4617	0.4513	0.4461	0.4385	0.4340
2.75	0.0992	0.0848	0.0802	0.0765	0.0746	0.0737	0.0724	0.0715
	0.1853	0.1587	0.1504	0.1440	0.1408	0.1392	0.1369	0.1355
	0.5054	0.4133	0.3859	0.3651	0.3551	0.3502	0.3430	0.3387
3.00	0.0673	0.0548	0.0507	0.0476	0.0460	0.0452	0.0440	0.0433
	0.1312	0.1084	0.1013	0.0957	0.0930	0.0917	0.0897	0.0885
	0.4090	0.3210	0.2957	0.2767	0.2676	0.2632	0.2567	0.2528
3.25	0.0449	0.0343	0.0309	0.0283	0.0270	0.0264	0.0254	0.0249
	0.0906	0.0714	0.0655	0.0609	0.0586	0.0575	0.0559	0.0549
	0.3186	0.2393	0.2171	0.2006	0.1928	0.1889	0.1833	0.1800
3.50	0.0296	0.0209	0.0182	0.0162	0.0152	0.0147	0.0140	0.0135
	0.0612	0.0455	0.0407	0.0371	0.0353	0.0344	0.0331	0.0323
	0.2393	0.1713	0.1527	0.1390	0.1325	0.1294	0.1248	0.1221
3.75	0.0194	0.0125	0.0104	0.0089	0.0082	0.0078	0.0073	0.0070
	0.0407	0.0282	0.0245	0.0216	0.0203	0.0196	0.0186	0.0180
	0.1736	0.1180	0.1030	0.0921	0.0869	0.0844	0.0808	0.0787
4.00	0.0126	0.0073	0.0058	0.0047	0.0042	0.0040	0.0036	0.0034
	0.0268	0.0170	0.0142	0.0121	0.0111	0.0106	0.0099	0.0095
	0.1223	0.0783	0.0667	0.0583	0.0544	0.0525	0.0498	0.0482
4.25	0.0082	0.0042	0.0031	0.0024	0.0021	0.0019	0.0017	0.0016
	0.0175	0.0101	0.0080	0.0065	0.0058	0.0055	0.0050	0.0048
	0.0840	0.0502	0.0416	0.0354	0.0325	0.0311	0.0291	0.0280
4.50	0.0054	0.0024	0.0017	0.0012	0.0010	0.0009	0.0008	0.0007
	0.0114	0.0058	0.0044	0.0034	0.0029	0.0027	0.0024	0.0023
	0.0566	0.0313	0.0250	0.0205	0.0185	0.0176	0.0162	0.0154
4.75	0.0035	0.0013	0.0009	0.0006	0.0005	0.0004	0.0003	0.0003
	0.0074	0.0033	0.0024	0.0017	0.0014	0.0013	0.0011	0.0010
	0.0375	0.0190	0.0145	0.0115	0.0101	0.0095	0.0085	0.0080
5.00	0.0023	0.0008	0.0004	0.0003	0.0002	0.0002	0.0001	0.0001
	0.0048	0.0019	0.0012	0.0008	0.0007	0.0006	0.0005	0.0004
	0.0247	0.0113	0.0082	0.0062	0.0053	0.0049	0.0043	0.0040

of H_1. Therefore it is necessary to restate the p'-value for negative effect sizes as the complement of the p'-value defined in (3.7). Hence,

$$(p')^c = 1 - p' = \Pr(P \leq p|H_1) = F_P(p|H_1) = 1 - \Phi(Z_p - \sqrt{n}\delta). \tag{3.20}$$

Consequently the revised critical values for negative $\sqrt{n}\delta$ are, from (3.7) and (3.20), obtained by simply subtracting from one the critical value corresponding to the equivalent $|\sqrt{n}\delta|$. Table A.1 therefore provides critical p-values, d, such that $\Pr(P \leq d|H_1) = \gamma$, hence the d values correspond to p-values with associated p'-values, $(p')^c$, equal to γ for the respective $\sqrt{n}\delta$.

Similarly when variances are estimated, (3.15) leads to the following lower-tail p'-value,

$$(p')^c = 1 - p' = \Pr(P \leq p|H_1) = F_P(p|H_1) = 1 - F_T(t_{p,\nu} - \sqrt{n}\hat{\delta}; \nu). \tag{3.21}$$

Again subtracting the entries in Table 3.4 from one yields the corresponding critical p-values for negative $\sqrt{n}\hat{\delta}$ provided in Table A.2. Just as usual for obtaining critical values, suitable interpolation should be performed when the exact $\sqrt{n}\delta$, $\sqrt{n}\hat{\delta}$ and ν values required are not explicitly listed.

3.7 Composite Alternative Hypotheses

So far, analysis has been restricted to simple forms of H_1. In practice the population parameter is unknown, otherwise why test for its true value? Hence it may not always be appropriate to invoke a simple alternative, rather it is better to model

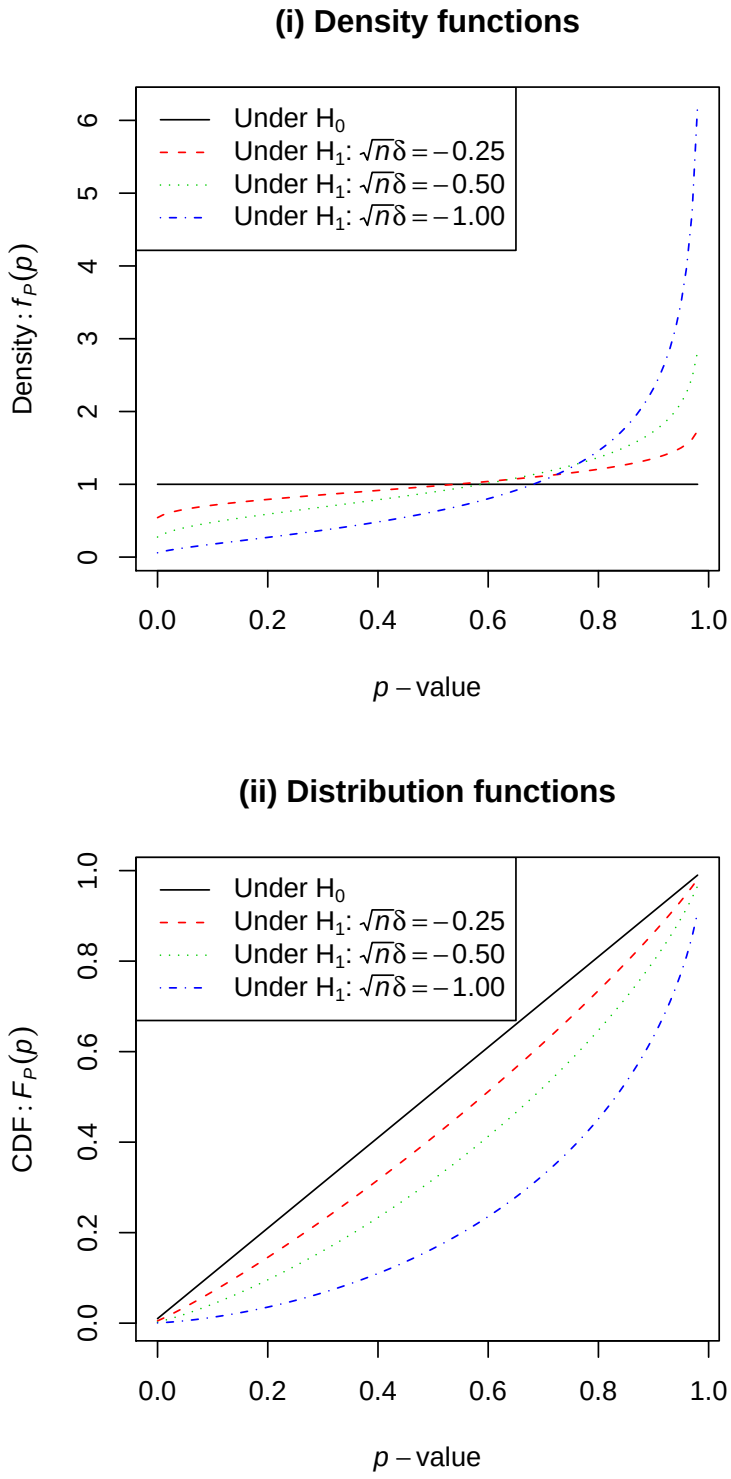


Figure 3.3: P-value density and distribution functions for simple forms of H_1 where the effect size function $\sqrt{n}\delta$ takes negative values.

the parameter as following some hypothesised distribution.[11] In this light, the effect size, δ, is now treated as a random variable over some prescribed interval of values reflecting the domain of the modelled distribution.

Different distribution families will naturally lead to different p-value distributional properties. As seen in the previous section, when the effect size parameter takes negative values, the p-value density, $f_P(p|\mathrm{H}_1)$, is left-skewed meaning small p-values are unlikely under the alternative. This phenomenon extends to composite alternative hypotheses when the domain of the modelled distribution for δ incorporates negative values. To illustrate this, let δ follow one of two distributions — the Gaussian and uniform will be considered. Hung, O'Neill, Bauer, and Köhne (1997) derive the respective p-value density and distribution functions. For the Gaussian case such that $\delta \sim N(\zeta, \omega^2)$ with sample size n,

$$\begin{aligned} f_P(p|\mathrm{H}_1) = g_{\zeta,\omega,n}(p) &= \left[\omega(n+\omega^{-2})^{1/2}\right]^{-1} \\ &\quad \times \exp\left\{-\frac{1}{2}\left[(\zeta/\omega)^2 - (\sqrt{n}Z_p + \zeta/\omega^2)^2/(n+\omega^{-2})\right]\right\}, \end{aligned} \tag{3.22}$$

$$F_P(p|\mathrm{H}_1) = G_{\zeta,\omega,n}(p) = 1 - \Phi\left\{(Z_p - \sqrt{n}\zeta)/(\omega^2 n + 1)^{1/2}\right\}, \tag{3.23}$$

where (3.22) and (3.23) denote the density and distribution functions, under H_1, respectively. For the uniform case such that $\delta \sim U[a, b]$ for arbitrary parameters a

[11] The argument previously advanced in defence of a simple H_1 related to the minimum parameter value required for commercial or scientific relevance, which is justified in certain circumstances.

and b with sample size n,

$$f_P(p|\mathrm{H}_1) = g_{a,b,n}(p) = (2\pi/n)^{1/2}\exp(0.5Z_p^2) \times \left\{\Phi(\sqrt{n}b - Z_p) - \Phi(\sqrt{n}a - Z_p)\right\}/(b-a), \quad (3.24)$$

$$F_P(p|\mathrm{H}_1) = G_{a,b,n}(p) = \left\{\sqrt{n}(b-a)\right\}^{-1} \times \int_{Z_p}^{\infty} \left\{\Phi(v - \sqrt{n}a) - \Phi(v - \sqrt{n}b)\right\} dv, \quad (3.25)$$

where (3.24) and (3.25) denote the density and distribution functions, under H_1, respectively.

Our interest concerns how the parametric specification affects the p-value distribution. It is noted that each distribution takes three parameters, namely ζ, ω and n for the Gaussian distribution and a, b and n for the uniform distribution.

3.7.1 Gaussian specification of δ in $\mathbf{H}_1$

Note that setting $\zeta = 0$ corresponds to δ degenerating to 0 as $\omega^2 \longrightarrow 0$. (3.23) reduces in this limit to $G_{0,0,n} = 1 - \Phi(Z_p) = 1 - (1-p) = p$, i.e. the continuous uniform distribution function over $[0, 1]$, which is consistent with the (simple) null hypothesis with $\mu = 0$.[12]

Clearly given the domain of the Gaussian distribution is the entire real line, $\mathbb{R}$, we have $-\infty < \zeta < \infty$ and $\omega^2 > 0$. Therefore the domain of a Gaussian variable always incorporates the entire real line, so the p-value density should exhibit large values for p-values close to both 0 and 1 for modest ζ. For illustrative purposes, Figure 3.4 plots the density and distribution functions for a variety of ζ and ω^2 values for sample sizes of $n = 100$. It can be seen that these densities depart from

[12]Recall that $\delta = \mu/\sigma$.

the conventional right-skewed distributions with both extremely small and large p-values under H_1 being highly likely. This is reflected in the right-hand side of the figure which displays the corresponding distribution functions. Whereas Figure 3.1, say, would be compatible with a fully concave distribution function, it is evident that as the p-value increases convexity is obtained resulting in substantial probabilities for observing p-values around both the lower and upper limits of the unit interval.

The specification of the two parameters, ζ and ω^2, in the Gaussian distribution of δ will affect the proportion of the distribution of δ which traverses negative values, thus affecting the likelihood of observing large p-values under H_1. *Ceteris paribus*, as ζ increases then this proportion reduces corresponding to a decrease in the height of $f_P(p|H_1)$ for large p-values. However full concavity of the distribution function is never achieved, since even as $\zeta \longrightarrow \infty$, there still exists some positive probability that $\delta < 0$, and hence slight convexity of $F_P(p|H_1)$ as $p \longrightarrow 1$.

It follows that the variance of δ, ω^2, also affects the distributional shape of $f_P(p|H_1)$ and hence the proportion of δ's probability distribution taking negative values. For $\zeta > 0$, larger ω^2 values increase this proportion, while for $\zeta < 0$ a higher δ variance reduces this proportion, but as mentioned above this proportion is always strictly positive resulting in a raised p-value density function towards the upper bound of the unit interval.

3.7.2 Uniform specification of δ in H_1

Under H_1, $\delta \sim U[a, b]$. Figure 3.5 plots the p-value density function for three pairs of $[a, b]$ values for $n = 100$. For $a = 0$ and $b = 2$ we obtain the familiar right-skewed characteristic as expected since this δ domain excludes negative values. However a left-skewed mirror image is obtained for $a = -2$ and $b = 0$ (note encompassing

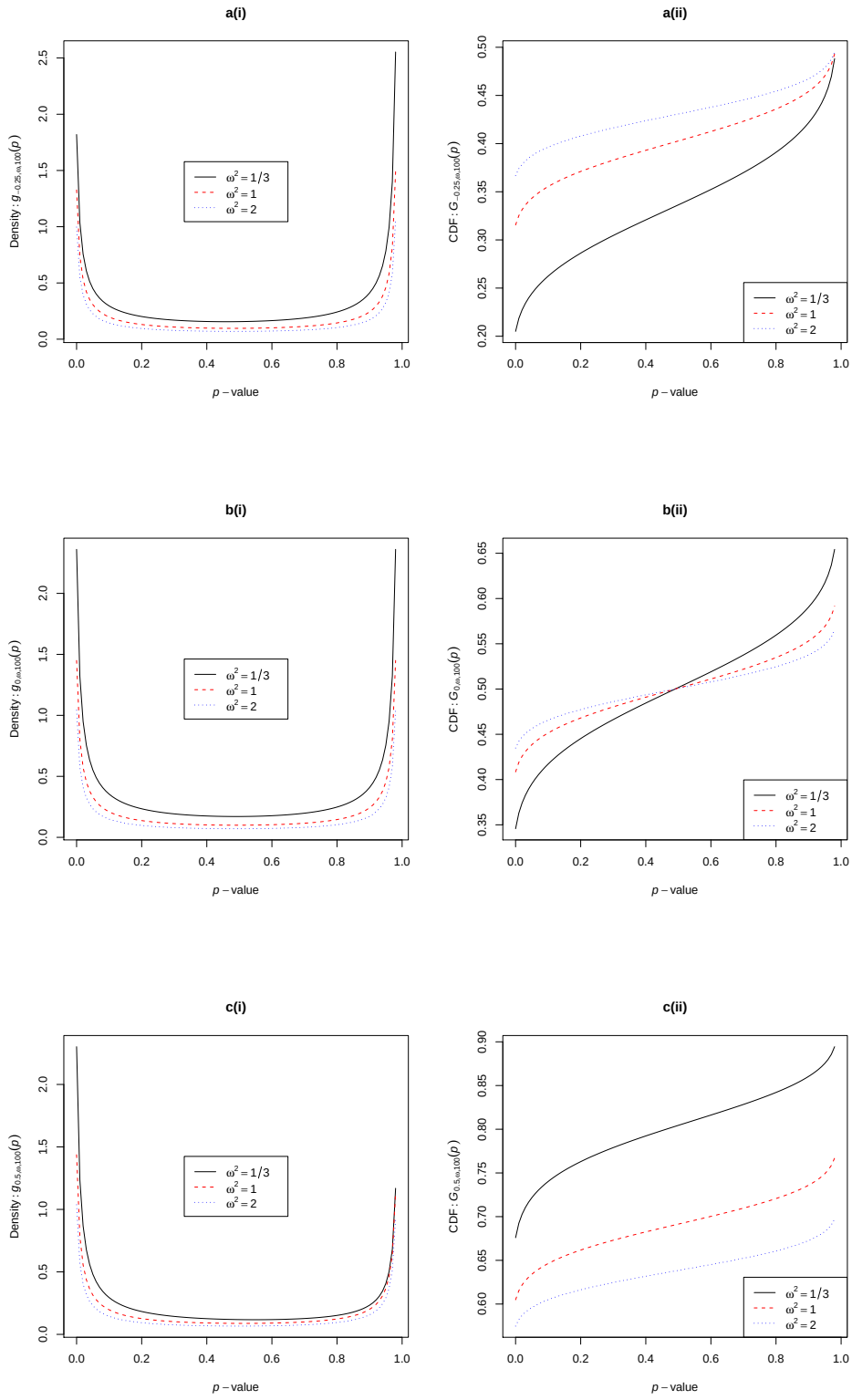


Figure 3.4: P-value density and distribution functions for composite forms of H_1 where the effect size parameter $\delta \sim N(\zeta, \omega^2)$ as per (3.22) and (3.23). a(i) = $g_{-0.25,\omega,100}(p)$, a(ii) = $G_{-0.25,\omega,100}(p)$, b(i) = $g_{0,\omega,100}(p)$, b(ii) = $G_{0,\omega,100}(p)$, c(i) = $g_{0.5,\omega,100}(p)$ and c(ii) = $G_{0.5,\omega,100}(p)$.

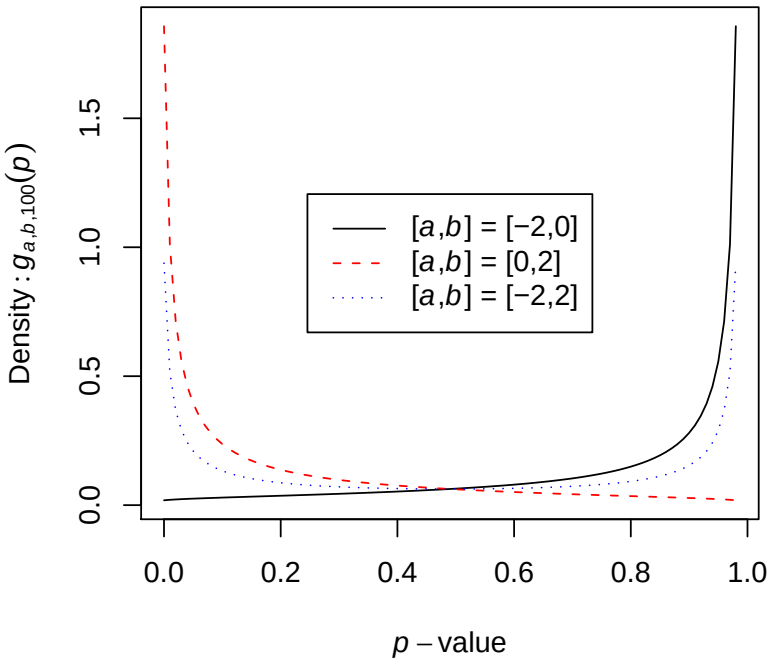


Figure 3.5: P-value density functions for composite forms of H_1 where the effect size parameter $\delta \sim U[a, b]$ as per (3.24).

exclusively negative δ values) reflecting the simple H_1 analogue as illustrated in Figure 3.3, implying that under the alternative, extremely large p-values are more likely. Finally for $a = -2$ and $b = 2$ we have a mixture of positive and negative δ values resulting in a distributional shape similar to that of the Gaussian cases presented in Figure 3.4.

3.7.3 Choice of δ specification in H_1

As a result, when the effect size parameter is hypothesised to be a random variable, depending on the domain specification of the distribution of δ, it is possible that small p-values could be more indicative of the null hypothesis rather than H_1 when negative δ values are feasible. Hung, O'Neill, Bauer, and Köhne (1997) also derive the p-value distribution when δ follows a lognormal distribution such that $\log(\delta)$ is Gaussian with mean θ and variance ν^2, replicated here for completeness as

$$f_P(p|\mathrm{H}_1) = g_{\theta,\nu,n}(p) = \int_{-\infty}^{\infty} \phi(Z_p - \sqrt{n}e^{\nu u+\theta})\phi(u)/\phi(Z_p)\ du, \quad (3.26)$$

$$F_P(p|\mathrm{H}_1) = G_{\theta,\nu,n}(p) = \int_{-\infty}^{\infty} \Phi(\sqrt{n}e^{\nu u+\theta} - Z_p)\phi(u)\ du, \quad (3.27)$$

where (3.26) and (3.27) denote the density and distribution functions, under H_1, respectively. However since the support of a lognormal distribution is $[0, \infty)$, the negative δ problem is not encountered and so no further consideration is given to this distribution here.

3.8 Conclusions

This chapter has focused on the distribution of the p-value under the alternative hypothesis, H_1. The motivation for doing so is that just because a given test statistic value is improbable under the null, it cannot automatically be inferred that the alternative specification is more likely. Consideration was given to the second-order p-value, p', to assess the plausibility of H_1.

For simple forms of H_1 testing for a non-zero effect, a simultaneous testing methodology extending the conventional classical and Fisherian procedures to incorporate testing of H_1 was outlined, highlighting the possibility that both or none of the two hypotheses might be rejected. In order to facilitate widespread adoption of this testing approach, critical value tables were produced, reflecting the popular use of such tables for numerous test statistic distributions. These tables provided critical p-values for a range of significance levels, γ, which can be used to determine, for a given sample size n and effect size δ, whether H_1 should be rejected. Derivation of the p-value distribution extended to the case of t-distributed test statistics was introduced for use in common testing situations when unknown population variances are estimated.

Another addition to the literature concerns the recognition that the sign of the effect size influences the shape of $f_P(p|H_1)$ such that for a simple alternative with $\delta < 0$, $f_P(p|H_1)$ is left-skewed mirroring the case for a positive δ. Such scenarios indicate that the conventional approach to hypothesis testing whereby small p-values constitute automatic rejection of H_0 and acceptance of H_1 are misguided since an improbable test statistic outcome under the null can be even more improbable under the alternative.

Attention then turned towards composite forms of H_1. By modelling the unknown effect size with a particular distribution, it was shown that when the support of the distribution included negative values (for example the Gaussian and some continuous uniform distributions), then departure from the typical right-skewed form of $f_P(p|H_1)$ resulted. Consequently this chapter has emphasised an important caveat when performing such hypothesis tests, specifically that large p-values can be indicative of H_1.

With this in mind, the next chapter will look towards the methodological unification of the different schools of hypothesis testing which will explicitly incorporate the p-value distribution under the alternative hypothesis. Although the Bayesian, Classical and Fisherian schools of hypothesis testing each have their advantages, a single approach consolidating the best of all the doctrines is clearly sought.

Chapter 4

Ménage à Trois Inference Style: Unifying Three Hypothesis Testing Doctrines

In recent years considerable attention in the literature has focused on the suitability of conventional null hypothesis significance testing (NHST), with a frustrating lack of agreement. For instance in wildlife research, Anderson, Burnham, and Thompson (2000) report on the increasing number of papers criticising the NHST approach, but Thompson's data cite numerous defences as well. A common complaint concerns the misuse of NHST, rather than the procedure itself. Incorrect interpretation of the test conclusions however is hardly justification for an embargo on NHST (as suggested in Schmidt (1996)), but rather simply a matter of researcher training.

This chapter seeks to extend previous attempts to provide a methodological unification of the different schools of hypothesis testing (Neyman-Pearson, Fisherian and Bayesian). Each school has its own merits, however each also suffers from limitations which are discussed. Attention focuses on the concept of conditional

frequentist testing which has been developed in recent years to help provide unity. New results presented here include a revised conditioning statistic taking into account the behaviour of the p-value under H_1 by considering its density, $f_P(p|\mathrm{H}_1)$. As a consequence, new critical p-value curves and surfaces are constructed to provide a quick-and-easy method for researchers to employ this conditional methodology, with computation limited to obtaining a conventional p-value and determining certain parameter values which are a product of the specification of H_1.

4.1 Bayesian Hypothesis Testing

Prior to the 1920s, statistical inference was foremost Bayesian, following on from the pioneering work of Bayes and Laplace. As a simple illustrative example, consider n independent Bernoulli trials used to test $\mathrm{H}_0 : \theta = \theta_0$ against $\mathrm{H}_1 : \theta \neq \theta_0$ for $0 < \theta < 1$. Prior probabilities for the two hypotheses, $\Pr(\mathrm{H}_i)$, $i = 0, 1$, are stated as well as the prior density for θ, $\pi(\theta)$.[1] Objective Bayesians would use default probabilities of 0.5 and a default prior density $\pi(\theta) = 1$ over $0 < \theta < 1$ for a typical Bernoulli problem. A subjective approach would choose probabilities and a density based on personal beliefs or real extraneous information. Once data, say x, have been collected, the posterior probabilities of the hypotheses can be computed, representing the posterior distribution. For example for the null hypothesis, from Bayes' theorem,

$$\Pr(\mathrm{H}_0|x) = \frac{\Pr(\mathrm{H}_0) f(x|\theta = \theta_0)}{\Pr(\mathrm{H}_0) f(x|\theta = \theta_0) + \Pr(\mathrm{H}_1) \int_{\{\theta \neq \theta_0\}} f(x|\theta)\pi(\theta) d\theta}, \tag{4.1}$$

[1] One can view this as a weight function permitting calculation of an average likelihood under the alternative hypothesis.

where $f(x|\theta)$ is the sample distribution of x, given parameter θ. Hence, it is necessary to be able to evaluate the integral analytically, or at least approximate it numerically by Monte Carlo methods. Equation (4.1) then gives a measure of the likelihood of H_0 taking into account $\pi(\theta)$. This contrasts and conflicts with a conventional p-value, an issue returned to later.

An alternative quantity to report is the *Bayes factor* which yields a measure of the odds of H_0 to H_1, given the data — essentially a weighted likelihood ratio. Formally the Bayes factor, say $B_{0,1}$ of H_0 to H_1, is the posterior odds ratio over the prior odds ratio,

$$B_{0,1} = \frac{\Pr(\mathrm{H}_0|x)/\Pr(\mathrm{H}_1|x)}{\Pr(\mathrm{H}_0)/\Pr(\mathrm{H}_1)} \tag{4.2}$$

$$= \frac{f(x|\theta=\theta_0)}{\int_{\theta\neq\theta_0} f(x|\theta)\pi(\theta)d\theta}. \tag{4.3}$$

The interpretation of the specification in (4.3) is as the likelihood of the data under H_0 divided by the average likelihood under H_1, with the advantage that the Bayes factor is independent of the prior hypothesis probabilities, and so reflects the observed data only. Clearly for the objective approach with $\Pr(\mathrm{H}_0) = \Pr(\mathrm{H}_1) = 0.5$, the Bayes factor is simply the posterior odds ratio. Given (4.3), the posterior probability of H_0 can alternatively be stated as

$$\Pr(\mathrm{H}_0|x) = \left[1 + \frac{\Pr(\mathrm{H}_1)}{\Pr(\mathrm{H}_0)} \cdot \frac{1}{B_{0,1}}\right]^{-1}. \tag{4.4}$$

Berger and Delampady (1987), for example, derive the Bayes factor for $\pi(\theta) \sim N(\theta_0, \tau^2)$ in contrast to the Cauchy $C(\theta_0, \tau^2)$ preferred by Jeffreys (1961), where τ^2

is a hyperparameter. Berger (1985) provides useful references in defence of objective priors in response to the frequentist criticism of Bayesian techniques requiring a prior specification.[2] Having equal prior probabilities is intuitively acceptable as representing objectivity due to the symmetry of the prior beliefs (despite the fact that just considering Bayes factors removes the need to even consider such probabilities), however there is no clear objective choice for $\pi(\theta)$.

In Berger and Delampady (1987) it is argued that $\pi(\theta)$ should be symmetric about θ_0 for a parameter space spanning the entire real line, and possibly be non-increasing in $|\theta - \theta_0|$ to avoid bias towards $\theta \neq \theta_0$. They note that the functional form of $\pi(\theta)$ is largely irrelevant, however in the Gaussian versus Cauchy specification, the scale factor τ is influential in both Bayes factor and posterior probabilities which means that τ must be specified, and for that matter specified subjectively since there is no obvious default, objective value. Of particular note however, are the 'automatic' Bayesian significance tests of Jeffreys (1961) (specifying a Cauchy $C(\theta_0, \sigma^2)$ prior) and Smith and Spiegelhalter (1980) (specifying a constant default prior) which, although not completely objective, do yield superior results vis-à-vis p-values.

Whereas non-Bayesians would be inclined to report a p-value and perhaps a confidence interval of likely values of the unknown parameter, a Bayesian approach would be to report the posterior $\Pr(\mathrm{H}_0|x)$ with, say, a 95% posterior credible interval for the parameter. So fundamentally, we have two competing statistics for point statistics for empirical conclusions, namely the p-value and the posterior probability, $\Pr(\mathrm{H}_0|x)$.[3] Although both seem intuitively appealing, it is possible to

[2]Berger and Berry (1988) note the disguised subjectivity within the frequentist ideology.

[3]The focus here will be on these summary statistics rather than confidence intervals and posterior credible intervals.

encounter an inferential conflict between p-values and the conditional measures of Bayes factors and posterior probabilities for two-sided tests, such as a small (i.e. significant) p-value occurring in parallel with a large $\Pr(\mathrm{H}_0|x)$. When such cases occur, p-values are very misleading resulting in an irreconcilability between p-values and posterior probabilities. In terms of quantities to report, such as reporting posterior probabilities for a range of (subjective) prior inputs, see Dickey (1973).

4.1.1 Example of inferential conflicts

Take as an example the interesting experiment investigating the presence of psychokinesis, that is the ability of the mind to influence matter. In 1987, an experiment by Schmidt, Jahn and Radin (see Jahn, Dunne, and Nelson (1987)) seemed to prove the existence of this phenomenon. Particles arrived at a quantum gate and the experiment was set up such that the probability of particles veering towards one of two directions was 0.5. Of the 104,900,000 independent Bernoulli trials, there were 53,263,000 successes providing allegedly strong evidence in favour of the paranormal, with the test of $\mathrm{H}_0 : \theta = 0.5$ yielding a p-value of 0.0003. So does this imply that the X-Files are true? Sadly, no. If the Bayesian approach outlined above is used, then $\Pr(\mathrm{H}_0|x) = 0.94$, so psychic ability is unlikely. Hence here the p-value is extremely misleading. Of course in practice we would not expect p to be exactly equal to $\Pr(\mathrm{H}_0|x)$, however although $p < \Pr(\mathrm{H}_0|x)$, the magnitude of the difference is particularly startling. Other examples of such a conflict can be found in Diamond and Forrester (1983). Note the focus here on two-sided tests, i.e. a simple or small interval null hypothesis being tested against $\mathrm{H}_1 : \Omega \setminus \Theta_0$.

Clearly the (very) large sample size used would easily yield a (very) small p-value when the sample proportion deviates even slightly from 0.5 due to the standard

error. Some departure from H_0 is likely to occur precisely because in any experiment there is likely to be some systematic deviation from the strict H_0 such as a calibration issue in the experimental design as well as the stochastic nature of the experimental particles. Consequently the p-value will be decreasing in n, the sample size. Therefore with a sample size of several million, even minor deviations from a strict H_0 will be statistically significant as a result of false positives.

So the Fisherian approach which produces a p-value, i.e. the probability of the observed outcome or a more extreme one, seems to be flawed. Much better, therefore, to report the likelihoods of all the different hypotheses assessing their strengths conditional on the data, as achieved in Bayesian testing. In essence the hypotheses are all in direct competition with one another[4] and the posterior probabilities allow the researcher to discriminate between them. Should no hypothesis emerge the 'winner', i.e. we have inconclusive results, then more data should be collected. Note that in practice it is never possible to be 100% certain in accepting or rejecting a particular hypothesis — an open mind must be maintained since new observations might cause a revision in the posterior probabilities culminating in a previously preferred hypothesis becoming less likely while the less endeared hypothesis might suddenly become in vogue. Initial data may be compatible with the sample distribution $f(x|\theta)$, however the true sampling distribution could be of a completely different functional form, but the data x might be compatible with both distributions by pure coincidence, unlike new observations (from the true distribution) which may be incompatible with $f(x|\theta)$.[5]

[4]In model choice problems, it is possible to have several hypotheses each representing a different model.

[5]This argument reflects the questions facing any theory, i.e. model assumptions and simplifications should be reasonable and the core theoretical implications should be reflected in the data. However this does not prove that a model is true, rather a completely different mechanism may have generated the data, but the incorrect model provides a good fit by pure coincidence,

A basic deficiency with Fisherian hypothesis testing is that it answers the question 'Given H_0 is true, what is the probability of these (or more extreme) data?' i.e. $\Pr(x|H_0)$, however what one really wants to answer is 'Given these data, what is the probability that H_0 is true?' i.e. $\Pr(H_0|x)$, that is the conditioning is reversed. The important point is that in general $\Pr(x|H_0) \neq \Pr(H_0|x)$.[6] The reason for the considerable disparity between $\Pr(H_0|x)$ and the p-value for two-sided tests stems from the conditioning set. The posterior probability takes into account only the data, while the p-value considers the probability of observing the data *or a more extreme result*. As Jeffreys (1980) commented,

> 'I have always considered the arguments for the use of P absurd. They amount to saying that a hypothesis that may or may not be true is rejected because a greater departure from the trial value was improbable; that is, that it has not predicted something that has not happened.' (p. 453)

Cohen (1994) provides an entertaining review and critique of the pitfalls of Fisherian and Neyman-Pearson testing citing the '*mechanical dichotomous decisions around a sacred .05 criterion*' (italics in original) and his 'temptation to call it statistical hypothesis inference testing', with an eye on its acronymous namesake.

Berger and Sellke (1987) investigate lower bounds on the posterior probability, $\Pr(H_0|x)$, for different priors for testing point null hypotheses, and include references to other papers testing Bayesian point nulls. They report that $p << \Pr(H_0|x)$ where

although the likelihood of this decreases with more data.

[6]Although Fisherian advocates may argue that the definition of a p-value as $\Pr(x|H_0)$ is no secret and hence it is foolish to treat a p-value as measuring $\Pr(H_0|x)$, the fact is that most practitioners are non-specialists who confuse the distinction between $\Pr(x|H_0)$ and $\Pr(H_0|x)$ — see Diamond and Forrester (1983). Therefore given this conflict for two-sided tests, the reporting of p-values inevitably leads to a culture of rejecting H_0 too liberally. As already mentioned, such Type I errors are by convention more intolerable than their Type II counterparts.

equality can only be achieved provided the prior is heavily biased in favour of H_1, for example $\Pr(H_1) = 0.85$, where the probability mass is symmetrically spread out to most favour H_1, can achieve a posterior probability of 0.05 for a two-sided z-statistic of 1.96. Clearly such a biased prior would be unpalatable to most, if not all, however a practitioner wishing to reject the null (if a 'significant' result was especially sought) can easily circumvent this perceived bias by just reporting the conventional p-value, citing 'standard practice'. Since these lower bounds all exceed the p-values regardless of prior choice, then it is not possible to dismiss this inferential conflict between p-values and conditional measures based on the subjective choice of $\pi(\theta)$. Edwards, Lindman, and Savage (1963) are considered the first to expose the magnitude of this irreconcilability, such that p-values are typically at least an order of magnitude less than conditional measures.

4.2 Unifying Bayesians and Frequentists

The previous section highlighted the conflict between classical p-values and conditional measures, namely Bayesian posterior probabilities and, through (4.4), Bayes factors. Researchers have sought to bridge the divide between the various schools of testing (Neyman-Pearson, Fisherian and Bayesian). Bayarri and Berger (2004) review achievements in developing a methodological, if not philosophical, union between the opposing camps citing the pedagogical benefits which inevitably result from consistent inference.[7] For a discussion concerning the adverse effects of divided methodologies, see Goodman (1999a) and Goodman (1999b). Synthesising

[7]Robinson and Wainer (2001) present a critique of null hypothesis significance testing (NHST) to educate researchers in the art of best practice. They conclude that NHST has its merits but should be treated as an adjunct to other forms, such as Bayesian testing when a probabilistic statement concerning the hypotheses is sought.

the best of both worlds is naturally appealing.

It should be noted that as far as *estimation* is concerned, frequentist and Bayesian approaches typically yield the same, or at least similar, results for common parametric problems involving continuous parameters, allowing the adoption of either frequentist or Bayesian interpretations. Yet despite frequentist estimation being effective, Bayesian tools should be implemented to assess estimator accuracy. Many frequentist methods require asymptotic approximations, and are also used in Bayesian cases (see LeCam (1986) and Schervish (1995) for further details), however unlike frequentist methodologies exact small sample solutions can be obtained for Bayesian procedures, often more easily than asymptotic methods.

4.2.1 Review of testing doctrines

Three distinct schools of hypothesis testing exist advocated by Neyman, Fisher and Jeffreys.[8] The trouble arises due to the considerable disagreement in the test results of simple, or small interval, null hypotheses reported by each method.[9] Efron and Gous (2001) consider the differences in the scales of evidence. For a historical review of the different approaches, see Carlson (1976), Savage (1976), Spielman (1978), Hall and Sellinger (1986), Zabell (1992) and Lehmann (1993). For completeness, a brief review of the different techniques and common criticisms of them is now provided.

[8]Interestingly, despite the ideological clash concerning testing, the schools reach agreement on estimation and confidence procedures (in terms of the numerical values to report), disagreeing only in the correct interpretation.

[9]Casella and Berger (1987) and Berger and Montera (1999) consider one-sided, i.e. composite null hypotheses with the former highlighting the similarity between results.

Neyman-Pearson approach

Both H_0 and H_1 need to be specified (in order to produce error probabilities and to assess test power). A pre-experimental significance level, α, is chosen which is used to define a critical region, C, and an appropriate test statistic, say T, is used. A simple decision rule follows. Reject H_0 when $t \in C$,[10] otherwise fail to reject. As appropriate, report Type I or Type II errors, α and β respectively. This approach is justified by way of the *frequentist principle*:

> In *repeated*[11] use of a statistical procedure, the long-run average actual error should not be greater than (and ideally should equal) the long-run average reported error.

An oft-cited criticism with the Neyman-Pearson approach is that the error probabilities are fixed *a priori*, and so fail to adequately reflect variation in test statistic values. Also H_1 must be specified to enable computation of Type II errors and consequently power functions, which rely on parameters which are typically unknown. Classical tests specify H_1 and choose n to achieve the desired power, but surely choice of the parameter under H_1, θ_1, is subjective, hence neutralising the critics of Bayesian methods who dispute the use of a prior distribution. A possible remedy is to use a prior distribution for θ_1 and consider average power with respect to this distribution. Also since the goal is to maximise power subject to the pre-experimental α, there exists an asymmetric treatment of the hypotheses.

[10] t is the observed realisation of the random variable T.

[11] An important consideration concerns the repeatability of an experiment, however this may not always be feasible, for example a nuclear war.

Fisher approach

Fisher's approach champions p-values, as defined in Chapter 2, as reflecting the strength of evidence against H_0. Unlike the Neyman-Pearson framework, the (subjective) specification of an alternative hypothesis is not required. The original Fisher approach advocated the replication of small studies, and so false negatives were considered costlier (to society) than false positives. The rationale for this is that a significant result would then be tested many more times, resulting in it being discarded if subsequent rejections failed to occur. A false negative on the other hand would be ignored from the start.

Criticisms include violation of the frequentist principle and the very definition of p-values, i.e. the questionable justification for providing the probability of the data 'or a more extreme value', as remarked upon above. A client is concerned with inference on his/her actual data, not hypothetical data by considering what might have been observed under repeated sampling.

The condition $p < \alpha$ acts as a screen for potentially useful innovations. The original idea of p-value testing in the context of a continuing series of experiments is intuitively sensible. Originally, inference was performed as follows: $p < 0.05$ identified an effect, $p > 0.2$ indicated no effect or one too small to be discovered in an experiment of the current size, inbetween these cases a revision to the experiment would be proposed. In practice most studies are 'single-shot' studies with no replications and any $p > 0.05$ is automatically ignored. However such one-off studies can be combined to form meta-analyses such as the Cochrane Collaboration. Also, because of the potentially high-cost consequences of rejection errors, it is unlikely that in practice high-stakes decisions would be based on a single study.

Jeffreys approach

Jeffreys favoured an alternative hypothesis which allows the Bayes factor, as per (4.3), to be specified. Inference can then be based on a balance-of-probabilities basis, whereby we reject H_0 if $B_{0,1} \leq 1$ and fail to reject if $B_{0,1} > 1$. (Recall the Bayes factor is a likelihood ratio, hence values sub-unity suggest that H_0 is less likely, hence its rejection.) In addition, objective posterior error probabilities are reported. If equal prior probabilities are used, i.e. $\Pr(\mathrm{H}_i|x) = 0.5$, $i = 0, 1$, then the posterior probabilities are,

$$\Pr(\mathrm{H}_0|x) = \frac{B_{0,1}}{1 + B_{0,1}} = \alpha(B_{0,1}), \tag{4.5}$$

$$\Pr(\mathrm{H}_1|x) = \frac{1}{1 + B_{0,1}} = \beta(B_{0,1}). \tag{4.6}$$

Intuitively, a fully accurate subjective prior distribution should result in optimal inferential decision-making. However to be *fully* accurate requires all prior beliefs to be incorporated which in principle means an infinite number of assessments, i.e. $F_\theta(\theta = k)\ \forall\ k \in \Theta \subseteq \mathbb{R}$, for distribution function F and parameter space Θ, need to be reflected in $\pi(\theta)$. Partially-elicited priors, for example $\pi(\theta)$ reflecting particular quartiles or moments, are problematic due to the omitted prior beliefs concerning the remainder of the distribution. Therefore it is appropriate to work with a class of prior distributions, Γ, encompassing the residual uncertainties. Hence use of an objective prior is prudent. However, is it really possible to have a completely impartial prior distribution? Adoption of different types of prior distributions may not achieve such impartiality, for example use of conjugate priors for analytical convenience and tractability. For a collective review of the different approaches, see Berger (2003).

4.3 Conditional Frequentist Testing

Although frequentist and Bayesian methods yield similar results in terms of estimation, this is not so for testing as characterised by the conflict between p-values and conditional measures. The problem with conventional frequentist testing is a lack of suitable conditioning. Berger, Brown, and Wolpert (1994) offer a helpful unification focusing on simple hypothesis tests following in the footsteps of Kiefer (1977) and also Brownie and Kiefer (1977) who propose the conditional confidence approach.[12]

The goal of conditional frequentist testing (CFT) is to obtain agreement over the numerical values to report when performing hypothesis tests, if not agreement in terms of interpretation (i.e. a methodological unification rather than a philosophical one), similar to estimation and confidence procedures.

The unconditional error probabilities α and β in the Neyman-Pearson world suffer from their inflexibility, i.e. Type I and Type II errors fail to distinguish between test statistic values on (or just inside) the critical region boundary and those values deep within it. To remedy this deficiency, Berger, Brown, and Wolpert (1994) recommend reporting the conditional error probabilities given in (4.5) and (4.6). Since $\alpha(B_{0,1})$ and $\beta(B_{0,1})$ are functions of the Bayes factor, the Bayesian influence has now been incorporated into the Neyman-Pearson framework. Birnbaum (1961) referred to these as 'intrinsic significance levels' providing a likelihoodist interpretation.

So the basic conditional test can be summarised as follows[13] for critical value c,

[12] Alternative approaches have been suggested, such as Hwang, Casella, Robert, Wells, and Farrell (1992).

[13] Berger, Brown, and Wolpert (1994), report that CFT can also be used for sequential testing, noting that the Bayes factor is not affected by the chosen stopping rule.

- If $B_{0,1} \leq c$, reject H_0 and report conditional error probability $\alpha(B_{0,1})$.
- If $B_{0,1} > c$, do not reject H_0 and report conditional error probability $\beta(B_{0,1})$.

If the Bayes factor is evaluated on a balance-of-probabilities basis, then set $c = 1$. From a decision-theoretic perspective, consider $\mathrm{H}_0 : \theta \in \Theta_0$ and $\mathrm{H}_1 : \theta \in \Theta_1 \equiv \Omega_\theta \setminus \Theta_0$, for parameter space Ω_θ. Consider the '0-1' loss function, L, on action space $\mathcal{A} = \{a_0, a_1\}$ where $a_0 =$ do not reject H_0 and $a_1 =$ reject H_0. Then,

$$L(\theta, a_0) = \begin{cases} 0 & \text{if} \quad \theta \in \Theta_0, \\ 1 & \text{if} \quad \theta \in \Theta_1, \end{cases} \qquad L(\theta, a_1) = \begin{cases} 1 & \text{if} \quad \theta \in \Theta_0, \\ 0 & \text{if} \quad \theta \in \Theta_1. \end{cases} \tag{4.7}$$

The Bayes decision rule is not to reject H_0, provided

$$\Pr(\theta \in \Theta_0 | x) > \Pr(\theta \in \Theta_1 | x). \tag{4.8}$$

Berger, Brown, and Wolpert (1994) do consider a 'no decision' region[14] for inconclusive values of $B_{0,1}$, i.e. $\epsilon < B_{0,1} < \nu$ for arbitrary constants ϵ and ν, typically such that $\epsilon < 1 < \nu$. However for simplicity it is easier to partition $B_{0,1} \in \mathbb{R}^+$ into solely 'reject' and 'not reject' sets and report a large conditional error probability if $\epsilon < B_{0,1} < \nu$. Readers can then readily interpret the conclusiveness of the test result based on this information themselves.

Up to this point, CFT unifies and satisfies frequentist, likelihoodist and Bayesian principles. It remains to explicitly incorporate p-values in order to offer a plausible methodological unification of hypothesis testing.

[14]Compare with the decision criteria presented in Chapter 3 Table 3.2.

4.3.1 Conditioning

Reid (1995) and Bjørnstad (1996) discuss conditioning, although the following basic example from Berger and Wolpert (1988) nicely highlights the benefits of conditional frequentism. Consider the observations X_i, $i = 1, 2$, such that

$$X_i = \begin{cases} \theta + 1 & \text{with probability} \quad \frac{1}{2}, \\ \theta - 1 & \text{with probability} \quad \frac{1}{2}. \end{cases} \tag{4.9}$$

Define a confidence set, $C(X_1, X_2)$, for the unknown parameter θ as

$$C(X_1, X_2) = \begin{cases} \frac{1}{2}(X_1 + X_2) & \text{if} \quad X_1 \neq X_2 \\ X_1 - 2 & \text{if} \quad X_1 = X_2, \end{cases} \tag{4.10}$$

which yields unconditional frequentist coverage of 0.75. However this can be considerably improved if we condition on the observed data. The sample mean provides the precise value of θ when $x_1 \neq x_2$ with probability 1, yet if $x_1 = x_2$, then all we know is that this observed value is $\theta + 1$ or $\theta - 1$. If we define a *conditioning statistic*, $S = |X_1 - X_2|$, then S can only take two values, i.e. 0 or 2. Hence conditional on S,

$$\Pr_\theta(\theta \in C(X_1, X_2) | S = 0) = \frac{1}{2} \tag{4.11}$$

$$\Pr_\theta(\theta \in C(X_1, X_2) | S = 2) = 1. \tag{4.12}$$

Such conditioning still satisfies frequentist as well as Bayesian ideology through the conditional error probabilities (4.5) and (4.6) which can also be viewed as

$$\begin{aligned}\alpha(B_{0,1}) &= \alpha(s) = \Pr(\text{Type I error}|S(X)=s) = \Pr_0(\text{Reject } \mathrm{H}_0|S(X)=s),\\ \beta(B_{0,1}) &= \beta(s) = \Pr(\text{Type II error}|S(X)=s) = \Pr_1(\text{Not reject } \mathrm{H}_0|S(X)=s).\end{aligned}$$

All that is required now is to introduce p-values into the methodology. Wolpert (1995) and Sellke, Bayarri, and Berger (2001) consider the conditioning statistic for simple hypotheses

$$S = \max\{p_0, p_1\} \tag{4.13}$$

where p_i is the p-value when testing hypothesis H_i against H_{1-i}, $i = 0, 1$. It follows that the decision rule should be

$$\begin{aligned}\text{If } p_0 \leq p_1 &: \text{ Reject } \mathrm{H}_0, \text{report } \alpha(s)\\ \text{If } p_0 > p_1 &: \text{ Do not reject } \mathrm{H}_0, \text{report } \beta(s).\end{aligned}$$

Of course, any strictly increasing function $\psi(p_i)$ would yield the same decision, hence the importance of the use of p-values in (4.13) is less than their interpretation as a measure of evidence in support of a hypothesis.

4.3.2 Alternative conditioning statistic, S

However the conditioning statistic in (4.13) requires two separate hypothesis tests: H_0 v. H_1 and H_1 v. H_0 in order to obtain p_0 and p_1 respectively. A new alternative to this approach presented here is to make use of the second-order p-value, p', as detailed in Chapter 3, Section 3.1. The advantage of doing this is that p' can be

computed directly from $p_0 = p$ as per (3.1) for the general case, which is easily applicable to specific test statistic distributions. Hence the proposed variant of (4.13) is

$$S = \max\{p, p'\}. \tag{4.14}$$

where p is the conventional p-value obtained from testing H_0 against H_1, and p' is the corresponding second-order p-value. Given the p-value density under H_1, $f_P(p|\mathrm{H}_1)$ from (2.4), is readily computable then there should be no additional computational burden of obtaining p' as opposed to p_1. Indeed, since most common hypothesis tests involve Gaussian and t-distributed test statistics, $f_P(p|\mathrm{H}_1)$ is already known for these instances as presented in Chapter 3.

4.3.3 Use of conditional error probabilities in S

The conditional error probabilities, $\alpha(s) = \alpha(B_{0,1})$ and $\beta(s) = \beta(B_{0,1})$, sum to one as evident from (4.5) and (4.6). Therefore it could be said that a conditioning statistic S with p-value arguments is redundant, as decision making could also be based on the conditional error probabilities instead which require the computation of the Bayes factor.

An intuitive decision rule would be to conclude in favour of the hypothesis which minimises the reported conditional error probability, i.e. $\alpha(B_{0,1})$ if H_0 is rejected, or $\beta(B_{0,1})$ otherwise. This leads to the following conditioning statistic,

$$\min\{\alpha(B_{0,1}), \beta(B_{0,1})\}. \tag{4.15}$$

So the corresponding decision rule would be:

$$\text{If } \alpha(B_{0,1}) < \beta(B_{0,1}) \quad : \quad \text{Reject } \mathrm{H}_0, \text{ report } \alpha(B_{0,1}) \tag{4.16}$$

$$\text{If } \alpha(B_{0,1}) \geq \beta(B_{0,1}) \quad : \quad \text{Do not reject } \mathrm{H}_0, \text{ report } \beta(B_{0,1}). \tag{4.17}$$

4.4 Critical p-value Curves and Surfaces

Although the concept of conditional frequentist testing is appealing, for a particular methodology to be widely employed in practice it is necessary to have a simple implementation along qualitative lines, i.e. a simple-to-understand reject or not reject rule.

This section develops a new concept of critical p-value curves and surfaces which can be constructed for simple and composite hypotheses respectively. The idea is to find the p-value which yields equality between the rejection regions under $f_P(p|\mathrm{H}_0)$ and $f_P(p|\mathrm{H}_1)$, p and p' respectively, for a particular effect size δ. Hence if $p = p'$, the researcher should conclude that each hypothesis is equally likely to be true, perhaps resulting in a randomisation or further testing. However, should $p \neq p'$ we should invoke the conditioning statistic in (4.14), but for ease the appropriate conclusion can be quickly established from the curve/surface. The following examples illustrate.

4.4.1 Simple hypotheses with standard Gaussian-distributed test statistics

For $\sqrt{n}\delta > 0$, Figure 3.1 illustrated the typical shape of $f_P(p|\mathrm{H}_1)$. Under H_1, the concentration of the probability density near zero provides the rationale for sufficiently small p-values to warrant rejection of H_0 (lower 'tail' of $f_P(p|\mathrm{H}_0)$) and

sufficiently large p-values to warrant rejection of H_1 (upper tail of $f_P(p|\mathrm{H}_1)$). To obtain the associated critical p-value curve, we solve

$$\begin{aligned} p &= p' && (4.18) \\ &= \Phi(Z_p - \sqrt{n}\delta) \quad \text{(from (3.7))} && (4.19) \end{aligned}$$

which rearranges to

$$\sqrt{n}\delta = Z_p - \Phi^{-1}(p). \tag{4.20}$$

For $\sqrt{n}\delta < 0$, $f_P(p|\mathrm{H}_1)$ is left-skewed, as per Figure 3.3 (i).[15] Consequently sufficiently large p-values warrant rejection of H_0 (upper 'tail' test of H_0), while sufficiently small p-values warrant rejection of H_1 (lower tail test of $f_P(p|\mathrm{H}_1)$). This translates into

$$1 - p = 1 - \Phi(Z_p - \sqrt{n}\delta) \quad \text{(from (3.20))} \tag{4.21}$$

which still rearranges to (4.20). Figure 4.1 plots the corresponding critical p-value curve. So all a researcher needs to do is obtain the conventional p-value from the test statistic in the usual way and determine $\sqrt{n}\delta$ using H_1. Using this information all that is required is to determine where the observed co-ordinates $(p, \sqrt{n}\delta)$ fall in relation to the critical p-value curve. If the point is on the curve itself, then $p = p'$ and so the hypotheses are equally plausible, however if the point departs from the curve, then it is appropriate to reject or not reject H_0 as indicated. Note the x-axis

[15] Note Figure 3.3 depicts t-distributed test statistics, but the standard Gaussian case is just the limiting distribution in terms of degrees of freedom.

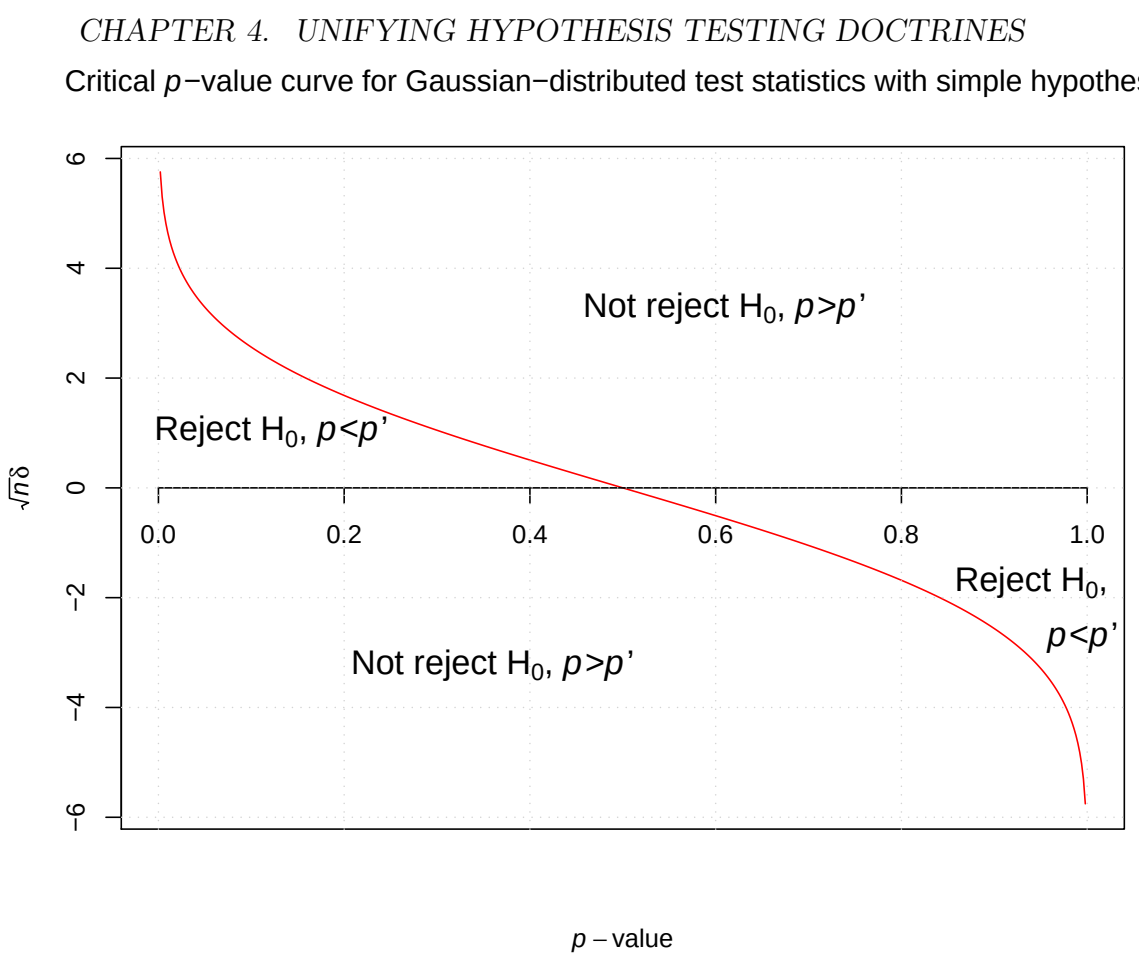


Figure 4.1: Critical p-value curve for standard Gaussian-distributed test statistics with simple forms for H_0 and H_1.

represents the divide between the decision rule. Formally, we have

$$\text{For } \sqrt{n}\delta > 0 : \begin{cases} \text{Reject } H_0, \text{report } \alpha(s) & \text{if} \quad \sqrt{n}\delta < Z_p - \Phi^{-1}(p) \\ \text{Do not reject } H_0, \text{report } \beta(s) & \text{if} \quad \sqrt{n}\delta > Z_p - \Phi^{-1}(p). \end{cases} \tag{4.22}$$

$$\text{For } \sqrt{n}\delta < 0 : \begin{cases} \text{Reject } H_0, \text{report } \alpha(s) & \text{if} \quad \sqrt{n}\delta > Z_p - \Phi^{-1}(p) \\ \text{Do not reject } H_0, \text{report } \beta(s) & \text{if} \quad \sqrt{n}\delta < Z_p - \Phi^{-1}(p). \end{cases} \tag{4.23}$$

Of course, having a graphical depiction of the critical p-value curve removes the need for any formal computation along the lines of (4.22) or (4.23), making

implementation of the methodology fast and simple.[16]

4.4.2 Simple hypotheses with t-distributed test statistics

Critical p-value curves and surfaces involve setting $p = p'$. $f_P(p|\mathrm{H}_0)$ is known to be uniform, while $f_P(p|\mathrm{H}_1)$ depends on the test statistic distribution. In the previous subsection, standard Gaussian test statistics were considered, however curves can be constructed for a variety of distributions. Here critical p-value curves are presented for t-distributed test statistics for various degrees of freedom for simple forms of H_1 in Figure 4.2. As can be seen there is little change in the position of the curve as the degrees of freedom are adjusted, with Figure 4.1 representing the limiting case.

For $p = p'$ in this environment, we seek

$$p = F_T(t_{p,\nu} - \sqrt{n}\hat{\delta}; \nu) \quad \text{(from (3.15))} \tag{4.24}$$

for $\sqrt{n}\hat{\delta} > 0$, with compliments for negative $\sqrt{n}\hat{\delta}$ comparable with (4.21) which still yields (4.24) upon rearrangement, analogous to the standard Gaussian case above.

Formally, the decision rule can be stated as

$$\text{For } \sqrt{n}\hat{\delta} > 0 : \begin{cases} \text{Reject } \mathrm{H}_0, \text{report } \alpha(s) & : \quad \sqrt{n}\hat{\delta} < t_{p,\nu} - T_\nu^{-1}(p) \\ \text{Do not reject } \mathrm{H}_0, \text{report } \beta(s) & : \quad \sqrt{n}\hat{\delta} > t_{p,\nu} - T_\nu^{-1}(p). \end{cases} \tag{4.25}$$

$$\text{For } \sqrt{n}\hat{\delta} < 0 : \begin{cases} \text{Reject } \mathrm{H}_0, \text{report } \alpha(s) & : \quad \sqrt{n}\hat{\delta} > t_{p,\nu} - T_\nu^{-1}(p) \\ \text{Do not reject } \mathrm{H}_0, \text{report } \beta(s) & : \quad \sqrt{n}\hat{\delta} < t_{p,\nu} - T_\nu^{-1}(p). \end{cases} \tag{4.26}$$

Note $t_{p,\nu}$ is the $(1-p)$-th percentile of a Student's t variable on ν degrees of freedom

[16] Obviously the conditional error probabilities, $\alpha(s)$ and $\beta(s)$ will need to be calculated, however the critical p-value curve itself is indicative of the level of significance of a particular p-value.

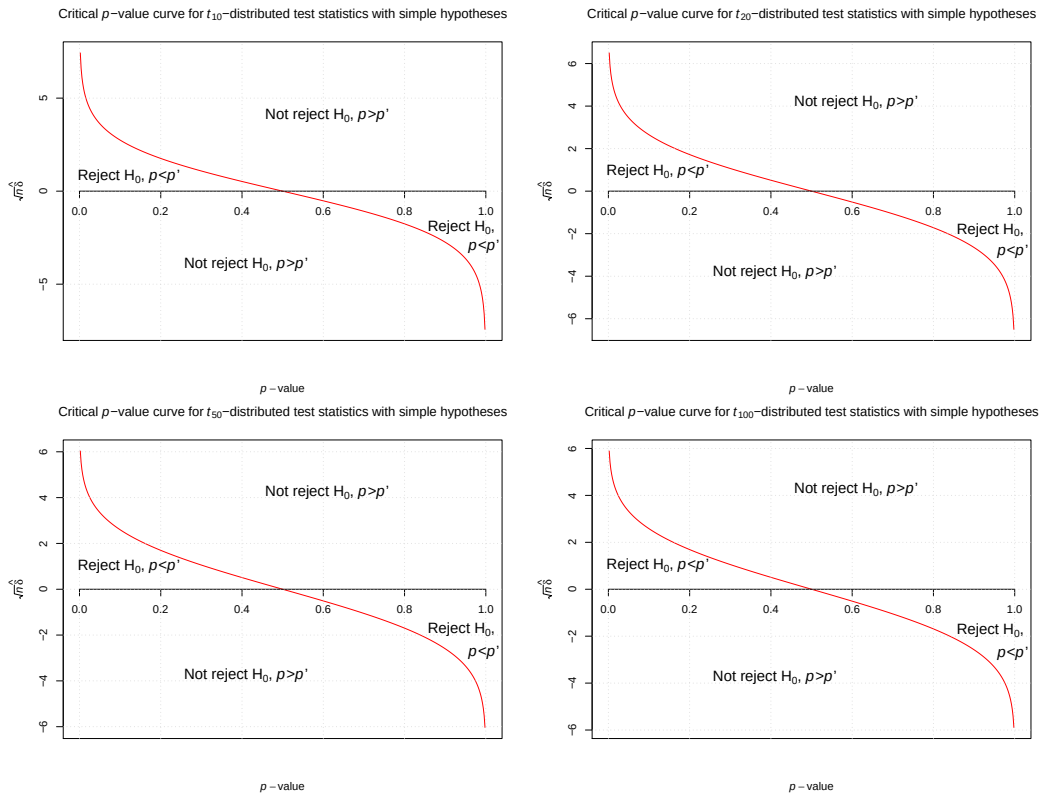


Figure 4.2: Critical p-value curve for t-distributed test statistics with 10, 20, 50 and 100 degrees of freedom with simple forms for H_0 and H_1.

(analogous to Z_p in the standard Gaussian case) and T_ν^{-1} is the quantile function for such a distribution (analogous to Φ^{-1}).

4.4.3 Critical p-value surfaces for composite alternative hypotheses

The critical p-value curves presented above are ideal for testing simple hypotheses. However when testing a null, say, of $\mathrm{H}_0 : \mu = 0$, very often we are interested in composite forms of H_1, for example $\mathrm{H}_1 : \mu \neq 0$. In Chapter 3 the results from Hung, O'Neill, Bauer, and Köhne (1997) concerning the densities of $f_P(p|\mathrm{H}_1)$ when the effect size δ is assumed to follow a probability distribution, such as the Gaussian case, i.e. $\delta \sim N(\zeta, \omega^2)$, were given.

This construction introduces two new parameters, namely ζ and ω^2. In order to provide a graphical depiction for when $p = p'$, it is possible to construct a critical p-value surface by controlling for one of these additional parameters. ζ will be chosen for this purpose.

As Figure 3.4 demonstrated, when the domain of δ under H_1 encompasses both positive and negative values (as is the case for the Gaussian distribution), $f_P(p|\mathrm{H}_1)$ has significant density concentration around both 0 and 1, achieving a minimum in the vicinity of 0.5. In order to accommodate these features it is necessary to re-state the subsets of the respective densities which define the p and p' regions.

Given we seek to reject H_0 when the observed p-value is sufficiently unlikely vis-à-vis H_1, it is necessary to associate extremely small and extremely large p-values with this region, due to the distribution of $f_P(p|\mathrm{H}_1)$. Similarly, p' will be associated with low probabilities of p under $f_P(p|\mathrm{H}_1)$ vis-à-vis $f_P(p|\mathrm{H}_0)$. Such a region exists around $p \approx 0.5$.

Surface for $\zeta = 0$

Controlling for the mean parameter ζ, by setting it to zero the distribution of δ under H_1 is symmetric about zero. To obtain equality between p and p', it is necessary to solve the following,

$$2p = F_P(1-p|\mathrm{H}_1) - F_P(p|\mathrm{H}_1) \quad \text{(from (3.23))}. \tag{4.27}$$

The left-hand-side value of $2p$ represents the rejection region p being comprised of two equal-sized tails (each of area p), the lower covering $[0, p]$ and the upper $[1-p, 1]$ with total area of $2p$. Meanwhile the right-hand-side specifies p', that is the probability of being between p and $1-p$ under H_1, whose area can be computed from the distribution function under H_1.

Consequently the critical p-value surface comprises a floor and a ceiling due to the dual-nature of unlikely p-values under H_0, namely the lower and upper tails. Figure 4.3 presents these. With respect to the floor (ceiling), as the p-value decreases, p decreases ($1-p$ increases), while p' increases, hence for p-values below the floor (above the ceiling) H_0 should be rejected, while it should not be rejected between the floor and ceiling. Formally,

$$\begin{aligned}
\text{Reject } \mathrm{H}_0, \text{report } \alpha(s) \quad \text{if} \quad & \frac{1}{2}\left(\Phi\left(\frac{Z_p - \sqrt{n}\zeta}{(\omega^2 n+1)^{1/2}}\right) - \Phi\left(\frac{Z_{1-p} - \sqrt{n}\zeta}{(\omega^2 n+1)^{1/2}}\right)\right) > p \\
\text{Do not reject } \mathrm{H}_0, \text{report } \beta(s) \quad \text{if} \quad & \frac{1}{2}\left(\Phi\left(\frac{Z_p - \sqrt{n}\zeta}{(\omega^2 n+1)^{1/2}}\right) - \Phi\left(\frac{Z_{1-p} - \sqrt{n}\zeta}{(\omega^2 n+1)^{1/2}}\right)\right) < p.
\end{aligned}$$

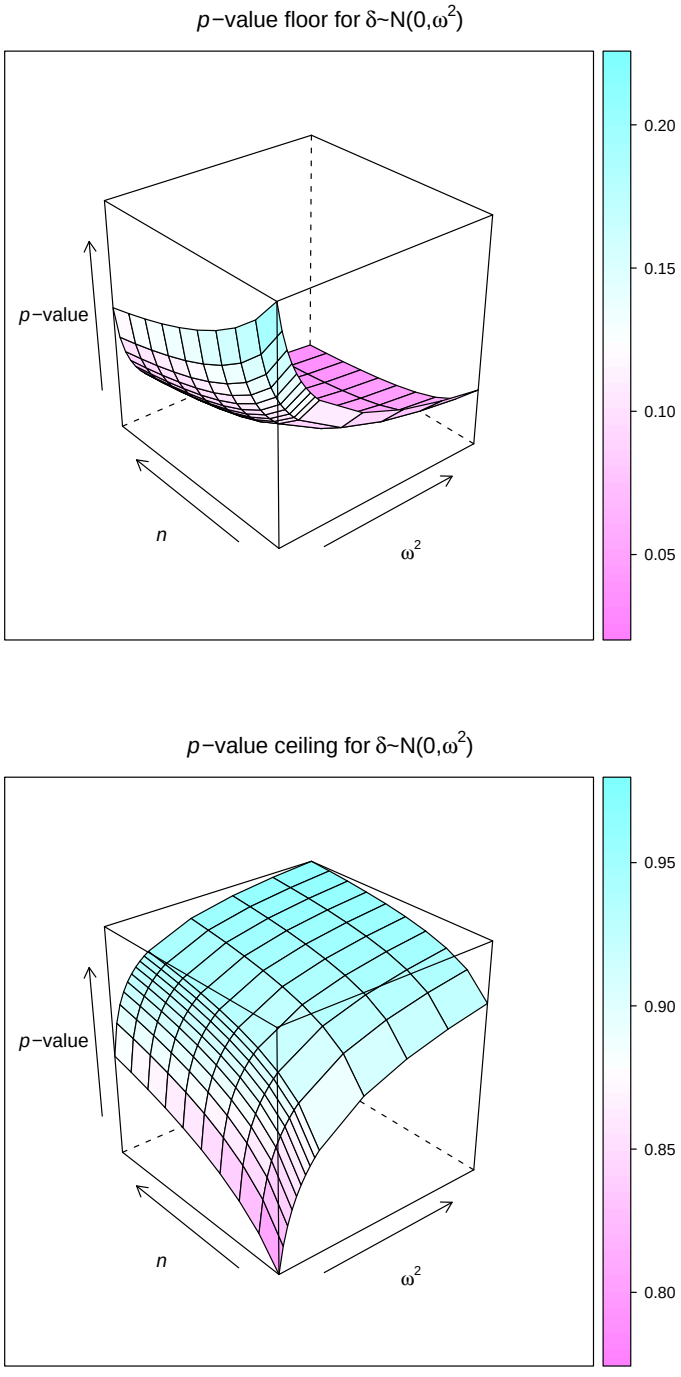


Figure 4.3: Critical p-value surface floor and ceiling for $\delta \sim N(0, \omega^2)$ under H_1.

Surface for non-zero ζ

For $\zeta \neq 0$, $f_P(p|\mathrm{H}_1)$ is no longer symmetric about $p = 0.5$, as demonstrated in Figure 3.4. Therefore the critical p-value floor/ceiling surfaces cannot simply be obtained using p and $1 - p$, as the density $f_P(p|\mathrm{H}_1)$ has different weights around 0 and 1. Therefore to equate p and p', we must solve

$$a + (1 - b) = F_p(b|\mathrm{H}_1) - F_p(a|\mathrm{H}_1), \text{ s.t. } a < m < b, \tag{4.28}$$

where $m = \arg_p f_p(p|\mathrm{H}_1)$. Note $\zeta = 0$ is simply a special case of (4.28) with $a = p$ and $b = 1-p$. Sample floors and ceilings for $\zeta = 0.5$ and $\zeta = -1$ are given in Figure 4.4. Note in particular the behaviour of the ceiling for small values of ω^2. Also, recall each point on the surface corresponds to a different distribution $f_P(p|\mathrm{H}_1)$ due to the specification of changing values for the ω^2 parameter. Interpretation of the surfaces in terms of when to reject H_0 is analogous to the $\zeta = 0$ case.

4.5 Conclusions

This chapter has sought to extend the methodological unification of the Neyman-Pearson, Fisherian and Bayesian schools of hypothesis testing. Although each doctrine has its merits, each also carries limitations, as discussed. To date, the concept of conditional frequentist testing has offered a plausible unification, however results in this chapter extend this methodology by explicitly considering the behaviour of the p-value under H_1 through $f_P(p|\mathrm{H}_1)$.

A variant of an oft-cited conditioning statistic has been proposed making use of the so-called second-order p-value, p', which is obtainable from the original

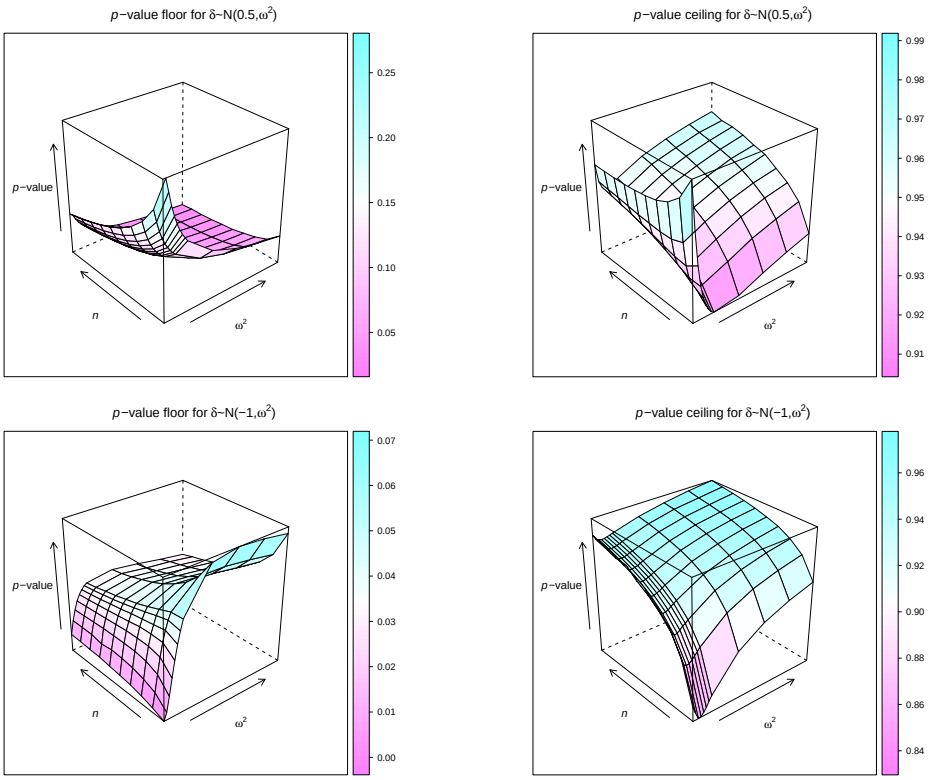


Figure 4.4: Critical p-value floors and ceilings for $\delta \sim N(0.5, \omega^2)$ and $\delta \sim N(-1, \omega^2)$ respectively under H_1.

p-value. By taking the maximum of p and p', the researcher can conclude in favour of the most plausible hypothesis, *conditional on the observed data*. By reporting the conditional error probability in conjunction with the reject/not reject decision, the end-user of the test result can decide the strength of the conclusion themselves.

In order to offer a simple-to-use framework of this methodology, the research above also presents new critical p-value curves and surfaces. By graphically displaying, for a range of parametric specifications relating to H_1, the p-value which results in equality between the p and p' rejection regions under H_0 and H_1 respectively, it is possible to quickly identify the correct decision in relation to whether to reject H_0 while taking into account the specification of H_1 and sample size.

Applying this methodology allows informed decision making, by accommodating the plausibility of *both* hypotheses for a given set of data. Just as in conventional hypothesis testing, inferential errors can occur but these are reflected in the more useful conditional error probabilities. Collectively, this approach helps in the quest for the holy grail of a unified inferential framework universally accepted by proponents of the various testing schools.

Chapter 5

Is Significance Significant? Estimating Inferential Decision Error Rates

Inferential decision errors come in two forms: a Type I error will be deemed a *false discovery*, while a Type II error will be labelled as a *missed discovery*. Ideally we seek to minimise the probability of both occurring, although the trade-off which exists between the two is well-documented. Therefore despite elimination of these errors being infeasible, accurate quantification of the two error rates at least is desirable.

Single hypothesis tests were the subject matter of the preceding chapters and for which the concepts of Type I and Type II errors are well-known. Our attention now turns toward the idea of multiple hypothesis testing. Suppose we simultaneously test m independent hypotheses. For each individual test we will conclude in favour of either the null or alternative hypothesis (ignoring second order p-value-induced indecision possibilities), though on each occasion we are liable to potentially commit a decision error.

Multiple hypothesis testing has received considerable attention in the literature in recent years, fuelled for example by gene expression microarray studies in the biological sciences when potentially thousands of hypotheses are run simultaneously. This chapter will propose a new improved technique for estimating these multiple testing error rates (specifically for large numbers of p-values) by synthesising established methodologies, through introducing the behaviour of p-values under H_1 — an approach not previously considered in the literature.

Before considering the concept of false discovery rates, we begin by revisiting the Bonferroni correction and some of its extensions. Here the idea of a collection of hypotheses satisfying free association can be established and extended to other techniques.

5.1 Bonferroni Procedure Family

The classic Bonferroni inequality establishes an upper bound on the familywise error rate (FWER). The FWER gives the probability that *at least one* false positive occurs after testing a hypothesis series, therefore this equates to FWER = $\Pr(\#\text{ false positives} \geq 1)$. Using the Bonferroni method, we can obtain FWER $\leq \alpha$ for m hypothesis tests by controlling each individual test such that p-values $\leq \frac{\alpha}{m}$ are deemed significant. For m hypothesis tests on the set of test statistics $T_1, \ldots, T_m$ with corresponding p-values $p_1, \ldots, p_m$ we have,

$$\Pr\left(\bigcup_{i=1}^{m}\left(p_i \leq \frac{\alpha}{m}\right)\right) \leq \alpha, \qquad 0 \leq \alpha \leq 1, \tag{5.1}$$

which guarantees an upper bound of α on the probability of rejecting at least one truly null hypothesis overall; equivalently $1-\alpha$ is the lower bound for the probability

of no false discoveries. As can be seen in (5.1), this equates to setting each individual hypothesis test to level $\frac{1}{m} \cdot \alpha$. Hence for $m = 1$, we return to a single α-level test.

Versions of the Bonferroni procedure aimed at improving power include Holm (1979) which comprises a sequentially rejective procedure. For the ordered p-values $p_1, \ldots, p_m$, with corresponding null hypotheses $\mathrm{H}_0^1, \ldots \mathrm{H}_0^m$, we reject H_0^i when

$$p_j \leq \frac{\alpha}{m - j + 1}, \tag{5.2}$$

for all $j = 1, \ldots, i$. (5.2) therefore controls the FWER in the strong sense because it tests all subset intersection hypotheses. In Lehmann and Romano (2005), Chapter 9 features the strong optimality conditions of the Holm procedure. Westfall and Young (1960) also sought to improve power by handling test statistic dependence. For large m though, these power improvements are restricted.

The desirability of the Bonferroni procedure of needing no distributional assumptions is offset however by its too conservative nature, particularly in the presence of highly correlated test statistics. In order to remedy this, Simes (1986) defines the modified Bonferroni procedure which weakly controls the FWER,

$$\Pr\left(\bigcup_{i=1}^{m}\left\{p_i \leq i \cdot \frac{\alpha}{m}\right\}\right) \leq \alpha, \tag{5.3}$$

so that H_0^i is rejected when its associated p-value, p_i, is no greater than $i \cdot \frac{\alpha}{m}$. For independent test statistics, hence independent p-values, this yields a Type I error with probability α. Simulations were run to show an upper bound of α for a range of multivariate distributions — normal, gamma and chi-squared.

Hommel (1988) provides an extended Simes procedure which strongly controls the FWER, making use of the closure principle of Marcus, Peritz, and Gabriel

(1976). In addition, Hochberg (1988) provides a sharper procedure than that of Holm (1979) which is a simplification of the extended Simes procedure. A summary of alternative improvements to the Bonferroni method which yield greater average test power can be found in Shaffer (1995).

5.2 False Discovery Rate

Use of classical single-inference procedures in a multiple-testing context leads to false positive significance rates and hence an exaggeration of reported positive results. The false positive significance rate is derived from the proportion of the results deemed significant for which the null hypothesis is actually true. As such, in order to extend the single hypothesis paradigm to the multi-test setting, it is helpful to use a compound error measure providing a threshold tolerance on the false positive rate, and subsequently develop a method for controlling the error rate whilst retaining test power.

The Bonferroni-related procedures outlined above concern controlling the familywise error rate when testing multiple hypotheses. However, the rather stringent testing of 'at least one' Type I error is for most purposes inadequate, as researchers are usually more concerned about the overall *proportion* of false discoveries. Benjamini and Hochberg (1995) proposed a revised approach to multiple significance testing. They consider a p-value step-up method in which the ordered p-values from m simple, independent hypothesis tests are given such that $p_1 \leq p_2 \leq \ldots \leq p_m$ with associated null hypotheses $\mathrm{H}_0^1, \mathrm{H}_0^2, \ldots, \mathrm{H}_0^m$. If we reject the

Table 5.1: Outcome scenarios for m hypothesis tests with significance level α

	Not reject null	Reject null	Total
Null true	$A(\alpha)$	$U(\alpha)$	m_0
Alternative true	$B(\alpha)$	$T(\alpha)$	m_1
	$R(\alpha)$	$W(\alpha)$	m

k-th p-value, p_k, then it follows that we reject all p_i, $\forall\ i \leq k$. So by defining

$$\hat{k} = \arg \max_{1 \leq k \leq m} \{k : p_k \leq \frac{k}{m} \cdot \alpha\}, \tag{5.4}$$

we reject $\mathrm{H}_0^1, \ldots, \mathrm{H}_0^{\hat{k}}$. No hypotheses are rejected if no such $\hat{k}$ exists. Given the subset of p-values $p_1, \ldots, p_{\hat{k}}$, the False Discovery Rate (FDR) is given by $\frac{m_0}{m_s} \cdot \alpha$ where m_0 factors are truly null, and m_s being the total number of significant results when testing m hypotheses at level α. In their paper, Benjamini and Hochberg (1995) show via simulations that the more relaxed control of Type I errors by way of the FDR is superior to the FWER resulting in greater power and therefore call for control of the FDR instead. Benjamini and Yekutieli (2001) extend this procedure to allow for highly correlated test statistics, i.e. positive regression dependence. Sarkar (2002) extends this work further.

5.2.1 Model set-up

Consider simultaneously testing m null hypotheses, m_0 of which are truly null, while the remaining $m - m_0 = m_1$ have a true alternative. This is represented in Table 5.1 along with the various possible outcomes from hypothesis tests with some designated significance level, α.

The values $U(\alpha)$, $T(\alpha)$ and $W(\alpha)$ are themselves random variables representing the realised quantity of false, genuine and total discoveries respectively, that is all

test statistics with p-values $\leq \alpha$. Naturally these values are monotonically non-decreasing functions of the significance level as when α is made larger, we increase our threshold for acceptable significant p-values and therefore we reject a greater (or at least the same) number of the tested hypotheses. A prudent error measure which accounts for the inclusion of false positives and genuine discoveries, $U(\alpha)$ and $T(\alpha)$ respectively, is

$$\frac{U(\alpha)}{U(\alpha)+T(\alpha)} = \frac{U(\alpha)}{W(\alpha)}. \tag{5.5}$$

Being stochastic variables though, it is appropriate to consider the FDR in terms of expected values. Assuming independence of p-values,

$$\text{FDR} = \text{E}\left[\frac{U(\alpha)}{U(\alpha)+T(\alpha)}\right] = \text{E}\left[\frac{U(\alpha)}{W(\alpha)}\right] \approx \frac{\text{E}[U(\alpha)]}{\text{E}[W(\alpha)]} \quad \text{for large } m. \tag{5.6}$$

Given the possibility that $P(W(\alpha)=0)>0$ is likely to exist in certain circumstances almost surely, (5.6) is undefined when $W(\alpha)=0$. Therefore for the sake of rigour, a non-negativity constraint must be imposed on the denominator to give the positive FDR (pFDR)

$$\text{pFDR} = \text{E}\left[\frac{U(\alpha)}{W(\alpha)}\middle| W(\alpha)>0\right]. \tag{5.7}$$

Benjamini and Hochberg (1995) proposed the following solution to the undefined problem which is the 'official' FDR definition

$$\text{FDR} = \text{E}\left[\frac{U(\alpha)}{W(\alpha)}\middle| W(\alpha)>0\right]\cdot \Pr(W(\alpha)>0). \tag{5.8}$$

Confusion between the two error measures can occur as noted by Zaykin, Young, and Westfall (2000) in response to a study by Weller, Song, Heyen, Lewin, and Ron (1998). However as m increases, $\Pr(W(\alpha)>0)\rightarrow 1$ asymptotically. This is

so because, assuming all m tested hypotheses are independent and truly null, then $\Pr(W(\alpha) = 0) = (1 - \alpha)^m$, which tends to zero as $m \to \infty$, since we are unable to reject any null hypothesis when all p-values are greater than α. Hence the difference between FDR and pFDR becomes trivial. Storey (2002) discusses further.

5.2.2 FDR estimation

In order to estimate the FDR, estimates of $\mathrm{E}[U(\alpha)]$ and $\mathrm{E}[W(\alpha)]$ are required. The total number of 'significant' factors, $W(\alpha)$, is readily obtained by counting the observed number of p-values less than or equal to α which can then be used as an estimate of $\mathrm{E}[W(\alpha)]$. Estimation of $\mathrm{E}[U(\alpha)]$ makes use of the fact that under the simple[1] null hypothesis, say, of a zero parameter, for example $\mathrm{H}_0 : \theta_i = 0$, $i = 1, \ldots, m$, test statistic p-values are uniformly distributed over $[0, 1]$, hence the $U(\alpha)$ p-values are uniformly distributed on $[0, \alpha]$. This argument was shown in Section 2.3.

Given a two-sided framework, then for an amount m_0 of hypotheses which are truly null, $\mathrm{E}[U(\alpha)] = m_0 \cdot \alpha$. This is because we would expect, on average, that this proportion of the sample parameter estimates would yield significant p-values when testing at level α by chance alone when H_0 is true. However, the quantity m_0 from m hypotheses is typically unknown and therefore needs a data-dependent estimate. Define $\pi_0 = \frac{m_0}{m}$ to be the *proportion* of the m hypotheses which are true under the null. Having obtained an estimate for the proportion of true null hypotheses in the

[1] As discussed previously, in order to establish the p-value distribution to be uniform under the null, H_0 must be simple, i.e. single-valued.

population, $\hat{\pi}_0$, the estimated FDR can be computed as

$$\widehat{\text{FDR}}(\alpha) = \frac{\hat{\pi}_0 \cdot m \cdot \alpha}{W(\alpha)} = \frac{\hat{\pi}_0 \cdot m \cdot \alpha}{\#\{p_i \leq \alpha; i = 1, \ldots, m\}}. \tag{5.9}$$

5.3 Missed Discoveries

As previously highlighted, false discoveries represent only half the story with regards to inferential decision errors. Missed discoveries (i.e. Type II errors) are also of concern, therefore a complementary statistic is required to quantify the extent of these. During exploratory analyses, researcher indifference to any missed discoveries can lead to significant (hidden) *opportunity costs* — to borrow from economists' vocabulary. For example, failure to identify a life-saving new drug has clear implications, although people would be completely unaware of the potential life-saving capacity being undetected and subsequently discarded.

Genovese and Wasserman (2002) introduce the dual notion of false non-rejections in multiple testing — the so-called false nondiscovery rate (FNR). This measures the proportion of non-significant hypotheses for which the alternative hypothesis is true. In terms of Table 5.1, the FNR can be thought of as an error measure based on both false negatives and genuine negatives, $B(\alpha)$ and $A(\alpha)$ respectively, denoted

$$\frac{B(\alpha)}{A(\alpha) + B(\alpha)} = \frac{B(\alpha)}{R(\alpha)}. \tag{5.10}$$

As with the FDR, dealing in random variables means expected values must be considered. Again assuming independent p-values,

$$\text{FNR} = \text{E}\left[\frac{B(\alpha)}{A(\alpha) + B(\alpha)}\right] = \text{E}\left[\frac{B(\alpha)}{R(\alpha)}\right] \approx \frac{\text{E}[B(\alpha)]}{\text{E}[R(\alpha)]} \quad \text{for large } m. \tag{5.11}$$

Because $\Pr(R(\alpha) = 0)$ may occur with strictly positive probability almost surely, (5.11) will be undefined when $R(\alpha) = 0$. Once again, a non-negativity constraint is required on the denominator to give the positive FNR (pFNR),

$$\text{pFNR} = \text{E}\left[\frac{B(\alpha)}{R(\alpha)}\middle| R(\alpha) > 0\right], \tag{5.12}$$

as per Storey (2003). Following Benjamini and Hochberg (1995) this can be extended to provide a robust FNR definition,

$$\text{FNR} = \text{E}\left[\frac{B(\alpha)}{R(\alpha)}\middle| R(\alpha) > 0\right] \cdot \Pr(R(\alpha) > 0). \tag{5.13}$$

However as m increases asymptotically, $\Pr(R(\alpha) > 0) \to 1$ implying the difference between FNR and pFNR becomes trivial.

5.3.1 FNR estimation

Estimation of the FNR therefore needs estimates of $\text{E}[B(\alpha)]$ and $\text{E}[R(\alpha)]$ in (5.11). The latter is easily obtainable by counting the observed number of p-values greater than the significance level, α. Of course in order to estimate the FDR for m hypotheses we already have an estimate for $\text{E}[W(\alpha)]$, the observed count $W(\alpha)$, and we require $\text{E}[R(\alpha)] + \text{E}[W(\alpha)] = m$. Therefore $\widehat{\text{E}[R(\alpha)]} = m - W(\alpha)$.

Similarly we have $\widehat{\text{E}[B(\alpha)]} = \widehat{\text{E}[R(\alpha)]} - \widehat{\text{E}[A(\alpha)]}$, where $\widehat{\text{E}[A(\alpha)]} = \hat{\pi}_0 \cdot m \cdot (1-\alpha)$ which makes use of the parameter estimator $\hat{\pi}_0$ used in the estimation of the FDR. So the FNR equivalent of (5.9) is

$$\widehat{\text{FNR}}(\alpha) = \frac{R(\alpha) - \hat{\pi}_0 \cdot m \cdot (1-\alpha)}{R(\alpha)} = 1 - \frac{\hat{\pi}_0 \cdot m \cdot (1-\alpha)}{\#\{p_i > \alpha; i = 1, \ldots, m\}}. \tag{5.14}$$

5.4 Estimation of π_0

Estimation of both false discovery and false nondiscovery rates, in (5.9) and (5.14) respectively, requires a point estimate of $\pi_0 = \frac{m_0}{m}$, the proportion of the m hypotheses that are true under H_0. This section will consider a new, alternative approach to deriving an estimator $\hat{\pi}_0$ taking into account knowledge of the p-value's distribution under H_1 — an approach not previously attempted.

Previous efforts at estimating π_0 have exploited the uniformity of p-values under H_0 and automatically assumed that under H_1 p-values tend to be clustered near zero. As has been examined in earlier chapters, this is not always the case – specifically when the effect size can take negative values in the H_1 space.

5.4.1 Histogram-motivated approach

Assuming p-values under H_1 have a distribution characteristic of Figure 3.1, one approach to estimating π_0, and subsequently the FDR and FNR, focuses on plotting a density histogram of computed p-values and exploiting the uniformity of these under H_0 to obtain a conservative estimate of π_0. Under the assumption of a right-skewed $f_P(p|H_1)$, it is expected that such a histogram would display a relatively flat appearance on the right-hand-side, consistent with the underlying uniform population distribution for p-values of test statistics true under the null hypothesis, $f_P(p|H_0)$.

In principle, (nearly) all p-values located at the upper end of the unit interval are representative of hypotheses for which the null is true. So it is anticipated that the density histogram is relatively flat above some defined nominal threshold level, say λ. Choice of the λ parameter can be made casually by histogram inspection

or heuristically using a data-driven bootstrap procedure in which we minimise the mean-squared error of the estimator, $\hat{\pi}_0$. Details of this procedure can be found in Storey, Taylor, and Siegmund (2004). This method for choosing an appropriate value for the λ tuning parameter is considered more suitable for limited numbers of p-values. Utilisation of a smoother method is employed in Storey and Tibshirani (2003) and offers an alternative approach for $\hat{\pi}_0$ estimation which is conducive to large m.

It is this latter context, i.e. large m, which will be the main focus here. The Storey and Tibshirani (2003) approach involved estimating π_0 for a range of λ values, such as $0, 0.01, 0.02, \ldots, 0.95$, through

$$\hat{\pi}_0 = \frac{\#\{p_j > \lambda\}}{m(1-\lambda)}, \quad j = 1, \ldots, m. \tag{5.15}$$

They note that the plot of $\hat{\pi}_0$ versus λ displays a trade-off between bias and variance when choosing the optimal λ. They also note that for "well formed p-values" the bias of $\hat{\pi}_0$ decreases as λ increases, and so it is concluded that it is sensible to estimate $\lim_{\lambda \to 1} \hat{\pi}_0 \equiv \hat{\pi}_0(\lambda = 1)$. To achieve this they fit a natural cubic spline with three degrees of freedom, $\hat{f}$, to the plot of $\hat{\pi}_0(\lambda)$ against λ to accommodate the bias/variance issue and use the estimator $\hat{\pi}_0 = \hat{f}(1)$.

Consequently, $\hat{\pi}_0$ is a function of λ and is thus given by

$$\hat{\pi}_0(\lambda) = \frac{\hat{R}(\lambda)}{m(1-\lambda)} = \frac{\#\{\hat{p}_i > \lambda; i = 1, \ldots, m\}}{m(1-\lambda)}. \tag{5.16}$$

This estimator is unbiased only if all p-values in the interval $(\lambda, 1]$ are from the truly null population. As the inclusion of a few truly alternative p-values is a possibility, it is likely that $\hat{\pi}_0$ will prove to be a conservative estimator, that is tend to overestimate

π_0.

This data-driven estimator is intuitively appealing for "well formed p-values" under H_1, but negative effect sizes in H_1 result in different shapes for the density $f_P(p|\mathrm{H}_1)$. Therefore it is possible that a density histogram of all m p-values will exhibit relatively flat behaviour in a location other than the right-hand-side. For example the density functions plotted in Figure 3.3 would be consistent with a mirror image about $p = 0.5$ with the left-hand-side displaying a flat appearance.

In such instances, the Storey and Tibshirani (2003) approach would need to be adjusted. This can be easily achieved by symmetry for simple forms of H_1 with negative δ. Now π_0 should be estimated for the range of λ values that is $0.05, 0.06, \ldots,$ 1. This would result in

$$\hat{\pi}_0 = \frac{\#\{p_j < \lambda\}}{m\lambda}, \quad j = 1, \ldots, m. \tag{5.17}$$

By symmetry, the same bias-variance trade-off exists in the plot of $\hat{\pi}_0$ against λ when choosing the optimal λ. Now the bias of $\hat{\pi}_0$ decreases as λ *decreases*, hence we seek $\lim_{\lambda \to 0} \hat{\pi}_0 \equiv \hat{\pi}_0(\lambda = 0)$. Natural cubic spline fitting is again applied, and we should take as our estimator $\hat{\pi}_0 = \hat{f}(0)$.

As things stand, this represents a trivial contribution to the literature. However a new alternative methodology is now considered which is motivated by this spline-fitting approach.

5.4.2 P-value plot regression approach

Schweder and Spjøtvoll (1982) present "P-value plots" (a simple transformation of (half-)normal plots) as a graphical technique for estimating the number of true

null (and therefore alternative) hypotheses when multiple simultaneous testing is performed. Such plots have many applications such as those shown for multiple equality of means, correlation coefficient and contingency table tests.

As previously mentioned on numerous occasions, regardless of the test statistic distribution, under H_0 all p-values are uniformly distributed, whereas under the 'conventional' alternative (i.e. a right-skewed $f_P(p|H_1)$), p-values will tend to be small because a false H_0 should be rejected more often than a true H_0 under a '$p < \alpha$' rejection rule. By letting N_p be the number of p-values strictly greater than p, and T_0 be the (unknown) number of true null hypotheses out of T, $T_0 \leq T$, for sufficiently large p (equivalently sufficiently small $(1-p)$),

$$\mathrm{E}(N_p) \approx T_0(1-p). \tag{5.18}$$

This is analogous to $\mathrm{E}[U(\alpha)] = m_0 \cdot \alpha$ in Section 5.2.2. Consequently there is an approximate linear relationship between $\mathrm{E}(N_p)$ and $(1-p)$ for non-small p.[2] Hence a P-value plot of $\mathrm{E}(N_p)$ against $(1-p)$ should allow the fitting of a line for large values of p (the left-hand side of the plot), the slope of which acts as an estimator for T_0. Of course large values of T_0 will assist line-fitting (the greater T_0/T, the fewer p-values from $f_P(p|H_1)$ to distort the linear relationship), though any p-value correlation will affect the plot variance.[3] For simplicity, we will restrict attention to independent p-values drawn either from $f_P(p|H_0) = \mathrm{Uniform}[0,1]$ or the same $f_P(p|H_1)$. The $T - T_0$ hypotheses true under the alternative, as stated above, will generally be small (for a right-skewed $f_P(p|H_1)$), hence plots are expected to deviate

[2] If *all* the p-values were consistent with H_0 then a P-value plot would approximate a straight line across the entire unit interval, consistent with an empirical distribution function of a continuous uniform distribution over $[0,1]$.

[3] For example, Schweder and Spjøtvoll (1982) note this problem when exhaustively comparing means from a one-way layout as the cross-correlation increases the sampling variation.

upwards from the line for large values of $(1-p)$.

Figure 5.1 displays P-value plots constructed by simulation for 1000 independent p-values drawn from either H_0 or H_1 for various parametric specifications. Plot (i) was generated by setting $\pi_0 = 1$, i.e. all p-values are drawn from test statistics for which H_0 is true. Given $f_P(p|\mathrm{H}_0) = 1$, this P-value plot (plotting N_p against $(1-p)$), thus depicts an empirical distribution function for the continuous uniform distribution over the unit interval. Plots (ii) and (iii) depict P-value plots for a mixture of p-values simulated from $f_P(p|\mathrm{H}_0)$ and $f_P(p|\mathrm{H}_1)$ for positive effect sizes, subject to the weighting factor π_0. Under H_1, the p-values are modelled as being from Gaussian-distributed test statistics with simple forms of H_1, i.e. positive effect sizes corresponding to distribution function (3.6). Given the positive effect sizes shown, the P-value plots display linear behaviour for small values of $(1-p)$, with the upper end of the plots diverging upwards due to the presence of numerous small p-values (equivalently large values of $(1-p)$).

Plot (iv) considers the impact of a negative effect size for a simple H_1 on a typical P-value plot. Now large p-values are likely under the alternative hypothesis, linearity of the P-value plot shifts to the upper end. Hence estimation in such circumstances requires line-fitting to this part of the plot. Therefore unlike plots (ii) and (iii), a best-fitting line would not be forced through the origin.

Finally, plots (v) and (vi) consider composite forms of H_1 with the effect size $\delta \sim N(\zeta, \omega^2)$. Given such forms of $f_P(p|\mathrm{H}_1)$ lead to the majority of the density being split amongst small and large p-values, this translates into linearity of the P-value plot being confined to the central portion.

Just as the histogram approach required the tuning parameter λ, so this regression approach needs to determine the linear range of the p-value plot in order

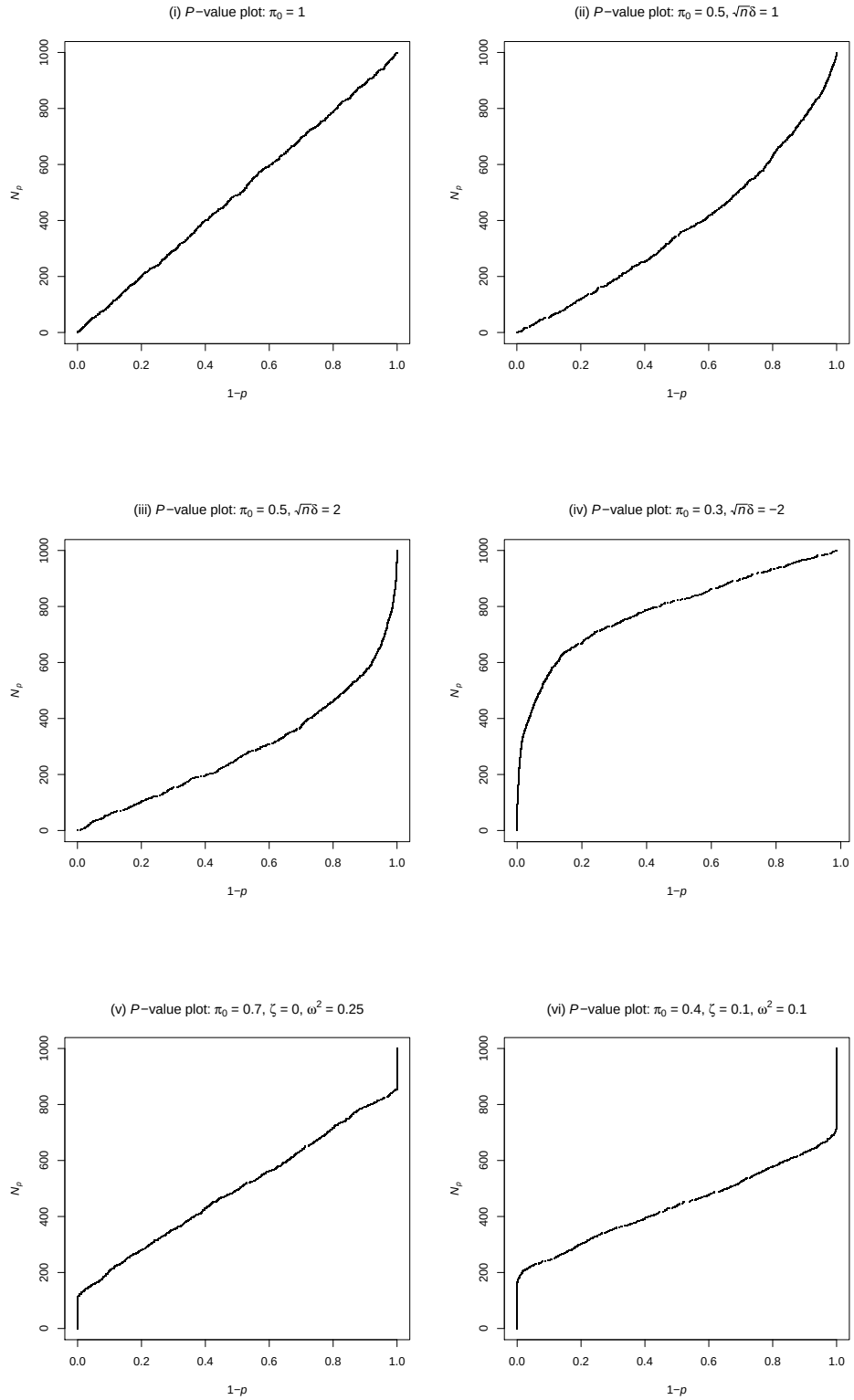


Figure 5.1: Simulated P-value plots of N_p against $(1-p)$, where N_p denotes the number of p-values strictly greater than p for 1000 p-values, for the following cases: (i) $\pi_0 = 1$, i.e. all p-values drawn from $U[0,1]$, (ii) $\pi_0 = 0.5$, $\sqrt{n}\delta = 1$, (iii) $\pi_0 = 0.5$, $\sqrt{n}\delta = 2$, (iv) $\pi_0 = 0.3$, $\sqrt{n}\delta = -2$ where (ii), (iii) and (iv) refer to Gaussian-distributed test statistics under a simple H_1, (v) $\pi_0 = 0.7$, $\zeta = 0.1$, $\omega^2 = 0.25$ and (vi) $\pi_0 = 0.4$, $\zeta = 0.1$, $\omega^2 = 0.1$ where (v) and (vi) refer to composite forms of H_1 where the effect size $\delta \sim N(\zeta, \omega^2)$.

to perform ordinary least squares (OLS) regression. This problem arises as there will be no obvious rigid break (akin to a structural break) and so local regression estimation is required. Schweder and Spjøtvoll (1982) themselves offer no formal approach, merely using a line "drawn by visual fit", although they acknowledged the possibility that the technique "could be formalized using some form of least squares fit". Here a new formal procedure will be outlined, motivated by the spline fitting of Storey and Tibshirani (2003).

The fact that Schweder and Spjøtvoll (1982) resort to fitting the line of best fit visually leaves scope for improvement by automating the choice of $(1-p)$ below which the p-value plot is considered linear[4] (just as Storey and Tibshirani (2003) sought to automate the choice of λ).

5.4.3 P-value plot spline-fitting algorithm for $\hat{\pi}_0$ estimation

A new estimator for π_0 is now proposed based on a spline-fitting procedure applied to regression estimation of P-value plots. The general algorithm for this new estimator is as follows. For m p-values, assumed for simplicity to be independent:

- Sort the p-values into order statistics, i.e. $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$.
- Determine $N_{p_{(i)}}$ for each $p_{(i)}$ (the number of p-values greater than $p_{(i)}$), such that $N_{p_{(i)}} = m - i$. Hence $N_{p_{(i)}} \in \{0, 1, \ldots, m-1\}$.
- Use OLS successively $n = m - 1$ times[5] to estimate the slope coefficient T_0 in

[4] Again, considering 'conventional' right-skewed $f_P(p|\mathrm{H}_1)$ density functions.

[5] A minimum of two points are required for OLS estimation. Suppression of the regression through the origin could lead the origin to be considered as a 'point', however without loss of generality this can be ignored here.

the model

$$N_{p_{(m+1-j)}} = \alpha + T_0(1 - p_{(m+1-j)}) + \epsilon_j, \qquad j = 1, \ldots, n, \qquad \epsilon_j \sim N(0, \sigma_\epsilon^2) \tag{5.19}$$

for $n = 2, \ldots, m$. This yields $m - 1$ separate estimates for T_0. Here for 'conventional' $f_P(p|\mathrm{H}_1)$ (that is as per Figure 3.1) the intercept is suppressed to zero ($\alpha = 0$) to force the line of best fit through the origin.

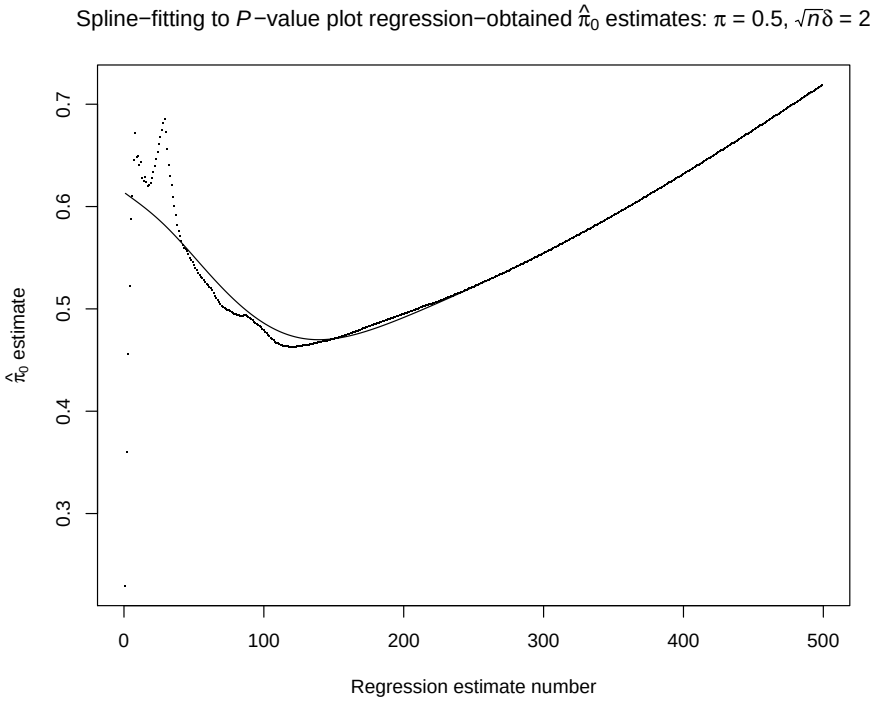
- Fit a cubic smoothing spline with seven[6] degrees of freedom, $\hat{f}$, to $\hat{T}_{0,k}/m$ against k, for $k = 1, \ldots, m-1$, where k denotes the k-th regression coefficient estimate.

- If it exists, obtain the global minimum of $\hat{f}$, $\hat{f}_{\min}$, and let $\hat{\pi}_0 = \hat{f}_{\min}$. If there are multiple local minima, choose the one corresponding to the largest k.

- Otherwise, obtain the saddle point of $\hat{f}$, $\hat{f}_{\text{saddle}}$, and let $\hat{\pi}_0 = \hat{f}_{\text{saddle}}$. If there are multiple saddle points, choose the one corresponding to the largest k.

This algorithm accommodates the spirit of the Storey and Tibshirani (2003) procedure by incorporating spline fitting. Figure 5.2 provides two simulated examples showing how the smoothing spline fits the regression-obtained $\hat{\pi}_0$ point estimate values. This provides a more pleasing fit than the analogous case presented in Storey and Tibshirani (2003) for their histogram-based approach which suffers from wide variability in $\hat{\pi}_0$ estimates as λ tends to 1 (0 in the modified method for $\sqrt{n}\delta < 0$).

[6]Seven degrees of freedom are chosen as this amount provides a pleasing visual fit. Storey and Tibshirani (2003) choose three degrees of freedom in their words to limit the natural cubic spline's curvature "to be like a quadratic function, which is suitable for our purposes". In a similar vein, seven degrees of freedom are suitable for this purpose.



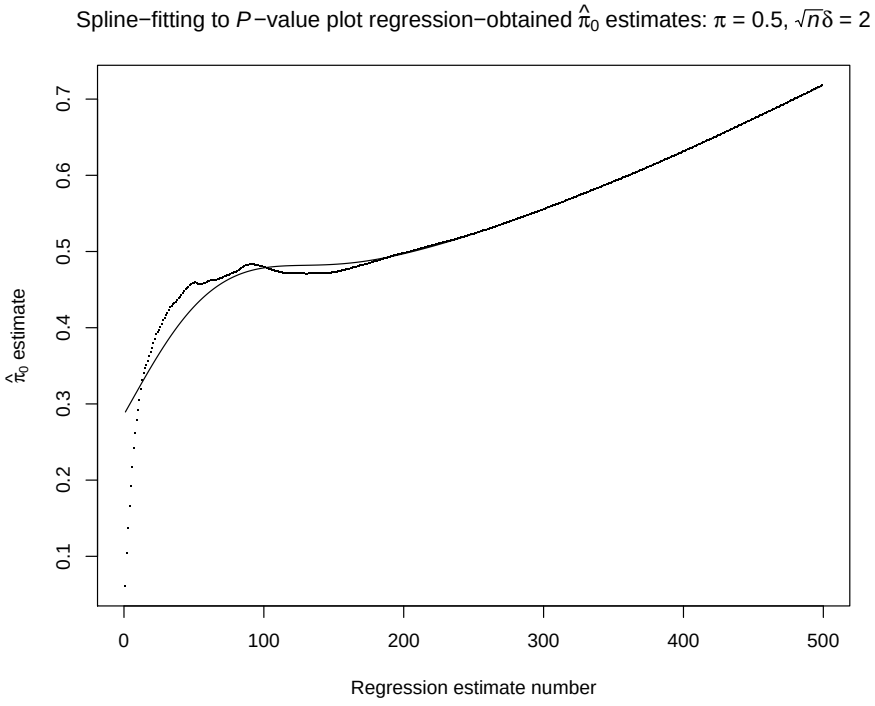


Figure 5.2: These plots show examples of simulated results of the proposed new spline-fitting algorithm. In each example $m = 500$, $\pi_0 = 0.5$, and $\sqrt{n}\delta = 2$. As can be seen, the smoothing spline proves to be a good fit and, following the criteria of the algorithm, both yield point estimates close to 0.5 — specifically 0.470 and 0.482 for the two plots respectively.

It remains to compare these two different estimation techniques for π_0 to determine our preferred estimator. Note attention will be restricted to simple forms of H_1 in order to facilitate a comparison. For composite forms of H_1, the behaviour of $f_P(p|H_1)$ has been well-documented in discussions above. In such instances the Storey and Tibshirani (2003) approach is not readily applicable since in its original form it assumes that $f_P(p|H_1)$ has the shape characteristic of Figure 3.1 (under $\sqrt{n}\delta > 0$), which can easily be transformed for negative $\sqrt{n}\delta$ as discussed above. However, given the p-value density under composite H_1, p-value histograms would not be relatively flat at either end of the unit interval, rather in the middle. Hence the spline-fitting procedure at limit of $\lambda = 1$ (for $\sqrt{n}\delta > 0$) or $\lambda = 0$ (for $\sqrt{n}\delta < 0$) is not appropriate for composite H_1.

Of course, the new regression-based estimator can readily accommodate this issue by taking the median $(1 - p)$ value as a starting point, and then iteratively running regressions taking the nearest $(1 - p)$ value each time. However, since the objective here is to evaluate the performance of two estimators, for which the Storey and Tibshirani (2003) approach is not valid in such circumstances, attention will be restricted to simple forms of H_1, for which direct comparisons can be made.

5.5 Evaluation of Estimators

In order to compare these two estimators for π_0 (either of which can then be used to compute false discovery and non-discovery rates via (5.9) and (5.14) respectively), Monte Carlo simulations are performed for a variety of $\sqrt{n}\delta$ values. Recall that for $\sqrt{n}\delta < 0$, the histogram-based approach of Storey and Tibshirani (2003) has to be modified in light of $f_P(p|H_1)$, as detailed above. Also, the regression-based

estimator, extending the P-value plots of Schweder and Spjøtvoll (1982), differs slightly for positive and negative $\sqrt{n}\delta$ values, since regression lines for $\sqrt{n}\delta > 0$ can have the intercept suppressed to zero as argued from Figure 5.1.

Figures 5.3 to 5.6 provide the simulation results when estimating a range of π_0 values from 0.3 to 0.8 across $0.5 \leq |\sqrt{n}\delta| \leq 5$. The simulated p-values replicate those from Gaussian-distributed test statistics with simple forms of H_0. For each of these values, the mean simulated point estimate is provided, along with the corresponding MSE value to allow a ready comparison of the two estimators. Tabulated values of the results can be found in Appendix B.

It can be seen that in general for $|\sqrt{n}\delta| \leq 2$ both estimators tend to be (quite substantially) positively biased. This is to be expected, since as $\sqrt{n}\delta \to 0$, $f_P(p|\mathrm{H}_1) \to f_P(p|\mathrm{H}_0)$, which of course is uniform over $[0, 1]$. The reason for this upward bias in $\hat{\pi}_0$ values is that the 'flat' region of both histogram and P-value plot is heavily contaminated with drawings from $f_P(p|\mathrm{H}_1)$ which both methods have difficulty distinguishing from p-values drawn from $f_P(p|\mathrm{H}_0)$, that is Uniform$[0, 1]$. As such these estimators will tend to overestimate π_0 for relatively small absolute values of $|\sqrt{n}\delta|$.

As $|\sqrt{n}\delta|$ increases, the more the 'flat' region of the histogram and P-value plot becomes cleansed of p-values from $f_P(p|\mathrm{H}_1)$, and therefore the spline-fitting procedures which are tuned to this region are more accurate in their estimation of π_0, on average. Figures 5.3 to 5.6 illustrate this, although it is noted that as $|\sqrt{n}\delta|$ increases, the Storey and Tibshirani (2003) approach does converge to an unbiased estimator of π_0. This contrasts with the new regression-based approach offered in this chapter, which tends to (slightly) underestimate the parameter — more so for negative $\sqrt{n}\delta$ — suggesting scope for a suitable bias correction method.

Clearly bias is not our only concern when evaluating estimator performance. Estimator variability is also of particular interest, hence we consider alongside the mean of the simulated estimates its mean-squared error (MSE), plotted on the right-hand-side of Figures 5.3 to 5.6.

In all cases, the MSE of both estimators improves (i.e. gets smaller) as $|\sqrt{n}\delta|$ increases, again this is consistent with expectations, since the larger $|\sqrt{n}\delta|$ the lower the likelihood for estimation error due to the 'flat' regions being almost entirely composed of p-values from truly null hypotheses resulting in minimal variation in $\hat{\pi}_0$.

Of particular interest here is the relative performance of the two estimators based on the MSE criterion. In short, for all π_0 values, the new regression-based estimator yields universally lower MSE values for larger values of $|\sqrt{n}\delta|$. That is, we would prefer this new estimator when $|\sqrt{n}\delta|$ is sufficiently large. As can be seen from the plots, this critical point where the new estimator outperforms the Storey and Tibshirani (2003) estimator does vary with π_0, but tends to be around $|\sqrt{n}\delta| = 1.5$.

5.6 Conclusions

Based on the simulation results presented, it can be concluded that neither estimator is uniformly preferred. Instead, the choice of estimator will depend on $\sqrt{n}\delta$ in these multiple hypothesis testing contexts, with $|\sqrt{n}\delta| > 1.5$ indicating that the new regression-based estimator should be employed, while $|\sqrt{n}\delta| \leq 1.5$ should lead to the Storey and Tibshirani (2003) methodology being used.

Of course, noting the bias in $\hat{\pi}_0$ estimates, further research can attempt to determine an appropriate bias correction method for both estimators. Consequently,

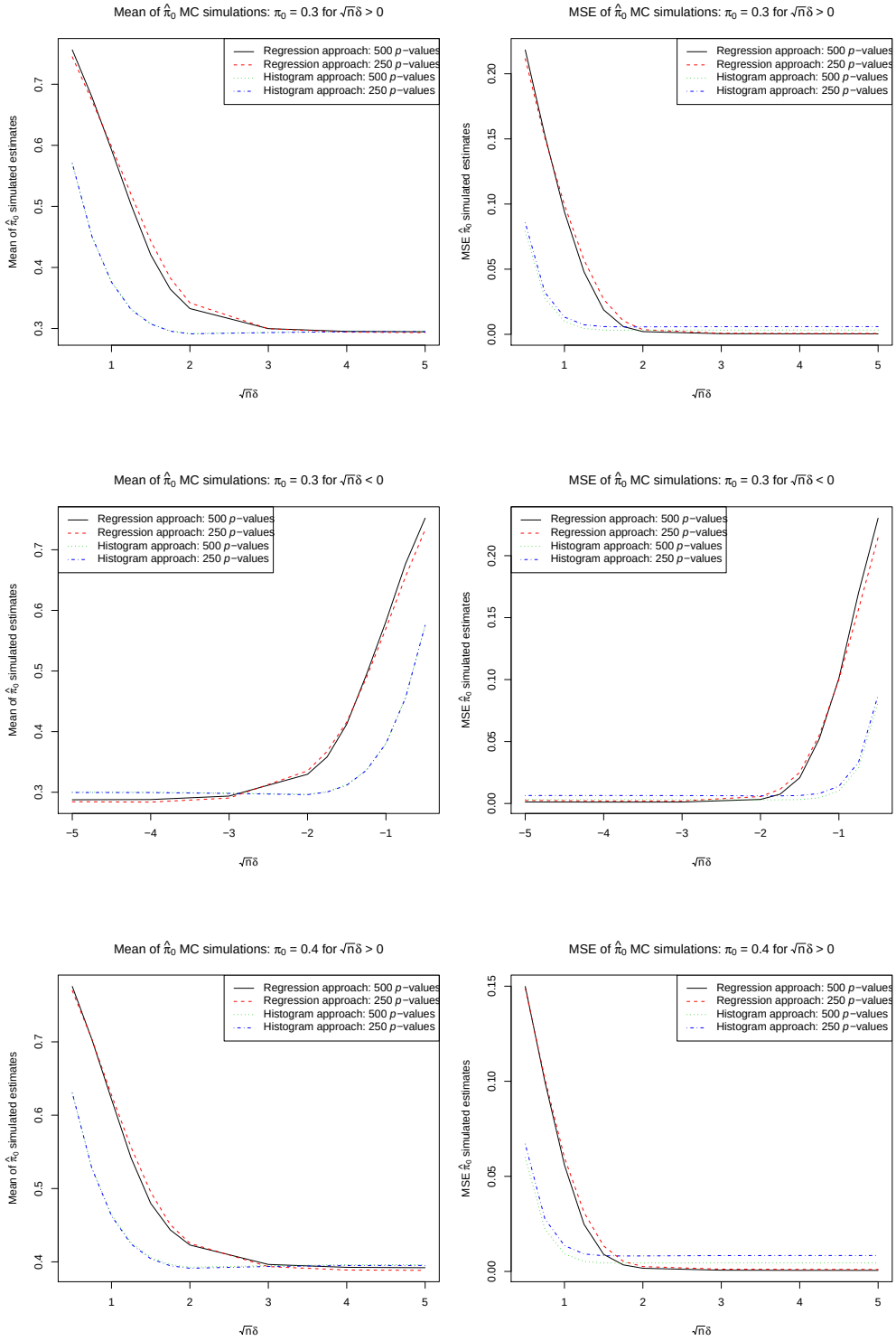


Figure 5.3: $\hat{\pi}_0$ estimation simulation results performed by 1000 Monte Carlo simulations comparing spline-fitting applied to regression based estimators (Regression approach) and to histogram-based estimators (Histogram approach). These methods, designed for large m, were applied to 250 and 500 p-values during each simulation for a range of $\sqrt{n}\delta$ values. Graphics provide mean $\hat{\pi}_0$ estimate in each case (LHS) and corresponding MSE (RHS).

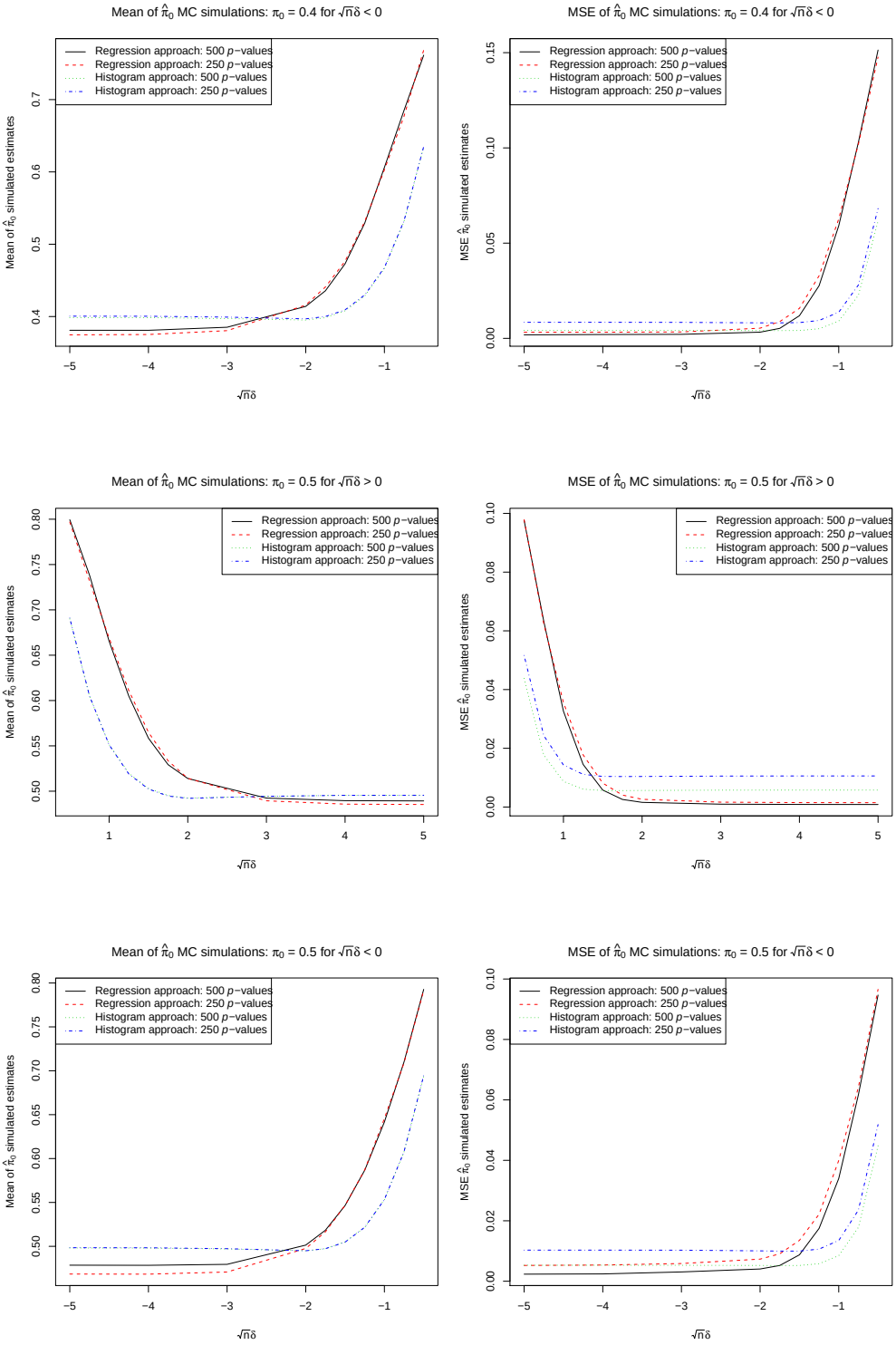


Figure 5.4: $\hat{\pi}_0$ estimation simulation results performed by 1000 Monte Carlo simulations comparing spline-fitting applied to regression based estimators (Regression approach) and to histogram-based estimators (Histogram approach). These methods, designed for large m, were applied to 250 and 500 p-values during each simulation for a range of $\sqrt{n}\delta$ values. Graphics provide mean $\hat{\pi}_0$ estimate in each case (LHS) and corresponding MSE (RHS).

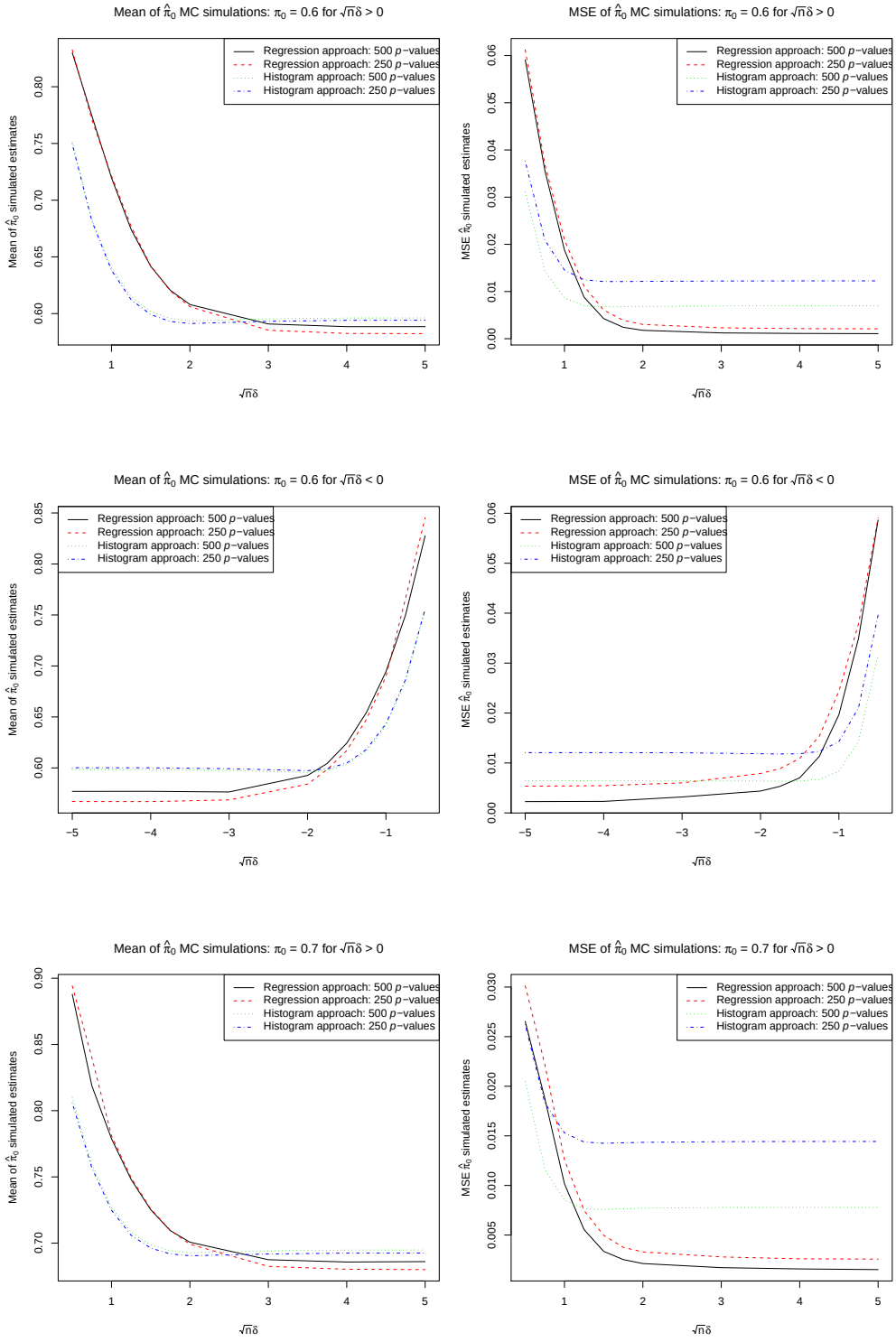


Figure 5.5: $\hat{\pi}_0$ estimation simulation results performed by 1000 Monte Carlo simulations comparing spline-fitting applied to regression based estimators (Regression approach) and to histogram-based estimators (Histogram approach). These methods, designed for large m, were applied to 250 and 500 p-values during each simulation for a range of $\sqrt{n}\delta$ values. Graphics provide mean $\hat{\pi}_0$ estimate in each case (LHS) and corresponding MSE (RHS).

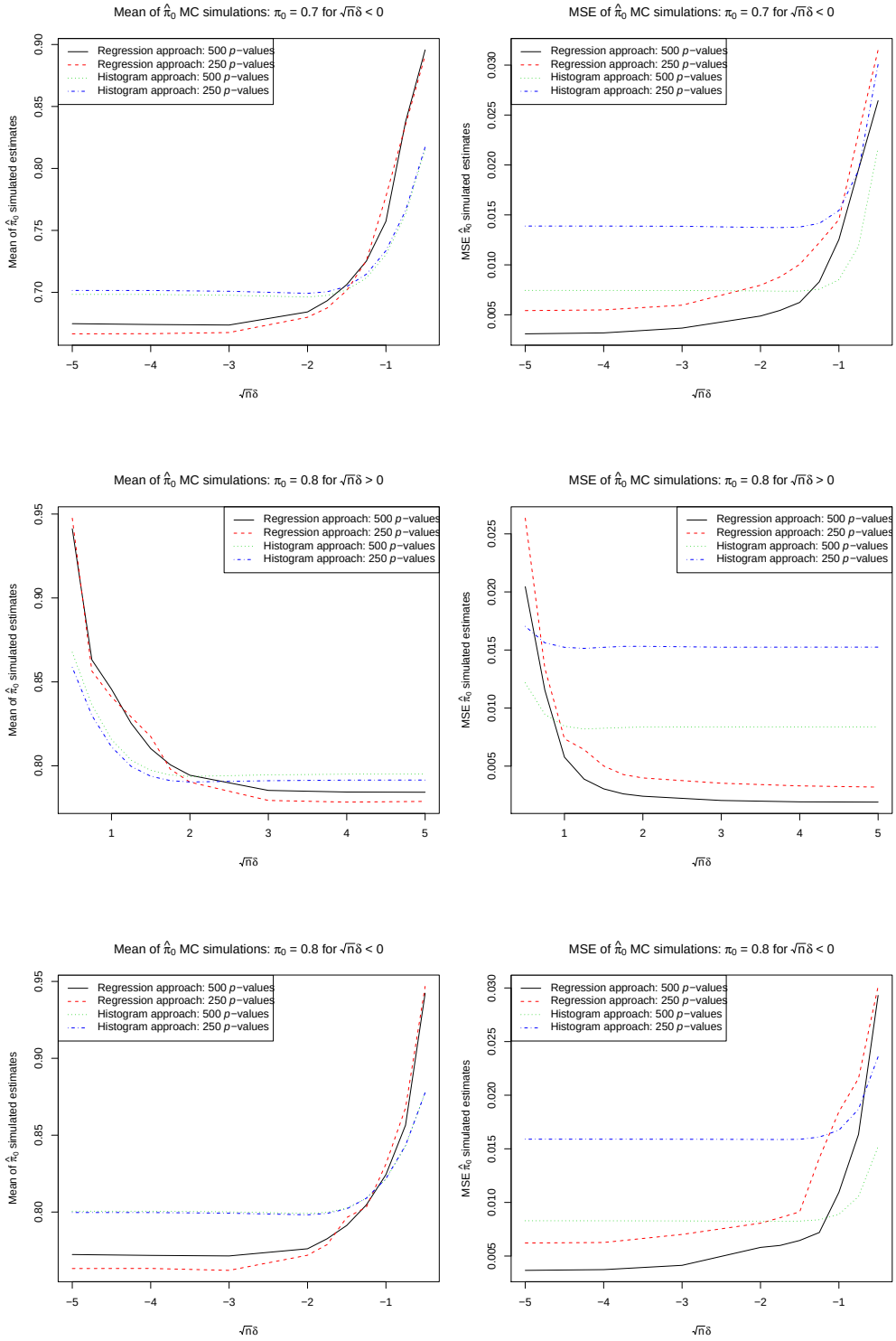


Figure 5.6: $\hat{\pi}_0$ estimation simulation results performed by 1000 Monte Carlo simulations comparing spline-fitting applied to regression based estimators (Regression approach) and to histogram-based estimators (Histogram approach). These methods, designed for large m, were applied to 250 and 500 p-values during each simulation for a range of $\sqrt{n}\delta$ values. Graphics provide mean $\hat{\pi}_0$ estimate in each case (LHS) and corresponding MSE (RHS).

this will lead to a reduction in the MSE. Given that the Storey and Tibshirani (2003) estimator appears to already be unbiased for large $|\sqrt{n}\delta|$, the MSE will not be affected, however there is certainly scope for bias reduction of the new estimator which, combined with its already low relative variance, would provide a very reliable estimator. Also, for smaller values of $|\sqrt{n}\delta|$, it is the excess bias of the new estimator relative to the Storey and Tibshirani (2003) one which contributes to the comparatively larger MSE, hence it is possible that bias reduction of the new estimator could yield a uniformly preferred estimator if the MSE reduction helps to lower the MSE below that of the Storey and Tibshirani (2003) estimator.

It should also be reported that simulation studies were also undertaken for $m = 100$ (recall m is the number of hypotheses being simultaneously tested). However the Storey and Tibshirani (2003) approach frequently failed under repeated simulation due to the failure of the spline-fitting procedure. This problem did not occur when the new estimator was used. Although Storey and Tibshirani (2003) state their estimator is suitable for testing 'large' numbers of hypotheses, 100 is hardly on the small side, hence the new estimator is advantageous down to a smaller m since it can at least return a point estimate in these cases.

Finally, recall the purpose of estimating π_0 in the first place is to allow the subsequent estimation of the FDR and FNR in (5.9) and (5.14) respectively. It therefore follows that an improved estimator for π_0 results in an improved estimator for these multiple hypothesis error rates.[7] Back to this chapter's title, we are in a position to state the following. Question: Is significance significant? Answer: Sometimes.

[7]Appendix C considers a somewhat related concept of multiple error rates along the lines of FDR and FNR applied to a legal framework.

Chapter 6

Conclusions

Here we summarise the key results of the preceding chapters and offer suggestions for future research.

6.1 Summary

The key motivation for this research was to investigate the true meaning of p-values, and ensure their correct interpretation when employed in practice. Their widespread use in empirical research is well-known and this emphasises the need to accurately elicit the full informational content of p-values.

Central to this work is the recognition that an improbable null hypothesis does not by default mean the alternative is more likely. This concept is consistent with exploring the distributional behaviour of p-values under both hypotheses. It was noted early on that p-values are random variables and their distributions under H_1 vary according to whether the alternative hypothesis is simple or composite, and the p-value density is also determined by the sample size and effect size.

The main contributions to the literature are:

- Presentation of a framework for simultaneously testing H_0 and H_1 based solely on the conventional p-value summarised with the construction of critical value tables to facilitate quick inferential decision making.

- Derivation of p-value densities under H_1, $f_P(p|H_1)$, for Student's t-distributed test statistics.

- Extending previous attempts at a methodological unification of the different schools of hypothesis testing, using alternative conditioning statistics.

- Construction of critical p-value curves and surfaces to permit test conclusions to be drawn using visual media.

- For multiple hypothesis testing, a new estimator for the proportion of hypotheses true under H_0 is given and its statistical performance, relative to another estimator, investigated. Under certain conditions the new estimator is preferred. This has subsequent implications for compound error measures.

Collectively the above contributions have endeavoured to 'bridge' different areas of the literature. P-values have been investigated by numerous authors, often from different perspectives. For example, although the behaviour of the p-value under H_1 was researched by Hung, O'Neill, Bauer, and Köhne (1997), its practical relevance was not widely explored. In contrast, Donahue (1999) used second-order p-values, but failed to generalise their implementation.

This thesis has attempted to consolidate and extend earlier thoughts in this domain. By no means is it claimed to be an exhaustive study. Indeed, there is considerable scope for further research into this area. Therefore, it is appropriate to outline some preliminary suggestions for the interested reader.

6.2 Suggestions for Future Research

The p-value density under H_1, $f_P(p|H_1)$, has been a core theme in this research. In the preceding chapters, the primary focus has been on Gaussian-distributed, and by extension Student's t-distributed, test statistics due to their prominent use in inferential testing.

Of course, this does not exhaust all test statistic distributions. Clearly, efforts need to be deployed researching other probability distributions. The χ^2 would be a good candidate, given its widespread use in statistics and econometrics. Also, the F distribution, due to its links with Student's t. Fortunately, all the concepts involved with assessing the plausibility of H_1 are readily applicable to other test statistic distributions, with the only differentiating factor being $f_P(p|H_1)$. The general results contained in Hung, O'Neill, Bauer, and Köhne (1997) make such future research readily achievable.

The multiple hypothesis testing work in Chapter 5 discussed the possibility of applying a bias correction method to the proposed new estimator. Obviously research into a suitable bias correction method is needed. In addition, the simplifying assumption of independent p-values was made such that, under H_1, all p-values were drawn from the same $f_P(p|H_1)$. Inevitably the feasibility of such an assumption in practice is open to criticism. Indeed, there is likely to be considerable heterogeneity of p-values, with clear implications for the π_0 estimation methodologies outlined above. This therefore represents another avenue for future research. Incorporating mixture distributions seems a sensible approach, as these provide a convenient framework to work with when sampling from heterogeneous populations.

6.3 Closing Remarks

Of course any testing procedure based on sample data is liable to yield incorrect conclusions. Given the absence of a perfect world, such errors, though unwanted, are inevitable. Our main objectives have to be to correctly assess the chances of error and to provide a suitable caveat to any inferential conclusions drawn. The use of p-values is well-established, but the general interpretation of them by researchers has room for refinement. So in response to the question 'To p, or not to p?', the conclusion is that we should indeed continue to p, though perhaps in a different way.

Appendix A

Critical Value Tables for Negative Effect Sizes

Table A.1 provides critical p-values, d, for testing $\mathrm{H}_0 : \theta = 0$ against $\mathrm{H}_1 : \theta = k$ for $k < 0$. Since $f_P(p|\mathrm{H}_1)$ is left-skewed, as illustrated in Figure 3.3, testing H_1 corresponds to a lower-tail test hence the interpretation of table values is such that $\Pr(P \leq d|\mathrm{H}_1) = \gamma$.

In the limit as $\sqrt{n}\delta \longrightarrow 0$, i.e. as $\theta|\mathrm{H}_1 \longrightarrow 0$, then this represents convergence to H_0. Therefore the p-value density under the alternative converges to that under the null, that is the continuous uniform distribution. This means that the lower bound on d as $\sqrt{n}\delta \longrightarrow 0$ is γ, akin to the p-value critical value α when testing H_0. Conversely as $\sqrt{n}\delta \longrightarrow \infty$, $f_P(p|\mathrm{H}_1)$ given by (2.5) degenerates on 1. Interpolate as appropriate for $\sqrt{n}\delta$ values not provided.

Table A.2 provides equivalent critical p-values when estimated population variances are used in the test statistic which therefore follows a Student's t distribution. Note the entries in both tables are simply the complement of those in Tables 3.3 and 3.4.

Table A.1: Critical p-values, d, for various levels of γ and selected values of $\sqrt{n}\delta$ when testing, for *known variance*, non-composite hypotheses featuring population mean parameter θ, i.e. $\mathrm{H}_0 : \theta = 0$ and $\mathrm{H}_1 : \theta = k$, such that $\Pr(P < d|\mathrm{H}_1) = \gamma$, where P is the one-tailed p-value of the test statistic under H_0. Under the null, the test statistic is to have a standard Gaussian distribution and under H_1 the test statistic is Gaussian with unit variance. n is sample size and $\delta = k/\sigma$ where $k < 0$ is the hypothesised parameter value of θ under H_1.

$\sqrt{n}\delta$ / γ	-0.25	-0.50	-0.75	-1.00	-1.25	-1.50	-1.75	-2.00	-2.25	-2.50	-2.75	-3.00
0.100	0.1511	0.2172	0.2975	0.3891	0.4874	0.5865	0.6803	0.7638	0.8336	0.8885	0.9290	0.9571
0.050	0.0815	0.1261	0.1854	0.2595	0.3465	0.4424	0.5419	0.6388	0.7275	0.8038	0.8655	0.9123
0.025	0.0436	0.0722	0.1131	0.1685	0.2389	0.3228	0.4168	0.5160	0.6141	0.7054	0.7852	0.8508
0.010	0.0189	0.0339	0.0575	0.0924	0.1409	0.2043	0.2822	0.3721	0.4696	0.5689	0.6641	0.7497
0.005	0.0100	0.0190	0.0339	0.0575	0.0924	0.1410	0.2044	0.2824	0.3723	0.4698	0.5691	0.6643

$\sqrt{n}\delta$ / γ	-3.25	-3.50	-3.75	-4.00	-4.25	-4.50	-4.75	-5.00	-5.25	-5.50	-5.75	-6.00
0.100	0.9755	0.9867	0.9932	0.9967	0.9985	0.9994	0.9997	0.9999	1.0000	1.0000	1.0000	1.0000
0.050	0.9458	0.9682	0.9824	0.9907	0.9954	0.9978	0.9990	0.9996	0.9998	0.9999	1.0000	1.0000
0.025	0.9015	0.9382	0.9633	0.9793	0.9890	0.9945	0.9974	0.9988	0.9995	0.9998	0.9999	1.0000
0.010	0.8222	0.8797	0.9227	0.9529	0.9728	0.9851	0.9923	0.9962	0.9983	0.9992	0.9997	0.9999
0.005	0.7499	0.8223	0.8798	0.9228	0.9530	0.9728	0.9852	0.9923	0.9963	0.9983	0.9992	0.9997

Table A.2: Critical p-values, d, for various levels of γ and selected values of $\sqrt{n}\hat{\delta}$ when testing, *for unknown variance*, non-composite hypotheses featuring population mean parameter θ, i.e. $\mathrm{H}_0 : \theta = 0$ and $\mathrm{H}_1 : \theta = k$ such that $\Pr(P < d|\mathrm{H}_1) = \gamma$, where P is the one-tailed p-value of the test statistic under H_0. Under the null, the test statistic has a t-distribution with $\nu = n - 1$ degrees of freedom and under H_1 the test statistic achieves the same distribution once $\sqrt{n}\hat{\delta}$ has been subtracted. n is sample size and $\hat{\delta} = k/S$ where $k < 0$ is the hypothesised parameter value of θ under H_1, and $S^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2}{n-1}$. Entries in the table give 10% (top), 5% (middle) and 1% (bottom) significance points respectively.

ν / $\sqrt{n}\hat{\delta}$	10	20	30	50	75	100	200	500
-0.25	0.1440	0.1475	0.1487	0.1497	0.1502	0.1504	0.1508	0.1510
	0.0746	0.0779	0.0791	0.0801	0.0805	0.0808	0.0812	0.0814
	0.0154	0.0169	0.0175	0.0181	0.0183	0.0185	0.0187	0.0188
-0.50	0.2018	0.2095	0.2120	0.2141	0.2152	0.2157	0.2165	0.2169
	0.1093	0.1175	0.1203	0.1226	0.1238	0.1244	0.1252	0.1258
	0.0235	0.0280	0.0298	0.0314	0.0322	0.0326	0.0332	0.0336
-0.75	0.2739	0.2857	0.2897	0.2928	0.2944	0.2952	0.2963	0.2970
	0.1565	0.1707	0.1755	0.1795	0.1815	0.1824	0.1839	0.1848
	0.0359	0.0453	0.0491	0.0523	0.0540	0.0548	0.0561	0.0569
-1.00	0.3588	0.3742	0.3792	0.3832	0.3852	0.3862	0.3877	0.3886
	0.2177	0.2385	0.2455	0.2511	0.2539	0.2553	0.2574	0.2587
	0.0541	0.0711	0.0777	0.0834	0.0863	0.0878	0.0901	0.0914
-1.25	0.4526	0.4703	0.4761	0.4807	0.4829	0.4841	0.4857	0.4867
	0.2931	0.3201	0.3289	0.3360	0.3395	0.3413	0.3439	0.3454
	0.0805	0.1079	0.1184	0.1271	0.1316	0.1339	0.1374	0.1395
-1.50	0.5496	0.5684	0.5745	0.5794	0.5817	0.5829	0.5847	0.5858
	0.3806	0.4122	0.4225	0.4305	0.4345	0.4365	0.4395	0.4412
	0.1175	0.1581	0.1730	0.1854	0.1916	0.1948	0.1995	0.2024
-1.75	0.6433	0.6622	0.6683	0.6731	0.6755	0.6767	0.6785	0.6796
	0.4757	0.5100	0.5209	0.5294	0.5336	0.5357	0.5388	0.5406
	0.1673	0.2228	0.2424	0.2583	0.2662	0.2702	0.2762	0.2798
-2.00	0.7279	0.7462	0.7521	0.7568	0.7591	0.7603	0.7620	0.7631
	0.5725	0.6070	0.6179	0.6264	0.6306	0.6326	0.6357	0.6375
	0.2313	0.3017	0.3254	0.3442	0.3536	0.3582	0.3652	0.3693

ν / $\sqrt{n}\hat{\delta}$	10	20	30	50	75	100	200	500
-2.25	0.7997	0.8169	0.8225	0.8270	0.8292	0.8303	0.8319	0.8329
	0.6645	0.6974	0.7077	0.7158	0.7197	0.7217	0.7246	0.7263
	0.3093	0.3919	0.4186	0.4394	0.4496	0.4546	0.4622	0.4666
-2.50	0.8571	0.8730	0.8782	0.8824	0.8844	0.8854	0.8869	0.8879
	0.7463	0.7764	0.7858	0.7931	0.7967	0.7985	0.8011	0.8027
	0.3987	0.4890	0.5169	0.5383	0.5487	0.5539	0.5615	0.5660
-2.75	0.9008	0.9152	0.9198	0.9235	0.9254	0.9263	0.9276	0.9285
	0.8147	0.8413	0.8496	0.8560	0.8592	0.8608	0.8631	0.8645
	0.4946	0.5867	0.6141	0.6349	0.6449	0.6498	0.6570	0.6613
-3.00	0.9327	0.9452	0.9493	0.9524	0.9540	0.9548	0.9560	0.9567
	0.8688	0.8916	0.8987	0.9043	0.9070	0.9083	0.9103	0.9115
	0.5910	0.6790	0.7043	0.7233	0.7324	0.7368	0.7433	0.7472
-3.25	0.9551	0.9657	0.9691	0.9717	0.9730	0.9736	0.9746	0.9751
	0.9094	0.9286	0.9345	0.9391	0.9144	0.9425	0.9441	0.9451
	0.6814	0.7607	0.7829	0.7994	0.8072	0.8111	0.8167	0.8200
-3.50	0.9704	0.9791	0.9818	0.9838	0.9848	0.9853	0.9860	0.9865
	0.9388	0.9545	0.9593	0.9629	0.9647	0.9656	0.9669	0.9677
	0.7607	0.8287	0.8473	0.8610	0.8675	0.8706	0.8752	0.8779
-3.75	0.9806	0.9875	0.9896	0.9911	0.9918	0.9922	0.9927	0.9930
	0.9593	0.9718	0.9755	0.9784	0.9797	0.9804	0.9814	0.9820
	0.8264	0.8820	0.8970	0.9079	0.9131	0.9156	0.9192	0.9213
-4.00	0.9874	0.9927	0.9942	0.9953	0.9958	0.9960	0.9964	0.9966
	0.9732	0.9830	0.9858	0.9879	0.9889	0.9894	0.9901	0.9905
	0.8777	0.9217	0.9333	0.9417	0.9456	0.9475	0.9502	0.9518
-4.25	0.9918	0.9958	0.9969	0.9976	0.9979	0.9981	0.9983	0.9984
	0.9825	0.9899	0.9920	0.9935	0.9942	0.9945	0.9950	0.9952
	0.9160	0.9498	0.9584	0.9646	0.9675	0.9689	0.9709	0.9720
-4.50	0.9946	0.9976	0.9983	0.9988	0.9990	0.9991	0.9992	0.9993
	0.9886	0.9942	0.9956	0.9966	0.9971	0.9973	0.9976	0.9977
	0.9434	0.9687	0.9750	0.9795	0.9815	0.9824	0.9838	0.9846
-4.75	0.9965	0.9987	0.9991	0.9994	0.9995	0.9996	0.9997	0.9997
	0.9926	0.9967	0.9976	0.9983	0.9986	0.9987	0.9989	0.9990
	0.9625	0.9810	0.9855	0.9885	0.9899	0.9905	0.9915	0.9920
-5.00	0.9977	0.9992	0.9996	0.9997	0.9998	0.9998	0.9999	0.9999
	0.9952	0.9981	0.9988	0.9992	0.9993	0.9994	0.9995	0.9996
	0.9753	0.9887	0.9918	0.9938	0.9947	0.9951	0.9957	0.9960

Appendix B

Monte Carlo Simulation Results

B.1 Estimation of π_0

Tabulated values below report Monte Carlo simulation results for estimating π_0, the proportion of tested hypotheses for which H_0 is true. Two estimators are employed (i) a new regression-based estimator (RBE) with its origins rooted in the P-value plots of Schweder and Spjøtvoll (1982) documented in Chapter 5, and (ii) the histogram-based estimator (HBE) of Storey and Tibshirani (2003), again detailed in Chapter 5. Graphical depictions of these results are presented in Figures 5.3 to 5.6.

Tables B.1 to B.6 provide results for estimating a specific value of π_0 from 0.3 to 0.8 for various values of $\sqrt{n}\delta$. These estimators are designed for large numbers of (independent) p-values when multiple hypothesis testing is performed. As such simulations were conducted for 500 and 250 p-values in each run for 1,000 simulations in each case. The values reported are the mean $\hat{\pi}_0$ estimate and the corresponding mean squared error (MSE).

Table B.1: π_0 is the parameter to be estimated. m = number of p-values per simulation. RBE = regression-based approach, HBE = histogram-based approach. Tabulated values are the mean estimates, $\hat{\pi}_0$, with their respective mean-squared errors in parentheses.

$\pi_0 = 0.3$ $\sqrt{n}\delta$	RBE $m = 500$	HBE $m = 500$	RBE $m = 250$	HBE $m = 250$
5	0.294625	0.295202	0.293257	0.294840
	(0.000362)	(0.003278)	(0.000636)	(0.005918)
4	0.295244	0.295087	0.294009	0.294729
	(0.000373)	(0.003277)	(0.000653)	(0.005916)
3	0.299778	0.293655	0.299640	0.293242
	(0.000467)	(0.003254)	(0.000783)	(0.005870)
2	0.332633	0.291843	0.341946	0.291242
	(0.002058)	(0.003158)	(0.003550)	(0.005712)
1.75	0.364344	0.296355	0.382732	0.295558
	(0.005971)	(0.003125)	(0.010330)	(0.005704)
1.50	0.420855	0.308015	0.443446	0.307412
	(0.018721)	(0.003245)	(0.026822)	(0.005923)
1.25	0.502636	0.332074	0.519437	0.330933
	(0.047878)	(0.004425)	(0.057021)	(0.007301)
1	0.592749	0.376631	0.598263	0.375934
	(0.093977)	(0.009834)	(0.099296)	(0.013224)
0.75	0.679392	0.451059	0.673363	0.451030
	(0.152922)	(0.027707)	(0.150809)	(0.032210)
0.50	0.756327	0.569598	0.744975	0.570707
	(0.218333)	(0.078954)	(0.211588)	(0.085701)
-0.50	0.752144	0.575230	0.733115	0.575372
	(0.230299)	(0.081669)	(0.214823)	(0.087500)
-0.75	0.676872	0.455849	0.655872	0.455717
	(0.169969)	(0.028847)	(0.156732)	(0.033240)
-1	0.581677	0.380738	0.570129	0.101085
	(0.101085)	(0.010177)	(0.100091)	(0.013941)
-1.25	0.493613	0.336334	0.488696	0.336026
	(0.052261)	(0.004547)	(0.055166)	(0.008034)
-1.50	0.412293	0.312427	0.415909	0.311623
	(0.020749)	(0.003138)	(0.025087)	(0.006433)
-1.75	0.358284	0.301054	0.367386	0.300352
	(0.007490)	(0.002938)	(0.011492)	(0.006152)
-2	0.329609	0.296846	0.335230	0.296027
	(0.003269)	(0.002951)	(0.005577)	(0.006152)
-3	0.293682	0.298805	0.290416	0.298020
	(0.001168)	(0.003042)	(0.002115)	(0.006392)
-4	0.287941	0.300234	0.283720	0.299428
	(0.001105)	(0.003066)	(0.002089)	(0.006442)
-5	0.287593	0.300339	0.284346	0.299537
	(0.001212)	(0.003068)	(0.002514)	(0.006445)

Table B.2: π_0 is the parameter to be estimated. m = number of p-values per simulation. RBE = regression-based approach, HBE = histogram-based approach. Tabulated values are the mean estimates, $\hat{\pi}_0$, with their respective mean-squared errors in parentheses.

$\pi_0 = 0.4$ $\sqrt{n}\delta$	RBE $m = 500$	HBE $m = 500$	RBE $m = 250$	HBE $m = 250$
5	0.392053	0.396220	0.388397	0.395094
	(0.000557)	(0.004530)	(0.001043)	(0.008377)
4	0.392544	0.396120	0.388938	0.395000
	(0.000569)	(0.004529)	(0.001068)	(0.008375)
3	0.396500	0.394873	0.393501	0.393701
	(0.000651)	(0.004501)	(0.001196)	(0.008331)
2	0.422951	0.392814	0.425292	0.391298
	(0.001671)	(0.004409)	(0.002562)	(0.008149)
1.75	0.443534	0.396390	0.450726	0.394660
	(0.003447)	(0.004384)	(0.005302)	(0.008131)
1.50	0.479798	0.405988	0.495153	0.404282
	(0.008969)	(0.004456)	(0.013398)	(0.008238)
1.25	0.541749	0.426301	0.556238	0.424187
	(0.024762)	(0.005263)	(0.031005)	(0.009194)
1	0.621745	0.464422	0.627614	0.462833
	(0.055975)	(0.009207)	(0.059821)	(0.013588)
0.75	0.703365	0.528090	0.703724	0.527491
	(0.099983)	(0.022310)	(0.101506)	(0.027655)
0.50	0.775620	0.629811	0.769932	0.630331
	(0.149885)	(0.059940)	(0.148898)	(0.066987)
-0.50	0.761467	0.634508	0.768546	0.635158
	(0.151381)	(0.061490)	(0.147676)	(0.068138)
-0.75	0.685203	0.531410	0.677096	0.532519
	(0.102949)	(0.022625)	(0.101942)	(0.028026)
-1	0.606155	0.466478	0.602884	0.467656
	(0.059100)	(0.009031)	(0.062725)	(0.013720)
-1.25	0.529143	0.428331	0.531378	0.429864
	(0.027616)	(0.005059)	(0.032856)	(0.009400)
-1.50	0.472852	0.407944	0.476519	0.409229
	(0.011885)	(0.004123)	(0.015740)	(0.008279)
-1.75	0.435601	0.398448	0.440865	0.400015
	(0.005245)	(0.004050)	(0.008797)	(0.008142)
-2	0.414025	0.395123	0.415822	0.396808
	(0.003254)	(0.004076)	(0.005351)	(0.008151)
-3	0.385187	0.397217	0.380487	0.399192
	(0.002108)	(0.004139)	(0.003363)	(0.008393)
-4	0.380969	0.398455	0.375144	0.400417
	(0.002017)	(0.004157)	(0.003229)	(0.008429)
-5	0.380983	0.398545	0.374686	0.400507
	(0.001849)	(0.004158)	(0.003204)	(0.008431)

Table B.3: π_0 is the parameter to be estimated. m = number of p-values per simulation. RBE = regression-based approach, HBE = histogram-based approach. Tabulated values are the mean estimates, $\hat{\pi}_0$, with their respective mean-squared errors in parentheses.

$\pi_0 = 0.5$ $\sqrt{n}\delta$	RBE $m = 500$	HBE $m = 500$	RBE $m = 250$	HBE $m = 250$
5	0.489270	0.495539	0.485326	0.495536
	(0.000857)	(0.005753)	(0.001477)	(0.010583)
4	0.489542	0.495455	0.485721	0.495456
	(0.000880)	(0.005753)	(0.001506)	(0.010580)
3	0.492404	0.494388	0.489413	0.494362
	(0.000967)	(0.005735)	(0.001659)	(0.010540)
2	0.514030	0.492470	0.514231	0.492088
	(0.001588)	(0.005647)	(0.002628)	(0.010399)
1.75	0.529251	0.495305	0.533500	0.494630
	(0.002574)	(0.005609)	(0.004071)	(0.010381)
1.50	0.558486	0.503101	0.564979	0.502396
	(0.005767)	(0.005607)	(0.008130)	(0.010421)
1.25	0.604987	0.519859	0.611125	0.519047
	(0.014508)	(0.006084)	(0.017771)	(0.011126)
1	0.665468	0.551543	0.668892	0.551155
	(0.032623)	(0.008708)	(0.035563)	(0.014315)
0.75	0.738593	0.604738	0.732080	0.605318
	(0.063242)	(0.017691)	(0.062406)	(0.024299)
0.50	0.799462	0.689974	0.796872	0.691006
	(0.097634)	(0.043796)	(0.097852)	(0.051627)
-0.50	0.792746	0.694062	0.791635	0.694670
	(0.094697)	(0.044828)	(0.096575)	(0.051825)
-0.75	0.710159	0.607655	0.709160	0.607991
	(0.061695)	(0.017812)	(0.064225)	(0.023525)
-1	0.641628	0.553271	0.645586	0.553308
	(0.033982)	(0.008459)	(0.039964)	(0.013477)
-1.25	0.586701	0.521568	0.587118	0.521729
	(0.017470)	(0.005787)	(0.022168)	(0.010629)
-1.50	0.546465	0.504995	0.546676	0.504772
	(0.008731)	(0.005221)	(0.013516)	(0.009972)
-1.75	0.518349	0.497400	0.516508	0.497323
	(0.005206)	(0.005194)	(0.009067)	(0.009974)
-2	0.501390	0.494918	0.497593	0.494957
	(0.004048)	(0.005227)	(0.007314)	(0.010049)
-3	0.479473	0.497014	0.470743	0.497313
	(0.003072)	(0.005273)	(0.005853)	(0.010271)
-4	0.478391	0.498036	0.468297	0.498348
	(0.002415)	(0.005290)	(0.005354)	(0.010295)
-5	0.478527	0.498110	0.468428	0.498423
	(0.002335)	(0.005292)	(0.005226)	(0.010296)

Table B.4: π_0 is the parameter to be estimated. m = number of p-values per simulation. RBE = regression-based approach, HBE = histogram-based approach. Tabulated values are the mean estimates, $\hat{\pi}_0$, with their respective mean-squared errors in parentheses.

$\pi_0 = 0.6$ $\sqrt{n}\delta$	RBE $m = 500$	HBE $m = 500$	RBE $m = 250$	HBE $m = 250$
5	0.588488	0.596213	0.582318	0.594131
	(0.001073)	(0.006955)	(0.002092)	(0.012258)
4	0.588469	0.596146	0.582506	0.594066
	(0.001116)	(0.006954)	(0.002144)	(0.012255)
3	0.590849	0.595285	0.585318	0.593198
	(0.001226)	(0.006940)	(0.002314)	(0.012227)
2	0.607999	0.593608	0.606099	0.591223
	(0.001773)	(0.006858)	(0.003017)	(0.012151)
1.75	0.620563	0.595698	0.619926	0.593070
	(0.002415)	(0.006801)	(0.003904)	(0.012117)
1.50	0.641714	0.601823	0.642156	0.599103
	(0.004263)	(0.006761)	(0.006054)	(0.012115)
1.25	0.674264	0.614991	0.676958	0.612284
	(0.008794)	(0.006983)	(0.011122)	(0.012491)
1	0.719785	0.640251	0.721532	0.638135
	(0.018795)	(0.008624)	(0.020969)	(0.014551)
0.75	0.774301	0.682899	0.771066	0.681597
	(0.035667)	(0.014353)	(0.036962)	(0.020883)
0.50	0.829931	0.751234	0.832525	0.749489
	(0.059104)	(0.031097)	(0.061213)	(0.037591)
-0.50	0.827655	0.754523	0.845235	0.755998
	(0.058656)	(0.031707)	(0.059087)	(0.039624)
-0.75	0.749951	0.684948	0.766343	0.686717
	(0.034932)	(0.014318)	(0.037855)	(0.021033)
-1	0.694348	0.641496	0.689487	0.643303
	(0.019666)	(0.008377)	(0.024361)	(0.014291)
-1.25	0.654220	0.616298	0.647498	0.618104
	(0.011318)	(0.006701)	(0.015358)	(0.012313)
-1.50	0.624306	0.603324	0.617060	0.604847
	(0.007029)	(0.006369)	(0.010911)	(0.011842)
-1.75	0.604319	0.597470	0.598387	0.599117
	(0.005324)	(0.006357)	(0.008873)	(0.011805)
-2	0.592618	0.595594	0.584108	0.597336
	(0.004376)	(0.006380)	(0.007899)	(0.011848)
-3	0.576453	0.597487	0.568658	0.599310
	(0.003197)	(0.006418)	(0.006019)	(0.012047)
-4	0.577055	0.598307	0.566974	0.600142
	(0.002308)	(0.006435)	(0.005451)	(0.012070)
-5	0.577009	0.598364	0.567054	0.600202
	(0.002262)	(0.006437)	(0.005358)	(0.012072)

Table B.5: π_0 is the parameter to be estimated. m = number of p-values per simulation. RBE = regression-based approach, HBE = histogram-based approach. Tabulated values are the mean estimates, $\hat{\pi}_0$, with their respective mean-squared errors in parentheses.

$\pi_0 = 0.7$ $\sqrt{n}\delta$	RBE $m = 500$	HBE $m = 500$	RBE $m = 250$	HBE $m = 250$
5	0.686051	0.694864	0.680018	0.692515
	(0.001514)	(0.007786)	(0.002573)	(0.014432)
4	0.685751	0.694814	0.680288	0.692464
	(0.001583)	(0.007786)	(0.002602)	(0.014431)
3	0.687496	0.694162	0.682475	0.691825
	(0.001711)	(0.007778)	(0.002791)	(0.014414)
2	0.700714	0.692760	0.699187	0.690485
	(0.002124)	(0.007705)	(0.003293)	(0.014342)
1.75	0.709461	0.694228	0.709706	0.691869
	(0.002524)	(0.007648)	(0.003760)	(0.014300)
1.50	0.725249	0.698614	0.725882	0.696265
	(0.003355)	(0.007583)	(0.004974)	(0.014249)
1.25	0.748127	0.708222	0.749504	0.705941
	(0.005546)	(0.007643)	(0.007472)	(0.014376)
1	0.778880	0.727201	0.781180	0.724911
	(0.010169)	(0.008519)	(0.012598)	(0.015301)
0.75	0.818718	0.759233	0.840045	0.756702
	(0.018756)	(0.011647)	(0.022193)	(0.018307)
0.50	0.887856	0.810069	0.894241	0.805879
	(0.026535)	(0.020459)	(0.030123)	(0.026198)
-0.50	0.895665	0.815427	0.891232	0.817290
	(0.026435)	(0.021743)	(0.031523)	(0.030090)
-0.75	0.837587	0.762982	0.835134	0.765869
	(0.019545)	(0.011874)	(0.023143)	(0.019449)
-1	0.757367	0.730420	0.777460	0.733640
	(0.012532)	(0.008510)	(0.014534)	(0.015442)
-1.25	0.725200	0.711635	0.725544	0.714629
	(0.008304)	(0.007555)	(0.012215)	(0.014151)
-1.50	0.706248	0.701927	0.701462	0.704714
	(0.006238)	(0.007373)	(0.010045)	(0.013787)
-1.75	0.693216	0.697574	0.687372	0.700495
	(0.005442)	(0.007365)	(0.008795)	(0.013730)
-2	0.684214	0.696239	0.679906	0.699224
	(0.004875)	(0.007388)	(0.007957)	(0.013743)
-3	0.673648	0.697692	0.667586	0.700902
	(0.003668)	(0.007430)	(0.005958)	(0.013864)
-4	0.674064	0.698308	0.666659	0.701521
	(0.003185)	(0.007443)	(0.005488)	(0.013883)
-5	0.674740	0.698350	0.666532	0.701564
	(0.003091)	(0.007446)	(0.005418)	(0.013884)

Table B.6: π_0 is the parameter to be estimated. m = number of p-values per simulation. RBE = regression-based approach, HBE = histogram-based approach. Tabulated values are the mean estimates, $\hat{\pi}_0$, with their respective mean-squared errors in parentheses.

$\pi_0 = 0.8$ $\sqrt{n}\delta$	RBE $m = 500$	HBE $m = 500$	RBE $m = 250$	HBE $m = 250$
5	0.784260	0.795054	0.778794	0.791525
	(0.001901)	(0.008379)	(0.003193)	(0.015248)
4	0.784375	0.795020	0.778375	0.791492
	(0.001923)	(0.008379)	(0.003299)	(0.015248)
3	0.785346	0.794604	0.779374	0.791121
	(0.002043)	(0.008378)	(0.003527)	(0.015241)
2	0.794382	0.793685	0.790302	0.790361
	(0.002408)	(0.008363)	(0.003981)	(0.015317)
1.75	0.800628	0.794552	0.797985	0.791178
	(0.002611)	(0.008325)	(0.004272)	(0.015312)
1.50	0.810272	0.797282	0.817465	0.793799
	(0.003044)	(0.008263)	(0.005014)	(0.015231)
1.25	0.825391	0.803409	0.829144	0.799685
	(0.003883)	(0.008200)	(0.006425)	(0.015141)
1	0.845662	0.815666	0.840988	0.811298
	(0.005774)	(0.008442)	(0.007353)	(0.015232)
0.75	0.863121	0.836255	0.856436	0.830009
	(0.011545)	(0.009460)	(0.013424)	(0.015618)
0.50	0.941254	0.867600	0.947555	0.858659
	(0.020455)	(0.012167)	(0.026354)	(0.017038)
-0.50	0.942355	0.877913	0.946758	0.877684
	(0.029305)	(0.015196)	(0.030144)	(0.023590)
-0.75	0.856525	0.842928	0.867745	0.843377
	(0.016342)	(0.010550)	(0.021564)	(0.018692)
-1	0.824353	0.821349	0.831276	0.821769
	(0.010943)	(0.008887)	(0.018460)	(0.016737)
-1.25	0.804618	0.808874	0.803143	0.808959
	(0.007195)	(0.008384)	(0.014142)	(0.016097)
-1.50	0.791291	0.802606	0.796465	0.802061
	(0.006453)	(0.008261)	(0.009123)	(0.015885)
-1.75	0.782636	0.799839	0.778826	0.799146
	(0.005989)	(0.008235)	(0.008575)	(0.015866)
-2	0.776115	0.799087	0.771918	0.798196
	(0.005806)	(0.008253)	(0.008063)	(0.015874)
-3	0.771456	0.800155	0.762065	0.799251
	(0.004134)	(0.008277)	(0.007017)	(0.015891)
-4	0.771806	0.800547	0.763326	0.799689
	(0.003734)	(0.008288)	(0.006264)	(0.015896)
-5	0.772367	0.800572	0.763266	0.799719
	(0.003664)	(0.008291)	(0.006216)	(0.015898)

Appendix C

Beyond Reasonable Doubt

C.1 Introduction

Criminal prosecutions endeavour to filter out the innocent from the guilty. In most countries the burden of proof required for a successful prosecution is "beyond reasonable doubt" but there is no standardised, quantifiable measurement of this. This precludes the possibility of miscarriages of justice due to the conviction criteria not being well-defined. Mixture distributions allow us to view defendants as being drawn from two heterogeneous populations, that is defendants are either truly guilty or truly innocent with certainty. Unfortunately this is private information, and the former have an incentive to plead not guilty. Given the two types of defendant, the distributions of the likelihood of innocence (based on court evidence) differ.

This appendix sets out a theoretical model of a typical judicial system. New measures based on compound error testing are presented, namely the false conviction rate (FCR) and false acquittal rate (FAR). Identification of the actual densities of the component distributions and level of proof required for "reasonable doubt" will allow estimates of FCR and FAR to be made. A potential estimation methodology is

subsequently discussed. State legislatures can then assess the acceptability of these conviction errors in light of the costs to society — principally compensation resulting from quashed convictions and repeat offences committed by the wrongly-acquitted. If judicial reform is required, limits on FCR and FAR can be imposed, which can be solved for a well-defined, standardised value for the threshold of reasonable doubt.

C.2 The Truth, the Whole Truth and Nothing but the Truth?

Traditionally, criminal justice systems have endeavoured to administer the law for the good of society. In an ideal world, the "bad guy" ends up behind bars while the unsullied law-abiding citizen enjoys his or her liberty. Occasionally things go wrong. History is littered with cases of miscarriages of justice — although the term is often associated with wrongful convictions, it can also be applied to errors of impunity whereby guilty parties walk free. However, the social costs of false convictions are generally perceived by society to outweigh those of false acquittals.[1]

Decision-making by juries in court has many parallels with classical hypothesis testing. Just as we assume a stated null hypothesis is true unless we have sufficiently strong evidence to reject it, in the legal setting all defendants are presumed innocent

[1]Apart from the obvious loss of freedom imposed by the incarceration of the innocent, there is a financial cost to society (specifically taxpayers) due to the monetary compensation often awarded as damages for wrongful convictions. However it could be argued that those who are wrongly-acquitted pose a (possibly greater) cost in terms of future offences that they may be tempted to commit, believing that they have an "untouchable" status (i.e. an air of impunity) as they have managed to evade justice. Also in the UK, if guilty of violent, indictable offences, repeat attacks could increase publicly-funded compensation to victims through the Criminal Injuries Compensation Authority. However in countries where some offences are punishable by the death penalty, a posthumous acquittal is a remedy of little consolation to the individual concerned.

until proven guilty proxying prior belief in the "null", here of innocence.[2]

Judicial errors of false convictions and false acquittals can be viewed analogously to Type I and Type II errors respectively which occur in conventional hypothesis testing. Just as researchers dislike making errors and seek to minimise them,[3] in its pursuit of justice an ideal court system would convict all the guilty and acquit all the innocent.

A significant handicap of modern criminal justice systems around the world is the criterion for deciding between guilt and innocence. A jury returns a guilty verdict if it believes the defendant committed the offence "beyond reasonable doubt". This qualitative approach clearly lends itself to biases attributable to subjective interpretation by jurors who may have different opinions as to the level of proof required to exceed reasonable doubt. Consequently this long-established, and long-accepted, benchmark of determining guilt has the potential to create a significant number of miscarriages of justice due to a lack of a standardised and quantifiable measurement of reasonable doubt.

An unambiguous and consistent threshold is necessary to ensure the oft-cited need, and demand, for *fair* trials, hence a quantitative evaluation of the current success of trial verdicts can allow an objective assessment of whether there is a need to reform the jury system. It is somewhat surprising that given the modern world's obsession with standardisation, for example technological and political integration, English as the global *lingua franca* etc., the archaic judicial process has remained virtually unchanged and unchallenged. Given the importance of a well-functioning

[2]This presumption of innocence is typical of legal systems with their roots in the Anglo-Saxon tradition. Some authoritarian regimes apply the presumption of guilt and hence the burden of proof of innocence falls on the defence. This discussion will concentrate on the fairer presumption of innocence approach.

[3]It is well-known however that there is a trade-off between Type I and Type II errors, with reduction of one being at the expense of an increase in the other.

judiciary to uphold the law, it is important to ensure the consistent application of justice.

Previous studies into the quantification of reasonable doubt include Simon and Mahan (1971), Tribe (1971) and Kagehiro and Stanton (1985). To the author's knowledge, little research has been carried out using a mixture distribution-based model. This paper will outline a method for assessing the effectiveness of current practice by viewing defendants as being drawn from two heterogeneous populations — the truly innocent and the truly guilty. As such it is appropriate to employ mixture distributions as a modelling approach. Difficulties in estimation of the model are discussed, and a potential methodology is outlined.

C.3 Court Trials as Hypothesis Tests

In order to mitigate the possibility of false convictions, a defendant is considered to be innocent until proven guilty. The burden of proof then falls on the prosecution to present sufficient evidence to convince a jury (consisting of reasonable, rational and impartial individuals) of a defendant's guilt. The defence counsel will present counter-arguments and offer evidence in favour of innocence. Therefore when guilty verdicts are reached, it can be assumed that the likelihood of guilt exceeds the threshold of reasonable doubt, i.e. *beyond* reasonable doubt.

Parallels with classical hypothesis testing can readily be drawn. We begin by assuming the declared null hypothesis, H_0, to be true. A relevant dataset is obtained and an appropriate test statistic, T, is constructed and its distribution *under the null* noted. The test statistic is then evaluated using the dataset to yield a real number. A significance level, α, is chosen and the corresponding critical value(s), c,[4] hence

[4] A one-tail test will have a single critical value, a two-tail test will have two. Clearly if the test

critical region, obtained. Should the sampled value of T lie in the critical region then the null hypothesis is rejected in favour of the alternative, H_1. Another approach is to compute the p-value associated with the sampled test statistic to provide a measure of the likelihood of obtaining the sample data given the null hypothesis is true. Small p-values therefore suggest improbable likelihoods, indicative of the alternative.

Just as it is stressed that we *never accept* a hypothesis but merely *reject* or *fail to reject* the null hypothesis (because under the null any test statistic value within the support of its distribution is *possible*), so trial verdicts of guilty and not guilty are based solely on court evidence (assuming no juror biases) hence there is either sufficient or insufficient evidence of guilt, hence no verdict can offer a definitive proof of guilt. This concept therefore precludes the possibility of decision errors. The analogies between hypothesis testing and criminal trials are summarised in Table C.1.

An important issue raised in Table C.1 is the significance level applied to criminal trials. Whereas in hypothesis testing a nominal level is chosen, such as 5%, there is no explicit equivalent for criminal trials. Instead we are restricted to the also nominal, but non-standardised, idea of reasonable doubt. Hence the equivalent term to α would be indifference between guilt and innocence.

At this point it is worth expressing the difference between criminal and civil trials. The former is the subject of this discussion and concerns the prosecution, on behalf of the state, of individuals charged with common law offences or cases where statute law has been broken. In the UK cases are brought by the Crown Prosecution

statistic's distribution is symmetric about zero, then comparison of the observed test statistic with the critical value can be performed in absolute terms, i.e. check whether $|T| \geq |c|$. Should the distribution be skewed when a two-tailed test is performed, for example the F distribution when testing equality of variances, then such a comparison is not possible.

Table C.1: Analogies between hypothesis tests and criminal trials.

	Hypothesis Test	Criminal Trial
Investigation	H_0 v. H_1	Not Guilty v. Guilty
Prior Belief	H_0	Not Guilty
Evaluation Method	Test Statistic, X_n	Prosecution / Defence Evidence
Significance Level	α, typically 5%	No Standardised Value
Critical Region	$C = \{x_n \in \mathbb{R} : \lvert x_n\rvert \geq \lvert c\rvert\}$	Beyond Reasonable Doubt (BRD)
Decision Criterion	Reject H_0 if $x_n \in C$	Guilty if Evidence BRD

Service (CPS) following police investigation when the CPS considers that there is sufficient evidence to secure a conviction. Offences can be categorised as either summary or indictable, representing minor and serious offences respectively. The former are usually heard in magistrates courts, and the latter in Crown Courts.

Civil lawsuits are brought by claimants, also referred to as plaintiffs, who seek a legal remedy (for example, the awarding of damages or issue of an injunction) to a particular grievance.[5] As such, civil actions are less serious than the criminal variety.[6] A consequence of this is the burden of proof required to be found liable (the civil equivalent of guilty). This is given as being "on the balance of probabilities" which has an obvious probabilistic interpretation of $\alpha = 50\%$.

An attempt to estimate the significance level associated with likelihood of guilt was made in Simon (1970). The method employed consisted of two student groups.

[5]For completeness, UK civil cases are heard in county and high courts, again commensurate with the level of the claimant's grievance.

[6]Having a criminal conviction is more socially unacceptable than being found liable in a civil case.

For a given trial, members of one group each selected whether they thought the defendant was guilty, while the others each gave opinions of the likelihood of guilt expressed as a percentage. Matching up the highest percentages with the guilty verdicts, similarly lowest likelihoods with innocent verdicts, the indifference point was found to be in the interval $[0.70, 0.74]$. Although not a definitive solution, it at least provides a ballpark figure.[7]

Hung juries then would represent the indifference point. This implies that a crude estimate for the CPS's threshold opinion on a defendant's likelihood of guilt is to be slightly less than Simon's. This is because if a jury's threshold was the same as the CPS's for each case, then there would never be any acquittals. Juries know *a priori* that there is at least some compelling evidence against the defendant, given the very fact the case has been sent to trial following prior appraisal of the evidence by the CPS. This results in jurors having to weigh up the evidence (especially in light of the defence counsel) to determine whether guilt is implied beyond reasonable doubt.

C.4 Defendant Distributions

A defendant is *either* truly guilty (G) *or* truly innocent (I) with probability 1, that is,

[7] Admittedly it is unlikely that this implied probabilistic burden of proof is consistent in practice across types of offence being tried. For the most serious offences, such as murder, juries may be more willing to accept a lower level of evidence, explicitly or implicitly, to ensure a conviction. Also, despite some individual jurors doubtless possessing idiosyncratic biases, be they positive or negative, it would be expected that these cancel out on average due to the jury consisting of (usually) twelve randomly selected members.

$$\Pr(G) = \begin{cases} 1 & \text{if truly guilty,} \\ 0 & \text{if truly innocent.} \end{cases}$$

$$\Pr(I) = \begin{cases} 1 & \text{if truly innocent,} \\ 0 & \text{if truly guilty.} \end{cases}$$

However this is private information which only the defendant knows with certainty. Guilty defendants have a clear incentive to plead not guilty[8] because they will be acquitted with positive probability (due to favourable defence evidence presented at trial). So assuming a fixed proportion of truly guilty defendants, from a jury's perspective we have a binomial set-up, where defendants are assumed guilty with probability w, and not guilty with probability $1 - w$.

Verdicts are delivered by juries who assess the likelihood of guilt, given the evidence. To maintain the analogy with hypothesis testing (i.e. given the prior belief of innocence), the likelihood of *innocence* will be considered instead, denoted by the random variable X. Because juries hear competing arguments (guilt-implying from the prosecution, innocence-implying from the defence), a jury's perceived likelihood of the defendant's innocence will be strictly greater than 0% and strictly less than 100%, despite the actual $\Pr(G)$ and $\Pr(I)$ above. As a proportion, this gives $x \in (0, 1)$.

Given the population of defendants are of two types, truly guilty and truly

[8]Guilty pleas imply no trial and may be viewed as the result of a game theory strategy, i.e. if the prior likelihood of acquittal is sufficiently small, it pays to plead guilty, show remorse and consequently receive a lighter sentence. Plea-bargaining is another possibility whereby a defendant pleads guilty to a lesser offence, ensuring a conviction for the CPS, while again the defender incurs a more lenient tariff during sentencing. Although the thought of a truly innocent defendant pleading guilty seems implausible, one of two situations may be sufficient conditions, (i) protecting the true offender (ii) although innocent, the evidence may be so compelling to a jury of guilt, that a guilty plea would mean a less severe sentence. For further details on the links between game theory and statistical decision theory, see Ferguson (1967) and Berger (1985).

innocent, the overall distribution of X can be modelled as a mixture distribution,

$$f_X(x;\boldsymbol{\Theta}) = wf_G(x;\boldsymbol{\theta}_G) + (1-w)f_I(x;\boldsymbol{\theta}_I), \qquad 0 < x < 1,\ 0 \leq w \leq 1. \tag{C.1}$$

Here, w is the mixture weight and represents the proportion of defendants who are truly guilty, again assumed constant. The component density functions for the truly guilty, f_G, and truly innocent, f_I, have associated parameter vectors $\boldsymbol{\theta}_G$ and $\boldsymbol{\theta}_I$ respectively, aggregated (with w) in $\boldsymbol{\Theta}$.

Assume guilty beyond reasonable doubt (BRD) represents a sufficiently small likelihood of innocence, with its upper limit proxied by the fixed proportion α. Thus the probability of conviction, $\Pr(C)$, and acquittal, $\Pr(A)$, can be stated respectively as

$$\Pr(C) = \int_0^{\alpha} wf_G(x;\boldsymbol{\theta}_G) + (1-w)f_I(x;\boldsymbol{\theta}_I)dx, \tag{C.2}$$

$$\Pr(A) = \int_{\alpha}^{1} wf_G(x;\boldsymbol{\theta}_G) + (1-w)f_I(x;\boldsymbol{\theta}_I)dx. \tag{C.3}$$

Our interest lies in the accuracy of this system, namely we seek to maximise $\Pr(C|G)$ and $\Pr(A|I)$ while minimising $\Pr(C|I)$ and $\Pr(A|G)$. These are subsequently given by,

$$\Pr(C|G) = \frac{\Pr(C\cap G)}{\Pr(G)} = \frac{\int_0^{\alpha} wf_G(x;\boldsymbol{\theta}_G)dx}{w} = \int_0^{\alpha} f_G(x;\boldsymbol{\theta}_G)dx, \tag{C.4}$$

$$\Pr(A|I) = \frac{\Pr(A\cap I)}{\Pr(I)} = \frac{\int_{\alpha}^{1}(1-w)f_I(x;\boldsymbol{\theta}_I)dx}{1-w} = \int_{\alpha}^{1} f_I(x;\boldsymbol{\theta}_I)dx, \tag{C.5}$$

$$\Pr(C|I) = \frac{\Pr(C\cap I)}{\Pr(I)} = \frac{\int_0^{\alpha}(1-w)f_I(x;\boldsymbol{\theta}_I)dx}{1-w} = \int_0^{\alpha} f_I(x;\boldsymbol{\theta}_I)dx, \tag{C.6}$$

$$\Pr(A|G) = \frac{\Pr(A\cap G)}{\Pr(G)} = \frac{\int_{\alpha}^{1} wf_G(x;\boldsymbol{\theta}_G)dx}{w} = \int_{\alpha}^{1} f_G(x;\boldsymbol{\theta}_G)dx. \tag{C.7}$$

However from a potential miscarriage of justice perspective, alternative complementary quantities can be introduced, namely the false conviction rate (FCR) and false acquittal rate (FAR). These error rates are a natural extension of the false discovery rate (FDR) developed in Benjamini and Hochberg (1995) and the false nondiscovery rate (FNR). In this context, FCR would provide a measure of the proportion of false convictions and the FAR the proportion of false acquittals. Should estimates of these measures prove politically unacceptable, then there would be a case for reform of the current judicial process.

Returning briefly to the analogy of classical hypothesis testing, previous studies have concentrated on the p-values of test statistic values. In this context, the p-values take the place of the likelihood of innocence, whose support is also constrained to the unit interval. Under the usual null hypothesis of a factor being insignificant, i.e. equal to zero, it is known that the distribution of these p-values is the uniform distribution, which corresponds to a component density of 1. Given n (independent) p-values in a sample of size N (where $n \leq N$) to be true under the null, and a significance level α, the FDR takes the form,

$$\text{FDR}(\alpha) = \frac{n \cdot \alpha}{\#\{p\text{-values} \leq \alpha\}}. \tag{C.8}$$

It can be seen from (C.8) that for a single significant (but truly null) p-value, FDR $= \alpha$, which is equivalent to the probability of a Type I error. The key point of these false rates is their application in *multiple* testing.

The FCR and FAR are functions of α,[9] and are defined below. For N

[9] A lower likelihood of innocence required to convict would, *ceteris paribus* lead to fewer convictions. Assuming false convictions and acquittals do not change proportionately, then this will affect the FCR and FAR.

independent[10] defendants, let N_I be the number of truly innocent and N_G be the number of truly guilty, such that $N_I + N_G = N$. Given the mixing weight, w, it follows that $N_I = (1-w)N$ and $N_G = wN$.

$$\text{FCR}(\alpha) = \frac{N_I \cdot \int_0^\alpha f_I(x;\boldsymbol{\theta}_I)dx}{N_G \cdot \int_0^\alpha f_G(x;\boldsymbol{\theta}_G)dx + N_I \cdot \int_0^\alpha f_I(x;\boldsymbol{\theta}_I)dx}, \tag{C.9}$$

$$\text{FAR}(\alpha) = \frac{N_G \cdot \int_\alpha^1 f_G(x;\boldsymbol{\theta}_G)dx}{N_G \cdot \int_\alpha^1 f_G(x;\boldsymbol{\theta}_G)dx + N_I \cdot \int_\alpha^1 f_I(x;\boldsymbol{\theta}_I)dx}. \tag{C.10}$$

The greater the values of FCR and FAR, the greater the lack of public confidence in the criminal justice system. When estimates of $f_G(x;\boldsymbol{\theta}_G)$, $f_I(x;\boldsymbol{\theta}_I)$ and w are known, then (C.9) and (C.10) can be solved for α when a politically-acceptable cap is placed on both FCR and FAR.[11] The value of this α solution would then serve as a well-defined benchmark to overcome the vagueness of BRD. If an independent assessment of the likelihood of innocence given the trial evidence was presented to a jury, then this could result in more accurate verdicts.

C.5 Estimation Issues

The theoretical framework outlined above provides a logical model to quantify miscarriages of justice. Of course in practice an empirical evaluation will be required to assess the merits of the status quo. As discussed, in order to solve for the optimal α, estimates of $f_G(x;\boldsymbol{\theta}_G)$, $f_I(x;\boldsymbol{\theta}_I)$ and w are necessary.

[10] The assumption of independence is reasonable as criminal trials cover unrelated offences.

[11] In order for law-makers to establish an appropriate cap on FCR, consideration of the compensation awarded to the wrongly-convicted needs to be taken into account. As for an FAR limit, consideration of the costs of repeat offences perpetrated by the wrongly-acquitted is relevant. In both cases, a measure of the public's lack of confidence in the judicial system to deliver correct verdicts should be included.

In order to estimate FCR and FAR we need data. Therefore a large random sample of trials[12] is required with observations of the defendant's true guilt status and the likelihood of innocence displayed by the court evidence. Evans, Osthus, and Spurrier (2006) discuss potential methods for acquiring data regarding a defendant's actual guilt and likelihood of guilt. They note the problem of sampling the actual defendants to ascertain their true guilt or innocence due to their incentive to always state innocence regardless. Therefore they suggest sampling defence counsels who, due to their close proximity to the defendant, are likely to know with high probability the actual guilt or innocence of their clients.[13] However, despite anonymous testing there is still a possibility that when asked defence lawyers will tend to state an opinion of innocence for fear of prejudicing their client if the survey responses were somehow traced back to identifiable defendants which might hinder subsequent appeals; also they are acting *for* the defendants and would want to be seen as doing everything possible to secure an acquittal.

Consequently, due to their neutral impartiality and extensive experience of hearing cases, judges are perhaps best-placed to offer unbiased (and hopefully broadly accurate) assessments of (i) true guilt or innocence of a defendant and (ii) the likelihood of innocence, x, given trial evidence. Therefore an informative survey would involve sampling several judges and for each case they hear, to give their opinion on the values of these random variables for each defendant. To allow for the possibility of different conviction rates for different offences, results should be categorised by type of offence. The following methodology should therefore be performed separately for each offence category.

[12] The court, not statistical, variety.

[13] As noted in Evans, Osthus, and Spurrier (2006), defence lawyers do not wish to know explicitly from the defendants whether they are truly guilty, as this can lead to the defence knowingly allowing their clients to perjure themselves.

Having obtained a sufficiently large data sample, estimation of $f_G(x;\boldsymbol{\theta}_G)$, $f_I(x;\boldsymbol{\theta}_I)$ and w would then be possible. Denoting estimators using the $\hat{}$ notation, a suitable estimator for w, i.e. $\hat{w}$, can be obtained from the sample proportion of perceived guilty defendants relative to all defendants.

The sample can then be partitioned by perceived guilt and innocence into two parts. A frequency density plot of the likelihood of innocence for each subset can be constructed, the shape of which would then suggest the appropriate probability distribution to model, i.e. representing the stylized facts of the empirical distributions. Given cases are only sent to trial if the prosecution authorities feel a conviction can be secured and that there will be mitigating and/or contradictory evidence from the defence, it is expected that such distributions will be positively-skewed, i.e. an elongated right tail as most of the mass will be concentrated around lower values of X. A truncated distribution,[14] would then be modelled.

Once the parametric form of the component densities $f_G(x;\boldsymbol{\theta}_G)$ and $f_I(x;\boldsymbol{\theta}_I)$ is known, maximum likelihood estimation should be performed to obtain the parameter vectors $\hat{\boldsymbol{\theta}}_G$ and $\hat{\boldsymbol{\theta}}_I$ to yield the estimated densities $\hat{f}_G$ and $\hat{f}_I$ respectively.

The final stage is then to compute the FCR and FAR and for the legislature to decide on the acceptability of these current levels. Should the FCR and FAR prove unacceptable, limits can be imposed, allowing (C.9) and (C.10) to be solved for a standardised conviction threshold (in terms of likelihood of innocence), α.

[14]The likelihood of innocence is restricted to the unit interval.

Table C.2: Summary trial data for England and Wales, 2001-2005. Source: Home Office Statistical Bulletin, Criminal Statistics 2005, England and Wales.

Description	2001	2002	2003	2004	2005
Magistrates trials (thousands)	1838	1925	2001	2023	1895
- conviction rate (%)	70.3	70.8	71.6	73.6	75.3
- of which indictable offences (thousands)	501	517	509	453	423
- conviction rate (%)	53.9	54.4	54.6	57.4	59.8
Total convictions quashed	3000	2977	2835	3044	3676
- proportion of appeals quashed (%)	23.7	24.9	24.1	24.2	28.7
Crown Court trials (thousands)	77	76	80	80	76
- conviction rate (%)	72.7	78.9	75.0	75.0	76.3
Total convictions quashed	135	166	178	261	233
- proportion of appeals quashed (%)	30.1	34.2	32.8	38.4	37.7

C.6 Feasibility of Assumptions

This discussion has made various *a priori* assumptions. For any model to be an accurate reflection of the real world, the assumptions need to be feasible and consistent with actual data. The Home Office Statistical Bulletin, Criminal Statistics 2005, England and Wales is a useful data source consisting of count data on convictions over recent years. Table C.2 provides some useful observations.

It can be seen that conviction rates (and, by default, acquittal rates) have remained fairly consistent over the period 2001-2005 across the different categories of trial. This suggests that there is a consistent application of BRD.[15]

An interesting observation concerns the level of appeals quashed between magistrates and Crown Court convictions. A higher proportion of Crown Court convictions are overturned on appeal. Assuming quashed convictions are a suitable proxy for false convictions,[16] this suggests greater accuracy of magistrates

[15]This though does not preclude absence of *systematic* biases (i.e. non-stochastic errors) in verdict decisions.

[16]In an ideal world, the appeal courts would ensure all wrong convictions are overturned, and

convictions over the Crown Courts.[17] Given that juries only give verdicts in Crown Courts, this implies magistrates convictions (given their experience) are of a higher quality, reinforcing the argument for neutral judges to be best-placed to offer opinions on a defendant's guilt and likelihood of innocence in the data collection exercise outlined previously.

Since there are several quashed convictions each year, with associated damages awarded, then there is a case for further research to obtain the necessary data to allow an empirical investigation. This is justified since the potential reduction in compensation could significantly outweigh the required research finance.

C.7 Conclusions

This discussion has sought to offer a model to examine the vagueness of the widely-applied criterion of "beyond reasonable doubt" used to assess a defendant's guilt. Considering the importance of judicial systems around the world to administer accurate verdicts, it is perhaps surprising that no formal, standardised benchmark exists.[18] Given the similarities with classical hypothesis testing, compound error testing analogous to false discovery and indifference rates is possible.

Viewing defendants as being drawn from two heterogeneous populations, truly guilty and truly innocent, false conviction and acquittal rates can be derived. A problem occurs in obtaining appropriate data to estimate the unknown component

only wrong convictions overturned.

[17]However, the seriousness of most Crown Court trials exceeds that of magistrates, hence prompting a greater proportion of appeals (whether ultimately successful or not). Also, it has already been mentioned that juries may use a more flexible threshold to ensure convictions for more serious offences, therefore increasing the probability of false convictions.

[18]This is in stark contrast to the concept of precedent in terms of sentencing guidelines which is well-documented.

densities and weights in the resulting mixture distribution. A potential solution is offered by sampling judges to offer their impartial opinions as to a defendant's true type (regardless of the verdict reached in court), and the likelihood of innocence based on the evidence presented at trial. The likelihood of innocence is considered, as opposed to the likelihood of guilt, to maintain the analogy with hypothesis testing by assuming prior belief in the null hypothesis. Here the maxim of "innocent until proven guilty" is the corresponding null applied.

Given miscarriages of justice (false convictions and false acquittals) occur, these should be minimised to ensure public confidence in the judiciary to deliver accurate verdicts. Government-, therefore tax-, funded compensation in the wake of quashed convictions, as well as the hidden costs of repeat offences by the wrongly-acquitted, are just cause to optimise the trial process. By establishing politically acceptable limits on false conviction and/or acquittal rates, the theoretical model can be solved for the optimal threshold of evidence required to secure a conviction.

Bibliography

Alberoni, F. (1962). Contribution to the study of subjective probability. Part i. *Journal of General Psychology 66*, 241–264.

Anderson, D. R., K. P. Burnham, and W. L. Thompson (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management 6*, 912–923.

Bahadur, R. R. (1960). Simultaneous comparison of the optimum and sign tests of a normal mean. In I. Olkin (Ed.), *Contributions to Probability and Statistics*, pp. 77–88. Stanford, CA: Stanford University Press.

Bayarri, M. J. and J. O. Berger (2004). The interplay of Bayesian and Frequentist analysis. *Statistical Science 19*, 58–80.

Becker, B. J. (1991). Small-sample accuracy of approximate distributions of functions of observed probabilities from t tests. *Journal of Educational Statistics 16*, 345–369.

Benjamini, Y. and Y. Hochberg (1995). Controlling the False Discovery Rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society 57*, 289–300.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics 29*, 1165–1188.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed.). New York: Springer Verlag.

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

Berger, J. O. and D. Berry (1988). Statistical analysis and the illusion of objectivity. *American Scientist 76*, 159–165.

Berger, J. O., L. D. Brown, and R. L. Wolpert (1994). A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing. *The Annals of Statistics 22*, 1787–1807.

Berger, J. O. and M. Delampady (1987). Testing precise hypotheses. *Statistical Science 2*, 317–335.

Berger, J. O. and J. Montera (1999). Default Bayes factors for non-nested hypothesis testing. *Journal of the American Statistical Association 94*, 542–554.

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

Berger, J. O. and R. L. Wolpert (1988). *The Likelihood Principle*. Hayward, California: Institute of Mathematical Statistics.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association 33*, 526–542.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association 37*, 325–335.

Birnbaum, A. (1961). On the foundations of statistical inference: binary experiments. *The Annals of Mathematical Statistics 32*, 414–435.

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association 57*, 269–306.

Bjørnstad, J. (1996). On the generalization of the likelihood function and the likelihood principle. *Journal of the American Statistical Association 91*, 791–806.

Bornstein, R. F. (1989). Exposure and affect: Overview and meta-analysis of research. *Psychological Bulletin 106*, 265–289.

Brownie, C. and J. Kiefer (1977). The ideas of conditional confidence in the simplest setting. *Communications in Statistics - Theory and Methods 6*, 691–751.

Carlson, R. (1976). The logic of tests of significance (with discussion). *Philosophy of Science 43*, 116–128.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review 48*, 378–399.

Casella, G. and R. L. Berger (1987). Reconciling Bayesian and Frequentist evidence in the one-sided testing problem (with discussion). *Journal of the American Statistical Association 82*, 106–111.

Cohen, J. (1994). The earth is round ($p<.05$). *American Psychologist 49*, 997–1003.

Cornfield, J. (1966). Sequential trials, sequential analysis and the likelihood principle. *The American Statistician 20*, 18–23.

Cox, D. R. (1977). The role of significance tests. *Scandinavian Journal of Statis-*

tics 4, 49–70.

David, H. A. (1998). First (?) occurrence of common terms in probability and statistics - A second list, with corrections. *The American Statistician 52*, 36–40.

DeGroot, M. H. (1973). Doing what comes naturally: Interpreting a tail area as a posterior probability or as a likelihood ratio. *Journal of the American Statistical Association 68*, 966–969.

Deming, W. E. (1943). *Statistical adjustment of data*. New York: Wiley.

Diamond, G. A. and J. S. Forrester (1983). Clinical trials and statistical verdicts: probable grounds for appeal. *Annals of Internal Medicine 98*, 385–394.

Dickey, J. M. (1973). Scientific reporting. *Journal of the Royal Statistical Society 35*, 285–305.

Donahue, R. M. J. (1999). A note on information seldom reported via the P value. *The American Statistician 53*, 303–306.

Edwards, W., H. Lindman, and L. J. Savage (1963). Bayesian statistical inference for psychological research. *Psychological Review 70*, 193–242.

Efron, B. and A. Gous (2001). Scales of evidence for model selection: Fisher versus Jeffreys (with discussion). In P. Lahiri (Ed.), *Model Selection*. Beachwood.

Evans, K., D. Osthus, and R. Spurrier (2006). Distributions of interest for quantifying reasonable doubt and their applications. *VERUM Working Paper, Valparaiso University*.

Ferguson, T. S. (1967). *Mathematical statistics: A decision-theoretic approach*. New York: Academic Press.

Gabriel, K. R. (1969). Simultaneous test procedures - some theory of multiple comparisons. *The Annals of Mathematical Statistics 40*, 224–250.

Gardner, M. J. and D. G. Altman (1986). Confidence intervals rather than P values; estimating rather than hypothesis testing. *British Medical Journal 292*, 746–750.

Genovese, C. and L. Wasserman (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society Series B (Statistical Methodology) 64*, 499–517.

Gibbons, J. D. and J. W. Pratt (1975). P-values: Interpretation and methodology. *The American Statistician 29*, 20–25.

Goodman, S. (1999a). Toward evidence-based medical statistics 1: the p-value fallacy. *Annals of Internal Medicine 130*, 995–1004.

Goodman, S. (1999b). Toward evidence-based medical statistics 2: the Bayes factor. *Annals of Internal Medicine 130*, 1005–1013.

Hall, P. and B. Sellinger (1986). Statistical significance: balancing evidence against doubt. *Australian Journal of Statistics 28*, 354–370.

Hedges, L. V. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin 93*, 563–573.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika 75*, 800–812.

Hodges, J. L. J. and E. L. Lehmann (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society 16*, 261–268.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandi-*

navian Journal of Statistics 6, 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika 75*, 383–386.

Huberty, C. J. (1987). On statistical testing. *Educational Researcher 16*, 4–9.

Hung, H. M. J., R. T. O'Neill, P. Bauer, and K. Köhne (1997). The behavior of the p-value when the alternative hypothesis is true. *Biometrics 53*, 11–22.

Hwang, J. T., G. Casella, C. Robert, M. T. Wells, and R. Farrell (1992). Estimation of accuracy in testing. *The Annals of Statistics 20*, 490–509.

Jahn, R. G., B. J. Dunne, and R. D. Nelson (1987). Engineering anomalies research. *Journal of Scientific Exploration 1*, 21–50.

Jeffreys, H. (1961). *Theory of Probability*. London: Oxford University Press.

Jeffreys, H. (1980). Some general points in probability theory. In A. Zellner (Ed.), *Bayesian analysis in econometrics and statistics*, pp. 451–454. Amsterdam: North-Holland.

Kagehiro, D. K. and W. C. Stanton (1985). Legal vs. quantified definitions of standards of proof. *Law and Human Behaviour 9*, 159–178.

Kaiser, H. F. (1960). Directional statistical decisions. *Psychological Review 67*, 160–167.

Kempthorne, O. (1976). Of what use are tests of significance and tests of hypotheses. *Communications in Statistics A5*, 763–777.

Kiefer, J. (1977). Conditional confidence statements and confidence estimators (with discussion). *Journal of the American Statistical Association 72*, 789–827.

Kimball, A. W. (1957). Errors of the third kind in statistical consulting. *Journal of the American Statistical Association 52*, 133–142.

Labovitz, S. (1968). Criteria for selecting a significance level: A note on the sacredness of .05. *American Sociologist 3*, 220–222.

Lambert, D. and W. J. Hall (1982). Asymptotic lognormality of p-values. *The Annals of Statistics 10*, 44–64.

Leahey, E. (2005). Alphas and asterisks: The development of statistical significance testing standards in sociology. *Social Forces 84*, 1–24.

LeCam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. New York: Springer Verlag.

Lehmann, E. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? *Journal of the American Statistical Association 88*, 1242–1249.

Lehmann, E. L. (1986). *Testing statistical hypotheses* (2nd ed.). New York: Wiley.

Lehmann, E. L. and J. P. Romano (2005). *Testing Statistical Hypotheses*. New York: Springer.

Leventhal, L. and C. Huynh (1996). Directional decisions for two-tailed tests: Power, error rates, and sample size. *Psychological Methods 1*, 278–292.

Lindley, D. V. and W. F. Scott (1984). *New Cambridge Statistical Tables*. Cambridge, UK: Cambridge University Press.

Marascuilo, L. A. and J. R. Levin (1970). Appropriate post hoc comparisons for interaction and nested hypotheses in analysis of variance designs: The elimination of Type-IV errors. *American Educational Research Journal 7*, 397–

421.

Marcus, R., E. Peritz, and K. R. Gabriel (1976). On closed testing procedures with special references to ordered analysis of variance. *Biometrika 63*, 655–660.

Mood, A. M., F. A. Graybill, and D. C. Boes (1974). *Introduction to the Theory of Statistics* (3rd ed.). New York: McGraw-Hill.

Mosteller, F. (1948). A k-sample slippage test for an extreme population. *The Annals of Mathematical Statistics 19*, 58–65.

Nagel, S. S. and M. Neef (1977). Determining an optimal level of statistical significance. In M. Guttentag (Ed.), *Evaluation studies review annual*, Volume 2, pp. 146–158. Beverly Hills: Sage.

Pearson, E. S. (1938). The probability integral transformation for testing goodness of fit and combining independent tests of significance. *Biometrika 30*, 134–148.

Peto, R., M. C. Pike, P. Armitage, N. E. Breslow, D. Cox, R. Howard, S. V. Mantel, K. McPherson, J. Peto, and P. G. Smith (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient, I: Introduction and design. *British Journal of Cancer 34*, 585–612.

Raiffa, H. (1968). *Decision Analysis: Introductory Lectures on Choices Under Uncertainty*. Reading: Addison-Wesley.

Reid, N. (1995). The roles of conditioning in inference. *Statistical Science 10*, 138–157.

Robinson, D. H. and H. Wainer (2001). On the past and future of null hypothesis significance testing. Technical report, Research Report RR-01-24, Educational Testing Service, Princeton.

Rosenthal, R. (1984). *Meta-analytic procedures for social research*. Beverly Hills: Sage.

Royall, R. M. (1986). The effect of sample size in the meaning of significance tests. *The American Statistician 40*, 313–315.

Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *The Annals of Statistics 30*, 239–257.

Savage, L. J. (1976). On rereading R. A. Fisher. *The Annals of Statistics 4*, 441–500.

Sawyer, A. G. and J. P. Peter (1983). The significance of statistical significance tests in marketing research. *Journal of Marketing Research 20*, 122–133.

Schervish, M. J. (1995). *Theory of Statistics*. New York: Springer.

Schervish, M. J. (1996). P values. What they are and what they are not. *The American Statistician 50*, 203–206.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods 1*, 115–129.

Schweder, T. and E. Spjøtvoll (1982). Plots of P-values to evaluate many tests simultaneously. *Biometrika 69*, 493–502.

Sellke, T., M. J. Bayarri, and J. O. Berger (2001). Calibration of p-values for testing precise null hypotheses. *The American Statistician 55*, 62–71.

Shaffer, J. (1995). Multiple hypothesis testing. *Annual Review of Psychology 46*, 561–584.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika 73*, 751–754.

Simon, R. J. (1970). Beyond a reasonable doubt: An experimental attempt at quantification. *Journal of Applied Behavioral Science 6*, 203–209.

Simon, R. J. and L. Mahan (1971). Quantifying burdens of proof: A view from the bench, the jury, and the classroom. *Law and Society Review 5*, 319–330.

Smith, A. F. M. and D. J. Spiegelhalter (1980). Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society 42*, 213–220.

Spielman, S. (1978). Statistical dogma and the logic of statistical testing. *Philosophy of Science 45*, 120–135.

Stallings, W. M. (1985). Mind your p's and alphas. *Educational Researcher 14*, 19–20.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society 64*, 479–498.

Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *The Annals of Statistics 31*, 2013–2035.

Storey, J. D., J. E. Taylor, and D. Siegmund (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society Series B 66*, 187–205.

Storey, J. D. and R. Tibshirani (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences 100*, 9440–9445.

Strube, M. J. (1985). Combining and comparing significance levels from noninde-

pendent hypothesis tests. *Psychological Bulletin 97*, 334–341.

Tribe, L. H. (1971). Trial by mathematics: Precision and ritual in the legal process. *Harvard Law Review 84*, 1329–1393.

Weller, J. I., J. Z. Song, D. W. Heyen, H. A. Lewin, and M. Ron (1998). A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. *Genetics 150*, 1699–1706.

Westfall, P. H. and S. S. Young (1960). *Resampling-based Multiple Testing: Examples and Methods for P-value Adjustment*. New York: Wiley.

Wolpert, R. L. (1995). Testing simple hypotheses. In H. H. Bock and W. Polasek (Eds.), *Studies in Classification, Data Analysis, and Knowledge Organization*, Volume 7. Heidelberg: Springer Verlag.

Zabell, S. (1992). R. A. Fisher and the fiducial argument. *Statistical Science 7*, 369–387.

Zaykin, D. V., S. S. Young, and P. H. Westfall (2000). Using the false discovery rate approach in the genetic dissection of complex traits: A response to Weller et al. *Genetics 154*, 1917–1918.

Zellner, A. (1984). Posterior odds ratios for regression hypotheses. In *Basic issues in econometrics*, pp. 275–305. Chicago: University of Chicago Press.