Asymmetry and other distributional properties in medical research data

by

Christopher Partlett

A thesis submitted to
The University of Birmingham
for the degree of
Doctor of Philosophy (PhD Statistics)

School of Mathematics
The University of Birmingham
June 2015

Abstract

The central theme of this thesis is to investigate the use of non-parametric methods for making inferences about a random sample with an unknown distribution function. The overarching aim is the development of new methods to make inferences regarding the nature of the unknown distribution to enhance medical research. Initially, the focus is exclusively on the asymmetry of a random variable. In particular, a recently proposed measure of asymmetry provides the foundation for the proposal and development of a new test for symmetry. The potential applications of the test and measure are applied to a number of medical research settings including randomised trials. Moreover, guidance is provided on its implementation, with particular emphasis on the problem of small sample estimation.

This investigation is then generalised to examine asymmetry across multiple studies. In particular, meta-analysis methods are used to synthesise information about the amount of asymmetry in several studies. Further, a detailed simulation study is carried out to investigate the impact of asymmetry on linear models and meta-analyses of randomised trials, in terms of the accuracy of the treatment effect estimate and the coverage of confidence and prediction intervals.

Finally, the scope of the investigation is widened to encompass the problem of comparing and synthesising information about the probability density function and cumulative distribution function, based on samples from multiple studies. The meta-analysis of the smooth distribution function estimate is then applied to propose new methods for conducting meta-analyses of diagnostic test accuracy, which have a number of merits compared to the existing methodology.

Acknowledgements

Firstly, I owe sincerest thanks to Professor Richard Riley for his invaluable help and guidance and for opening the door to the potential real life applications of the theory discussed here. I am also indebted to Dr. Prakash Patil for his help in developing and expanding the statistical theory that appears in this thesis.

I also wish to thank the staff and postgraduate students at the School of Mathematics as well as the School of Health and Population Sciences for making these past years some of the most enjoyable of my life. I would like to give a special mention to Dr. Yemisi Takwoingi for her support in the last few months and valuable insight into some of the problems encountered in the penultimate chapter.

I am also immensely grateful for the support of my friends and family, but in particular to Mum, Dad, Steve, and Rebecca.

Finally, I gratefully acknowledge the financial support of the School of Mathematics and EPSRC.

Contents

Chapter 1

Introduction

1.1 Overview of thesis

The central theme of this thesis is to investigate the use of non-parametric methods for making inferences about a random sample with an unknown distribution function, with particular focus on application to medical research settings. The overarching aim is the development of new methods to make inferences regarding the nature of the unknown distribution, with particular emphasis on effectively measuring the asymmetry of the random variable and to test for symmetry in these cases.

Distributional assumptions play a fundamental role in almost all statistical analyses. Indeed, the primary aim of many statistical analyses is to draw conclusions (and quantify the uncertainty of these conclusions) regarding the distribution of a random sample. This amounts to, at the most fundamental level, determining the distribution function from which the random sample is independently drawn. The classical approach is to narrow the investigation by assuming a specific parametric form for this distribution function. Usually, this reduces the problem to determining the mean and the variance of a specific family of distributions.

Likewise, symmetric random variables are important for the development and application of statistical theory. In particular, symmetry is an important assumption for many statistical models. For example, symmetry assumptions are essential in deriving many point or interval estimations of location parameters. Furthermore, a wide range of statistical techniques, many

of which have applications in medical statistics, rely on assumptions of normality and hence symmetry. For example, linear regression models assume that residuals are normally distributed, and assessing the symmetry of the residual distribution is an important precursor in assessing the normality of the residuals. The problem of assumption validation is particularly pertinent in medical statistics where model assumptions are often not validated in practice, or at the very least not referred to in the literature when the results are presented.

As we have already mentioned, our principle focus is to investigate measuring the asymmetry of the unknown distribution, and to test for symmetry. Recent research by Patil et al. [88] has produced new measures of asymmetry, which have been shown to effectively quantify the amount of asymmetry. We propose a new test based upon one such measure $\widehat{\eta}$. We derive the asymptotic distribution of the test statistic and we analyse the performance of the proposed test through the use of a simulation study.

We also show that the measure $\widehat{\eta}$ has a number of viable applications in medical statistics. For example, we show that it is an effective tool for validation of model assumptions. Furthermore, we show that $\widehat{\eta}$ can play a crucial role in identifying when to transform the data, what type of transformation is likely to be effective, and evaluating whether the transformation has been a success. $\widehat{\eta}$ can even be used as a summary statistic in randomised trials, alongside the mean and variance to assess the baseline similarities of treatment and control groups.

The problem of small sample estimation of η is also identified and addressed, by discussing a new procedure for making inferences about η, which has a number of advantages when there are only a small number of observations. Both methods are applied to the more general problem of assessing the asymmetry in multiple trials. This methodology has a number of genuine applications in medical research, principally, allowing the comparison and synthesis of information about the distribution of multiple (possibly heterogeneous) samples from a single population.

We conclude our investigation of asymmetry with an extensive analysis of how ignoring the presence of asymmetry impacts the accuracy of the inferences drawn from a number of statistical models, and to what extent they are robust to departures from symmetry.

Finally, we move away from the investigation of the asymmetry of an unknown distribution

and consider, more generally, a method for synthesising information about the distribution as a whole. That is, we widen the scope of our investigation to encompass the problem of comparing and synthesising information about the probability density function and cumulative distribution function, based on several samples from heterogeneous (but similar) populations.

We proceed in section 1.2 with a thorough introduction of the concepts of symmetry, asymmetry and skewness in the context of random variables. In particular, we clearly define symmetric random variables and discuss some of their elementary properties. In section 1.3 we introduce a few examples of statistical tests and models, which rely on the assumption of symmetry or normality to draw effective inferences. In section 1.4 we introduce a variety of non-parametric methods, which are used to analyse the distribution of a random sample in a more general sense. In section 1.5 we summarise the aims of the thesis and in section 1.6 we provide an outline of the remaining chapters.

1.2 Symmetry, asymmetry and skewness

1.2.1 Symmetry

The concept of symmetry has existed in a wide variety of contexts in both science and art for thousands of years. Indeed, it is often considered the pinnacle of aesthetics in art and architecture. Furthermore, it is fundamental in understanding the structure of macroscopic and microscopic bodies in biology and chemistry. Serfling [110] gives a wonderful description of symmetry, drawing on its revered properties as a 'principle of order' and 'an abstraction of balance, harmony and perfection.'

The idea of symmetry in a single univariate random variable is very well understood. First, we provide the definition of a symmetric random variable given by Hershkorn and Chapman [52] and then discuss some of the elementary properties of symmetric random variables. Unless otherwise stated, all random variables X are assumed to be continuous.

Definition 1.1 *A random variable X with cumulative distribution function F is said to be*

symmetric, if there exists a real number θ such that

$$F(\theta - x) = 1 - F(\theta + x) \quad \forall x \in \mathbb{R}.$$

If the probability density function f exists then Lemma 1.2 establishes an equivalent definition of symmetry.

Lemma 1.2 *If X is a symmetric, absolutely continuous random variable X with continuous density f and centre of symmetry θ, then*

$$f(\theta - x) = f(\theta + x) \quad \forall x \in \mathbb{R}.$$

Proof. By Definition 1.1,

$$F(\theta - x) = 1 - F(\theta + x) \quad \forall x \in \mathbb{R}.$$

Then, via differentiation with respect to x,

$$f(\theta - x) = f(\theta + x) \quad \forall x \in \mathbb{R}.$$

□

Examples of symmetric random variables include the Normal and Cauchy distribution, shown in Figure 1.1. We shall now state and prove some of the elementary properties of symmetric random variables, beginning with the following theorem.

Theorem 1.3 *Let X be a symmetric random variable with distribution function F. Then the centre of symmetry θ is the median of the random variable X.*

Proof. Because the random variable X is symmetric we have $F(\theta - x) = 1 - F(\theta + x) \quad \forall x \in \mathbb{R}$. Then, by setting $x = 0$ we obtain

$$F(\theta) = 1 - F(\theta) \Rightarrow F(\theta) = \frac{1}{2}.$$

Hence, θ is the median. □

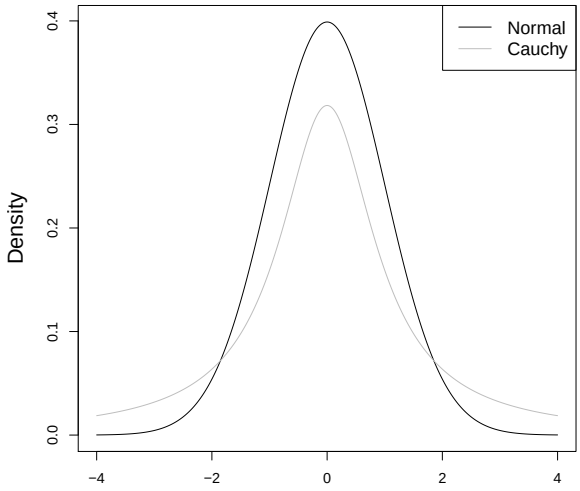

Figure 1.1: The density functions of the symmetric Normal and Cauchy random variables.

Theorem 1.4 *Let X be a symmetric random variable with density function f. Then, providing the mean $\mathrm{E}[X]$ exists, the centre of symmetry θ is also equal to the mean.*

Proof. Without loss of generality take $\theta = 0$, as we can always define a random variable $Y = X - \theta$ so that the centre of symmetry is equal to zero. First, observe that

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} xf(x)dx = \int_{-\infty}^{0} xf(x)dx + \int_{0}^{\infty} xf(x)dx$$

Now we apply the transformation $y = -x$ to the left integral to obtain

$$\mathrm{E}[X] = \int_{\infty}^{0} (-y)f(-y)(-dy) + \int_{0}^{\infty} xf(x)dx.$$

We can now apply the hypothesis of symmetry $f(-x) = f(x)$ for all x. Doing so, we obtain

$$\mathrm{E}[X] = \int_{\infty}^{0} xf(x)dx + \int_{0}^{\infty} xf(x)dx = -\int_{0}^{\infty} xf(x)dx + \int_{0}^{\infty} xf(x)dx = 0.$$

Hence, in general $\mathrm{E}[X] = \theta$. □

1.2.2 Asymmetry

Whilst the idea of symmetry is very easily understood, quantifying the absence of symmetry is a more subtle problem. It is trivial to define an asymmetric random variable as one which does not fulfil Definition 1.1. Some examples of asymmetric random variables are the Log-Normal, Folded Normal and Exponential random variables, which can be seen in Figure 1.2.

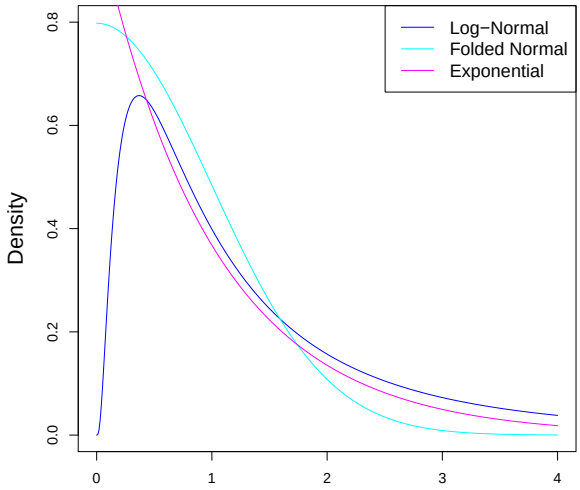

Figure 1.2: The density functions of the asymmetric Log-Normal, Folded Normal and Exponential random variables.

However, intuitively it is believed that asymmetry is something that can be measured. When presented with two similar density curves, it is usually possible to provide some rationale on why one is more or less asymmetric than the other density curve. However, it is difficult to find a mathematical expression to calibrate or quantify the amount of asymmetry.

MacGillivray [76] actually proposes an asymmetry measure by modifying an existing measure for skewness, a property which we discuss in more detail in section 1.2.3. The quantile measure of skewness for a continuous random variable, with distribution function F and median θ, is given by

$$\gamma(u; F) = \frac{F^{-1}(1-u) + F^{-1}(u) - 2\theta}{F^{-1}(1-u) - F^{-1}(u)}.$$

MacGillivray [76] proposes the following adaptation to the measure so that

$$\Gamma(F) = \sup_{\alpha \leq u \leq \frac{1}{2}} \gamma(u; F),$$

for some $\alpha > 0$ provides a suitable measure of asymmetry. MacGillivray shows that Γ maintains a reasonable ordering with respect to asymmetry, and that Γ can be interpreted as the "minimum standardised difference between F and a symmetric distribution about θ". However, a drawback to this measure is that the parameter α means that $\alpha \times 100\%$ of the probability mass at each tail is not included in the measure. Since the tails obviously play a crucial role in identifying and quantifying asymmetry this is a significant shortcoming.

More recently, Boshnakov [11] proposed an interesting alternative for measuring the asymmetry of unimodal density functions. For a given unimodal density function f, Boshnakov defines l_{incr} and l_{decr} as the length of the regions where f is increasing and deceasing respectively. Also by letting $l = l_{incr} + l_{decr}$, it is readily verified that for a symmetric density we have

$$\frac{l_{incr}}{l} = \frac{l_{decr}}{l} = \frac{1}{2}.$$

Boshnakov [11] provides several contenders for a coefficient (or index) of asymmetry. These are given by

$$r_+ = \mathrm{E}\left[\frac{l_{decr}}{l}\right],$$

$$r_- = \mathrm{E}\left[\frac{l_{incr}}{l}\right],$$

and

$$r_a = r_+ - r_-.$$

It is clear that for symmetric densities that $r_+ = r_- = \frac{1}{2}$ and $r_a = 0$. Boshnakov [11] comments that the above measures provide a useful indication of the amount of asymmetry present. However, the condition that f must be a unimodal density is a very restricting one.

Li and Morris [73] provide a simpler and more intuitive measure, provided the density func-

tion f exists. For a density function f with mean μ, Li and Morris propose

$$s_1(f) = \int_{-\infty}^{\infty} |f(\mu + x) - f(\mu - x)| dx,$$

as a measure for asymmetry. Clearly $0 \leq s_1 \leq 2$ for all f and $s_1 = 0$ if and only if f is symmetric.

There are other measures that one can propose using the current tests of symmetry. Whilst these have not been formally proposed as measures of asymmetry, due to their use in tests of symmetry, one expects that they should at least capture departure from symmetry. For example for a distribution function F, Boos [10] uses the sample version of

$$s_2(F) = \int_{-\infty}^{\infty} (F(\theta + x) + F(\theta - x) - 1)^2 dx,$$

where θ is the median, to define a test for symmetry. Alternatively Rothman and Woodroofe [105] use the sample version of

$$s_3(F) = \int_{-\infty}^{\infty} (F(\theta + x) + F(\theta - x) - 1)^2 dF(x),$$

to test for symmetry. Note that these measures are zero for symmetric densities and are positive for asymmetric density functions.

1.2.3 Skewness

Due to the difficultly involved in quantifying asymmetry, very often skewness is used to make inferences about the symmetry of random variables. Indeed, these two concepts are very closely related, primarily because skewness is zero for a symmetric density. Furthermore, the direction of the skew provides an indication of the direction of asymmetry. Therefore, the use of skewness is valid upto a certain point. Since one of the themes of this thesis is borne out of the fact that skewness and symmetry are different concepts, it is pertinent that we define skewness and compare it with asymmetry.

Pearson [91] first proposed measuring skewness using

$$\gamma_1 = \frac{\mu - M}{\sigma},$$

for a univariate random variable with mean μ, mode M and variance σ^2. Two other measures of skewness (also attributed to Pearson) are given by

$$\gamma_2 = \frac{\mu - \theta}{\sigma},$$

and

$$\gamma_3 = \frac{\mu_3}{\sigma^3},$$

where θ is the median and μ_3 is the third central moment,

$$\mu_3 = \mathrm{E}\left[(X - \mu)^3\right].$$

Another common measure of skewness, proposed by David and Johnson [22], is the quantile measure which we introduced in the previous subsection,

$$\gamma(u; F) = \frac{F^{-1}(1-u) + F^{-1}(u) - 2\theta}{F^{-1}(1-u) - F^{-1}(u)}.$$

Alternatively, van Zwet [122] takes a different approach and formalises the concept of skewness by proposing the following ordering of random variables based on skewness.

Definition 1.5 *Let F and G be two distribution functions, and let f and g be their respective density functions. Then g is more skewed to the right than f if and only if $G^{-1}(F)$ is convex on the support of F. In van Zwet's notation this is denoted by $F \underset{c}{<} G$.*

From here, van Zwet is able to show that the operator $\underset{c}{<}$ can be used to order probability distributions based on the amount of skewness present. Our interest is to order probability distributions according to their size of asymmetry and it is reasonable to expect this to be different from the ordering based upon skewness. Indeed, there is a problem with simply substituting

asymmetry for skewness as, whilst they are similar, they are not equivalent properties. Indeed, the measures of skewness are designed to highlight the tail behaviour of density functions. This can lead to discrepancy between skewness and asymmetry. For example, consider the following density functions

$$f(x) = \begin{cases} \exp(-x) & \text{if} \quad 0 < x < \infty \\ 0 & \text{otherwise,} \end{cases}$$

and

$$g(x) = \begin{cases} (\alpha - 1)(x+1)^{-\alpha} & \text{if} \quad 0 < x < \infty \\ 0 & \text{otherwise.} \end{cases}$$

Here α is a real parameter greater than one. First, note that their respective distribution functions are given by

$$F(x) = \begin{cases} 1 - \exp(-x) & \text{if} \quad 0 < x < \infty \\ 0 & \text{otherwise,} \end{cases}$$

and

$$G(x) = \begin{cases} 1 - (x+1)^{-\alpha+1} & \text{if} \quad 0 < x < \infty \\ 0 & \text{otherwise.} \end{cases}$$

Then using van Zwet's definition of skewness (Definition 1.5) we see that

$$G^{-1}(F(x)) = \exp\left(\frac{x}{\alpha - 1}\right) - 1,$$

on the interval $0 < x < \infty$. This is clearly a convex function on the support of F and, therefore, in van Zwet's notation we have that $F \underset{c}{<} G$. That is, g is more skewed to the right than f. This is to be expected, as g has a heavier right-hand tail than f. However, from Figure 1.3 it is clear that the probability mass is more evenly spread under g, suggesting that g is less asymmetric than f. Hence, skewness cannot always be considered a reliable substitute for asymmetry.

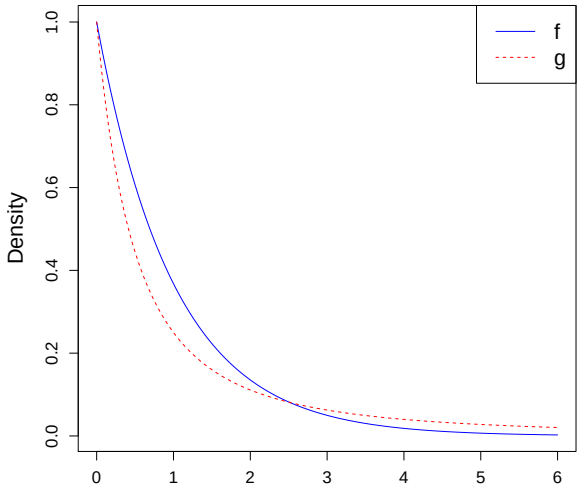

Figure 1.3: Plots of the density functions f and g with $\alpha = 2$.

1.2.4 Testing for symmetry

The null hypothesis of a test for symmetry, in the case where the probability density function f exists, is

$$H_0 : f(\theta - x) = f(\theta + x) \; \forall x \in \mathbb{R}.$$

There are several approaches to test the above hypothesis, and these tests can be separated into two approaches. They either apply direct use of the definition for symmetry (Definition 1.1 or 1.2) or involve exploiting the relationship between asymmetry and skewness. Most importantly, however, these test statistics are constructed such that they capture the departure from symmetry, and not the size of asymmetry, if the alternative is true. As a result, as the amount of asymmetry increases, these tests do not necessarily exhibit an increase in power. A key aim of the thesis is to demonstrate and address this issue, and we return to this subject in Chapter 2. In the next section we discuss a couple of commonly used statistical methods which rely, either implicitly or explicitly, on the assumption of symmetry.

1.3 Symmetry and other distributional assumptions in statistical tests and models

There are a vast array of statistical methods which rely on the assumption of asymmetry, either explicitly or implicitly. For example, the one sample Wilcoxon signed-rank test for the location of the median relies on the assumption that the data are drawn from a symmetric population. Often the assumption of symmetry is less explicit. For example, many other commonly used statistical models rely on assumptions of normality, which implicitly assumes that the data are symmetric.

1.3.1 Wilcoxon signed-rank test

The one sample Wilcoxon signed-rank test can be thought of as the non-parametric equivalent of Student's t-test [128]. For a single sample $X_1, \cdots, X_n$, it tests the null hypothesis that the median θ is equal to some known value θ_0. That is, it tests

$$H_0 : \theta = \theta_0,$$

using the test statistic,

$$W = \left| \sum_{i=1}^{n} \operatorname{sign}(X_i - \theta_0) R_i \right|, \tag{1.6}$$

where R_i is the rank of $|X_i - \theta_0|$. The test does not assume a specific probability distribution for the data, however, it does assume that,

1. The data are independently sampled from the same distribution.
2. The data are measured on an ordinal scale.
3. The distribution is symmetric about the median θ.

This test is not at all robust to departures from symmetry and will provide misleading results when asymmetry is present in the data [123]. Indeed, it is often posed as a test *for symmetry* about a median θ, something we shall elaborate on later in Chapter 2 when we discuss tests for symmetry in more detail.

1.3.2 Linear models

Now consider the following simple linear model applied to two groups of data,

$$Y_j = \beta_0 + \beta_1 X_j + e_j, \quad j = 1, \cdots, n,$$

where Y_j is some continuous response and X_j is a group identifier and can be equal to 0 or 1. For example, in a clinical trial setting $X_j = 0$ could represent a placebo, while $X_j = 1$ could represent a new treatment. It is assumed that

1. The residuals e_j are normally distributed.
2. The variances are homogeneous in both groups.
3. The errors are independent.

It is often stated that linear models are robust to small departures from normality [82]. As a result, the inferences that one makes with the model should not be too badly biased by the existence of some slight skewness. However, for more serious departures from normality (e.g. heavily skewed data) the usual inferences that one draws from the model (e.g. treatment effect estimates or confidence intervals) might be more seriously compromised.

In the next section we introduce several commonly used non-parametric methods which are utilised to analyse the distribution of a random variable from an unknown population.

1.4 Non-parametric methods for analysing the distribution of data

The primary aim of many statistical analyses is to draw conclusions (and quantify the uncertainty of these conclusions) regarding the distribution of a random sample. This amounts to, at the most fundamental level, determining the distribution function from which the random sample is independently drawn. The classical approach narrows the investigation by a considerable degree by assuming a specific parametric form for this distribution function. Usually, this reduces the problem to determining the mean and the variance of a specific family of distributions.

Typically this distribution is chosen to be the Normal distribution, which is commonly the most mathematically tractable route.

The principle weakness of the parametric approach is that the results are sensitive to the parametric assumption, and misspecification of the parametric distribution can lead to erroneous conclusions. One way to alleviate this problem is to adopt more flexible parametric models, for example, those which include additional parameters that allow for skewness or additional weight in the tails of the distribution [18].

An alternative approach is to adopt so called non-parametric approaches, which don't assume a specific functional form for the distribution or density function. These methods are particularly useful when there are no clear theoretical grounds for assuming a single parametric form. The key difference with non-parametric methods is that it allows the data to 'speak for itself' without restrictive assumptions about the shape of the underlying distribution of density function.

For example, let $X_1, \cdots, X_n$ be a random sample from a random variable with distribution function F and probability density function f. Then the empirical distribution function,

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}[X_i < x],$$

is a truly non-parametric estimate of the distribution function $F(x)$. $F_n(x)$ is an extremely efficient point estimate of $F(x)$, however, one could argue that $F_n(x)$ is not a very good estimate of the curve $F(x)$. Indeed, the resultant curve is an increasing step function with change points at the points X_i. It could be reasoned, therefore, that $F_n(x)$ is guilty of excessively pandering to the data. This may lead to over-fitting to chance data points.

One way to avoid this problem is to apply a smoothing method. A smoothing method utilises all the available information in the neighbourhood of the point x to generate an estimate of the function. This process commonly reduces the uncertainty in the estimate at x, at the cost of introducing a certain amount of bias. It also allows the generation of smooth curves over a range of values. For example, consider the kernel density estimate of f,

$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right),$$

where K is a kernel density and h is the bandwidth. The bandwidth defines the width of the bin about x, within which the data X_i can influence the estimate of $f(x)$. As a consequence the bandwidth h directly controls the degree of smoothing.

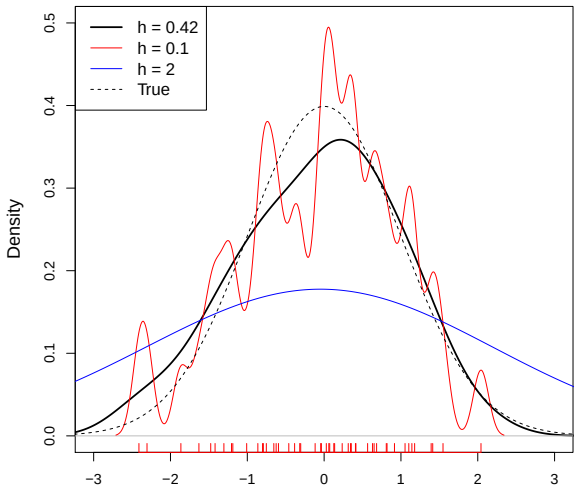

Figure 1.4: Kernel density estimates of a standard Normal sample of size $n = 50$ for a range of bandwidths h, compared with the true density curve.

Non-parametric kernel density estimates were first pioneered by Rosenblatt [104] and Parzen [87]. Under suitable smoothness conditions (including the existence of f'', the second derivative of f) it is readily calculated that the bias of $\widehat{f}(x)$ is

$$\mathrm{E}\left[\widehat{f}(x)\right] - f(x) = \frac{1}{2} f''(x) h^2 \int_{\mathbb{R}} x^2 K(u) du + O\left(h^4\right),$$

and the variance is

$$\mathrm{Var}\left[\widehat{f}(x)\right] = \frac{f(x)}{nh} \int_{\mathbb{R}} K^2(u) du + o\left(\frac{1}{nh}\right).$$

Hence, it is clear that as h increases the variance is reduced by the order $\frac{1}{h}$, but the bias increases quadratically. Thus, h is a smoothing parameter that directly controls the trade off between

bias and variance. This is exhibited in Figure 1.4, which shows the kernel density estimates of a Normal sample of size $n = 50$ using a Normal kernel K and a range of bandwidths h. It is apparent that the bandwidth $h = 2$ is too large and results in an 'over-smoothed' density estimate $\widehat{f}$ with has low variance but excessive bias. By contrast, the kernel density estimate with bandwidth $h = 0.1$ is 'under-smoothed' and is overly sensitive to stochastic variability in the sample. In other words, in this case the density estimate has reduced the bias, but at the cost of excessive variance in the estimate. Finally, it is clear from the figure that the density estimate using the bandwidth $h = 0.42$ provides a good approximation to the true density function and adequately balances the bias and variance.

The mean integrated square error (MISE) summarises the bias and variance in a single measure and is given by

$$\text{MISE}\left(\widehat{f}(x)\right) = \frac{1}{nh}\int_{\mathbb{R}} K^2(u)du + \frac{1}{4}h^4\left(\int_{\mathbb{R}} u^2K(u)dx\right)^2\left(\int_{\mathbb{R}}\left(f''(x)\right)^2dx\right)^2 + o\left(h^4\right) + o\left(\frac{1}{nh}\right).$$

The optimal bandwidth, that minimises the asymptotic MISE, is

$$h_{\text{opt}} = \frac{\left(\int_{\mathbb{R}} K^2(u)du\right)^{\frac{1}{5}}}{\left(\int_{\mathbb{R}} u^2K(u)du\right)^{\frac{2}{5}}\left(\int_{\mathbb{R}}\left(f''(x)\right)^2dx\right)^{\frac{1}{5}}n^{\frac{1}{5}}}.$$

Of course, this optimal bandwidth cannot be used in practice, as it depends on the unknown density f via its second derivative f''. As a result, there are a number of data based procedures to select the bandwidth h. The simplest procedure, suggested by Silverman [114], is to use the bandwidth which is optimal for Gaussian f and K,

$$h^*_{\text{opt}} = \left(\frac{4\widehat{\sigma}^5}{3n}\right)^{\frac{1}{5}},$$

where $\widehat{\sigma}$ is the standard deviation of the sample. This is precisely the method used to obtain the optimal bandwidth $h = 0.42$ in Figure 1.4, and it provides a good estimate of the optimal bandwidth when the assumptions of Normal f and K hold. Alternative methods, which are more robust to violations of this assumption, include the plug-in method [111] and cross validation

[46].

Epanechnikov [32] demonstrated that the optimal kernel K, in terms of mean square error, is given by

$$K_{\text{opt}}(u) = \frac{3}{4}(1-u^2)\mathrm{I}\left[-1 \leq u < 1\right].$$

However, it is also shown by Wand and Jones [124] that the efficiency of more conventional kernels, such as the Normal and uniform densities, remains relatively close to the Epanechnikov kernel.

The distribution function $F(x)$ can also be estimated using the smooth estimate,

$$\begin{aligned}
\widehat{F}(x) &= \int_{-\infty}^{x} \widehat{f}(u)du \\
&= \int_{-\infty}^{x} \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{u-X_i}{h}\right)du \\
&= \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{x} K\left(\frac{u-X_i}{h}\right)du \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\frac{x-X_i}{h}} K\left(v\right)dv \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{K}\left(\frac{x-X_i}{h}\right),
\end{aligned}$$

where $\mathbb{K}(x) = \int_{-\infty}^{x} K(u)du$ is a cumulative distribution kernel function.

In this thesis we will focus predominantly on kernel density estimation, however, there are a variety of other non-parametric methods for estimating a density function. For example, one can use wavelet estimation [26], spline smoothing [42], or local polynomial smoothing [106].

1.5 Aims of the thesis

The overarching aim of this thesis is to investigate new methods for making inferences about the nature of an unknown distribution based on a random sample, with particular attention paid to applications in medical research. More specifically, the aims are outlined as follows.

- Develop a new test for symmetry based on a recently proposed measure of asymmetry, which improves upon the existing methods.

- Explore this measure of asymmetry in a number of applications relating to the analysis of randomised control trials.
- Modify the measure of asymmetry to improve upon its small sample properties.
- Extend this approach to measure asymmetry in several studies investigating the same population, by developing a formal meta-analysis model for the asymmetry measure.
- Determine the impact of ignoring departures from symmetry on the inferences drawn from statistical models, with special attention paid to prediction intervals that are derived from meta-analyses.
- Expand the investigation further to consider a meta-analysis of a non-parametric estimate of the density or distribution function.
- Develop new methods for conducting meta-analyses of diagnostic test accuracy studies, which have a number of desirable features compared to the existing methodology.

1.6 Outline of the upcoming chapters in this thesis

In Chapter 2 we introduce and discuss some of the commonly used tests for symmetry. Through the use of a simulation study we identify that the existing tests do not have power which accurately reflects the magnitude of asymmetry in the sample. As a result, we propose a new test based on a recently proposed measure of asymmetry $\widehat{\eta}$, which has been shown to effectively capture the amount of asymmetry. We derive the asymptotic properties of this new test statistic and compare the power of the new test with the existing methods.

In Chapter 3 we discuss the potential applications of the new test, with a particular emphasis on informing the analysis of randomised trials in medical research. Moreover, with the help of a large data set consisting of multiple randomised control trials investigating hypertension, we also show that the measure $\widehat{\eta}$ is an effective summary statistic. Indeed, $\widehat{\eta}$ can be used for making inferences such as assessing baseline imbalance in clinical trials and checking for asymmetry in linear model residuals. Furthermore, $\widehat{\eta}$ can be a useful aid in preconditioning data. That is, it can be used to objectively determine whether asymmetry within data is sufficient to require a

normalising transform, inform what sort of transformation is likely to be effective, and provide an objective measure of whether the transformation has been a success.

In Chapter 4 we demonstrate that it is not appropriate to assume a Normal distribution for $\widehat{\eta}$ for small samples, before proposing a new estimation procedure which is more robust to small samples. Furthermore, we derive the asymptotic distribution of this new measure and demonstrate that this distribution is more robust to small samples. We also introduce and discuss the bootstrap, to assist in the calculation of the standard error of the new measure and thereby allow us to construct accurate confidence intervals. We also appraise the new methodology using a simulation study and a couple of real data examples.

In Chapter 5 we investigate measuring asymmetry in data across several studies with the aim of obtaining an 'overall' measure of asymmetry in the underlying population. In particular, we explore the potential pitfalls of naively pooling data across all studies (and ignoring the original clustering), before proposing a more formal meta-analysis model to summarise the asymmetry in a robust fashion.

In Chapter 6 we carry out a simulation study to assess what effect, if any, the presence of skewed data has on a number of common statistical models. In particular, we discuss the potential pitfalls of ignoring violations of symmetry or normality assumptions.

In Chapter 7 we generalise the methods of Chapter 5 to allow for the synthesis of information about the entire density or distribution function of several studies. We also provide a detailed demonstration of one potential application of this method, namely, conducting meta-analyses of diagnostic test accuracy studies. In particular, we propose a novel method that has a number of desirable properties when compared to the existing methodology in this area.

In Chapter 8 we present the conclusions of this thesis, and discuss the areas for future work.

Chapter 2

η - A measure of asymmetry and a new test of symmetry

2.1 Introduction

As we have previously discussed, symmetric random variables are important for the development and application of statistical theory. In particular, symmetry is an important assumption for many statistical models. For example, symmetry assumptions are often essential to the derivation of point or interval estimates of location parameters. In non-parametric statistics such as the Wilcoxon signed-rank test, given in equation (1.6) in the previous chapter, it is fundamental to assume that the data are drawn from a symmetric population [128]. In the latter case one should investigate whether that the data are indeed symmetric, before proceeding with the test. Furthermore, a wide range of statistical techniques rely on the assumption of symmetry implicitly, through the assumption of normality. For example, linear regression models assume that residuals are normally distributed. Assessing the symmetry of the residual distribution is an important precursor in assessing the normality of the residuals. As a result, there are a wealth of options for the statistician wishing to conduct tests of symmetry on a set of data and we discuss some of these in section 2.2.

In this chapter we examine the power of several existing tests of symmetry and, in doing so, motivate the development of a new test. In section 2.2 we carry out a simulation study to

examine the power of the existing tests of symmetry. In section 2.3 we introduce a recently proposed measure of asymmetry η, which has been shown to adequately quantify the amount of asymmetry [88]. Using this new measure we construct a new test, which aims to improve on the power of the existing tests. We discuss the asymptotic properties of the new test statistic and finally, we carry out a simulation study and compare the power of the new test with the existing tests. In section 2.4 we provide a short worked example of how the test can be applied to test for symmetry in a real data set. In section 2.5 we provide some concluding remarks on the measure η and the new test.

Aims of the chapter:

- Examine the power of existing tests for symmetry.
- Propose and develop a new test for symmetry T_n.
- Compare the performance of T_n with existing tests.

2.2 Testing symmetry

2.2.1 Introduction

In this section we carry out a simulation study on several existing tests of symmetry, namely, the tests proposed by Cabilio and Masaro [17], Antille et al. [3], Randles et al. [95], Gupta [43], and McWilliams [81]. We study the power of these tests using simulated data from a number of different distributions with varying degrees of asymmetry. For example, Figure 2.1 shows the density functions of some of these random variables, namely, the Normal, Cauchy, Normal mixtures, Log-Normal, Folded Normal, and Exponential distributions. The Normal mixtures are constructed using

$$pN(0,1) + (1-p)N(2,2),$$

for $p = 0.945$, 0.872, 0.773, and 0.606.

It is clear from the figure that the Normal and Cauchy densities are symmetric about zero,

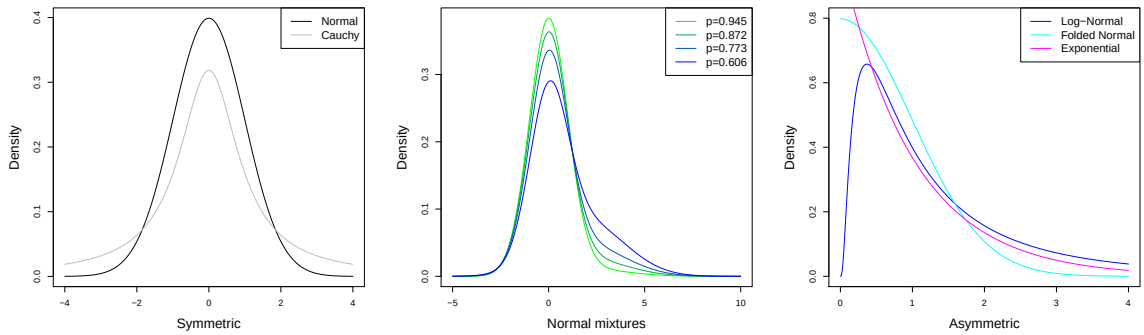

Figure 2.1: The left plot shows the symmetric Normal and Cauchy density curves. The middle plot shows the density curves of Normal mixtures of the form $pN(0,1) + (1-p)N(2,2)$, for $0 < p < 1$. The rightmost plot shows the three 'highly' asymmetric densities which, in order of increasing asymmetry, are Log-Normal, Folded Normal and Exponential.

whilst the other density functions are clearly asymmetric. However, we are entitled to ask, 'which of these asymmetric densities is the *most* asymmetric?' In this case, it is possible to obtain a visual impression of the size of asymmetry present in the random variables. For example consider the middle plot in Figure 2.1, which shows four Normal mixture densities. As p decreases the $N(2,2)$ population has more of an effect on the mixture density and the curve is skewed to the right. It is clear that as p decreases from 1 towards 0.5, the resultant density becomes more asymmetric. Note how, in the rightmost plot containing the most extreme cases, the Log-Normal density is even further skewed to the right and as a result, it is reasonable to say that it is even more asymmetric than the Normal mixtures. Further, the Folded Normal and the Exponential density represent an even more extreme example of asymmetry as they have no left tail whatsoever. Observe that the Folded Normal curve has a 'more even spread' of probability mass compared to the exponential curve, hence one can reason that a Folded Normal random variable is not as asymmetric as an Exponential random variable.

Thus, for the random variables given above we can arrive at the following ordering of asymmetry, based on visual interpretation.

$$\text{Normal} <_a \text{Normal mixtures} <_a \text{Log-Normal} <_a \text{Folded Normal} <_a \text{Exponential},$$

where the binary operator $<_a$ represents the sentence "appears to be less asymmetric than". An 'ideal' test statistic would have power which reflects this increasing departure from symmetry.

2.2.2 Existing tests for symmetry

Consider a random sample $X_1, \cdots, X_n$ identically drawn from a probability distribution. Then, Cabilio and Masaro [17] propose a test based on sample skewness,

$$S_1 = \sqrt{n}\ \frac{\overline{x} - \tilde{\theta}}{s},$$

where $\overline{x}$ and $\tilde{\theta}$ are the sample mean and sample median respectively and s is the sample standard deviation. The simple rationale behind this statistic is the necessary condition that for a symmetric continuous population the mean is equal to the median. Thus, significantly large values of $|S_1|$ are indicative of departure from symmetry. The main focus of S_1 is detecting departure from symmetry and and not the quantification of asymmetry.

Another test is suggested by Antille et al. [3], who define the following test statistic based on ranks,

$$\mathcal{R}(\alpha) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} G_\alpha \left(\frac{R\left(|X_i - \tilde{\theta}|\right)}{2(n+1)} \right) \text{sign}\left(X_i - \tilde{\theta}\right),$$

where $G_\alpha(x) = \min(x, \frac{1}{2} - \alpha)$ and $R(X_i)$ is defined as the rank of X_i among the X_is. Antille et al. [3] propose a test based on $\mathcal{R}(\alpha)$, and determine the asymptotic properties of the test statistic. For simplicity we only consider $\alpha = 0$ and denote $S_2 = \mathcal{R}(0)$. For symmetric random variables one expects that the distances $\left|X_i - \tilde{\theta}\right|$ should be of a similar magnitude on both sides of the median. In this case, one also expects the ranks $R\left(|X_i - \tilde{\theta}|\right)$ to be evenly distributed on both sides of θ. Thus, under the null hypothesis of symmetry S_2 is very close to zero and hence one rejects the null for large values of $|S_2|$.

Alternatively, Randles et al. [95] define the following 'triples' test,

$$S_3 = \frac{1}{3}\binom{n}{3}^{-1} \sum_{i<j<k} \Big[\text{sign}(X_i + X_j - 2X_k) + \text{sign}(X_i + X_k - 2X_j) + \text{sign}(X_j + X_k - 2X_i)\Big].$$

A triple of observations (X_i, X_j, X_k) is defined as a right triple if the middle observation is closer

to the smallest observation than it is to the largest observation, and vice-versa for a left triple. Thus, defined in this way, S_3 is a constant multiple of the difference between the proportion of right and left triples. As a result, $E[S_3] = 0$ when the underlying distribution is symmetric. Suggesting that the class of asymmetric distributions for which $E[S_3] = 0$ is small, Randles et al. [95] use S_3 to test for symmetry.

Gupta [43] details the classical test of skewness based on

$$S_4 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^3}{\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right)^{\frac{3}{2}}}.$$

Again, as with S_1 the rationale behind this test statistic is that a symmetric population has zero skewness. Indeed, the theoretical analogue of S_4 is precisely Pearson's skewness measure γ_3 introduced in Chapter 1. Thus, significantly large values of $|S_4|$ are indicative of departure from symmetry. However, as with the other test statistics discussed here the main focus of this test statistic is to identify departure from symmetry and and not to quantify the asymmetry.

Finally, we consider the test proposed by McWilliams [81], based on a runs statistic. Suppose now that $X_1, X_2, \cdots, X_n$, denote a random sample with median 0. Further, let $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ denote the sample values ordered from smallest to largest according to their absolute value, but retaining their sign, and let Δ_i indicate the sign of $X_{(i)}$, $i = 1, 2, \cdots, n$, by way of defining $\Delta_i = 1$ when $X_{(i)} > 0$ and zero otherwise. Then define

$$S_5 = 1 + I_2 + I_3 + \cdots + I_n,$$

where

$$I_k = \begin{cases} 0 & \text{if} \quad \Delta_k = \Delta_{k-1} \\ 1 & \text{if} \quad \Delta_k \neq \Delta_{k-1} \end{cases}, \quad k = 2, \cdots, n$$

which counts the number of runs in the sequence $\{\Delta_i\}$. For symmetric random variables the probability of two successive observations $X_{(k-1)}$ and X_k having the same sign is equal to $\frac{1}{2}$.

Hence, under the null hypothesis of symmetry, $S_5 - 1$ has a binomial distribution with parameters $n-1$ and $\frac{1}{2}$. By contrast, for asymmetric distributions one may be more likely to obtain a run of successive observations with the same sign. This results in a larger number of the I_k being equal to zero and, thus, one rejects the null hypothesis if S_5 falls in the lower tail of the null distribution.

With the exception of S_5 all of the test statistics discussed here are asymptotically normally distributed. For further details regarding the sampling distributions of $S_1, \cdots, S_5$ refer to Table 2.3 in the next section. In the next subsection we conduct a simulation study to assess the power of the tests discussed here.

2.2.3 Simulation study

Aims and methods

We now approximate the power (i.e. calculate the empirical power) of the tests proposed by Cabilio and Masaro [17], Antille et al. [3], Randles et al. [95], Gupta [43] and McWilliams [81] for a range of different distributions. In particular, we consider the symmetric Normal distribution; the Normal mixtures; as well as the the extreme Log-Normal, Folded Normal and Exponential distributions. We also supplement these random variables with the symmetric Cauchy distribution, as well as several other classes of asymmetric distributions, namely, the Skew Normal distributions proposed by Azzalini [4]; the Sinh-arcsinh distribution proposed by Jones and Pewsey [64]; and the skewed distributions introduced by Fernandez and Steel [33]. The Skew Normal distribution with parameter λ has density function

$$SN(z;\lambda) = 2\phi(z)\Phi(\lambda z), \quad -\infty < z < \infty,$$

where ϕ and Φ are the standard Normal density and distribution functions respectively. When $\lambda = 0$ this reduces to the symmetric standard Normal distribution. When $\lambda > 0$ the distribution is skewed to the right and when $\lambda < 0$ the distribution is skewed to the left.

The Sinh-arcsinh distribution has density function

$$SAS(z;\varepsilon,\delta) = \frac{1}{\sqrt{2\pi}} \frac{\delta C_{\varepsilon,\delta}(z)}{\sqrt{1+z^2}} \exp\left\{-\frac{1}{2} S^2_{\varepsilon,\delta}(z)\right\},$$

where

$$C_{\varepsilon,\delta}(x) = \cosh\left[\delta \sinh^{-1}(x) - \varepsilon\right],$$

and

$$S_{\varepsilon,\delta}(x) = \sinh\left[\delta \sinh^{-1}(x) - \varepsilon\right].$$

Here $\varepsilon \in \mathbb{R}$ plays the role of a skewness parameter, while $\delta > 0$ controls the weight of the tails.

The skewed Fernandez and Steel distribution has density function

$$FAS(z;\gamma) = \frac{2}{\gamma + \frac{1}{\gamma}} \left\{ f\left(\frac{z}{\gamma}\right) \mathrm{I}[z \geq 0] + f(\gamma z)\, \mathrm{I}[z < 0] \right\},$$

for a unimodal density function f and some $\gamma \in (0, \infty)$. This distribution is symmetric when $\gamma = 1$ and is asymmetric whenever $\gamma \neq 1$.

Figure 2.2 shows the probability density functions of the Skew Normal, Sinh-arcsinh and Fernandez and Steel distributions for a range of the skewness parameters λ, ε and γ.

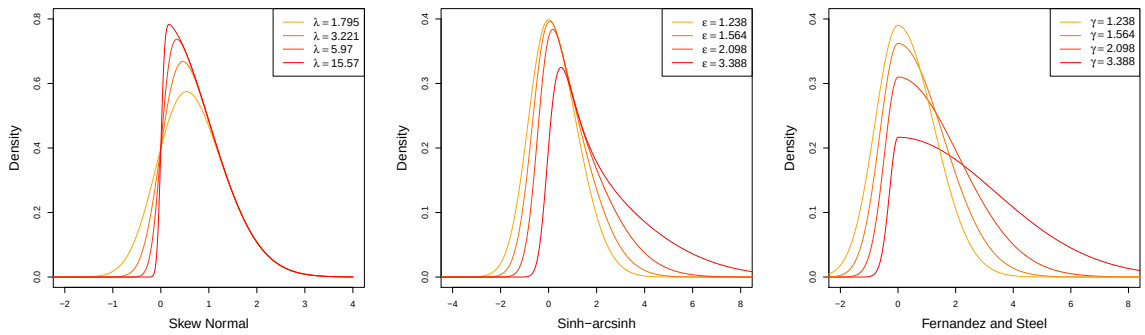

Figure 2.2: The probability density functions of the Skew Normal, Sinh-arcsinh and Fernandez and Steel distribution for a range of skewness parameters.

For the simulation study we consider the Skew Normal distributions with $\lambda = 1.2135$, 1.795, 2.429, 3.221, 4.310, 5.970, 8.890, 15.570 respectively; the Sinh-arcsinh distribution with $\delta = 1$

and $\varepsilon = 0.1$, 0.203, 0.311, 0.430, 0.565, 0.727, 0.939, 1.263; the Fernandez and Steel distribution where f is the standard Normal probability density function and $\gamma = 1.111$, 1.238, 1.385, 1.564, 1.791, 2.098, 2.557, 3.388.

We simulate samples of varying sizes ($n = 30, 50$ and 70) from each of the probability models using the statistical package R. We simulate each sample $10,000$ times and calculate the test statistics each time to obtain a large sample from the sampling distributions of the test statistics. In each sample the null hypothesis of symmetry is accepted or rejected at the level $\alpha = 0.05$, based on the value of these statistics. The critical value, at which to reject the null hypothesis of symmetry, is determined from the asymptotic distribution of the test statistics, given in Table 2.3. Finally, the empirical powers (the proportion of rejections) of each of the tests are reported. Table 2.1 outlines the simulation procedure.

We present the empirical powers of the test based on sample skewness S_1 proposed by Cabilio and Masaro [17]; the test based on ranks S_2 suggested by Antille et al. [3]; the triples test S_3 proposed by Randles et al. [95]; the classical test of skewness S_4 presented by Gupta [43]; and runs test S_5 proposed by McWilliams [81]. Table 2.2 gives a brief summary of the test statistics. Further details regarding the sampling distributions of $S_1, \cdots, S_5$ are given in Table 2.3.

Step 1	Simulate samples of size $n = 30, 50$ and 70 from each of the distributions.
Step 2	Calculate the test statistics S_1, S_2, S_3, S_4 and S_5 based on these samples.
Step 3	If the test statistics fall inside the respective rejection region of the test (see Table 2.3) then reject the null hypothesis of symmetry at the 5% level.
Step 4	Repeat **Steps 1-3** $10,000$ times and report the proportion of rejections (empirical power).

Table 2.1: Step by step guide to the simulation study for the existing tests of symmetry.

Reference	Test statistic
Cabilio and Masaro [17]	$S_1 = \sqrt{n}\,\dfrac{\overline{x} - \tilde{\theta}}{s}$
Antille et al. [3]	$S_2 = \dfrac{1}{\sqrt{n}} \displaystyle\sum_{i=1}^{n} G_0\left(\dfrac{R\left(\lvert X_i - \tilde{\theta}\rvert\right)}{2(n+1)}\right) \operatorname{sign}\left(X_i - \tilde{\theta}\right)$
Randles et al. [95]	$S_3 = \dfrac{1}{3}\binom{n}{3}^{-1} \displaystyle\sum_{i<j<k} \Big[\operatorname{sign}(X_i + X_j - 2X_k) + \operatorname{sign}(X_i + X_k - 2X_j) + \operatorname{sign}(X_j + X_k - 2X_i)\Big]$
Gupta [43]	$S_4 = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^3}{\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right)^{\frac{3}{2}}}$
McWilliams [81]	$S_5 = 1 + I_2 + I_3 + \cdots + I_n$

Table 2.2: Summary of the existing tests for symmetry used in the simulation study.

$\mathbf{S_1}$	$S_1 \xrightarrow{\text{L}} N(0, \sigma_1^2)$, where $\sigma_1^2 = 1 + \dfrac{1}{e_{m,\overline{x}}(F)} - \dfrac{2}{\sqrt{e_{m,\overline{x}}(F)}}\text{E}\left\lvert\dfrac{X-\mu}{\sigma}\right\rvert$, and $e_{m,\overline{x}}(F) = 4\sigma^2 f^2(0)$.
$\mathbf{S_2}$	$S_2 \xrightarrow{\text{L}} N(0, \sigma_2^2)$, where $\sigma_2^2 = \dfrac{1}{48} + \dfrac{1}{4}\left(A - \dfrac{1}{2}\right)$, and $A = \dfrac{1}{f(0)}\displaystyle\int_0^1 f(F^{-1}(u))du.$
$\mathbf{S_3}$	$S_3 \xrightarrow{\text{L}} N(0, \sigma_3^2)$, where $\sigma_3^2 = n\binom{n}{3}^{-1}\displaystyle\sum_{c=1}^{3}\binom{3}{c}\binom{n-3}{3-c}\text{Var}\left[f_c^*(X_1, \cdots, X_c)\right]$, and $f_c^*(x_1, \cdots, x_c) = \text{E}\left[f^*(x_1, \cdots, x_c, X_{c+1}, \cdots X_3)\right], \qquad c = 1, 2, 3$ and $f^*(X_i, X_j, X_k) = \dfrac{1}{3}\Big[\text{sign}(X_i + X_j - 2X_k) + \text{sign}(X_i + X_k - 2X_j)$ $+\text{sign}(X_j + X_k - 2X_i)\Big]$.
$\mathbf{S_4}$	$S_4 \xrightarrow{\text{L}} N(0, \sigma_4^2)$, where $\sigma_4^2 = \dfrac{y_6 - 6y_2y_4 + 9y_2^3}{ny_2^3}$, and $y_j = \text{E}\left[X^j\right]$.
$\mathbf{S_5}$	$S_5 \sim \text{Bin}\left(n-1, \dfrac{1}{2}\right)$.

Table 2.3: Sampling distributions for the test statistics $S_1, \cdots, S_5$ under the null hypothesis of symmetry.

Results

Let NM1, NM2, NM3, NM4 denote the Normal mixtures with $p = 0.945$, 0.872, 0.773, 0.606 respectively, and let SN1-SN8, denote the Skew Normal distribution with $\lambda = 1.2135$, 1.795, 2.429, 3.221, 4.310, 5.970, 8.890, 15.570 respectively. Let SAS1-SAS8 denote the Sinh-arcsinh distribution with $\delta = 1$ and $\varepsilon = 0.1$, 0.203, 0.311, 0.430, 0.565, 0.727, 0.939, 1.263 respectively. Let FAS1-FAS8 denote the Fernandez and Steel distribution with $\gamma = 1.111$, 1.238, 1.385, 1.564, 1.791, 2.098, 2.557, 3.388 respectively. The empirical powers are shown in Table 2.4 and Table 2.5. The tables also include a column entitled η, which is a measure of the asymmetry in the distribution, something we will define in detail in section 2.3.

In Table 2.5, observe that for S_1 the test has nominal empirical power for the symmetric Normal distribution, in keeping with the set level of 0.05. For $n = 30$ the empirical power is 0.036 rising to 0.04 for $n = 70$. As expected, the power steadily increases for the Normal mixtures and Skew Normal distributions, however, the amount of power is not related to the amount of asymmetry as determined by our previous 'visual inspection'. For example, for $n = 30$ in Table 2.4 the test S_1 has power equal to 0.282 and 0.709 for the Folded Normal and Log-Normal distributions respectively, although we understand the Folded Normal distribution to be the more asymmetric distribution.

The performance of the test S_2 is very similar to S_1. For the symmetric distributions the test has nominal power and this slowly increases for the Normal mixtures and Skew Normal distributions. However, in the very asymmetric cases there is very little relation between the power of the test and the amount of asymmetry in the distribution. For example, in Table 2.4 for $n = 70$, although the Log-Normal density is less asymmetric than Exponential density, the power of S_2 does not reflect this with values of 0.704 and 0.603 respectively. To reiterate, whilst the test has good rejection levels for non-symmetric distributions, its power does not reflect the size of asymmetry.

For the asymmetric distributions the test S_3 does achieve very high empirical power. Also, although it is not perfect, it does appear to capture the size of the asymmetry more accurately than S_1 and S_2. However, the test does not appear to be a conservative. Indeed, the test has

η	Dist.	S_1			S_2			S_3		
		$n=30$	$n=50$	$n=70$	$n=30$	$n=50$	$n=70$	$n=30$	$n=50$	$n=70$
0	N	0.036	0.042	0.040	0.016	0.030	0.035	0.082	0.068	0.065
0	C	0.033	0.029	0.023	0.004	0.011	0.012	0.108	0.079	0.075
0.1	NM1	0.040	0.055	0.064	0.015	0.027	0.031	0.113	0.125	0.134
0.2	NM2	0.078	0.110	0.159	0.017	0.031	0.041	0.200	0.284	0.353
0.3	NM3	0.128	0.238	0.338	0.023	0.050	0.074	0.369	0.519	0.645
0.4	NM4	0.212	0.368	0.509	0.047	0.109	0.173	0.483	0.693	0.835
0.1	SN1	0.043	0.046	0.052	0.018	0.032	0.036	0.100	0.100	0.106
0.2	SN2	0.062	0.075	0.100	0.026	0.039	0.058	0.152	0.186	0.229
0.3	SN3	0.086	0.135	0.185	0.031	0.062	0.092	0.224	0.327	0.427
0.4	SN4	0.130	0.209	0.286	0.042	0.087	0.139	0.332	0.494	0.638
0.5	SN5	0.184	0.295	0.394	0.066	0.133	0.198	0.438	0.629	0.770
0.6	SN6	0.224	0.353	0.471	0.082	0.165	0.243	0.524	0.734	0.867
0.7	SN7	0.251	0.393	0.523	0.097	0.190	0.275	0.588	0.790	0.903
0.8	SN8	0.267	0.407	0.548	0.106	0.207	0.306	0.617	0.826	0.932
0.1	SAS1	0.042	0.053	0.053	0.020	0.033	0.040	0.101	0.098	0.100
0.2	SAS2	0.057	0.084	0.097	0.023	0.047	0.059	0.149	0.178	0.216
0.3	SAS3	0.091	0.133	0.180	0.034	0.062	0.095	0.227	0.312	0.413
0.4	SAS4	0.141	0.226	0.310	0.050	0.100	0.151	0.332	0.487	0.627
0.5	SAS5	0.211	0.344	0.459	0.071	0.158	0.226	0.476	0.677	0.815
0.6	SAS6	0.292	0.471	0.615	0.097	0.215	0.322	0.615	0.826	0.923
0.7	SAS7	0.398	0.623	0.761	0.141	0.309	0.440	0.742	0.920	0.978
0.8	SAS8	0.516	0.732	0.863	0.198	0.394	0.562	0.843	0.967	0.995
0.1	FAS1	0.044	0.048	0.054	0.022	0.030	0.044	0.099	0.090	0.096
0.2	FAS2	0.056	0.079	0.099	0.024	0.051	0.064	0.141	0.173	0.207
0.3	FAS3	0.090	0.127	0.167	0.034	0.070	0.096	0.208	0.293	0.376
0.4	FAS4	0.120	0.185	0.244	0.046	0.092	0.129	0.292	0.432	0.552
0.5	FAS5	0.154	0.242	0.326	0.058	0.125	0.176	0.390	0.571	0.696
0.6	FAS6	0.191	0.305	0.399	0.074	0.150	0.221	0.472	0.662	0.814
0.7	FAS7	0.223	0.356	0.457	0.087	0.177	0.255	0.536	0.745	0.871
0.8	FAS8	0.245	0.387	0.511	0.099	0.196	0.281	0.590	0.799	0.908
0.91	LN	0.709	0.916	0.976	0.263	0.514	0.704	0.973	0.999	1.000
0.95	FN	0.282	0.426	0.552	0.114	0.213	0.309	0.633	0.837	0.936
1	EXP	0.610	0.841	0.940	0.217	0.450	0.603	0.917	0.992	0.999

Table 2.4: Empirical power of the tests based on S_1, S_2 and S_3 for a variety of density functions and sample sizes.

η	Dist.	S_4			S_5		
		$n = 30$	$n = 50$	$n = 70$	$n = 30$	$n = 50$	$n = 70$
0	N	0.027	0.034	0.041	0.016	0.030	0.031
0	C	0.024	0.013	0.009	0.014	0.025	0.029
0.1	NM1	0.033	0.045	0.061	0.015	0.030	0.037
0.2	NM2	0.088	0.160	0.244	0.017	0.046	0.051
0.3	NM3	0.187	0.354	0.523	0.027	0.061	0.085
0.4	NM4	0.275	0.499	0.669	0.042	0.100	0.137
0.1	SN1	0.039	0.052	0.068	0.017	0.030	0.035
0.2	SN2	0.058	0.102	0.159	0.020	0.035	0.045
0.3	SN3	0.098	0.198	0.307	0.026	0.048	0.065
0.4	SN4	0.146	0.298	0.467	0.036	0.076	0.103
0.5	SN5	0.198	0.407	0.603	0.052	0.116	0.149
0.6	SN6	0.247	0.500	0.693	0.074	0.152	0.217
0.7	SN7	0.284	0.553	0.750	0.092	0.194	0.289
0.8	SN8	0.305	0.593	0.779	0.122	0.255	0.367
0.1	SAS1	0.038	0.048	0.067	0.015	0.029	0.038
0.2	SAS2	0.054	0.099	0.143	0.019	0.033	0.042
0.3	SAS3	0.093	0.186	0.284	0.027	0.052	0.069
0.4	SAS4	0.154	0.310	0.467	0.036	0.077	0.100
0.5	SAS5	0.223	0.449	0.635	0.052	0.113	0.166
0.6	SAS6	0.304	0.571	0.761	0.080	0.180	0.262
0.7	SAS7	0.383	0.662	0.823	0.128	0.285	0.392
0.8	SAS8	0.446	0.722	0.844	0.204	0.432	0.591
0.1	FAS1	0.037	0.045	0.063	0.015	0.032	0.033
0.2	FAS2	0.057	0.095	0.144	0.018	0.039	0.046
0.3	FAS3	0.085	0.164	0.262	0.024	0.051	0.068
0.4	FAS4	0.131	0.270	0.422	0.037	0.071	0.092
0.5	FAS5	0.180	0.366	0.556	0.050	0.102	0.140
0.6	FAS6	0.226	0.466	0.659	0.061	0.142	0.196
0.7	FAS7	0.260	0.525	0.728	0.091	0.193	0.274
0.8	FAS8	0.294	0.576	0.770	0.114	0.250	0.355
0.91	LN	0.303	0.353	0.392	0.329	0.663	0.831
0.95	FN	0.321	0.609	0.792	0.138	0.314	0.454
1	EXP	0.409	0.587	0.677	0.306	0.629	0.808

Table 2.5: Empirical power of the tests based on S_4 and S_5 for a variety of density functions and sample sizes.

an estimated type-I error rate of 0.082 for a sample of size $n = 30$ from a Normal population and 0.075 for a substantial sample of size $n = 70$ from a Cauchy distribution.

For the classical test of skewness S_4 the test has nominal power for the symmetric distributions, although the test appears to be overly conservative for the Cauchy case. Again, for the asymmetric distributions the test fails to capture the asymmetry present in the most asymmetric distributions. For example, in Table 2.5 when $n = 70$ the test has empirical power 0.792 for the Folded Normal distribution, but has much less power (0.677) to detect asymmetry for the Exponential distribution.

As with the previous tests S_5 achieves nominal empirical power for the symmetric Normal and Cauchy distributions. There is also a steady increase in power through the asymmetric Normal mixtures and Skew Normal distributions. Once again, however, the empirical power does not perfectly reflect the size of asymmetry. For example, when $n = 70$ the test has empirical powers of 0.831 and 0.454 for the Log-Normal and Folded Normal distributions.

Hence, we have demonstrated that the tests considered here all fail to generate power that is representative of the quantity of asymmetry in the underlying distribution. In the next section we investigate the effectiveness of several tests that are optimal for a certain class of asymmetric distributions.

2.2.4 Other tests for symmetry

There are many other methodologies for testing symmetry and for further details on these different methods refer to Hollander [56] and the references therein. For the tests that are under investigation here, we demonstrated that the tests do not have power which is reflective of the size of asymmetry. That is, while the tests have good rejection levels for non-symmetric distributions their power does not increase as asymmetry increases. This is because the primary rationale for the test statistics that are proposed in the literature to test for symmetry is to detect departure from symmetry, rather than the quantification of asymmetry. It is not practical to demonstrate this point, through simulations or otherwise, for all other tests of symmetry. However, based on the discussion given in section 2 of Patil, Patil and Bagkavos [88] one can conclude that the tests of symmetry proposed in Butler [16], Rothman and Woodroofe [105] and Boos [10] will

also fail to have power, which is an increasing function of the amount of asymmetry. That is, it is shown that the population versions of the test statistics do not effectively quantify asymmetry in a certain class of distributions. The reason for this, as mentioned above, is that these test statistics are also designed primarily to capture the departure from symmetry as opposed to the dual purpose of capturing and quantifying the departure from symmetry. Recent research has led to the development of new measures aimed at quantifying the size of asymmetry and the main subject of the rest of this chapter is to explore the use of one such measure to test symmetry, and investigate whether this new test has power which is superior to the tests discussed here.

2.3 Measuring asymmetry

2.3.1 A recently proposed measure of asymmetry

Intuitively it is believed that asymmetry is something that can be measured. When presented with two similar density curves, most people are able to provide some rationale on why one is more or less asymmetric than the other density curve (it was precisely this type of reasoning that generated our $<_a$ orderings in the previous section.) Despite this, it is a challenge to find a mathematical expression to effectively calibrate or quantify the amount of asymmetry.

Several measures of asymmetry have been proposed and we introduced some examples in Chapter 1, however, we identified that each of these limits the class of density functions in one way or another. For further details, see MacGillivray [76] and Boshnakov [11] and the references therein. For a more general discussion on measuring asymmetry refer to Patil, Patil and Bagkavos [88] (hereafter referred to as PPB). In PPB [88] their proposal is to measure asymmetry using

$$\eta(X) = \begin{cases} -\text{Corr}(f(X), F(X)) & \text{if} \quad 0 < \text{Var}(f(X)) < \infty \\ 0 & \text{if} \quad \text{Var}(f(X)) = 0, \end{cases}$$

where X is a continuous random variable, with continuous function f as its probability density function and F as its distribution function. Note that η is appropriately defined to avoid division by zero when $\text{Var}(f(X)) = 0$, as is the case for the symmetric uniform distribution. PPB [88]

show that this user-friendly measure effectively quantifies asymmetry in a number of different distributions. The measure is based on the fact that, for a symmetric random variable X with continuous probability density function f and cumulative distribution function F,

$$\mathrm{Cov}(f(X), F(X)) = 0.$$

We present the proof of this result below.

Theorem 2.1 *If X is a symmetric random variable with probability density function f and cumulative distribution function F then $\eta(X) = 0$.*

Proof. It is sufficient to prove that $\mathrm{Cov}(f(X), F(X)) = 0$, or that

$$\mathrm{E}[f(X)F(X)] = \mathrm{E}[f(X)]\mathrm{E}[F(X)].$$

Now, since $F(X) \sim U(0,1)$,

$$\mathrm{E}[F(X)] = \frac{1}{2}.$$

Once again, we take the centre of symmetry to be zero without loss of generality. Because X is a symmetric random variable we have $F(-x) = 1 - F(x)$ and $f(x) = f(-x)$ for all x. Using these identities gives

$$\begin{aligned}
\mathrm{E}[f(X)F(X)] &= \int_{-\infty}^{\infty} F(x)f^2(x)dx \\
&= \int_{-\infty}^{0} F(x)f^2(x)dx + \int_{0}^{\infty} F(x)f^2(x)dx \\
&= \int_{0}^{\infty} F(-y)f^2(-y)dy + \int_{0}^{\infty} F(x)f^2(x)dx \\
&= \int_{0}^{\infty} [1 - F(y)]f^2(y)dy + \int_{0}^{\infty} F(x)f^2(x)dx \\
&= \int_{0}^{\infty} f^2(x)dx \\
&= \frac{1}{2}\int_{-\infty}^{\infty} f^2(x)dx \\
&= \mathrm{E}[f(X)]\mathrm{E}[F(X)].
\end{aligned}$$

Hence, for a symmetric random variable X we have $\eta(X) = 0$. □

Table 2.6 gives some examples of the measure η applied to some commonly encountered distributions, including some of the distributions discussed earlier in the chapter. In particular, the table gives the value of η for the Normal, Uniform, t, Log-Normal, Folded Normal and Exponential distributions. Firstly, observe that $\eta = 0$ for the symmetric distributions. Second, the amount of asymmetry in the Folded Normal and Exponential distributions is independent of the location and scale parameters that are used. On the other hand, for the Log-Normal distribution (which is a transformation of the Normal (μ, σ^2) random variable), both parameters influence the amount of asymmetry.

Distribution	η
Normal (μ, σ^2)	0
Uniform (a, b)	0
t (k)	0
Log-Normal (μ, σ^2)	$\frac{3\left(2e^{\mu+\frac{1}{4\sigma^2}} \cdot D_\sigma - e^{\frac{1}{4}\sigma^2}\right)}{\sqrt{2\sqrt{3}e^{\frac{2}{3}\sigma^2} - 3e^{\frac{1}{2}\sigma^2}}}$
Folded Normal (μ, σ^2)	$\frac{3\pi - 12\tan^{-1}(\frac{1}{2}\sqrt{2})}{\pi\sqrt{2\sqrt{3}-3}} \approx 0.95$
Exponential (λ)	1

Table 2.6: The measure of η for a variety of commonly encountered distributions. The asymmetry in the Log-Normal distribution depends on both parameters μ and σ^2 as well as the probability $D_\sigma = P\left[-\infty < Z_1 < \frac{Z_2}{\sqrt{2}} - \frac{1}{2\sigma}, -\infty < Z_2 < \infty\right]$.

It is not always simple or convenient to compute an exact solution for η. For example, an exact expression for the Normal mixtures as a function of p is too long to be discussed here, but for details see the appendix of PPB [88]. Figure 2.3 shows the numerical approximation of η for a range of the parameters in the Normal mixtures (p), Skew Normal distribution (λ), Sinh-arcsinh distribution (ε) and the Fernanadez and Steel distribution (γ). Observe that the η identifies that the Normal mixtures are almost symmetric whenever p is close to either 0 or 1,

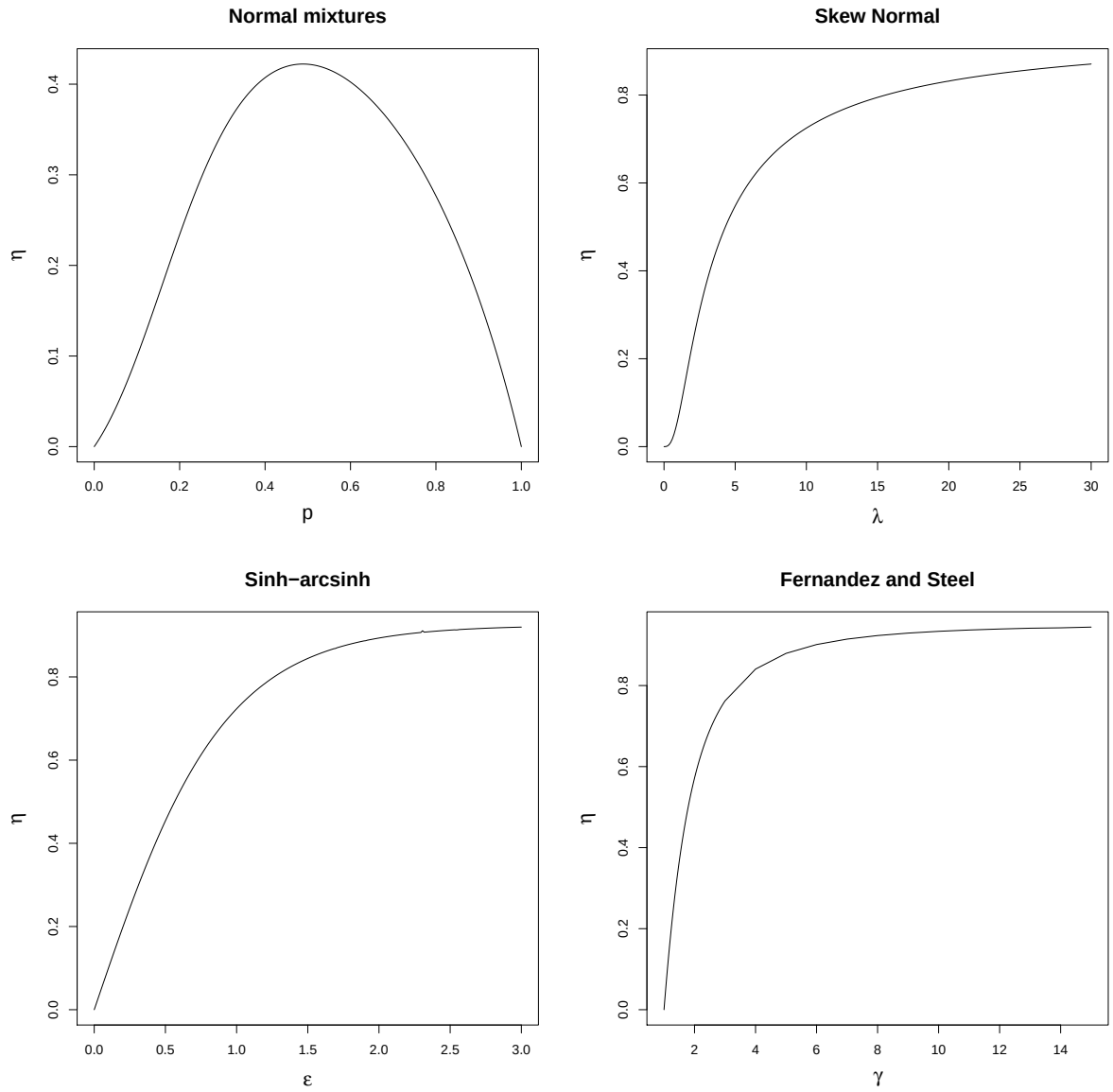

Figure 2.3: The amount of asymmetry η as a function of the parameters in the Normal mixtures (p), Skew Normal distribution (λ), Sinh-arcsinh distribution (ε) and the Fernandez and Steel distribution (γ).

and consequently the mixture is very close to $N(0,1)$ or $N(2,2)$. For the other distributions, as one increases the skewness parameter, the amount of asymmetry η also increases appropriately.

In a recent article Patil, Bagkavos and Wood [89] discuss a stronger measure

$$\eta_s(X) = -\frac{1}{2}\text{sign}(\rho_1) \max_{0.5\leq p\leq 1} \left|\rho_p + \rho_p^*\right|,$$

where

$$\rho_p = \frac{2\sqrt{3}}{p} \frac{\int_{-\infty}^{\xi_p} f^2(x)F(x)dx - \frac{p}{2}\int_{-\infty}^{\xi_p} f^2(x)dx}{\left[p\int_{-\infty}^{\xi_p} f^3(x)dx - \left(\int_{-\infty}^{\xi_p} f^2(x)dx\right)^2\right]^{1/2}},$$
$$\rho_p^* = \frac{2\sqrt{3}}{p} \frac{-\int_{\xi_{1-p}}^{\infty} f^2(x)F(x)dx + \frac{p}{2}\int_{\xi_{1-p}}^{\infty} f^2(x)dx}{\left[p\int_{\xi_{1-p}}^{\infty} f^3(x)dx - \left(\int_{\xi_{1-p}}^{\infty} f^2(x)dx\right)^2\right]^{1/2}},$$

and $F(\xi_p) = p$. Now the condition $\eta_s = 0$ is, both, necessary and sufficient for the symmetry. Unfortunately, a drawback of the stronger measure η_s is a loss of the 'user-friendly' aspect of η. Thus we propose to use η to devise a test for symmetry. But before that, the next subsection gives a brief description regarding the estimation of η.

2.3.2 Estimating η

PPB [88] construct three competing estimates of η. These are based upon calculating the sample correlation using different estimates for f and F. For example $f(X_i)$ is estimated using kernel smoothing,

$$\widehat{f}_{(i)}(X_i) = \frac{1}{n-1}\frac{1}{h}\sum_{j\neq i}^{n} K\left(\frac{X_j - X_i}{h}\right),$$

where K is a kernel density and h is the bandwidth. Now $F(X_i)$ is estimated by

$$\widehat{F}_{(i)}(X_i) = \frac{1}{n-1}\sum_{j\neq i}^{n} \text{I}[X_j < X_i].$$

Note that the estimates $\widehat{f}_{(i)}$ and $\widehat{F}_{(i)}$ are dependent on i, but for brevity this is suppressed $\widehat{f}_{(i)}(X_i) = \widehat{f}(X_i)$ and $\widehat{F}_{(i)}(X_i) = \widehat{F}(X_i)$. We will use the following estimator of η,

$$\widehat{\eta} = -\frac{\sum_{i=1}^{n} \widehat{f}(X_i)\widehat{F}(X_i) - n\left(\overline{\widehat{f}}\right)\left(\overline{\widehat{F}}\right)}{\sqrt{\left(\sum_{i=1}^{n}(\widehat{f}(X_i))^2 - n\overline{\widehat{f}}^2\right)\left(\sum_{i=1}^{n}(\widehat{F}(X_i))^2 - n\overline{\widehat{F}}^2\right)}}, \tag{2.2}$$

where $\overline{\widehat{f}} = \frac{1}{n}\sum_i \widehat{f}(X_i)$ and $\overline{\widehat{F}} = \frac{1}{n}\sum_i \widehat{F}(X_i)$. It was shown via simulation that $\widehat{\eta}$ is the most effective estimator of η considered by PPB [88]. Furthermore, PPB state that standard methods can be used to show the consistency of this estimate.

When calculating $\widehat{\eta}$ in practice it is helpful to note that the correlation is unchanged by altering the order of the data, so suppose that the random sample $X_1, \cdots, X_n$ is ordered from smallest to largest. Further, let $\underline{X} = (X_1, \cdots, X_n)$. Now, we have

$$\widehat{F}(\underline{X}) = \left(0, \frac{1}{n-1}, \frac{2}{n-1}, \cdots, 1\right).$$

Moreover, because $\widehat{\eta}$ is invariant under transformations of location and scale, the measure is equivalent to the sample correlation between $\widehat{f}(\underline{X})$ and $\underline{U} = (1, 2, 3, \cdots, n)$. Figure 2.4 shows the four Normal mixtures NM1, NM2, NM3 and NM4, which are associated with values of η equal to $0.1, 0.2, 0.3$ and 0.4. Figure 2.5 demonstrates how the measure works by showing a scatter-plot of $\widehat{f}(\underline{X})$ against $\underline{U}$, along with a line of best fit. In this setting $\widehat{\eta}^2$ is equivalent to the coefficient of determination r^2, where r is the sample correlation coefficient. It is clear from Figure 2.5 that as the amount of asymmetry increases, r^2 increases. Furthermore, the direction of the asymmetry has a direct effect on the slope of the line of best fit. As a result, asymmetry to the right results in a negative gradient and hence, a positive value of $\widehat{\eta}$. On the other hand, asymmetry to the left results in a positive slope and hence, a negative value for $\widehat{\eta}$.

2.3.3 A new test statistic T_n and its asymptotic distribution

Since η has already been shown to be an effective measure of asymmetry it seems only natural to suggest using $\widehat{\eta}$ as test statistic for testing symmetry. Therefore we extend the work of PPB

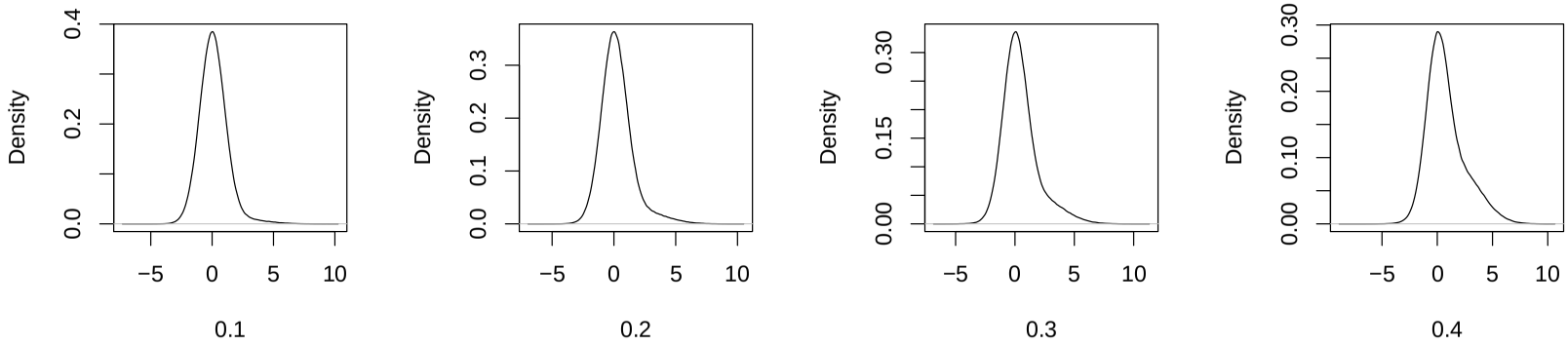

Figure 2.4: Normal mixture densities with $\eta = 0.1, 0.2, 0.3$ and 0.4.

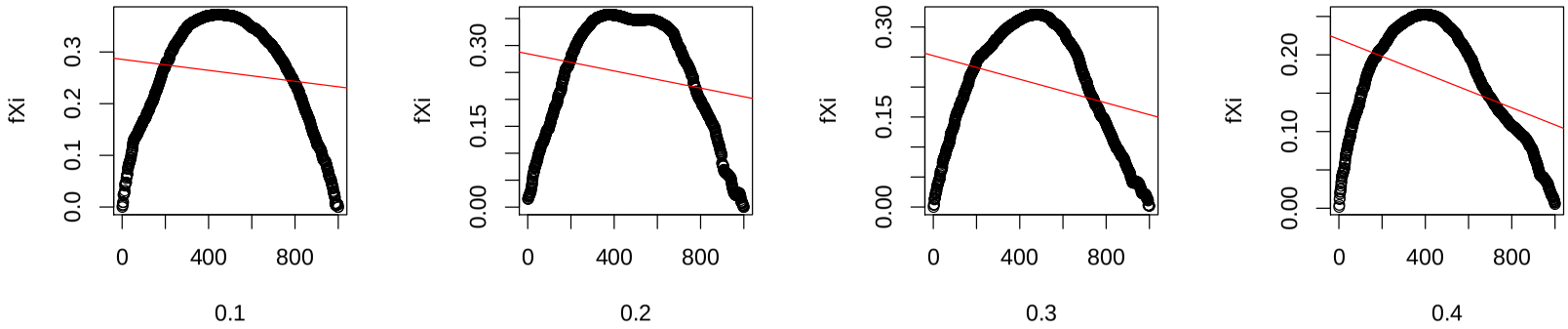

Figure 2.5: $\widehat{\text{Corr}}(\widehat{f}, \widehat{F})$ of a sample of 1000 observations from increasingly asymmetric densities as shown in Figure 2.4.

[88] to investigate the use of $\widehat{\eta}$ as a test statistic. For example, the standardised test statistic would be

$$T_n := \sqrt{n}\frac{\widehat{\eta}}{\sqrt{\widehat{\sigma}^2}},$$

where $\widehat{\sigma}^2$ is the estimate of the variance of $\sqrt{n}\widehat{\eta}$. Before we conduct a power analysis of T_n, we first derive the asymptotic distribution of $\widehat{\eta}$.

The asymptotic distribution of $\widehat{\eta}$ is established in Theorem 2.3 below. For that, we require the following assumptions:

A1 Assume that $\mathrm{E}[f^2(X)] < \infty$.

A2 The kernel function K is smooth, has bounded support and is of bounded variation.

A3 The bandwidth $h \sim n^{-\gamma}$ for $\frac{1}{4} \leq \gamma < \frac{1}{2}$.

Theorem 2.3 *Let $X_1, \cdots, X_n$ be a random sample from a continuous probability density function $f(x)$ and distribution function $F(x)$ and further suppose that assumptions* **A1**, **A2** *and* **A3** *all hold. Then as $n \to \infty$*

$$\sqrt{n}\,[\widehat{\eta} - \eta] \xrightarrow{L} N(0, \sigma^2),$$

where

$$\sigma^2 = \mathrm{Var}\left[\frac{2}{\sqrt{\nu_f \nu_F}}\left(f(X)F(X) - \frac{1}{2}f(X)\right) + \int_X^\infty \frac{f(y)^2}{\sqrt{\nu_f \nu_F}}dy \right.$$
$$\left. + \eta\left\{\frac{\left(F(X) - \frac{1}{2}\right)^2}{2\nu_F} + \frac{(f(X) - \mu_f)^2}{2\nu_f} + \int_X^\infty \frac{\left(F(y) - \frac{1}{2}\right)}{\nu_F} f(y)dy + \frac{(f(X) - \mu_f)f(X)}{\nu_f}\right\}\right], \tag{2.4}$$

where ν_f and ν_F denote $\mathrm{Var}(f(X))$ *and* $\mathrm{Var}(F(X))$ $\left(= \frac{1}{12}\right)$ *respectively, and $\mu_f = \mathrm{E}[f(X)]$.*

Note that, although $\widehat{\eta}$ uses non-parametric density estimates, one is able to obtain a parametric convergence rate n^{-1}, similar to the estimation of integrated squared density derivatives

discussed by Hall and Marron [45]. Below we present a detailed proof of Theorem 2.3, which is based on determining the first order asymptotic linear expansion of $\widehat{\eta}$, before applying a general result given by Giné and Mason [38]. The proof also requires repeated application of Slutsky's lemma [115].

Proof. Recall that

$$\widehat{\eta} = -\widehat{\mathrm{Corr}}\left(\widehat{f}, \widehat{F}\right) = -\frac{\sum_i \left(\widehat{f}_i - \overline{\widehat{f}}\right)\left(\widehat{F}_i - \overline{\widehat{F}}\right)}{\sqrt{\sum_i \left(\widehat{f}_i - \overline{\widehat{f}}_i\right)^2}\sqrt{\sum_i \left(\widehat{F}_i - \overline{\widehat{F}}_i\right)^2}},$$

where $\widehat{f}_i = \widehat{f}(X_i)$ and $\widehat{F}_i = \widehat{F}(X_i)$. This is an estimate of the population correlation coefficient,

$$\eta = -\mathrm{Corr}\left(f(X), F(X)\right) = -\frac{\mathrm{E}[f(X)F(X)] - \mathrm{E}[f(X)]\mathrm{E}[F(X)]}{\sqrt{\mathrm{Var}[f(X)]\mathrm{Var}[F(X)]}}.$$

To ease the notation, let

$$\begin{aligned}
\nu_{fF} &= \mathrm{E}[f(X)F(X)] - \mathrm{E}[f(X)]\mathrm{E}[F(X)] \\
\widehat{\nu}_{fF} &= \frac{1}{n}\sum_i \left(\widehat{f}_i - \overline{\widehat{f}}\right)\left(\widehat{F}_i - \overline{\widehat{F}}\right) = \frac{1}{n}\sum_i \left(\widehat{f}_i\right)\left(\widehat{F}_i - \overline{\widehat{F}}\right) = \frac{1}{n}\sum_i \widehat{f}_i\left(\widehat{F}_i - \frac{1}{2}\right) \\
\nu_f &= \mathrm{Var}[f(X)] \\
\widehat{\nu}_f &= \frac{1}{n}\sum_i \left(\widehat{f}_i - \overline{\widehat{f}}\right)^2 \\
\nu_F &= \mathrm{Var}[F(X)] = \frac{1}{12} \\
\widehat{\nu}_F &= \frac{1}{n}\sum_i \left(\widehat{F}_i - \overline{\widehat{F}}\right)^2 = \frac{1}{n}\sum_i \left(\widehat{F}_i - \frac{1}{2}\right)^2.
\end{aligned}$$

Then,

$$\widehat{\eta} = -\frac{\widehat{\nu}_{fF}}{\sqrt{\widehat{\nu}_f \widehat{\nu}_F}} \quad \text{and} \quad \eta = -\frac{\nu_{fF}}{\sqrt{\nu_f \nu_F}}.$$

Firstly, observe

$$\begin{aligned}\sqrt{n}(\widehat{\eta}-\eta) &= -\sqrt{n}\left(\frac{\widehat{\nu}_{fF}}{\sqrt{\widehat{\nu}_f\widehat{\nu}_F}}-\frac{\nu_{fF}}{\sqrt{\nu_f\nu_F}}\right)\\ &= -\sqrt{n}\left(\frac{\widehat{\nu}_{fF}}{\sqrt{\widehat{\nu}_f\widehat{\nu}_F}}-\frac{\nu_{fF}}{\sqrt{\widehat{\nu}_f\widehat{\nu}_F}}+\frac{\nu_{fF}}{\sqrt{\widehat{\nu}_f\widehat{\nu}_F}}-\frac{\nu_{fF}}{\sqrt{\nu_f\nu_F}}\right)\\ &= -\sqrt{n}\frac{1}{\sqrt{\widehat{\nu}_f\widehat{\nu}_F}}\left(\widehat{\nu}_{fF}-\nu_{fF}\right)+\sqrt{n}\frac{\nu_{fF}}{\sqrt{\widehat{\nu}_f\widehat{\nu}_F}\sqrt{\nu_f\nu_F}}\left(\sqrt{\widehat{\nu}_f\widehat{\nu}_F}-\sqrt{\nu_f\nu_F}\right). \qquad (2.5)\end{aligned}$$

Ignoring the sign of the first term on the right handside of equation (2.5) rewrite

$$\begin{aligned}\sqrt{n}\frac{1}{\sqrt{\widehat{\nu}_f\widehat{\nu}_F}}\left(\widehat{\nu}_{fF}-\nu_{fF}\right) &= \sqrt{n}\frac{1}{\sqrt{\nu_f\nu_F}}\left(\widehat{\nu}_{fF}-\nu_{fF}\right)+\sqrt{n}\frac{1}{\sqrt{\widehat{\nu}_f\widehat{\nu}_F}}\left(\widehat{\nu}_{fF}-\nu_{fF}\right)\\ &\quad -\sqrt{n}\frac{1}{\sqrt{\nu_f\nu_F}}\left(\widehat{\nu}_{fF}-\nu_{fF}\right)\\ &= \sqrt{n}\frac{1}{\sqrt{\nu_f\nu_F}}\left(\widehat{\nu}_{fF}-\nu_{fF}\right)\\ &\quad +\sqrt{n}\frac{1}{\sqrt{\widehat{\nu}_f\widehat{\nu}_F}\sqrt{\nu_f\nu_F}}\left(\widehat{\nu}_{fF}-\nu_{fF}\right)\left(\sqrt{\nu_f\nu_F}-\sqrt{\widehat{\nu}_f\widehat{\nu}_F}\right).\end{aligned}$$

Claim 1 *$\sqrt{n}\left(\widehat{\nu}_{fF}-\nu_{fF}\right)$ converges in law to a Normal distribution with finite variance.*

We shall return to the matter of proving Claim 1 later. First, observe that from Hall and Marron [45] it follows that $\widehat{\nu}_f$ and $\widehat{\nu}_F$ converge in probability to ν_f and ν_F. Therefore, using this fact and Claim 1 we have

$$\sqrt{n}\frac{1}{\sqrt{\widehat{\nu}_f\widehat{\nu}_F}}\left(\widehat{\nu}_{fF}-\nu_{fF}\right)=\sqrt{n}\frac{1}{\sqrt{\nu_f\nu_F}}\left(\widehat{\nu}_{fF}-\nu_{fF}\right)+o_p(1). \qquad (2.6)$$

Write the second term in equation (2.5) as

$$\begin{aligned}
&\sqrt{n}\frac{\nu_{fF}}{\sqrt{\widehat{\nu}_f\widehat{\nu}_F}\sqrt{\nu_f\nu_F}}\left(\sqrt{\widehat{\nu}_f\widehat{\nu}_F}-\sqrt{\nu_f\nu_F}\right)\\
&\qquad=\sqrt{n}\frac{\nu_{fF}\left(\widehat{\nu}_f\widehat{\nu}_F-\nu_f\nu_F\right)}{\sqrt{\widehat{\nu}_f\widehat{\nu}_F}\sqrt{\nu_f\nu_F}\left(\sqrt{\widehat{\nu}_f\widehat{\nu}_F}+\sqrt{\nu_f\nu_F}\right)}\\
&\qquad=\sqrt{n}\nu_{fF}\frac{\widehat{\nu}_f\widehat{\nu}_F-\nu_f\nu_F}{2\nu_f\nu_F\sqrt{\nu_f\nu_F}}\\
&\qquad\qquad+\sqrt{n}\nu_{fF}\left(\widehat{\nu}_f\widehat{\nu}_F-\nu_f\nu_F\right)\frac{2\nu_f\nu_F-\sqrt{\widehat{\nu}_f\widehat{\nu}_F}\left(\sqrt{\widehat{\nu}_f\widehat{\nu}_F}+\sqrt{\nu_f\nu_F}\right)}{2\nu_f\nu_F\sqrt{\nu_f\nu_F}\sqrt{\widehat{\nu}_f\widehat{\nu}_F}\left(\sqrt{\widehat{\nu}_f\widehat{\nu}_F}+\sqrt{\nu_f\nu_F}\right)}\\
&\qquad=\sqrt{n}\nu_{fF}\frac{\widehat{\nu}_f\widehat{\nu}_F-\nu_f\nu_F}{2\nu_f\nu_F\sqrt{\nu_f\nu_F}}+o_p(1),
\end{aligned}\tag{2.7}$$

again, using the fact that $\widehat{\nu}_f$ and $\widehat{\nu}_F$ converge in probability to ν_f and ν_F. Furthermore,

$$\begin{aligned}
\sqrt{n}\left(\widehat{\nu}_f\widehat{\nu}_F-\nu_f\nu_F\right)&=\sqrt{n}\left(\nu_f(\widehat{\nu}_F-\nu_F)+\nu_F(\widehat{\nu}_f-\nu_f)\right)+\sqrt{n}\left(\widehat{\nu}_f-\nu_f\right)\left(\widehat{\nu}_F-\nu_F\right)\\
&=\sqrt{n}\left(\nu_f(\widehat{\nu}_F-\nu_F)+\nu_F(\widehat{\nu}_f-\nu_f)\right)+o_p(1).
\end{aligned}\tag{2.8}$$

Hence using equations (2.6), (2.7), and (2.8), rewrite (2.5) as

$$\begin{aligned}
\sqrt{n}(\widehat{\eta}-\eta)&=-\sqrt{n}\left(\frac{1}{\sqrt{\nu_f\nu_F}}\left(\widehat{\nu}_{fF}-\nu_{fF}\right)-\frac{\nu_{fF}}{2\nu_f\nu_F\sqrt{\nu_f\nu_F}}\left(\nu_f(\widehat{\nu}_F-\nu_F)+\nu_F(\widehat{\nu}_f-\nu_f)\right)\right)+o_p(1)\\
&=-\sqrt{n}\left(\frac{1}{\sqrt{\nu_f\nu_F}}\left(\widehat{\nu}_{fF}-\nu_{fF}\right)-\frac{\nu_{fF}}{2\nu_F\sqrt{\nu_f\nu_F}}\left(\widehat{\nu}_F-\nu_F\right)-\frac{\nu_{fF}}{2\nu_f\sqrt{\nu_f\nu_F}}\left(\widehat{\nu}_f-\nu_f\right)\right)\\
&\qquad+o_p(1).
\end{aligned}$$

The leading order term in the expansion for $\sqrt{n}(\widehat{\eta}-\eta)$ is composed of three parts. We show that the linear combination is asymptotically Normal by applying Theorem 1 given by Giné and Mason [38]. The statement and assumptions of this Theorem are given in detail in Appendix A. Firstly, observe that

$$\begin{aligned}
\sqrt{n}\widehat{\nu}_f&=\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{f}_i-\overline{\widehat{f}}\right)^2\\
&=\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{f}_i-\mu_f\right)^2+o_p(1),
\end{aligned}$$

where $\mu_f = \mathrm{E}[f(X)]$. Hence, defining

$$\tilde{\nu}_f = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{f}_i - \mu_f\right)^2,$$

it is clear that

$$\begin{aligned}\sqrt{n}(\widehat{\eta}-\eta) &= -\sqrt{n}\left(\frac{1}{\sqrt{\nu_f\nu_F}}\left(\widehat{\nu}_{fF}-\nu_{fF}\right) - \frac{\nu_{fF}}{2\nu_F\sqrt{\nu_f\nu_F}}\left(\widehat{\nu}_F-\nu_F\right) - \frac{\nu_{fF}}{2\nu_f\sqrt{\nu_f\nu_F}}\left(\tilde{\nu}_f-\nu_f\right)\right) + o_p(1)\\ &= -\sqrt{n}\widehat{\Theta} + o_p(1),\end{aligned}$$

where

$$\widehat{\Theta} = \frac{1}{\sqrt{\nu_f\nu_F}}\left(\widehat{\nu}_{fF}\right) - \frac{\nu_{fF}}{2\nu_F\sqrt{\nu_f\nu_F}}\left(\widehat{\nu}_F\right) - \frac{\nu_{fF}}{2\nu_f\sqrt{\nu_f\nu_F}}\left(\tilde{\nu}_f\right).$$

Claim 2 $\sqrt{n}\widehat{\Theta} \xrightarrow{L} N(0,\sigma^2)$.

The proof of the Theorem will be complete if we prove Claim 1 and Claim 2. Since the proof of Claim 1 and 2 are similar we prove Claim 2, whilst Claim 1 follows similarly. Observe that $\widehat{\Theta}$ is in the form

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{\phi}\left(\widehat{f}(X_i),\widehat{F}(X_i)\right),$$

and is an estimator of

$$\begin{aligned}\Theta &= \int_{-\infty}^{\infty}\left\{\frac{f(x)F(x)-\frac{1}{2}f(x)}{\sqrt{\nu_f\nu_F}} - \frac{\nu_{fF}}{2\nu_F\sqrt{\nu_f\nu_F}}\left(F(x)-\frac{1}{2}\right)^2 - \frac{\nu_{fF}}{2\nu_f\sqrt{\nu_f\nu_F}}(f(x)-\mu_f)^2\right\}f(x)dx\\ &= 0.\end{aligned}$$

Therefore we can apply Theorem 1 of Giné and Mason [38] to show that $\widehat{\Theta}$ is asymptotically Normal once we have verified the conditions *I-VIII* of the theorem. Details of these conditions can be found in Appendix A. Conditions *I, VI, VII* and *VIII* hold directly from the assumptions **A1**, **A2** and **A3**. Also, under assumption **A3** condition *II* holds with $H = f$ as the suitable measurable function. To verify the remaining conditions note that, in Giné and Mason's notation, we have that

$$\psi(x, F(x), f(x)) = \frac{f(x)F(x) - \frac{1}{2}f(x)}{\sqrt{\nu_f \nu_F}} - \frac{\nu_{fF}}{2\nu_F\sqrt{\nu_f\nu_F}}\left(F(x) - \frac{1}{2}\right)^2 - \frac{\nu_{fF}}{2\nu_f\sqrt{\nu_f\nu_F}}(f(x) - \mu_f)^2$$
$$\psi(x, y_0, y_1) = \frac{y_0 y_1 - \frac{1}{2}y_1}{\sqrt{\nu_f \nu_F}} - \frac{\nu_{fF}}{2\nu_F\sqrt{\nu_f\nu_F}}\left(y_0 - \frac{1}{2}\right)^2 - \frac{\nu_{fF}}{2\nu_f\sqrt{\nu_f\nu_F}}(y_1 - \mu_f)^2.$$

Further,

$$\psi_m(x) = \frac{\partial}{\partial y_m}\psi(x, y_0, y_1)\big|_{(x,F(x),f(x))}.$$

Hence,

$$\begin{aligned}
\psi_0(x) = \frac{\partial}{\partial y_0}\psi(x, y_0, y_1) &= \frac{y_1}{\sqrt{\nu_f\nu_F}} - \frac{\nu_{fF}}{\nu_F\sqrt{\nu_f\nu_F}}\left(y_0 - \frac{1}{2}\right) \\
&= \frac{f(x)}{\sqrt{\nu_f\nu_F}} - \frac{\nu_{fF}}{\nu_F\sqrt{\nu_f\nu_F}}\left(F(x) - \frac{1}{2}\right), \\
\psi_1(x) = \frac{\partial}{\partial y_1}\psi(x, y_0, y_1) &= \frac{y_0 - \frac{1}{2}}{\sqrt{\nu_f\nu_F}} - \frac{\nu_{fF}}{\nu_f\sqrt{\nu_f\nu_F}}(y_1 - \mu_f) \\
&= \frac{F(x) - \frac{1}{2}}{\sqrt{\nu_f\nu_F}} - \frac{\nu_{fF}}{\nu_f\sqrt{\nu_f\nu_F}}(f(x) - \mu_f).
\end{aligned}$$

Therefore *III* and *IV* hold under the assumption **A1**. Further, define

$$\begin{aligned}
\xi(X_i) &= \psi(X_i) - \mathrm{E}\left[\psi(X)\right] \\
&= \frac{f(X)F(X) - \frac{1}{2}f(X)}{\sqrt{\nu_f\nu_F}} - \frac{\nu_{fF}}{2\nu_F\sqrt{\nu_f\nu_F}}\left(F(X) - \frac{1}{2}\right)^2 - \frac{\nu_{fF}}{2\nu_f\sqrt{\nu_f\nu_F}}(f(X) - \mu_f)^2, \\
\xi_0(X_i) &= \int_{X_i}^{\infty}\left\{\frac{f(y)}{\sqrt{\nu_f\nu_F}} - \frac{\nu_{fF}}{\nu_F\sqrt{\nu_f\nu_F}}\left(F(y) - \frac{1}{2}\right)\right\}f(y)dy - \frac{1}{2}\frac{\mu_f}{\sqrt{\nu_f\nu_F}}, \\
\chi_1(y) &= \psi_1(y)f(y) \\
&= \frac{F(y)f(y) - \frac{1}{2}f(y)}{\sqrt{\nu_f\nu_F}} - \frac{\nu_{fF}}{\nu_f\sqrt{\nu_f\nu_F}}(f(y) - \mu_f)f(y).
\end{aligned}$$

Hence, condition *V* is also satisfied.

$$\begin{aligned}
\xi_1(X_i) &= \chi_1(X_i) - \mathrm{E}\left[\chi_1(X)\right] \\
&= \frac{F(X)f(X) - \frac{1}{2}f(X)}{\sqrt{\nu_f\nu_F}} - \frac{\nu_{fF}}{\nu_f\sqrt{\nu_f\nu_F}}(f(X) - \mu_f)f(X).
\end{aligned}$$

Finally, define

$$
\begin{aligned}
Y &= \xi(X) + \xi_0(X) + \xi_1(X) \\
&= \frac{f(X)F(X) - \frac{1}{2}f(X)}{\sqrt{\nu_f \nu_F}} - \frac{\nu_{fF}}{2\nu_F\sqrt{\nu_f\nu_F}}\left(F(X) - \frac{1}{2}\right)^2 - \frac{\nu_{fF}}{2\nu_f\sqrt{\nu_f\nu_F}}(f(X) - \mu_f)^2 \\
&\quad + \int_X^\infty \left\{ \frac{f(y)}{\sqrt{\nu_f\nu_F}} - \frac{\nu_{fF}}{\nu_F\sqrt{\nu_f\nu_F}}\left(F(y) - \frac{1}{2}\right)\right\} f(y)dy - \frac{1}{2}\frac{\mu_f}{\sqrt{\nu_f\nu_F}} \\
&\quad + \frac{F(X)f(X) - \frac{1}{2}f(X)}{\sqrt{\nu_f\nu_F}} - \frac{\nu_{fF}}{\nu_f\sqrt{\nu_f\nu_F}}(f(X) - \mu_f)f(X).
\end{aligned}
$$

Hence, we conclude under the assumptions **A1**-**A3**, that $\sqrt{n}\left(\widehat{\Theta} - \Theta\right)$ is asymptotically normally distributed with mean zero and variance

$$
\begin{aligned}
\sigma^2 &:= \mathrm{Var}(Y) = \mathrm{Var}\left[\frac{2}{\sqrt{\nu_f\nu_F}}\left(f(X)F(X) - \frac{1}{2}f(X)\right) + \int_X^\infty \frac{f(y)^2}{\sqrt{\nu_f\nu_F}}dy - \frac{1}{2}\frac{\mu_f}{\sqrt{\nu_f\nu_F}}\right. \\
&\quad \left. - \nu_{fF}\left\{\frac{\left(F(X) - \frac{1}{2}\right)^2}{2\nu_F\sqrt{\nu_f\nu_F}} + \frac{(f(X) - \mu_f)^2}{2\nu_f\sqrt{\nu_f\nu_F}} + \int_X^\infty \frac{\left(F(y) - \frac{1}{2}\right)}{\nu_F\sqrt{\nu_f\nu_F}}f(y)dy + \frac{(f(X) - \mu_f)f(X)}{\nu_f\sqrt{\nu_f\nu_F}}\right\}\right] \\
&= \mathrm{Var}\left[\frac{2}{\sqrt{\nu_f\nu_F}}\left(f(X)F(X) - \frac{1}{2}f(X)\right) + \int_X^\infty \frac{f(y)^2}{\sqrt{\nu_f\nu_F}}dy\right. \\
&\quad \left. - \frac{\nu_{fF}}{\sqrt{\nu_f\nu_F}}\left\{\frac{\left(F(X) - \frac{1}{2}\right)^2}{2\nu_F} + \frac{(f(X) - \mu_f)^2}{2\nu_f} + \int_X^\infty \frac{\left(F(y) - \frac{1}{2}\right)}{\nu_F}f(y)dy + \frac{(f(X) - \mu_f)f(X)}{\nu_f}\right\}\right] \\
&= \mathrm{Var}\left[\frac{2}{\sqrt{\nu_f\nu_F}}\left(f(X)F(X) - \frac{1}{2}f(X)\right) + \int_X^\infty \frac{f(y)^2}{\sqrt{\nu_f\nu_F}}dy\right. \\
&\quad \left. + \eta\left\{\frac{\left(F(X) - \frac{1}{2}\right)^2}{2\nu_F} + \frac{(f(X) - \mu_f)^2}{2\nu_f} + \int_X^\infty \frac{\left(F(y) - \frac{1}{2}\right)}{\nu_F}f(y)dy + \frac{(f(X) - \mu_f)f(X)}{\nu_f}\right\}\right].
\end{aligned}
$$

□

Therefore, in conclusion, we have shown that the test statistic,

$$T_n := \sqrt{n}\frac{\widehat{\eta}}{\sqrt{\widehat{\sigma}^2}},$$

is approximately standard Normal, where $\widehat{\sigma}^2$ is an appropriate estimate of the variance σ^2. The expression for the variance is generally a complicated one to evaluate in practice, but it simplifies a great deal under the null hypothesis ($\eta = 0$). In the next section we discuss several procedures to estimate σ^2.

2.3.4 Estimating the variance σ^2

We now provide details of how to estimate the variance that appears in the test statistic T_n. Recall,

$$\sigma^2 = \text{Var}\left[\frac{2}{\sqrt{\nu_f \nu_F}}\left(f(X)F(X) - \frac{1}{2}f(X)\right) + \int_X^\infty \frac{f(y)^2}{\sqrt{\nu_f \nu_F}}dy \right.$$
$$\left. + \eta\left\{\frac{\left(F(X) - \frac{1}{2}\right)^2}{2\nu_F} + \frac{(f(X) - \mu_f)^2}{2\nu_f} + \int_X^\infty \frac{\left(F(y) - \frac{1}{2}\right)}{\nu_F} f(y)dy + \frac{(f(X) - \mu_f)f(X)}{\nu_f}\right\}\right],$$

One approach is to estimate using the Monte Carlo method replacing f and F with $\widehat{f}$ and $\widehat{F}$ respectively. Using this method the integrals in the variance expression can be calculated using some numerical method.

One can also simplify matters by computing the null variance (i.e. assuming that the random variable X is symmetric) where $\eta = 0$. In this case σ^2 reduces to

$$\sigma_0^2 = \frac{1}{\nu_f \nu_F}\text{Var}\left(2\left\{f(X)F(X) - \frac{1}{2}f(X)\right\} + \int_X^\infty f^2(y)dy\right).$$

To ease the notation somewhat let

$$\Phi_1(X) = \int_X^\infty f^2(y)dy.$$

Hence, for symmetric random variables we have

$$\begin{aligned}
\sigma_0^2 &= \frac{1}{\nu_f \nu_F} \mathrm{Var}\left[2\left\{f(X)F(X) - \frac{1}{2}f(X)\right\} + \Phi_1(X)\right] \\
&= \frac{1}{\nu_f \nu_F}\left(\mathrm{Var}\left[2\left\{f(X)F(X) - \frac{1}{2}f(X)\right\}\right] + \mathrm{Var}\left[\Phi_1(X)\right]\right. \\
&\qquad \left. + 2\mathrm{Cov}\left(2\left\{f(X)F(X) - \frac{1}{2}f(X)\right\}, \Phi_1(X)\right)\right) \\
&= \frac{1}{\nu_f \nu_F}\left(\mathrm{E}\left[4\left\{f(X)F(X) - \frac{1}{2}f(X)\right\}^2\right] + \mathrm{Var}\left[\Phi_1(X)\right]\right. \\
&\qquad \left. + 4\mathrm{E}\left[\left\{f(X)F(X) - \frac{1}{2}f(X)\right\}\Phi_1(X)\right]\right) \\
&= \frac{1}{\nu_f \nu_F}\Big(4\mathrm{E}\left[f(X)^2F(X)^2\right] + \mathrm{E}\left[f(X)^2\right] - 4\mathrm{E}\left[f(X)^2F(X)\right] + \mathrm{E}\left[\Phi_1(X)^2\right] - \left[\mathrm{E}\Phi_1(X)\right]^2 \\
&\qquad + 4\mathrm{E}\left[f(X)F(X)\Phi_1(X)\right] - 2\mathrm{E}\left[f(X)\Phi_1(X)\right]\Big).
\end{aligned}$$

Further, observe that by a simple change of variables

$$\begin{aligned}
\mathrm{E}\left[\Phi_1(X)\right] &= \mathrm{E}\left[\int_{u=X}^{\infty} f^2(u)du\right] \\
&= \int_{y=-\infty}^{\infty} f(y)\int_{u=y}^{\infty} f^2(u)dudy \\
&= \int_{u=-\infty}^{\infty} f^2(u)\int_{y=-\infty}^{u} f(y)dydu \\
&= \int_{u=-\infty}^{\infty} f^2(u)F(u)du \\
&= \mathrm{E}\left[f(X)F(X)\right],
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\left[f(X)\Phi_1(X)\right] &= \mathrm{E}\left[f(X)\int_{u=X}^{\infty} f^2(u)du\right] \\
&= \int_{y=-\infty}^{\infty} f^2(y)\int_{u=y}^{\infty} f^2(u)dudy \\
&= \int_{u=-\infty}^{\infty} f^2(u)\int_{y=-\infty}^{u} f^2(y)dydu \\
&= \int_{u=-\infty}^{\infty} f^2(u)\left\{\int_{y=-\infty}^{\infty} f^2(y)dy - \int_{y=u}^{\infty} f^2(y)dy\right\}du \\
&= \int_{u=-\infty}^{\infty} f^2(u)\left\{\mathrm{E}[f(X)] - \int_{y=u}^{\infty} f^2(y)dy\right\}du \\
&= \mathrm{E}[f(X)]^2 - \mathrm{E}\left[f(X)\Phi_1(X)\right].
\end{aligned}$$

Thus, under the null hypothesis

$$\begin{aligned}
\mathrm{E}\left[\Phi_1(X)\right] &= \frac{1}{2}\mathrm{E}\left[f(X)\right] = \mathrm{E}\left[f(X)F(X)\right], \\
2\mathrm{E}\left[f(X)\Phi_1(X)\right] &= \left[\mathrm{E}f(X)\right]^2 = 4\left[\mathrm{E}f(X)F(X)\right]^2.
\end{aligned}$$

Hence,

$$\sigma_0^2 = \frac{1}{\nu_f\nu_F}\left\{4m_{22} + m_{20} - 4m_{21}\right\} + 4\left\{\mathrm{E}\left[f(X)F(X)\Phi_1(X)\right] - m_{11}^2\right\} + \left\{\mathrm{E}[\Phi_1(X)^2] - m_{11}^2\right\},$$

where

$$m_{ij} = \mathrm{E}\left[f(X)^i F(X)^j\right].$$

For a random sample we can readily estimate m_{ij} using

$$\widehat{m}_{ij} = \frac{1}{n}\sum_{k=1}^{n} \widehat{f}(X_k)^i \widehat{F}(X_k)^j.$$

Even in this greatly reduced form, the presence of the terms involving $\Phi_1(X)$ means that the expression for the variance is a complex one to evaluate in practice. In general, one could carry out a numerical integration technique using the estimated density function $\widehat{f}(x)$ in place of $f(x)$. Alternatively, in most situations one is primarily interested in whether samples are taken from a Normal population. If we add the additional assumption (under the null hypothesis) that X

is a normally distributed random variable we obtain

$$\int_x^\infty f^2(y)dy = \frac{1}{4\sqrt{\pi}\sigma_X}\text{cerf}\left(\frac{x-\mu_X}{\sigma_X}\right),$$

where μ_X and σ_X are the mean and variance of the random variable X respectively and

$$\text{cerf}(z) = \frac{2}{\sqrt{\pi}}\int_z^\infty e^{-t^2}dt,$$

is the complementary error function. This allows for an additional simplification to the null variance and removes the need to carry out a potentially computationally expensive numerical integration technique. More generally, one can define

$$Y_i = \left[\frac{2}{\sqrt{\widehat{\nu}_f\widehat{\nu}_F}}\left(\widehat{f}(X_i)\widehat{F}(X_i) - \frac{1}{2}f(X_i)\right) + \frac{1}{\sqrt{\widehat{\nu}_f\widehat{\nu}_F}}\widehat{\Phi}_1(X_i) + \widehat{\eta}\left\{\frac{\left(\widehat{F}(X_i)-\frac{1}{2}\right)^2}{2\widehat{\nu}_F} + \frac{\left(\widehat{f}(X_i)-\overline{\widehat{f}}\right)^2}{2\widehat{\nu}_f} + \widehat{\Phi}_2(X_i) + \frac{\left(\widehat{f}(X_i)-\overline{\widehat{f}}\right)\widehat{f}(X_i)}{\widehat{\nu}_f}\right\}\right],$$

where $\widehat{\Phi}_1(x)$ is a numerical approximation of the integral $\Phi_1(x)$ estimating f with $\widehat{f}$, and $\widehat{\Phi}_2(x)$ is a numerical approximation of

$$\int_x^\infty \frac{\left(F(y)-\frac{1}{2}\right)}{\nu_F}f(y)dy,$$

estimating f and F by $\widehat{f}$ and $\widehat{F}$. It is then possible to estimate σ^2 using

$$\widehat{\sigma}^2 = \widehat{\text{Var}}(Y) = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y})^2. \tag{2.9}$$

Hence, in summary, for large n the distribution of

$$T_n = \sqrt{n}\frac{\widehat{\eta}}{\sqrt{\widehat{\sigma}^2}},$$

is approximately standard Normal, where, $\widehat{\sigma}^2$ is an appropriate estimate of

$$\sigma^2 = \text{Var}\left[\frac{2}{\sqrt{\nu_f \nu_F}}\left(f(X)F(X) - \frac{1}{2}f(X)\right) + \int_X^\infty \frac{f(y)^2}{\sqrt{\nu_f \nu_F}}dy \right.$$
$$\left. + \eta\left\{\frac{\left(F(X) - \frac{1}{2}\right)^2}{2\nu_F} + \frac{(f(X) - \mu_f)^2}{2\nu_f} + \int_X^\infty \frac{\left(F(y) - \frac{1}{2}\right)}{\nu_F} f(y)dy + \frac{(f(X) - \mu_f)f(X)}{\nu_f}\right\}\right].$$

In the next section we analyse the empirical power of T_n through the use of a simulation study.

2.3.5 Power analysis of T_n

Methods

We subject the new test statistic to a similar simulation study as detailed in section 2.2. Once again, the test statistic is generated $m = 10,000$ times from samples of size $n = 30, 50$ and 70. In each case the density and distribution function is estimated using a Normal kernel and the bandwidth is estimated using the simple rule of thumb given by Silverman [114]. The asymptotic variance σ^2 is estimated from the sample using $\widehat{\sigma}^2$ defined in equation (2.9), where the numerical integration of Φ_1 and Φ_2 is carried out using a simple rectangle approximation with a grid of 512 points. Therefore T_n is asymptotically standard Normal and, thus, we reject the null hypothesis at level 0.05 if the absolute value of the test statistic is greater than $Z_{0.975} \approx 1.96$. As before, the empirical powers (the proportion of rejections) of each of the tests are reported. Table 2.7 summarises the simulation procedure.

Step 1	Simulate samples of size $n = 30, 50$ and 70 from each of the distributions.
Step 2	Calculate the test statistic T_n based on these samples, estimating σ^2 using equation (2.9).
Step 3	If $\lvert T_n \rvert > Z_{0.975}$ then the null hypothesis of symmetry is rejected at the 5% level.
Step 4	Repeat **Steps 1-3** $10,000$ times and report the proportion of rejections (empirical power).

Table 2.7: Step by step guide to the simulation study for the newly proposed test T_n.

η	Dist.	T_n		
		$n = 30$	$n = 50$	$n = 70$
0	N	0.040	0.037	0.038
0	C	0.037	0.036	0.035
0.1	NM1	0.049	0.055	0.059
0.2	NM2	0.075	0.113	0.147
0.3	NM3	0.128	0.229	0.321
0.4	NM4	0.202	0.392	0.573
0.1	SN1	0.046	0.054	0.060
0.2	SN2	0.068	0.100	0.127
0.3	SN3	0.105	0.179	0.261
0.4	SN4	0.167	0.296	0.425
0.5	SN5	0.251	0.446	0.591
0.6	SN6	0.359	0.574	0.737
0.7	SN7	0.443	0.681	0.841
0.8	SN8	0.506	0.757	0.883
0.1	SAS1	0.044	0.050	0.061
0.2	SAS2	0.069	0.096	0.130
0.3	SAS3	0.109	0.186	0.257
0.4	SAS4	0.170	0.311	0.436
0.5	SAS5	0.270	0.477	0.655
0.6	SAS6	0.408	0.671	0.833
0.7	SAS7	0.581	0.838	0.948
0.8	SAS8	0.738	0.936	0.987
0.1	FAS1	0.047	0.046	0.063
0.2	FAS2	0.070	0.098	0.128
0.3	FAS3	0.103	0.172	0.247
0.4	FAS4	0.170	0.281	0.402
0.5	FAS5	0.242	0.405	0.556
0.6	FAS6	0.326	0.541	0.687
0.7	FAS7	0.409	0.645	0.794
0.8	FAS8	0.482	0.729	0.866
0.91	LN	0.908	0.991	0.999
0.95	FN	0.550	0.790	0.910
1	EXP	0.866	0.981	0.998

Table 2.8: Empirical power of the test based on T_n for a variety of density functions and sample sizes.

Results

The results of the simulation are detailed in Table 2.8. For the test based on T_n we observe a steady increase in power for the asymmetric families, whilst for the remaining asymmetric distributions (Log-Normal, Folded Normal, Exponential) the test achieves a high level of power. Furthermore, whilst not perfect, the amount of power is more closely related to the amount of asymmetry. This is to be expected since the tests are based on η, which has previously been identified as a more effective measure of the magnitude of asymmetry. For example, for $n = 70$ the empirical power of T_n for the Log-Normal and Folded Normal distribution is 0.999 and 0.910 respectively. Comparing this with the respective power of the existing tests, reported in Table 2.4 and Table 2.5, it is apparent that T_n is considerably more powerful than S_1 (0.976 and 0.552), S_2 (0.704 and 0.309), S_4 (0.392 and 0.792) and S_5 (0.831 and 0.454), and is comparable to S_3 (1.000 and 0.936).

In fact, T_n consistently out performs the existing tests S_1, S_2, S_4 and S_5 in terms of power. Further, whilst S_3 has marginally higher power than T_n, recall that S_3 is not conservative. Indeed, for a Normal sample, T_n has an estimated type-I error of 0.040 for $n = 30$ compared to 0.082 for S_3. For a sample of $n = 30$ from the Cauchy distribution this difference is even more stark with a type-I error estimate of 0.037 for T_n compared to 0.108 for S_3. Hence, the additional power achieved by S_3 is somewhat artificial if the test is unable to maintain the level α under the null hypothesis.

In the next section we provide a short worked example to demonstrate the usefulness of T_n when applied to a real data set.

2.4 Real data example

We now evaluate the usefulness of our test in a more practical setting by applying it to a real data set. We consider a reasonably large set of data collected for a randomised trial investigating hypertension (one of several trials collated by Wang et al. [125]), which we will introduce in the next chapter. For example, we consider the EWPH trial, which consists of 172 patients [2]. The main focus of this study was to investigate the effect of an antihypertensive treatment versus

placebo at reducing mortality rates in elderly patients (above the age of 60). However, a number of covariates were also collected. For example, every patient had their age recorded and 164 had their body mass index (BMI) recorded. After pooling the treatment and control groups together the BMI data appears predominantly symmetric. In contrast, the age data appears significantly skewed to the right. This is unsurprising, since this specific trial focused on elderly patients and therefore we observe an abrupt cut-off in age at 60 on the left-hand tail, with a few outliers above 90 in the right-hand tail. This can be seen visually in Figure 2.6, which shows the density estimate of the BMI and age data. However, testing for symmetry will allow us to make an objective judgement with regards to the asymmetry inherent in the data.

Suppose that our principle interest is to test whether the BMI data are being drawn from a symmetric population. Pooling the treatment and control groups let $X_1, X_2, \cdots, X_{164}$ denote the BMI readings, and let η_b denote the value η for the BMI in this population. The null hypothesis is

$$H_0 : \eta_b = 0,$$

which is indicative of symmetry in the BMI readings, whilst the alternative hypothesis is

$$H_1 : \eta_b \neq 0,$$

which implies asymmetry. It is readily calculated using equation (2.2) that the estimate of η_b for the BMI data is

$$\widehat{\eta}_b = 0.077.$$

Recall that a value of $\eta = 0$ indicates symmetry, thus this small value suggests that the BMI data are being drawn from a near-symmetric population. This supports the visual evidence in the left-hand plot of Figure 2.6, which shows the density estimate of the data.

Using equation (2.9) we estimate the variance of $\sqrt{n_b}\widehat{\eta}_b$ to be

$$\widehat{\sigma}_b^2 = 3.41,$$

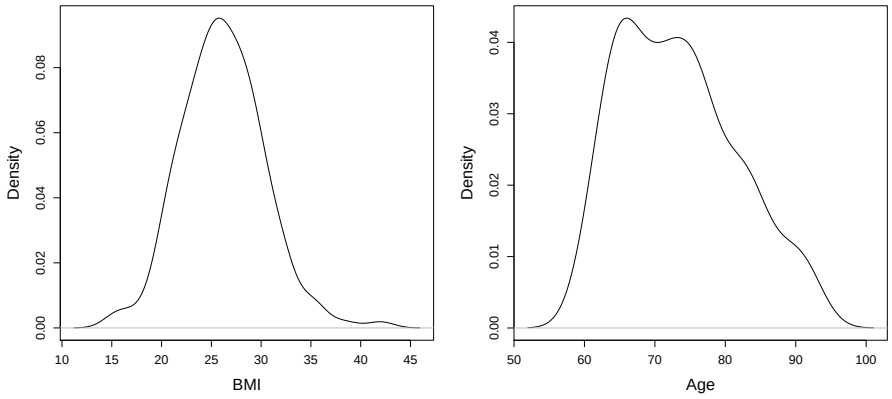

Figure 2.6: Density estimate of the BMI and age data in the EWPH trial.

where $n_b = 164$. Recall, by Theorem 2.3 we have

$$\sqrt{n}\left(\widehat{\eta} - \eta\right) \xrightarrow{\mathrm{L}} N\left(0, \sigma^2\right).$$

Hence,

$$\widehat{\eta} \sim N\left(\eta, \frac{\sigma^2}{n}\right),$$

approximately. Thus, under the null hypothesis we have

$$\widehat{\eta}_b \sim N\left(0, \frac{\sigma_b^2}{n_b}\right),$$

approximately. From this it is readily determined that the probability (p-value) of obtaining a value of $\widehat{\eta}_b$ at least as large as we did (0.077) under the null hypothesis is 0.30. Hence, as one should expect, there is not significant evidence to reject the null hypothesis of symmetry at the 5% level.

We can also obtain a confidence interval by observing that

$$P\left[Z_{-\frac{\alpha}{2}} < \sqrt{n}\frac{\eta - \widehat{\eta}}{\sigma} < Z_{\frac{\alpha}{2}}\right] \approx 1 - \alpha.$$

Therefore, the approximate $100(1-\alpha)\%$ confidence interval for η is given by

$$\widehat{\eta} \pm Z_{\frac{\alpha}{2}} \frac{\widehat{\sigma}_b}{\sqrt{n_b}}.$$

The standard error of $\widehat{\eta}_b$ is

$$\text{s.e.}(\widehat{\eta}_b) = \frac{\widehat{\sigma}_b}{\sqrt{n_b}} = \frac{1.85}{\sqrt{164}} = 0.144.$$

Therefore, the 95% confidence interval of η_b is given by

$$(-0.206, 0.360)\,.$$

Observe that this interval contains the null value of 0, backing up the p-value conclusion that we cannot reject the null hypothesis of symmetry.

Now consider the problem of testing whether the age data are being drawn from a symmetric population. This time, let η_a denote the value η for the distribution of ages in this population. Once again, using equation (2.2), it is readily calculated that the estimate of η_a for the age data is

$$\widehat{\eta}_a = 0.594.$$

As expected, this is a much larger estimate of η_a, which is indicative of a more asymmetric population. The estimated variance of $\sqrt{n_a}\widehat{\eta}_a$ using (2.9) is given by

$$\widehat{\sigma}_a^2 = 4.25,$$

where $n_a = 172$. Under the null hypothesis of symmetry

$$\widehat{\eta}_a \sim N\left(0, \frac{\sigma_a^2}{n_a}\right),$$

approximately, which gives a p-value of < 0.0001 indicating there is considerable evidence against the null hypothesis of symmetry. Once again, this supports the visual evidence in the right-hand

plot of Figure 2.6, which shows the density estimate of the age data. It is clear that the age data shows moderate skewness to the right, which is confirmed by $\widehat{\eta}_a = 0.594$. Also, on this occasion we determine the 95% confidence interval for η_a to be

$$(0.286, 0.902),$$

indicating that there is substantial evidence that the age data are drawn from a population which is asymmetric to the right.

2.5 Discussion

In this chapter some of the existing tests of symmetry have been appraised and shown to perform well at detecting departure from symmetry. However, an undesirable feature was identified, which seems to have been overlooked in the existing tests of symmetry. Namely, that the tests failed to reject the symmetry hypothesis with greater power for the most visually asymmetric distributions. This trait was exhibited though visual inspection and a simulation study. The reason for this feature is principally because, until recently, there was no measure of asymmetry which adequately measured the size of asymmetry. However, a recently proposed measure η, which has been shown to effectively measure the size of asymmetry, was introduced and discussed. By considering sample estimates of η, a new test for symmetry was proposed. Furthermore, the asymptotic properties of this test were determined and the test was compared with the existing tests in a simulation study. Finally, we discussed a short worked example, applying the new test to set of real data.

In conclusion, it was shown that η provides a useful test for symmetry, improving on the other tests discussed in this chapter. While the test did not achieve the initial aim of always achieving greater power in the most asymmetric cases, it did improve upon the performance of the existing tests of symmetry. Indeed, for most cases, the test based on $\widehat{\eta}$ rejects most often for the most asymmetric distributions. Moreover, it was demonstrated that the test achieves a greater power than most of the existing tests, whist maintaining the set level of 0.05 under the null hypothesis. In fact, the only test to outperform the newly proposed test was the 'triples'

test proposed by Randles et al. [95] and this particular test failed to maintain the set level for the symmetric random variables.

Another advantage of the test based on $\widehat{\eta}$ is the flexibility afforded by the smoothing involved in the estimate of $f(x)$. For the simulation study in this chapter we have only considered using a simple kernel density estimate with the most basic data-based estimate of h, however, one can readily tailor the choice of the kernel function K and the bandwidth estimate h to the particular problem in question. For example, one possibility is to account for known boundaries in the support of $f(x)$ using a boundary modification to the kernel density estimate. Indeed, if the support of $f(x)$ does not consist of the whole real line then the usual kernel density function estimate (employed throughout this chapter) is biased near the boundary [98]. For example, the density estimates of the Log-Normal, Folded Normal, and Exponential distribution are all subject to this boundary bias in the region of zero. This bias is particularly serious for the Folded Normal and Exponential distributions, where the density function is not continuous at zero. This naturally leads to bias in the estimate $\widehat{\eta}$ and has consequences for the resulting test for symmetry. In fact, it is likely that this bias is the main cause of the lower than expected power in the Folded Normal and Exponential distributions. Provided that the location of the boundary is known, this boundary bias can be reduced through the use of boundary kernel density estimation [63]. Hence, appropriately used in the estimate of $f(x)$, this can lead to improvements in the estimate of $\widehat{\eta}$ and power of the test.

Up to this point we have only discussed estimating the variance of $\widehat{\eta}$ through the use of the derived asymptotic distribution, however, other methods can be employed to obtain an estimate of the variance. For example, bootstrapping can be used to make inferences, such as estimating the variance or generating confidence intervals, for almost any test statistic [31]. This will be explored in more detail in Chapter 4 when we investigate applying $\widehat{\eta}$ in smaller samples, where the asymptotic distribution may be inappropriate.

Moreover, recent research by Patil, Bagkavos and Wood [89] has given rise to a new stronger measure of asymmetry η_s, which is based on an analogous necessary and sufficient condition for symmetry. This stronger measure also has the property that $|\eta_s(X)| \geq |\eta(X)|$ and so a test

based on η_s could potentially improve upon the power of the test discussed here.

In summation, the test based on $\widehat{\eta}$ provides a valid alternative to the existing tests. In the next chapter we discuss the potential applications of the measure η in more detail, including applying $\widehat{\eta}$ to inform the analysis of randomised control trials.

> ***N.B.*** *The content of this chapter is the subject of a research paper by Partlett and Patil, which has been accepted for publication in the Annals of the Institute of Statistical Mathematics.*

Chapter 3

Applying $\widehat{\eta}$ to inform the analysis of randomised trials

3.1 Introduction

The results of Chapter 2 established how to use $\widehat{\eta}$ as a test statistic for testing symmetry. Indeed, when dealing with a reasonably large data set we are able to approximate the sampling distribution of $\widehat{\eta}$ as discussed in the previous chapter. This allows us to use the measure $\widehat{\eta}$ as a test statistic as well as a summary statistic that measures asymmetry. The focus of this chapter is to consider the application of $\widehat{\eta}$ in both contexts, for the analysis of randomised trials. This is illustrated by using the hypertension trials collated by Wang et al. [125], which we introduce in detail in section 3.2.

We discuss the possible applications of $\widehat{\eta}$ in three separate contexts. Firstly, we apply $\widehat{\eta}$ to test and measure for baseline imbalance between the treatment and control arms of a randomised control trial. This is an important assumption when carrying out analyses of randomised trials and, when violated, it may lead to misleading inferences regarding the treatment effect. Secondly, we apply $\widehat{\eta}$ to test for symmetry in the residuals of the analysis of covariance (ANCOVA) model, used to estimate a treatment effect in a trial. Symmetry is a simple prerequisite for normality and as such, any departure from symmetry indicates a departure from normality. Hence, we discuss the benefits of applying $\widehat{\eta}$ to validate normality assumptions alongside more conventional

methods, such as the QQ-plot [129]. Thirdly, we consider how $\widehat{\eta}$ can provide an aid for carrying out transformations of the covariate or response variable. For example, if the residuals do not appear to be coming from a Normal population, one reason could be that the covariate or response data itself is non-Normal. In this case, a possible remedy is to transform the response or covariate data to obtain a sample which conforms more closely to the normality assumption. We show that the coefficient $\widehat{\eta}$ has a role in identifying which variables to transform, objectively determining whether the transformation has been successful, and even providing an indication as to what sort of transformation may be most appropriate.

The chapter is outlined as follows. In section 3.2 we introduce randomised control trials, discuss how they are set up, and what they aim to achieve. We also introduce a collection of data comprising of ten randomised control trials investigating hypertension treatment, collated by Wang et al. [125]. In section 3.3 we use $\widehat{\eta}$ as an additional summary statistic to test the homogeneity of the treatment and control arms of the hypertension data. In section 3.4 we apply $\widehat{\eta}$ to test the symmetry of the residuals in several ANCOVA models. Section 3.5 examines the use of $\widehat{\eta}$ to influence and appraise the effect of variable transformation on the residual distribution in cases where the residuals show a significant departure from symmetry. In section 3.6 we briefly discuss some of the limitations of $\widehat{\eta}$ in this context, before offering some concluding remarks in section 3.7.

Aims of the chapter:

- Investigate the use of $\widehat{\eta}$ to aid in the analysis of randomised control trials.
- Demonstrate that $\widehat{\eta}$ can be a useful measure in all stages of the transformation of skewed data:
  - **Identifying** whether a transformation is necessary (and for which variables).
  - **Deciding** what type of transformation to apply.
  - **Determining** whether the transformation has be successful at reducing the amount of asymmetry.

- An in depth appraisal of $\widehat{\eta}$, including its advantages and limitations in this context.

3.2 Randomised control trials

A randomised control trial is very much the gold standard for clinical trials. Firstly, patients which are representative of the population at risk are recruited randomly into two or more groups. The separate groups are subject to identical conditions and undergo the same follow up, save for the treatment (or treatments) under consideration. The aim of randomisation is to ensure that the only difference between the groups is the intervention of interest. That is, for a sufficiently large sample size, randomisation eliminates the effect of confounding factors. The patients in the groups are followed up to determine whether there is a significant difference in the outcome of interest between the groups. For example, suppose we are interested in testing the effect of a new treatment on blood pressure. Firstly, we randomly allocate patients into two groups, one of which receives the new treatment and one of which receives a placebo. At the end of follow-up the patients have their blood pressure recorded and the mean blood pressure is compared between the two groups, to estimate the effect of the treatment.

Wang et al. [125] collated data from ten such randomised control trials. In the original paper the authors used individual patient data (IPD) to perform a quantitative overview of trials in hypertension to investigate the effect of lowering of systolic blood pressure (SBP) and diastolic blood pressure (DBP) on cardiovascular outcome. They selected randomised control trials that tested antihypertensive drugs against a placebo or no treatment. Information was collected about $28,581$ patients from 10 different studies. The study measured a range of different covariate factors, including age, sex, and body mass index (BMI). In the later chapters we consider this collection of data as a meta-analysis, however, here it is convenient for our illustrations to consider each of the trials individually.

3.3 Comparing the distribution of control and treatment samples

When analysing a randomised control trial, it is important that the control and treatment samples are representative of the same population and that they are similar in terms of baseline characteristics (apart from treatment) to minimise the effect of confounding factors. If there are baseline differences between the control and treatment groups then the estimated treatment effect may be misleading or biased. Indeed, it may simply be a reflection of baseline imbalance, rather than a genuine treatment effect. An advantage of η is that it is characteristic which can be readily compared between two populations. Hence, its estimate $\widehat{\eta}$ can be used to compare two samples.

For example, consider the effect of hypertensive treatments at lowering blood pressure. We can use $\widehat{\eta}$ to determine whether the initial blood pressures in the control and treatment groups are distributed similarly. Indeed, to evaluate the treatment effect the initial blood pressure data in the control and treatment samples should be drawn from the same population. As a result, we expect that that the amount of asymmetry will be the same (along with other features, such as the mean and variance). Moreover, because the two samples are independent and $\widehat{\eta}$ is approximately normally distributed, we can test whether η is the same in the two samples in a statistically rigorous framework.

Our null hypothesis is, for a given trial, that the initial blood pressure (either systolic or diastolic) in the treatment and control samples are both drawn out of the same population with density function f and cumulative distribution F. In this case the population has an unknown quantity of asymmetry measured by η. Let η_T and η_C denote the values of η in the treatment and control populations respectively.

Suppose that the treatment and control arms contain n_T and n_C individuals respectively. Under null, $\eta_T = \eta_C$ and one should expect that the estimate of η in the control sample $\widehat{\eta}_C$ and the estimate in the treatment sample $\widehat{\eta}_T$ will be very similar. Hence, we set up a test of

$$H_0 : \theta = \eta_T - \eta_C = 0,$$

against

$$H_1 : \theta \neq 0.$$

Under null, $\widehat{\eta}_T$ and $\widehat{\eta}_C$ are independent and approximately normally distributed with variances, $\frac{\sigma^2}{n_T}$ and $\frac{\sigma^2}{n_C}$ respectively where, using the results of the previous chapter,

$$\sigma^2 = \text{Var}\left[\frac{2}{\sqrt{\nu_f \nu_F}}\left(f(X)F(X) - \frac{1}{2}f(X)\right) + \int_X^\infty \frac{f(y)^2}{\sqrt{\nu_f \nu_F}}dy \right.$$
$$\left. + \eta\left\{\frac{\left(F(X) - \frac{1}{2}\right)^2}{2\nu_F} + \frac{(f(X) - \mu_f)^2}{2\nu_f} + \int_X^\infty \frac{\left(F(y) - \frac{1}{2}\right)}{\nu_F} f(y)dy + \frac{(f(X) - \mu_f)f(X)}{\nu_f}\right\}\right],$$

where ν_f and ν_F denote $\text{Var}(f(X))$ and $\text{Var}(F(X))\left(= \frac{1}{12}\right)$ respectively, and $\mu_f = \text{E}[f(X)]$. It is important to note that we are implicitly assuming that the variance of $\widehat{\eta}$ is equal in both groups (a stronger assumption is that $f(x)$ and $F(x)$ are the same in both groups), however, this assumption may be unnecessarily restrictive. Indeed, the variance of $\widehat{\theta}$ can be readily calculated when this assumption is relaxed. For this particular setting, and under the null hypothesis, we expect the two groups to be relatively homogeneous and so this assumption is plausible in this case. Indeed, additional calculations (not included here) show that the results are not sensitive to this assumption for this example.

Hence, we can conclude that

$$\widehat{\theta} = \widehat{\eta}_T - \widehat{\eta}_C \sim N\left(0, \left\{\frac{1}{n_T} + \frac{1}{n_C}\right\}\sigma^2\right),$$

approximately. We estimate the variance σ^2 using the pooled estimate,

$$\widehat{\sigma}^2 = \frac{(n_T - 1)\widehat{\sigma}_t^2 + (n_C - 1)\widehat{\sigma}_c^2}{n_T + n_C - 2},$$

where $\widehat{\sigma}_t^2$ and $\widehat{\sigma}_c^2$ are the estimates of σ^2 using equation (2.9) in the treatment and control groups respectively.

It is important to note that some authors warn against the testing for baseline imbalance,

because there is no guarantee that the study is sufficiently powered to detect a difference and, theoretically, randomisation should be adequate to avoid an imbalance anyway [109]. In any case, imbalance is better resolved by adjusting for potential confounders in the analysis. Here, we discuss tests for baseline imbalance to illustrate a potential application of $\widehat{\eta}$ which may sometimes be useful, even if it is not recommended for routine use.

Trial	Systolic			Diastolic		
	θ	p-value	n	θ	p-value	n
ANBP	0.008	0.455	1542	0.017	0.417	1542
COOP	−0.028	0.431	349	−0.084	0.310	349
EWPH	0.026	0.454	172	−0.060	0.422	172
HDFP	0.008	0.019	4797	0.016	0.338	4798
MRC1	−0.003	0.469	6991	−0.005	0.455	6991
MRC2	−0.049	0.295	2651	−0.026	0.285	2651
SHEP	−0.005	0.426	4736	−0.123	0.004	4736
STOP	0.001	0.498	268	−0.047	0.450	268
SYCH	−0.074	0.098	2391	0.045	0.196	2391
SYSE	0.057	0.082	4695	0.044	0.158	4695
Pooled	−0.018	0.080	28592	0.013	0.148	28593

Table 3.1: Testing for a difference in η for initial blood pressures between treatment and control groups ($\theta = \widehat{\eta}_T - \widehat{\eta}_C$).

Table 3.1 gives the results of the test of $\eta_C = \eta_T$ in each of the 10 hypertension trials. We see that in most of the cases there isn't a significant difference between η_C and η_T. There are a couple of exceptions, however, in most cases there are relatively low levels of significance. In two out of 20 tests we find a significant result at the 5% level, which is perhaps not surprising. Also, observe that these 'significant' results both occurred in trials with over 4000 individuals, where there is power to detect even a slight difference in the value of η.

This shows that there is very little difference in the amount of asymmetry between the treatment and control groups, which suggests there is homogeneity between the two arms. Figures 3.1 - 3.10 show the density estimates of the systolic and diastolic blood pressure data in the treatment and control data. Note that there appears to be substantial asymmetry in the majority of studies, and also a reasonable amount of variability between studies. For example, the diastolic blood pressure in HDFP trial (in both groups) displays a reasonable amount of asymmetry to the right, whereas the same diastolic blood pressure in the MRC2 trial displays a

large degree of asymmetry to the left. However, crucially, in each trial the density estimates of the treatment and control appear to be very similar, which ratifies the conclusion of the tests. In the next section we apply $\widehat{\eta}$ to test for symmetry in the residuals of the ANCOVA model.

3.4 Testing for symmetry in the residuals of the ANCOVA model

Section 3.3 compared asymmetry between two randomised groups to verify that they were balanced in terms of baseline characteristics. Another key component of randomised trials is to estimate the treatment effect (the difference in outcome response between treatment and control groups) at the end of the trial, using a statistical model. Such models often make normality assumptions, for example, it is commonly assumed that the variables or residuals of the model are normally distributed.

When handling data in practice, it is extremely unlikely that the variables (and as a consequence, model residuals) are sampled from a truly symmetric Normal population. Indeed, when collecting data in practice the number of error-inducing factors can be so numerous and complex that in reality we fully expect that the data are drawn from a population which is not strictly Normal. This discrepancy between theory and practice motivated Geary [37] to suggest that, "Normality is a myth; there never has, and never will be, a Normal distribution." For the most part this is not a serious issue, as for many commonly used tests, such as the t-test for testing the equality of means, the tests are robust to small departures from normality [93]. This presents the discerning statistician, who wants to validate normality or symmetry assumptions of their model, with a problem. This is because, when dealing with large data sets, tests for normality (or symmetry) will often produce significant results (i.e. those indicating departure from the model assumptions). However, this doesn't mean that the underlying data are not 'close to' normality or 'nearly symmetric'. This demonstrates the value of simple visual evidence, such as Q-Q plots or histograms, which instantly give a good idea of how closely the data conforms to a Normal population. The downside with this approach of course, is the fact that interpretation of the visual evidence is inherently subjective. We propose a less subjective 'test' for symmetry by placing ourselves somewhere between these two approaches. The measure of asymmetry $\widehat{\eta}$ introduced in the previous chapter gives an impression of the amount of

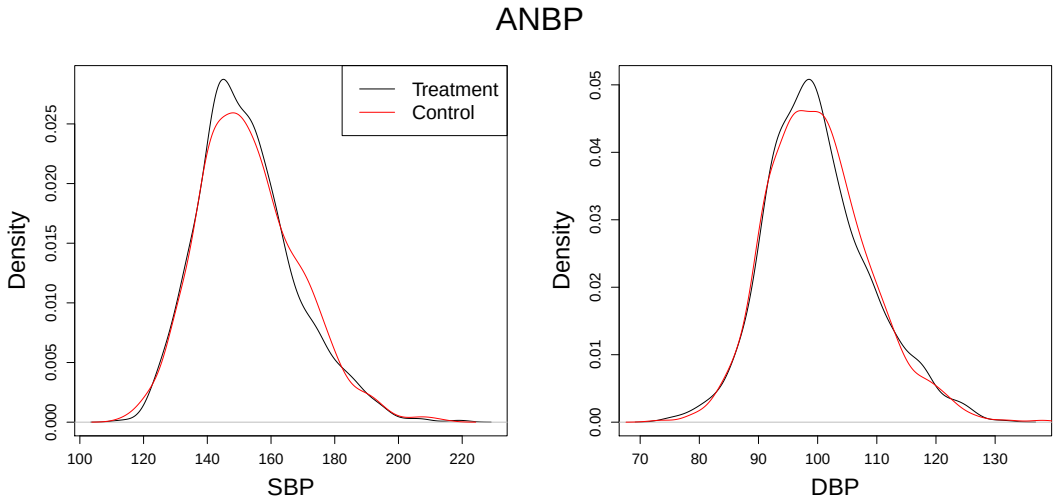

Figure 3.1: Comparing the treatment and control samples for initial systolic and diastolic blood pressure in the ANBP study.

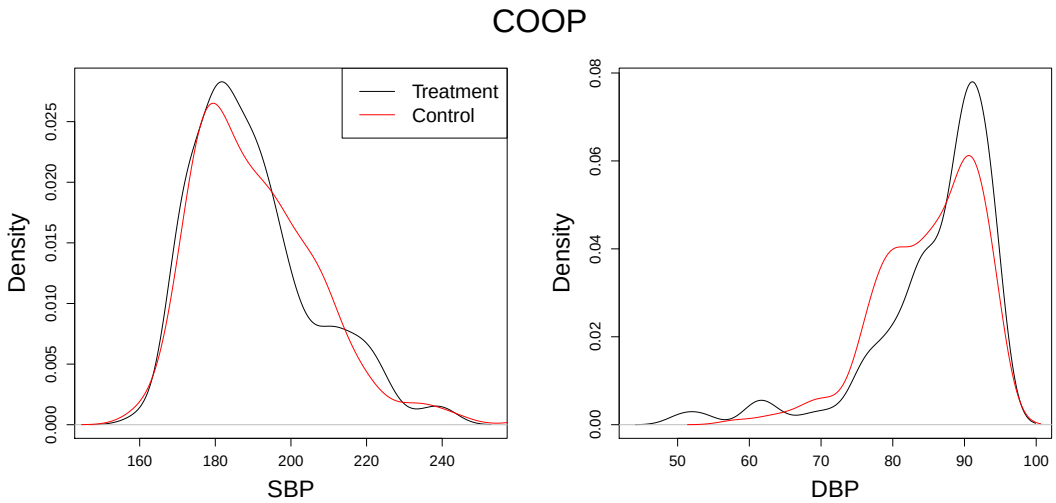

Figure 3.2: Comparing the treatment and control samples for initial systolic and diastolic blood pressure in the COOP study.

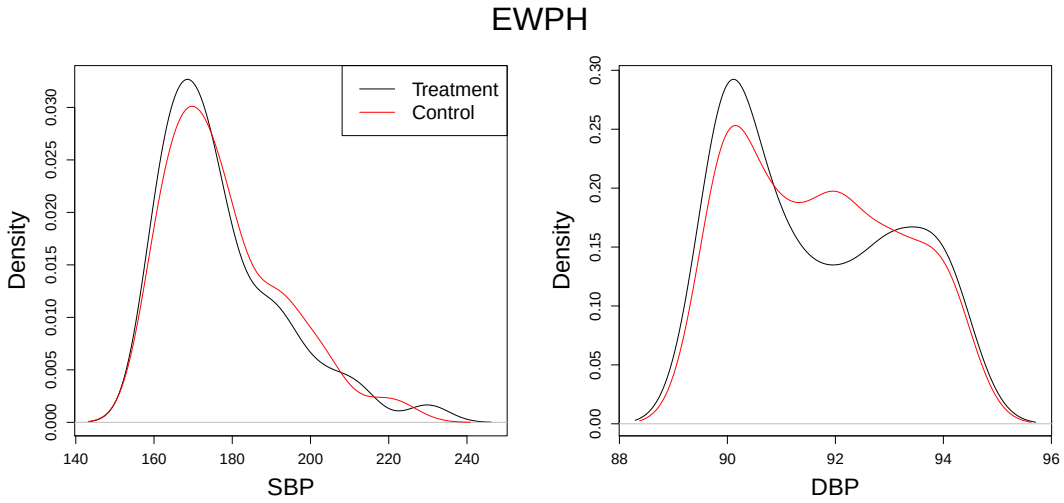

Figure 3.3: Comparing the treatment and control samples for initial systolic and diastolic blood pressure in the EWPH study.

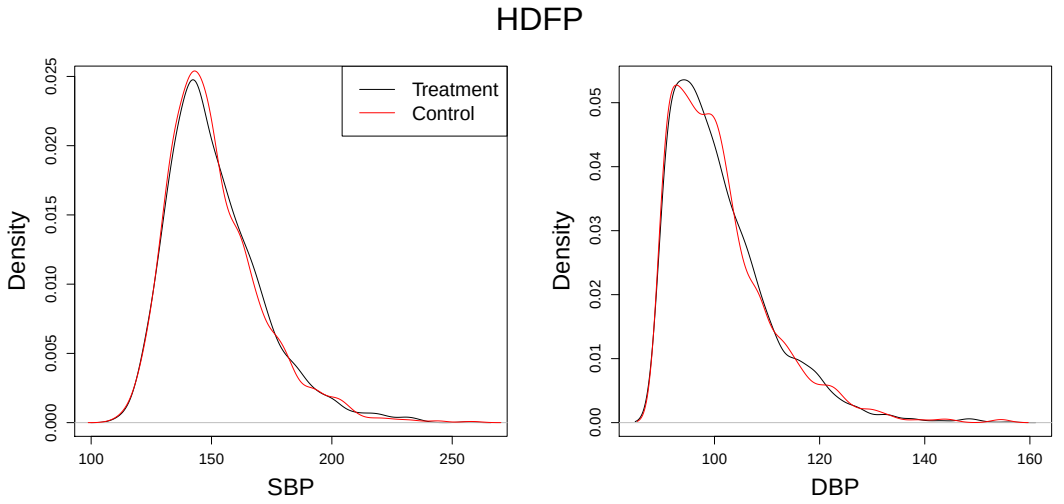

Figure 3.4: Comparing the treatment and control samples for initial systolic and diastolic blood pressure in the HDFP study.

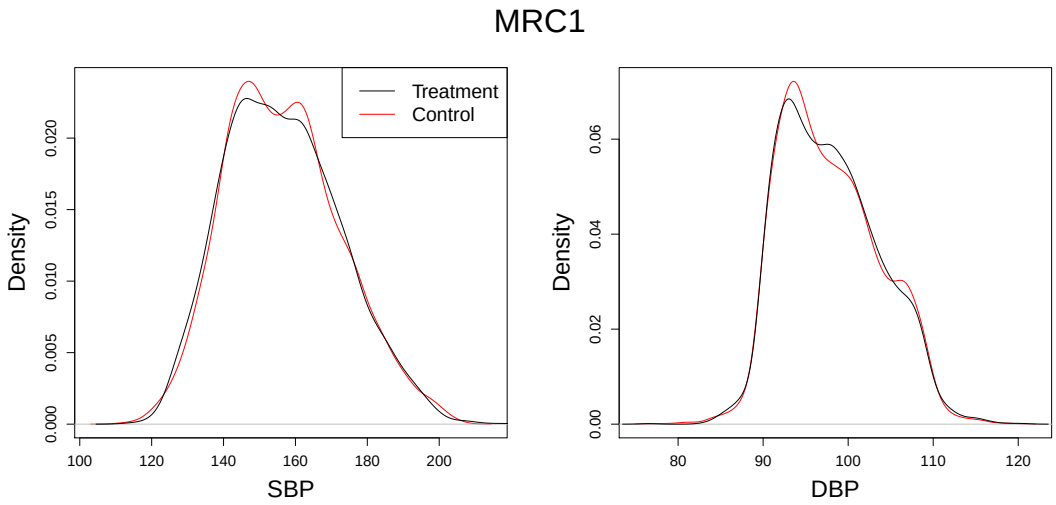

Figure 3.5: Comparing the treatment and control samples for initial systolic and diastolic blood pressure in the MRC1 study.

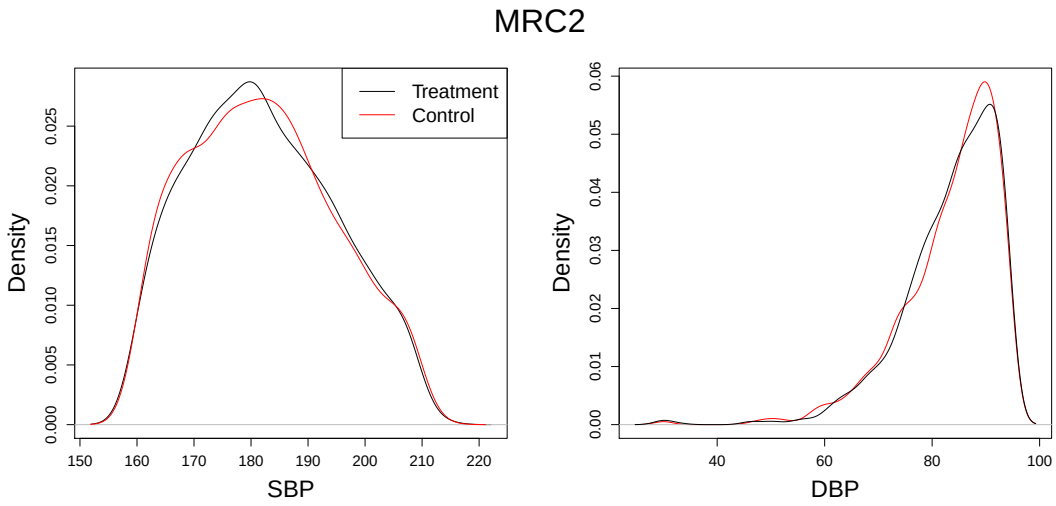

Figure 3.6: Comparing the treatment and control samples for initial systolic and diastolic blood pressure in the MRC2 study.

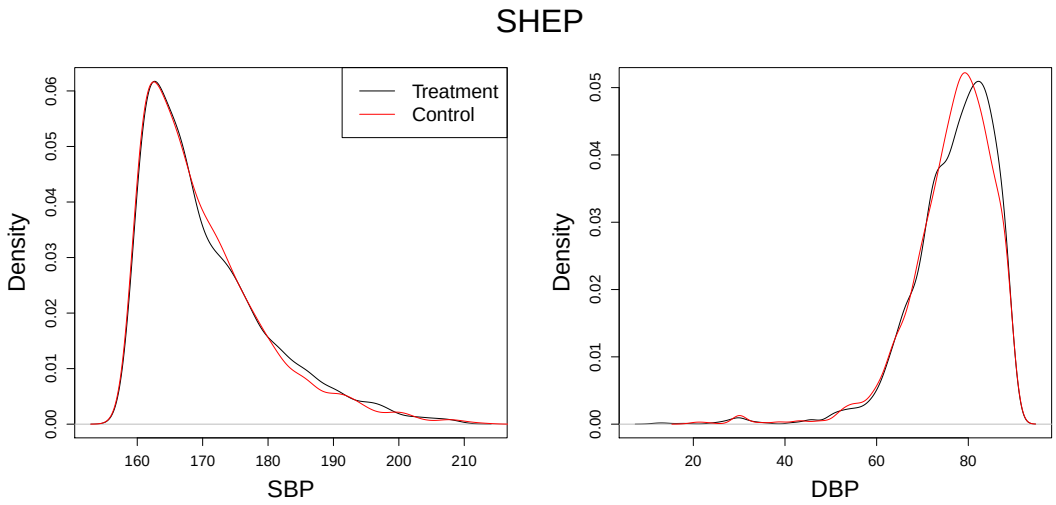

Figure 3.7: Comparing the treatment and control samples for initial systolic and diastolic blood pressure in the SHEP study.

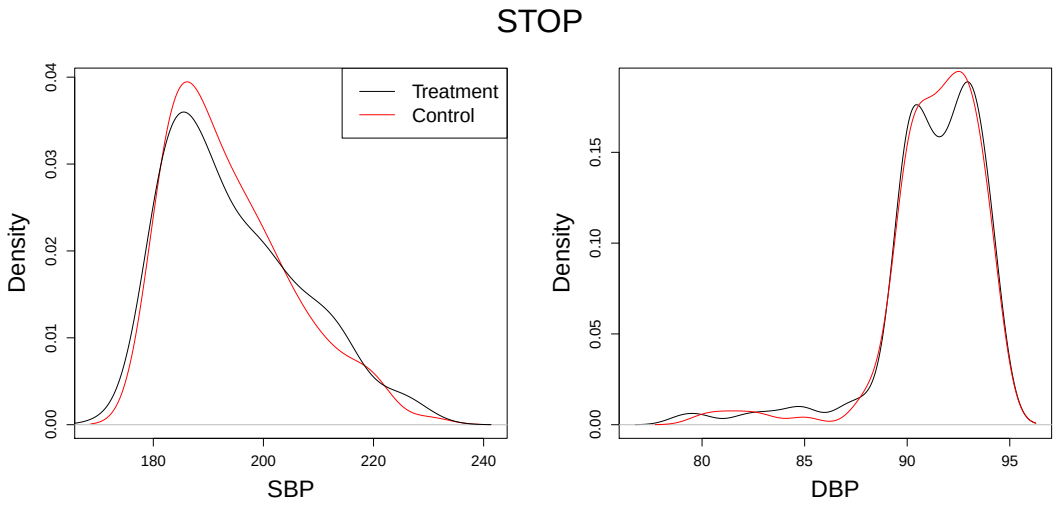

Figure 3.8: Comparing the treatment and control samples for initial systolic and diastolic blood pressure in the STOP study.

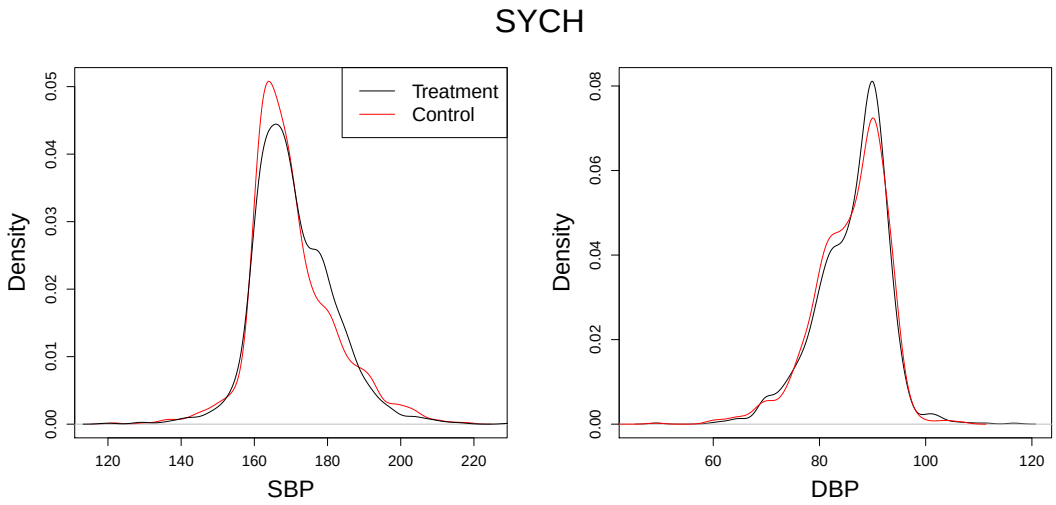

Figure 3.9: Comparing the treatment and control samples for initial systolic and diastolic blood pressure in the SYCH study.

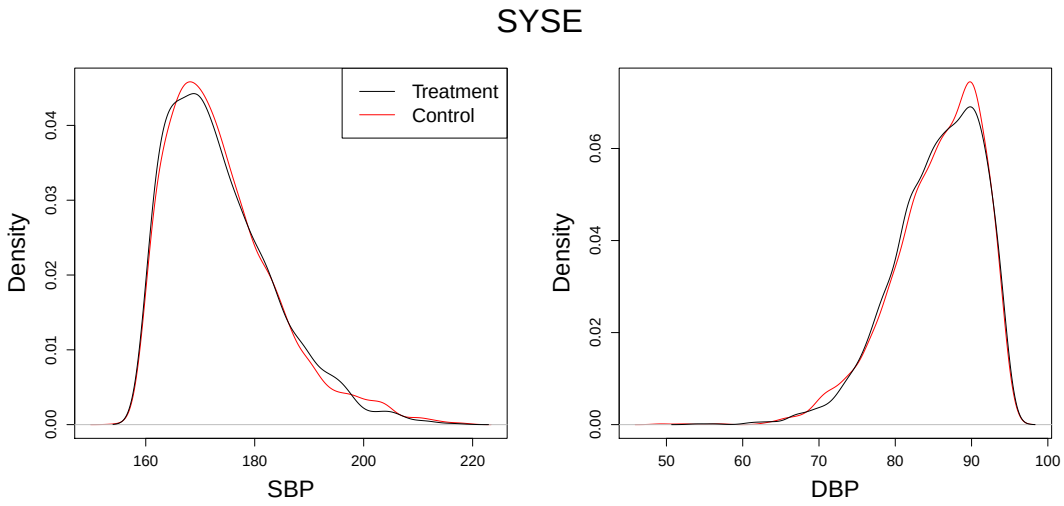

Figure 3.10: Comparing the treatment and control samples for initial systolic and diastolic blood pressure in the SYSE study.

asymmetry in the underlying sample. Therefore, it can be used to inform just how closely the sample conforms to normality. Also, using the distributional properties of $\widehat{\eta}$ outlined in Chapter 2 one can construct an approximate confidence interval for $\widehat{\eta}$. Hence, using the measure one can make a more objective decision as to whether the sample displays acceptable departure from symmetry. Alongside a Q-Q plot this provides a reasonable method for deciding whether or not there is excessive violation of the model assumptions.

We consider implementing this idea for the analysis of the hypertension trials using ANCOVA. Using the hypertension data included in the meta-analysis by Wang et al. [125] we propose the following ANCOVA model for each of the ten trials $i = 1, 2, \cdots, 10$,

$$Y_{F_{ij}} = \alpha_{_i} + \beta_{1_i} Y_{B_i} + \beta_{2_i} X_{_{ij}} + e_{_{ij}}, \qquad j = 1, \cdots, n_i,$$

where n_i is the sample size in study i, $Y_{F_{ij}}$ is the final blood pressure in millimetres of mercury (mm Hg) for patient j in study i. Similarly, Y_{B_i} denotes the initial blood pressure (in mm Hg), $X_{_{ij}}$ represents an indicator variable equal to 1 for the treatment group and 0 for the control group, while $e_{_{ij}}$ denotes the residual associated with patient j in study i. Furthermore, β_{1i} denotes the average effect of a unit increase in the baseline blood pressure on the resultant blood pressure in study i and β_{2i} denotes the treatment effect in trial i, correcting for the initial blood pressure. That is, the average effect of the treatment on the resultant blood pressure in study i, holding the initial blood pressure fixed. Table 3.2 shows the treatment effect estimates $\widehat{\beta}_{2i}$ and the 95% confidence interval for the treatment effect in the ANCOVA model. For example, in the EWPH trial the estimated treatment effect on systolic blood pressure is -12.88. This implies that the average effect of the treatment is to reduce blood pressure by 12.88 mm Hg. It is clear that the treatment effect is consistently below zero across all studies, which indicates that the treatment is effective at reducing blood pressure.

When conducting an analysis of covariance one makes a number of fundamental assumptions, including the assumption that the residuals of the model are normally distributed. One of the most common methods of validating this assumption involves the use of QQ-plots, that is, plotting the sample quantiles of the residuals against the theoretical quantiles that one would

Trial	Systolic		Diastolic	
	TE	95% CI	TE	95% CI
ANBP	−6.66	[−8.33; −4.99]	−2.99	[−4.01; −1.97]
COOP	−14.17	[−18.44; −9.89]	−7.87	[−10.23; −5.52]
EWPH	−12.88	[−19.22; −6.54]	−6.01	[−8.64; −3.38]
HDFP	−8.71	[−9.77; −7.64]	−5.11	[−5.74; −4.48]
MRC1	−8.70	[−9.44; −7.97]	−4.64	[−5.08; −4.20]
MRC2	−10.60	[−12.10; −9.10]	−5.56	[−6.40; −4.73]
SHEP	−11.36	[−12.43; −10.29]	−3.98	[−4.51; −3.44]
STOP	−17.93	[−22.68; −13.18]	−6.54	[−8.79; −4.28]
SYCH	−6.55	[−7.81; −5.29]	−2.08	[−2.72; −1.44]
SYSE	−10.26	[−11.12; −9.39]	−3.49	[−3.90; −3.09]

Table 3.2: ANCOVA results including mean treatment effect (TE) and 95% confidence intervals (CI).

expect from a Normal sample. A QQ-plot of a normally distributed sample would be approximately linear. However, this approach relies on the subjective opinion of the researcher as to what constitutes 'approximately linear'. We propose using $\widehat{\eta}$, alongside a QQ-plot, to test for the symmetry of the random sample and thus, we can establish a quantification of how far the sample departs from a symmetric Normal population.

In Table 3.3 we report the value of $\widehat{\eta}$ for the residuals in each trial, along with the approximate 95% confidence interval in each case. We also report the p-value, that is, the probability of observing such a value (or one at least of extreme) if the data were in fact perfectly symmetric. In the vast majority of cases it is shown that it is highly unlikely that the residuals are coming from a completely symmetric population, as the p-values are often less than 0.05. This is precisely what one should expect in such large randomised control trials, where such a large sample size means that the test based on $\widehat{\eta}$ has the power to detect even modest departures from symmetry. For example, most studies have over 1000 patients in total. The important thing to observe is that the value of $\widehat{\eta}$ is small in most of the cases, indicating only a small departure from symmetry (and hence only a slight violation of the normality assumption).

Figures 3.11 and 3.12 show the QQ-plots for the residuals from the models for systolic and diastolic blood pressures respectively. The QQ-plots compare the sample quantiles of the observed standardised residuals with the theoretical quantiles of the standard Normal distribution.

Trial	Systolic				Diastolic			
	$\widehat{\eta}$	95% CI	p-value	n	$\widehat{\eta}$	95% CI	p-value	n
ANBP	0.18	$[0.09; 0.26]$	< 0.001	1530	0.06	$[-0.03; 0.15]$	0.1	1530
COOP	-0.08	$[-0.25; 0.08]$	0.16	349	0.18	$[0.01; 0.35]$	0.01	349
EWPH	0.05	$[-0.27; 0.37]$	0.38	172	-0.24	$[-0.54; 0.05]$	0.05	172
HDFP	0.22	$[0.18; 0.26]$	< 0.001	4797	0.13	$[0.08; 0.18]$	< 0.001	4798
MRC1	0.14	$[0.10; 0.17]$	< 0.001	6991	-0.01	$[-0.05; 0.02]$	0.23	6991
MRC2	0.06	$[0.00; 0.12]$	0.02	2651	0.05	$[-0.02; 0.11]$	0.07	2650
SHEP	0.10	$[0.05; 0.15]$	< 0.001	4736	-0.07	$[-0.12; -0.02]$	< 0.001	4736
STOP	-0.23	$[-0.44; -0.02]$	0.02	268	-0.05	$[-0.25; 0.15]$	0.32	268
SYCH	0.06	$[-0.01; 0.13]$	0.05	2390	-0.03	$[-0.10; 0.04]$	0.19	2391
SYSE	-0.05	$[-0.10; 0.00]$	0.03	4695	-0.26	$[-0.30; -0.22]$	< 0.001	4695

Table 3.3: Testing for the symmetry of the residuals in the ANCOVA model using $\widehat{\eta}$ and T_n.

It is apparent that the majority of QQ-plots display a near linear trend indicating a reasonable agreement between the quantiles of the observed residuals with the Normal residuals. In the studies where $\widehat{\eta}$ is unusually large, with absolute value greater than 0.2 (HDFP and STOP for systolic blood pressure, and EWPH and SYSE for diastolic blood pressure), we see that the QQ-plots also identify asymmetry in the form of a non-linear trend between the sample quantiles and the theoretical Normal quantiles. It is also worthy of note that the QQ-plots for the diastolic and systolic blood pressure in the ANBP trial also display a small anomaly in the form of a vertical kink in the lower tail. This is evidence of a bimodal structure in the underlying distribution, which may go undetected by the measure of asymmetry $\widehat{\eta}$. Figures 3.13 and 3.14 show the density estimates of the residuals in the the models of systolic and diastolic blood pressures respectively. These figures provide further evidence of the small departures from symmetry in some of the trials, as well as confirming the presence of a bimodal residual distribution for the ANBP data.

As we have already identified, there are four trials which contain residuals which display potentially important asymmetry (an absolute value of $\widehat{\eta} > 0.2$). These are the HDFP and STOP trials in the systolic case and EWPH and SYSE in the diastolic case. When residuals fail to conform to a Normal distribution it is common practice to transform the dependent or independent variables to rectify this problem. In the next section we discuss the potential applications of $\widehat{\eta}$ to conducting transformations of asymmetric data.

3.5 Correcting for asymmetry

As we have already alluded to, violations of the normality assumption can potentially compromise the estimation of coefficients and the calculation of confidence intervals when analysing linear models [13]. It is sometimes the case that the error distribution is 'skewed' by the presence of outliers in the data. Whilst it is common practice to remove outliers from the data, one should be careful not to throw away meaningful data. Since parameter estimation is typically based on the minimization of squared error, a few extreme observations can exert a disproportionate influence on parameter estimates, as noted by Huber [57]. Furthermore, exact inference, such as the calculation of confidence intervals and various significance tests for coefficients are based on

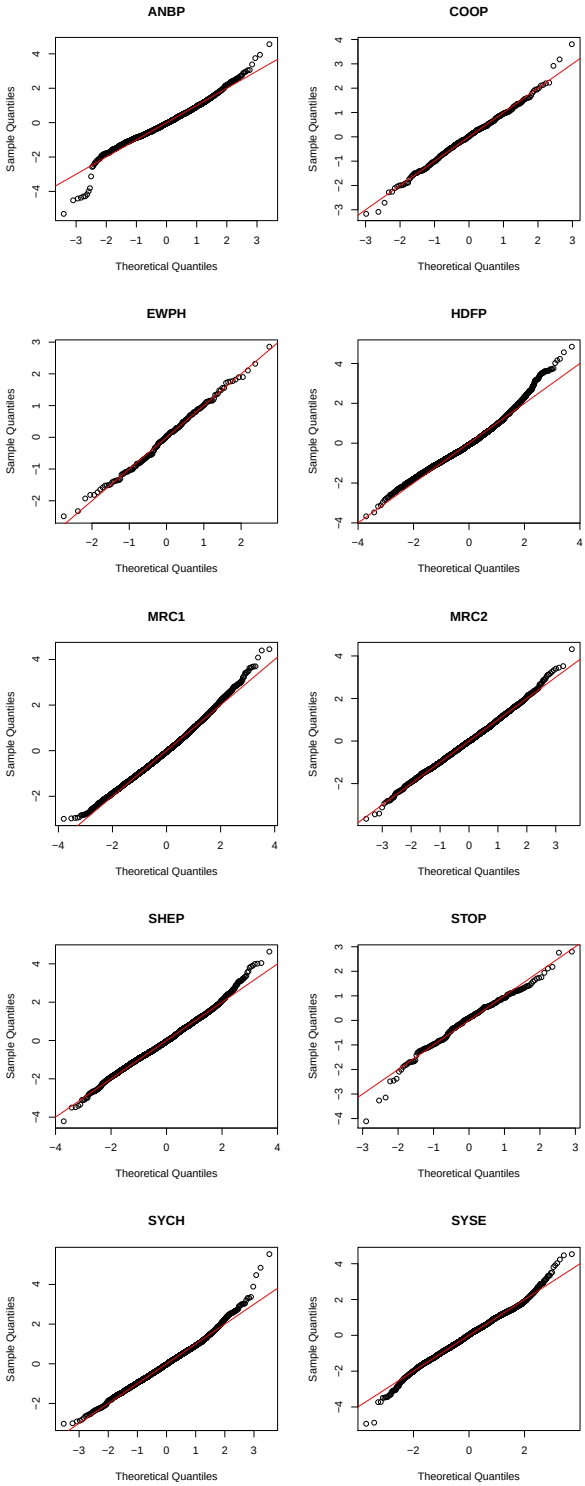

Figure 3.11: QQ-plots of the residuals for the ANCOVA model (systolic blood pressure).

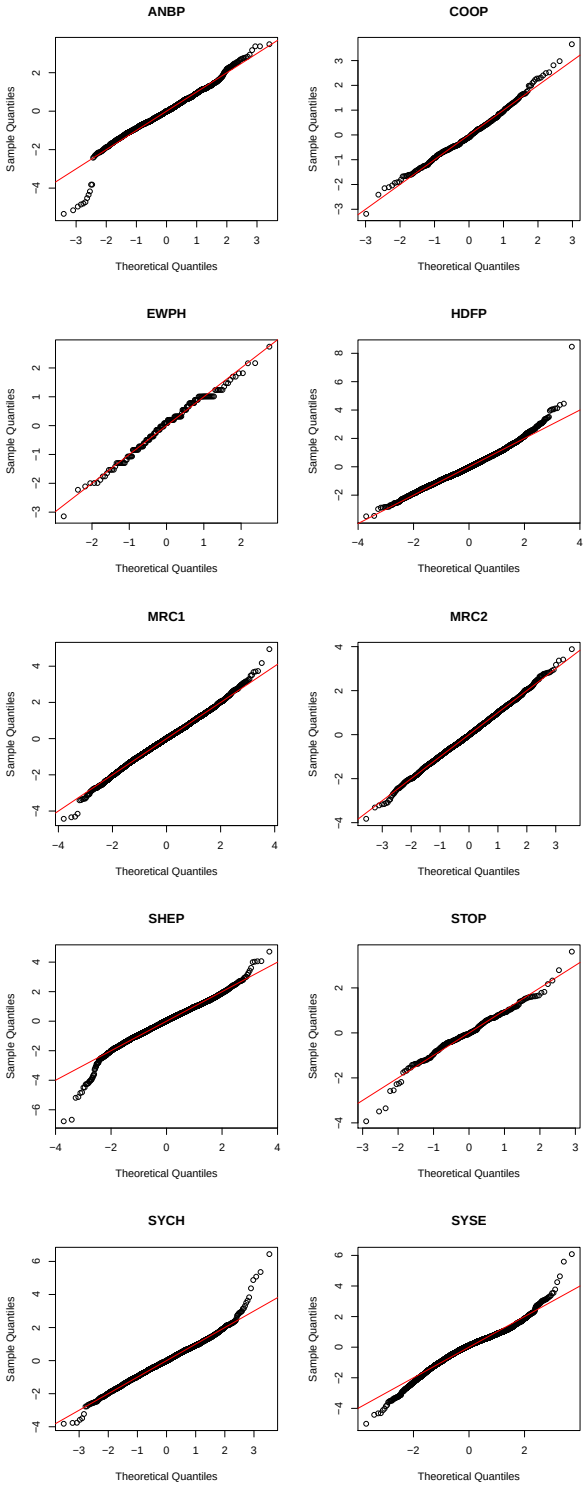

Figure 3.12: QQ-plots of the residuals for the ANCOVA model (diastolic blood pressure).

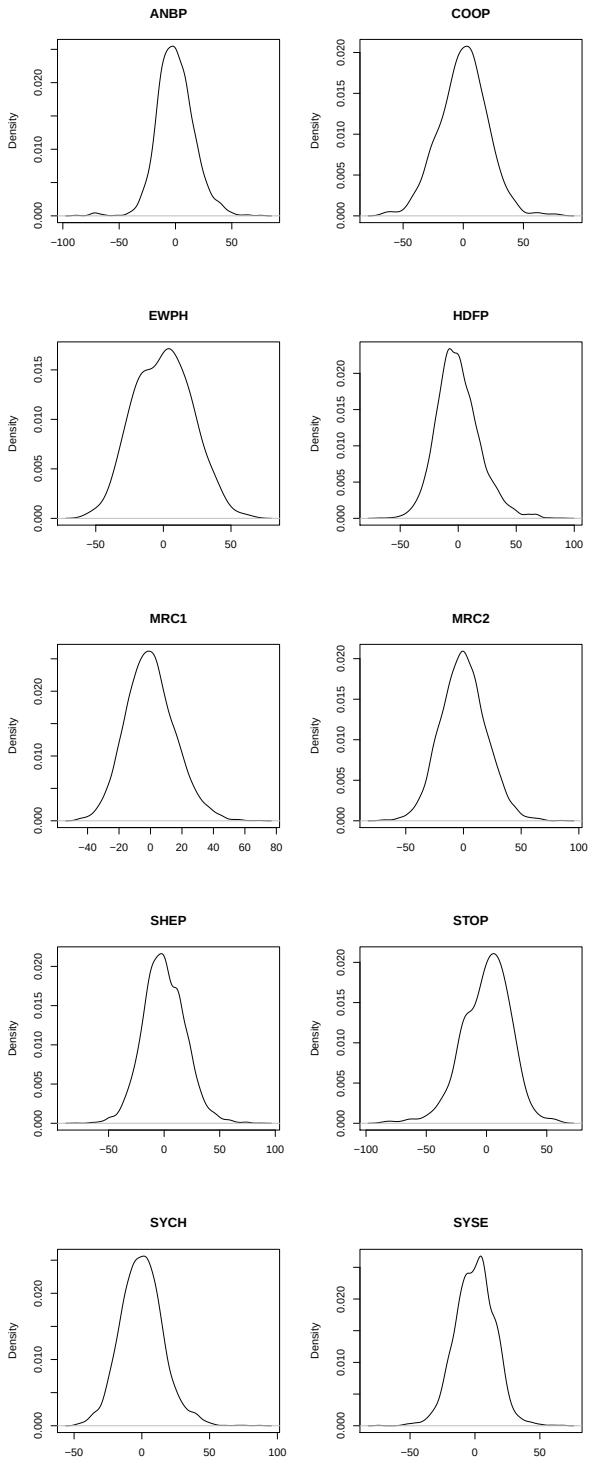

Figure 3.13: Density estimate of the residuals for the ANCOVA model (systolic blood pressure).

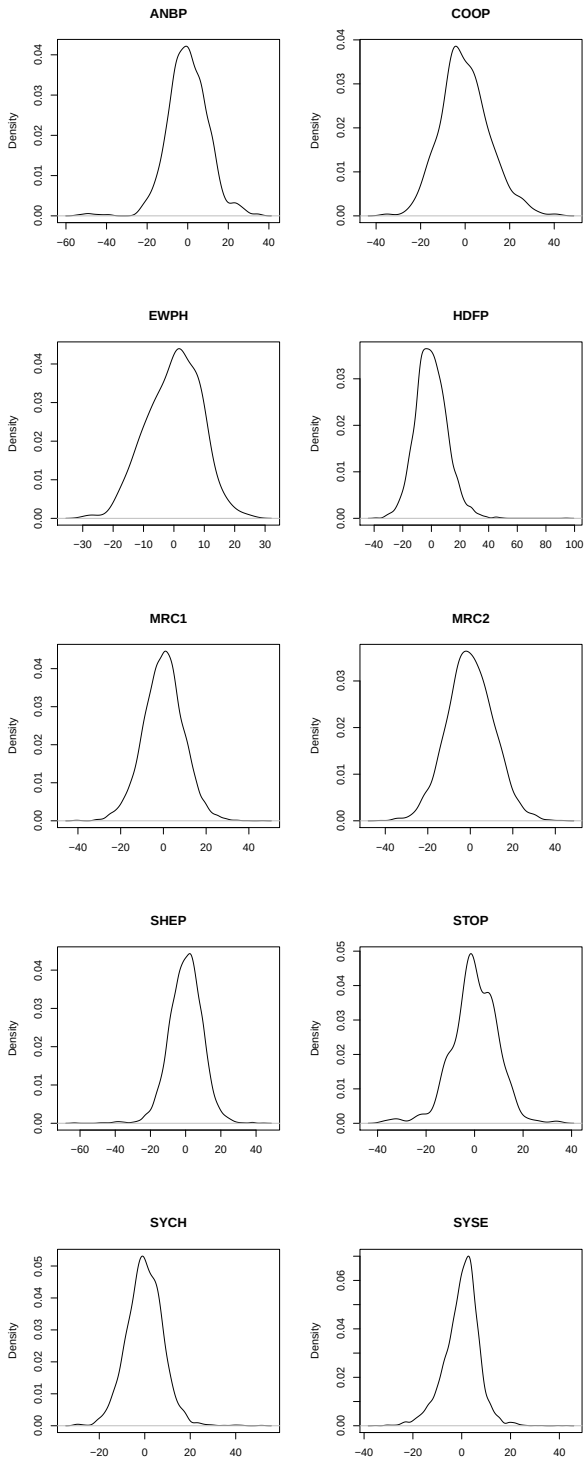

Figure 3.14: Density estimate of residuals for the ANCOVA model (diastolic blood pressure).

the assumption that the errors are normally distributed. If the error distribution is significantly non-Normal, confidence intervals may be too wide or too narrow [108]. The assumption of normality in residuals is often violated because the distributions of the dependent or independent variables are themselves significantly non-Normal, and in such cases it is common practice to transform the dependent or independent variables to rectify this problem.

One of the advantages of $\widehat{\eta}$ is that it can play a role in selecting an appropriate transformation, because the sign indicates the direction of the asymmetry. When $\widehat{\eta} > 0$ this implies the data are skewed towards the right and so we must apply a transformation which 'pulls in' the extreme values. Examples of such transformations include the logarithmic and square root transformations. On the other hand, if $\widehat{\eta} < 0$ then this indicates the data are skewed to the left and so there is a need to 'pull up' the smaller values. Such transformations could involve performing an exponential or power transformation.

We now illustrate this using the hypertension data once more, and examine whether transformation of the final blood pressure data improves the normality of the residuals from the ANCOVA model. Our first step is to select an appropriate transformation for the response variable, based on the value of $\widehat{\eta}$. After attempting a number of different transformations on the final blood pressure data it appears that for positive $\widehat{\eta}$ the logarithmic transform have the most appreciable effect on the asymmetry, whilst for the negative $\widehat{\eta}$ squaring the data appears most effective. Table 3.4 shows the effect of applying the appropriate transformation to the response variable. It is clear that the measure of asymmetry $\widehat{\eta}$ is markedly reduced, particularly in the HDFP, STOP, and EWPH trials. In fact applying a logarithmic transform to the systolic blood pressure data in the HDFP trial reduces the value of $\widehat{\eta}$ from 0.22 to 0.10. Figures 3.15 to 3.18 show the visual improvement in the symmetry of the residuals.

The results of this section clearly demonstrate that $\widehat{\eta}$ can be a useful tool in governing transformation of the covariate or response data in our ANCOVA model. Indeed, $\widehat{\eta}$ has a role to play in all stages of the data transformation process. Firstly, it provides an objective measure of whether data display inordinate departure from symmetry. Secondly, it provides a useful guide as to what sort of transformation might be the most effective. Finally, it can be used to

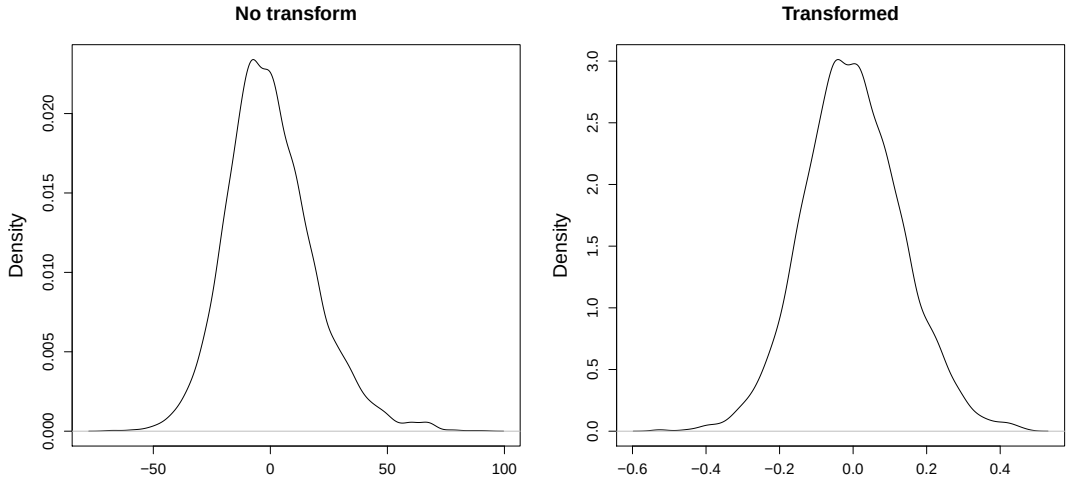

Figure 3.15: Density estimation of the raw residuals in the HDFP trial and the residuals after carrying out a logarithmic transform on the response variable (systolic blood pressure).

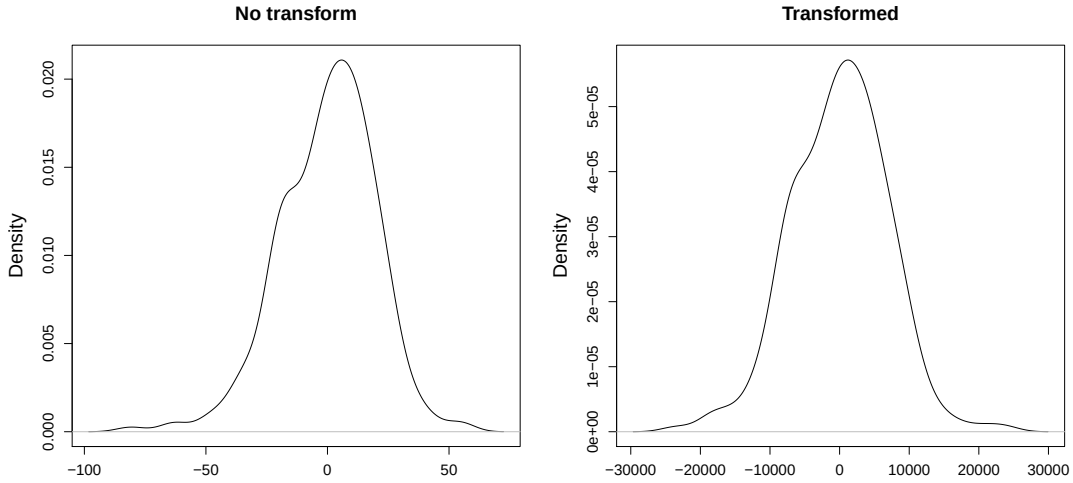

Figure 3.16: Density estimation of the raw residuals in the STOP trial and the residuals after squaring the response variable (systolic blood pressure).

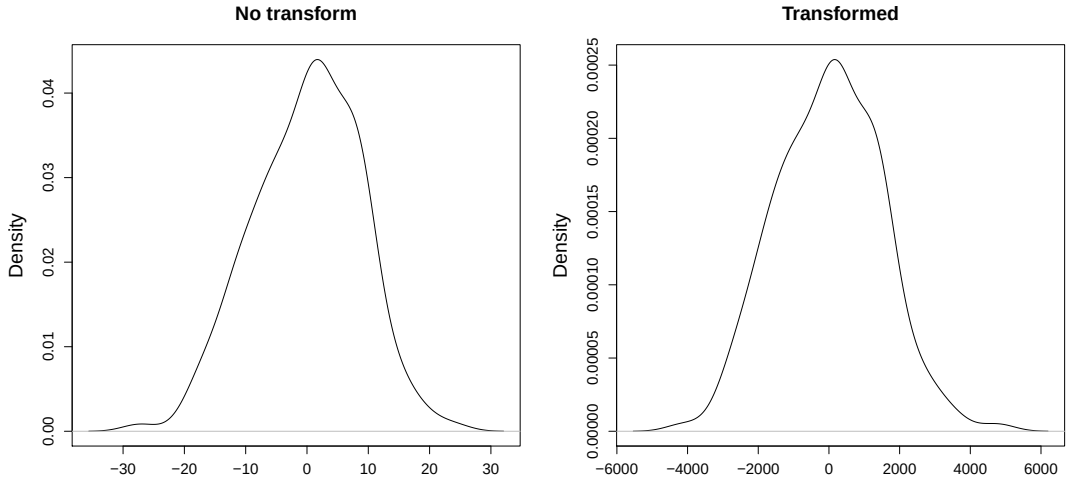

Figure 3.17: Density estimation of the raw residuals in the EWPH trial and the residuals after squaring the response variable (diastolic blood pressure).

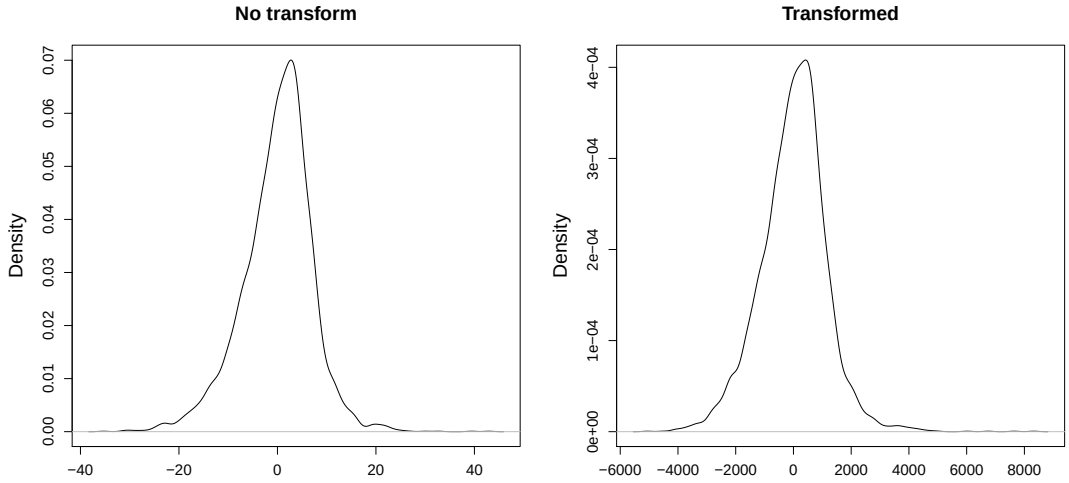

Figure 3.18: Density estimation of the raw residuals in the SYSE trial and the residuals after squaring the response variable (diastolic blood pressure).

Trial	Residual $\widehat{\eta}$ pre-transform	Transformation	$\widehat{\eta}$ after response transformation
HDFP (systolic)	$0.22\ (0.18, 0.26)$	Log	$0.10\ (0.05, 0.15)$
STOP (systolic)	$-0.23\ (-0.44, 0.01)$	Square	$-0.10\ (-0.32, 0.14)$
EWPH (diastolic)	$-0.24\ (-0.57, 0.04)$	Square	$-0.09\ (-0.43, 0.20)$
SYSE (diastolic)	$-0.26\ (-0.30, -0.22)$	Square	$-0.18\ (-0.22, -0.13)$

Table 3.4: Effect of applying an appropriate transformation to the response variable on the asymmetry of the residuals in the ANCOVA model.

objectively determine whether the transformation has been successful by providing a numerical value to quantify the difference in asymmetry before and after the transformation.

3.6 Limitations of testing for normality with $\widehat{\eta}$

It is important to note that the Normal distribution of $\widehat{\eta}$ described in Chapter 2 is asymptotic. As a result, we are reliant on large samples to accurately conduct tests and report p-values based on $\widehat{\eta}$. For small samples, the actual value of $\widehat{\eta}$ remains a reasonable measure of the asymmetry of the data, but because we are estimating f and F we do require a reasonable number of data points to build an accurate description of the underlying density and distribution function. Further, one must be mindful of the fact that $\widehat{\eta}$ is a measure of asymmetry, not normality. Hence, even if $\widehat{\eta}$ is close to zero the sample need not represent a Normal population. We can use the hypertension data once again to demonstrate this fact. In the hypertension data some of the studies specifically include more elderly patients, for example MRC2 and STOP. On the other hand, some studies are interested in younger patients, for example, ANBP and HDFP. As a result when we pool together all of the studies we can see that the age data are described by a clearly bimodal population. This can be seen in Figure 3.19, which shows the density estimate of the pooled age data. Interestingly however, due to the incidental symmetry of this bimodal sample, $\widehat{\eta}$ is actually very small, and the results can be seen in Table 3.5. In fact, we are unable to reject the null hypothesis of symmetry even with such a large sample size. This example demonstrates that $\widehat{\eta}$ alone is not a suitable indicator of the normality of a set of data. Rather,

one should also consult a histogram and QQ-plot before jumping to the conclusion that a given sample is Normal just because $\widehat{\eta}$ is small. Figure 3.20 shows the QQ-plot for this set of data and it is immediately clear that the age data are not Normal.

	Treatment			Control		
	$\widehat{\eta}$	p-value	n	$\widehat{\eta}$	p-value	n
Age	0.01	0.25	14459	0.01	0.36	14122

Table 3.5: Testing symmetry of age data pooled across studies.

3.7 Discussion

In this chapter it was shown that, for a reasonably large data set, $\widehat{\eta}$ has a number of useful applications as a test statistic, as well as a summary statistic. Using the hypertension data set collated by Wang et al. [125] one is able to test for baseline imbalance between the treatment and control samples in each trial. Furthermore, $\widehat{\eta}$ can also be used to test the symmetry of residuals in ANCOVA models, which is an important assumption of the model. Furthermore, it was shown that $\widehat{\eta}$ can play a crucial role in identifying when to transform the data, what type of transformation is likely to be effective, and evaluating whether the transformation has been a success. In conclusion, it appears that $\widehat{\eta}$ has the potential to be a very useful tool in a number of different important applications.

Of course, while we have exclusively considered using $\widehat{\eta}$, there are many other measures of skewness or asymmetry that would be viable for the applications discussed here. For example, in Chapter 1 we discussed measures of asymmetry proposed by MacGillivray [76], Boshnakov [11], and Li and Morris [73], as well as the multiple measures of skewness proposed by Pearson [91] and the quantile measure of skewness proposed by David and Johnson [22]. However, each of these measures have a number of shortcomings when compared with $\widehat{\eta}$. For example, the method given by MacGillivray [76] disregards a certain amount of the probability mass in the tails of the random variable. This is a significant drawback when measuring the asymmetry of a random variable, as the tails obviously provide crucial information in that regard. Moreover, the measures of asymmetry proposed by Boshnakov [11] are only applicable to unimodal distributions, and so are only valid for a reduced class of random variables.

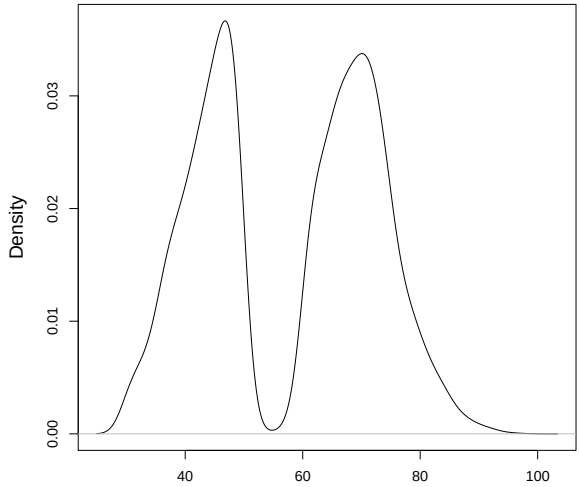

Figure 3.19: Density estimate of pooled age data.

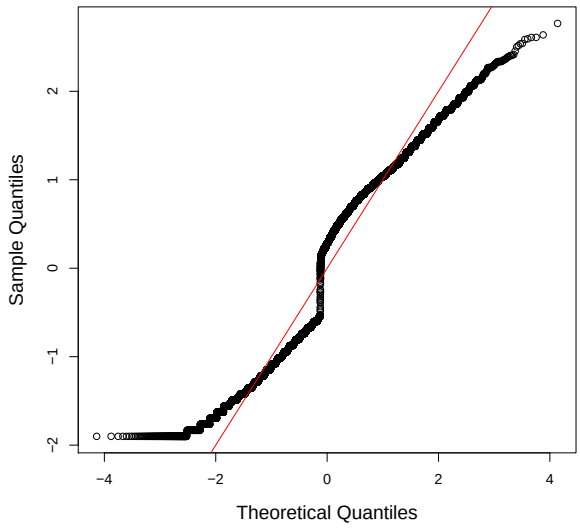

Figure 3.20: QQ-plot of age data.

Furthermore, the measures of skewness discussed here all fail to effectively calibrate asymmetry in comparison with η. This is demonstrated in Chapter 2 where we consider two tests of symmetry based on Pearson's skewness coefficients. For example, the test proposed by Cabilio and Masaro [17] is based on the sample version of γ_1, while the test detailed by Gupta [43] is based on the sample version of γ_3. It was established in Chapter 2 that each of these tests have power which is inferior to the test based on $\widehat{\eta}$.

It was also established that one of the shortcomings of $\widehat{\eta}$ is that the distribution given in Chapter 2 is asymptotic. As a result, one requires a reasonably large sample to apply the Normal approximation. Large samples were available in the hypertension trials, but other situations (e.g. observational studies of rare diseases) may involve small numbers of patients. In the subsequent chapter we demonstrate that it is not appropriate to assume a Normal distribution for $\widehat{\eta}$ in small samples, before proposing a new estimation procedure which is more robust for small samples.

In summary,

- $\widehat{\eta}$ can be used to compare the baseline characteristics of a randomised control trial, either as an additional summary statistic or to statistically test for a difference in the distributions (though such routine testing is not advised).
- $\widehat{\eta}$ can provide a helpful insight in guiding and evaluating the transformation of skewed data.
- $\widehat{\eta}$ can also be used to assess the normality of a set of data, provided it is used alongside an appropriate visual aid, such as a QQ-plot.
- Small samples may compromise the accuracy of confidence intervals and p-values, but $\widehat{\eta}$ still provides a helpful measure of asymmetry provided there are sufficient samples to accurately estimate the density or distribution function.

Chapter 4

An investigation into the small sample properties of $\widehat{\eta}$

4.1 Introduction

Up to this point, we have only considered using $\widehat{\eta}$ to measure asymmetry or test for symmetry in large samples. As we have previously identified, the distribution of $\widehat{\eta}$ calculated in Chapter 2 is asymptotic and so requires a reasonably large sample to apply the Normal approximation. In this chapter we demonstrate that it is not appropriate to assume a Normal distribution for $\widehat{\eta}$ for small samples, before proposing a new estimation procedure which is theoretically more robust for small samples. We then compare the performance of this new procedure with $\widehat{\eta}$ in small and large samples by way of a simulation study, as well as two examples using real data.

In section 4.2 we estimate the sampling distribution of $\widehat{\eta}$ for small samples and reveal the limitations of applying $\widehat{\eta}$ in this case. This is the motivation for a transformed measure which is shown to have better distributional properties for small samples. In section 4.3 we derive the asymptotic distribution of this new measure and demonstrate that this distribution is more accurate for small samples. In section 4.4 we introduce and discuss the bootstrap, to assist in the calculation of the standard error of the new measure and thereby allow us to construct accurate confidence intervals. In section 4.5 we use a simulation study to compare the raw and transformed measure by assessing the accuracy the confidence intervals for η obtained by

using the two methods, where the variance is estimated using both bootstrapping and asymptotic theory. In section 4.6 we compare the effectiveness of the two measures at quantifying asymmetry using two real data examples. Finally, section 4.7 summarises the results of this chapter.

Aims of the chapter:

- Demonstrate the limitations of $\widehat{\eta}$ in small samples.
- Develop a new measure $\widehat{\zeta}$ that has better small sample properties.
- Compare the performance of the two measures using real and simulated data.

4.2 The sampling distribution of $\widehat{\eta}$ for small data sets

Recall that η is defined as the correlation coefficient between $f(X)$ and $F(X)$. Likewise, $\widehat{\eta}$ can be regarded as the corresponding sample estimate of this correlation coefficient. Tjostheim [120] notes that the small sample estimation of the sample correlation coefficient is somewhat problematic. Indeed, Tjostheim states that it is "well established that the sampling distribution of the sample correlation coefficient is appreciably skewed" even for quite substantial sample sizes. Soper et al. [116] previously carried out an extensive investigation into the distribution of the sample correlation coefficient in small samples. They conclude that the distribution is substantially non-Normal for samples of 25 and 50, irrespective of the value of the true correlation coefficient ρ. Further, while there is a better approximation for $n = 100$ when $\rho < 0.4$, there is still a substantial deviation for larger values of ρ.

Therefore, for small samples the Normal approximation that was derived in Chapter 2 is invalidated, as are the proposed estimates of the variance. To clarify, whilst the estimated value of η is not significantly compromised, small samples present a barrier to calculating accurate confidence intervals. As a result, small samples will also have an adverse effect on the performance of our test based on $\widehat{\eta}$, proposed in Chapter 2. In particular, we can no longer guarantee the accuracy of the calculated p-values. One suggestion is to apply the asymptotic theory with the t distribution in place of the Normal distribution, to account for the uncertainty in the

estimate of the variance of $\widehat{\eta}$. However, the variance estimate proposed in Chapter 2 (equation (2.9)) is generally already conservative, as we will show in section 4.5. Also, the t distribution fails to account for asymmetry in the small sample distribution of $\widehat{\eta}$.

Figures 4.1 and 4.2 show the estimated sampling distribution of $\widehat{\eta}$ based on $10,000$ small samples ($n = 15$) from several random variables. It is clear that the sampling distribution of $\widehat{\eta}$ is particularly non-Normal when the samples are taken from asymmetric random variables (when η is closer to the extreme value of 1). In particular, the sampling distribution is heavily skewed to the left as $\widehat{\eta}$ cannot take values larger than one. In the next section we propose how to transform $\widehat{\eta}$ to reduce the skewness in the sampling distribution.

4.3 ζ - The Fisher Z-transformation of η

The Fisher Z-transform of the sample correlation coefficient r,

$$Z(r) = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right),$$

is known to be a better approximation to normality.

Fisher [35] proposed the Z-transformation to stabilise the variance of the sample correlation coefficient r. That is, the transformation ensures that the variance of $Z(r)$ is similar for all values of r. Moreover, Fisher [36] showed that for small samples $Z(r)$ follows a Normal distribution more closely than r. Fisher comments that for just 8 observations "when these correlations are distributed over the scale of r, the (density) curves are far from Normal even when (the true correlation coefficient) ρ is equal to 0, and when $\rho = 0.8$ they become extremely skew." By contrast, Fisher observes that in the Z-transformed scale the density curve "appears symmetrical" and furthermore, "the entire curve is identical for all values of ρ."

Indeed, simulations appear to suggest that the finite sample behaviour of $Z(\widehat{\eta})$ is better than $\widehat{\eta}$, that is, $Z(\widehat{\eta})$ appears to follow a Normal distribution more closely than $\widehat{\eta}$ for small samples. Figures 4.3 and 4.4 show the estimated sampling distribution of $Z(\widehat{\eta})$ based on $10,000$ small samples ($n = 15$) from several random variables. In this case the sampling distribution is significantly improved compared to $\widehat{\eta}$ in Figures 4.1 and 4.2, despite still showing some slight

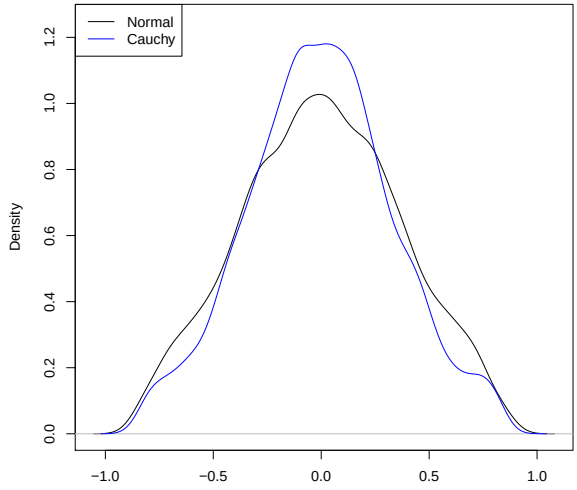

Figure 4.1: Density estimate of the sampling distribution of $\widehat{\eta}$ based on $10,000$ small samples of size $n = 15$ from Normal and Cauchy random variables.

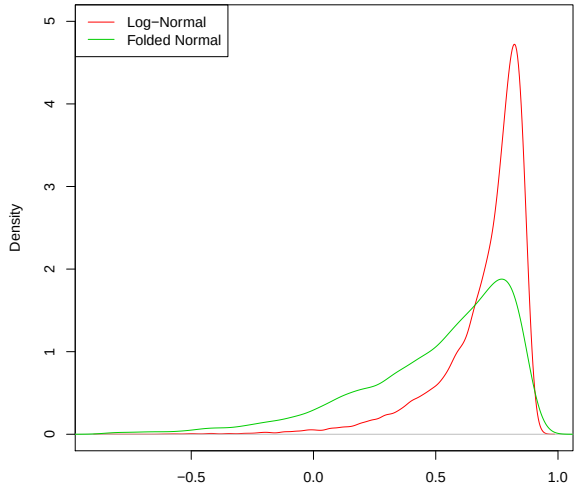

Figure 4.2: Density estimate of the sampling distribution of $\widehat{\eta}$ based on $10,000$ small samples of size $n = 15$ from the Log-Normal and Folded Normal distributions.

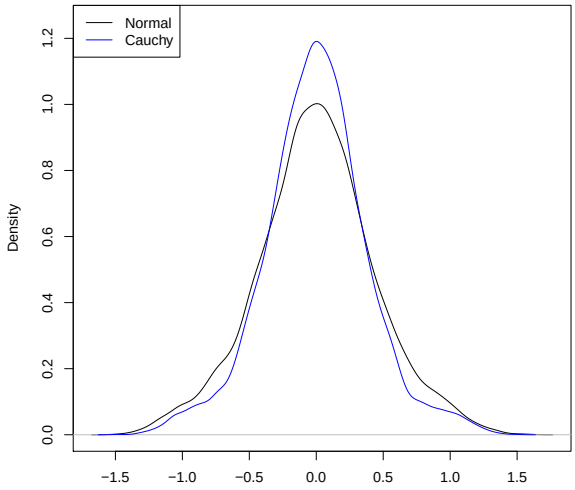

Figure 4.3: Density estimate of the sampling distribution of the Fisher Z-transformed measure $Z(\widehat{\eta})$ based on $10,000$ small samples of size $n = 15$ from Normal and Cauchy distributions.

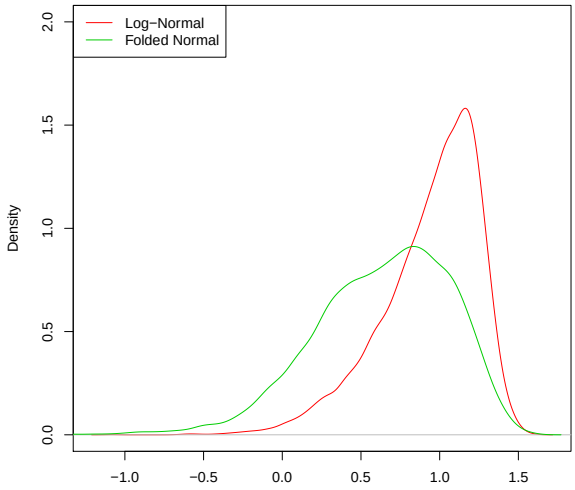

Figure 4.4: Density estimate of the sampling distribution of the Fisher Z-transformed measure $Z(\widehat{\eta})$ based on $10,000$ small samples of size $n = 15$ from the Log-Normal and Folded Normal distributions.

skewness to the left for the asymmetric random variables.

Therefore, let

$$\zeta = Z(\eta),$$

and

$$\widehat{\zeta} = Z(\widehat{\eta}).$$

The asymptotic distribution of $\widehat{\zeta}$ is outlined in the following theorem.

Theorem 4.1 *Let $X_1, \cdots, X_n$ be a random sample from a continuous probability density function $f(x)$ and distribution function $F(x)$. Then, under the assumptions of Theorem 2.3, as $n \to \infty$*

$$\sqrt{n} \cdot \left[\widehat{\zeta} - \zeta\right] \xrightarrow{L} N(0, \tau^2),$$

with

$$\tau^2 = \frac{\sigma^2}{(1-\eta^2)^2},$$

and

$$\sigma^2 = \mathrm{Var}\left[\frac{2}{\sqrt{\nu_f \nu_F}}\left(f(X)F(X) - \frac{1}{2}f(X)\right) + \int_X^\infty \frac{f(y)^2}{\sqrt{\nu_f \nu_F}}dy \right.$$
$$\left. + \eta\left\{\frac{\left(F(X) - \frac{1}{2}\right)^2}{2\nu_F} + \frac{(f(X) - \mu_f)^2}{2\nu_f} + \int_X^\infty \frac{\left(F(y) - \frac{1}{2}\right)}{\nu_F} f(y)dy + \frac{(f(X) - \mu_f)f(X)}{\nu_f}\right\}\right],$$

where ν_f and ν_F denote $\mathrm{Var}(f(X))$ *and* $\mathrm{Var}(F(X))$ $\left(= \frac{1}{12}\right)$ *respectively, and $\mu_f = \mathrm{E}[f(X)]$.*

The proof follows almost directly from Theorem 2.3.

Proof. Recall that from Theorem 2.3,

$$\sqrt{n} \cdot [\widehat{\eta} - \eta] \xrightarrow{\mathrm{L}} N(0, \sigma^2).$$

Also, it is readily calculated that

$$Z'(r) = \frac{1}{1-r^2}.$$

The delta method [86] states that, if there is a sequence of random variables X_n such that

$$\sqrt{n} \cdot [X_n - \theta] \xrightarrow{\text{L}} N(0, \varsigma^2),$$

then

$$\sqrt{n} \cdot [g(X_n) - g(\theta)] \xrightarrow{\text{L}} N\left(0, \left[g'(\theta)\right]^2 \varsigma^2\right),$$

for any function g with at least one continuous derivative and $g'(\theta) \neq 0$.

Hence, an application of the delta-method gives

$$\sqrt{n} \cdot \left[\widehat{\zeta} - \zeta\right] \xrightarrow{\text{L}} N(0, \tau^2). \qquad \square$$

One can readily estimate τ^2 by simply transforming a suitable estimate of σ^2, for example, $\widehat{\sigma}^2$ given in equation (2.9) in Chapter 2. In this case, the estimate of τ^2 is

$$\widehat{\tau}^2 = \frac{\widehat{\sigma}^2}{(1 - \widehat{\eta}^2)^2}. \tag{4.2}$$

Therefore the standard error of $\widehat{\zeta}$ is given by

$$\text{s.e.}\left(\widehat{\zeta}\right) = \frac{\widehat{\tau}}{\sqrt{n}}. \tag{4.3}$$

Now, instead of calculating a confidence interval for η directly, we can instead calculate a confidence interval for ζ and transform back using

$$Z^{-1}(x) = \frac{\exp(2x) - 1}{\exp(2x) + 1} = \tanh(x),$$

to obtain confidence interval for η. Now, the advantage of this approach is that it ensures that we do not obtain any confidence intervals for η which include values outside of $[-1, 1]$. However, it is worth noting that the distribution outlined in Theorem 4.1 is also asymptotic and, in fact, carrying out the delta-method creates another source of error. Therefore the estimate of the variance of $\widehat{\zeta}$, given in equation (4.2), is more imprecise than the variance estimate of $\widehat{\eta}$. We

demonstrate this in section 4.5 when we compare the effectiveness of the two approaches using simulated data. Hence, although we are able to reduce the skewness in the sampling distribution of the measure, it comes at a price. Namely, that estimating the variance by conventional methods becomes more difficult. This motivates an alternative procedure, bootstrapping, to estimate the variance. The next section is dedicated to introducing bootstrapping and discussing how to apply it to help in estimating the variance of $\widehat{\zeta}$.

4.4 Bootstrapping

As noted by Efron and Tibshirani [31], the bootstrap is a data based simulation method for statistical inference, which can be used to make inferences about almost any test statistic, such as estimating the variance and generating confidence intervals. The use of the term bootstrap derives from the phrase to pull oneself up by one's bootstrap. The simple principle behind the bootstrap is to re-sample from the available data and thus approximate the sampling distribution of the statistic of interest.

In the context of testing for symmetry or measuring asymmetry non-parametric bootstrapping proceeds as follows. Given a random sample from an unknown population with an unknown amount of asymmetry η, draw with replacement another sample from it and repeat until the same sample size is obtained. Repeating this process a large number of times, say 1000 times, one can obtain 1000 bootstrap samples. Then η can be estimated in each replicate and the 1000 estimates provide an approximate sampling distribution of $\widehat{\eta}$, which can be used to make inferences about η.

Our interest is to construct accurate confidence intervals for η and ζ, which can be asymmetric for small samples. Suppose that $\underline{X}_0 = (X_1, X_2, \cdots, X_n)$ is the random sample of interest, which is drawn from a population with an unknown quantity of asymmetry η_0. Further, let $\widehat{\eta} = \eta(\underline{X}_0)$ be the observed estimate of η in the original sample $\underline{X}_0$. Let $\underline{\tilde{X}}_1, \underline{\tilde{X}}_2, \cdots, \underline{\tilde{X}}_R$ be the surrogate data sets that are re-sampled from the original data $\underline{X}_0$. Further, suppose that $\tilde{\eta}_1, \tilde{\eta}_2, \cdots, \tilde{\eta}_R$ are the corresponding estimates of η in the surrogate data sets. This collection of estimates can then be used to make inferences about η. For example, one can obtain an estimate of the variance of $\widehat{\eta}$ or construct an approximate 95% confidence interval.

Approximate confidence intervals based on bootstrapping were pioneered by Efron [28]. The basic $100(1-\alpha)\%$ bootstrap confidence interval for η is given by

$$\left[2\widehat{\eta} - \tilde{\eta}_{\left(1-\frac{\alpha}{2}\right)};\ 2\widehat{\eta} - \tilde{\eta}_{\frac{\alpha}{2}}\right], \tag{4.4}$$

where $\tilde{\eta}_\alpha$ is the α quantile of the bootstrap sample, $\tilde{\eta}_1, \tilde{\eta}_2, \cdots, \tilde{\eta}_R$ [23]. Similarly, the corresponding $100(1-\alpha)\%$ confidence interval for ζ is

$$\left[2\widehat{\zeta} - \tilde{\zeta}_{\left(1-\frac{\alpha}{2}\right)};\ 2\widehat{\zeta} - \tilde{\zeta}_{\frac{\alpha}{2}}\right], \tag{4.5}$$

where $\tilde{\zeta}_\alpha$ is the α quantile of the bootstrap sample, $\tilde{\zeta}_1, \tilde{\zeta}_2, \cdots, \tilde{\zeta}_R$.

In fact, since their first implementation a number of other competing methods for generating bootstrap confidence intervals have been developed [30]. Aside from the basic confidence interval defined in equation (4.4) other methods include: the percentile bootstrap (based directly on the quantiles of the surrogate estimates $\tilde{\eta}_i$); the bootstrap-t; and the bias corrected bootstrap [29]. Efron and Tibshirani [31] state that "the Normal and t percentage points are symmetric about zero, and as a consequence the resulting intervals are symmetric about the point estimate. In contrast, the bootstrap-t percentiles can be asymmetric about 0, leading to intervals which are longer on the left or right."

However, it is widely known that the sample correlation coefficient is problematic to handle with the ordinary bootstrap. In fact, Hall [44] goes as far as describing the problem of constructing a confidence interval for a correlation coefficient as the "smoking gun" of bootstrap methods. Indeed, Hall comments that "when used without a variance-stabilizing transformation, percentile-t fails spectacularly, producing intervals that have poor coverage accuracy and fail to respect the range of the coefficient. This seems to be case regardless of the choice of the variance estimate."

For this reason, we construct rough bootstrap confidence intervals for $\widehat{\eta}$ and $\widehat{\zeta}$ using equations (4.4) and (4.5). In the next section we apply this method to generate 95% confidence intervals for η using $\widehat{\eta}$ and $\widehat{\zeta}$ and compare their accuracy with the confidence intervals constructed under

the asymptotic theory, with particular focus on the small sample case.

4.5 Simulation study comparing the estimation procedures

4.5.1 Methods

We compare the bootstrap approach with the asymptotic theory by generating confidence intervals for η using simulated data from a range of distributions and determining the coverage of these confidence intervals. In particular, we consider data sampled from Normal, Cauchy, Uniform, Normal mixtures, Skew Normal, Log-Normal, Folded Normal and Exponential populations. For both η and ζ we simulate data from each distribution with a relatively small sample size ($n = 15$) and a more substantial sample size ($n = 100$). In each case we calculate the 95% confidence intervals using the asymptotic theory given in Theorem 2.3 and Theorem 4.1 (using a Normal distribution), and calculate a competing confidence interval using bootstrapping with 1000 replicates by applying equations (4.4) and (4.5). For each of these approaches we repeat the process $10,000$ times to determine the coverage, that is, the number of confidence intervals that contain the true value. Table 4.1 summarises the simulation procedure.

Step 1	Simulate samples of size $n = 15$ and $n = 100$ from each of the distributions.
Step 2	Calculate confidence intervals for η and ζ using the asymptotic theory in Theorem 2.3 and Theorem 4.1 and bootstrap confidence intervals using equations (4.4) and (4.5).
Step 3	For each method, determine whether the confidence interval contains the true value of η and ζ.
Step 4	Repeat **Steps 1-3** $10,000$ times and report coverage for each method.

Table 4.1: Step by step guide to the simulation study comparing the coverage of the confidence intervals.

4.5.2 Results

Table 4.2 shows the coverage results for η and ζ with sample size $n = 100$. In this case it appears that the Normal asymptotic theory using $\widehat{\eta}$ does a reasonable job of capturing the distribution of $\widehat{\eta}$, with the coverage not deviating too far from the expected value of 0.95 in the conventional cases. That is, the coverage is close to 0.95 for the Normal and Normal mixtures and is still

reasonably close for the Cauchy distribution and even the Exponential distribution. However, the confidence intervals based on $\widehat{\eta}$ do perform poorly for several of the more pathological cases, with a very low coverage of 0.88 and 0.85 for the Uniform and Folded Normal distributions and a very high coverage of 0.99 for the Log-Normal distribution. Furthermore, it is clear that the confidence intervals are not improved when using $\widehat{\zeta}$ with asymptotic theory. It is also worthy of note that in this case, when the data are exponentially distributed, the confidence intervals constructed using $\widehat{\zeta}$ never contain the true value of $\eta = 1$. This is because, when the confidence intervals are constructed in this fashion, the confidence intervals are always entirely between -1 and 1.

Using bootstrapping to construct the confidence intervals for $\widehat{\eta}$ produces confidence intervals which are too small with coverage around 0.91 for most cases. This reflects the concerns raised by Hall and Efron regarding the use of the bootstrap to estimate the sampling distribution of the correlation coefficient. On the other hand, using the variance stabilising Z-transformation and considering $\widehat{\zeta}$ along with bootstrapping appears to produce very good results in most cases. Indeed, the coverage is very close to 0.95 on the whole, with the exception of the uniform distribution and the heavily skewed distributions.

Table 4.3 reports the coverage of the 95% confidence interval for η and ζ using a much smaller sample of $n = 15$. At first glance it appears a similar story to the large sample case. Indeed, using $\widehat{\eta}$ with asymptotic theory seems to perform reasonably well on the whole, particularly for the Normal and Cauchy distributions and mildly skewed Normal mixtures. In the other cases the confidence intervals are generally overly conservative, with the exception of the Folded Normal distribution. By contrast, using $\widehat{\zeta}$ with bootstrapping produces confidence intervals which are too small, with coverage around 0.91 for most of the conventional distributions. An advantage of the confidence intervals generated by $\widehat{\zeta}$, which goes undetected by only considering the coverage, is that it produces suitably asymmetric confidence intervals for the asymmetric distributions and will never produce a confidence interval containing impossible values outside of the range of $[-1, 1]$.

Unsurprisingly, using the asymptotic variance of $\widehat{\zeta}$ generates very poor confidence intervals

in every case considered here. Also, using bootstrapping with $\widehat{\eta}$ generates extremely erratic confidence intervals which are frequently too short.

The results of this section show that, even for rather small samples, applying the asymptotic theory $\widehat{\eta}$ produces the best approximate confidence intervals in terms of coverage. However, these confidence intervals have two notable disadvantages. Firstly, the confidence intervals constructed using asymptotic theory are necessarily symmetric and therefore fail to capture the asymmetry in the small sample distribution. Secondly, it is possible to construct confidence intervals which include values outside the range of $[-1, 1]$. In these cases we have shown that confidence intervals constructed using $\widehat{\zeta}$ and bootstrapping resolve these issues, but do not attain the same level of coverage and generally appear to be too short.

For the symmetric populations under consideration here, the best coverage results are obtained by using the asymptotic variance for $\widehat{\eta}$, irrespective of sample size. Hence, it is recommended to use this approach to test for symmetry in small or large samples, however, it is imperative that the sample is large enough to provide an accurate estimate of the underlying density or distribution function.

However, when η is close to the -1 or 1 the sampling distribution of $\widehat{\eta}$ is appreciably skewed, therefore, more care should be taken when constructing confidence intervals. Nevertheless, when the sample size is reasonably large ($n > 30$), one can still apply the asymptotic theory directly on $\widehat{\eta}$. For smaller samples, based on the coverage results in Table 4.3, it is still recommended to apply the asymptotic theory for $\widehat{\eta}$, truncating the confidence intervals at -1 or 1 if necessary. However, in these cases, one should also check how discrepant the results are to the $\widehat{\zeta}$ approach using bootstrapping. If the two approaches are excessively disparate then one might conclude that there is insufficient data to accurately assess the asymmetry and, thus, further research with bigger samples is necessary. In the next section we reinforce these ideas by considering a couple of examples of calculating $\widehat{\zeta}$ using real data.

η	Distribution	Asymptotic		Bootstrapping	
		$\widehat{\eta}$	$\widehat{\zeta}$	$\widehat{\eta}$	$\widehat{\zeta}$
0	Normal	0.959	0.971	0.896	0.936
0	Cauchy	0.966	0.973	0.924	0.950
0	Uniform	0.883	0.957	0.643	0.763
0.1	NM1	0.959	0.971	0.906	0.944
0.2	NM2	0.958	0.968	0.909	0.944
0.3	NM3	0.964	0.972	0.918	0.946
0.4	NM4	0.969	0.969	0.918	0.941
0.5	SN5	0.961	0.951	0.904	0.918
0.6	SN6	0.964	0.937	0.902	0.899
0.7	SN7	0.962	0.917	0.906	0.869
0.8	SN8	0.953	0.870	0.915	0.811
0.91	Log-Normal	0.990	0.920	0.970	0.783
0.95	Folded Normal	0.850	0.365	0.887	0.245
1	Exponential	0.969	0.000	0.907	0.000

Table 4.2: Coverage of 95% confidence intervals for η based on $10,000$ samples of size $n = 100$. Confidence intervals are constructed using $\widehat{\eta}$ and $\widehat{\zeta}$, estimating the variance using asymptotic theory and bootstrapping with $R = 1000$ replicates.

η	Distribution	Asymptotic		Bootstrapping	
		$\widehat{\eta}$	$\widehat{\zeta}$	$\widehat{\eta}$	$\widehat{\zeta}$
0	Normal	0.953	1.000	0.766	0.911
0	Cauchy	0.960	1.000	0.826	0.952
0	Uniform	0.906	1.000	0.571	0.769
0.1	NM1	0.954	0.999	0.771	0.913
0.2	NM2	0.963	0.997	0.780	0.913
0.3	NM3	0.971	0.993	0.785	0.905
0.4	NM4	0.974	0.987	0.789	0.903
0.5	SN5	0.951	0.961	0.759	0.882
0.6	SN6	0.943	0.937	0.755	0.865
0.7	SN7	0.942	0.904	0.768	0.845
0.8	SN8	0.926	0.832	0.819	0.822
0.91	Log-Normal	0.979	0.866	0.950	0.846
0.95	Folded Normal	0.885	0.553	0.855	0.550
1	Exponential	0.959	0.000	0.952	0.000

Table 4.3: Coverage of 95% confidence intervals for η based on $10,000$ samples of size $n = 15$. Confidence intervals are constructed using $\widehat{\eta}$ and $\widehat{\zeta}$, estimating the variance using asymptotic theory and bootstrapping with $R = 1000$ replicates.

4.6 Real data example

4.6.1 A small data set

We begin by considering a small set of data from an observational study investigating the use of parathyroid hormone (PTH) to predict the onset of hypocalcemia after a thyroidectomy. This is part of a collection of studies on the topic collated by Noordzij et al. [85], which we will introduce in the next chapter. In particular, we consider the preoperative PTH levels in patients who did not go on to develop hypocalcemia in the study by Lam and Kerr [71]. Figure 4.5 shows the histogram of the data, which appears to demonstrate some asymmetry to the right.

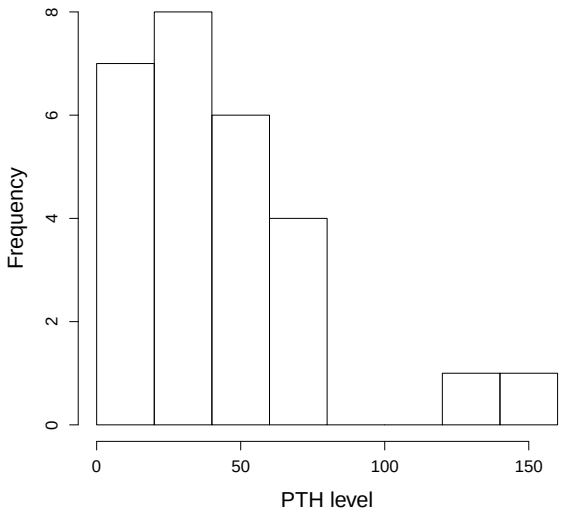

Figure 4.5: Histogram of the preoperative PTH levels in patients who did not go on to develop hypocalcemia in the study by Lam and Kerr [71].

Now, suppose that the main interest is to measure the asymmetry in this particular set of data, for example, one might require an estimate of the asymmetry to inform or guide a normalising transformation. For this particular set of data there are only 27 individuals so we may expect that the confidence interval for η is not likely to be especially precise. Let η_s be the measure of asymmetry η in this case. If we estimate η_s for this particular data set we obtain

$$\widehat{\eta}_s = 0.69,$$

which suggests quite substantial asymmetry to the right. However, the confidence interval for η_s constructed using the asymptotic theory given in Theorem 2.3, is given by

$$[-0.09; 1.47].$$

The first thing to observe is that this confidence interval is very wide but, more than that, the confidence interval includes numbers outside the range $[-1, 1]$. As we have already mentioned η_s is a correlation coefficient, therefore, it is impossible that the value of η_s could fall outside $[-1, 1]$. Thus, it is clear in this case that asymptotic theory has broken down and that the constructed confidence interval has lost some of its meaning. By truncating the interval at 1 we obtain the approximate confidence interval

$$[-0.09; 1].$$

However, if we apply the new small sample approach we find that

$$\widehat{\zeta}_s = Z\left(\widehat{\eta}_s\right) = 0.84,$$

and the confidence interval for $\zeta_s = Z(\eta_s)$, constructed using equation (4.4) and 1000 bootstrap replicates, is

$$[0.05; 1.63].$$

Now, transforming back using the inverse transformation $Z^{-1}(x) = \tanh(x)$ we obtain the following confidence interval for η_s,

$$[0.05; 0.93].$$

Hence, by using the transformed measure $\widehat{\zeta}$ along with a bootstrap estimate of the variance we have produced a slightly shorter approximate confidence interval. However, recall that the confidence intervals constructed using this method are generally too short. Since the result is not overly dissimilar to the truncated confidence interval constructed using the asymptotic theory, and appealing to the results of the simulation study, we conclude that $[-0.09; 1]$ is a reasonably

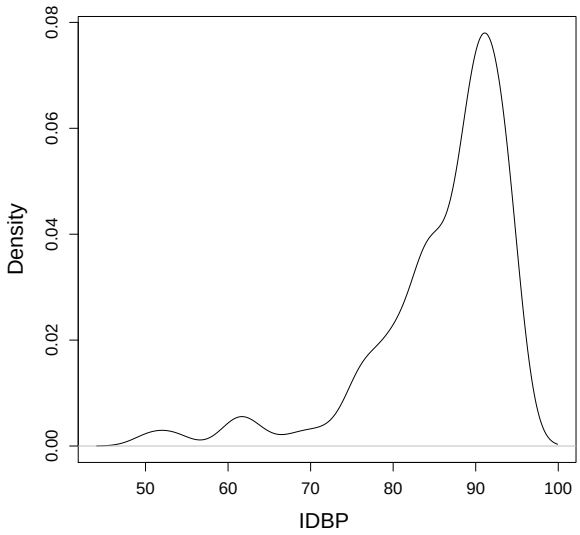

Figure 4.6: Density estimate of the initial diastolic blood pressure (IDBP) in the treatment group of the COOP trial.

accurate 95% confidence interval. Clearly, however, there is large uncertainty in the findings and further research with larger samples is required to produce more precise results.

4.6.2 A large trial

We now revisit the hypertension data collated by Wang et al. [125] to investigate the effectiveness of $\widehat{\zeta}$ for measuring asymmetry in larger data sets. In particular, we focus on the initial diastolic blood pressure in the treatment group of the COOP trial ($n = 150$). Figure 4.6 shows the density estimate of the data, which this time demonstrates asymmetry to the left.

In this case, as the sample size is larger, we expect that the asymptotic theory will be more effective at approximating the confidence interval for η. Let η_l denote the measure of asymmetry in the underlying population. It is readily calculated that

$$\widehat{\eta}_l = -0.83,$$

which indicates substantial asymmetry to the left. Moreover, the confidence interval for η_l is

given by

$$[-0.99; -0.66].$$

On the other hand, by applying the transformed method we obtain $\widehat{\zeta}_l = -1.2$ and the confidence interval for $\zeta_l = Z(\eta_l)$, constructed using bootstrapping with 1000 bootstrap replicates, is

$$[-1.42; -0.98].$$

Once again, if we transform back to acquire a confidence interval for η_l then we obtain

$$[-0.89; -0.75].$$

Hence, even in the large sample case we obtain a much tighter confidence interval. However, this is certainly due in part to the fact that the coverage of the 95% confidence intervals constructed using the $\widehat{\zeta}$ approach is lower than 0.95. Regardless, it is clear from either confidence interval that large asymmetry exists in this data.

4.7 Discussion

In this chapter we revealed the limitations of applying $\widehat{\eta}$ to small samples. Principally, we showed that the asymptotic Normal distribution does not provide a good approximation. This was the motivation for a transformed measure $\widehat{\zeta}$ which appeared to follow a Normal distribution more closely for small samples. We also introduced and discussed the bootstrap to facilitate the construction of confidence intervals for this new measure, which appear to be fairly accurate for the more conventional distributions considered here. We carried out a simulation study to compare the accuracy of the confidence intervals for η using bootstrapping and asymptotic theory. It was shown that, on the whole, confidence intervals constructed directly using $\widehat{\eta}$ with variance estimation based on the asymptotic theory provide the best the confidence intervals in terms of coverage. On the other hand, $\widehat{\zeta}$ can be used to generate confidence intervals for η which are more meaningful, in the sense that they are suitably asymmetric and do not extend beyond the interval $[-1;\ 1]$. We proposed an effective compromise, namely, to generate the

confidence intervals using $\widehat{\eta}$, but truncate the confidence limits at -1 or 1. This approach attains reasonable accuracy in terms of coverage. However, it is constructed by assuming that the sampling distribution of $\widehat{\eta}$ is approximately Normal, an untenable assumption when η is close to -1 or 1 and the sample size is small. Finally, we compared the effectiveness of the two procedures using two real data sets, one small and one large. It was shown that, despite discrepant upper and lower bounds, the conclusions about asymmetry were very similar irrespective of which approach was used to derive the confidence intervals. In the first, both confidence intervals revealed substantial uncertainty indicating the need for more data; in the second, both gave strong evidence of a large degree of asymmetry.

In summary, for the symmetric populations the best coverage results are obtained by using the asymptotic variance for $\widehat{\eta}$, irrespective of sample size. If the principal interest is to test for symmetry, then one is only really concerned with the sampling distribution under the null hypothesis of symmetry. Hence, for testing symmetry it is recommended to use $\widehat{\eta}$ with the asymptotic variance estimate, regardless of the sample size. Nonetheless, to ensure an accurate estimate of η it is imperative that the sample is large enough to provide an accurate estimate of the underlying density or distribution function.

However, it was shown that when η is close to the -1 or 1 the sampling distribution is appreciably skewed, therefore more care should be taken when constructing confidence intervals. When the sample size is reasonably large ($n > 30$), it is recommended that one applies the asymptotic theory directly on $\widehat{\eta}$. For smaller samples, it is suggested to apply the asymptotic theory for $\widehat{\eta}$, truncating the confidence intervals at -1 or 1 if necessary. However, one should also check how discrepant the results are to the $\widehat{\zeta}$ approach using bootstrapping, and, if the two approaches are unreasonably incongruent then one can conclude that there are too few samples to accurately assess the asymmetry in the population.

In conclusion, the methods discussed in this chapter provide a useful alternative for small samples, however, there are a number of limitations which must be addressed. Indeed, the problem of constructing accurate confidence intervals for η in the small sample setting has not been satisfactorily resolved. In particular, the issues encountered are similar to those identified

by Hall [44] in the context of constructing bootstrap confidence intervals for the correlation coefficient. Hence, the problem of drawing inferences from small samples remains something of a 'smoking gun' for η, and something that merits further research. It was shown that, while using bootstrapping with $\widehat{\zeta}$ succeeds in producing asymmetric confidence intervals, they are generally too narrow. As a result, for small samples, the recommendation is to prioritise the original asymptotic approach, truncating the confidence intervals at -1 or 1 if necessary.

Another limitation is the fact that we have only considered using non-parametric bootstrapping with the simple confidence intervals given in equation (4.4). In fact, there are a huge variety of methods for performing bootstrapping, as well as the subsequent calculation of confidence intervals. For instance, other bootstrap methods include the parametric bootstrap or the smoothed bootstrap [27]. Additionally, it is also possible to construct confidence intervals that adjust for bias [29]. However, these procedures are often more computationally intensive and, consequently, their implementation is not so widespread.

In the next chapter we discuss how one can apply $\widehat{\eta}$ and $\widehat{\zeta}$ to carry out a meta-analysis of η with the aim of comparing the distribution of data across several studies.

Recommendations for applying $\widehat{\eta}$:

- For large samples ($n \geq 30$) the asymptotic theory for $\widehat{\eta}$ seems to produce accurate confidence intervals across a range of distributions.
- Even for small samples ($n < 30$) the asymptotic theory for $\widehat{\eta}$ seems to produce the most accurate confidence intervals in terms of coverage, despite the fact that the sampling distribution of $\widehat{\eta}$ is somewhat more skewed than $\widehat{\zeta}$.
- Confidence intervals for η, which are constructed using $\widehat{\zeta}$, allow for asymmetry in the sampling distribution of $\widehat{\eta}$ when η is close to -1 or 1. Further, they are guaranteed to be between $[-1, 1]$, but do not perform as well in terms of coverage.
- On balance, it is recommended to calculate confidence intervals using the asymptotic theory of $\widehat{\eta}$ and appropriately crop the confidence intervals that extend beyond the

range $[-1, 1]$.

- For symmetric distributions the sampling distribution of $\widehat{\eta}$ appears to be approximately Normal, even for samples as small as $n = 15$. Therefore, in the context of testing for symmetry, p-values should not be too badly compromised when the sample size is small.

- Irrespective of what method we use, it is imperative that there are enough sample points to reasonably estimate the density function f and the distribution function F. The requisite number of points for this will vary depending on the shape of the underlying population.

- If confidence intervals are very wide then it indicates that further data collection is required to enable more precise inferences.

Chapter 5

Analysing the asymmetry of data across several studies

5.1 Introduction

As shown in the previous chapter, a common problem when dealing with small samples is that it can often be difficult to assess the distributional properties of the data. For example, for small samples the data may appear to be considerably asymmetric or fail to conform to a Normal distribution, but this could be due to chance. This, in turn, leads to uncertainty about whether or not it is appropriate to transform the data. Indeed, in medical statistics it is not uncommon that, for a number of small studies which are investigating the same treatment or intervention, different authors come to disparate conclusions about the appropriate scale [54]. When individual patient data are available, one might be tempted to pool the data across all studies to assess the underlying distribution. However, in this chapter we show that this approach can produce misleading results, and propose alternative methodologies.

In this chapter we consider how to examine and combine asymmetry estimates of a particular distribution, when multiple studies are available. In section 5.2 we introduce the fixed effect and random effects meta-analysis models before conducting a brief review of the existing literature on the meta-analysis of correlation studies. In section 5.3 we investigate measuring the asymmetry of data across several studies with the aim of obtaining an 'overall' measure of the asymmetry

in the underlying population. In particular, we explore the potential pitfalls of simply pooling the data from every study, before proposing a meta-analysis of the asymmetry measure η to combat this problem. In section 5.4 we examine employing the meta-analysis of η on a much smaller data set, a collection of comparatively small observational studies. In this case, we show that a meta-analysis of ζ can be used to obtain more meaningful conclusions. In section 5.5 we extend this technique to identify distributional differences between the treatment and control arms across several studies. In section 5.6 we summarise the results of the chapter.

Aims of the chapter:

- Investigate the problem of measuring the asymmetry in a population with several samples from heterogeneous studies.
- Identify the dangers of simply pooling studies together to investigate asymmetry.
- Develop a meta-analysis approach to solve this problem using the measures of asymmetry, $\widehat{\eta}$ and $\widehat{\zeta}$.
- Explore the possible applications of these techniques in small and large data sets.

5.2 Meta-analysis

5.2.1 Fixed effect meta-analysis

As the name suggests, a meta-analysis is an analysis of analyses. Glass [40] defines a meta-analysis as "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings." In essence, a meta-analysis seeks to combine the results of several independent studies that are considered to be similar enough to be synthesised in a meaningful manner. In the medical literature it is commonplace for a number of different studies to investigate the effect of the same treatment or intervention. This provides something of a dilemma to a clinician as studies can often give different or even contradictory results. A meta-analysis provides a way to collate the results of a number of studies in a systematic way

and provide a single overall treatment effect estimate, which is of practical use to a clinician. An obvious advantage of carrying out a meta-analysis is that it increases the available sample size, which reduces the uncertainty in the treatment effect estimate and increases statistical power. It also allows for investigation of the similarities and differences between individual studies and their effect sizes.

Suppose there are k studies investigating a specific treatment or intervention. Further, suppose that each study reports the treatment effect estimate $\widehat{\theta}_i$ and its variance σ_i^2 for $i = 1, \cdots, k$. The fixed effect model assumes homogeneity across studies and, more specifically, that each study is estimating the same effect θ. Therefore, in the parametric framework detailed by Whitehead and Whitehead [127], the fixed effect model has the form

$$\widehat{\theta}_i \sim N(\theta, \sigma_i^2).$$

The first challenge when synthesising treatment effect estimates from multiple studies is to account for the different levels of uncertainty in each study. Typically, some studies will have a larger sample size than others and so the each individual estimate has a different standard error. The inverse-variance method provides greater weighting to the estimates with a smaller variance,

$$w_i = \frac{1}{\sigma_i^2},$$

where σ_i^2 is the variance of $\widehat{\theta}_i$. The maximum likelihood estimate of the overall effect estimate is the following weighted average of the individual treatment effect estimates $\widehat{\theta}_i$,

$$\widehat{\theta} = \frac{\sum_{i=1}^{k} w_i \widehat{\theta}_i}{\sum_{i=1}^{k} w_i}.$$

The overall effect estimate $\widehat{\theta}$ has variance

$$\text{Var}\left(\widehat{\theta}\right) = \frac{1}{\sum_i w_i}.$$

Hence, an approximate $100(1-\alpha)\%$ confidence interval for θ is given by

$$\widehat{\theta} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{1}{\sum_i w_i}}.$$

As we have mentioned, $\widehat{\theta}$ is the maximum likelihood estimate of the treatment effect and as such, it is asymptotically unbiased and asymptotically Normal. In this way, $\widehat{\theta}$ provides the best possible unbiased estimate of the treatment effect θ, using the evidence from all studies. However, the estimate $\widehat{\theta}$ is based on the assumption that all of the studies are estimating the same underlying treatment effect θ.

Many consider the fixed effect model to be an over-simplification. For example, DerSimonian and Laird [25] identify that, given the many potential sources of heterogeneity between trials (e.g. location, dosage, characteristics of the patients), it is unlikely that all the trials are measuring the same treatment effect. A more realistic assumption is that each of the trials are estimating a slightly different treatment effect, which itself is randomly drawn from a Normal population. This gives rise to random effects meta-analysis.

5.2.2 Random effects meta-analysis

When there is evidence of between study heterogeneity, one can modify the meta-analysis to account for the variability between studies. A random effects meta-analysis assumes that the underlying treatment effects θ_i vary across studies, following a Normal distribution with mean θ and variance τ^2. The model can be expressed as

$$\widehat{\theta_i} \sim N(\theta_i, \sigma_i^2),$$

where

$$\theta_i \sim N(\theta, \tau^2).$$

Now, to estimate the overall effect one must account for the within study variances σ_i^2 and the between study variance τ^2. Hence the maximum likelihood estimate of θ in the random effects

model is

$$\widehat{\theta}_r = \frac{\sum_{i=1}^{k} w_i^* \widehat{\theta}_i}{\sum_{i=1}^{k} w_i^*},$$

where

$$w_i^* = \frac{1}{\sigma_i^2 + \tau^2}.$$

A number of methods have been proposed to estimate the between study variance τ^2. For example, the most common approach is given by DerSimonian and Laird [25]. They propose the following unbiased, method of moments estimate for τ^2,

$$\widehat{\tau}^2 = \frac{Q - (k-1)}{\sum_i w_i - \frac{\sum_i w_i^2}{\sum_i w_i}},$$

where

$$Q = \sum_{i=1}^{k} w_i \left(\theta_i - \widehat{\theta}\right)^2,$$

where $\widehat{\theta}$ is the overall effect estimate from the fixed effect meta-analysis. Another useful measure for assessing the amount of between study variation is

$$I^2 = 100\% \times \frac{Q - (k-1)}{Q}.$$

The measure, proposed by Higgins and Thompson [53], describes the percentage of the total variation which is due to between study heterogeneity. This can be observed by noting that

$$I^2 = 100\% \times \frac{\widehat{\tau}^2}{\widehat{\tau}^2 + \widehat{\sigma}^2},$$

where $\widehat{\tau}^2$ is the DerSimonian and Laird estimate of τ^2 and

$$\widehat{\sigma}^2 = \frac{(k-1)\sum_i w_i}{\left(\sum_i w_i\right)^2 - \sum_i w_i^2},$$

is an estimate of a 'typical' within study variance,

The overall effect estimate $\widehat{\theta}_r$ has variance

$$\text{Var}\left(\widehat{\theta}_r\right) = \frac{1}{\sum_i w_i^*}.$$

Hence, an approximate $100(1-\alpha)\%$ confidence interval for θ is given by

$$\widehat{\theta}_r \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{1}{\sum_i w_i^*}}.$$

The random effects approach assumes a distribution of true treatment effects across studies and $\widehat{\theta}$ provides an estimate of the average treatment effect across all studies. One must be mindful of the amount of heterogeneity between studies when interpreting this 'average' effect, as individual studies may have a true treatment effect which is far from the average value. As a result, some authors also suggest reporting a prediction interval for a new study treatment effect [100, 55]. The $100(1-\alpha)\%$ prediction interval is given by

$$\widehat{\theta}_r \pm t_{k-2;\frac{\alpha}{2}} \sqrt{\widehat{\tau}^2 + \text{Var}\left(\widehat{\theta}_r\right)},$$

where $\widehat{\tau}^2$ is the estimate of the between study heterogeneity.

5.2.3 Meta-analysis of a correlation coefficient

Our interest is to conduct a meta-analysis of $\widehat{\eta}$, which is effectively a sample correlation coefficient. Indeed, recall that

$$\eta = \text{Cor}(f(X), F(X)),$$

while the estimate $\widehat{\eta}$ is the corresponding sample correlation coefficient. There is a wealth of literature on the meta-analysis of correlation coefficients, something which is particular relevant in psychology studies. For example, Hunter and Schmidt [58] propose carrying out a meta-analysis on the raw correlation coefficient. As noted by Pigott et al. [92], it is common practice to assume that the variance of the study specific correlation r_i is known and is given by

$$v_i = \frac{(1-r_i^2)^2}{n_i},$$

where n_i is the within study sample size. However, as observed by Hedges and Olkin [51], the more conventional approach is to carry out a meta-analysis after applying the Fisher Z-transformation to the correlation coefficients. In this case the within study variance of the Z-transformed correlations is given by

$$v_i = \frac{1}{n_i - 3}.$$

Field [34] carries out an extensive Monte Carlo comparison of these meta-analysis approaches and concludes that both methods break down for a small number of studies and sample sizes. In particular, the Schmidt-Hunter method underestimates the summary correlation, whilst the Hedges-Olkin method tends to over estimate the summary correlation when the true correlation is greater than 0.5.

In the next section we investigate the problem of measuring asymmetry in data across several studies with the aim of obtaining an 'overall' measure of asymmetry in the underlying population. In particular, we explore the potential pitfalls of simply pooling data across all studies, before proposing a meta-analysis on the asymmetry measure $\widehat{\eta}$ to combat this problem.

5.3 Meta-analysis of $\widehat{\eta}$ to quantify asymmetry of a random variable across multiple studies

When testing for symmetry, for example, in a covariate or response variable across multiple studies it is important to note that it is not sufficient to pool the data and test for symmetry assuming the data are all from a single trial. Rather, one should evaluate the symmetry in each of the trials separately. The reason for this is that, even when the data are sampled from several asymmetric populations, the pooled samples can appear to be symmetric, and thus give a misleading impression of symmetry. This phenomena is exhibited by revisiting the hypertension data. In particular, we determine the value of $\widehat{\eta}$ for the initial diastolic blood pressure data in the treatment group for each trial separately as well as the value of $\widehat{\eta}$ obtained after naively pooling the data from all trials. The values of $\widehat{\eta}$ can be seen in Table 5.1.

What we observe seems somewhat counter intuitive. The data are asymmetric in the majority

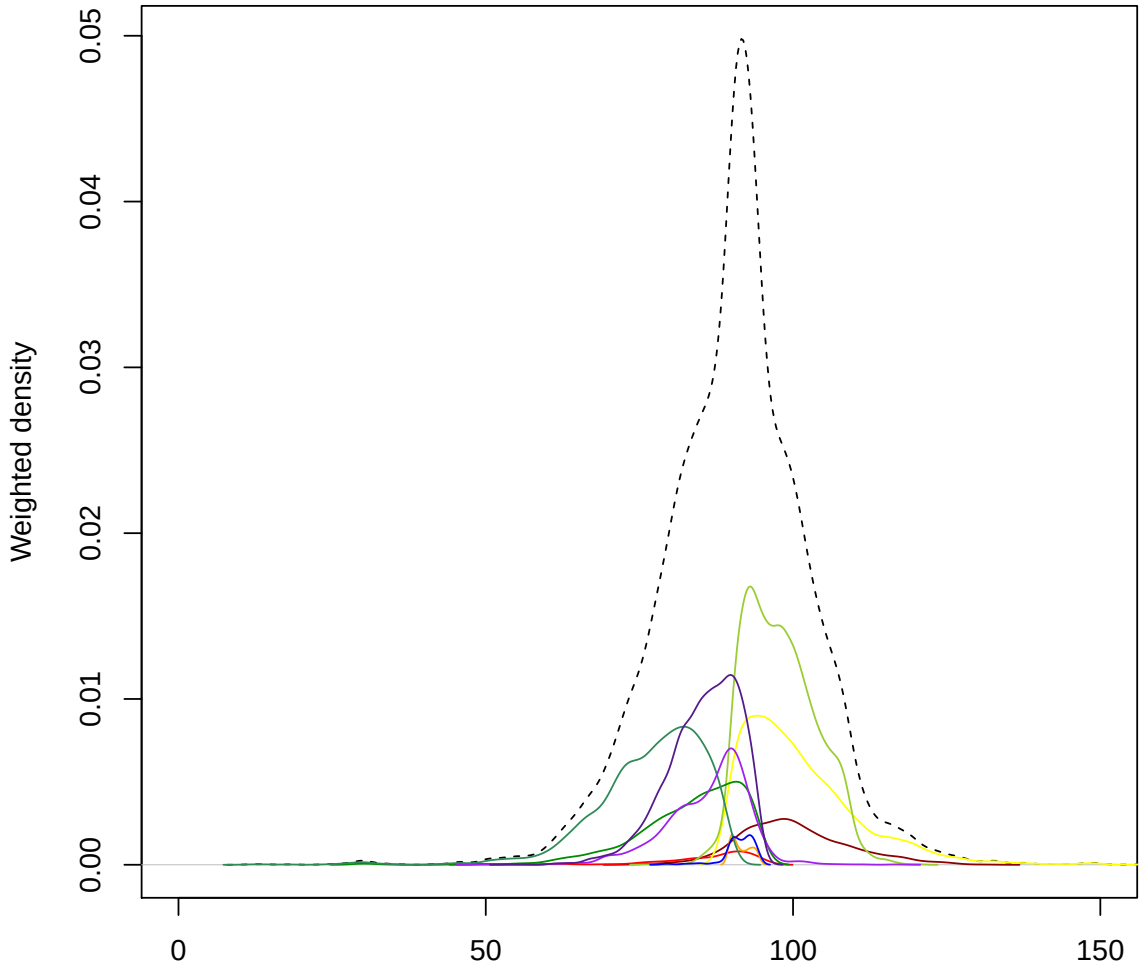

Figure 5.1: Density estimates of individual trials weighted with respect to their sample size (coloured) and the density estimate of pooled data (black-dashed) naively assuming all the data are from one 'mega' trial.

Trial	$\widehat{\eta}$	95% CI	s.e.	n
ANBP	0.27	$[0.16; 0.38]$	0.06	780
COOP	-0.83	$[-0.99; -0.66]$	0.08	150
EWPH	0.84	$[0.44; 1.24]$	0.20	90
HDFP	0.81	$[0.75; 0.86]$	0.02	2427
MRC1	0.51	$[0.45; 0.58]$	0.03	3546
MRC2	-0.80	$[-0.87; -0.73]$	0.03	1314
SHEP	-0.63	$[-0.69; -0.57]$	0.03	2365
STOP	-0.51	$[-0.88; -0.13]$	0.19	137
SYCH	-0.50	$[-0.57; -0.43]$	0.04	1252
SYSE	-0.59	$[-0.66; -0.53]$	0.03	2398
Naive pooled result	-0.08	$[-0.10; -0.07]$	0.01	14459

Table 5.1: The value of $\widehat{\eta}$ with 95% confidence intervals (CI) for initial diastolic blood pressure data in the treatment group, along with the standard error (s.e.) of $\widehat{\eta}$. The results are given for each trial separately and also by pooling all the trials together assuming all the data are from one 'mega' trial.

of individual cases, for example, for the COOP and MRC2 trials $\widehat{\eta} = -0.83$ and $\widehat{\eta} = -0.80$ respectively, signifying these trials are highly skewed to the left. On the other hand, for the EWPH and HDFP trials $\widehat{\eta} = 0.84$ and $\widehat{\eta} = 0.81$ respectively, indicating these trials are highly skewed to the right. However, upon pooling the data we obtain a very symmetric set of data with $\widehat{\eta} = -0.08$. This apparent 'paradox' is resolved in Figure 5.1, which shows the individual study specific density estimates as well as the density estimate obtained by naively pooling all the data together. Further, the individual density estimates have been weighted according to their respective sample size so that one obtains an appreciation of how much each trial contributes to the overall pooled data set. We observe that, because the different skewed trials have different means, when the data are pooled we observe a near symmetric density curve. That is to say, the asymmetry inherent in individual trials is masked when pooling the data together. The upshot of this is that when evaluating the symmetry or normality of the data, it is insufficient to pool the data and test for symmetry. Instead, it is better practice to evaluate each trial separately. For further details on why it is inappropriate to ignore clustering when pooling data across multiple studies refer to Abo-Zaid et al. [1].

This motivates a more formal, more statistically rigorous, meta-analysis of $\widehat{\eta}$ using the meth-

ods detailed in section 5.2.1 and 5.2.2, with the aim of determining the size of asymmetry in the underlying distribution. The meta-analysis methods in section 5.2.1 and 5.2.2 require, for each study, an estimate of the measure of interest along with its variance.

Recall from Theorem 2.3 in Chapter 2 that

$$\sqrt{n} \cdot [\widehat{\eta} - \eta] \xrightarrow{\mathrm{L}} N(0, \sigma^2),$$

where n is the sample size and

$$\sigma^2 = \mathrm{Var}\left[\frac{2}{\sqrt{\nu_f \nu_F}}\left(f(X)F(X) - \frac{1}{2}f(X)\right) + \int_X^\infty \frac{f(y)^2}{\sqrt{\nu_f \nu_F}}dy \right.$$
$$\left. + \eta\left\{\frac{\left(F(X) - \frac{1}{2}\right)^2}{2\nu_F} + \frac{(f(X) - \mu_f)^2}{2\nu_f} + \int_X^\infty \frac{\left(F(y) - \frac{1}{2}\right)}{\nu_F} f(y)dy + \frac{(f(X) - \mu_f)f(X)}{\nu_f}\right\}\right],$$

where ν_f and ν_F denote $\mathrm{Var}(f(X))$ and $\mathrm{Var}(F(X))\left(= \frac{1}{12}\right)$ respectively, and $\mu_f = \mathrm{E}[f(X)]$. Hence, we have that approximately

$$\mathrm{Var}\left(\widehat{\eta}\right) = \frac{\sigma^2}{n}.$$

Further, recall that one can estimate σ^2 from the data using

$$\widehat{\sigma}^2 = \widehat{\mathrm{Var}}(Y) = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2, \tag{5.1}$$

where

$$Y_i = \left[\frac{2}{\sqrt{\widehat{\nu}_f \widehat{\nu}_F}}\left(\widehat{f}(X_i)\widehat{F}(X_i) - \frac{1}{2}f(X_i)\right) + \frac{1}{\sqrt{\widehat{\nu}_f \widehat{\nu}_F}}\widehat{\Phi}_1(X_i)\right.$$
$$\left. + \widehat{\eta}\left\{\frac{\left(\widehat{F}(X_i) - \frac{1}{2}\right)^2}{2\widehat{\nu}_F} + \frac{\left(\widehat{f}(X_i) - \overline{\widehat{f}}\right)^2}{2\widehat{\nu}_f} + \widehat{\Phi}_2(X_i) + \frac{\left(\widehat{f}(X_i) - \overline{\widehat{f}}\right)\widehat{f}(X_i)}{\widehat{\nu}_f}\right\}\right],$$

where $\widehat{f}$ and $\widehat{F}$ are the kernel density estimate and the empirical distribution function estimate

respectively. Moreover, recall that $\widehat{\Phi}_1(x)$ is a numerical approximation of the integral

$$\Phi_1(x) = \int_x^\infty f^2(y)dy,$$

using $\widehat{f}(x)$ as an estimate of the curve $f(x)$, and $\widehat{\Phi}_2(x)$ is a numerical approximation of

$$\int_x^\infty \frac{\left(F(y) - \frac{1}{2}\right)}{\nu_F} f(y)dy,$$

estimating f and F by $\widehat{f}$ and $\widehat{F}$.

Hence, provided each study has a sufficiently large sample size we can approximate the standard error using

$$\text{s.e.}\left(\widehat{\eta}\right) = \frac{\widehat{\sigma}}{\sqrt{n}}.$$

As a result, when we have a number of different trials which are composed of samples from the same population, one can conduct a meta-analysis of $\widehat{\eta}$ to identify whether there is a common η to describe the underlying population. In Chapter 4 it was shown that this Normal approximation is only valid for reasonably large samples ($n > 30$). Therefore, ideally we need every study in the meta-analysis to be sufficiently large, however, generally smaller samples will be given a lower weighting.

Suppose we have k independent studies of size n_i from the random variables $X^{(i)}$, for $i = 1, \cdots, k$. That is, let $X_1^{(i)}, X_2^{(i)}, \cdots, X_{n_i}^{(i)}$ be a random sample from $X^{(i)}$. Then each study has a study specific quantity of asymmetry

$$\eta^{(i)} = \text{Corr}\left(f_i\left(X^{(i)}\right), F_i\left(X^{(i)}\right)\right),$$

where f_i and F_i are probability density function and cumulative distribution function for $X^{(i)}$ respectively. The study specific measure of asymmetry $\eta^{(i)}$ is readily estimated from the data using

$$\widehat{\eta}^{(i)} = -\frac{\sum_{j=1}^{n_i} \widehat{f}_i\left(X_j^{(i)}\right)\widehat{F}_i\left(X_j^{(i)}\right) - n_i\overline{\widehat{f}}_i\overline{\widehat{F}}_i}{\sqrt{\left(\sum_{j=1}^{n_i}\left(\widehat{f}_i\left(X_j^{(i)}\right)\right)^2 - n_i\overline{\widehat{f}}_i^2\right)\left(\sum_{i=1}^{n_i}\left(\widehat{F}_i\left(X_j^{(i)}\right)\right)^2 - n_i\overline{\widehat{F}}_i^2\right)}}, \quad \text{for } i = 1, \cdots, k.$$

where $\overline{\widehat{f}}_i = \frac{1}{n_i}\sum_j \widehat{f}_i\left(X_j^{(i)}\right)$, $\overline{\widehat{F}}_i = \frac{1}{n_i}\sum_j \widehat{F}_i\left(X_j^{(i)}\right)$ and $\widehat{f}_i$ and $\widehat{F}_i$ are the study specific kernel density estimate and empirical cumulative distribution function respectively. The optimal bandwidth for the kernel density estimate $\widehat{f}_i$ can be estimated using the the data from study i.

Then, using Theorem 2.3 in Chapter 2 it is readily verified that

$$\widehat{\eta}^{(i)} \sim N\left(\eta_i, \frac{\sigma_i^2}{n_i}\right)$$

approximately, where σ_i^2 is estimated using equation (5.1) using the data in study i. Hence, we have defined the study specific estimate of η and its within study variance. Now, let us specify the meta-analysis model.

The fixed effect model assumes that the σ_i^2 are known and that every estimate $\widehat{\eta}^{(i)}$ is estimating the same underlying measure of asymmetry. That is

$$\eta^{(i)} = \eta, \quad \text{for } i = 1, \cdots, k.$$

Therefore, the fixed effect meta-analysis model is given by

$$\widehat{\eta}^{(i)} \sim N\left(\eta, \frac{\sigma_i^2}{n_i}\right), \tag{5.2}$$

where η denotes the overall measure of asymmetry in the underlying population. The random effects meta-analysis model assumes that the study specific $\eta^{(i)}$ are randomly drawn from a Normal population with mean η^o and variance τ_1^2. That is, we assume

$$\begin{aligned} \widehat{\eta}^{(i)} &\sim N\left(\eta^{(i)}, \frac{\sigma_i^2}{n_i}\right) \\ \eta^{(i)} &\sim N\left(\eta^o, \tau_1^2\right), \end{aligned} \tag{5.3}$$

Study	$\widehat{\eta}$	95% CI	$\%W_f$	$\%W_r$	s.e.	n
ANBP	0.27	$[0.16; 0.38]$	4.94	10.12	0.06	780
COOP	-0.83	$[-0.99; -0.66]$	2.26	10.05	0.08	150
EWPH	0.84	$[0.44; 1.24]$	0.37	9.40	0.20	90
HDFP	0.81	$[0.75; 0.86]$	21.47	10.17	0.02	2427
MRC1	0.51	$[0.45; 0.58]$	15.62	10.16	0.03	3546
MRC2	-0.80	$[-0.87; -0.73]$	12.95	10.16	0.03	1314
SHEP	-0.63	$[-0.69; -0.57]$	16.58	10.16	0.03	2365
STOP	-0.51	$[-0.88; -0.13]$	0.41	9.47	0.19	137
SYCH	-0.50	$[-0.57; -0.43]$	12.03	10.16	0.04	1252
SYSE	-0.59	$[-0.66; -0.53]$	13.38	10.16	0.03	2398
Overall (fixed)	-0.10	$[-0.12; -0.07]$				
Overall (random)	-0.15	$[-0.59; 0.29]$				

Table 5.2: Meta-analysis of $\widehat{\eta}$ for initial diastolic blood pressure data in the treatment group ($I^2 = 99.7\%$). $\%W_f$ and $\%W_r$ denote the percentage weights of each study in the fixed and random effects meta-analysis respectively.

where η^o denotes the overall measure of asymmetry in the underlying population and τ_1^2 is the between study variance in the measure of asymmetry.

Table 5.2 shows the results of a meta-analysis on the values of $\widehat{\eta}$ for initial diastolic blood pressure data in the treatment group, whilst the results are displayed graphically in a forest plot in Figure 5.2. In this case the majority of the trials have a large number of patients, with all but one study having sample size greater than 100.

Unsurprisingly, the studies display a vast amount of heterogeneity. Indeed, the Q statistic rejects homogeneity between studies with a p-value < 0.0001 and the I^2 statistic reveals that 99.7% (with CI $[99.6\%; 99.8\%]$) of the variability in the values of $\widehat{\eta}$ can be attributed to between study heterogeneity. Therefore the fixed effect model (5.2), which gives an overall estimate of $\widehat{\eta} = -0.10$ (with CI $[-0.12; -0.07]$) is not appropriate here. It is more accurate to conclude that different studies are drawing diastolic blood pressure data from different populations (with different values of η). Fitting the random effects model (5.3) suggests that these separate values of η are normally distributed about a mean value, estimated as -0.15 (with CI $[-0.59; 0.29]$). The between study variance (estimated using DerSimonian and Laird's method of moments) is $\widehat{\tau}_1^2 = 0.49$ (with $I^2 = 99.7\%$ $[99.6\%; 99.8\%]$). In fact, the amount of between study variability is so extreme that the approximate 95% prediction interval for the value of η for a new study

is $[-1.85, 1.55]$. This 'prediction interval' includes the entire range of possible range of η, which can only take values between -1 and 1.

It could be argued, after observing the forest plot, that the values of η can be split into two distinct clusters. Indeed, there is a cluster of studies with asymmetry to the left ($\widehat{\eta} < 0$) and another cluster with asymmetry to the right ($\widehat{\eta} > 0$), but there are few samples which appear to be symmetric (with $\widehat{\eta}$ close to 0). This demonstrates that the amount of asymmetry could be drastically different from one setting to another. This motivates further analyses, such as meta-regression and subgroup analyses, to determine the source of this apparent heterogeneity between the studies.

In the next section we investigate applying a meta-analysis of $\widehat{\eta}$ on a much smaller data set, a collection of comparatively small observational studies. In this case, we show that a meta-analysis of $\widehat{\zeta}$ can be useful for obtaining more meaningful conclusions.

5.4 Meta-analysis of $\widehat{\eta}$ for small samples

As we have already discussed, a meta-analysis of $\widehat{\eta}$ is of particular interest when individual studies have small sample sizes, as it can be particularly difficult to determine the underlying distribution in these cases. However, there is a problem with applying a meta-analysis of $\widehat{\eta}$ in this setting. Indeed, the distribution of $\widehat{\eta}$ calculated in Chapter 2 is asymptotic and so requires a reasonably large sample to apply the Normal approximation. In Chapter 4 it was demonstrated that, for a single trial, the Normal approximation breaks down for small samples. In this section we investigate the implications for a meta-analysis of $\widehat{\eta}$.

In order to demonstrate this problem we consider a set of much smaller studies. Rather than dealing with large randomised control trials, we consider a meta-analysis of smaller observational studies. In particular, we use the parathyroid hormone (PTH) data collated by Noordzij et al. [85]. Noordzij et al. gathered data from nine observational studies investigating the use of PTH to make early predictions about the onset of hypocalcemia after a thyroidectomy. Postoperative hypocalcemia is a common complication following a thyroidectomy, but, unfortunately hypocalcemia is not usually present until 24 to 48 hours after surgery. As result, it is common practice to keep patients under observation for this time. However, 70% of these patients will

not go on to develop hypocalcemia. As noted by Noordzij et al., this puts unnecessary strain on healthcare resources if a simple laboratory test is able to accurately classify patients at an early stage. Thus, there is an appreciable interest in the accurate early prediction of hypocalcemia. In the original paper the authors obtained individual patient data (IPD) for 457 patients across the nine studies. All nine studies reported the preoperative PTH level along with the PTH level during at least one of three time periods: 0-20 minutes, 1-2 hours, and 6 hours.

Figure 5.3 shows the histograms of the preoperative PTH levels for patients who did not go on to develop hypocalcemia. Two things are clear from the histograms. Firstly, as we have already mentioned, the number of individuals in each group is very small. Secondly, it is very difficult to tell whether the data are sampled from a symmetric distribution, or a Normal distribution, or indeed whether they are drawn from the same distribution. To attempt to answer these questions we conduct a meta-analysis of $\widehat{\eta}$ to compare and synthesise information about the amount of asymmetry in these studies.

The results of the meta-analysis for $\widehat{\eta}$ on the preoperative PTH data are shown in the top half of Table 5.3 and the forest plot is given in Figure 5.4. The first thing to note is that there appears to be homogeneity in the asymmetry across the studies, with between study variance estimated as $\widehat{\tau}_1^2 = 0$ using DerSimonian and Laird's method of moments. However, there is a problem applying a meta-analysis of $\widehat{\eta}$ in this case. As we have already mentioned, the distribution of $\widehat{\eta}$ determined in Chapter 2 is asymptotic. In Chapter 4 it was demonstrated that, for a single trial, the Normal approximation breaks down for small samples. This has implications for a meta-analysis of $\widehat{\eta}$ and this can lead to erroneous results. Indeed, observe that some of the confidence intervals calculated in Table 5.3 extend beyond the range $[-1, 1]$. For example, for the studies by Lam and Kerr [71] (Lam2003) and Lombardi et al. [74] (Lombardi2004) the 95% confidence interval for η is given by $[-0.03;\ 1.40]$ and $[-0.36;\ 1.20]$ respectively. However, as we have already discussed, η is a correlation coefficient and therefore it is certainly bounded between -1 and 1. In this case, the sample sizes are so small that the Normal approximation is simply not appropriate.

In the previous chapter it was recommended to construct approximate confidence intervals

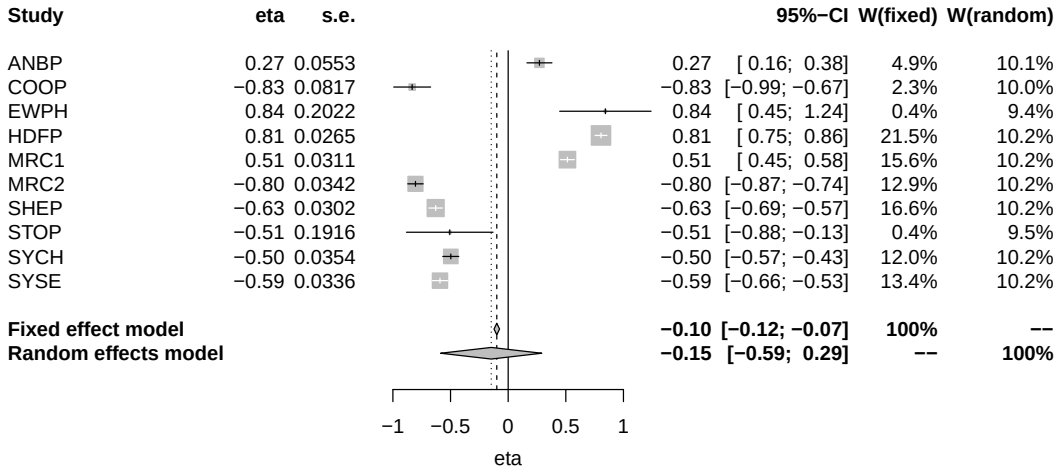

Figure 5.2: Forest plot of the meta-analysis of $\widehat{\eta}$ for initial diastolic blood pressure data in the treatment group ($I^2 = 99.7\%$).

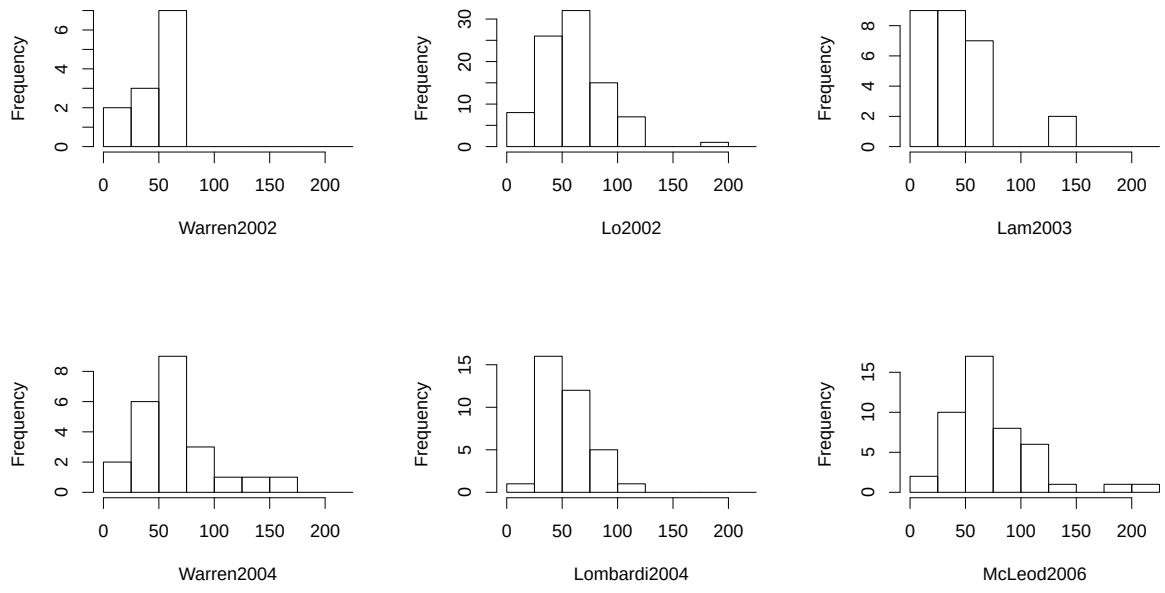

Figure 5.3: Histograms of the preoperative PTH levels in patients who did not go on to develop hypocalcemia for each of the six studies.

Study	$\widehat{\eta}$	95% CI	$\%W_f$	$\%W_r$	s.e.	n
Warren2002	-0.17	$[-0.89; 0.54]$	11.6	11.6	0.36	12
Lo2002	0.34	$[-0.12; 0.81]$	27.33	27.33	0.24	89
Lam2003	0.69	$[-0.03; 1.40]$	11.48	11.48	0.37	27
Warren2004	0.32	$[-0.29; 0.92]$	16.1	16.1	0.31	23
Lombardi2004	0.42	$[-0.36; 1.20]$	9.69	9.69	0.40	35
McLeod2006	0.40	$[-0.10; 0.90]$	23.8	23.8	0.25	46
Overall (fixed)	0.34	$[0.10; 0.58]$				
Overall (random)	0.34	$[0.10; 0.58]$				

Study	$\widehat{\zeta}$	95% CI	$\%W_f$	$\%W_r$	s.e.	n
Warren2002	-0.17	$[-1.19; 0.84]$	5.77	5.77	0.52	12
Lo2002	0.36	$[-0.10; 0.82]$	28.27	28.27	0.23	89
Lam2003	0.84	$[0.06; 1.63]$	9.68	9.68	0.40	27
Warren2004	0.33	$[-0.24; 0.89]$	18.57	18.57	0.29	23
Lombardi2004	0.44	$[-0.30; 1.18]$	10.81	10.81	0.38	35
McLeod2006	0.43	$[-0.04; 0.90]$	26.91	26.91	0.24	46
Overall (fixed)	0.40	$[0.15; 0.64]$				
Overall (random)	0.40	$[0.15; 0.64]$				
η	0.38	$[0.15; 0.56]$				

Table 5.3: Meta-analysis of $\widehat{\eta}$ and $\widehat{\zeta}$ on preoperative PTH levels in patients who did not go on to develop hypocalcemia ($I^2 = 0\%$). $\%W_f$ and $\%W_r$ denote the percentage weights of each study in the fixed and random effects meta-analysis respectively. For the meta-analysis of $\widehat{\zeta}$, the table also displays the overall estimate of η (and the corresponding confidence interval) obtained by applying the inverse Z-transformation.

using $\widehat{\eta}$ using an estimate of the asymptotic variance. However, when the number of observations was small ($n < 30$) it was suggested to suitably truncate any confidence intervals that extended beyond the endpoints -1 and 1. However, this truncation method cannot be easily generalised to a meta-analysis involving these offending studies, as the meta-analysis methods use standard errors (and not upper and lower confidence interval bounds) to determine the weights. Thus, when attempting to synthesise information about the amount of asymmetry η using a meta-analysis, it is especially important to check how the results compare to a corresponding small sample approach, which utilises a standard error estimate appropriate for small samples. Hence, for small samples we propose validating the meta-analysis result for η using a meta-analysis for ζ.

Recall from Theorem 4.1 in Chapter 4 that

$$\sqrt{n} \cdot \left[\widehat{\zeta} - \zeta\right] \xrightarrow{\text{L}} N(0, \upsilon^2),$$

where

$$\upsilon^2 = \frac{\sigma^2}{(1-\eta^2)^2}.$$

Let $\zeta^{(i)}$ denote the study specific Z-transformed measure of asymmetry for study i. Further, let $\widehat{\zeta}^{(i)}$ denote the estimate of $\zeta^{(i)}$ in study i. Then, the fixed effect model for ζ is given by

$$\widehat{\zeta}^{(i)} \sim N\left(\zeta, \frac{\upsilon_i^2}{n_i}\right), \tag{5.4}$$

where ζ denotes the overall measure of symmetry in the underlying population and υ_i^2 denotes the study specific variance.

The analogous random effects meta-analysis model assumes that the study specific $\zeta^{(i)}$ are randomly drawn from a Normal population with mean ζ and variance τ_2^2. That is, we assume

$$\begin{aligned} \widehat{\zeta}^{(i)} &\sim N\left(\zeta^{(i)}, \frac{\upsilon_i^2}{n_i}\right) \\ \zeta^{(i)} &\sim N\left(\zeta, \tau_2^2\right), \end{aligned} \tag{5.5}$$

where ζ denotes the overall measure of asymmetry in the underlying population and τ_2^2 is the between study variance in the measure of asymmetry.

Recall from equation (4.3) in Chapter 4 that the standard error can be approximated using

$$\text{s.e.}\left(\widehat{\zeta}\right) = \frac{\widehat{\sigma}}{\sqrt{n}\,(1-\widehat{\eta}^2)}.$$

However, in the previous chapter we also demonstrated that this estimate is generally very poor. As a result, we may alternatively estimate the variance of $\widehat{\zeta}$ using bootstrapping, which was shown to be more effective. We now redo the analysis for the PTH data using the meta-analysis of $\widehat{\zeta}$, estimating the within study variance $\frac{\upsilon_i^2}{n_i}$ using bootstrapping with $R = 1000$ replications.

The bottom half of Table 5.3 shows the meta-analysis of $\widehat{\zeta}$ for the preoperative PTH levels in individuals who did not go on to develop hypocalcemia, whilst Figure 5.5 shows the forest plot. Once again, there is little evidence of heterogeneity with $\widehat{\tau}_2^2 = 0$ using DerSimonian-Laird, and the overall fixed effect $\widehat{\zeta}$ is given by 0.40 with confidence interval $[0.15, 0.64]$. Transforming back gives an overall asymmetry measure of $\widehat{\eta} = 0.38$ with confidence interval $[0.15, 0.56]$. Comparing this with the meta-analysis directly on $\widehat{\eta}$ in Table 5.3, the confidence interval is slightly narrower. This is not surprising, since in the simulation study in Chapter 4 the confidence intervals were found to be tighter when using $\widehat{\zeta}$ with bootstrapping.

Table 5.4 summarises the results of the two meta-analysis approaches. As expected, the summary confidence interval constructed by carrying out the meta-analysis of $\widehat{\zeta}$ is asymmetric and tighter than the confidence interval constructed using a meta-analysis of $\widehat{\eta}$. There is only a very slight difference in the summary result. Indeed, the summary $\widehat{\eta}$ is lower when based on a meta-analysis of $\widehat{\eta}$ compared to $\widehat{\zeta}$. There are a number of possible explanations for this. First, the meta-analyses use different variance estimates and consequently assign different weights to each study. Indeed, we estimated the variance of $\widehat{\eta}$ using the asymptotic theory, whilst the variance of $\widehat{\zeta}$ was determined via the bootstrap. Generally speaking, the variance of the study specific $\widehat{\zeta}$ will be greater than the variance of the study specific than $\widehat{\eta}$. On top of this, because of the random sampling involved in calculating the bootstrap estimate of the variance, there are likely to be differences in the weights induced by chance. All of these factors contribute to the slight difference in the summary estimate of $\widehat{\eta}$ but, crucially, the results are not radically disparate. In particular, we conclude that there is statistically significant asymmetry to the right in the underlying population. In particular, the magnitude of asymmetry is fairly moderate and is not likely greater than $\eta = 0.6$. The advantage of applying a meta-analysis of $\widehat{\zeta}$ is that the study specific confidence intervals will be interpretable and, as a result, the final summary confidence interval is guaranteed to be contained within the interval $[-1, 1]$.

Hence, taking into consideration the coverage results in Chapter 4, there is once again a tension regarding which approach to take. Similar to Chapter 4, we generally recommend that the meta-analysis be performed directly on $\widehat{\eta}$ using the asymptotic variance. However, if there

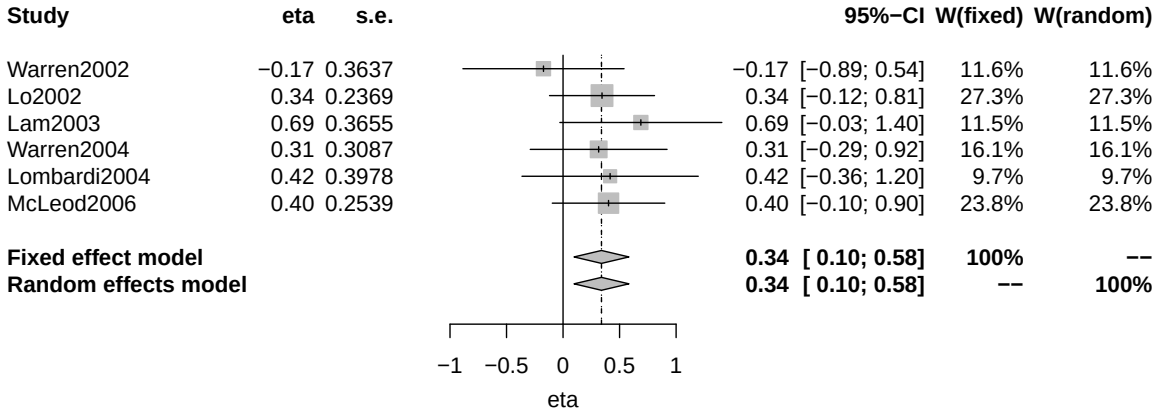

Figure 5.4: Forest plot of the meta-analysis of $\widehat{\eta}$ for preoperative PTH levels in patients who did not go on to develop hypocalcemia ($I^2 = 0\%$).

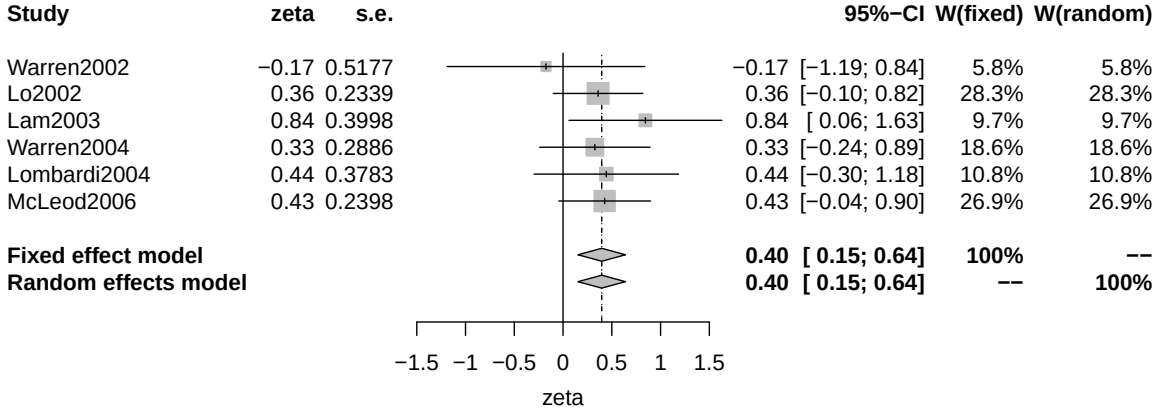

Figure 5.5: Forest plot of the meta-analysis of $\widehat{\zeta}$ for preoperative PTH levels in patients who did not go on to develop hypocalcemia ($I^2 = 0\%$).

M-a method	Overall $\widehat{\eta}$ [95% CI]	I^2
η	$0.34\ [0.10; 0.58]$	0
ζ	$0.38\ [0.15; 0.56]$	0

Table 5.4: Comparing the fixed effect summary results for η using the meta-analysis of $\widehat{\eta}$ and $\widehat{\zeta}$ on preoperative PTH levels in individuals who did not go on to develop hypocalcemia.

are several studies with a relatively small number of observations ($n_i < 30$), which could be influential in the meta-analysis result, then it is suggested to compare the meta-analysis of $\widehat{\eta}$ with a meta-analysis of $\widehat{\zeta}$, estimating the variance of $\widehat{\zeta}$ via bootstrapping. If the two approaches do not differ by an excessive margin then, as with the example given above, we can be relatively confident in the accuracy of the meta-analysis of $\widehat{\eta}$. On the other hand, if the two approaches provide vastly disparate results, then we are forced to conclude that there are simply too few studies or observations within studies to accurately assess the asymmetry in the underlying population. Thus, in such a situation, further studies with larger sample sizes are required to accurately determine the distribution of the underlying population.

In the next section we return to the hypertension data and extend this technique to assess whether there are distributional differences between the treatment and control arms of the trials.

5.5 Using $\widehat{\eta}$ to examine the distributional differences between treatment and control arms

Up to this point, this chapter has focused on summarising and quantifying heterogeneity in the asymmetry of a random variable with samples across multiple studies. In this section we apply the meta-analysis of $\widehat{\eta}$ to identify distributional differences between the treatment and control arms of a randomised control trial across several studies. In particular, suppose we are interested in comparing the systolic blood pressure of the patients. More specifically, we shall try to ascertain whether there is a significant difference between η for systolic blood pressure in the treatment and control groups at two times, before and after the intervention. As the trials are randomised we should expect, prior to treatment, the two populations of systolic blood pressure to be relatively similar. As a result, the amount of asymmetry $\widehat{\eta}$ should be similar between the two groups. In section 3.3 of Chapter 3 we investigated applying $\widehat{\eta}$ to test this hypothesis for individual trials. We now apply the meta-analysis methods outlined in the previous sections to perform this analysis across every study simultaneously and thereby synthesise and compare the results. By contrast, we have no reason to suspect that the two samples (and indeed $\widehat{\eta}$) will be homogeneous post-treatment.

Now suppose that we have k studies comprising of two arms, treatment and control, each containing the same continuous outcome measure. Let $\eta_T^{(i)}$ and $\eta_C^{(i)}$ denote the amount of asymmetry in the continuous variable for the treatment and control arms of study i respectively. Further, let $\theta^{(i)} = \eta_T^{(i)} - \eta_C^{(i)}$ denote the difference in the amount of asymmetry between the arms in study i. We estimate $\theta^{(i)}$ using

$$\widehat{\eta}_T^{(i)} - \widehat{\eta}_C^{(i)},$$

where $\widehat{\eta}_T^{(i)}$ and $\widehat{\eta}_C^{(i)}$ are the estimates of η in the treatment and control groups of study i respectively. Then, using Theorem 2.3 in Chapter 2 it is readily verified that

$$\widehat{\theta}^{(i)} = \widehat{\eta}_T^{(i)} - \widehat{\eta}_C^{(i)} \sim N\left(\theta^{(i)}, \left\{\frac{1}{n_{T;i}} + \frac{1}{n_{C;i}}\right\}\sigma_i^2\right),$$

approximately, where $n_{T;i}$ and $n_{C;i}$ are the sample sizes in the treatment and control groups respectively. We estimate the variance σ_i^2 using the pooled estimate,

$$\widehat{\sigma}_i^2 = \frac{(n_{T;i}-1)\widehat{\sigma}_{t;i}^2 + (n_{C;i}-1)\widehat{\sigma}_{c;i}^2}{n_{T;i} + n_{C;i} - 2},$$

where $\widehat{\sigma}_{t;i}^2$ and $\widehat{\sigma}_{c;i}^2$ are the estimates of σ_i^2 using equation (5.1) in the treatment and control groups of study i respectively.

It is important to note that we are implicitly assuming that the variances of $\widehat{\eta}_T^{(i)}$ and $\widehat{\eta}_C^{(i)}$ are equal. This is perhaps a reasonable assumption provided that $\widehat{\eta}$ is similar in both groups across all studies, but otherwise may be overly restrictive. Indeed, the variance of $\widehat{\theta}^{(i)}$ can be readily calculated when the assumption of equal variances is relaxed. However, additional calculations show that, for the examples considered here, the results are not overly sensitive to this assumption even when there are some small differences in $\widehat{\eta}$ between the treatment and control groups in some studies.

In this case, the fixed effect model assumes that the within study variances are known and that every estimate of the difference in asymmetry $\widehat{\theta}^{(i)}$ is estimating the same underlying difference. That is

$$\theta^{(i)} = \theta, \quad \text{for } i = 1, \cdots, k.$$

Therefore, the fixed effect meta-analysis model for the difference in asymmetry is given by

$$\widehat{\theta}^{(i)} \sim N\left(\theta, \omega_i^2\right), \tag{5.6}$$

where θ denotes the overall difference in the measure of asymmetry between the treatment and control populations and

$$\omega_i^2 = \left\{\frac{1}{n_{T;i}} + \frac{1}{n_{C;i}}\right\}\sigma_i^2$$

The random effects meta-analysis model assumes that the study specific $\theta^{(i)}$ are randomly drawn from a Normal population with mean θ and variance τ_3^2. That is, we assume

$$\begin{aligned} \widehat{\theta}^{(i)} &\sim N\left(\theta^{(i)}, \omega_i^2\right) \\ \theta^{(i)} &\sim N\left(\theta, \tau_3^2\right), \end{aligned} \tag{5.7}$$

where θ denotes the overall difference in the measure of asymmetry between the treatment and control populations, and τ_3^2 is the between study variance.

Table 5.5 shows the results of the meta-analysis of θ prior to treatment, whilst Table 5.6 shows the results of the meta-analysis after treatment. There is no evidence of between study heterogeneity in the value of $\widehat{\eta}$ prior to treatment with, $\widehat{\tau}_3^2 = 0$. Figure 5.6 shows the forest plot obtained by fitting the fixed effect model (5.6) on the difference in $\widehat{\eta}$ between the treatment and control groups for pre-treatment systolic blood pressure. The average difference in the amount of asymmetry θ is 0 (with CI $[-0.03; 0.04]$). In conclusion, we see that there is no evidence of any difference in $\widehat{\eta}$ between the treatment and control groups pre-treatment. This reaffirms the results of our tests in Chapter 3, that prior to treatment the two arms are homogeneous, at least in terms of asymmetry.

Conversely, after treatment there is a substantial amount of between study variability in θ with I^2 as high as 78.8% (with CI $[61.3\%; 88.3\%]$). This implies that 78% of the total variability in the effect estimates can be attributed to between study heterogeneity. Figure 5.7 shows the forest plot obtained by fitting the random effects meta-analysis model (5.7) on the difference in $\widehat{\eta}$ between the treatment and control groups for post-treatment systolic blood pressure. The 95%

Study	$\widehat{\eta}_T - \widehat{\eta}_C$	95% CI	$\%W_f$	$\%W_r$	s.e.	n
ANBP	0.010	$[-0.14; 0.16]$	4.98	4.98	0.08	1530
COOP	-0.023	$[-0.36; 0.31]$	1.01	1.01	0.17	350
EWPH	0.035	$[-0.46; 0.53]$	0.45	0.45	0.26	172
HDFP	0.008	$[-0.07; 0.08]$	19.62	19.62	0.04	4797
MRC1	-0.003	$[-0.09; 0.08]$	14.93	14.93	0.04	6991
MRC2	-0.040	$[-0.21; 0.13]$	3.71	3.71	0.09	2651
SHEP	0.001	$[-0.06; 0.06]$	28.8	28.8	0.03	4736
STOP	0.004	$[-0.36; 0.37]$	0.84	0.84	0.19	268
SYCH	-0.075	$[-0.19; 0.04]$	8.79	8.79	0.06	2391
SYSE	0.049	$[-0.03; 0.13]$	16.87	16.87	0.04	4695
Overall (fixed)	0.00	$[-0.03; 0.04]$				
Overall (random)	0.00	$[-0.03; 0.04]$				

Table 5.5: Meta-analysis of the difference between $\widehat{\eta}$ in the treatment and control arms for systolic blood pressure data prior to treatment ($I^2 = 0\%$). $\%W_f$ and $\%W_r$ denote the percentage weights of each study in the fixed and random effects meta-analysis respectively.

Study	$\widehat{\eta}_T - \widehat{\eta}_C$	95% CI	$\%W_f$	$\%W_r$	s.e.	n
ANBP	0.141	$[-0.02; 0.30]$	4.88	10.18	0.08	1530
COOP	-0.017	$[-0.33; 0.30]$	1.31	5.30	0.16	349
EWPH	0.029	$[-0.48; 0.54]$	0.50	2.55	0.26	172
HDFP	0.063	$[-0.02; 0.14]$	20.66	13.76	0.04	4798
MRC1	-0.007	$[-0.08; 0.07]$	22.59	13.89	0.04	6991
MRC2	0.232	$[0.11; 0.36]$	8.6	11.94	0.06	2651
SHEP	0.213	$[0.12; 0.31]$	14.64	13.17	0.05	4736
STOP	0.282	$[-0.10; 0.66]$	0.90	4.08	0.19	268
SYCH	0.148	$[0.02; 0.28]$	7.50	11.56	0.07	2391
SYSE	0.324	$[0.24; 0.41]$	18.41	13.58	0.04	4695
Overall (fixed)	0.14	[0.11; 0.18]				
Overall (random)	0.15	[0.06; 0.24]				

Table 5.6: Meta-analysis of the difference between $\widehat{\eta}$ in the treatment and control arms for systolic blood pressure data after treatment ($I^2 = 78.8\%$). $\%W_f$ and $\%W_r$ denote the percentage weights of each study in the fixed and random effects meta-analysis respectively.

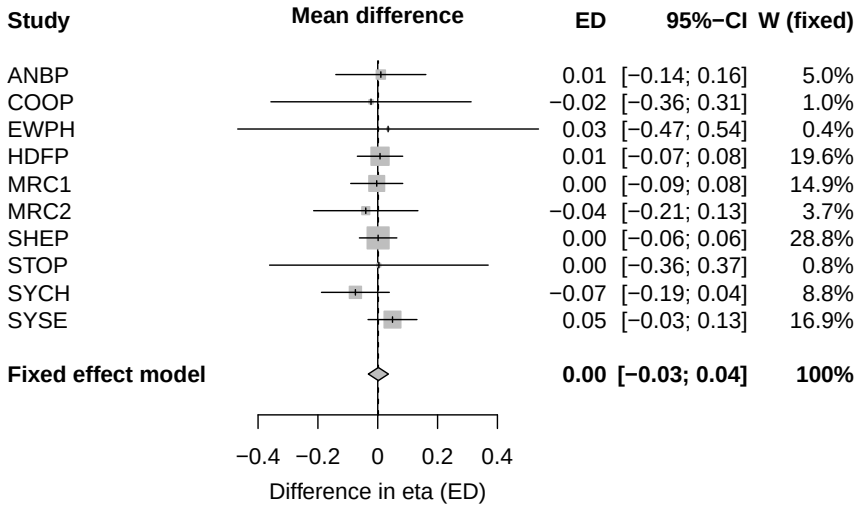

Figure 5.6: Fixed effect meta-analysis conducted on the difference in $\widehat{\eta}$ (ED) between the treatment and control of pre-treatment systolic blood pressure ($I^2 = 0\%$).

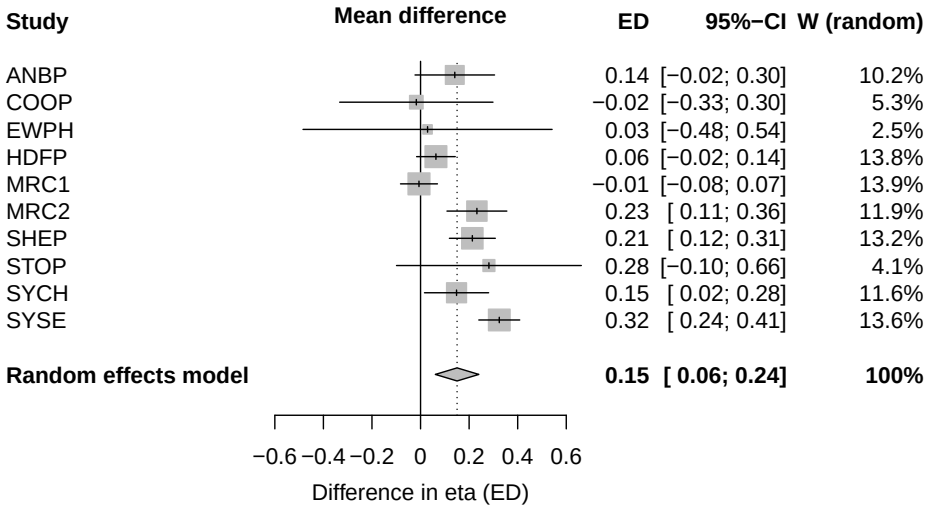

Figure 5.7: Random effects meta-analysis fit using DerSimonian-Laird on the difference $\widehat{\eta}$ (ED) between the treatment and control of post-treatment systolic blood pressure ($I^2 = 78.8\%$).

confidence interval for the average post-treatment difference $\widehat{\theta}$ is [0.06; 0.24], which suggests that η is larger in the treatment arm than in the control arm post-treatment. Intuitively, perhaps this is what one should expect, as there is likely to be little or no change in the control group after treatment, whilst it is plausible that the treatment not only reduces blood pressure, but also alters the distribution in terms of the amount of asymmetry. Indeed, variations in an individual's response to treatment can induce changes in the sampling distribution of the blood pressure readings. The 95% prediction interval for the post-treatment difference θ for a new study is $[-0.14; 0.44]$, which suggests that it is also plausible that the post-treatment difference in $\widehat{\eta}$ may be zero or even negative in a future study.

5.6 Discussion

In this chapter we introduced the fixed effect and random effects meta-analysis models before conducting a brief review of the existing literature on the meta-analysis of correlation studies. This provided the foundation for investigating the problem of measuring asymmetry in a population based on several studies, with the aim of obtaining an 'overall' measure of asymmetry in the underlying population and determining a prediction interval for the asymmetry in a new study. Initially, we explored the potential pitfalls of naively pooling data across all studies and analysing it as if it were a single data set. This was the motivation for proposing a meta-analysis on the asymmetry measure $\widehat{\eta}$ to combat this problem. We also applied the technique using a much smaller data set and, in this case, we show that $\widehat{\zeta}$ can be used to obtain suitably asymmetric confidence intervals. Next, we extended this technique to identify distributional differences between the treatment and control arms of a randomised control trial across several studies. In particular, we demonstrated that the method can be used to assess the baseline similarity in $\widehat{\eta}$ across multiple studies simultaneously, as well as investigating the change in the distribution post-treatment.

To summarise, when every study is reasonably large (say, $n > 30$), it is recommended that meta-analyses be performed directly on $\widehat{\eta}$. However, if there are a number of studies with insufficient sample sizes to ensure the accuracy of the asymptotic distribution of $\widehat{\eta}$, it is recommend that the meta-analysis results be compared to the small sample procedure based on

$\widehat{\zeta}$. That is, it was shown that $\widehat{\eta}$ may still produce meaningful results (especially if the offending studies receive a lower weighting in the meta-analysis), but the results should be the subject of a sensitivity analysis, by additionally considering a meta-analysis of $\widehat{\zeta}$. In general, caution is advised when applying either method if there are a small number of samples or studies, as the investigations of Field [34] into meta-analyses of correlation coefficients attest. In this situation neither method is ideal and, therefore, further research studies with large sample sizes are the priority.

In conclusion we have shown that a meta-analysis of $\widehat{\eta}$ is a useful technique for determining the overall asymmetry in a population, when we have access to a number of different studies from that population. It also provides a platform to compare and contrast the distribution of different groups over several studies on the same treatment. In the next chapter we perform a simulation study to assess the effect of ignoring violations of the symmetry or normality assumption on the inferences one draws from several statistical models.

Key findings and recommendations

- When every study is reasonably large (say, $n > 30$), it is recommended that meta-analysis be performed directly on $\widehat{\eta}$.
- If there are a number of studies with insufficient sample sizes to ensure the accuracy of the asymptotic theory, then the results should be the subject of a sensitivity analysis by alternatively considering a meta-analysis of $\widehat{\zeta}$.
- If the majority of studies have small sample sizes, further research is required to obtain larger sample sizes across the population of interest.
- To ensure accurate estimation of η every study must have a sufficiently large sample size to construct a reasonable estimate of the density function f and the distribution function F. The requisite number of points for this will vary depending on the shape of the underlying population.

Chapter 6

The effect of violating symmetry and normality assumptions in statistical models

6.1 Introduction

As we have previously identified, there are a number of statistical tests which assume that the data are sampled from a symmetric population. For example, the Wilcoxon signed-rank test, used to test for differences between two samples with unknown distribution functions, relies on the assumption that the data are drawn from a symmetric population [128]. Many other commonly used statistical tests or models rely on the assumption of symmetry implicitly, via the normality assumption. For example, the t-test assumes that the data are drawn from a Normal distribution [119], while linear models assume that the residuals are sampled from a normally distributed population [82]. Furthermore, the random effects meta-analysis model introduced in the previous chapter assumes a Normal distribution on two levels. That is, it assumes that the study specific treatment effect estimates $\widehat{\theta}_i$ are each normally distributed about a 'true' study specific effect θ_i, and that these random effects are also normally distributed about an overall treatment effect. In these cases assessing the symmetry of the distribution is an essential step towards evaluating the normality. In this chapter we analyse to what extent some of the

commonly used statistical methods are robust to departures from symmetry or normality.

The outline of the chapter is as follows. In section 6.2 we investigate to what extent violations of the normality assumption impinge the inferences we are able to draw from linear models. In the context of clinical trials, we discuss the effect on the sampling distribution of the treatment effect estimate, as well as the corresponding confidence intervals. Further, we investigate the impact of asymmetric or skewed data on the predictive capabilities of the linear model. In section 6.3 we extend this investigation to meta-analyses and determine how the presence of asymmetry in the random effects distribution impacts the accuracy of confidence and prediction intervals. In section 6.4 we also include an example to illustrate the effect that asymmetry in the random effects has on the construction of prediction intervals in this context. In section 6.5 we summarise the results of the chapter and provide some recommendations regarding the use of linear models and meta-analyses when there is asymmetry present in the residual or random effects distribution respectively.

Aims of the chapter:

- Investigate the effect of asymmetric data on the accuracy of common inferences made from linear models.
- Explore the impact of asymmetry in the random effects on the inferences drawn from meta-analyses.
- Propose a number of recommendations regarding the use of linear models and meta-analyses when there is asymmetry present in the residual or random effects distribution respectively.

6.2 The effect of asymmetric data on linear models used to analyse randomised control trials

6.2.1 Introduction

Recall from Chapter 1 the following simple linear model applied to two groups of data,

$$\begin{aligned} &Y_j = \beta_0 + \beta_1 X_j + e_j, \quad j = 1, \cdots, n, \\ &e_j \sim N(0, \sigma^2), \end{aligned} \tag{6.1}$$

where Y_j is some continuous response and X_j is the group identifier and can be equal to 0 or 1. It is assumed that

1. The residuals e_j are normally distributed.
2. The variances are homogeneous in both groups.
3. The errors are independent.

In practice it is common for one or more of the assumptions are violated. For example, the number of error inducing factors can be so numerous and complex that in reality we should not expect the data to reflect a perfectly Normal sample. As a result, in reality the residuals of model (6.1) are rarely going to follow a Normal distribution exactly. For the most part this is not a serious issue, as it is well established that linear models are robust to small departures from normality [82]. As a result, the inferences that one draws from the model should not be too badly biased by the existence of some slight skewness. However, for more serious departures from normality (e.g. heavily skewed data) the usual inferences that one draws from the model (e.g. treatment effect estimates or confidence intervals) may well be seriously compromised.

Previous work by Lumley et al. [75] has shown that t-test approach is surprisingly robust to even extreme departures from normality. In this section we conduct a similar analysis in the context of linear models that are used to estimate a treatment effect estimate of a randomised trial. In particular, we analyse the effect of asymmetric data on the sampling distribution of the treatment effect estimate, as well as investigating the accuracy of the corresponding confidence

intervals for the treatment effect. Further, we also investigate the effect of asymmetric data on the predictive power of linear models by examining the the accuracy of prediction intervals for future patients.

6.2.2 Methods

The sampling distribution of the treatment effect

Firstly, we assess the effect of asymmetric residuals on the sampling distribution of $\widehat{\beta}_1$, the treatment effect estimate, by simulating two sets of data resulting in heavily skewed residuals from the Log-Normal distribution. In particular, we generate two artificial Log-Normal samples (Y_j), each of size n, which have a difference in mean equal to one and fit the simple linear model (6.1) using all $2n$ data points. Under the normality assumption we have that

$$\frac{\widehat{\beta}_1 - \beta_1}{s_{\widehat{\beta}}} = \frac{\widehat{\beta}_1 - 1}{s_{\widehat{\beta}}} \sim t_{2n-2},$$

where t_{2n-2} is the t distribution with $2n-2$ degrees of freedom, and $s_{\widehat{\beta}}$ is the standard error of $\widehat{\beta}_1$ given by

$$s_{\widehat{\beta}} = \sqrt{\frac{\frac{1}{2n-2}\sum_i e_i^2}{\sum_i (x_i - \overline{x})^2}} = \sqrt{\frac{\frac{1}{2n-2}\sum_i e_i^2}{2n\frac{1}{4}}} = \sqrt{\frac{\frac{1}{n-1}\sum_i e_i^2}{2n}},$$

and e_i are the residuals of the model. By repeatedly simulating the asymmetric Log-Normal data and fitting the above linear model we obtain a random sample of $\widehat{\beta}_1$. We also perform the same analysis using normally distributed data and compare the results to assess the impact of the asymmetric Log-Normal data.

Coverage of confidence intervals

On top of this, we investigate the effect of asymmetry on the confidence intervals obtained for the treatment effect estimates using the linear model (6.1). Montgomery et al. [82] state that the $100(1-\alpha)\%$ confidence interval for the treatment effect of a simple linear regression is given by

$$\widehat{\beta}_1 \pm t_{\alpha/2;n-2}\sigma\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{s_{xx}}}, \tag{6.2}$$

where n is the total sample size and $s_{xx} = \sum_i (x_i - \overline{x})^2$.

We simulate $10,000$ replications of the simple linear model (6.1) with baseline $\beta_0 = 0$ and true treatment effect $\beta_1 = 1$ with residuals drawn from a variety of distributions. In particular, we consider residuals drawn from Normal, Normal mixtures, Skew Normal, Log-Normal, Folded Normal and Exponential populations. We then construct the $100(1-\alpha)\%$ confidence interval for the treatment effect using a range of levels α, namely, $\alpha = 0.5, 0.1, 0.05, 0.01$. For each simulated model we determine whether or not this confidence interval contains the true value and thereby calculate the coverage.

Coverage of prediction intervals

Next, we consider the effect of asymmetric data on the accuracy of prediction intervals. Given an observation for the independent variable $X = x_0$ Montgomery et al. [82] give the $100(1-\alpha)\%$ prediction interval for a simple linear regression as

$$\widehat{y}_0 \pm t_{\alpha/2;n-2}\sigma\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{s_{xx}}}, \tag{6.3}$$

where $\widehat{y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$. Observe, that the form is very similar to the confidence intervals, but with an additional term to account for the extra variability.

We assess whether the predictions from linear models are sensitive to the to the distribution of the data by, firstly, constructing the fitted model using a random sample from a particular distribution. Again, we consider residuals drawn from Normal, Normal mixtures, Skew Normal, Log-Normal, Folded Normal and Exponential populations. Next, we generate a prediction interval based on that model, before simulating a new random point from the distribution and determine the predicted value for y,

$$\widehat{y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0.$$

Finally, we determine whether or not the predicted value is contained within the constructed prediction interval. For each distribution, we repeat this process $m = 10,000$ times and determine the coverage of the prediction intervals. The simulation study is outlined in Table 6.1.

Step 1a	For each given distribution, generate the 'control' sample by simulating n observations from the residual distribution.
Step 1b	Generate the 'treatment' sample by simulating n observations from the residual distribution and adding on the 'true' treatment effect $\beta_1 = 1$.
Step 2	Fit the linear model (6.1) using all $2n$ data points, naively assuming that the residuals are normally distributed.
Step 3a	Extract the estimate of the treatment effect $\widehat{\beta}_1$. Also, construct the confidence interval for β_1 using equation (6.2), as well as the prediction interval for a new individual (treatment and control) using equation (6.3).
Step 3b	Check to see whether the confidence intervals contain the true overall treatment effect $\beta_1 = 1$. Also, simulate a new observation for a new individual (either treatment or control) using the 'true' residual distribution and determine whether the corresponding prediction interval contains the new observation.
Step 4	Repeat **Steps 1-3b** $10,000$ times and report the coverage of the confidence and prediction intervals.

Table 6.1: Outline of the simulation study assessing the impact of asymmetry in the residuals of linear models on the coverage of confidence and prediction intervals.

6.2.3 Results

The sampling distribution of the treatment effect

Figure 6.1 shows the density estimate for the values of $\frac{\widehat{\beta}_1}{s_{\widehat{\beta}}}$ obtained by fitting the linear model (6.1) $10,000$ times with $n = 5, 15, 30, 100$ observations in each arm, using Normal residuals and asymmetric Log-Normal residuals. The dotted line shows the t distribution with $2n - 2$ degrees of freedom, which is the distribution of $\widehat{\beta}_1$ under the normality assumption. It is clear that $\widehat{\beta}_1$ does not conform exactly to this curve when the normality assumption has been violated, however, the error is significantly reduced as the sample size increases. As a result, for small sample sizes confidence intervals based upon this assumption will be compromised. On the other hand, when the sample size is reasonably large $\widehat{\beta}_1$ is approximately normally distributed by the Central Limit Theorem, even when the normality assumption is violated. This can be seen in the far right density in Figure 6.1 where $n = 100$.

Consequently, for small n it is clear that the sampling distribution of $\widehat{\beta}_1$ is altered by viola-

tions of the normality assumption, however, the pertinent question is 'How seriously does this effect the inferences that we draw from the model?'. To answer this question we investigate the effect of asymmetric residuals on the coverage of the confidence intervals constructed under the assumption of normality.

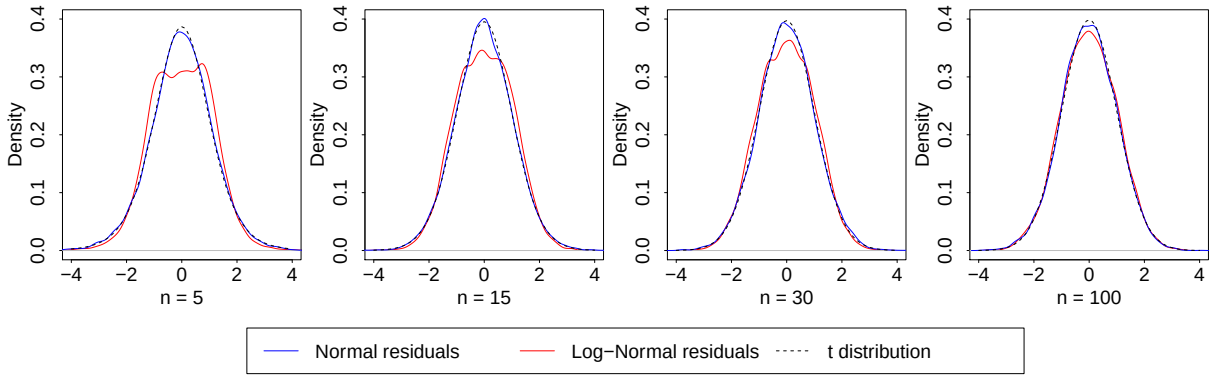

Figure 6.1: The density estimate of $\frac{\widehat{\beta}_1}{s_{\widehat{\beta}}}$ for Log-Normal residuals ($\eta \approx 0.91$) based on $10,000$ samples. The dotted line shows the hypothesised t distribution whilst the red line shows the distribution of $\frac{\widehat{\beta}_1}{s_{\widehat{\beta}}}$ for Normal residuals.

Coverage of confidence intervals

Table 6.2 shows the coverage of the confidence intervals constructed using the linear model (6.1) with residuals drawn from several different populations and a range of sample sizes. As expected, for the symmetric Normal distribution roughly $100(1-\alpha)\%$ of the confidence intervals contain the true value. For the most part, the coverage is not greatly altered when the residuals are sampled from the the asymmetric distributions. However, when the residuals are drawn from the heavy tailed Log-Normal samples the coverage deviates from the expected value of $1-\alpha$. This is most pronounced for the smaller samples ($n < 30$) and $\alpha = 0.5$. Indeed, for a sample of $n = 15$ from a Log-Normal population the 50% confidence interval for the treatment effect has coverage of 0.452 and so is considerably shorter than expected. However, this disparity is reduced as n increases. Moreover, it is also less noticeable for the more conventional confidence intervals ($\alpha = 0.1, 0.05$). For example, for a sample of $n = 15$ from a Log-Normal population the 95% confidence interval for the treatment effect has coverage of 0.9608, and is only marginally too wide.

		$\alpha = 0.5$					$\alpha = 0.1$				
η	Dist.	$n = 3$	$n = 8$	$n = 15$	$n = 30$	$n = 100$	$n = 3$	$n = 8$	$n = 15$	$n = 30$	$n = 100$
0	N	0.493	0.508	0.495	0.491	0.503	0.898	0.903	0.900	0.900	0.899
0.1	NM1	0.489	0.486	0.493	0.496	0.501	0.906	0.903	0.904	0.904	0.897
0.2	NM2	0.477	0.496	0.494	0.490	0.503	0.911	0.903	0.905	0.898	0.896
0.3	NM3	0.475	0.492	0.498	0.505	0.514	0.907	0.896	0.900	0.900	0.907
0.4	NM4	0.484	0.494	0.495	0.499	0.507	0.913	0.901	0.900	0.902	0.901
0.5	SN5	0.497	0.507	0.502	0.496	0.500	0.908	0.900	0.900	0.905	0.901
0.6	SN6	0.495	0.496	0.501	0.498	0.500	0.894	0.903	0.898	0.902	0.905
0.7	SN7	0.497	0.493	0.501	0.497	0.499	0.903	0.903	0.896	0.898	0.903
0.8	SN8	0.483	0.489	0.494	0.501	0.499	0.908	0.902	0.901	0.903	0.902
0.91	LN	0.433	0.440	0.452	0.477	0.475	0.929	0.915	0.915	0.907	0.899
0.95	FN	0.501	0.493	0.482	0.501	0.495	0.901	0.900	0.900	0.906	0.898
1	EXP	0.469	0.476	0.488	0.494	0.495	0.916	0.908	0.903	0.902	0.899
		$\alpha = 0.05$					$\alpha = 0.01$				
η	Dist.	$n = 3$	$n = 8$	$n = 15$	$n = 30$	$n = 100$	$n = 3$	$n = 8$	$n = 15$	$n = 30$	$n = 100$
0	N	0.950	0.948	0.949	0.949	0.947	0.991	0.990	0.989	0.990	0.990
0.1	NM1	0.954	0.951	0.953	0.952	0.953	0.991	0.994	0.992	0.990	0.991
0.2	NM2	0.959	0.950	0.949	0.948	0.952	0.994	0.994	0.993	0.990	0.992
0.3	NM3	0.958	0.958	0.953	0.953	0.947	0.992	0.992	0.992	0.990	0.990
0.4	NM4	0.956	0.949	0.951	0.953	0.951	0.992	0.992	0.989	0.991	0.991
0.5	SN5	0.949	0.950	0.950	0.946	0.947	0.990	0.992	0.989	0.991	0.989
0.6	SN6	0.952	0.951	0.950	0.953	0.946	0.989	0.991	0.991	0.990	0.991
0.7	SN7	0.951	0.950	0.955	0.949	0.950	0.989	0.992	0.992	0.991	0.991
0.8	SN8	0.949	0.951	0.955	0.951	0.950	0.987	0.991	0.990	0.989	0.991
0.91	LN	0.966	0.966	0.961	0.958	0.954	0.992	0.996	0.996	0.995	0.994
0.95	FN	0.952	0.952	0.952	0.951	0.951	0.988	0.992	0.990	0.990	0.992
1	EXP	0.957	0.958	0.955	0.950	0.952	0.992	0.995	0.993	0.992	0.990

Table 6.2: The coverage of the $100(1-\alpha)\%$ confidence interval for the treatment effect, constructed using the linear model (6.1) with residuals simulated from a range of populations (Dist.) and n observations in each arm.

Hence, we observe that when we have an extremely small sample drawn from an asymmetric population, failure to account for this may produce confidence intervals which are too wide or too short. However, in most cases the discrepancy is still small for 95% confidence intervals and, moreover, this effect is reduced as the sample size increases. Indeed, the Central Limit Theorem ensures that the sampling distribution of the treatment effect estimate will approach a Normal distribution for sufficiently large n, regardless of the asymmetry inherent in the underlying data.

In summary, the results indicate that, generally speaking, the confidence intervals for the treatment effect estimate are extremely robust to departures from symmetry. This is particularly evident for the 95% confidence intervals (as opposed to the shorter unconventional 50% confidence intervals). Indeed, even for the extremely asymmetric distributions considered here, the coverage of the 95% confidence intervals only display minor departures from 0.95.

Further, it is important to note that some of the distributions considered here are extreme examples (for example, the Log-Normal distribution). In reality, most clinical trials are unlikely to contain such large departures from symmetry and, moreover, the majority of randomised control trials do not contain fewer than 30 patients. As a result 'realistic' asymmetry will not seriously effect the treatment effect estimate. The same may not be true if one's interest is to use the model to make predictions about future patients. Next, we present the corresponding coverage results for the prediction intervals.

Coverage of prediction intervals

The coverage of the $100(1-\alpha)\%$ prediction intervals are given in Table 6.3. As one should expect, when the residuals are sampled from the symmetric Normal distribution, roughly $100(1-\alpha)\%$ of the prediction intervals contain the newly simulated value. Once again, for the asymmetric distributions, the most striking departures occur for the 50% confidence intervals where the residuals are sampled from a Log-Normal distribution. Indeed, for $n = 15$ the coverage of the prediction interval is much too large (0.648). Moreover, this actually worsens as n increases and for $n = 100$ the 50% confidence interval is far too large with coverage given by 0.739. This is a trend emulated by the Normal mixtures which have reasonable coverage for small n, but this coverage steadily grows as we increase the sample size. By contrast, for the Skew Normal data

		$\alpha = 0.5$					$\alpha = 0.1$				
η	Dist.	$n = 3$	$n = 8$	$n = 15$	$n = 30$	$n = 100$	$n = 3$	$n = 8$	$n = 15$	$n = 30$	$n = 100$
0	N	0.500	0.503	0.499	0.500	0.489	0.901	0.902	0.906	0.906	0.901
0.1	NM1	0.508	0.530	0.539	0.542	0.543	0.901	0.905	0.909	0.909	0.918
0.2	NM2	0.514	0.542	0.562	0.567	0.583	0.893	0.901	0.904	0.913	0.915
0.3	NM3	0.521	0.553	0.560	0.569	0.577	0.893	0.899	0.907	0.905	0.910
0.4	NM4	0.517	0.514	0.524	0.540	0.536	0.901	0.900	0.905	0.902	0.910
0.5	SN5	0.500	0.497	0.489	0.496	0.494	0.900	0.905	0.910	0.917	0.914
0.6	SN6	0.501	0.496	0.480	0.479	0.485	0.896	0.900	0.915	0.918	0.917
0.7	SN7	0.503	0.495	0.487	0.475	0.480	0.899	0.911	0.913	0.927	0.923
0.8	SN8	0.509	0.485	0.480	0.468	0.477	0.896	0.908	0.921	0.920	0.928
0.91	LN	0.548	0.608	0.648	0.688	0.739	0.882	0.912	0.928	0.935	0.941
0.95	FN	0.506	0.485	0.479	0.473	0.472	0.901	0.917	0.922	0.919	0.929
1	EXP	0.534	0.531	0.541	0.542	0.527	0.887	0.915	0.918	0.919	0.927
		$\alpha = 0.05$					$\alpha = 0.01$				
η	Dist.	$n = 3$	$n = 8$	$n = 15$	$n = 30$	$n = 100$	$n = 3$	$n = 8$	$n = 15$	$n = 30$	$n = 100$
0	N	0.950	0.950	0.951	0.948	0.950	0.989	0.991	0.991	0.989	0.991
0.1	NM1	0.946	0.945	0.947	0.954	0.955	0.985	0.981	0.980	0.981	0.981
0.2	NM2	0.939	0.940	0.937	0.949	0.947	0.985	0.974	0.977	0.973	0.973
0.3	NM3	0.934	0.938	0.940	0.939	0.940	0.980	0.978	0.974	0.974	0.973
0.4	NM4	0.939	0.944	0.945	0.941	0.945	0.983	0.983	0.977	0.979	0.982
0.5	SN5	0.946	0.947	0.951	0.952	0.954	0.986	0.986	0.982	0.985	0.983
0.6	SN6	0.942	0.947	0.947	0.955	0.954	0.985	0.984	0.984	0.981	0.984
0.7	SN7	0.945	0.951	0.951	0.953	0.955	0.982	0.979	0.978	0.982	0.982
0.8	SN8	0.938	0.947	0.952	0.954	0.952	0.984	0.981	0.982	0.985	0.982
0.91	LN	0.916	0.930	0.937	0.949	0.958	0.959	0.954	0.961	0.967	0.973
0.95	FN	0.942	0.954	0.951	0.952	0.951	0.983	0.984	0.979	0.981	0.984
1	EXP	0.924	0.939	0.943	0.946	0.945	0.972	0.967	0.970	0.969	0.971

Table 6.3: The coverage of the $100(1-\alpha)\%$ prediction interval for the response variable in a new individual, constructed using the linear model (6.1) with residuals simulated from a range of populations (Dist.) and n observations in each arm.

the coverage steadily declines away from 0.5 for increasing sample size. This seems to suggest that, unlike the previous problem of constructing confidence intervals for the treatment effect, one cannot guarantee that the prediction intervals will be suitable for large samples.

This is emphasised in Figure 6.2 which shows the appropriately centred and scaled density functions for the Normal, SN5, SN8 and Log-Normal distributions. It also includes the location

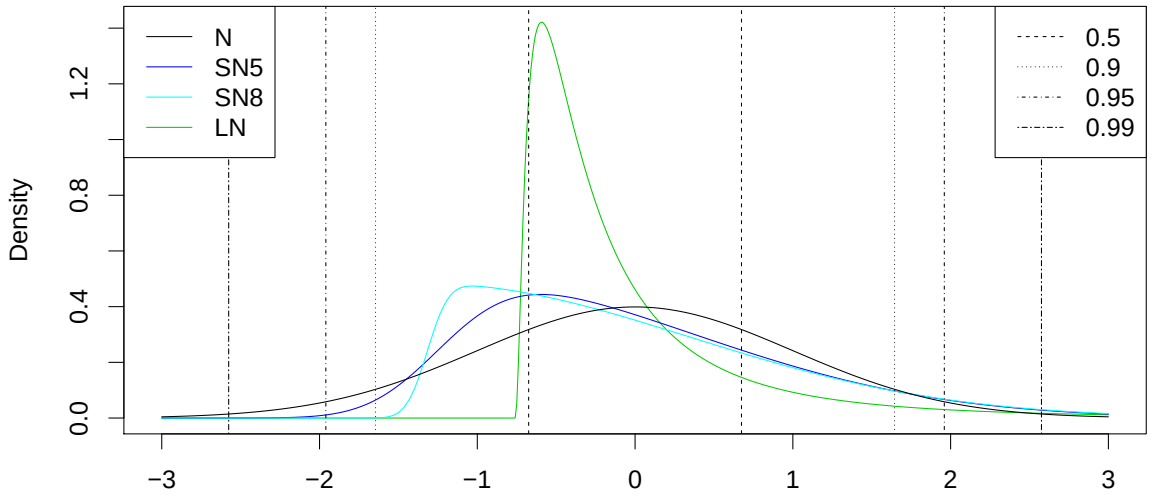

Figure 6.2: Centred and scaled density functions of the Normal, SN5, SN8, and Log-Normal distributions along with the location of the $100\frac{\alpha}{2}$ and $100\left(1-\frac{\alpha}{2}\right)$ percentiles of the Normal distribution for $\alpha = 0.5$, 0.1, 0.05 and 0.01.

of the $100\frac{\alpha}{2}\%$ and $100\left(1-\frac{\alpha}{2}\right)\%$ percentiles of the Normal distribution for $\alpha = 0.5$, 0.1, 0.05 and 0.01. If a new observation is sampled from a Normal distribution then these percentiles correspond to the expected endpoints of the $100(1-\alpha)\%$ prediction interval. If, however, the new observation is not drawn from a Normal distribution, it is clear that a prediction interval constructed under the assumption of symmetry will fail to adequately capture the variability in the new observation. It just so happens that all of these distributions have approximately 95% of their mass between the 2.5% and 97.5% percentiles of the Normal distribution. By contrast, there are large differences in the proportion of probability mass between the 25% and 75% percentiles of the Normal distribution. For example, the Log-Normal distribution has approximately 82% of its mass in this region, whilst the SN8 distribution has approximately

47% of its mass in this region. These differences explain the stark contrast in the coverage of the 50% confidence interval in Table 6.3.

Again, it must be pointed out that this discrepancy is less exaggerated for the wider prediction intervals, in particular the 95% and 99% prediction intervals, which only show mild departures from 0.95 or 0.99 for very small samples. Indeed, even for the heavily asymmetric Log-Normal distribution with a very small sample number of observations in each arm ($n = 5$) the coverage of the 95% prediction interval is 0.916, which is only a slight deviation from the expected value, whilst for $n = 100$ the 99% prediction interval has a coverage of 0.973.

In summary, although the coverage of the more conventional 95% and 99% prediction intervals appears to be reasonable even when the residual distribution is asymmetric, the results for the shorter prediction intervals highlight the fact that the impact on probabilistic inferences is not yet clear. Indeed, it was shown that the 50% prediction intervals can be extremely sensitive to asymmetric data. Moreover, in the context of clinical trials more care should be taken when interpreting prediction intervals than confidence intervals, as one can no longer appeal to the Central Limit Theorem for larger samples. In the next section we extend this investigation to meta-analyses and we begin by reintroducing fixed effect and random effects meta-analysis.

6.3 The impact of asymmetry in the random effects of meta-analyses

6.3.1 Introduction

Suppose that we have k studies investigating a specific treatment or intervention and that each study reports the treatment effect estimate $\widehat{\theta}_i$ and its variance σ_i^2, where $i = 1, \cdots, k$. Recall from section 5.2.2 in Chapter 5, that when there is evidence of between study heterogeneity, one can perform a random effects meta-analysis to account for the variability between studies. A random effects meta-analysis assumes that the treatment effects θ_i vary across studies, following a Normal distribution with mean θ and variance τ^2. In particular, the conventional parametric approach to random effects meta-analysis, outlined by Whitehead and Whitehead [127] has the form

$$\begin{aligned} \widehat{\theta}_i &\sim N(\theta_i, \sigma_i^2), \quad i = 1, \cdots, k, \\ \theta_i &\sim N(\theta, \tau^2). \end{aligned} \tag{6.4}$$

Recall, that the $(1-\alpha)100\%$ confidence interval for θ is conventionally calculated by assuming that the σ_i^2 are known. In this case

$$\widehat{\theta} \pm Z_{\frac{\alpha}{2}} \sqrt{\text{Var}\left(\widehat{\theta}\right)}, \tag{6.5}$$

where $\widehat{\theta}$ is the maximum likelihood estimate of θ, and $Z_{\frac{\alpha}{2}}$ is the $\frac{\alpha}{2}$ quantile of the standard Normal distribution. However, it was noted by Knapp and Hartung [67] that this form for the confidence interval does not account for the uncertainty in the estimate of the variance of $\widehat{\theta}$. Instead, Knapp and Hartung suggest using

$$\widehat{\theta} \pm t_{k-1;\frac{\alpha}{2}} \sqrt{\text{Var}_\iota\left(\widehat{\theta}\right)}, \tag{6.6}$$

where $t_{k-1;\frac{\alpha}{2}}$ is the $\frac{\alpha}{2}$ quantile of the t distribution with $k-1$ degrees of freedom and $\text{Var}_\iota\left(\widehat{\theta}\right)$ is the appropriately modified variance estimate

$$\text{Var}_\iota\left(\widehat{\theta}\right) = \frac{1}{k-1} \frac{\sum_{i=1}^n w_i^* \left(\widehat{\theta}_i - \widehat{\theta}\right)^2}{\sum_{i=1}^n w_i^*},$$

where $w_i^* = \frac{1}{\sigma_i^2 + \tau^2}$ [50]. This alternative variance estimate facilitates the use of the t distribution in equation (6.6) because

$$\frac{\widehat{\theta}}{\sqrt{\text{Var}_\iota\left(\widehat{\theta}\right)}} = \frac{\sqrt{\sum_{i=1}^n w_i^*} \cdot \widehat{\theta}}{\sqrt{\frac{1}{k-1} \sum_{i=1}^n w_i^* \left(\widehat{\theta}_i - \widehat{\theta}\right)^2}},$$

where $\sqrt{\sum_{i=1}^n w_i^*} \cdot \widehat{\theta}$ is approximately standard Normal and $\sum_{i=1}^n w_i^* \left(\widehat{\theta}_i - \widehat{\theta}\right)^2$ follows a chi-squared distribution with $k-1$ degrees of freedom.

The $(1-\alpha)100\%$ prediction interval, proposed by Higgins et al. [55], is

$$\widehat{\theta} \pm t_{k-2;\frac{\alpha}{2}} \sqrt{\widehat{\tau}^2 + \text{Var}\left(\widehat{\theta}\right)}, \tag{6.7}$$

where $\widehat{\tau}^2$ is the estimate of the between study heterogeneity.

As we have already identified, the random effects meta-analysis model assumes symmetry at two levels. Firstly, it assumes that the study specific estimate of the treatment effect $\widehat{\theta}_i$ is normally distributed about the 'true' study specific effect θ_i. Secondly, it is assumed that these random effects are also normally distributed about an overall treatment effect. Baker and Jackson [6] state that normality is "the default assumption made about unknown distributions... having the merit of simplicity and often being motivated by the Central Limit Theorem (CLT)." Indeed, the CLT unquestionably suggests that, for reasonably large studies, the estimated effect $\widehat{\theta}_i$ should be approximately normally distributed about the true study effect θ_i. However, as noted by Baker and Jackson, "the CLT does not really imply anything about the distribution of the random effects," and that one can only appeal to the CLT in this case "with the vague idea that the unknown source of variation between studies might be the sum of several factors."

The results of the previous section exemplify the strength of the Central Limit Theorem in the context of estimating the treatment effect from a linear model. Thus, in this section we focus our investigation on asymmetry in the random effects distribution, and the impact this can have on the results and interpretation of the meta-analysis.

6.3.2 Methods

In order to obtain the most accurate perspective of the impact of asymmetry on meta-analyses we simulate data in two stages. Firstly, for a given treatment effect we first generate k random effects from a particular distribution, which has variance τ^2. This allows for the possibility of generating asymmetry in the random effects distribution. Secondly, for each of the k studies, we simulate Normal data (i.e. the continuous outcome response value) for the control and treatment arms of the individual studies, with sample size n in each arm and a difference in means determined by the random effects.

Thus, in this case the treatment effects θ_i are the mean differences in each study, and the 'known' within study variances σ_i^2 are estimated from the continuous response. We then fit the random effects model (6.4) using restricted maximum likelihood estimation (REML), naively assuming normality at both levels. We extract the model estimates of θ and τ and construct confidence intervals for θ naively assuming that overall treatment effect is normally distributed, using both equation (6.5) and equation (6.6). Similarly, we construct prediction intervals for a new treatment effect by assuming that the random effects are normally distributed and using equation (6.7).

The random effects are generated from a Normal distribution and a range of increasingly asymmetric distributions. In particular, we consider random effects drawn from Normal, Normal mixtures, Skew Normal, Log-Normal, Folded Normal and Exponential populations.

For every distribution we: vary the number of studies k; use a range of sample sizes n; use a true overall treatment effect $\theta = 1$; use a fixed baseline $\theta_0 = 0$; fix the variance of the continuous response at $\sigma^2 = 1$; and set the between study variance to be $\tau^2 = 1$. It is important to note that the within study variance is given by

$$\sigma_i^2 = \frac{2\sigma^2}{n},$$

hence $\tau^2 >> \sigma_i^2$. We also tested the sensitivity of the results to a range of different parameters for θ, σ^2, τ^2 and found the results to be relatively consistent provided that $\tau^2 >> \sigma_i^2$. We generate each meta-analysis $10,000$ times using the R package meta and extract the model estimates of the overall treatment effect θ and the between study standard deviation τ. We also construct the $100(1-\alpha)\%$ confidence intervals for the overall treatment effect (using Z and t quantiles), as well as the prediction interval for the treatment effect of a new study. In particular, we consider $\alpha = 0.5, 0.1, 0.05$ and 0.01. At each iteration we check to see whether the constructed confidence intervals contain the true treatment effect and thereby calculate the coverage (the proportion of confidence intervals containing the true treatment effect). Similarly, for each iteration we simulate a new treatment effect from the predictive distribution and determine whether the prediction interval contains this new treatment effect. In this way we also calculate the coverage

of the $100(1-\alpha)\%$ prediction interval. The simulation study is outlined in Table 6.4.

Step 1	For a given random effects distribution, simulate k random effects with mean $\theta = 1$ and variance $\tau^2 = 1$.
Step 2	For each study, simulate two normally distributed samples of size n with equal variance $\sigma^2 = 1$, and with a difference in means determined by the random effect θ_i. In particular, the 'control' arm comprises a sample of size n from a Normal distribution with mean 0 and variance σ^2, whilst the 'treatment' arm consists of a sample of size n from a Normal distribution with mean θ_i and variance σ^2.
Step 3	Fit the meta-analysis model (6.4) using the R package meta with REML, naively assuming normality at both levels.
Step 4a	Extract the estimates of the overall treatment effect θ and the between study standard deviation τ. Also, construct confidence intervals for θ using equations (6.5) and (6.6), as well as the prediction interval for a new treatment effect using equation (6.7).
Step 4b	Check to see whether the confidence intervals contain the true overall treatment effect $\theta = 1$. Also, simulate a new treatment effect from the true random effects distribution and determine whether the prediction interval contains the new treatment effect.
Step 5	Repeat **Steps 1-4b** $10,000$ times and report the coverage of the confidence and prediction intervals.

Table 6.4: Outline of the simulation study assessing the impact of asymmetry in the random effects of meta-analyses.

6.3.3 Results

Coverage of confidence intervals derived using the Normal distribution

Tables 6.5, 6.6 and 6.7 display the coverage of the $100(1-\alpha)\%$ confidence intervals for the treatment effect, where the confidence intervals are constructed by assuming a Normal distribution and using equation (6.5), for $n = 30, 100$ and 1000 respectively. It is apparent that, regardless of the number of observations in each study, if k is small then the coverage of the confidence intervals is uniformly poor across all distributions. Indeed, for each level α, the confidence intervals are consistently too short when $k < 10$. There is some evidence that asymmetry in the random effects may exacerbate this effect, as the coverage is especially poor when the random effects are highly asymmetric. For example, in Table 6.7 (where $n = 1000$), the coverage of the 95% confidence based on $k = 7$ studies with Log-Normal random effects is 0.817 compared to 0.903

for normally distributed random effects. Both fall some way short of the expected coverage of 0.95, but the discord is much greater in the asymmetric case.

On the other hand, this discrepancy is much smaller when there are a large number of studies ($k = 100$), even if there are only a small number of observations n in each arm. For example, in Table 6.5 (where $n = 30$) the coverage for every distribution is reasonably good when there are $k = 100$ studies. Indeed, the 95% confidence interval almost every distribution achieves good coverage, with coverage of 0.931 for Log-Normal random effects and 0.950 for Normal random effects. Therefore, we conclude that, in the most commonly encountered situations, asymmetry in the random effects does not have a major impact on the confidence intervals for the average treatment effect. Whilst it is true that the confidence intervals are generally poor for $k < 10$, this is also the case when the random effects are normally distributed and so this is more likely to be a problem caused by having too few studies.

Coverage of confidence intervals derived using the t distribution

Tables 6.8, 6.9 and 6.10 display the coverage of the $100(1 - \alpha)\%$ confidence intervals for the treatment effect, where the confidence intervals are constructed by assuming a t distribution and using equation (6.6), for $n = 30, 100$ and 1000 respectively. In this case, even when the number of studies k is small, then the confidence intervals have appropriate coverage. Once again, when there are few studies, there is some evidence that the coverage is more seriously effected when there is substantial asymmetry in the random effects. For example, in Table 6.10 (where $n = 1000$), the coverage of the 95% confidence interval based on $k = 5$ studies with Log-Normal random effects is 0.856 compared to 0.950 for normally distributed random effects.

However, when there are a moderate number of studies, confidence intervals based on t are reasonably accurate regardless of the random effects distribution. For example, in Table 6.8 (where $n = 30$) the coverage for every distribution is reasonably good when there are $k = 10$ studies. For example, the coverage of the 95% confidence interval for Folded Normal random effects is 0.937, compared to 0.952 for normally distributed random effects. Therefore, once again, it appears that asymmetry in the random effects does not have a major impact on the confidence intervals for the average treatment effect.

		$\alpha = 0.5$					$\alpha = 0.1$				
η	Dist.	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$
0	N	0.438	0.456	0.480	0.488	0.498	0.767	0.822	0.850	0.864	0.896
0.1	NM1	0.421	0.453	0.460	0.468	0.500	0.772	0.831	0.850	0.865	0.895
0.2	NM2	0.411	0.442	0.464	0.454	0.498	0.761	0.822	0.838	0.856	0.899
0.3	NM3	0.415	0.437	0.454	0.471	0.497	0.756	0.813	0.840	0.846	0.891
0.4	NM4	0.420	0.453	0.466	0.480	0.497	0.754	0.807	0.830	0.848	0.894
0.5	SN5	0.440	0.458	0.462	0.480	0.500	0.757	0.816	0.842	0.853	0.894
0.6	SN6	0.427	0.457	0.473	0.485	0.492	0.745	0.813	0.843	0.858	0.896
0.7	SN7	0.435	0.463	0.472	0.482	0.493	0.744	0.807	0.831	0.858	0.898
0.8	SN8	0.434	0.463	0.473	0.481	0.494	0.744	0.801	0.830	0.854	0.895
0.91	LN	0.372	0.408	0.427	0.429	0.483	0.705	0.764	0.790	0.809	0.881
0.95	FN	0.434	0.450	0.470	0.485	0.503	0.735	0.807	0.833	0.857	0.897
1	EXP	0.401	0.428	0.440	0.459	0.495	0.712	0.775	0.816	0.832	0.886
		$\alpha = 0.05$					$\alpha = 0.01$				
η	Dist.	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$
0	N	0.815	0.871	0.897	0.921	0.950	0.884	0.938	0.958	0.972	0.987
0.1	NM1	0.820	0.888	0.906	0.921	0.945	0.891	0.939	0.956	0.969	0.987
0.2	NM2	0.819	0.880	0.901	0.910	0.949	0.893	0.934	0.951	0.964	0.989
0.3	NM3	0.820	0.860	0.890	0.909	0.946	0.884	0.934	0.950	0.959	0.987
0.4	NM4	0.806	0.863	0.890	0.911	0.946	0.877	0.925	0.947	0.960	0.988
0.5	SN5	0.798	0.869	0.889	0.906	0.948	0.867	0.923	0.949	0.963	0.988
0.6	SN6	0.793	0.860	0.887	0.905	0.948	0.866	0.920	0.945	0.958	0.988
0.7	SN7	0.795	0.867	0.885	0.908	0.946	0.861	0.914	0.942	0.956	0.986
0.8	SN8	0.786	0.861	0.882	0.903	0.944	0.860	0.911	0.941	0.958	0.985
0.91	LN	0.763	0.808	0.838	0.855	0.931	0.833	0.878	0.897	0.914	0.976
0.95	FN	0.787	0.860	0.885	0.901	0.949	0.855	0.917	0.937	0.957	0.987
1	EXP	0.766	0.830	0.856	0.876	0.943	0.830	0.887	0.912	0.937	0.984

Table 6.5: The coverage of the $100(1 - \alpha)\%$ confidence interval for the overall treatment effect constructed using the Normal distribution (6.5) and based on a random effects meta-analysis with k studies consisting of $n = 30$ observations in each arm. The random effects are drawn from a variety of random effects distributions (Dist.), but fitted using equation (6.4) with REML, naively assuming normality at both levels.

		$\alpha = 0.5$					$\alpha = 0.1$				
η	Dist.	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$
0	N	0.443	0.462	0.480	0.477	0.488	0.758	0.824	0.847	0.861	0.898
0.1	NM1	0.421	0.452	0.459	0.466	0.499	0.768	0.828	0.840	0.866	0.895
0.2	NM2	0.411	0.440	0.456	0.469	0.498	0.759	0.817	0.834	0.858	0.892
0.3	NM3	0.411	0.446	0.455	0.474	0.500	0.750	0.813	0.829	0.851	0.890
0.4	NM4	0.411	0.451	0.469	0.483	0.503	0.740	0.802	0.828	0.848	0.894
0.5	SN5	0.431	0.470	0.471	0.478	0.504	0.750	0.809	0.844	0.857	0.896
0.6	SN6	0.423	0.467	0.473	0.480	0.496	0.741	0.801	0.837	0.855	0.895
0.7	SN7	0.428	0.452	0.468	0.484	0.509	0.735	0.801	0.838	0.857	0.895
0.8	SN8	0.434	0.463	0.474	0.488	0.490	0.735	0.808	0.826	0.852	0.897
0.91	LN	0.378	0.407	0.431	0.435	0.486	0.669	0.740	0.767	0.801	0.878
0.95	FN	0.432	0.468	0.474	0.473	0.485	0.724	0.808	0.828	0.856	0.892
1	EXP	0.401	0.425	0.448	0.469	0.504	0.697	0.770	0.800	0.821	0.890
		$\alpha = 0.05$					$\alpha = 0.01$				
η	Dist.	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$
0	N	0.814	0.880	0.906	0.919	0.948	0.880	0.939	0.962	0.973	0.989
0.1	NM1	0.822	0.880	0.901	0.920	0.944	0.882	0.935	0.960	0.968	0.988
0.2	NM2	0.817	0.878	0.901	0.913	0.945	0.885	0.939	0.950	0.971	0.985
0.3	NM3	0.811	0.857	0.887	0.898	0.943	0.875	0.932	0.953	0.959	0.988
0.4	NM4	0.794	0.861	0.887	0.908	0.947	0.869	0.919	0.942	0.960	0.986
0.5	SN5	0.789	0.866	0.889	0.906	0.946	0.856	0.919	0.946	0.959	0.987
0.6	SN6	0.786	0.858	0.891	0.905	0.947	0.850	0.917	0.944	0.957	0.989
0.7	SN7	0.783	0.857	0.885	0.904	0.946	0.844	0.919	0.939	0.954	0.988
0.8	SN8	0.787	0.857	0.882	0.905	0.944	0.844	0.913	0.938	0.951	0.987
0.91	LN	0.728	0.798	0.807	0.849	0.934	0.796	0.851	0.880	0.903	0.971
0.95	FN	0.783	0.849	0.885	0.908	0.943	0.844	0.908	0.942	0.953	0.986
1	EXP	0.742	0.808	0.846	0.875	0.942	0.804	0.878	0.904	0.929	0.980

Table 6.6: The coverage of the $100(1 - \alpha)\%$ confidence interval for the overall treatment effect constructed using the Normal distribution (6.5) and based on a random effects meta-analysis with k studies consisting of $n = 100$ observations in each arm. The random effects are drawn from a variety of random effects distributions (Dist.), but fitted using equation (6.4) with REML, naively assuming normality at both levels.

		$\alpha = 0.5$					$\alpha = 0.1$				
η	Dist.	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$
0	N	0.433	0.475	0.464	0.480	0.496	0.768	0.824	0.852	0.864	0.897
0.1	NM1	0.420	0.453	0.462	0.473	0.495	0.762	0.830	0.850	0.864	0.896
0.2	NM2	0.408	0.441	0.458	0.460	0.504	0.760	0.824	0.839	0.860	0.888
0.3	NM3	0.402	0.434	0.451	0.466	0.512	0.750	0.812	0.827	0.847	0.892
0.4	NM4	0.413	0.448	0.470	0.488	0.502	0.735	0.803	0.831	0.856	0.895
0.5	SN5	0.428	0.463	0.463	0.479	0.496	0.740	0.807	0.837	0.857	0.897
0.6	SN6	0.433	0.457	0.470	0.482	0.505	0.735	0.800	0.842	0.857	0.896
0.7	SN7	0.430	0.463	0.465	0.481	0.496	0.734	0.800	0.830	0.854	0.902
0.8	SN8	0.429	0.463	0.465	0.479	0.503	0.733	0.802	0.832	0.850	0.896
0.91	LN	0.373	0.404	0.422	0.440	0.487	0.671	0.739	0.761	0.789	0.878
0.95	FN	0.439	0.459	0.470	0.471	0.492	0.740	0.802	0.828	0.859	0.892
1	EXP	0.398	0.442	0.453	0.459	0.491	0.696	0.767	0.805	0.827	0.889
		$\alpha = 0.05$					$\alpha = 0.01$				
η	Dist.	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$
0	N	0.810	0.873	0.903	0.925	0.945	0.876	0.935	0.959	0.970	0.990
0.1	NM1	0.813	0.879	0.903	0.919	0.946	0.884	0.940	0.960	0.971	0.988
0.2	NM2	0.815	0.876	0.895	0.908	0.945	0.879	0.937	0.953	0.968	0.986
0.3	NM3	0.795	0.860	0.882	0.905	0.943	0.874	0.931	0.948	0.959	0.990
0.4	NM4	0.787	0.861	0.886	0.905	0.945	0.857	0.918	0.940	0.952	0.987
0.5	SN5	0.790	0.863	0.885	0.905	0.944	0.853	0.921	0.945	0.959	0.988
0.6	SN6	0.787	0.861	0.887	0.906	0.944	0.851	0.922	0.942	0.958	0.987
0.7	SN7	0.788	0.851	0.886	0.899	0.946	0.843	0.915	0.939	0.957	0.986
0.8	SN8	0.778	0.857	0.887	0.899	0.941	0.841	0.910	0.935	0.956	0.988
0.91	LN	0.720	0.778	0.817	0.843	0.924	0.784	0.844	0.873	0.898	0.971
0.95	FN	0.782	0.851	0.886	0.902	0.946	0.833	0.909	0.934	0.949	0.987
1	EXP	0.745	0.815	0.844	0.873	0.942	0.804	0.870	0.901	0.925	0.981

Table 6.7: The coverage of the $100(1 - \alpha)\%$ confidence interval for the overall treatment effect constructed using the Normal distribution (6.5) and based on a random effects meta-analysis with k studies consisting of $n = 1000$ observations in each arm. The random effects are drawn from a variety of random effects distributions (Dist.), but fitted using equation (6.4) with REML, naively assuming normality at both levels.

		$\alpha = 0.5$					$\alpha = 0.1$				
η	Dist.	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$
0	N	0.498	0.500	0.502	0.507	0.511	0.896	0.902	0.897	0.901	0.905
0.1	NM1	0.496	0.488	0.485	0.487	0.499	0.905	0.904	0.905	0.903	0.898
0.2	NM2	0.480	0.479	0.472	0.485	0.490	0.905	0.897	0.895	0.893	0.901
0.3	NM3	0.475	0.481	0.484	0.491	0.509	0.895	0.888	0.887	0.886	0.901
0.4	NM4	0.475	0.478	0.492	0.495	0.497	0.883	0.884	0.885	0.887	0.896
0.5	SN5	0.497	0.493	0.489	0.497	0.500	0.881	0.880	0.887	0.893	0.900
0.6	SN6	0.489	0.492	0.499	0.496	0.502	0.877	0.888	0.886	0.891	0.895
0.7	SN7	0.498	0.491	0.494	0.500	0.498	0.882	0.878	0.882	0.888	0.901
0.8	SN8	0.498	0.502	0.493	0.496	0.494	0.876	0.881	0.881	0.886	0.893
0.91	LN	0.448	0.447	0.451	0.457	0.480	0.838	0.828	0.838	0.843	0.883
0.95	FN	0.492	0.497	0.496	0.496	0.507	0.869	0.874	0.883	0.888	0.901
1	EXP	0.464	0.473	0.479	0.491	0.502	0.843	0.845	0.856	0.867	0.890
		$\alpha = 0.05$					$\alpha = 0.01$				
η	Dist.	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$
0	N	0.949	0.952	0.948	0.952	0.953	0.990	0.990	0.989	0.991	0.990
0.1	NM1	0.952	0.952	0.954	0.952	0.947	0.991	0.991	0.991	0.990	0.988
0.2	NM2	0.952	0.947	0.949	0.945	0.950	0.991	0.991	0.989	0.988	0.989
0.3	NM3	0.947	0.942	0.936	0.937	0.947	0.988	0.985	0.984	0.985	0.987
0.4	NM4	0.939	0.936	0.938	0.940	0.945	0.987	0.986	0.984	0.984	0.986
0.5	SN5	0.937	0.932	0.938	0.945	0.951	0.987	0.982	0.983	0.985	0.989
0.6	SN6	0.933	0.939	0.933	0.941	0.948	0.984	0.984	0.980	0.983	0.988
0.7	SN7	0.932	0.930	0.932	0.939	0.950	0.986	0.981	0.980	0.980	0.988
0.8	SN8	0.934	0.932	0.930	0.932	0.944	0.985	0.980	0.978	0.978	0.988
0.91	LN	0.909	0.888	0.892	0.896	0.932	0.979	0.963	0.955	0.956	0.978
0.95	FN	0.924	0.924	0.930	0.937	0.948	0.981	0.976	0.977	0.979	0.989
1	EXP	0.905	0.897	0.908	0.915	0.943	0.978	0.963	0.963	0.963	0.983

Table 6.8: The coverage of the $100(1 - \alpha)\%$ confidence interval for the overall treatment effect constructed using the t distribution (6.6) and based on a random effects meta-analysis with k studies consisting of $n = 30$ observations in each arm. The random effects are drawn from a variety of random effects distributions (Dist.), but fitted using equation (6.4) with REML, naively assuming normality at both levels.

		$\alpha = 0.5$					$\alpha = 0.1$				
η	Dist.	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$
0	N	0.504	0.504	0.495	0.499	0.504	0.902	0.902	0.900	0.902	0.899
0.1	NM1	0.498	0.489	0.486	0.491	0.499	0.906	0.902	0.899	0.901	0.899
0.2	NM2	0.476	0.481	0.477	0.484	0.493	0.901	0.898	0.892	0.894	0.895
0.3	NM3	0.471	0.469	0.472	0.478	0.492	0.896	0.891	0.879	0.886	0.896
0.4	NM4	0.474	0.486	0.495	0.492	0.503	0.884	0.879	0.883	0.890	0.897
0.5	SN5	0.484	0.496	0.490	0.495	0.496	0.877	0.884	0.888	0.890	0.902
0.6	SN6	0.502	0.499	0.497	0.500	0.508	0.869	0.881	0.884	0.896	0.901
0.7	SN7	0.499	0.497	0.495	0.497	0.513	0.874	0.883	0.884	0.885	0.899
0.8	SN8	0.490	0.503	0.492	0.493	0.512	0.870	0.880	0.880	0.882	0.905
0.91	LN	0.441	0.426	0.440	0.448	0.483	0.818	0.809	0.819	0.831	0.873
0.95	FN	0.499	0.499	0.494	0.499	0.497	0.865	0.871	0.881	0.885	0.902
1	EXP	0.459	0.470	0.476	0.473	0.496	0.826	0.833	0.850	0.854	0.889
		$\alpha = 0.05$					$\alpha = 0.01$				
η	Dist.	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$
0	N	0.951	0.954	0.949	0.950	0.951	0.990	0.989	0.989	0.988	0.990
0.1	NM1	0.952	0.953	0.951	0.954	0.951	0.990	0.990	0.990	0.991	0.990
0.2	NM2	0.949	0.948	0.947	0.946	0.945	0.990	0.990	0.989	0.988	0.986
0.3	NM3	0.949	0.942	0.935	0.939	0.949	0.989	0.987	0.985	0.984	0.988
0.4	NM4	0.941	0.934	0.933	0.936	0.947	0.991	0.985	0.983	0.981	0.990
0.5	SN5	0.933	0.936	0.938	0.937	0.952	0.985	0.984	0.980	0.981	0.990
0.6	SN6	0.928	0.930	0.934	0.943	0.950	0.984	0.981	0.981	0.985	0.990
0.7	SN7	0.931	0.931	0.929	0.934	0.950	0.984	0.979	0.978	0.980	0.989
0.8	SN8	0.923	0.930	0.929	0.933	0.950	0.980	0.979	0.974	0.979	0.989
0.91	LN	0.887	0.868	0.874	0.882	0.926	0.970	0.948	0.942	0.942	0.974
0.95	FN	0.920	0.923	0.929	0.935	0.948	0.977	0.975	0.976	0.977	0.987
1	EXP	0.890	0.889	0.898	0.905	0.943	0.968	0.955	0.954	0.959	0.984

Table 6.9: The coverage of the $100(1 - \alpha)\%$ confidence interval for the overall treatment effect constructed using the t distribution (6.6) and based on a random effects meta-analysis with k studies consisting of $n = 100$ observations in each arm. The random effects are drawn from a variety of random effects distributions (Dist.), but fitted using equation (6.4) with REML, naively assuming normality at both levels.

		$\alpha = 0.5$					$\alpha = 0.1$				
η	Dist.	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$
0	N	0.499	0.500	0.509	0.502	0.499	0.904	0.893	0.901	0.901	0.897
0.1	NM1	0.491	0.484	0.490	0.490	0.501	0.905	0.903	0.902	0.902	0.896
0.2	NM2	0.480	0.472	0.483	0.471	0.493	0.901	0.892	0.894	0.891	0.899
0.3	NM3	0.476	0.477	0.481	0.492	0.506	0.901	0.887	0.885	0.890	0.895
0.4	NM4	0.473	0.491	0.491	0.494	0.501	0.889	0.882	0.882	0.882	0.898
0.5	SN5	0.503	0.497	0.494	0.502	0.491	0.886	0.879	0.883	0.888	0.897
0.6	SN6	0.496	0.496	0.487	0.502	0.504	0.875	0.881	0.885	0.880	0.900
0.7	SN7	0.499	0.495	0.493	0.491	0.492	0.871	0.870	0.879	0.885	0.898
0.8	SN8	0.501	0.489	0.501	0.494	0.492	0.867	0.868	0.882	0.887	0.901
0.91	LN	0.436	0.434	0.438	0.453	0.483	0.798	0.806	0.809	0.824	0.880
0.95	FN	0.503	0.494	0.496	0.496	0.497	0.866	0.876	0.877	0.890	0.899
1	EXP	0.464	0.468	0.474	0.465	0.497	0.817	0.836	0.841	0.856	0.893
		$\alpha = 0.05$					$\alpha = 0.01$				
η	Dist.	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$
0	N	0.952	0.950	0.952	0.950	0.947	0.990	0.991	0.989	0.990	0.990
0.1	NM1	0.952	0.951	0.953	0.955	0.948	0.990	0.991	0.991	0.991	0.989
0.2	NM2	0.951	0.946	0.947	0.943	0.947	0.990	0.989	0.988	0.987	0.986
0.3	NM3	0.949	0.941	0.938	0.936	0.947	0.992	0.987	0.985	0.982	0.986
0.4	NM4	0.942	0.935	0.933	0.934	0.946	0.988	0.983	0.983	0.980	0.989
0.5	SN5	0.939	0.930	0.931	0.938	0.946	0.987	0.982	0.979	0.981	0.990
0.6	SN6	0.928	0.931	0.930	0.931	0.948	0.983	0.978	0.976	0.980	0.988
0.7	SN7	0.929	0.922	0.928	0.930	0.947	0.982	0.976	0.976	0.978	0.989
0.8	SN8	0.921	0.923	0.930	0.931	0.949	0.980	0.975	0.977	0.976	0.988
0.91	LN	0.870	0.856	0.862	0.871	0.929	0.963	0.939	0.933	0.934	0.974
0.95	FN	0.920	0.926	0.925	0.936	0.949	0.980	0.977	0.973	0.977	0.988
1	EXP	0.884	0.888	0.891	0.904	0.942	0.964	0.952	0.952	0.954	0.985

Table 6.10: The coverage of the $100(1-\alpha)\%$ confidence interval for the overall treatment effect constructed using the t distribution (6.6) and based on a random effects meta-analysis with k studies consisting of $n = 1000$ observations in each arm. The random effects are drawn from a variety of random effects distributions (Dist.), but fitted using equation (6.4) with REML, naively assuming normality at both levels.

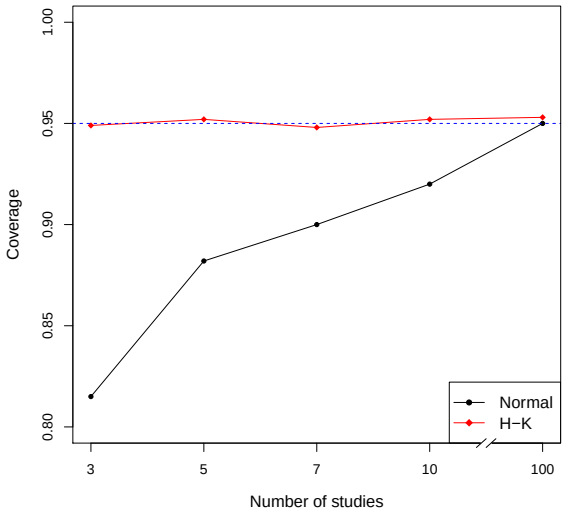

Figure 6.3: Comparing the coverage of 95% confidence intervals constructed using equation (6.5) (based on the Normal distribution) and those using equation (6.6) (based on the t distribution), using a random effects meta-analysis with k normally distributed random effects and k studies consisting of $n = 30$ observations in each arm.

It is apparent, comparing the coverage of the confidence intervals constructed using the Normal distribution and those using the Knapp and Hartung method based on the t distribution, that the latter consistently performs better in terms of coverage. Figure 6.3 compares the coverage of the 95% confidence intervals constructed using the two methods for normally distributed random effects. Using the Normal confidence intervals, it is clear that one requires at least 10 studies before the coverage provides an acceptable approximation to 0.95 (namely, 0.921), and that the confidence intervals are consistently anti-conservative. On the other hand, the Hartung-Knapp confidence intervals are generally much closer to 0.95, even for just $k = 3$ studies.

Coverage of prediction intervals

Tables 6.11, 6.12 and 6.13 display the coverage of the $100(1-\alpha)\%$ prediction intervals for a new treatment effect, where the prediction intervals are constructed by assuming a t distribution and using equation (6.7), for $n = 30, 100$ and 1000 respectively. It is clear that, when k is small, constructing accurate prediction intervals is particularly problematic when there is asymmetry in

the random effects. Indeed, when there are only $k = 3$ studies the prediction intervals are wildly over conservative, to such a degree that for many of the symmetric or asymmetric random effects distributions the 95% and 99% confidence intervals have coverage equal to 1. This is perhaps not a surprise, as it difficult to make accurate predictions about a future treatment effect based on only a few studies.

On the other hand, when the number of studies is large ($k = 100$) the coverage of the prediction intervals is much better, in most cases. For example, as seen in Table 6.13 (where $n = 1000$), if the meta-analysis is comprised of $k = 100$ studies then the coverage of the 50% prediction intervals is equal to 0.494 for Normal random effects. For a Normal random effects distribution with at least 30 individuals in each study, generally at least 5 studies are required to ensure the prediction intervals have suitable coverage. However, for Log-Normal random effects the prediction interval is too wide with coverage of 0.65. By contrast, for the highly asymmetric SN8 distribution ($\eta = 0.8$) the prediction interval is too short with coverage of 0.462. Therefore, it is apparent that the prediction intervals are sensitive to the random effects distribution, however, the differences are most marked when $\alpha = 0.5$ and they are less obvious for the 95% and 99% prediction intervals. Indeed, for these wider prediction intervals, the coverage is relatively good, but there is still some slight deviation for the most asymmetric random effects distributions. For example, when $k = 7$ and $n = 1000$ the coverage for the 95% prediction interval based on Log-Normal random effects is 0.925 compared to 0.958 for the Normal distribution. However, this disparity is less noticeable when $k = 100$ as the coverage of the 95% prediction interval is 0.953 and 0.950 for the Log-Normal and Normal distribution respectively. Figure 6.4 displays these results graphically and compares the coverage of 50% and 95% prediction intervals constructed assuming normally distributed random effects, for Normal and Log-Normal random effects.

Although the 95% prediction intervals display reasonable coverage even when the true distribution was asymmetric, the results for the shorter prediction intervals highlight the fact that the impact on probabilistic inferences is not yet clear. Next, we examine this problem in more detail through the use of an example, and assesses the consequences for clinically relevant inferences.

		$\alpha = 0.5$					$\alpha = 0.1$				
η	Dist.	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$
0	N	0.561	0.511	0.504	0.507	0.499	0.959	0.910	0.900	0.894	0.897
0.1	NM1	0.579	0.521	0.516	0.526	0.548	0.957	0.905	0.896	0.898	0.910
0.2	NM2	0.578	0.530	0.525	0.540	0.574	0.958	0.899	0.896	0.893	0.915
0.3	NM3	0.587	0.535	0.538	0.547	0.575	0.954	0.897	0.896	0.896	0.907
0.4	NM4	0.576	0.531	0.520	0.532	0.530	0.958	0.905	0.899	0.904	0.904
0.5	SN5	0.575	0.499	0.493	0.485	0.493	0.959	0.911	0.901	0.902	0.919
0.6	SN6	0.572	0.513	0.490	0.488	0.477	0.952	0.905	0.904	0.909	0.920
0.7	SN7	0.579	0.508	0.498	0.487	0.480	0.956	0.906	0.906	0.909	0.922
0.8	SN8	0.574	0.493	0.494	0.488	0.472	0.953	0.904	0.907	0.914	0.929
0.91	LN	0.590	0.542	0.554	0.567	0.646	0.946	0.874	0.884	0.900	0.936
0.95	FN	0.576	0.487	0.492	0.479	0.459	0.956	0.900	0.904	0.912	0.930
1	EXP	0.592	0.528	0.526	0.517	0.530	0.945	0.888	0.893	0.899	0.922
		$\alpha = 0.05$					$\alpha = 0.01$				
η	Dist.	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$
0	N	0.991	0.956	0.951	0.949	0.950	1.000	0.988	0.989	0.989	0.989
0.1	NM1	0.989	0.946	0.942	0.941	0.951	1.000	0.987	0.984	0.981	0.981
0.2	NM2	0.990	0.943	0.935	0.936	0.945	1.000	0.984	0.979	0.979	0.974
0.3	NM3	0.987	0.940	0.939	0.936	0.942	1.000	0.984	0.979	0.975	0.973
0.4	NM4	0.987	0.953	0.943	0.944	0.948	1.000	0.987	0.983	0.981	0.978
0.5	SN5	0.990	0.950	0.940	0.947	0.954	1.000	0.987	0.985	0.982	0.985
0.6	SN6	0.990	0.943	0.941	0.943	0.950	1.000	0.985	0.982	0.980	0.982
0.7	SN7	0.989	0.948	0.940	0.947	0.954	1.000	0.984	0.983	0.984	0.982
0.8	SN8	0.988	0.948	0.944	0.947	0.950	1.000	0.983	0.985	0.982	0.983
0.91	LN	0.985	0.909	0.910	0.923	0.951	1.000	0.962	0.956	0.956	0.967
0.95	FN	0.987	0.947	0.948	0.947	0.955	1.000	0.981	0.979	0.978	0.981
1	EXP	0.985	0.923	0.926	0.927	0.945	1.000	0.970	0.965	0.965	0.973

Table 6.11: The coverage of the $100(1 - \alpha)\%$ prediction interval for a new treatment effect constructed using equation (6.7) and based on a random effects meta-analysis with k studies consisting of $n = 30$ observations in each arm. The random effects are drawn from a variety of random effects distributions (Dist.), but fitted using equation (6.4) with REML, naively assuming normality at both levels.

		$\alpha = 0.5$					$\alpha = 0.1$				
η	Dist.	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$
0	N	0.571	0.509	0.499	0.492	0.485	0.969	0.922	0.907	0.903	0.901
0.1	NM1	0.580	0.517	0.532	0.528	0.544	0.965	0.914	0.900	0.903	0.914
0.2	NM2	0.583	0.538	0.530	0.542	0.578	0.963	0.907	0.893	0.899	0.916
0.3	NM3	0.589	0.540	0.549	0.548	0.567	0.962	0.908	0.892	0.902	0.910
0.4	NM4	0.586	0.519	0.521	0.519	0.531	0.966	0.914	0.901	0.901	0.905
0.5	SN5	0.578	0.515	0.502	0.499	0.489	0.963	0.908	0.913	0.912	0.915
0.6	SN6	0.585	0.513	0.499	0.486	0.480	0.962	0.915	0.911	0.912	0.921
0.7	SN7	0.588	0.503	0.496	0.487	0.473	0.960	0.910	0.909	0.915	0.925
0.8	SN8	0.590	0.502	0.485	0.480	0.474	0.958	0.910	0.913	0.915	0.921
0.91	LN	0.603	0.552	0.554	0.573	0.649	0.934	0.889	0.890	0.907	0.936
0.95	FN	0.586	0.494	0.496	0.481	0.471	0.957	0.912	0.909	0.915	0.920
1	EXP	0.597	0.534	0.515	0.517	0.534	0.944	0.894	0.905	0.906	0.930
		$\alpha = 0.05$					$\alpha = 0.01$				
η	Dist.	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$
0	N	0.989	0.966	0.958	0.959	0.951	1.000	0.993	0.992	0.990	0.989
0.1	NM1	0.990	0.959	0.947	0.944	0.953	1.000	0.991	0.987	0.978	0.980
0.2	NM2	0.989	0.952	0.943	0.939	0.943	1.000	0.990	0.980	0.978	0.976
0.3	NM3	0.989	0.947	0.940	0.936	0.940	1.000	0.987	0.980	0.979	0.973
0.4	NM4	0.990	0.953	0.941	0.941	0.940	1.000	0.987	0.983	0.982	0.978
0.5	SN5	0.988	0.957	0.952	0.947	0.953	1.000	0.991	0.987	0.985	0.983
0.6	SN6	0.987	0.956	0.947	0.950	0.953	1.000	0.990	0.987	0.984	0.983
0.7	SN7	0.987	0.954	0.948	0.951	0.956	1.000	0.991	0.987	0.983	0.982
0.8	SN8	0.987	0.954	0.949	0.949	0.953	1.000	0.988	0.983	0.981	0.982
0.91	LN	0.980	0.922	0.919	0.926	0.949	1.000	0.967	0.960	0.958	0.968
0.95	FN	0.986	0.949	0.951	0.950	0.951	1.000	0.987	0.986	0.983	0.980
1	EXP	0.980	0.933	0.928	0.935	0.943	1.000	0.974	0.968	0.967	0.970

Table 6.12: The coverage of the $100(1 - \alpha)\%$ prediction interval for a new treatment effect constructed using equation (6.7) and based on a random effects meta-analysis with k studies consisting of $n = 100$ observations in each arm. The random effects are drawn from a variety of random effects distributions (Dist.), but fitted using equation (6.4) with REML, naively assuming normality at both levels.

		$\alpha = 0.5$					$\alpha = 0.1$				
η	Dist.	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$
0	N	0.587	0.513	0.497	0.502	0.494	0.979	0.927	0.915	0.906	0.900
0.1	NM1	0.581	0.528	0.526	0.533	0.544	0.974	0.914	0.905	0.903	0.912
0.2	NM2	0.586	0.534	0.538	0.537	0.575	0.970	0.906	0.900	0.897	0.918
0.3	NM3	0.595	0.538	0.538	0.544	0.569	0.966	0.906	0.897	0.895	0.909
0.4	NM4	0.587	0.527	0.516	0.532	0.530	0.968	0.910	0.904	0.903	0.902
0.5	SN5	0.587	0.518	0.501	0.498	0.485	0.971	0.917	0.915	0.911	0.914
0.6	SN6	0.586	0.515	0.489	0.485	0.481	0.972	0.918	0.915	0.910	0.923
0.7	SN7	0.584	0.514	0.490	0.487	0.476	0.968	0.919	0.916	0.915	0.927
0.8	SN8	0.590	0.499	0.494	0.479	0.462	0.969	0.913	0.915	0.915	0.926
0.91	LN	0.607	0.551	0.565	0.564	0.650	0.945	0.901	0.901	0.910	0.936
0.95	FN	0.587	0.508	0.483	0.475	0.469	0.968	0.919	0.913	0.916	0.921
1	EXP	0.613	0.536	0.523	0.526	0.532	0.955	0.899	0.908	0.912	0.923
		$\alpha = 0.05$					$\alpha = 0.01$				
η	Dist.	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$	$k = 3$	$k = 5$	$k = 7$	$k = 10$	$k = 100$
0	N	0.993	0.964	0.958	0.953	0.950	0.999	0.995	0.993	0.991	0.990
0.1	NM1	0.990	0.958	0.945	0.947	0.953	1.000	0.990	0.986	0.983	0.979
0.2	NM2	0.991	0.958	0.938	0.939	0.940	0.999	0.990	0.981	0.978	0.973
0.3	NM3	0.989	0.951	0.938	0.936	0.942	1.000	0.990	0.979	0.974	0.977
0.4	NM4	0.993	0.955	0.946	0.941	0.942	1.000	0.991	0.983	0.979	0.983
0.5	SN5	0.990	0.963	0.948	0.949	0.952	1.000	0.993	0.989	0.985	0.982
0.6	SN6	0.992	0.957	0.955	0.951	0.952	1.000	0.992	0.987	0.986	0.983
0.7	SN7	0.990	0.957	0.954	0.953	0.951	1.000	0.990	0.986	0.983	0.979
0.8	SN8	0.988	0.956	0.950	0.948	0.955	1.000	0.991	0.987	0.982	0.981
0.91	LN	0.980	0.924	0.925	0.930	0.953	0.999	0.969	0.963	0.958	0.973
0.95	FN	0.989	0.960	0.951	0.951	0.953	1.000	0.989	0.984	0.982	0.982
1	EXP	0.982	0.935	0.931	0.939	0.949	0.999	0.977	0.968	0.966	0.968

Table 6.13: The coverage of the $100(1-\alpha)\%$ prediction interval for a new treatment effect constructed using equation (6.7) and based on a random effects meta-analysis with k studies consisting of $n = 1000$ observations in each arm. The random effects are drawn from a variety of random effects distributions (Dist.), but fitted using equation (6.4) with REML, naively assuming normality at both levels.

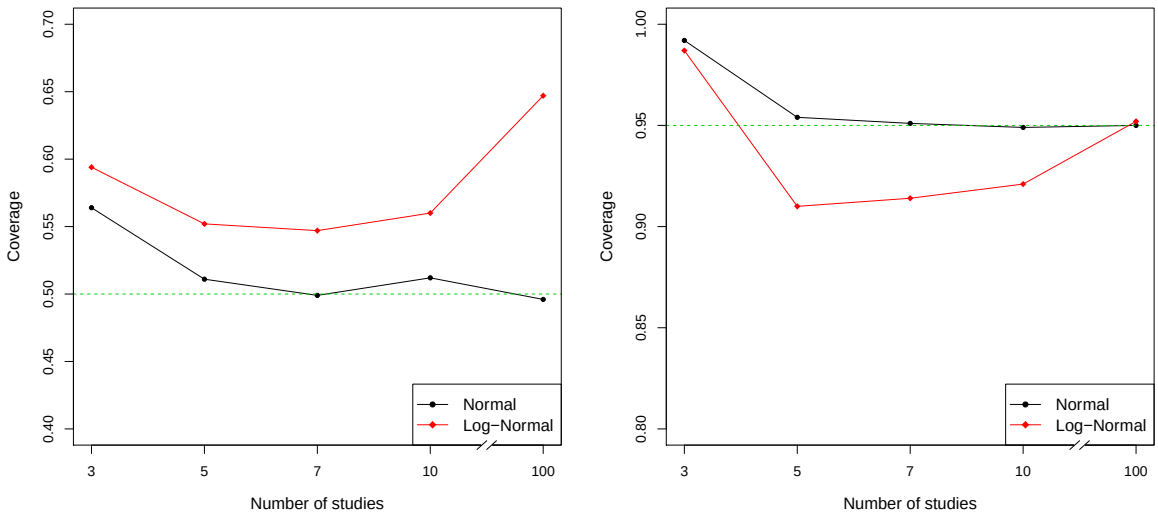

Figure 6.4: Comparing the coverage of 50% (left) and 95% (right) prediction intervals constructed using equation (6.7) (assuming Normal random effects) for Normal and Log-Normal random effects, based on a random effects meta-analysis with k studies consisting of $n = 30$ observations in each arm.

6.4 Impact on probabilistic inferences

In the simulation study in the previous section we assessed the accuracy of confidence and prediction intervals based on their coverage. It was shown that, provided there are a sufficiently large number of studies with a reasonably large sample size ($n \geq 30$), the 95% confidence and prediction intervals perform well in terms of coverage, irrespective of the random effects distribution. However, a 95% confidence (or prediction) interval can contain the overall (or predicted) treatment effect roughly 95% of the time, but still fail to provide useful insight regarding the range of possible values. If, for example, the underlying sampling distribution is appreciably skewed then a symmetric confidence or prediction interval can provide adequate coverage without providing a true reflection of the probable range of values. This is akin to the problem of selecting the most suitable credible interval in a Bayesian framework. Indeed, the traditional equal-tail credible interval is only sensible when the underlying distribution is symmetric. When calculating a credible interval for a skewed posterior density, it is recommended to consider the highest posterior density (HPD) interval [12]. This ensures that the resultant interval will contain the

region with the highest posterior probability.

For a random effects meta-analysis, this is more likely to be a problem for prediction intervals than confidence intervals. Indeed, since estimation of the overall treatment effect is based on an average across a number of studies, provided there are a reasonable number of studies, the CLT ensures that the sampling distribution of the overall treatment effect is approximately Normal even if the random effects are appreciably asymmetric. On the other hand, when predicting a treatment effect for a future study, asymmetry in the random effects cannot be so easily disregarded. Indeed, the results of the simulation study for the 50% prediction interval expose the problems caused by a substantially asymmetric random effects distribution, even when there are a large number of studies.

For example, recall Figure 6.2, which shows the centred and scaled density functions for the Normal, SN5, SN8 and Log-Normal distributions. All of these distributions have approximately 95% of their mass between the 2.5% and 97.5% percentiles of the Normal distribution. Thus, a prediction interval constructed by naively assuming a Normal distribution will have coverage close to 95% even if the actual predictive distribution is given by one of the more heavily asymmetric distributions. This is precisely what was observed in the simulation study in the previous section. However, although the naive 95% prediction interval performs as expected in terms of coverage, it is not particularly informative about the possible distribution of future treatment effects. Indeed the symmetric prediction interval conceals the fact that a new treatment effect is far more likely to be found in the right-hand side of the prediction interval.

In order to further emphasise the impact of incorrectly specifying the random effects distribution in this context we consider an example used by Lee and Thompson [72]. Lee and Thompson suggest modelling the random effects using the flexible skewed distribution proposed by Fernandez and Steel [33]. Recall from Chapter 2 that the Fernandez and Steel distribution has density function

$$FAS(z;\gamma) = \frac{2}{\gamma + \frac{1}{\gamma}} \left\{ f\left(\frac{z}{\gamma}\right) \mathrm{I}[z \geq 0] + f\left(\gamma z\right) \mathrm{I}[z < 0] \right\},$$

for any symmetric unimodal density function f and some $\gamma \in (0, \infty)$. When $\gamma = 1$ the dis-

tribution is symmetric, while $\gamma > 1$ and $\gamma < 1$ correspond to asymmetry to the right and left respectively. In particular, Lee and Thompson suggest setting f to be a Normal probability density function to obtain the three-parameter skewed Normal distribution or, alternatively, setting f to be the probability density function of the t distribution (with k degrees of freedom) to obtain the even more flexible four-parameter skewed t distribution, which allows for the possibility of heavier tails.

To illustrate their method Lee and Thompson apply it to a meta-analysis of 70 trials investigating the effectiveness of fluoride toothpaste conducted between 1954 and 1994, collated by Marinho et al. [77]. In this instance the treatment effect is the 'prevented fraction' of decayed and filled dental surfaces. That is, the difference in the mean change in the control and treatment groups from baseline, divided by the mean change in the control group. The treatment effect (and 95% confidence interval) obtained using the original analysis is given by 0.24 $[0.21; 0.28]$.

Lee and Thompson demonstrate that there is clear evidence that the random effects distribution is appreciably skewed. Indeed, Table 6.14 shows the parameter estimates obtained by applying the flexible method given by Lee and Thompson to the meta-analysis. Firstly observe that, for the flexible models that include a skewness parameter γ, this skewness parameter is clearly non-zero. Moreover, there is a clear drop in the deviance information criterion (DIC), a measure of the relative appropriateness of the model, for the more flexible models [117]. In particular, there is a drop of nearly 10 in the DIC for the skewed t compared to the skewed Normal model, indicating a better fit in this case. Further, Figure 6.5 shows the estimated predictive distributions for a new treatment effect by assuming several different parametric forms for the random effects distribution. It is clear from the bottom two plots that introducing a skewness parameter has greatly altered the predictive distribution, which suggests that the random effects are not normally distributed.

Lee and Thompson [72] did not explicitly derive prediction intervals, and so we extend their work by deriving equal-tail prediction intervals for a new treatment effect using the random effects distributions described above. Table 6.15 compares the overall treatment effect and the $50\%, 90\%, 95\%$ and 99% prediction intervals obtained by applying the flexible model proposed

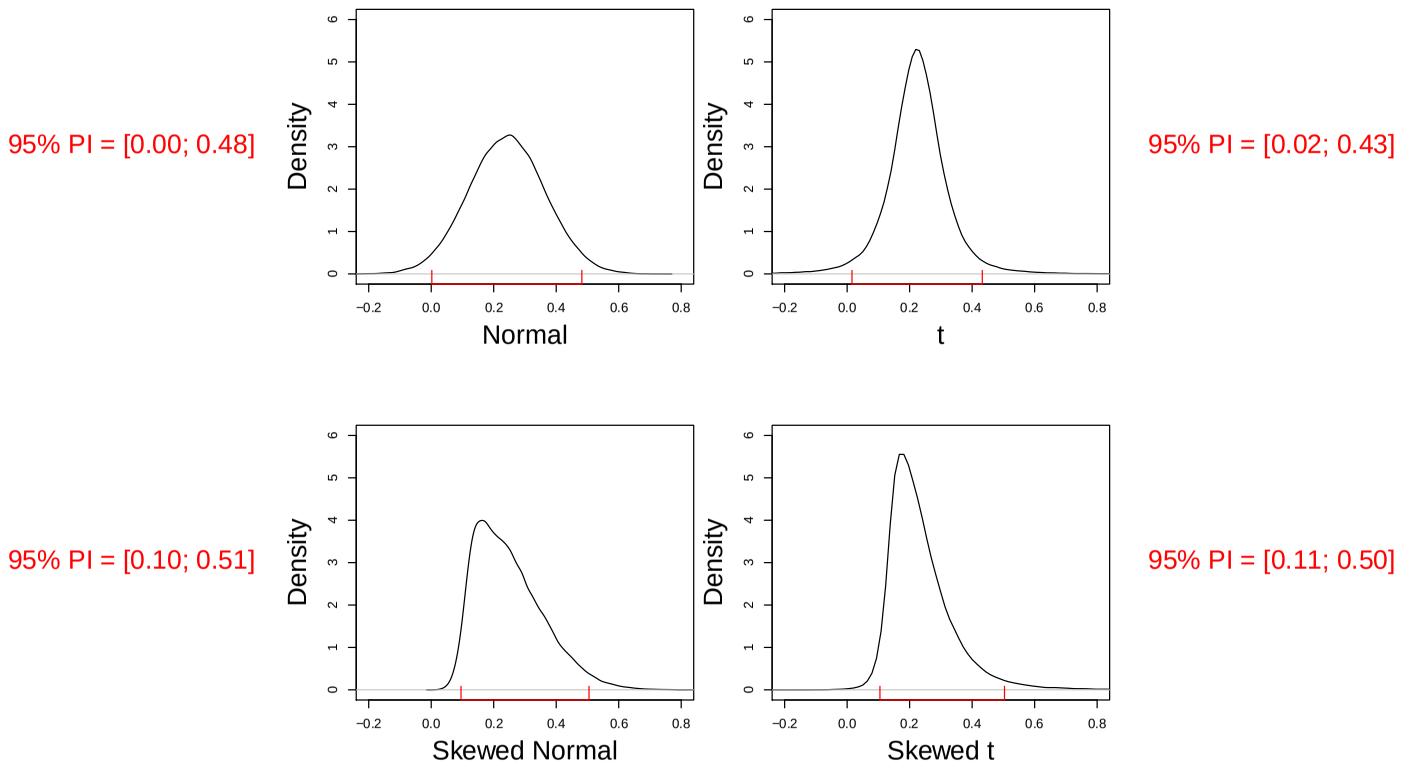

Figure 6.5: Predictive distributions for a treatment effect in a new trial investigating the effectiveness of fluoride toothpaste, by assuming several different parametric forms for the random effects distribution. The plots also include the 95% equal-tail prediction interval.

	μ	σ^2	k	γ	DIC
Normal	0.242 (0.017)	0.015 (0.003)	-	-	−139.2
t	0.224 (0.014)	0.012 (0.005)	3.332 (1.885)	-	−146.2
Skewed Normal	0.250 (0.015)	0.012 (0.003)	-	2.219 (0.593)	−143.5
Skewed t	0.238 (0.015)	0.012 (0.006)	4.005 (4.028)	1.951 (0.567)	−148.7

Table 6.14: Parameter estimates (and standard errors) obtained by applying the flexible method given by Lee and Thompson [72] to a meta-analysis of 70 trials investigating the effectiveness of fluoride toothpaste by assuming several different parametric forms for the random effects distribution.

Random effects distribution	Overall TE (95% CI)	Prediction interval			
		50%	90%	95%	99%
Normal	0.24 [0.21; 0.28]	[0.16; 0.33]	[0.04; 0.44]	[0.00; 0.48]	[−0.07; 0.56]
t	0.22 [0.20; 0.25]	[0.17; 0.28]	[0.07; 0.38]	[0.02; 0.43]	[−0.14; 0.59]
Skewed Normal	0.25 [0.22; 0.28]	[0.17; 0.32]	[0.11; 0.46]	[0.10; 0.51]	[0.07; 0.61]
Skewed t	0.24 [0.21; 0.27]	[0.17; 0.28]	[0.12; 0.43]	[0.11; 0.50]	[0.06; 0.72]

Table 6.15: The overall treatment effect (TE) and 95% confidence interval (CI), as well as the 50%, 90%, 95% and 99% equal-tail prediction intervals for the effectiveness of fluoride toothpaste obtained using the flexible parametric model proposed by Lee and Thompson [72] utilising various assumptions for the random effects distribution.

Random effects distribution	$P[\theta^* > 0]$	$P[\theta^* > 0.1]$	$P[\theta^* > 0.2]$
Normal	0.976	0.878	0.635
t	0.979	0.918	0.625
Skewed Normal	1.000	0.968	0.613
Skewed t	0.999	0.980	0.573

Table 6.16: The probability of a positive treatment effect θ^* in a new trial investigating the effectiveness of fluoride toothpaste by assuming several different parametric forms for the random effects distribution.

by Lee and Thompson [72] and the results obtained by naively assuming that the random effects are normally distributed. More specifically, we use the parameter estimates to derive equal-tail prediction intervals for the flexible random effects distributions. It is apparent that, while the estimate of the overall treatment effect (and its corresponding confidence interval) is not altered dramatically, the prediction intervals are sensitive to the choice of the random effects distribution. For example, if one assumes normality the 95% prediction interval is $[0.00; 0.48]$, however, if one assumes the skewed t distribution the 95% prediction interval is $[0.11; 0.50]$. This reinforces the argument that, while the simulation study suggests that prediction intervals may not be too adversely effected in terms of coverage, there are still repercussions for the inferences that are drawn from the predictive distribution. This is particularly relevant in a Bayesian framework, where one may be interested in making direct probability statements about the treatment effect of a new study (e.g. finding the probability that the treatment might be harmful in a new study, or determining the probability of a new treatment effect being clinically meaningful).

Indeed, Table 6.16 reports the probability that a new treatment effect is greater than zero. The probability that fluoride toothpaste will be effective in a new population ($\theta* > 0$) is 0.976 when assuming normality, but 0.999 when assuming a skewed t distribution. The difference is larger if more stringent clinical criteria for success are used. For example, the probability the treatment effect is at least 0.1 in a new population is 0.878 when assuming normality, and 0.980 when assumed a skewed t distribution. Thus, it is clear that probabilistic inferences such as these are sensitive to the random effects assumption.

6.5 Discussion

6.5.1 Key findings and recommendations

In this chapter we evaluated and discussed the effects of violating symmetry and normality assumptions in several statistical models through the use of simulation studies. In particular, we analysed how robust linear models and meta-analysis models are to violations of the normality assumption by simulating data from a number of asymmetric probability distributions. The key

findings and recommendations are presented in the box below, and discussed in detail thereafter.

Key findings and recommendations

- For linear models
  - Confidence intervals for the treatment effect estimate are not seriously effected unless there are radical departures from symmetry in the residuals and there are fewer than 8 individuals in each arm.
  - Prediction intervals for the response value of a new patient are compromised for asymmetric data even for substantial sample sizes, especially when considering unusually narrow confidence intervals (e.g. 50%). On the contrary, conventional intervals (95% and 99%) appear to be relatively robust to departures from symmetry.
  - However, probability statements regarding a new individual are likely to be compromised.
- For random effects meta-analyses
  - The number of studies is a far bigger contributor to the inaccuracy of confidence intervals than asymmetry in the random effects.
  - Confidence intervals for the average treatment effect constructed assuming normality (using Z_α) are frequently too short for all distributions (irrespective of whether the random effects are symmetric or asymmetric) when the number of studies k is small ($k < 10$).
  - Using the approach suggested by Knapp and Hartung [67] one obtains coverage which is consistently close to 0.95, even when the number of studies is small.
  - Prediction intervals for the treatment effect of a new study are generally overly conservative when $k = 3$ irrespective of the asymmetry in the random effects distribution.

- Narrow prediction intervals (e.g. 50%) are inaccurate when there is substantial asymmetry in the random effects distribution, even when k is large ($k = 100$). However, in terms of coverage, the 95% and 99% prediction intervals are fairly robust to asymmetry in the random effects distribution, especially when there are at least 5 studies and the within study sample size is at least 30.
- Clinically relevant inferences, such as determining the probability of a new treatment effect being above or below a specific value, are likely to be severely impinged by the presence of asymmetry in the random effects. Such inferences are not uncommon in a Bayesian framework and are used to determine the probability of a new treatment being harmful or having an effect which is clinically meaningful.
- Thus, when making probability statements or predictions about treatment effects in a new population, greater attention is needed regarding the asymmetry in the random effects distribution. This requires a reasonable number of studies, and in situations where the normality distribution cannot be verified, prediction intervals and probabilistic statements should be reported more cautiously.

Key findings for linear models

It was shown that, generally speaking, the conventional inferences that one draws from linear models (e.g. treatment effect estimates and confidence intervals) are robust even to extreme departures from symmetry. However, when the sample size is extremely small ($n < 10$) asymmetry in the residuals of a linear model may lead to a skewed sampling distribution for the treatment estimate, which impacts ones ability to draw accurate inferences from the model. Fortunately, most randomised control trials do not contain so few patients and when $n \geq 30$ this does not appear to be a problem.

This was reflected in the coverage of the confidence intervals for the treatment effect, constructed using a linear model. Indeed, it was shown that the confidence intervals for the treatment effect estimate are extremely robust to departures from symmetry. This is particularly

evident for the 95% confidence intervals. Indeed, even for the extremely asymmetric distributions considered here, the coverage of the 95% confidence intervals only display minor departure from 0.95. Indeed, the robustness of the t-test approach to departures from normality is well established [9, 39, 108]. Further, the conclusions drawn from our investigation, that even extreme asymmetry exerts very little influence on the coverage of confidence intervals, is in line with existing research by Lumley et al. [75].

On the surface it appears that prediction intervals for future observations are similarly unaffected for liner models. Indeed, the more conventional prediction intervals (95% or 99%) were shown to be surprisingly robust to departures from symmetry in terms of coverage. However, it was shown that shorter prediction intervals (50%) could be seriously erroneous with the coverage actually worsening for larger sample sizes. Now, while it is unlikely that one would concerned about a 50% prediction interval in practice, this discrepancy in coverage identifies that predictive distribution is far from the assumed Normal distribution. This casts doubt of the apparent validity of the 95% prediction intervals and, moreover, will have obvious ramifications if one is interested in answering questions of a 'Bayesian nature' (i.e. finding the probability of a newly treated individual displaying a clinically important difference in the response).

Key findings for meta-analyses

For a meta-analysis fitted using REML, it is apparent that asymmetry in the random effects distribution does not appear to be a substantial contributor to the accuracy of confidence intervals. However, it was observed that the deviation from the expected coverage was worse if the random effects displayed extreme asymmetry. In particular, confidence intervals had poor coverage when the number of studies was small ($k < 10$) and this seemed to be the case irrespective of the random effects distribution. Generally speaking, it was observed that for small numbers of studies, the confidence intervals tended to be too short using the Normal distribution. This finding was also observed for meta-analyses fit using DerSimonian and Laird's method of moments by Brockwell and Gordon [14]. Brockwell and Gordon [14] suggest calculating confidence intervals for the treatment effect using a 'quantile approximation method', that is, alternative quantiles motivated by a simulation study. However, Jackson and Bowden [59] show that this approach

is sensitive to the study parameters and the distribution of within study variances. Hence, for meta-analyses with fewer than five studies, they advise using Normal quantiles for the primary analysis, but to use the quantile approximation method as part of a sensitivity analysis.

In summary, the number of studies or the within study sample sizes are appeared to be far bigger contributors to the inaccuracy that we observe in the confidence intervals. This conclusion coincides with the findings of Kontopantelis and Reeves [68, 69] who conduct extensive simulation studies comparing the performance of different meta-analysis methods when there is asymmetry in the random effects. This finding is also concurrent with the results of McCulloch et al. [80] who investigate the impact of incorrectly specifying the random effects distribution in generalised linear mixed effects models. They conclude that the maximum likelihood approach displays a "large degree of robustness ... for a wide variety of commonly encountered situations." Moreover, it appeared that increasing the sample size did not significantly improve matters if the number of studies was small.

In these cases the most suitable method appears to be the approach suggested by Knapp and Hartung [67], which makes small sample adjustments to the variance estimates and constructs the confidence interval based on the t distribution. This method produces a slightly wider confidence interval than the conventional method, which reflects the uncertainty in the estimates of the within study variance. However, this procedure generally produced confidence intervals that were too wide in cases where there were a small number of studies (e.g. $k = 3$). This is in line with the conclusions of Cornell et al. [20] who note that the method may overestimate the amount of uncertainty in some cases, particularly when dealing with 5 or fewer studies.

Further, for a meta-analysis fitted using REML, the simulation study appeared to suggest that asymmetry in the random effects doesn't appear to be a serious problem for the construction of 95% or 99% prediction intervals. By contrast, if one is interested in making predictions over shorter intervals (50%), then asymmetry in the random effects distribution can seriously compromise these inferences. These results suggest that the predictive distribution is far from the assumed Normal distribution and, hence, this will have obvious consequences if one is interested in determining the clinical significance of a new treatment effect or answering questions of a

'Bayesian nature' (i.e. finding the probability of a new treatment effect being above or below a certain value). Indeed, it was revealed that prediction interval coverage isn't necessarily always an informative measure, particularly if the interval fails to contain the highest probability mass of the effect distribution. This was reaffirmed by an investigation into the impact of asymmetry on probabilistic inferences, considering an example data set used by Lee and Thompson. Indeed, to guarantee an informative prediction interval it is imperative to account for the asymmetry in the random effects distribution. This conclusion is supported by Karabatsos et al. [65], who propose a Bayesian non-parametric meta-analysis model, which can accommodate a wider range of effects distributions. Indeed they state that the Normal random effects model may be adequate for estimation of the overall effect size, but for prediction, "they surely are not if the effect-size distribution exhibits non-normal behaviour."

6.5.2 Limitations

It is worth noting that in this chapter we have exclusively considered the effect of asymmetry on the treatment effect (in terms of the difference in the mean between control and treatment groups) and the accuracy of confidence intervals and prediction intervals constructed about some mean value. However, when there is a substantial amount of asymmetry in the underlying data, it is debatable whether the mean provides the most informative measure of the centre of the data. For example, in a heavily asymmetric population, a few extreme values will exert an inordinate influence on the location of the mean. This may result in the mean being located a long way from the majority of the data. Hence, while we may be able to construct accurate confidence or prediction intervals about this mean, it may not be the best measure of the centre of the data. For example, when there is extreme asymmetry in the data, it can be argued that the median provides a better measure of the centre of the data (as a few extreme values in one direction do not influence on the estimate of the median).

Also, our principle focus has been to assess the impact of asymmetry on the inferences of statistical models. As a result, we neglected investigation of the impact of other departures from the normality assumption, for example, excessively heavy tailed distributions. In fact, these symmetric departures from the normality assumption have the potential to have equally,

if not more, serious implications for the inferences drawn from statistical models [108].

Further, random effects meta-analysis simulation study has a number of specific limitations. In particular, we only consider the situation where there is a large degree of heterogeneity. That is, where the between study variability is much larger than the within study variability. Further, we have also only considered the situation where the size of the studies is relatively balanced, which means that there is very little variation in the within study variance across studies. Further research is required to determine whether the key findings and recommendations discussed here are sensitive to these features.

Also, in addition to the method suggested by Knapp and Hartung [67], there are numerous other methods for accounting for uncertainty in the variance estimate of $\widehat{\theta}$. For example, Sidik and Jonkman [112, 113] suggest a modification to the Knapp and Hartung method, while Kenward and Roger [66] suggest correcting the conventional REML estimate for small samples. More recently, Noma [84] proposed several different confidence intervals based on Bartlett-type corrections [21]. A thorough investigation of all of these methods is required to ensure the relevance of the key findings and recommendations discussed here.

Another limitation to the simulation study in this chapter is that we have only reported the results based on meta-analyses fitted using REML. We also investigated the impact of asymmetric random effects distributions on the inferences drawn from a meta-analysis model fitted using DerSimonian and Laird's method of moments. The results were found to be relatively consistent with the REML results presented here, however, there are numerous alternatives to fit the meta-analysis. For example, one can use profile likelihood [49] or hierarchical Bayesian models [55].

6.5.3 Conclusion

In summation, it was demonstrated that the common inferences that one draws from linear models or meta-analyses, particularly confidence intervals for the average treatment effect, are robust to departures from symmetry. However, it is worth noting that in both settings, if the aim is to make predictions about future patients or treatments then it is possible that these predictions may be compromised by the presence of asymmetry, either in the residual or random

effects distribution. This conclusion is in line with existing work by Grilli and Rampichini [41] who investigate specification of random effects in in multilevel models. They conclude that "for parameter estimation, the consequences of misspecification are usually minor" whereas "appropriate selection is crucial for valid predictions of random effects".

Further, prediction intervals may have suitable coverage even when the normality assumption is inappropriate, but, they may still be unhelpful clinically. In particular, researchers are better making probabilistic statements (for example, the probability the treatment is effective in a new population) and this requires a critical examination of departures from normality, to assess the impact on such probabilities. In other words, researchers should not routinely use prediction intervals based on normality without checking if the normality assumption is correct, something which will generally require a large number of studies. In situations with small numbers of studies, prediction intervals may be helpful to reveal the uncertainty in predictions, but should be interpreted cautiously.

Additionally, if one is interested in making more unconventional inferences regarding the prediction of future observations (using a linear model or a meta-analysis) then it is recommended to account for the asymmetry. For example, if one suspects that there is substantial asymmetry in the response variable of a linear model, then it may be prudent to make apply a transformation to the data before making inferences about future observations. Similarly, if there is evidence of asymmetry in the random effects of a meta-analysis then it is recommended to apply a more flexible method to account for this asymmetry. For example, Cornell et al. [20] recommend the use of profile likelihood, which allows for the possibility of asymmetric intervals. Alternatively, one can apply a flexible Bayesian model, for example, those proposed by Lee and Thompson [72] or Karabatsos et al. [65], to account for this asymmetry. Cornell et al. also acknowledge the usefulness of flexible Bayesian methods, however, they note that these "require knowledge of Bayesian software and perform best with informed choice of a prior distribution." Indeed, Higgins et al. [55] state that by using "MCMC methods in a Bayesian framework, parametric distributions for the random effects offer the same opportunities for inference as a Normal distribution." They also comment that complex models particularly lend themselves to meta-

analyses with a large number of large studies, where there may be substantial evidence against the assumption of normally distributed random effects.

6.5.4 Next steps for thesis

So far, the investigations in this thesis have focused on developing and appraising methods for measuring asymmetry or testing for symmetry, as well as investigating the impact that asymmetric distributions can have on statistical models that naively assume a symmetric Normal distribution. For example, in this chapter we investigated how sensitive linear models are to asymmetric data, as well as examining the robustness of random effects meta-analyses to asymmetry in the random effects. Prior to that, we developed a new test for symmetry based on a recently proposed measure of asymmetry $\widehat{\eta}$; exhibited the usefulness of $\widehat{\eta}$ in the analysis of randomised control trials; investigated the small sample properties of $\widehat{\eta}$; and proposed a meta-analysis of $\widehat{\eta}$ to assess the asymmetry in several studies investigating the same population. However, if determining the distribution of the underlying population is our principal interest, we needn't restrict ourselves to just considering the location of the mean or the amount of asymmetry (or skewness, kurtosis or any other single summary statistic). In the next chapter we further generalise this approach and discuss how to synthesise and compare information about the entire density function.

Chapter 7

Meta-analysis of a function and its applications to analysing diagnostic test accuracy

7.1 Introduction

The primary goal of this thesis is to develop methods for making inferences about the nature of the distribution of a given random sample from an unknown distribution. Up to this point, we have focused exclusively on testing for symmetry or measuring asymmetry. For example, in Chapter 2 we developed a new test for symmetry based on a recently proposed measure of asymmetry $\widehat{\eta}$. In Chapter 3 $\widehat{\eta}$ was applied to analysing individual randomised control trials. In Chapter 5 we extended this approach to carry out a meta-analysis of $\widehat{\eta}$ to assess the distribution of several studies investigating the same population. Now, in this current chapter we further generalise this approach to consider a meta-analysis of the density function f and distribution function F as a whole. After all, if determining the distribution of the underlying population is our principal interest, we needn't restrict ourselves to just considering the amount of asymmetry (or skewness, kurtosis or any other single summary statistic).

This meta-analysis of a density or distribution function provides the foundation to develop a new type of meta-analysis for use in analysing diagnostic test accuracy, when the test depends

on dichotomising a continuous random variable. A key area of research in medicine is the evaluation of continuous measures as diagnostic tests, where the aim is to correctly identify diseased individuals as test positive (say, high values of the continuous factor) and non-diseased as test negative (low values). To do this, a threshold is required to define high (above threshold) and low (below threshold) values. In particular, when individual patient data is available we can use estimates of the density and distribution function to estimate the distribution of the continuous random variable. As a result, we are able to estimate the test accuracy at any threshold, by determining the probability of obtaining a value above or below this point. In fact, because of the smooth nature of our estimates for f and F we are not even constrained in our choice of threshold by the data. That is to say, there is nothing to prevent estimation of the test's accuracy at a 'hypothetical' threshold in a region where the data are sparse or even missing.

The goal of this chapter is to develop these new methods for analysing diagnostic test accuracy and compare them with existing methods. In section 7.2 we introduce the concept of a meta-analysis of a density estimate. After presenting the asymptotic theory we carry out a worked example using the PTH data introduced in the Chapter 5. In section 7.3 we introduce multivariate meta-analysis and apply this approach to improve upon the meta-analysis of a density estimate. In section 7.4 we briefly introduce the concept of diagnostic tests and some of the existing methods for analysing diagnostic test accuracy (DTA). We then discuss how to extend the meta-analysis of function to develop several novel approaches to conducting meta-analyses of DTA. In section 7.5 we carry out a simulation study to compare the accuracy of our newly proposed methods with some of the existing methods for performing meta-analyses of DTA. In section 7.6 we discuss the results of the chapter.

Aims of the chapter:

- Develop new methods for comparing and synthesising information about the distribution of a specific population, based on data from multiple studies.
- Apply this methodology to propose new techniques for performing meta-analyses of

DTA.

- Compare these new methods with some of the existing procedures for conducting meta-analyses of DTA using both a real life case study and a simulation study.

7.2 Meta-analysis of density estimates

7.2.1 Introduction

When comparing the distribution of data across a number of studies one has a variety of summary statistics at their disposal. Consider, for instance, the parathyroid hormone (PTH) data introduced in section 5.4 of Chapter 5. Table 7.1 shows the sample mean, variance, skewness, and kurtosis, as well as $\widehat{\eta}$ and the sample size n, for the preoperative PTH data in the individuals who did not got on to develop hypocalcemia. Table 7.2 shows precisely the same information for the individuals who did got on to develop hypocalcemia. Each of these summary statistics can be used to assess whether there is homogeneity between the two groups. It is of course commonplace to compare samples based on their mean, and in Chapter 5 we compared the two groups based on their asymmetry coefficient $\widehat{\eta}$.

In fact, the meta-analysis methods discussed in Chapter 5 to synthesise information about the amount of asymmetry η, can also be utilised along with a variety of other summary statistics to check for differences in the distribution of data across a number of studies. For example, one can readily apply a meta-analysis to synthesise and compare information about the mean, variance, skewness or kurtosis.

However, when one is assessing the distribution of a single sample, it is natural to construct a histogram or a density estimate to assess the underlying distribution. When there are multiple (and possibly heterogeneous) samples available, meta-analyses could also be extended to synthesise and compare estimates of the density function.

Indeed, the kernel density estimate $\widehat{f}(x)$ is approximately normally distributed at every point

Study	Mean	Variance	Skewness	Kurtosis	$\widehat{\eta}$	n
Warren2002	50.41	303.92	-0.59	2.59	-0.17	12
Lo2002	60.17	865.73	1.09	5.66	0.34	89
Richards2003	-	-	-	-	-	0
Lam2003	45.56	1099.79	1.83	6.23	0.69	27
Payne0305	-	-	-	-	-	0
Warren2004	65.41	1233.70	1.13	4.25	0.31	23
Lombardi2004	52.87	483.22	0.50	2.62	0.42	35
McLeod2006	75.81	1515.79	1.41	5.60	0.40	46
Pooled	60.48	1041.28	1.40	6.31	0.38	232

Table 7.1: Summary statistics for the preoperative PTH levels in the individuals who did not got on to develop hypocalcemia.

Study	Mean	Variance	Skewness	Kurtosis	$\widehat{\eta}$	n
2002Warren	54.52	830.97	-0.30	1.99	-0.06	4
2002Lo	47.97	410.57	-0.64	2.37	-0.75	11
2003Richards	-	-	-	-	-	0
2003Lam	51.83	770.88	0.11	1.62	0.38	12
0305Payne	-	-	-	-	-	0
2004Warren	99.70	3041.77	0.48	1.50	0.84	3
2004Lombardi	43.44	391.99	1.13	4.09	0.61	16
2006McLeod	71.41	604.52	1.01	3.75	0.41	14
Pooled	56.03	803.48	1.04	5.10	0.20	60

Table 7.2: Summary statistics for the preoperative PTH levels in the individuals who went on to develop hypocalcemia.

x,

$$\widehat{f}(x) \sim N\left(f(x), \sigma^2_{\widehat{f}}\right),$$

where the asymptotic variance, calculated by Bartlett [7], is

$$\sigma^2_{\widehat{f}} = \frac{1}{nh} f(x) \int_R K^2(y)dy.$$

Alternatively, one can perform bootstrapping to estimate the variance. Thus, given multiple random samples from similar (but possibly heterogeneous) populations, it is possible to construct an overall summary estimate of the density estimate. Next, we outline a fixed effect and random effects model for precisely this scenario.

7.2.2 Meta-analysis models for $\widehat{f}(x)$

Suppose we have k independent studies of size n_i from the random variables $X^{(i)}$, for $i = 1, \cdots, k$. That is, let $X_1^{(i)}, \cdots X_{n_i}^{(i)}$ be a random sample from the random variable $X^{(i)}$. Further, suppose that $X^{(i)}$ has density function $f^{(i)}(x)$. The study specific density function $f^{(i)}(x)$ is readily estimated from the data in study i using

$$\widehat{f}^{(i)}(x) = \frac{1}{n_i h_i} \sum_{j=1}^{n_i} K\left(\frac{X_j^{(i)} - x}{h_i}\right), \tag{7.1}$$

where the optimal bandwidth h_i can be estimated using the the data in study i. Then, the kernel density estimate $\widehat{f}^{(i)}(x)$ is approximately normally distributed at every point x,

$$\widehat{f}^{(i)}(x) \sim N\left(f^{(i)}(x), \sigma^2_{\widehat{f}^{(i)}}\right),$$

where the asymptotic variance is given by

$$\sigma^2_{\widehat{f}^{(i)}} = \frac{1}{n_i h_i} f^{(i)}(x) \int_R K^2(y)dy.$$

Again, for brevity we write $\sigma^2_{\widehat{f}^{(i)}}$ in place of $\sigma^2_{\widehat{f}^{(i)}(x)}$, but the asymptotic variance clearly depends on the value of x. The standard error of $\widehat{f}^{(i)}(x)$ is

$$\text{s.e.}\left[\widehat{f}^{(i)}(x)\right] = \sqrt{\frac{1}{n_i h_i}\widehat{f}^{(i)}(x)\int_R K^2(y)dy}. \tag{7.2}$$

The fixed effect model assumes that the $\sigma^2_{\widehat{f}^{(i)}}$ are known and that every estimate $\widehat{f}^{(i)}(x)$ is estimating the same underlying probability density function

$$f^{(i)}(x) = f^o(x), \quad \text{for } i = 1, \cdots, k.$$

Therefore, the fixed effect meta-analysis model is given by

$$\widehat{f}^{(i)}(x) \sim N\left(f^o(x), \sigma^2_{\widehat{f}^{(i)}}\right). \tag{7.3}$$

By contrast, the random effects meta-analysis model assumes that the study specific $f^{(i)}(x)$ are randomly drawn from a Normal population with mean $f^o(x)$ and variance τ^2_f. That is, we assume

$$\begin{aligned} \widehat{f}^{(i)}(x) &\sim N\left(f^{(i)}(x), \sigma^2_{\widehat{f}^{(i)}}\right) \\ f^{(i)}(x) &\sim N\left(f^o(x), \tau^2_f\right), \end{aligned} \tag{7.4}$$

where $f^o(x)$ denotes the overall density function of the underlying population and τ^2_f is the between study variance in the density function estimate at the point x. Note that, for brevity, we write τ^2_f in place of $\tau^2_{f(x)}$, but the between study variance will typically depend on x.

Thus, given multiple random samples from similar (but possibly heterogeneous) populations, it is possible to construct an overall summary estimate of the density estimate $\widehat{f}^o(x)$ as well as an approximate confidence interval for the overall density function. By constructing this at a number of points we can build up a non-parametric representation of the distribution underpinning multiple studies, as well as provide confidence bounds.

To summarise, the overall density $\widehat{f}^o(x)$ can be calculated as follows,

1. Pool data across studies to determine the range of x.

2. Choose the number of points p to estimate in this range $(x_1, x_2, \cdots, x_p)$.

3. For each study, calculate the density estimate $\widehat{f}^{(i)}(x)$ at each x_i using equation (7.1) (where the bandwidth h_i is based on the data in study i).

4. For each density estimate, calculate the standard error at every point x_i using equation (7.2).

5. Use this to perform either a fixed effect (7.3) or random effects (7.4) meta-analysis at each x_i to obtain $\widehat{f}^{o}(x_i)$.

6. Finally, interpolate the summary points $\widehat{f}^{o}(x_i)$ to obtain a summary density estimate $\widehat{f}^{o}(x)$.

A simple R function was written to perform the above steps. In the next subsection we provide some examples using the PTH data.

7.2.3 Examples

Here, we investigate applying the fixed and random effects meta-analysis models to estimate the density function of the preoperative PTH data. Figure 7.1 shows the summary density estimates obtained by fitting the fixed effect meta-analysis model (7.3) for diseased and non-diseased individuals. For both cases we estimate the density at 512 equally spaced points covering the range of the data (-21.08 to 190.28 for the diseased group and -13.25 to 239.25 for the non-diseased group) to ensure a suitably smooth density function in each case. The figure also includes 95% point-wise confidence bands. That is, the curves obtained by interpolating the upper and lower endpoints of the 95% confidence interval at each point. Similarly, Figure 7.2 shows the summary density estimates obtained by fitting the random effects meta-analysis model (7.4) at the same points. Moreover, Figure 7.3 shows the value of I^2 obtained my fitting the random effects meta-analysis (using DerSimonian and Laird's method of moments) at each point. This reveals that the biggest variability between studies occurs in the tails of the density,

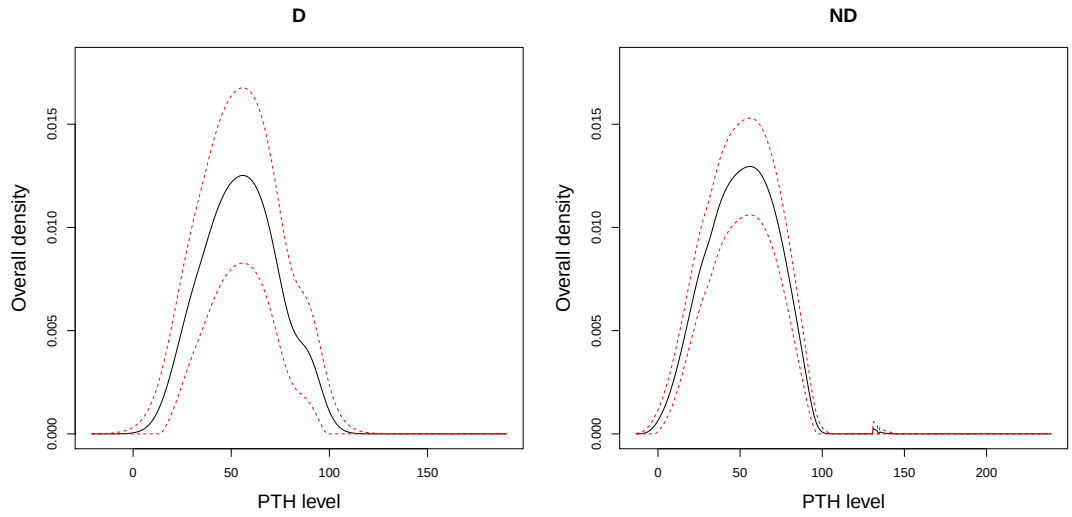

Figure 7.1: Fixed effect meta-analysis of the density estimate $\widehat{f}(x)$ for preoperative PTH data in diseased (D) and non-diseased (ND) individuals, along with 95% confidence bands.

whereas close to the mode there seems to be agreement across studies. Similarly, beyond the tails there is no between study variability as all studies estimate the density to be zero.

Figures 7.4 and 7.5 show the fixed and random effects summary density estimate, alongside the individual study density estimates. There seems to be reasonable agreement between the fixed and random effects methods. Moreover, the overall density functions appear to be sensible summaries of the individual study specific density estimates. However, there are a number of issues with the estimate of $f^o(x)$ described above and in the next subsection we highlight the limitations of this approach.

7.2.4 Issues surrounding $\widehat{f}^o(x)$

Performing a meta-analysis of a density function poses a number of challenges which we shall attempt to address in this subsection. Firstly, because we perform a meta-analysis at each point x_i separately, we cannot, in general, guarantee that the resulting summary function will be a bona-fide density estimate. In fact, the approach outlined above will generally systematically underestimate the density function and so the summary density estimate will not integrate to one. Indeed, if the sample sizes across studies are well balanced then, for every point on the density curve, higher weights will generally be given to the to the smaller density estimate (where

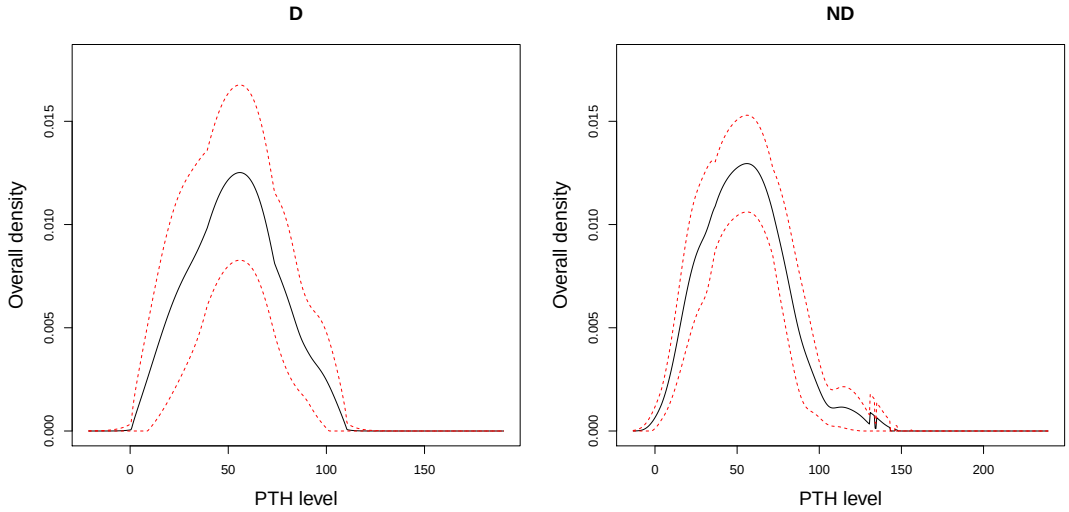

Figure 7.2: Random effects meta-analysis of the density estimate $\widehat{f}(x)$ for preoperative PTH data in diseased (D) and non-diseased (ND) individuals, along with 95% confidence bands.

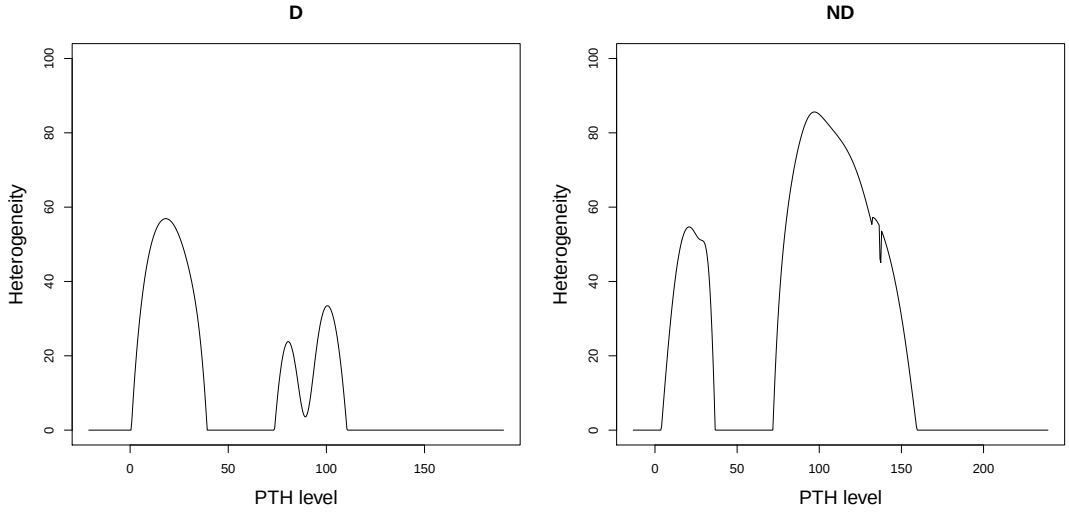

Figure 7.3: Heterogeneity (I^2) present in the random effects meta-analysis fit using DerSimonian and Laird's method of moments.

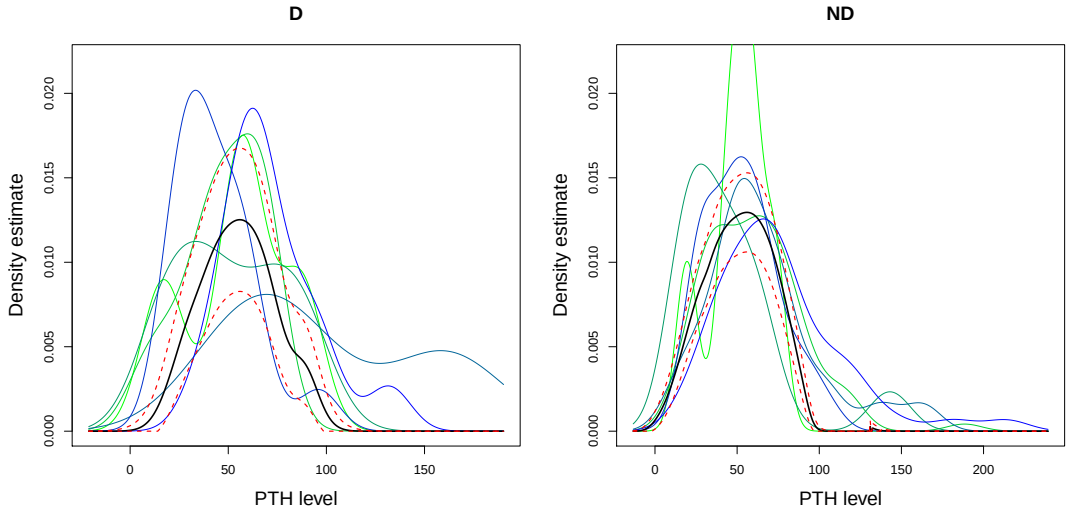

Figure 7.4: Summary density estimates from a fixed effect meta-analysis along with the individual study density estimates for diseased (D) and non-diseased (ND) groups.

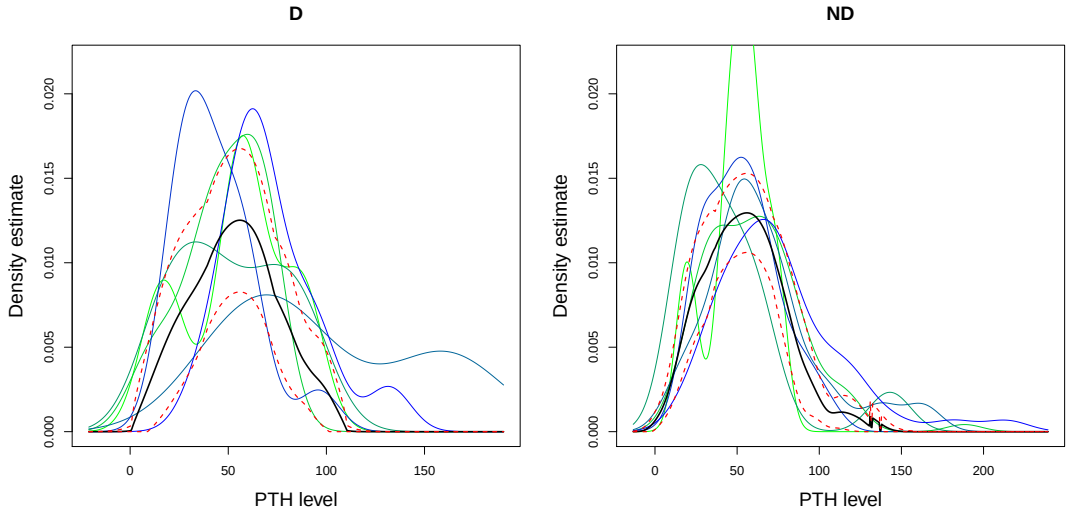

Figure 7.5: Summary density estimate from a random effects meta-analysis along with the individual study density estimates for diseased (D) and non-diseased (ND) groups.

the standard error is lower). This can therefore lead to the summary density estimate being underestimated at all points under consideration. As a result we propose scaling the summary density estimate in order to ensure that one obtains a genuine density function (i.e. it integrates to one). That is, the overall 'average' density function is given by

$$\tilde{f}^o(x) = \frac{\widehat{f}^o(x)}{\int \widehat{f}^o(u)du}.$$

Another issue surrounding $\widehat{f}^o(x)$ is that the estimates of the density at individual points are naturally correlated with one another. Performing a separate meta-analysis at each point ignores this potentially important correlation. In the next section we introduce multivariate meta-analysis to attempt to address this issue. Indeed, considering every point in a single multivariate meta-analysis allows for the possibility of correlation between points.

7.3 Multivariate meta-analysis

7.3.1 Introduction

Multivariate meta-analysis makes it possible to synthesise information regarding multiple outcomes whilst allowing for the possibility of correlation within, as well as between, studies. Jackson et al. [61] provide an extensive overview of the current state of multivariate meta-analysis, identifying a number of advantages and disadvantages. For example, an obvious advantage of carrying out a multivariate meta-analysis is that it increases the available information, and allows for the borrowing of strength from other variables [99]. In the context of medical statistics, this reduces the uncertainty in the treatment effect estimate and increases statistical power. On the other hand, a multivariate meta-analysis can involve estimation of many more parameters, which increases the computational complexity.

Consider a meta-analysis model consisting of k studies with a continuous response vector $\underline{Y}_{ij}$. The within study model has the form

$$\underline{Y}_{ij}|\underline{\mu}_i \sim N(\underline{\mu}_i, S_i),$$

where $\underline{\mu}_i$ is the vector of study specific treatment effects, and the matrix S_i is the covariance matrix of $\underline{Y}_{ij}$. As with the univariate case, the matrices S_i are estimated using the individual patient data within each study [101]. For a random effects model the study specific treatment effects μ_i are distributed according to

$$\underline{\mu}_i \sim N(\underline{\mu}, \Sigma),$$

where $\underline{\mu}$ is the vector of overall treatment effects and Σ is the between study covariance matrix. Generally Σ will be unstructured, but it is possible to simplify the model by assuming a specific structure for Σ (e.g. one can assume that all between study correlations are equal.)

The principal difficulty in performing a random effects multivariate meta-analysis lies in estimating the between study covariance matrix Σ. There are a variety of methods that can be used to calculate the estimate $\widehat{\Sigma}$. For example, Jackson et al. [60] recently extended the DerSimonian and Laird univariate estimate of τ^2 to the multivariate setting. More conventional methods involve estimating Σ using maximum likelihood estimation or restricted maximum likelihood (REML).

Once the between study variance has been estimated, the estimate of the overall treatment effect $\widehat{\mu}$ is given by

$$\underline{\widehat{\mu}} = \left(\sum_{i=1}^{k} \left(S_i + \widehat{\Sigma} \right)^{-1} \right)^{-1} \left(\sum_{i=1}^{k} \left(S_i + \widehat{\Sigma} \right)^{-1} \underline{\widehat{\mu}}_i \right),$$

where $\underline{\widehat{\mu}}_i$ is the estimate of μ_i and $\widehat{\Sigma}$ is the estimate of the between study covariance matrix. The overall effect estimate $\widehat{\mu}$ is approximately normally distributed with variance

$$\text{Var}\left(\underline{\widehat{\mu}}\right) = \left(\sum_{i=1}^{k} \left(S_i + \widehat{\Sigma} \right)^{-1} \underline{\mu}_i \right)^{-1}.$$

Therefore, by applying this result, one can readily obtain approximate confidence intervals for each effect of interest, and even joint confidence regions for two or more of the effects of interest.

7.3.2 Multivariate meta-analysis of f

Until now we have considered a separate meta-analysis for each point of interest on the support of the density function. However, this approach ignores the very real possibility of correlations between points on the density curve (both within and between studies). One way to address this issue is to apply a multivariate meta-analysis and estimate the density curve at all points simultaneously.

It is readily verified that the covariance between $\widehat{f}(x)$ and $\widehat{f}(y)$ is

$$\begin{aligned} S(x,y) := \mathrm{Cov}\left(\widehat{f}(x), \widehat{f}(y)\right) &= \frac{1}{n^2}\mathrm{Cov}\left(\sum_i K_h(X_i - x), \sum_j K_h(X_j - y)\right) \\ &= \frac{1}{n^2}\sum_i \sum_j \mathrm{Cov}\left(K_h(X_i - x), K_h(X_j - y)\right) \\ &= \frac{1}{n}\mathrm{Cov}\left(K_h(X - x), K_h(X - y)\right). \end{aligned}$$

This can be approximated by observing

$$\begin{aligned} S(x,y) &= \frac{1}{n}\left(\mathrm{E}\left[K_h(X-x)K_h(X-y)\right] - \mathrm{E}\left[K_h(X-x)\right]\mathrm{E}\left[K_h(X-y)\right]\right) \\ &= \frac{1}{nh}\mathrm{E}\left[\frac{1}{h}K\left(\frac{X-x}{h}\right)K\left(\frac{X-y}{h}\right)\right] + O\left(\frac{1}{n}\right) \\ &= \frac{1}{nh}\int_R K\left(z + \frac{y-x}{h}\right)K(z) + \left\{f(y) + O(h^2)\right\}dz + O\left(\frac{1}{n}\right) \\ &= \frac{1}{nh}f(y)\int_R K\left(z + \frac{y-x}{h}\right)K(z)\,dz + O\left(\frac{1}{n}\right). \end{aligned}$$

This motivates the following first order approximation,

$$S^{(1)}(x,y) = \frac{1}{nh}f(y)\int_R K\left(z + \frac{y-x}{h}\right)K(z)\,dz.$$

Now, if K has compact support $[-1,1]$, then we have

$$S^{(1)}(x,y) = \begin{cases} \dfrac{1}{nh}f(y)\displaystyle\int_R K\left(z + \frac{y-x}{h}\right)K(z)\,dz & \text{if } |y-x| > h \\ 0 & \text{if } |y-x| < h. \end{cases}$$

This form for the within study covariance matrix is rather insightful, as it clearly shows that there is non-zero correlation for sufficiently close x and y, namely $|x - y| < h$. On the other hand, outside this range the density estimates do not seriously correlate with one another.

However, $S^{(1)}(x, y) \geq 0$ for all x and y. Hence, by only considering the first order approximation we dismiss the possibility of negative correlations. Therefore, alternatively one can use the second order approximation,

$$S^{(2)}(x, y) = \frac{1}{nh} f(y) \int_R K\left(z + \frac{y - x}{h}\right) K(z)\, dz - \frac{1}{n} f(x) f(y) + O\left(\frac{h}{n}\right).$$

Using either of these approaches one can estimate the within study covariance matrix $S(x, y)$. However, by using approximate estimates of the within study variance and covariance we cannot guarantee that the resultant covariance matrix is positive definite. This is a problem because, in order to ensure the numerical stability and convergence of the multivariate model, we require that the estimates of the within study covariance matrices are all positive definite [62, 79]. In cases where the covariances matrices fail to be positive definite, this numerical instability can be remedied in two ways. One approach involves diagonalising the matrix and setting all off-diagonal covariance terms to zero. However, this removes the within study correlations that we are keen to represent in our estimate of $f(x)$. An alternative is to 'augment' the matrix so that the negative eigenvalues are made small and positive. For example, in the context of multivariate meta-analysis, Trikalinos et al. [121] suggest using ridge regression to choose these 'regularising constants'.

In the next section we discuss how our proposed meta-analysis of a density function can be modified to aid in the analysis of diagnostic test accuracy.

7.4 Applying to diagnostic test accuracy

7.4.1 Introduction to diagnostic test accuracy

Put simply, a diagnostic test is any medical procedure which attempts to aid in the diagnosis or detection of disease. For example, in Chapters 4 and 5 as well as this chapter, we conducted analyses on a collection of observational studies investigating the use of parathyroid hormone

(PTH) to predict the onset of hypocalcemia after a thyroidectomy [85]. Recall that postoperative hypocalcemia is a common complication following a thyroidectomy, but, unfortunately hypocalcemia is not usually present until 24 to 48 hours after surgery. As result, it is common practice to keep patients under observation for this time. However, 70% of these patients will not go on to develop hypocalcemia. As noted by Noordzij et al. [85], this puts unnecessary strain on healthcare resources if a simple laboratory test is able to accurately classify patients just hours after surgery.

The main hypothesis of these studies is that postoperative PTH levels are substantially lower in patients who go onto develop hypocalcemia. Hence, an informative test for hypocalcemia would be to classify an individual as positive if their PTH levels were low, and negative otherwise. This motivates the idea of a threshold, below which individuals are classified as positive and above which individuals are classified as negative. Of course, the downside with applying a diagnostic test of this kind, as opposed to relying on the gold standard test (that is, the best possible test, which is assumed to be 100% accurate at classifying diseased and non-diseased individuals - in this case, a 48 hour monitoring period) is that the diagnostic test will not necessarily be 100% accurate. As a result, for most diagnostic tests, there will be some error associated with the positive/negative classification.

Suppose that we have a test based on a continuous marker X, which classifies an individual as positive if $X < \delta$ for some threshold δ. Then, two common measures of test error are given by sensitivity (the probability of a diseased individual testing positive) and specificity (the probability of a non-diseased individual testing negative). Or, more formally in this case

$$\text{Sens}_\delta = P[\text{+ve}|\text{D}] = P[X < \delta|\text{D}] \quad \text{and} \quad \text{Spec}_\delta = P[\text{-ve}|\text{ND}] = P[X > \delta|\text{ND}].$$

In practice these measures can be estimated using a simple sample proportion. For example, for a given threshold, we can construct the following 2×2 table.

	Diseased	Non-diseased
Positive	n_{11}	n_{12}
Negative	n_{21}	n_{22}
	n_1	n_0

The above 2×2 table provides a concise representation of the number of true positives n_{11}, true negatives n_{22}, false positives n_{12} and false negatives n_{21} associated with a given test. Further, n_1 and n_0 denote the total number of diseased and non-diseased individuals as determined by the reference standard (or gold standard test).

In this case the sensitivity of the test can be estimated using

$$\widehat{\text{Sens}}_\delta = \widehat{p}_1 = \frac{n_{11}}{n_1}, \tag{7.5}$$

while specificity is similarly estimated using

$$\widehat{\text{Spec}}_\delta = \widehat{p}_0 = \frac{n_{22}}{n_0}. \tag{7.6}$$

Alternatively, the positive likelihood ratio LR_+ expresses how many times more likely a positive result is in the diseased group compared to the non-diseased group. Assuming that the test is informative, the positive likelihood ratio is greater than one and is defined as

$$LR_+ = \frac{P[\text{+ve}|\text{D}]}{P[\text{+ve}|\text{ND}]} = \frac{\text{Sens}}{1 - \text{Spec}}.$$

Similarly, the negative likelihood ratio expresses how many times less likely a negative result is in diseased individuals compared to non-diseased individuals and is given by

$$LR_- = \frac{P[\text{-ve}|\text{D}]}{P[\text{-ve}|\text{ND}]} = \frac{1 - \text{Sens}}{\text{Spec}}.$$

This information can be synthesised into a single measure of the accuracy of the test. The diagnostic odds ratio (DOR) expresses how many times higher the odds of a positive result in a diseased individual are compared to a non-diseased individual. While the DOR is not as

clinically meaningful as estimates of sensitivity or specificity it does, however, readily lend itself to a univariate meta-analysis setting as it summarises the test accuracy in a single value.

All the measures of accuracy mentioned above are sensitive to the choice of the threshold δ, used to classify an individual as positive or negative. For example, for the PTH data discussed above, as we increase the threshold the sensitivity of our test increases, but the specificity of our test decreases. Therefore, the threshold δ needs to be carefully chosen to adequately balance these two types of error.

One way to visualise the trade-off between sensitivity and specificity is through the use of a receiver operating characteristic (ROC) curve. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1 − specificity) for a variety of thresholds. For example, returning to the PTH example again, if we set the threshold very low then sensitivity and (1 − specificity) will both be close to zero (so we find ourselves in the bottom left corner of the graph). Conversely, for a very large threshold the sensitivity and (1 − specificity) will be close to one. In the intermediate range the ROC curve will be an increasing function between these two extreme regions.

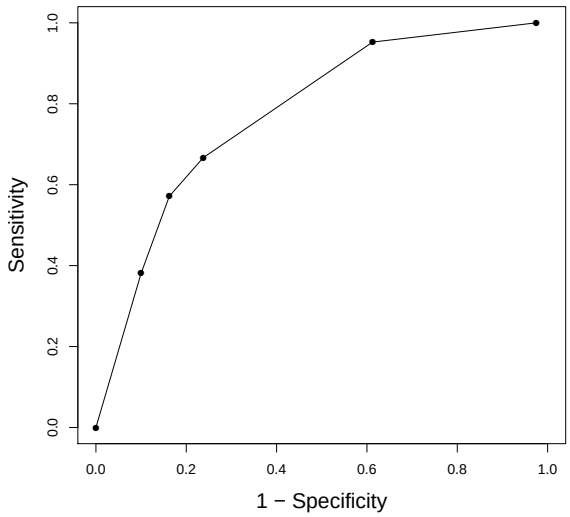

Figure 7.6: An example ROC curve describing the accuracy of a test for postoperative hypocalcemia based on PTH levels 1-2 hours after surgery, using the study carried out by Payne et al. [90]. The curve is constructed using linear interpolation between the true positive rates and false positive rates for the thresholds $\delta = 0, 10, 15, 20, 50, 100$.

For example, Figure 7.6 shows the ROC curve for a subset of the PTH data, using the study carried out by Payne et al. [90]. In particular, it provides a graphical representation of the effectiveness of using PTH levels 1-2 hours after surgery to predict the onset of hypocalcemia. This particular ROC curve is based on a simple linear interpolation between the true positive rates and false positive rates for the thresholds $\delta = 0, 10, 15, 20, 50, 100$.

It is commonplace for multiple studies to investigate the accuracy of a specific diagnostic test. Indeed, as we have already seen, this is exactly the case with the problem of testing for hypocalcemia using PTH. Noordzij et al. [85] collated the results of 9 different studies investigating the accuracy of such tests. In the next subsection we discuss some of the existing procedures for carrying out a meta-analysis of diagnostic test accuracy.

7.4.2 Meta-analysis of diagnostic test accuracy

In the context of diagnostic test accuracy, a meta-analysis aims to calculate and compare estimates of the diagnostic accuracy of a test over a number of studies and investigate the variability of results between these studies. There are a number of challenges to conducting a meta-analysis in this context. For example, variation in thresholds may lead to correlation of sensitivity and specificity estimates between studies. In fact, Putter et al. [94] acknowledge that this correlation may also arise 'implicitly', owing to differences in laboratory equipment or the method of measurement. Moreover, some studies may report multiple thresholds or certain thresholds might be missing in some studies [102].

There are a number of different meta-analysis methods that aim to produce an overall estimate of at least one of the summary estimates discussed in the previous section (e.g. sensitivity, specificity, or DOR). The simplest approach is to carry out a separate meta-analysis for each estimate, however, this ignores the possibility of correlation between the measures of accuracy. For example, as noted by Deeks [24], failing to account for the natural trade-off between sensitivity and specificity can cause underestimation of test accuracy.

Alternatively, if there is information about multiple thresholds or even individual patient data available, there are several meta-analysis methods which aim to generate a summary ROC (sROC) curve. For example, one of the earliest models to take this approach is the Moses-

Littenberg method, given by Moses et al. [83]. More recently, Hamza et al. [48] follow a multivariate random effects approach to generate a summary ROC curve, and Putter et al. [94] extend the approach of Hamza et al. using a meta-analysis model which was previously used for survival curves. Alternatively, Martínez-Camblor [78] proposes a completely different approach, estimating the summary ROC curve from a weighted average of each study specific ROC curve. Each of these are obtained using interpolation between the available thresholds, and this gives rise to a 'fully non-parametric method' for estimating the summary ROC.

We are particularly interested here with the case where individual patient data are available. In this case, for each study, one can construct a 2×2 table at every threshold and calculate sensitivity and specificity estimates. Then, at each threshold it is possible to model sensitivity and specificity directly using a bivariate meta-analysis model.

Suppose therefore, we have k studies containing the individual patient data and with study specific sensitivities $p_1^{(i)}$ and specificities $p_0^{(i)}$. That is, let $n_0^{(i)}$ and $n_1^{(i)}$ be the number of non-diseased and diseased individuals in study i. Further, let $\widehat{p}_1^{(i)}$ and $\widehat{p}_0^{(i)}$ be the estimates of sensitivity and specificity given in equations (7.5) and (7.6), based on the 2×2 table for study i. Reitsma et al. [97] suggest modelling the logit transformed sensitivity and specificity, which are assumed to approximately follow a bivariate Normal distribution. In this case, the within study model is given by

$$\begin{pmatrix} \text{logit}\left(\widehat{p}_1^{(i)}\right) \\ \text{logit}\left(\widehat{p}_0^{(i)}\right) \end{pmatrix} \sim N\left(\begin{pmatrix} \text{logit}\left(p_1^{(i)}\right) \\ \text{logit}\left(p_0^{(i)}\right) \end{pmatrix}, \begin{pmatrix} s_1^2 & 0 \\ 0 & s_0^2 \end{pmatrix}\right), \tag{7.7}$$

where

$$s_1^2 = \frac{1}{n_1^{(i)}\widehat{p}_1^{(i)}\left[1-\widehat{p}_1^{(i)}\right]} \text{ and } s_0^2 = \frac{1}{n_0^{(i)}\widehat{p}_0^{(i)}\left[1-\widehat{p}_0^{(i)}\right]},$$

and

$$\text{logit}(x) = \log\left(\frac{x}{1-x}\right).$$

At the between study level, it is assumed that the logit transformed study specific sensitivity and specificity follow a bivariate Normal distribution. That is,

$$\begin{pmatrix} \text{logit}\left(p_1^{(i)}\right) \\ \text{logit}\left(p_0^{(i)}\right) \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_D \\ \mu_N \end{pmatrix}, \sum\right), \tag{7.8}$$

where μ_D and μ_N are the overall logit-sensitivity and logit-specificity respectively, and

$$\sum = \begin{pmatrix} \sigma_D^2 & \sigma_{DN} \\ \sigma_{DN} & \sigma_N^2 \end{pmatrix}$$

is the between study variance matrix, where σ_D^2 and σ_N^2 define the between study variability in the logit-sensitivity and logit-specificity respectively, and σ_{DN} is the covariance between the two.

A disadvantage of this approach is that it requires an ad-hoc continuity correction in studies that have a zero in the 2×2 table. Moreover, the Normal approximation in equation (7.7) requires a reasonably large number of samples within each study, and is less reliable when the true sensitivity or specificity are close to one. Chu and Cole [19] improve upon this method by proposing a generalised linear mixed model, which removes these two requirements. The main difference is that, within a single study, they model the number of true positives $n_{11}^{(i)}$ and true negatives $n_{22}^{(i)}$ using the exact binomial distribution. Namely,

$$\begin{aligned} n_{11}^{(i)} &\sim \text{Bin}\left(n_1^{(i)}, p_1^{(i)}\right), \\ n_{22}^{(i)} &\sim \text{Bin}\left(n_0^{(i)}, p_0^{(i)}\right), \end{aligned} \tag{7.9}$$

where $n_1^{(i)}$ and $n_0^{(i)}$ are the total number of diseased and non-diseased individuals in study i, and $p_1^{(i)}$ and $p_0^{(i)}$ denote the sensitivity and specificity in study i. Within studies, n_{11} and n_{22} are assumed to be independent as they are based on separate sets of patients. At the between study level, the model is unchanged and uses equation (7.8), which assumes that the logit transformed study specific sensitivity and specificity follow a bivariate Normal distribution. Chu and Cole [19] show that this model can be fit using a mixed effects logistic regression model and that it is considerably more accurate for sparse data.

In the next section we explain how our meta-analysis of $\widehat{f}$ can be extended to analyse

diagnostic test accuracy.

7.4.3 Applying meta-analysis of $\widehat{f}$ to analyse diagnostic test accuracy

In this section we apply the meta-analysis of $\widehat{f}$ to develop a new method for assessing the accuracy of diagnostic tests. Provided that one has access to the individual patient data, it is possible to apply this method to determine the average distribution of the marker in diseased and non-diseased patients over several studies. It is then possible to use these average distributions to calculate the sensitivity and specificity for any threshold of interest. To calculate the sensitivity and specificity we require an estimate of the cumulative distribution function. For this reason, we modify the method proposed in section 7.2 to perform a meta-analysis on the smooth distribution function estimate $\widehat{F}$. Recall from Chapter 1 that

$$\begin{aligned}
\widehat{F}(x) &= \int_{-\infty}^{x} \widehat{f}(u)du \\
&= \int_{-\infty}^{x} \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{u - X_i}{h}\right) du \\
&= \frac{1}{nh} \sum_{i=1}^{n} \int_{-\infty}^{x} K\left(\frac{u - X_i}{h}\right) du \\
&= \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\frac{x - X_i}{h}} K\left(v\right) dv \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathbb{K}\left(\frac{x - X_i}{h}\right),
\end{aligned}$$

where $\mathbb{K}(x) = \int_{-\infty}^{x} K(u)du$ is a cumulative distribution kernel function. For the theoretical properties of this estimator refer to Reiss [96] and the references therein. Crucially, Watson and Leadbetter [126] proved that $\widehat{F}(x)$ is asymptotically normally distributed. Indeed,

$$\sqrt{n} \cdot \left[\widehat{F}(x) - F(x)\right] \xrightarrow{\text{L}} N\left(0, F(x)[1 - F(x)]\right). \tag{7.10}$$

Hence, just as before, it is possible to carry out a meta-analysis of $\widehat{F}(x)$ for each value of x of interest by using

$$\text{s.e.}\left(\widehat{F}(x)\right) = \sqrt{\frac{1}{n}\widehat{F}(x)\left[1-\widehat{F}(x)\right]}. \tag{7.11}$$

Therefore, we now outline how one can calculate the overall summary estimate of $\widehat{F}$ in the non-diseased groups $\widehat{F}_0^o$ and the summary estimate of $\widehat{F}$ in the diseased group $\widehat{F}_1^o$. Using these estimates one can readily obtain an estimate of sensitivity or specificity at any desired threshold. Hall et al. [47] propose precisely this method for generating ROC curves in a single diagnostic test accuracy study. We generalise their approach to perform similar analyses in a meta-analysis setting.

Suppose we have individual patient data from k independent diagnostic test accuracy studies based on a continuous maker. That is, let $X_1^{(i)}, \cdots, X_{n_0^{(i)}}^{(i)}$ be a random sample from the continuous marker in non-diseased group in study i. Similarly, let $Y_1^{(i)}, \cdots, Y_{n_1^{(i)}}^{(i)}$ be an analogous random sample for the diseased group. For study i, let $\widehat{F}_0^{(i)}$ and $\widehat{F}_1^{(i)}$ denote the estimate of $\widehat{F}$ in the non-diseased group and diseased group respectively. That is,

$$\begin{aligned}
\widehat{F}_0^{(i)}(x) &= \frac{1}{n_0^{(i)}} \sum_{j=1}^{n_0^{(i)}} \mathbb{K}\left(\frac{x - X_j^{(i)}}{h_{0;i}}\right), \\
\widehat{F}_1^{(i)}(x) &= \frac{1}{n_1^{(i)}} \sum_{j=1}^{n_1^{(i)}} \mathbb{K}\left(\frac{x - Y_j^{(i)}}{h_{1;i}}\right),
\end{aligned} \tag{7.12}$$

where $h_{0;i}$ and $h_{1;i}$ are the bandwidths for the control and treatment group in study i. Then, supposing we classify an individual as diseased if the marker $X < \delta$, we can estimate the sensitivity and specificity within study i using

$$\widehat{\text{Sens}}^{(i)} = \tilde{p}_1^{(i)} = \widehat{F}_1^{(i)}(\delta),$$

and

$$\widehat{\text{Spec}}^{(i)} = \tilde{p}_0^{(i)} = 1 - \widehat{F}_0^{(i)}(\delta).$$

Using equation (7.11) the approximate standard error for $\widehat{F}_j^{(i)}(x)$ is given by

$$\text{s.e.}\left(\widehat{F}_j^{(i)}(x)\right) = \sqrt{\frac{1}{n_j^{(i)}}\widehat{F}_j^{(i)}(x)\left[1-\widehat{F}_j^{(i)}(x)\right]}, \tag{7.13}$$

for $j = 0, 1$ and $i = 1, 2, \cdots, k$. To avoid the possibility of obtaining a standard error equal to zero, it is possible to carry out the ad-hoc continuity correction, akin to the method by Reitsma et al. [97]. Hence, the within study model for study i at a particular threshold is given by

$$\begin{pmatrix} \tilde{p}_1^{(i)} \\ \tilde{p}_0^{(i)} \end{pmatrix} \sim N\left(\begin{pmatrix} p_1^{(i)} \\ p_0^{(i)} \end{pmatrix}, S_i\right),$$

where $p_1^{(i)}$ and $p_0^{(i)}$ are the sensitivity and specificity for study i and S_i is the within study covariance matrix for study i,

$$S_i = \begin{pmatrix} \frac{1}{n_1^{(i)}}\widehat{F}_1^{(i)}(\delta)\left[1-\widehat{F}_1^{(i)}(\delta)\right] & 0 \\ 0 & \frac{1}{n_0^{(i)}}\widehat{F}_0^{(i)}(\delta)\left[1-\widehat{F}_0^{(i)}(\delta)\right] \end{pmatrix}.$$

The off-diagonal terms of S_i are zero, as it is assumed there is no correlation between the estimates of sensitivity and specificity, due to the fact they are estimated from different patients.

At the between study level, we assume that the study specific sensitivity and specificity are distributed according to

$$\begin{pmatrix} p_1^{(i)} \\ p_0^{(i)} \end{pmatrix} \sim N\left(\begin{pmatrix} F_1^o(\delta) \\ 1-F_0^o(\delta) \end{pmatrix}, \Psi\right),$$

where $F_1^o(\delta)$ and $1 - F_0^o(\delta)$ are the overall sensitivity and specificity respectively, and Ψ is the between study covariance matrix,

$$\Psi = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{12} & \psi_{22} \end{pmatrix}$$

where ψ_{11} and ψ_{22} are the between study variance in sensitivity and specificity, and ψ_{12} is the between study covariance between the two. Therefore, to summarise, the bivariate smooth

distribution function meta-analysis model is given by

$$\begin{aligned}\begin{pmatrix}\tilde{p}_1^{(i)}\\ \tilde{p}_0^{(i)}\end{pmatrix} &\sim N\left(\begin{pmatrix}p_1^{(i)}\\ p_0^{(i)}\end{pmatrix}, S_i\right),\\ \begin{pmatrix}p_1^{(i)}\\ p_0^{(i)}\end{pmatrix} &\sim N\left(\begin{pmatrix}F_1^o(\delta)\\ 1-F_0^o(\delta)\end{pmatrix}, \Psi\right).\end{aligned} \tag{7.14}$$

However, the Normal approximation in equation (7.10) will be particularly stretched when sensitivity or specificity are close to 1, which is typically the region of interest for test accuracy studies. As a result, we also propose a bivariate meta-analysis on the logit transformed smoothed distribution function estimate. This model has the form

$$\begin{aligned}\begin{pmatrix}\text{logit}\left(\tilde{p}_1^{(i)}\right)\\ \text{logit}\left(\tilde{p}_0^{(i)}\right)\end{pmatrix} &\sim N\left(\begin{pmatrix}\text{logit}\left(p_1^{(i)}\right)\\ \text{logit}\left(p_0^{(i)}\right)\end{pmatrix}, S_i^*\right),\\ \begin{pmatrix}\text{logit}\left(p_1^{(i)}\right)\\ \text{logit}\left(p_0^{(i)}\right)\end{pmatrix} &\sim N\left(\begin{pmatrix}\text{logit}\left(F_1^o(\delta)\right)\\ \text{logit}\left(1-F_0^o(\delta)\right)\end{pmatrix}, \Psi^*\right),\end{aligned} \tag{7.15}$$

where

$$S_i^* = \begin{pmatrix}\frac{1}{n_1^{(i)}\widehat{F}_1^{(i)}(\delta)\left[1-\widehat{F}_1^{(i)}(\delta)\right]} & 0\\ 0 & \frac{1}{n_0^{(i)}\widehat{F}_0^{(i)}(\delta)\left[1-\widehat{F}_0^{(i)}(\delta)\right]}\end{pmatrix}.$$

and Ψ^* is the between study covariance matrix,

$$\Psi = \begin{pmatrix}\psi_{11}^* & \psi_{12}^*\\ \psi_{12}^* & \psi_{22}^*\end{pmatrix}$$

where ψ_{11}^* and ψ_{22}^* are the between study variance in logit-sensitivity and logit-specificity, and ψ_{12} is the between study covariance between the two.

Indeed, the logit transformed bivariate model is also especially appealing as there are a number of direct comparisons with the existing method proposed by Reitsma et al. [97], which we investigate in the next subsection.

It is also possible to incorporate the possibility of correlations between the estimates of $\widehat{F}(\delta)$ across different values of δ. That is, analyse all thresholds simultaneously and account for correlations between the multiple thresholds. Indeed, Riley et al. [102] recently proposed an extension of the method by Reitsma et al. [97] for just this purpose. It was shown by Rosenblatt [104] that

$$\mathrm{Cov}\left(F_n(x), F_n(y)\right) = \frac{1}{n}\left[F(\min(x,y)) - F(x)F(y)\right],$$

and similarly it is readily calculated that the covariance between $\widehat{F}(x)$ and $\widehat{F}(y)$ is

$$\begin{aligned}
\mathrm{Cov}\left(\widehat{F}(x), \widehat{F}(y)\right) &= \mathrm{Cov}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{K}_h\left(x - X_i\right), \frac{1}{n}\sum_{j=1}^{n}\mathbb{K}_h\left(y - X_j\right)\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}\left(\mathbb{K}_h\left(x - X_i\right), \mathbb{K}_h\left(y - X_j\right)\right) \\
&= \frac{1}{n}\mathrm{Cov}\left(\mathbb{K}_h\left(x - X\right), \mathbb{K}_h\left(y - X\right)\right), \\
&= \frac{1}{n}\left[\mathrm{E}[\mathbb{K}_h(x - X)\mathbb{K}_h(y - X)] - \mathrm{E}[\mathbb{K}_h(x - X)]\mathrm{E}[\mathbb{K}_h(y - X)]\right] \\
&= \frac{1}{n}\left[F(\min(x,y)) - F(x)F(y)\right] + O\left(\frac{h^2}{n}\right),
\end{aligned}$$

where $\mathbb{K}_h(x) = \mathbb{K}\left(\frac{x}{h}\right)$. However, incorporating the cross-threshold correlations into our model greatly increases the number of variables in the multivariate meta-analysis and, therefore, fitting this more complicated model is much more computationally intensive. Moreover, in this case estimation of the within study variation is also plagued by the same problems as the multivariate meta-analysis of $\widehat{f}$, which we identified in section 7.3. That is, it is difficult to guarantee the positive definiteness of the within study covariance matrix. As a result, for the remainder of the chapter we restrict our analysis to fitting the bivariate models (7.14) and (7.15) at each threshold separately. Indeed, the bivariate model is also appealing as there are a number of direct comparisons with the existing approach proposed by Reitsma et al. [97], which we investigate in the next subsection.

7.4.4 Comparisons with the conventional approach

The newly proposed methods outlined above are directly comparable with the method proposed by Reitsma et al. [97] when analysing the accuracy of a diagnostic test based on a continuous outcome with individual patient data. Indeed, recall that the standard approach would be to generate a 2×2 table at every threshold and calculate sensitivity and specificity from these. In fact, this approach is identical to considering the empirical distribution function estimates for the diseased and non-diseased individuals. For example, recall the following 2×2 table based on dichotomising a continuous marker at a given threshold δ.

	Diseased	Non-diseased
Positive	n_{11}	n_{12}
Negative	n_{21}	n_{22}
	n_1	n_0

It is clear that

$$\widehat{\text{Sens}}_\delta = \widehat{p}_1 = \frac{n_{11}}{n_1} = F_{n;1}(\delta),$$

and

$$\widehat{\text{Spec}}_\delta = \widehat{p}_0 = \frac{n_{22}}{n_0} = 1 - F_{n;0}(\delta),$$

where $F_{n;1}$ and $F_{n;0}$ are the empirical distribution function estimates for the random samples in the diseased and non-diseased individuals respectively. Recall that the empirical distribution function estimate is given by

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \text{I}[X_i < x].$$

Also, while it is true that

$$F_n(x) \sim \text{Bin}(n, F(x)),$$

there is also the analogous result that

$$\sqrt{n} \left[F_n(x) - F(x) \right] \xrightarrow{\text{L}} N \left(0, F(x)[1 - F(x)] \right),$$

and

$$\sqrt{n}\left[\text{logit}\left(F_n(x)\right) - \text{logit}\left(F(x)\right)\right] \xrightarrow{\text{L}} N\left(0, \frac{1}{F(x)[1-F(x)]}\right).$$

In fact, this is precisely the approximation used by Reitsma et al. [97] in developing their bivariate model, using the logit transformed F_n. Hence, our newly proposed logit transformed method is actually a generalisation of the Reitsma method. Indeed, as we let the bandwidth h shrink to zero, $\widehat{F}(x)$ converges to $F_n(x)$. This is an obvious corollary of the result

$$\mathbb{K}\left(\frac{x - X_i}{h}\right) \to \text{I}\left[X_i < x\right],$$

as $h \to 0$. We present a short proof of this result.

Proof. Suppose that $\mathbb{K}$ has compact support $[-1, 1]$. Then

$$\mathbb{K}\left(y\right) = \begin{cases} 0 & \text{if } y < -1 \\ G(y) & \text{if } -1 < y < 1 \\ 1 & \text{if } y > 1 \end{cases},$$

where G represents an arbitrary increasing function on $[-1, 1]$. Hence,

$$\mathbb{K}\left(\frac{x - X_i}{h}\right) = \begin{cases} 0 & \text{if } X_i > x + h \\ G\left(\frac{x - X_i}{h}\right) & \text{if } x - h < X_i < x + h \\ 1 & \text{if } X_i < x - h \end{cases}.$$

Therefore,

$$\begin{aligned} \mathbb{K}\left(\frac{x - X_i}{h}\right) &\to \begin{cases} 0 & \text{if } X_i > x \\ 1 & \text{if } X_i < x \end{cases} \\ &= \text{I}\left[X_i < x\right], \end{aligned}$$

as $h \to 0$. □

This is also apparent in Figure 7.7 which shows $\widehat{F}(x)$ for a range of bandwidths, as well as the empirical distribution function estimate $F_n(x)$, based on a simulated Normal sample of size $n = 30$. In this particular case the optimal bandwidth (based on the rule by Silverman [114]) for the corresponding density function is given by $h = 0.52$. It is clear that, as the bandwidth h is decreased from this optimal value, the smooth distribution function estimate $\widehat{F}(x)$ approaches the empirical distribution function $F_n(x)$. This is an encouraging result, in terms of the implementation of our method, as it means that we cannot 'under-smooth' our estimates of sensitivity and specificity. Indeed, as we let the study specific bandwidths h_i converge to zero, our logit transformed method approaches the method proposed by Reitsma et al. [97]. On the other hand, we must be cautious not to select a bandwidth which is too large to avoid 'over-smoothing'. In Figure 7.7 the curve for $h = 3$ represents a wildly over-smoothed estimate of F. In this case, the bandwidth parameter is so large that the information in the underlying data is lost and the distribution function estimate is dominated by the kernel $\mathbb{K}$.

7.4.5 Issues surrounding $\widehat{F}_o(x)$

There are a number of issues with applying a meta-analysis directly to the smooth distribution function estimate. Because we apply a separate weighting at every point, the resulting function need not necessarily be an increasing function and thus, the summary ROC curve may be ill-defined. One way to ensure that the overall estimate of $\widehat{F}(x)$ is an increasing function of x is to generate the distribution function estimate in two stages. For example, it is possible to first calculate a summary density estimate $\widehat{f}$ using the meta-analysis techniques outlined in the previous section, before numerically integrating this function to arrive at an estimate of F which is surely increasing. Indeed, one can estimate the density function for the non-diseased individuals $\tilde{f}_o^{(0)}(x)$ and diseased individuals $\tilde{f}_o^{(1)}(x)$, before estimating the sensitivity and specificity using

$$\text{Sens} = \int_{-\infty}^{\delta} \tilde{f}_o^{(1)}(u)du, \quad \text{Spec} = \int_{\delta}^{\infty} \tilde{f}_o^{(0)}(u)du.$$

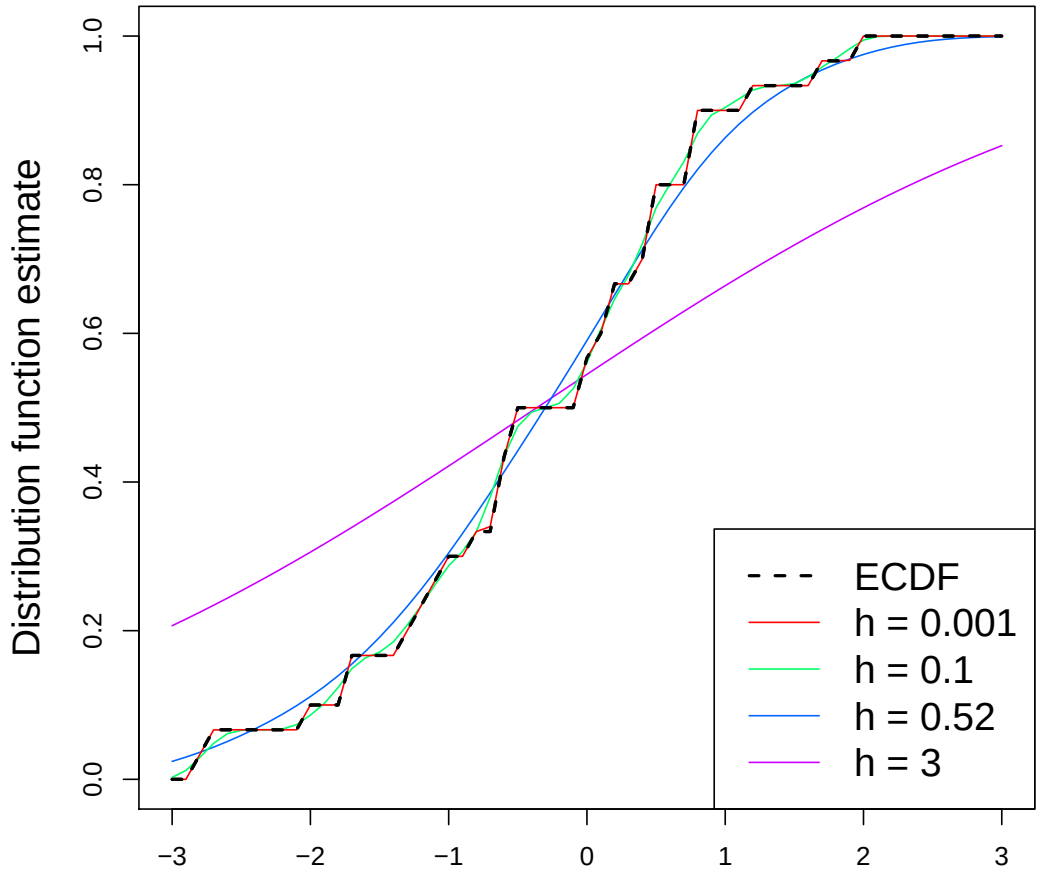

Figure 7.7: Distribution function estimates using the empirical distribution function estimate $F_n(x)$ (dotted line) and the smooth distribution function estimate $\widehat{F}(x)$ with bandwidth h (coloured lines), based on a Normal sample of size $n = 30$. As the bandwidth h is decreased, $\widehat{F}(x)$ approaches $F_n(x)$.

However, due to the complicated nature of this estimate, it is more difficult to determine the resulting standard errors for the sensitivity and specificity. As a result, for the purposes of evaluating diagnostic test accuracy we focus on applying meta-analysis directly on $\widehat{F}$.

7.4.6 Real data example

We now revisit the parathyroid hormone (PTH) meta-analysis data introduced at the start of this section to investigate applying the bivariate smooth distribution function models to a real life data set. Figures 7.8 and 7.9 show the summary smooth distribution function estimate obtained by applying the bivariate meta-analysis model to the PTH levels 1-2 hours after surgery, on the raw and logit scales respectively. The figures also include the individual study estimates and the overall curve appears to be a sensible average of these curves.

Figure 7.10 shows the corresponding sensitivity and specificity estimates obtained by fitting the model on the raw and logit scale, while 7.11 shows the ROC curve obtained by applying both these methods. It is apparent that there is a reasonable disparity between the two methods. This can likely be attributed to the fact that this particular collection of data is made up of several small studies and so the Normal approximation is inappropriate for both scales.

Recall that, when individual patient data are available, the usual approach for evaluating diagnostic test accuracy over multiple studies is to generate a 2×2 table at every threshold. It is then possible to perform multiple bivariate meta-analyses at every threshold using the model proposed by Reitsma et al. [97] or the bivariate binomial method by Chu and Cole [19]. We now compare our newly proposed methods with these existing procedures.

Figure 7.12 compares the overall sensitivity and specificity estimates generated using our newly proposed smooth distribution function methods with the conventional 2×2 table approaches given by Reitsma et al. and Chu and Cole for PTH levels 1-2 hours after surgery. Further, Figure 7.13 compares the corresponding summary ROC curves generated by all of these methods. It is immediately clear that the distinctive feature of the newly proposed methods is that they produce smooth sensitivity, specificity and ROC curves, even in a sparse data set such as this. This is in stark contrast to the existing methodologies that both produce something close to a step function, with jumps at thresholds where at least one individual changes

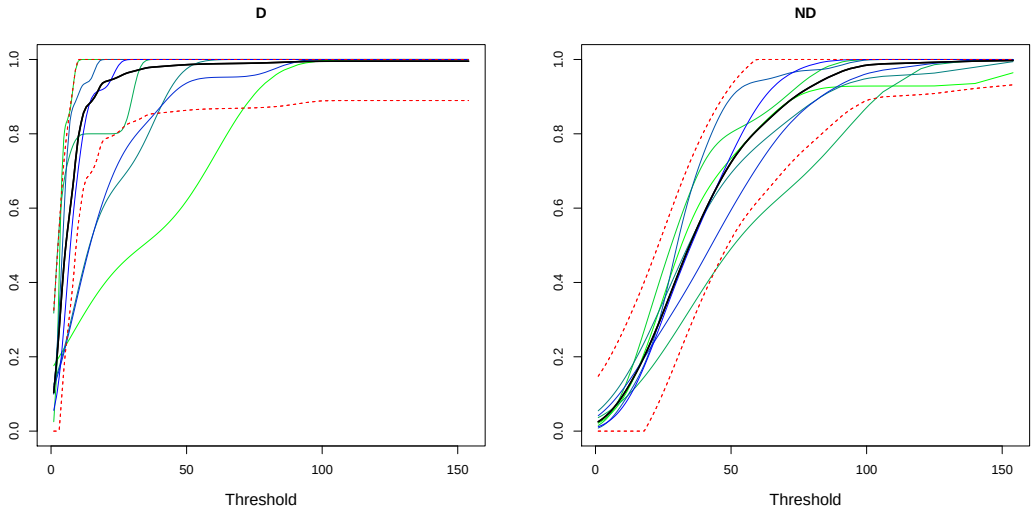

Figure 7.8: Smooth distribution function estimates obtained by fitting the bivariate model on the raw scale along with the individual study estimates for the PTH levels 1-2 hours after surgery for diseased (left) and non-diseased (right) individuals.

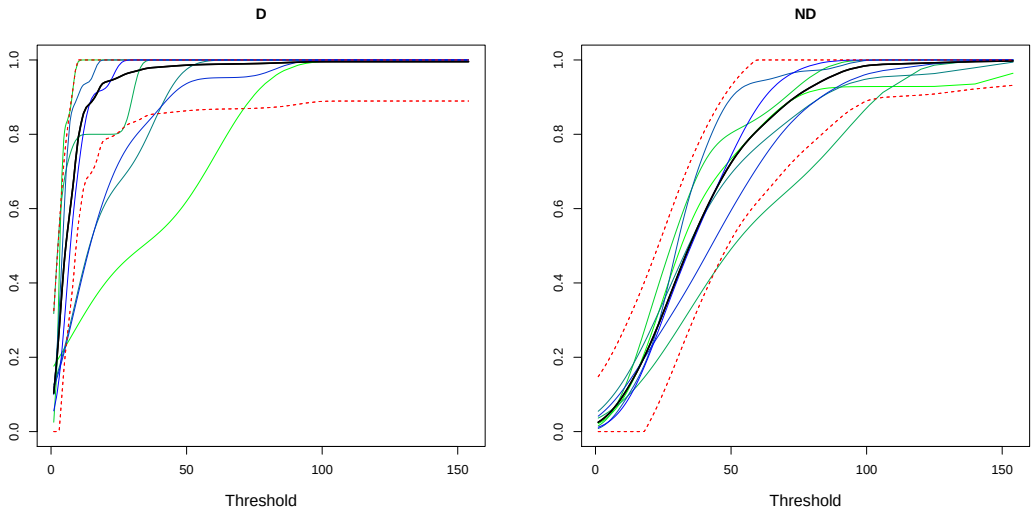

Figure 7.9: Smooth distribution function estimates obtained by fitting the bivariate model on the logit scale along with the individual study estimates for the PTH levels 1-2 hours after surgery for diseased (left) and non-diseased (right) individuals.

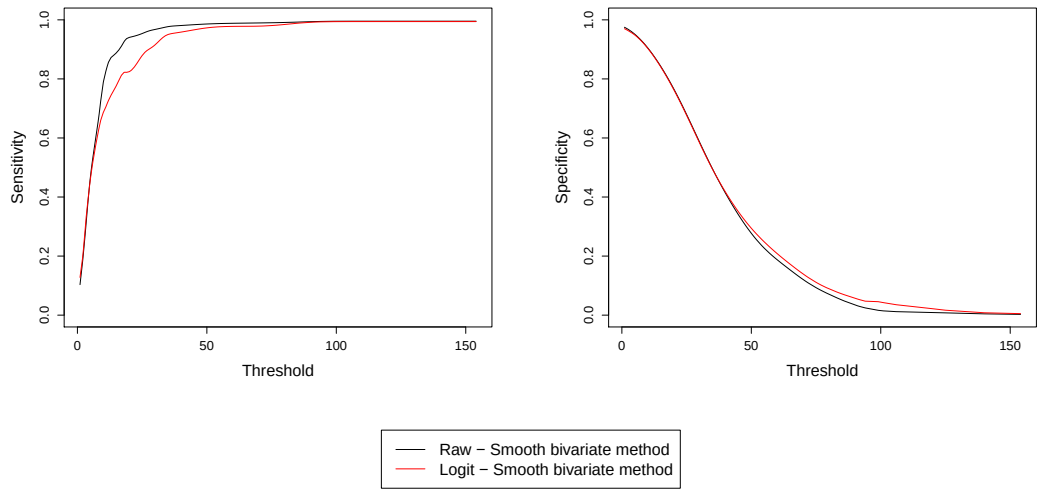

Figure 7.10: Overall sensitivity and specificity estimates obtained by applying the bivariate smooth distribution function method on the raw and logit scale for the PTH levels 1-2 hours after surgery.

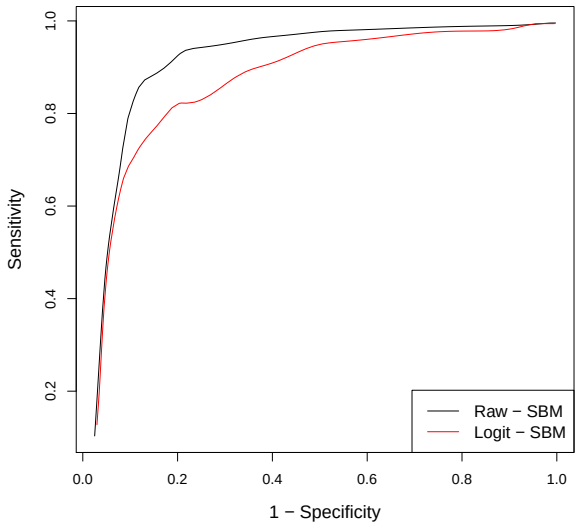

Figure 7.11: ROC curves obtained by applying the smooth bivariate method on the raw (Raw-SBM) and logit (Logit-SBM) scale for the PTH levels 1-2 hours after surgery.

classification (from positive to negative or vice versa).

It is also apparent from these figures that the sensitivity and ROC curve generated using the method by Reitsma et al. is not an increasing function. Furthermore, it is less noticeable for the Chu and Cole method, but this also generates an improper sensitivity and ROC curve. This elucidates the fact that the drawbacks of our newly proposed methods, highlighted in section 7.4.5, are not limited to our method, but are also facets the existing procedures too. Moreover, this example demonstrates that our smooth methods may actually improve upon the existing methods in this regard.

It is also clear from the figures that the actual summary estimates at each threshold vary considerably across the different methods. Therefore, the question remains as to which is method provides the most reliable estimate of test accuracy. In the next section we compare all of these methods in a more thorough fashion by conducting a simulation study designed to investigate the accuracy of these proposed methods at analysing diagnostic test accuracy.

7.5 Simulation study

7.5.1 Introduction

We compare our new methods for carrying out meta-analyses of diagnostic test accuracy with the existing method given by Reitsma et al. [97], and Chu and Cole [19] using a simulation study. The study will generate a continuous response for diseased and non-diseased individuals over several studies and assess the accuracy of the techniques over several thresholds. The simulation study will be used to answer a number of pertinent questions:

1. Principally, how do our newly proposed methods perform compared with the existing methods by Reitsma et al. and Chu and Cole for analysing diagnostic test accuracy? In particular we compare

    - Bias in the sensitivity and specificity meta-analysis results.
    - Empirical standard error of the overall sensitivity and specificity estimates.
    - Mean square error of the overall sensitivity and specificity estimates.

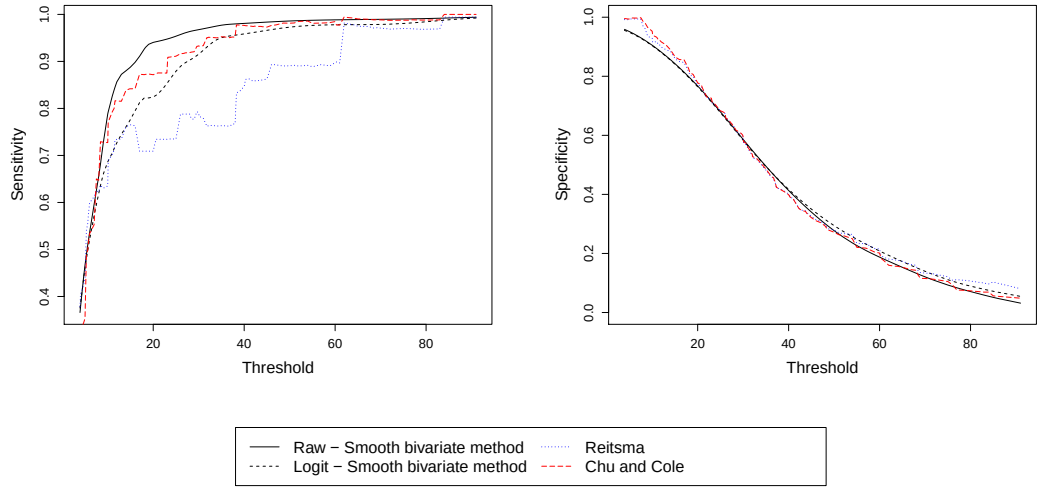

Figure 7.12: Comparison of the overall sensitivity and specificity estimates generated using the smooth bivariate method on the raw and logit scales with the conventional 2×2 table approaches for PTH levels 1-2 hours after surgery.

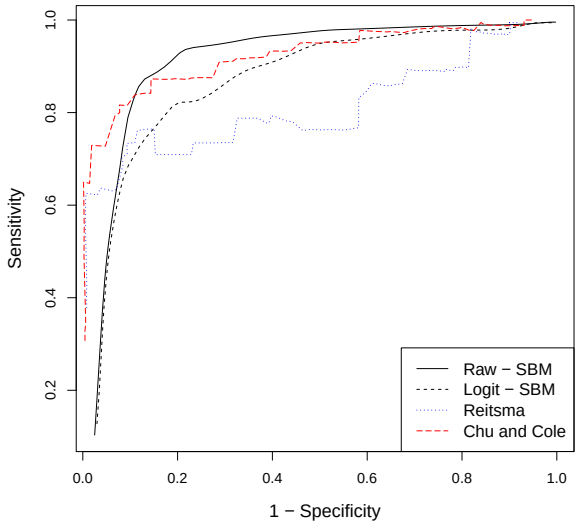

Figure 7.13: Comparison of the summary ROC curves generated using the smooth bivariate method on the raw (Raw-SBM) and logit (Logit-SBM) scales with the conventional 2×2 table approaches for PTH levels 1-2 hours after surgery.

- Coverage of the confidence intervals for overall sensitivity and specificity.

2. In how many of these meta-analyses do we observe impossible sensitivity, specificity, or ROC curves? That is, how frequently do we obtain a non-monotonic overall sensitivity or specificity curve?

3. How much variability is there in the between study variance of sensitivity and specificity across thresholds?

7.5.2 Methods

The simulation study will involve the following fixed parameters governing the distribution of the continuous response:

M The overall mean of the continuous response in healthy individuals.

X The average difference between diseased and non-diseased individuals ($M+X$ is the overall mean of the continuous response in diseased individuals).

κ^2 The between study variance of the mean of the continuous response.

σ_i^2 The within study variance of the continuous response in study i.

In order to choose these fixed distributional parameters for the simulation study, we use the PTH data as a template. In particular, we consider the PTH level 1-2 hours after surgery, which maximises the amount of available data (299 individuals). As the individual patient data are available for this dataset, it is readily calculated that the overall means of the non-diseased and diseased group are 40.84 and 10.90 respectively. Therefore, in our simulation study we set the overall average for the non-diseased group to be 40 and the overall treatment effect to be -25.

The within study standard deviation in the PTH data varies from 15 to 25, however, in order to calculate the true sensitivity and specificity it is convenient to set the within study variance as equal across all studies. By fitting a random effects meta-analysis it is readily calculated that, for this dataset, $I^2 = 14.5\%$ [0%; 58.1%]. We are particularly interested with the case where there is a reasonably large level of heterogeneity between studies. As a result, we choose

$\sigma_i^2 = 20^2 = 400$ for every study i, and select a value of $\kappa^2 = 100$ to reflect this. Therefore, to summarise, we make the following choices for the fixed parameters:

$$M = 40, X = -25, \kappa^2 = 100 \text{ and } \sigma_i^2 = 400 \quad \forall i.$$

The simulation study will also involve varying the following parameters:

k The number of studies.

n_i The total number of diseased and non-diseased individuals in study i.

π The prevalence of the disease.

In particular, we consider $k = 5, 10$ and 20 studies and a range of random sample sizes. For example,

$$n_i \sim U(20, 50) \text{ and } U(100, 300).$$

Both these choices for sample size fall well within the boundaries of real life diagnostic test accuracy studies, based on a review of the recent literature by Bochmann et al. [8]. Indeed, Bochmann et al. found the median sample size to be 122.5 with substantial variation across studies. We will also investigate the effectiveness of both methods for high and low prevalence ($\pi = 0.5$ and 0.25). Again, these choices were informed by another review of the recent diagnostic test accuracy literature by Bachmann et al. [5], who found that the median prevalence was 43% with interquartile range 27% to 61%. Another important parameter is the number of thresholds T to consider. Since our method relies on the availability of individual patient data we can chose any threshold of interest and, therefore, we use a reasonably large number of thresholds. In order to ensure the convergence of all methods, we restrict ourselves to thresholds which give sensitivity and specificity greater than or equal to 0.5. In particular, we consider $T = 26$ equally spaced thresholds over the range $[M + X; M] = [15, 40]$. This facilitates the construction of sensitivity, specificity and ROC curves over this range.

7.5.3 Simulation model

Let n_i^0 and n_i^1 denote the number of non-diseased and diseased individuals respectively in study i. Further, let t_{ij}^0 denote the continuous response of non-diseased individual j in study i where $j = 1, \cdots, n_i^0$ and $i = 1, \cdots, k$. Similarly, let t_{ij}^1 denote the continuous response of diseased individual j in study i where $j = 1, \cdots, n_i^1$ and $i = 1, \cdots, k$. Then,

$$t_{ij}^0 \sim N(\mu_i, \sigma_i^2),$$

and

$$t_{ij}^1 \sim N(\mu_i + X, \sigma_i^2),$$

where

$$\mu_i \sim N(M, \kappa^2).$$

Alternatively, in contracted form

$$t_{ij}^d \sim N(M + dX, \sigma_i^2 + \kappa^2), \qquad d = 0, 1. \tag{7.16}$$

Using this model we generate t_{ij}^d for all patients and all studies, before applying the new and existing methods to calculate estimates of sensitivity and specificity. By assuming that the study specific variances $\sigma_i^2 = \sigma^2$ are equal, we can calculate the true overall sensitivity and specificity at every threshold δ,

$$\text{Sens}_\delta = P\left[X < \delta | \text{D}\right] = F_1(\delta), \ \text{Spec}_\delta = P\left[X > \delta | \text{ND}\right] = 1 - F_0(\delta),$$

where F_1 and F_0 are the Normal cumulative distribution functions with means $M + X$ and M respectively, and variance $\sigma^2 + \kappa^2$. Hence, it is possible to directly compare the test accuracy estimates of sensitivity and specificity with the true overall sensitivity and specificity. The simulation study is outlined in Table 7.3.

Step 1	Randomly generate the $k = 5$ study specific sample sizes $n_i \sim U(100, 300)$.
Step 2	Generate the individual patient data (IPD) for each study using equation (7.16) with $M = 40, X = -25, \sigma_i^2 = 400$ and $\kappa^2 = 100$. In particular, assuming a prevalence of $\pi = 0.5$ simulate πn_i observations for the diseased group and $(1 - \pi)n_i$ observations for the non-diseased group (rounding to the nearest integer where necessary).
Step 3	At each threshold $T = 15, 16, \cdots 40$ generate a 2×2 table using the IPD.
Step 4	At each threshold, fit the smooth bivariate method on the raw (Raw-SBM) and logit (Logit-SBM) scale along with the existing bivariate methods by Reitsma et al. [97] and Chu and Cole [19].
Step 5a	At each threshold, extract the summary estimates of sensitivity and specificity for each method along with the estimates of between study variance for sensitivity and specificity τ^2.
Step 5b	At each threshold, generate maximum likelihood confidence intervals for sensitivity and specificity and check to see whether they contain the true sensitivity or specificity.
Step 5c	For each method, evaluate the monotonicity of the resulting sensitivity and specificity curves. That is, check to see whether the sensitivity (specificity) is non-decreasing (non-increasing) with the threshold.
Step 6	Repeat **Steps 1-5c** $10,000$ times and, for each method, calculate the average bias, empirical standard error, mean square error, and average τ at each threshold. Also report the coverage of sensitivity and specificity for each threshold. Further, calculate the average ROC curve and the total number of invalid sensitivity and specificity curves for each method.
Step 7	Repeat **Steps 1-6** for $k = 10$ and 20 as well as for low prevalence ($\pi = 0.25$) and smaller sample sizes, $n_i \sim U(20, 50)$.

Table 7.3: Step by step guide to the simulation study.

7.5.4 Results

Large within study sample size

First, we compare the performance of the smooth distribution function method (on the raw and logit scales) with the existing methods by Reitsma et al. [97] and Chu and Cole [19] for relatively large samples in each study. Figure 7.14 shows the bias in the estimates of the overall sensitivity and specificity using each of the methods, where the overall within study sample sizes are drawn from $n_i \sim U(100, 300)$, and the prevalence is taken to be $\pi = 0.5$. That is, the total number of observations per study n_i is divided equally the treatment and control arms, appropriately rounding if necessary. It is apparent that the logistic regression method (Chu and Cole) is by far the least biased. In fact, the method is almost completely unbiased. By contrast, all other methods are downwardly biased for the thresholds under consideration here. Of the two newly proposed methods, applying the method on the raw scale is most effective at minimising bias. However, both smooth methods are substantially more biased than the existing methods.

Figure 7.15 shows the empirical standard error in the estimates of the overall sensitivity and specificity using the smooth distribution function method (on the raw and logit scales) along with the existing methods by Reitsma et al. and Chu and Cole. The new methods radically improve on the variability in the estimates, something which is quite typical of smoothing methods. The smoothing methods incorporate more information from the surrounding thresholds, reducing the variability of the estimate, but at the expense of adding bias. Both newly proposed methods perform very closely in terms of empirical standard error.

Figure 7.16 shows the mean square error (MSE) in the estimates of the overall sensitivity and specificity using the smooth distribution function method (on the raw and logit scales) along with the existing methods by Reitsma et al. and Chu and Cole. The mean square error is extremely variable across thresholds, which is reflective of the changeable bias and variance. The new methods out perform the existing methods in the regions of low bias (when sensitivity or specificity are close to 0.5). Alternatively, for thresholds where the sensitivity or specificity is close to 0.9, typically the region of interest, the existing methods outperform the new methods.

Figure 7.17 shows the coverage of the 95% confidence intervals for the overall sensitivity

and specificity using the smooth distribution function method (on the raw and logit scales) along with the existing methods by Reitsma et al. and Chu and Cole. The existing methods are much more consistent in terms of coverage, although, both existing methods produce overly conservative confidence intervals. The new methods appear to produce confidence intervals that are either overly conservative or too short at the extreme ends of the thresholds.

Figure 7.18 shows the square-root of the between study variance in sensitivity and specificity for the smooth distribution function method (on the raw and logit scales) along with the existing methods by Reitsma et al. and Chu and Cole. It is important to note that the between study variance is not reported on the same scale for every method. Indeed, all the methods report the between study variance on the logit scale with the exception of the new method on the raw scale. The key point is that the largest between study variability occurs when the sensitivity or the specificity are close to 0.9.

Table 7.4 shows the number of invalid overall sensitivity and specificity curves obtained using the smooth distribution function method (on the raw and logit scales) along with the existing methods by Reitsma et al. and Chu and Cole. It is clear that when there are only a small number of studies ($k = 5$) the new methods are far less likely to produce an invalid sensitivity or specificity curve. Indeed, of the $10,000$ simulated meta-analyses, the new methods don't generate any invalid sensitivity or specificity curves. However, the existing methods improve in this regard as the number of studies is increased.

Figure 7.19 shows the average ROC curve obtained by averaging the overall sensitivity and specificity across all $10,000$ iterations of the simulated data, along with the 'true' summary ROC curve. This reinforces the earlier discovery that the Chu and Cole method is almost completely unbiased, as there is very little difference between the ROC curve generated using this approach and the 'true' summary ROC curve. Further, in this large sample setting, the summary ROC curve generated by using the approximate method by Reitsma et al. does not deviate far from the 'true' summary ROC. By contrast, the newly proposed methods are noticeably more biased, with the both procedures undervaluing the performance of the test.

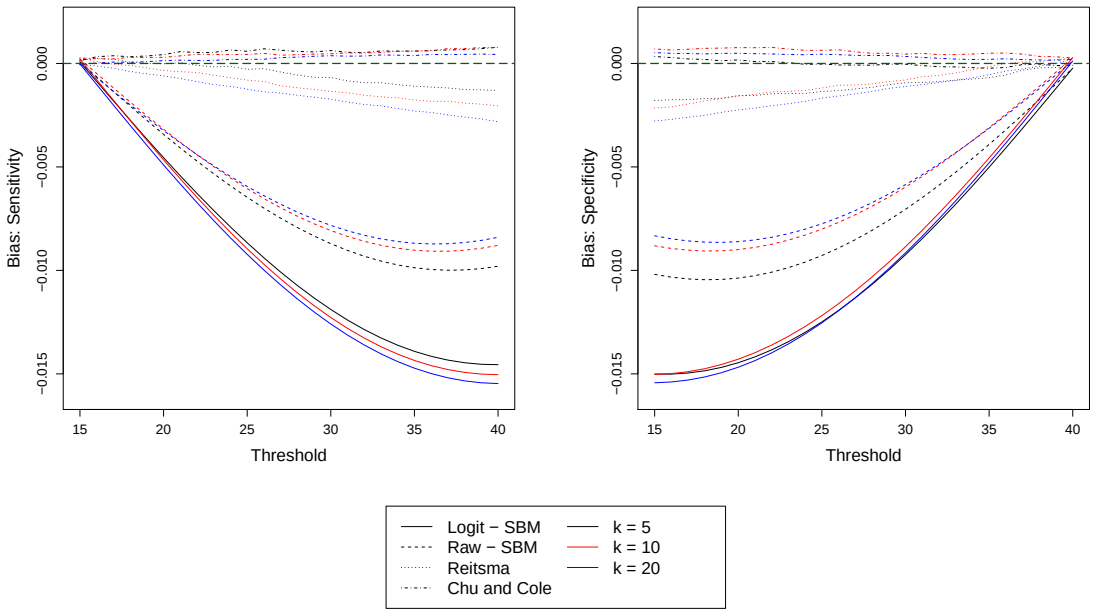

Figure 7.14: Bias in the estimates of the overall sensitivity and specificity using the smooth bivariate method on the raw (Raw-SBM) and logit (Logit-SBM) scales along with the existing methods by Reitsma et al. and Chu and Cole, based on k studies with sample sizes drawn from $U(100, 300)$ and a prevalence $\pi = 0.5$.

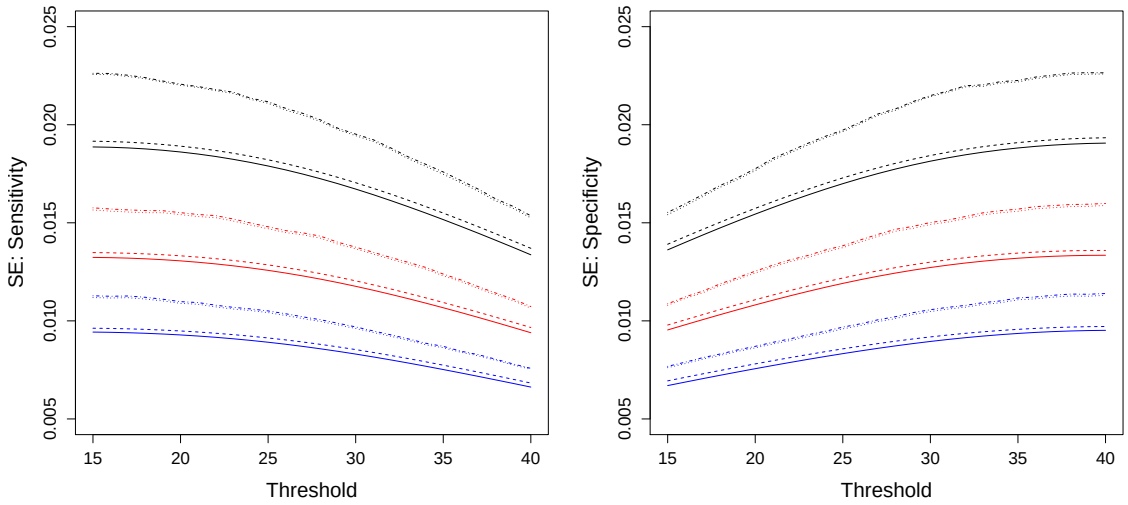

Figure 7.15: Empirical standard error (SE) in the estimates of the overall sensitivity and specificity using the smooth bivariate method on the raw (Raw-SBM) and logit (Logit-SBM) scales along with the existing methods by Reitsma et al. and Chu and Cole, based on k studies with sample sizes drawn from $U(100, 300)$ and a prevalence $\pi = 0.5$.

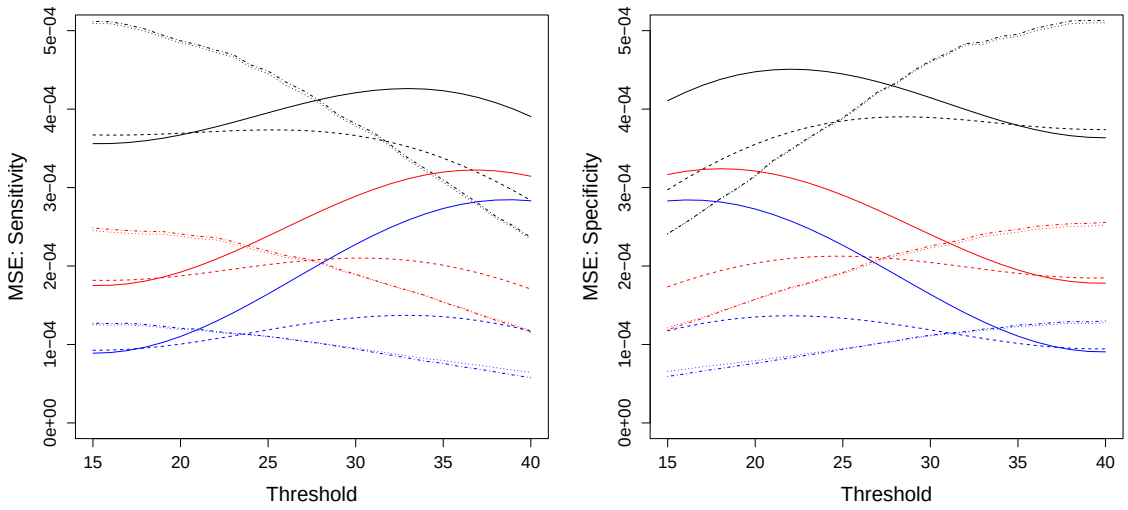

Figure 7.16: Mean square error (MSE) of the estimates of overall sensitivity and specificity using the smooth bivariate method on the raw (Raw-SBM) and logit (Logit-SBM) scales along with the existing methods by Reitsma et al. and Chu and Cole, based on k studies with sample sizes drawn from $U(100, 300)$ and a prevalence $\pi = 0.5$.

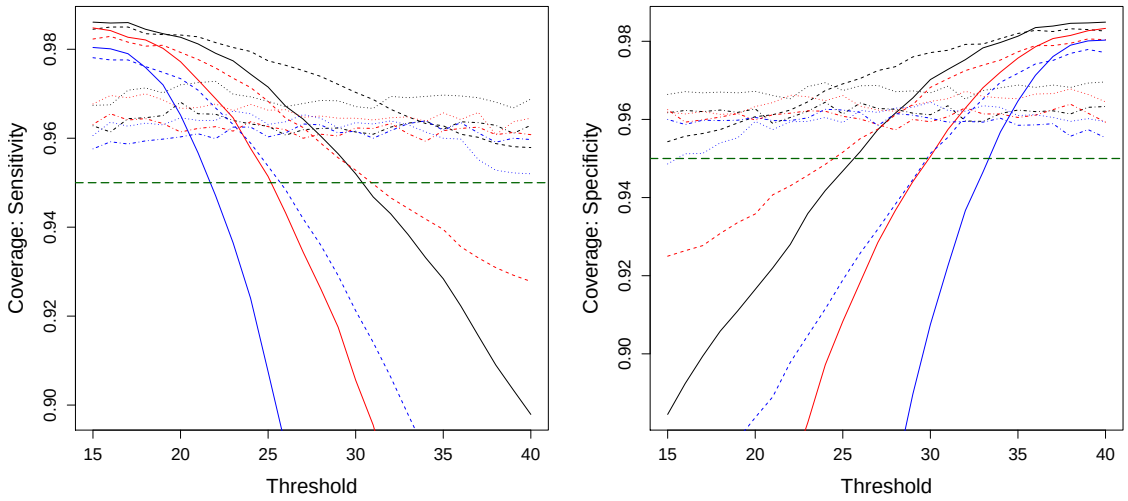

Figure 7.17: Coverage of the 95% confidence intervals for overall sensitivity and specificity using the smooth bivariate method on the raw (Raw-SBM) and logit (Logit-SBM) scales along with the existing methods by Reitsma et al. and Chu and Cole, based on k studies with sample sizes drawn from $U(100, 300)$ and a prevalence $\pi = 0.5$.

Method	Sensitivity			Specificity		
	$k = 5$	$k = 10$	$k = 20$	$k = 5$	$k = 10$	$k = 20$
Logit-SBM	0	0	0	0	0	0
Raw-SBM	0	0	0	0	0	0
Reitsma	270	2	0	276	4	0
Chu & Cole	240	8	0	189	1	0

Table 7.4: The number of invalid summary sensitivity and specificity curves for the smooth bivariate method on the raw (Raw-SBM) and logit (Logit-SBM) scales along with the existing methods by Reitsma et al. and Chu and Cole, based on k studies with sample sizes drawn from $U(100, 300)$ and a prevalence $\pi = 0.5$.

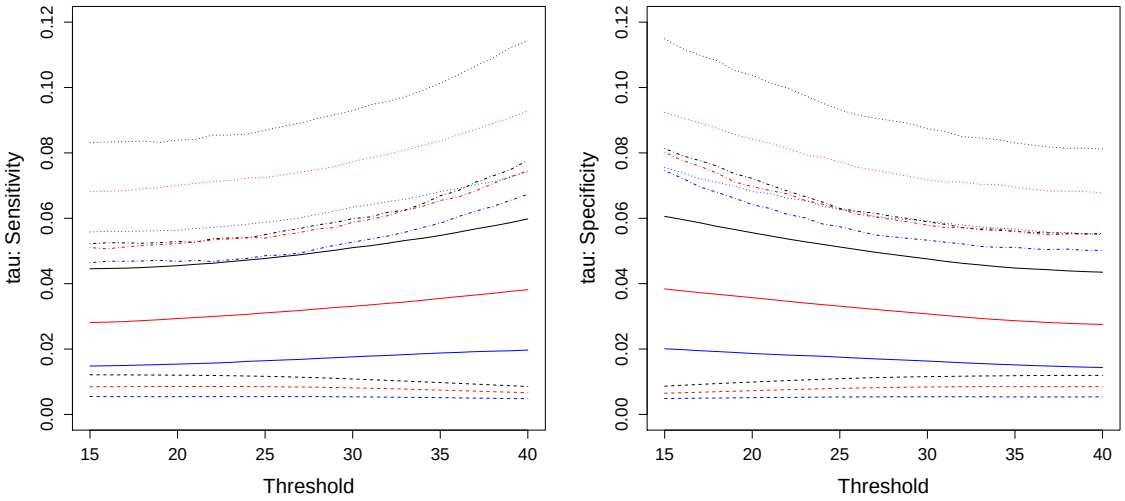

Figure 7.18: Comparing the square-root of the between study variance in sensitivity and specificity for the smooth bivariate method on the raw (Raw-SBM) and logit (Logit-SBM) scales along with the existing methods by Reitsma et al. and Chu and Cole, based on k studies with sample sizes drawn from $U(100, 300)$ and a prevalence $\pi = 0.5$.

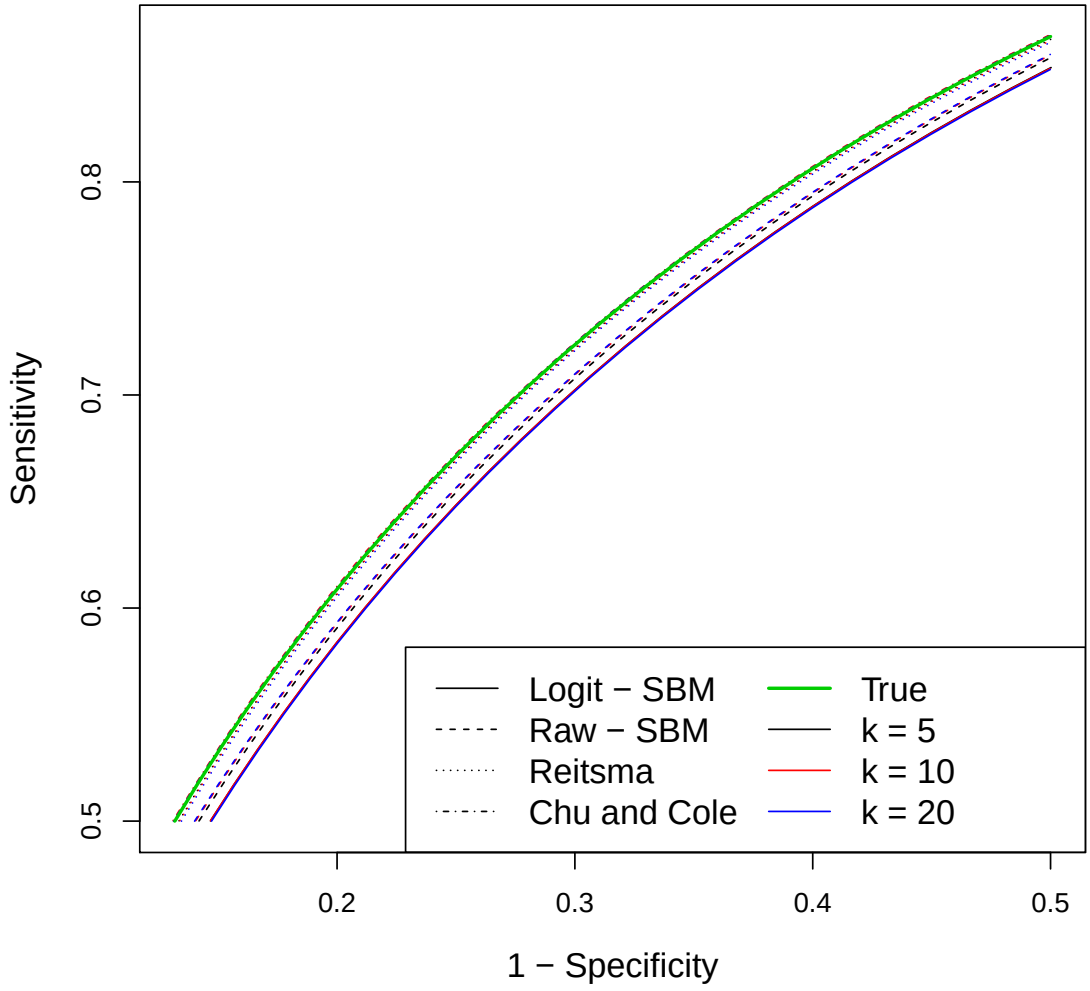

Figure 7.19: Average ROC curve based on the 10,000 simulation study results for the smooth bivariate method on the raw (Raw-SBM) and logit (Logit-SBM) scales along with the existing methods by Reitsma et al. and Chu and Cole, based on k studies with sample sizes drawn from $U(100, 300)$ and a prevalence $\pi = 0.5$.

Large within study sample size with lower prevalence

Next, we compare the performance of the smooth distribution function method (on the raw and logit scales) with the existing methods by Reitsma et al. [97] and Chu and Cole [19], but assuming a lower level of prevalence π. Figures 7.20 to 7.25 and Table 7.5 show the same information where the overall within study sample sizes are drawn from $n_i \sim U(100, 300)$, but for a prevalence of $\pi = 0.25$. That is, the total number of observations per study n_i is divided between the treatment and control arms in the ratio of 1:3, appropriately rounding where necessary.

The comparisons between the methods don't change for lower prevalence, however, the estimates of sensitivity become more imprecise for every method. For example, the bias, empirical standard error, and mean square error in the overall sensitivity estimate all increase. Moreover, for the existing methods, the number of invalid sensitivity curves increases more than 10 fold when $k = 5$ and $k = 10$. For example, for the Chu and Cole method, when $\pi = 0.5$ the number of invalid sensitivity curves increases from 240 and 8 to 2558 and 233 for $k = 5$ and $k = 10$ respectively. By contrast, the newly proposed smooth methods do not show such a staggering increase. In fact, when $\pi = 0.25$ there are only 2 invalid sensitivity curves for the logit transformed method, both of which occur for $k = 5$. In all other cases the new methods produce no invalid sensitivity or specificity curves whatsoever.

Method	Sensitivity			Specificity		
	$k = 5$	$k = 10$	$k = 20$	$k = 5$	$k = 10$	$k = 20$
Logit-SBM	2	0	0	0	0	0
Raw-SBM	0	0	0	0	0	0
Reitsma	3485	273	1	33	0	0
Chu & Cole	2558	233	7	24	0	0

Table 7.5: The number of invalid summary sensitivity and specificity curves for the smooth bivariate method on the raw (Raw-SBM) and logit (Logit-SBM) scales along with the existing methods by Reitsma et al. and Chu and Cole, based on k studies with sample sizes drawn from $U(100, 300)$ and a prevalence $\pi = 0.25$.

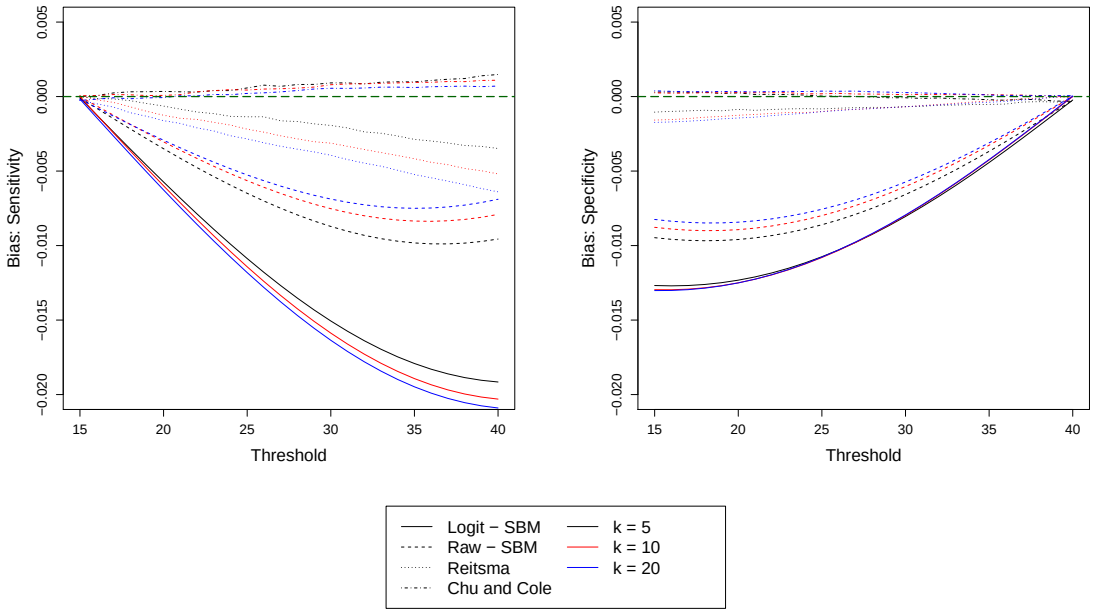

Figure 7.20: Bias in the estimates of the overall sensitivity and specificity using the smooth bivariate method on the raw (Raw-SBM) and logit (Logit-SBM) scales along with the existing methods by Reitsma et al. and Chu and Cole, based on k studies with sample sizes drawn from $U(100, 300)$ and a prevalence $\pi = 0.25$.

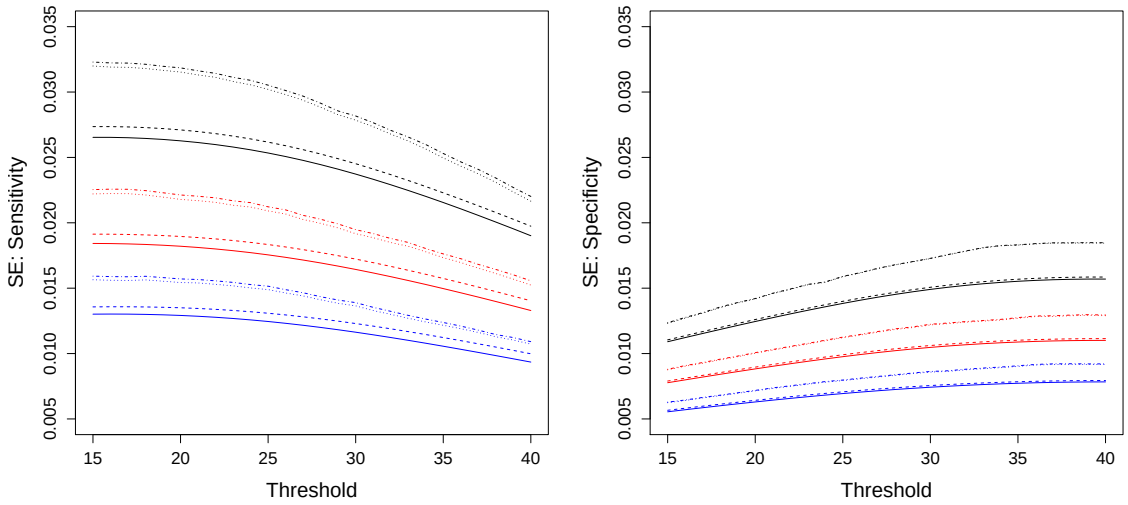

Figure 7.21: Empirical standard error (SE) in the estimates of the overall sensitivity and specificity using the smooth bivariate method on the raw (Raw-SBM) and logit (Logit-SBM) scales along with the existing methods by Reitsma et al. and Chu and Cole, based on k studies with sample sizes drawn from $U(100, 300)$ and a prevalence $\pi = 0.25$.

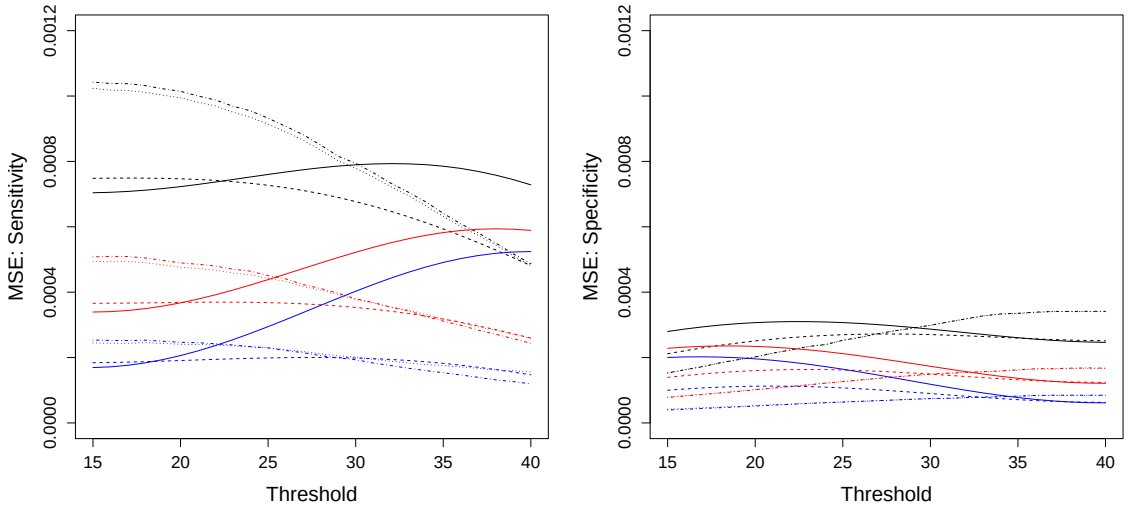

Figure 7.22: Mean square error (MSE) of the estimates of overall sensitivity and specificity using the smooth bivariate method on the raw (Raw-SBM) and logit (Logit-SBM) scales along with the existing methods by Reitsma et al. and Chu and Cole, based on k studies with sample sizes drawn from $U(100, 300)$ and a prevalence $\pi = 0.25$.

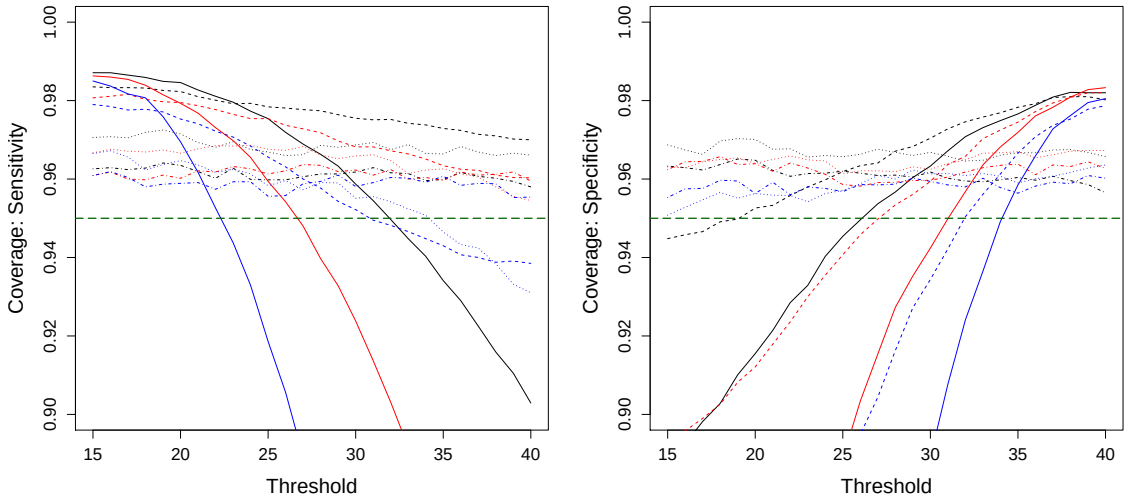

Figure 7.23: Coverage of the 95% confidence intervals for overall sensitivity and specificity using the smooth bivariate method on the raw (Raw-SBM) and logit (Logit-SBM) scales along with the existing methods by Reitsma et al. and Chu and Cole, based on k studies with sample sizes drawn from $U(100, 300)$ and a prevalence $\pi = 0.25$.

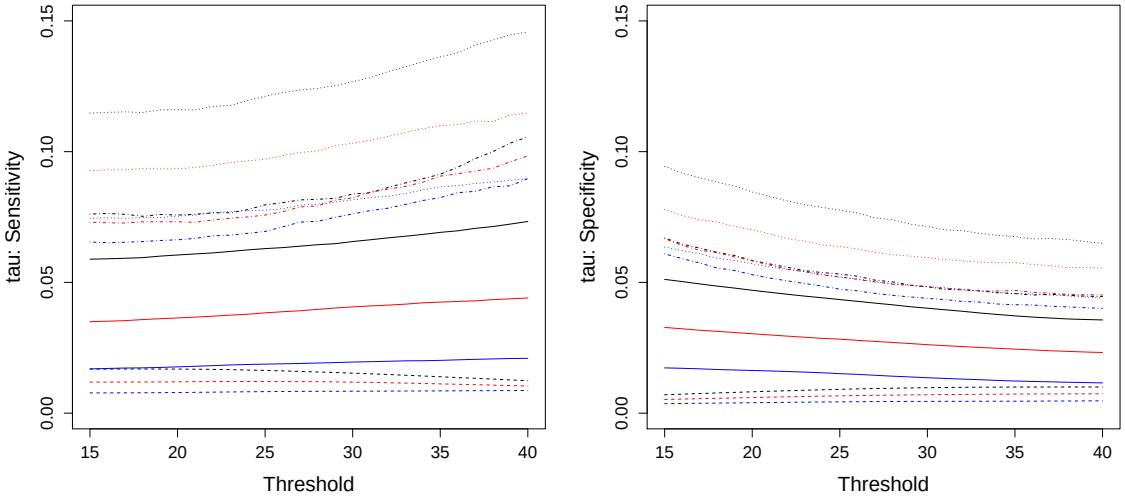

Figure 7.24: Comparing the square-root of the between study variance in sensitivity and specificity for the smooth bivariate method on the raw (Raw-SBM) and logit (Logit-SBM) scales along with the existing methods by Reitsma et al. and Chu and Cole, based on k studies with sample sizes drawn from $U(100, 300)$ and a prevalence $\pi = 0.25$.

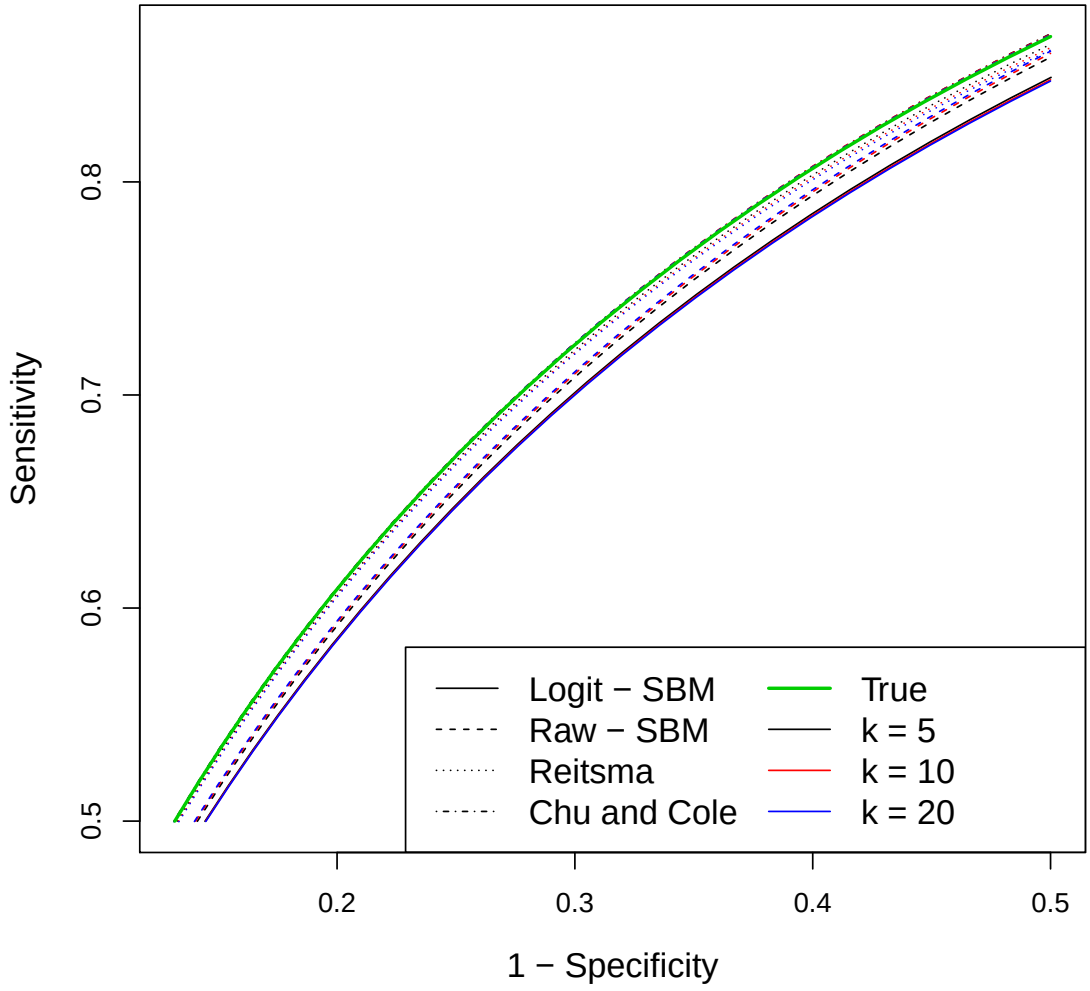

Figure 7.25: Average ROC curve based on the 10,000 simulation study results for the smooth bivariate method on the raw (Raw-SBM) and logit (Logit-SBM) scales along with the existing methods by Reitsma et al. and Chu and Cole, based on k studies with sample sizes drawn from $U(100, 300)$ and a prevalence $\pi = 0.25$.

Small within study sample size

Finally, we compare the performance of the smooth distribution function method (on the raw and logit scales) with the existing methods by Reitsma et al. [97] and Chu and Cole [19] for comparatively smaller studies. Figures 7.26 to 7.31 and Table 7.6 show the same information, but where the overall within study sample sizes are drawn from $n_i \sim U(20, 50)$, and the prevalence is taken to be $\pi = 0.5$. Thus, again the total study specific sample size n_i is divided equally the treatment and control arms, appropriately rounding if necessary.

For example, Figure 7.26 shows the bias in the estimates of the overall sensitivity and specificity using the smooth distribution function method (on the raw and logit scales) in the small sample case, along with the existing methods by Reitsma et al. [97] and Chu and Cole [19]. Firstly, it is apparent that for sparse data such as this, the superiority of the Chu and Cole method over the approximate method by Reitsma et al. is truly emphasised. Indeed, the approximate method becomes increasingly biased as sensitivity or specificity is increased. By contrast, the newly proposed method on the raw scale performs exceptionally well in terms of bias. In fact, the method is comparable to Chu and Cole's exact method. Again, applying the new methodology on the logit scale does not appear to be advantageous and produces by far the most biased estimates of sensitivity and specificity.

Figure 7.27 shows the empirical standard error in the estimates of the overall sensitivity and specificity for all the methods in the small sample case. Once more, the newly proposed smooth methods improve on the variability in the estimates. This time it appears that the logit transformed method performs best, in terms of producing estimates of sensitivity and specificity with less variability

Figure 7.28 shows the MSE in the estimates of the overall sensitivity and specificity obtained using each method for sparse data. Again, the MSE is sensitive to the threshold of interest, which is reflective of the changeable bias and variance. For example, the new method on the logit scale out performs the other methods in the regions of low bias (when sensitivity or specificity are close to 0.5). However, for thresholds where the sensitivity or specificity is close to 0.9, typically the region of interest, the new method on the raw scale consistently outperforms the

other methods. This is because, for small samples, the new method is competitive with existing methods in terms of bias and provides an improvement in terms of variance.

Figure 7.29 shows the coverage of the 95% confidence intervals for the overall sensitivity and specificity, applying each method to small samples. As for the large sample case, the new method on the logit scale appears to produce extremely erratic confidence intervals that are either overly conservative or too short at the extreme ends of the thresholds. However, in contrast to the large sample scenario, the new method on the raw scale seems to compete quite closely with the existing methods. However, this new method and the method by Reitsma et al. appear to be overly conservative for lower values of sensitivity and specificity. On the other hand, the exact method by Chu and Cole appears to be extremely consistent in terms of coverage.

Figure 7.30 shows the square-root of the between study variance in sensitivity and specificity for every method. As before, the between study variance is not reported on the same scale for every method. Again, all the methods report the between study variance on the logit scale with the exception of the new method on the raw scale. With the exception of the Chu and Cole method, there appears to be considerably less variability in between study variance across thresholds when n is small. In particular, the newly proposed method (on either scale) appear to produce an almost flat line, which suggests that the between study variability of the sensitivity and specificity estimate is constant across thresholds. This possibly reflects the impact of the smoothing implemented by the new methods. Indeed, one expects there to be less variability across thresholds, as an estimate of sensitivity or specificity at a given threshold is actually uses information from neighbouring thresholds. Hence, this can result in the variability being 'spread out' across thresholds.

Table 7.6 shows the number of invalid overall sensitivity and specificity curves obtained using every method when the data are sparse. It is clear that, as with the large sample case, the new smooth methods radically out perform the existing methods in this regard. In particular, for sparse data the new method on the raw scale clearly out performs all the others. For example, when there are only $k = 5$ studies of sparse data, the raw method produces only 1 invalid specificity curve. This is less than the new method on the logit scale (209) and a considerable

improvement on the methods by Reitsma et al. (9551) and Chu and Cole (8561).

Figure 7.31 shows the average ROC curve obtained by averaging the overall sensitivity and specificity across all $10,000$ iterations of the simulated data for $n \sim U(20, 50)$, along with the 'true' summary ROC curve. This reinforces the earlier discovery that the new method on the raw scale and the existing method by Chu and Cole are almost completely unbiased. Indeed, there is very little difference between the ROC curve generated using these approaches and the 'true' summary ROC curve. Further, in this small sample setting, it is noticeable that the approximate method by Reitsma et al. underestimates the performance of the test, as it deviates somewhat from the 'true' summary ROC. However, the newly proposed method on the logit scale remains noticeably more biased, seriously undervaluing the performance of the test.

Method	Sensitivity			Specificity		
	$k = 5$	$k = 10$	$k = 20$	$k = 5$	$k = 10$	$k = 20$
Logit-SBM	219	7	9	209	9	0
Raw-SBM	5	1	0	1	2	0
Reitsma	9517	7006	1160	9551	6983	1616
Chu & Cole	8305	5484	1001	8561	5670	1029

Table 7.6: The number of invalid summary sensitivity and specificity curves for the smooth bivariate method on the raw (Raw-SBM) and logit (Logit-SBM) scales along with the existing methods by Reitsma et al. and Chu and Cole, based on k studies with sample sizes drawn from $U(20, 50)$ and a prevalence $\pi = 0.5$.

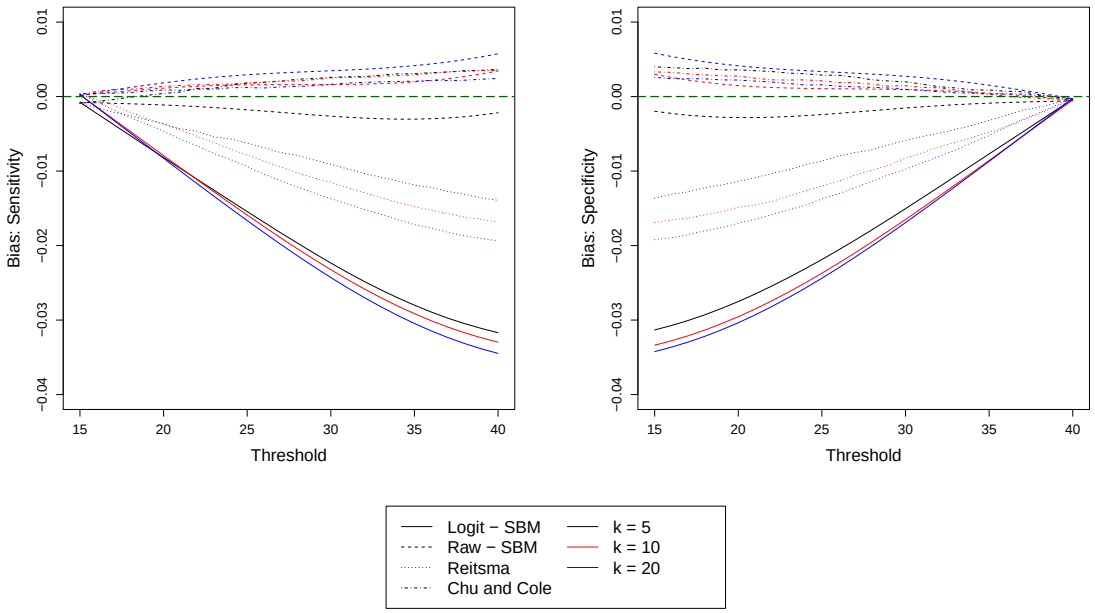

Figure 7.26: Bias in the estimates of the overall sensitivity and specificity using the smooth bivariate method on the raw (Raw-SBM) and logit (Logit-SBM) scales along with the existing methods by Reitsma et al. and Chu and Cole, based on k studies with sample sizes drawn from $U(20, 50)$ and a prevalence $\pi = 0.5$.

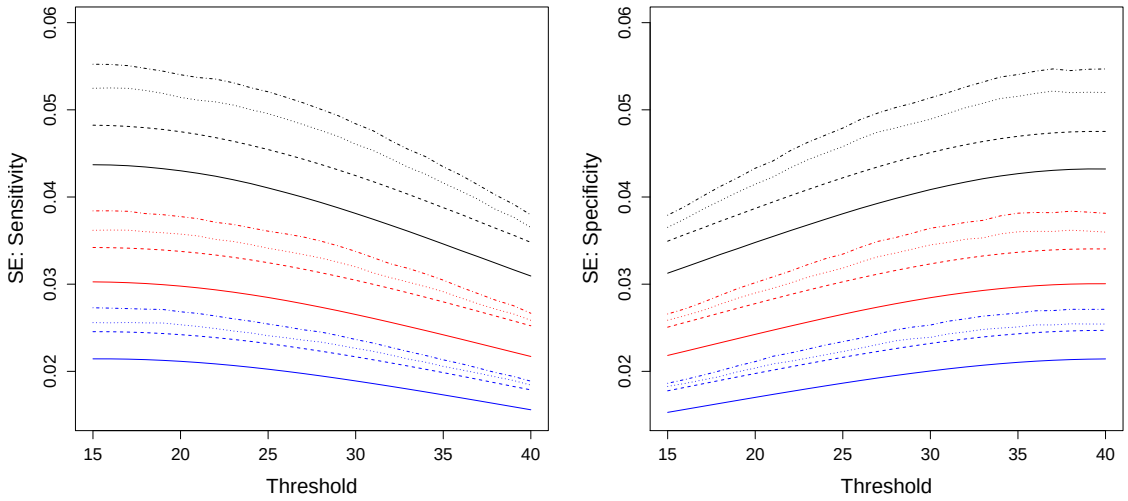

Figure 7.27: Empirical standard error (SE) in the estimates of the overall sensitivity and specificity using the smooth bivariate method on the raw (Raw-SBM) and logit (Logit-SBM) scales along with the existing methods by Reitsma et al. and Chu and Cole, based on k studies with sample sizes drawn from $U(20, 50)$ and a prevalence $\pi = 0.5$.

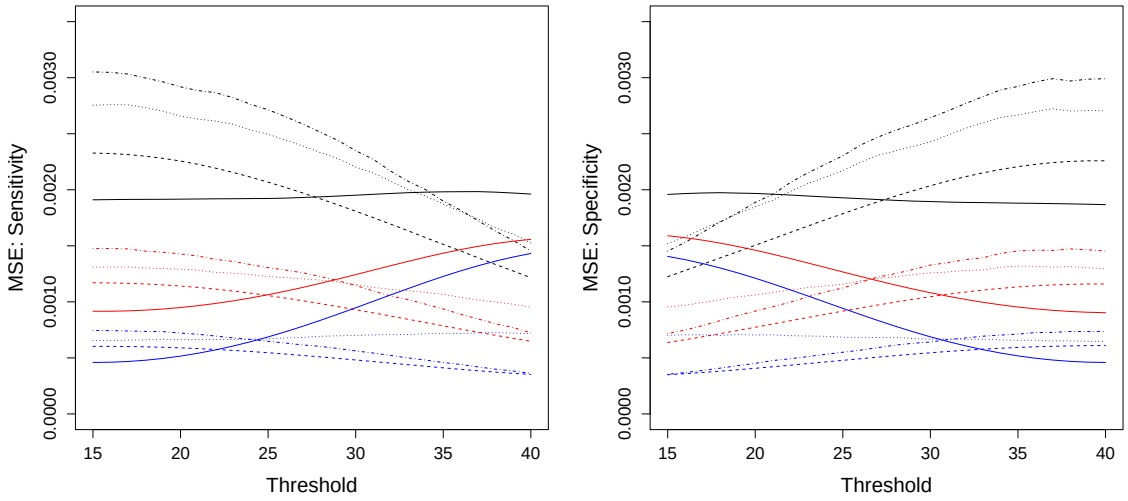

Figure 7.28: Mean square error (MSE) of the estimates of overall sensitivity and specificity using the smooth bivariate method on the raw (Raw-SBM) and logit (Logit-SBM) scales along with the existing methods by Reitsma et al. and Chu and Cole, based on k studies with sample sizes drawn from $U(20, 50)$ and a prevalence $\pi = 0.5$.

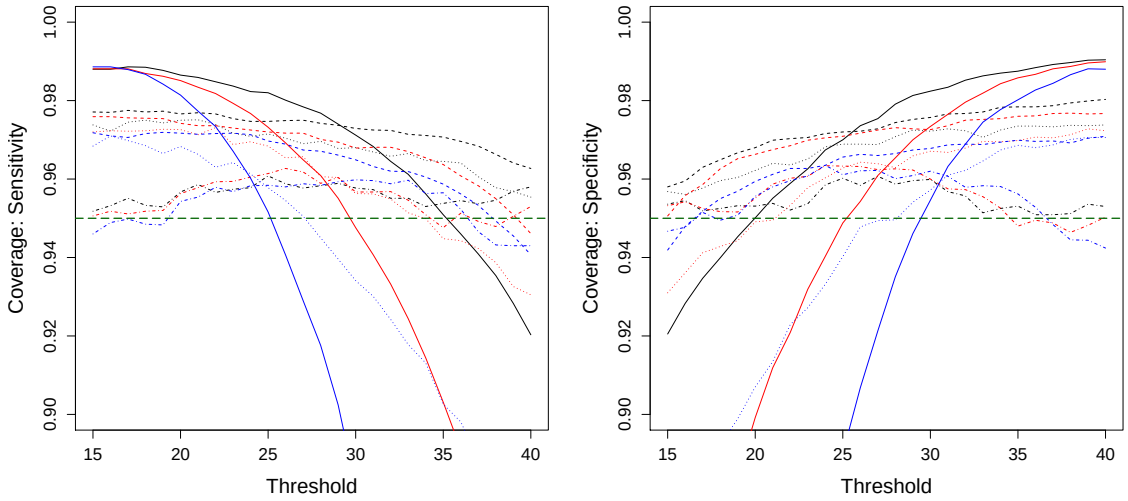

Figure 7.29: Coverage of the 95% confidence intervals for overall sensitivity and specificity using the smooth bivariate method on the raw (Raw-SBM) and logit (Logit-SBM) scales along with the existing methods by Reitsma et al. and Chu and Cole, based on k studies with sample sizes drawn from $U(20, 50)$ and a prevalence $\pi = 0.5$.

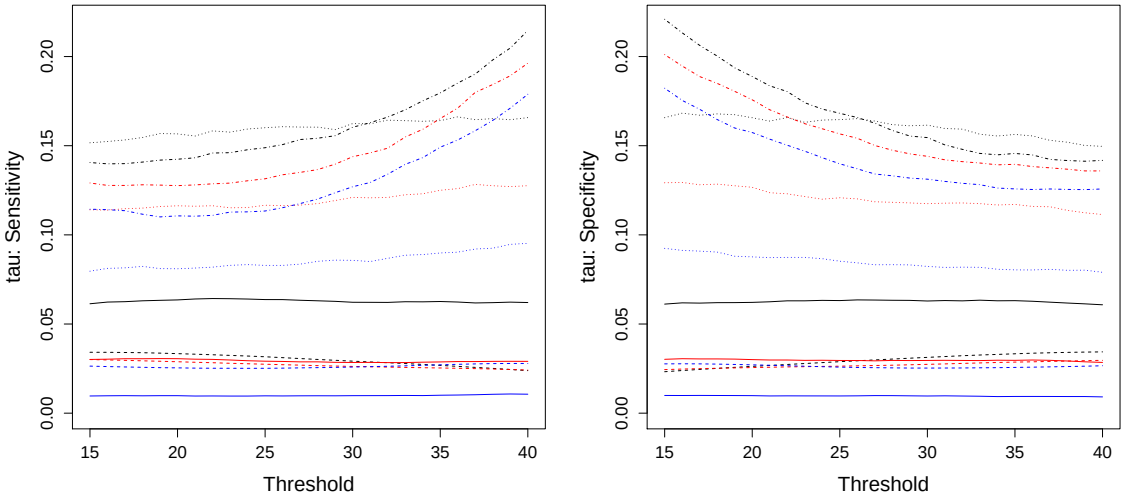

Figure 7.30: Comparing the square-root of the between study variance in sensitivity and specificity for the smooth bivariate method on the raw (Raw-SBM) and logit (Logit-SBM) scales along with the existing methods by Reitsma et al. and Chu and Cole, based on k studies with sample sizes drawn from $U(20, 50)$ and a prevalence $\pi = 0.5$.

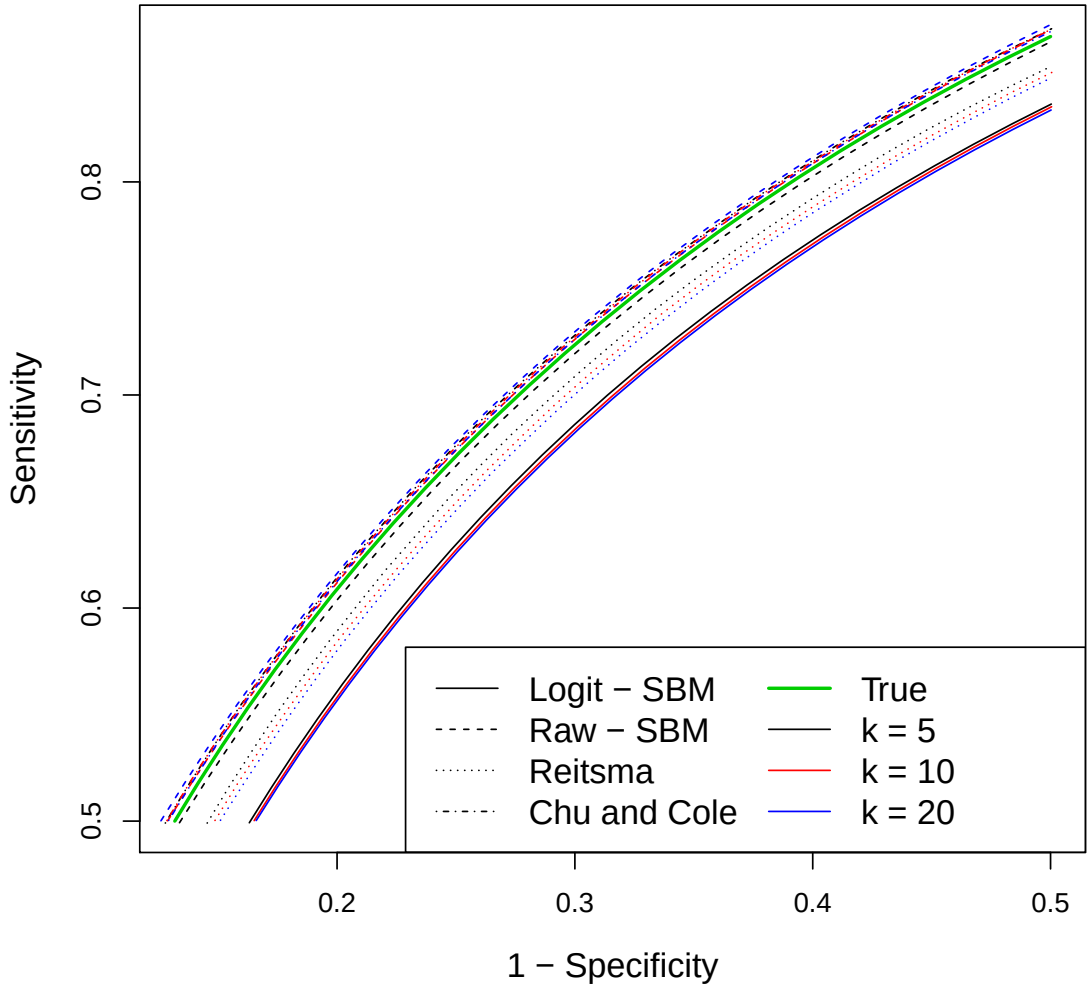

Figure 7.31: Average ROC curve based on the 10,000 simulation study results for the smooth bivariate method on the raw (Raw-SBM) and logit (Logit-SBM) scales along with the existing methods by Reitsma et al. and Chu and Cole, based on k studies with sample sizes drawn from $U(20, 50)$ and a prevalence $\pi = 0.5$.

7.6 Discussion

7.6.1 Key findings

In this chapter we introduced the concept of a meta-analysis of a density estimate, and discussed how it could be applied to the analysis of DTA studies. We started by proposing a simple univariate model to estimate a density function at a single point of interest, when there are multiple studies from the same population. Next, we introduced multivariate meta-analysis and showed how this could be used to improve upon the meta-analysis of a density estimate by considering every point on the curve simultaneously. After introducing the concept of diagnostic tests, we discussed how this methodology could be directly applied to develop several novel approaches to perform meta-analyses of DTA. We also introduced some of the existing methods for analysing DTA and then compared these with our newly proposed models using a real data example as well as a simulation study. The key findings and recommendations are summarised in the box below and discussed in more detail thereafter.

Key findings

- Bias in the estimates of sensitivity and specificity
  - The logistic regression method (Chu and Cole) performs best in terms of bias and is approximately unbiased in all the scenarios considered here.
  - All other methods are downwardly biased for the thresholds under consideration here.
  - Of the new methods, the method on the raw scale outperforms the method on the logit scale in terms of bias.
  - For small sample sizes, the new method on the raw scale competes closely with the Chu and Cole method in terms of bias.
- Empirical standard error of the sensitivity and specificity estimates

  - The new methods radically improve upon the existing methods in terms of the precision of the sensitivity and specificity estimates (quite typical of smoothing methods). Both newly proposed methods perform very closely in terms of empirical standard error.

- Mean square error (MSE) of the sensitivity and specificity estimates
  - MSE varies depending on threshold, which is reflective of the changeable bias and variance.
  - For large samples, the new methods out perform existing methods in the regions of low bias (when sensitivity or specificity are close to 0.5) and vice-versa when sensitivity and specificity are close to 0.9.
  - For small samples, the new method on the raw scale performs best in terms of MSE.

- Coverage of confidence intervals for sensitivity and specificity
  - For large samples, the new methods are either overly conservative or too short at the extreme ends of the thresholds.
  - The existing methods are much more consistent in terms of coverage, however, they are slightly overly conservative.
  - For sparse data, the new method on the raw scale is competitive with the existing methods in terms of coverage.

- Between study variability in sensitivity and specificity across thresholds
  - Different methods report between study variability on different scales (all methods use the logit scale, with the exception of the new method on the raw scale).
  - Biggest variation in sensitivity or specificity occurs when sensitivity or specificity are are close to 0.9.

  - For small samples there is far less variation across thresholds, particularly for the smooth methods.

- Number of invalid overall sensitivity or specificity curves
  - New methods produce far fewer invalid overall sensitivity and specificity (and therefore sROC) curves.
  - For small samples, the new method on the raw scale performs best in this regard.

Recommendations Given individual patient data from multiple studies:

- Use the approach of Chu and Cole at each threshold simultaneously, as this produces unbiased estimates and suitable coverage.
- Occasionally, the sROC this produces may not be well-defined (strictly increasing) as sensitivity and specificity are being estimated completely independently at neighbouring thresholds and thus are not constrained.
- If a smooth ROC curve is of interest, one might therefore utilise the approach of Hamza et al. [48] that extends the Chu and Cole method to fit smooth functions across studies. However, this may not always be a viable option and, especially in sparse data, the newly proposed smoothing method (on the raw scale) therefore provides a useful alternative.
- When applying the new methodology it is important to validate this approach using the exact method by Chu and Cole. If the methods are in relatively close agreement then the new method can be used estimate the sensitivity and specificity at any threshold. On the other hand, if the results of the two methods are excessively disparate then one is forced to conclude that the smooth method does not provide sufficiently accurate estimates of sensitivity and specificity.

The newly proposed methods have a number of advantages over the existing methods. The

most notable improvement is that the new methods generate a smooth ROC curve, which allows the sensitivity or specificity to be calculated at any threshold (even if there is little/no data available data at that point). This is in stark contrast to the existing methods proposed by Reitsma et al. [97] and [19], which generate unsightly step functions. For these methods, there is only a change in sensitivity and specificity at a threshold where at least one individual changes from being positive to negative (or vice versa). This is problematic if the data are particularly sparse for certain thresholds where one might be interested in estimating sensitivity and specificity. For example, suppose that for a given test one is interested in obtaining an estimate of the sensitivity when the threshold is set at $\theta = 30$. Further, suppose that due to the sparseness of the available data, the closest thresholds for which there is a change in a diseased individual's classification are $\theta = 20$ and 35. In this scenario, rigorously using the sensitivity curve generated by applying the methods given by Reitsma et al. or Chu and Cole will give the same sensitivity as at $\theta = 20$ or 35. Of course, this is a naive approach to follow, as we know that the sensitivity should be bound between the sensitivity at $\theta = 20$ and $\theta = 35$. Unfortunately, one is then faced with the dilemma of appropriately selecting an estimate on this range. For example, should one take the average sensitivity, or should one account for the relative closeness of the neighbouring thresholds?

The advantage of the newly proposed methods, based on the smooth distribution function, is that this quandary is neatly averted by providing a suitably smooth estimate of sensitivity and specificity. Indeed, the method borrows strength from thresholds in a neighbourhood about the threshold of interest θ, by allowing the estimates of sensitivity or specificity at neighbouring thresholds to contribute to the estimates of sensitivity or specificity at θ. This smoothing of the data introduces bias, but reduces the variability in the estimates of sensitivity and specificity. This trade off is controlled by the bandwidth parameter h. At one extreme, when $h \to \infty$ the newly proposed methods have minimal variance, but large bias. Conversely, when the bandwidth is small the newly proposed methods will have lower bias, but larger variance. In fact, when $h \to 0$, the newly proposed method on the logit scale reduces to the method by Reitsma et al. [97]. The added flexibility afforded by the smoothing parameter h can be regarded as a considerable

advantage of this method.

Another advantage of the newly proposed smooth methods, revealed by both the simulation study and the real data example, is that the smooth methods are far less likely to generate an invalid ROC curve. It was demonstrated that the existing methods can often produce a non-increasing ROC curve, particularly when the data are excessively sparse. This creates a number of problems when trying to draw inferences from the ROC curve (or indeed the corresponding sensitivity and specificity curves).

As we have already alluded to, one of the disadvantages of smoothing the data is that it introduces bias to the estimates of sensitivity and specificity. This is something that was particularly evident from the large sample simulation results of section 7.5. Indeed, when the study specific sample sizes n_i were sampled from $U(100, 300)$, the existing methods consistently outperformed the smooth methods in terms of bias. However, it is worthy of note that this disparity in bias between the new and existing procedures is less clear for smaller sample sizes. Indeed, when n was sampled from $U(20, 50)$ the new method on the raw scale was actually less biased than the method proposed by Reitsma et al. and competed closely with the exact method of Chu and Cole.

Another limitation of the newly proposed methodology is the requirement of individual patient data, so that all thresholds are available in all studies. Indeed, this is likely to be an extremely restrictive requirement in practice, since aggregated data is much more readily available. By limiting a meta-analysis to studies that publish individual patient data, one is likely to omit a large amount of useful aggregated data. By contrast, the methods given by Reitsma et al. and Chu and Cole can readily utilise aggregated data. Indeed, incorporating aggregated data is often an issue for conducting a meta-analysis of diagnostic test accuracy, as different studies will often report different thresholds. Riley et al. [102] comment that missing thresholds often cause the binomial method proposed by Hamza et al. [48] to not converge. For this reason, they extend the method of Reitsma et al. [97] as it can readily handle studies with missing thresholds. Moreover, one can jointly model all thresholds and obtain smooth, well-defined, ROC curves. However, the simulation study in this chapter indicates that this

method is likely to produce downwardly biased estimates of sensitivity and specificity, due to the need for the Normal approximation. Although, this may be accepted if one can use more of the evidence and produce a well-defined ROC curve. Riley et al. [103] also propose a sensitivity analysis to examine the impact of missing thresholds that retains the binomial modelling, which relies on imputed data.

7.6.2 Recommendations

It is clear that the exact binomial modelling approach of Chu and Cole should be the default method when individual patient data are available. Indeed, where possible it is recommended to carry out the approach of Chu and Cole at each threshold simultaneously, as this produces unbiased estimates and suitable coverage. Occasionally, however, the sROC this produces may not be well-defined (strictly increasing) as sensitivity and specificity are being estimated completely independently of neighbouring thresholds and thus are not suitably constrained. If a smooth ROC curve is of interest, one might therefore utilise the approach of Hamza et al. [48] that extends the Chu and Cole method to fit smooth functions across studies. However, this may not always be a viable option and, especially for sparse data, the newly proposed smooth method (on the raw scale) therefore provides a useful alternative.

Indeed, the desirable features of the new methods discussed here indicate that the new approaches provide a helpful addition to the existing methodology for analysing diagnostic tests. The smooth methods are particularly useful when the data are very sparse or there are gaps in the data at the threshold of interest, as they allow the construction of smooth curves for sensitivity and specificity. In these cases we recommend applying the new methodology on the raw scale alongside the exact method by Chu and Cole, to validate the results of the smooth method. If the two methods are in agreement then one can use the smooth method to estimate the sensitivity and specificity at any threshold, a luxury not afforded by the Chu and Cole method. If, on the other hand, the results of the two methods are excessively disparate then one is forced to conclude that the smooth method does not provide sufficiently accurate estimates of sensitivity and specificity.

7.6.3 Possible extensions

This chapter has focused primarily on the performance of two new methodologies for analysing diagnostic test accuracy. In particular, we considered two bivariate models based on the smooth distribution function estimate, one on the raw scale and the other on the logit scale. However, there are number of possible extensions to both the methodology and the analysis conducted here. We discussed, briefly, the possibility of performing a multivariate meta-analysis at every threshold simultaneously. This would allow for the additional borrowing of strength from neighbouring thresholds, by allowing for the very real possibility of correlations in the estimates of sensitivity and specificity between thresholds [99].

As previously stated, the primary focus of this chapter has been to investigate the usefulness of a meta-analysis of a density or distribution function to the analysis of DTA studies. However, the methodology discussed in section 7.2 can be readily adapted for use in a number of different settings. For example, one can modify the methods to estimate the hazard rate or survival function when there are individual patient data available from multiple heterogeneous studies. Similarly, the methods can also be generalised to perform non-parametric curve estimation in a regression setting where there are multiple heterogeneous data sets.

In the next, and final, chapter we discuss the conclusions of the thesis as a whole and identify some areas for further research.

Chapter 8

Conclusions and future work

8.1 Overview of thesis

The overarching theme of this thesis has been to develop non-parametric methodology, with the principal aim of making inferences about a random sample with an unknown distribution function. In fact, the thesis can be partitioned in a number of different ways. For example, Chapter 2 is driven by developing the mathematical theory therein, while the subsequent chapters are motivated by genuine applications in medical statistics. To look at it another way, Chapters 2 to 6 focus exclusively on the importance of symmetry and asymmetry in statistical models, whereas Chapter 7 broadens the scope of the investigation to assessing the distribution of data in a more general sense. Meta-analysis and, more generally, the analysis of medical research data have also been recurring themes.

In particular, we started by discussing how to effectively measure asymmetry and test for symmetry, and introduced a recently proposed measure for asymmetry $\widehat{\eta}$. Using this measure we developed a new test for symmetry, before evaluating its improvement on existing methods. We then discussed how the measure (and test) can be used to inform effectual analysis of statistical models of relevance to medical research. We also considered a new measure, which was shown to provide a valuable aid when making small sample inferences about η. Both of these procedures were then applied to the problem of assessing the asymmetry in multiple trials, by considering a meta-analysis of the measures of asymmetry. Next, we conducted a simulation study to analyse

what effect ignoring violations of symmetry assumptions can have on some of the most commonly used statistical models. Finally, we broadened our investigation to develop methods to synthesise information about the entire distribution of several samples, by proposing a meta-analysis of the density or distribution function. This led to the development of a novel method for synthesising test accuracy studies.

8.2 Key findings from each chapter

This thesis has contributed towards the development of new statistical methodology and novel medical statistics applications. The key contributions are listed in the box below and summarised in more detail thereafter.

Chapter 2

- Revealed that existing tests of symmetry do not have power which is reflective of the size of asymmetry.
- Proposed a new test for symmetry, based on a recently proposed measure of asymmetry η
- The power of the new test was shown to better reflect the size of asymmetry, and to improve upon the power of most existing tests.

Chapter 3

- Demonstrated the applicability of the asymmetry measure $\widehat{\eta}$ (and the corresponding test for symmetry) in a number of situations of relevance to medical research.

Chapter 4

- Investigated the small sample performance of the asymmetry measure $\widehat{\eta}$ and proposed an alternative measure $\widehat{\zeta}$ theoretically better suited to small samples.
- For symmetric distributions the sampling distribution of $\widehat{\eta}$ appears to be approximately Normal, even for samples as small as $n = 15$. Therefore, in the context

of testing for symmetry, p-values should not be too badly compromised when the sample size is small.

- The most accurate confidence intervals (in terms of coverage) are obtained by using the asymptotic variance for $\widehat{\eta}$.

Chapter 5

- Explored the problem of assessing the asymmetry in multiple studies.
- When every study is reasonably large (say, $n > 30$), it is recommended that meta-analysis be performed directly on $\widehat{\eta}$.
- If there are a number of studies with insufficient sample sizes to ensure the accuracy of the asymptotic theory, then the results should be the subject of a sensitivity analysis by also considering a meta-analysis of $\widehat{\zeta}$.

Chapter 6

- Investigated the impact of asymmetry on linear models and meta-analyses in the context of medical statistics.
- Conventional inferences (i.e. estimating an average treatment effect and confidence interval) are generally robust to asymmetry (in the residuals or the random effects) for linear models and meta-analyses.
- Clinically relevant inferences (i.e. prediction of future observations/effects or probability statements) are potentially seriously compromised by asymmetry.

Chapter 7

- Extended our investigations to synthesise information about the density and distribution function as a whole.
- Explored the potential applications of this method to meta-analyses of diagnostic test accuracy.

- Where possible, the exact binomial approach of Chu and Cole [19] should be the default, as this produces unbiased estimates and suitable coverage.
- New methodology can be useful, particularly for sparse data where a smooth ROC curve is desired.

In Chapter 2 it was shown that, while there are a wealth of options for a statistician wishing to test for symmetry on a set of data, some of the most commonly used tests of symmetry do not have power which is reflective of the size of asymmetry. That is, while the tests have good rejection levels for non-symmetric distributions, their power does not increase as asymmetry increases. This is because the primary rationale for the test statistics that are proposed in the literature to test the symmetry is to detect the departure from symmetry, rather than the quantification of the asymmetry. As a result, tests of symmetry based upon these statistics do not necessarily generate power that is representative of the departure from the null hypothesis of symmetry. Recent research by Patil et al. [88] has produced new measures of asymmetry, which have been shown to effectively quantify the amount of asymmetry. We proposed a new test based upon one such measure $\widehat{\eta}$ and derived the asymptotic distribution of the test statistic, before analysing the performance of this proposed test through the use of a simulation study. Our primary aim for Chapter 2 was to develop a new test for symmetry that improves upon the existing methods. This new test, based on a recently proposed measure of asymmetry, was shown to be an extremely competitive test. Indeed, a simulation study revealed that the new test exhibits an increase in power compared to most tests. Moreover, it was also shown that the new test had power that better reflected the amount of asymmetry in the underlying data. This work is the subject of a research paper by Partlett and Patil, which is currently under review for publication in the Annals of the Institute of Statistical Mathematics.

In Chapter 3 we discussed the potential applications of the new test, with a particular emphasis on informing the analysis of randomised trials in medical research. Moreover, with the help of a large data set consisting of multiple randomised control trials investigating hypertension, it was also shown that the measure $\widehat{\eta}$ is an effective summary statistic. Indeed, $\widehat{\eta}$ can be

used for making inferences such as assessing baseline imbalance in clinical trials and checking for asymmetry in linear model residuals. In addition, it was revealed that $\widehat{\eta}$ can be a useful aid in preconditioning data. That is, it can be used to objectively determine whether asymmetry within data is sufficient to require a normalising transform, inform what sort of transformation is likely to be effective, and provide an objective measure of whether the transformation has been a success. However, it was also established that one of the shortcomings of $\widehat{\eta}$ is that the distribution given in Chapter 2 is asymptotic. As a result, one requires a reasonably large sample to apply the Normal approximation.

In Chapter 4 we identified and addressed the problem of small sample estimation of η. In particular, it was demonstrated that it is not appropriate to assume a Normal distribution for $\widehat{\eta}$ in small samples. This provided the motivation for a new measure $\widehat{\zeta}$, which was subsequently shown to be more informative for small samples. We derived the asymptotic distribution of this new measure and demonstrated that this distribution is more robust to small samples. That is, $\widehat{\zeta}$ appeared to follow a Normal distribution more closely for finite samples. We also introduced and discussed the bootstrap to assist in the calculation of the standard error and confidence intervals for the new measure. Yet, a simulation study revealed that, irrespective of sample size, the best coverage results were obtained by using the asymptotic variance for $\widehat{\eta}$, truncating the confidence intervals at -1 or 1 if necessary. Hence, one might check how discrepant the results are to the $\widehat{\zeta}$ approach using bootstrapping, and, if the two approaches are unreasonably incongruent then one can conclude that there are too few samples to accurately assess the asymmetry in the population.

In Chapter 5 both $\widehat{\eta}$ and $\widehat{\zeta}$ were applied to the more general problem of assessing the asymmetry in multiple studies. That is, we investigated measuring asymmetry in data across several studies with the aim of obtaining an 'overall' measure of asymmetry in the underlying population. In particular, we explored the potential pitfalls of simply pooling data across all studies and treating it as a single study, before proposing a meta-analysis on the asymmetry measures to combat this problem. A meta-analysis of the asymmetry measures $\widehat{\eta}$ and $\widehat{\zeta}$ was shown to have a number of genuine applications in medical statistics, principally, allowing the comparison

and synthesis of information about the distribution of multiple (possibly heterogeneous) samples from a single population. It was shown that when the sample size is reasonably large in every study, one can carry out the meta-analysis directly on the asymmetry measure $\widehat{\eta}$ using the asymptotic theory to calculate the variance. However, if several studies have small samples then the asymptotic theory begins to breaks down. In these cases, it was demonstrated that $\widehat{\eta}$ may still be appropriate, but the best course of action is to examine whether the conclusions change by performing a meta-analysis of $\widehat{\zeta}$. If there is a reasonable agreement between the methods then the confidence intervals based on $\widehat{\eta}$ were shown (by the simulation study in Chapter 4) to produce the most accurate confidence intervals in terms of coverage. If the results of the two methods are excessively discrepant then this indicates that further research is needed to draw firm conclusions regarding the amount of asymmetry in the underlying population.

We concluded our investigation of asymmetry in Chapter 6 with an extensive analysis into the effect that ignoring asymmetry in the data has on the accuracy of the inferences drawn from a number of popular statistical models, and assessed to what extent these methods are robust to departures from symmetry. In particular, we discussed the potential pitfalls of ignoring violations of symmetry or normality assumptions in linear models and meta-analyses. The principal aim here was to formulate some key recommendations regarding the sample size or number of studies required to perform accurate analyses in this case. It was demonstrated that the common inferences that one draws from linear models or meta-analyses, particularly confidence intervals of the treatment effect estimate, are robust to departures from symmetry. However, it is worth noting that in both settings, if the aim is to make predictions about future patients or treatments then it is possible that these predictions may be compromised by the presence of asymmetry, either in the residual or random effects distribution. Indeed, while prediction intervals may have suitable coverage when the normality assumption is inappropriate, they may still be unhelpful clinically. Therefore, if one is interested in making clinically relevant inferences regarding the prediction of future observations (using linear models or meta-analyses) then it is recommended to account for the asymmetry. For example, if one suspects that there is substantial asymmetry in the response variable of a linear model, then it may be prudent to

make apply a transformation to the data before making inferences about future observations. Similarly, if there is evidence of asymmetry in the random effects of a meta-analysis then it is recommended to apply a more flexible method to account for this asymmetry, for example, those proposed by Lee and Thompson [72] or Karabatsos et al. [65].

In Chapter 7, we moved away from the investigation of the asymmetry of an unknown distribution and considered, more generally, a method for synthesising information about the distribution as a whole. That is, we widened the scope of our investigation to encompass the problem or comparing and synthesising information about the probability density function and cumulative distribution function, based on several samples from similar (but possibly heterogeneous) populations. That is to say, we generalised the methods of Chapter 5 to allow for the synthesis of information about the entire density or distribution function of several studies. We also provided a detailed demonstration of one potential application of this method, namely, conducting meta-analyses of diagnostic test accuracy studies. In particular, we proposed a couple of novel methods that have a number of desirable properties when compared to the existing methodology in this area. The ultimate objective of Chapter 7 was to develop these new methods for conducting meta-analyses of diagnostic test accuracy studies, and compare their accuracy with the existing methods proposed by Reitsma et al. [97] and Chu and Cole [19]. It was demonstrated that the exact binomial modelling approach of Chu and Cole should be the default method when individual patient data are available, as this produces unbiased estimates and suitable coverage. Occasionally, however, the sROC this produces may not be well-defined (strictly increasing) as sensitivity and specificity are being estimated completely independently of neighbouring thresholds and thus are not suitably constrained. If a smooth ROC curve is of interest, one might therefore utilise the approach of Hamza et al. [48] that extends the Chu and Cole method, however, for sparse data this may not always be a viable option. Thus, the newly proposed smooth method on the raw scale provides a useful alternative, especially for sparse data. In these cases it is recommended to apply the new methodology on the raw scale alongside the exact method by Chu and Cole, to validate the results of the smooth method. If the two methods are in agreement then one can use the smooth method to estimate the sensitivity and

specificity at any threshold, a luxury not afforded by the Chu and Cole method. If, on the other hand, the results of the two methods are excessively disparate then one is forced to conclude that the smooth method does not provide sufficiently accurate estimates of sensitivity and specificity. This is because, just like the method proposed by Reitsma et al., the newly proposed smooth methods are biased.

The main conclusion of this thesis is that asymmetry plays a crucial role in statistical models. In particular, it is important to validate symmetry and other distributional assumptions. Indeed, the simulation study in Chapter 6 identified that some of the commonly used statistical models (e.g. linear models and meta-analyses) are vulnerable to departures for symmetry, with predictions particularly influenced. Thus, when one suspects that these distributional assumptions have been violated, it is imperative to assess the exact nature of the underlying distribution. This accentuates the importance of the methods outlined in Chapters 5 and 7, which aim to synthesise information about the distribution underpinning multiple studies.

8.3 Publications submitted and in-progress

The contents of Chapter 2 (along with elements of Chapter 4) is the subject of a research paper by Partlett and Patil entitled "Measuring Asymmetry and Testing Symmetry", which is currently under review for publication in the Annals of the Institute of Statistical Mathematics.

Furthermore, a paper entitled "Random effects meta-analysis and the consequences of ignoring asymmetry", which expands the meta-analysis simulation study in Chapter 6, is at an advanced draft stage. Additionally, Chapter 5 and Chapter 7 will also make interesting methodology papers and drafts are currently being prepared for later this year.

8.4 Future work

A consequence of the reasonably broad nature of this thesis is that there are a wide variety of possibilities to extend it. For example, one extension could be to consider developing a test based on the strong measure of asymmetry η_s, which was briefly introduced in Chapter 2. To recap, Patil, Bagkavos and Wood [89] discuss a stronger measure η_s, where the condition $\eta_s = 0$ is, both, necessary and sufficient for the symmetry. A drawback of the stronger measure η_s is

a loss of the 'user-friendly' aspect of η. Despite this, further investigation into the estimators of η_s and their asymptotic properties is of genuine interest as Patil, Bagkavos and Wood show that $\eta_s \geq \eta$. Therefore, a test based on an estimate of η_s has the potential to be more powerful than than the test developed in Chapter 2. Given the already impressive performance of the test based on $\widehat{\eta}$ this could be an extremely valuable extension. Aside from testing for symmetry it would also be of interest to see how η_s lends itself to analysis of randomised trials.

Another possible extension is to investigate the usefulness of other bootstrap methods for drawing inferences based on $\widehat{\eta}$ in small samples. For example, in this thesis we focused on the non-parametric bootstrap with simple bootstrap confidence intervals, but, it may be possible to approximate the sampling distribution of $\widehat{\eta}$ more effectively by applying other bootstrap methods, such as the parametric bootstrap or the smoothed bootstrap [27]. Additionally, one may be able to improve the upon the coverage results of Chapter 4 by constructing bootstrap confidence intervals that adjust for bias [29]. In a similar vein, it would also be of interest to assess the small sample properties of $\widehat{\eta}_s$.

Furthermore, supposing that one is able to demonstrate the asymptotic normality of the stronger measure $\widehat{\eta}_s$, the measure would readily lend itself to the meta-analysis methods discussed in Chapter 5. It would also be of interest to perform a more rigorous analysis of these meta-analysis methods for varying levels of heterogeneity via a simulation study. In addition, one could readily extend the meta-analysis results of this chapter to perform meta-analyses of the classical skewness measures, for example, those proposed by Pearson [91] and introduced in Chapter 1. In particular, it would be interesting to compare these results with the meta-analysis of $\widehat{\eta}$, as well as determine whether or not the conclusions that are drawn about the overall distribution of a collection of studies are sensitive to the measure of skewness or asymmetry that is used to compare them.

For the simulation study in Chapter 6 it would of interest to open the study up to a wider class of asymmetric distributions, akin to those considered in Chapter 2. Moreover, this simulation study could also be extended to investigate the effect of unchecked asymmetry on the bias and standard error of the model estimates and the impact on prediction error for linear and

meta-analysis models. In fact, these extensions are being currently being pursued in a draft of a paper titled "Random effects meta-analysis and the consequences of ignoring asymmetry", based on the second half of Chapter 6. It would also be interesting to see whether the results generalise to more complicated meta-regression models [118] or even multivariate models. In particular, one could assess the impact of asymmetry in the data or the random effects on the estimation of covariates or between study covariances and the subsequent implications for clinical interpretations. Equally pressing is a more thorough analysis of the performance of the flexible Bayesian meta-analysis methods, which are frequently lauded in this thesis [72, 78]. While these methods have been appraised by Lee and Thompson [72] and Martínez-Camblor [78] in the context of skewed data, a more meticulous appraisal of their performance for a wide range of asymmetric distributions would further elucidate their robustness to asymmetry. Similarly, an in-depth evaluation of the robustness of Bayesian linear regression to asymmetric data would be equally worthy of attention [15]. Alternatively, for random effects meta-analysis, one could appraise the performance of profile likelihood [49], recommended by Cornell et al. [20], for the broad range of asymmetric random effects distributions considered here.

Another obvious extension to Chapter 6 would be to investigate methods for examining, in a statistically rigorous fashion, whether there is asymmetry in the random effects. Methods for evaluating the asymmetry in this context should account for the fact that there is likely to be only a small number of studies. Moreover, it is imperative that the methods guard against finding artificial patterns in the data.

As we have previously alluded to, in Chapter 7 we carried out a detailed demonstration of one potential application to our meta-analysis of a function, namely, conducting meta-analyses of diagnostic test accuracy studies. In particular, we proposed a couple of novel methods that have a number of desirable properties when compared to the existing methodology in this area. However, there are number of possible extensions to the methodology. For example, the possibility of performing a multivariate meta-analysis at every threshold simultaneously was briefly raised, but not pursued in detail. This would allow for the additional borrowing of strength from neighbouring thresholds by allowing for the very real possibility of correlations in the estimates

of sensitivity and specificity between thresholds [48, 102, 103]. In addition, we also briefly discussed the possibility of performing a two-stage estimate of the distribution function. It would, of course, be of considerable interest to investigate the feasibility of both of these methods in more detail.

Moreover, the newly proposed meta-analysis methods were only compared with two examples from the existing methodology, namely, the methods proposed by Reitsma et al. [97] and Chu and Cole [19]. In fact, there are a wealth of other meta-analysis methods which have their roots in analysing diagnostic test accuracy, and could also be compared with the new methods. For example, Rutter and Gatsonis [107] describe a multilevel model for fitting a hierarchical summary ROC (HSROC) curve and Kuss et al. [70] recently proposed a method based on beta-binomial distributions.

Furthermore, the meta-analysis methods in Chapter 7 can be readily generalised for use in a number of different settings. For example, a meta-analysis of a function can be used to estimate the hazard rate or survival function when one has access to individual patient data available from multiple, possibly heterogeneous, studies. The methods can also be applied in an even more general setting, namely, to perform non-parametric regression curve estimation in the case where there are multiple, possibly heterogeneous, data sets. Indeed, both of these topics would make for fascinating research topics in there own regard.

8.5 Conclusion

This thesis has provided an important contribution to the theory and application of examining distributional assumptions and properties, with particular emphasis on quantifying and examining asymmetry. We also investigated several applications using single and multiple research studies. Though many issues remain for future work, it is hoped that the findings will help improve the awareness and application of measuring asymmetry and encourage more consideration of distributional assumptions when making predictions and utilising data from multiple studies.

Appendix A
Giné and Mason proof

A.1 Introduction

Let $X_1, \cdots, X_n$ be a random sample from a population with distribution function F. Further, let $F^{(j)}(x) = \frac{d^j}{dx^j}F(x)$. Giné and Mason [38] study the uniform in bandwidth behaviour of estimators of

$$T(F) = \int_R \phi\left(x, F(x), F^{(1)}(x), \cdots, F^{(r)}(x)\right) dF(x),$$

which have the form

$$T_n(h) = \frac{1}{n}\sum_{i=1}^{n} \phi\left(X_i, F_{n,i}(X_i), F^{(1)}_{n,h_1,i}(X_i), \cdots, F^{(r)}_{n,h_r,i}(X_i)\right),$$

where $h = (h_1, \cdots, h_r), h_i > 0$, is a vector of bandwidths and

$$F_{n,i}(X_i) = \frac{1}{n-1}\sum_{1\le i\ne j\le n} K_0^{(-1)}(X_i - X_j),$$

with $K_0^{(-1)}(x) = I\{x > 0\}$, and for $1 \le m \le r$,

$$F^{(m)}_{n,h_m,i}(X_i) = \frac{1}{n-1}\frac{1}{h^m}\sum_{1\le i\ne j\le n} K_m^{(m-1)}\left(\frac{X_i - X_j}{h}\right), \quad i = 1, \cdots, n,$$

where $K_m^{(0)} = K_m$ is an L_1 kernel, $m-1$ times differentiable with $K_m^{(m-1)}(x) = \frac{d^{m-1}K_m(x)}{dx^{m-1}}$ satisfying

$$\int_R K_m(u)du = 1.$$

Define the sequence of processes in $\vec{\lambda} = (\lambda_1, \cdots, \lambda_r), 0 < a \le \lambda_m \le b < \infty$, for $m = 1, \cdots r$, by

$$v_n\left(\vec{\lambda}\right) = \sqrt{n}\left\{T_n(\vec{\lambda} \circ \mathbf{h_n}) - T(F)\right\},$$

where $\{\mathbf{h_n}\}$ is a sequence of vectors $\mathbf{h_n} = (h_{1,n}, \cdots h_{r,n})$ with positive coordinates converging to zero and

$$\vec{\lambda} \circ \mathbf{h_n} = (\lambda_1 h_{1,n}, \cdots, \lambda_r h_{r,n}).$$

Giné and Mason construct i.i.d. random variables $Y_1, Y_2, \cdots, Y_n$ with mean 0 and finite variance, such that

$$\sup_{\vec{\lambda}\in[a,b]^r} \left| v_n\left(\vec{\lambda}\right) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i \right| = o_p(1). \tag{A.1}$$

A.2 Notation

Towards defining the random variables $Y_1, Y_2, \cdots, Y_n$, let

$$\varphi(x) = \varphi\left(x, F(x), F^{(1)}(x), \cdots, F^{(r)}(x)\right),$$

and

$$\varphi_m(x) = \frac{\partial}{\partial y_m} \varphi(x, y_0, y_1, \cdots, y_r)|_{(y_0,y_1,\cdots,y_r)=(F(x),F^{(1)}(x),\cdots,F^{(r)}(x))}.$$

Let

$$\zeta(X_i) = \varphi(X_i) - \mathrm{E}\varphi(X) = \varphi(X_i) - T(F),$$

$$\zeta_0(X_i) = \int_{X_i}^{\infty} \varphi_0(y) f(y) dy - \int_R F(y)\varphi_0(y) f(y) dy,$$

and, for $m = 1, \cdots, r$, let

$$\chi_m(y) = \frac{d^{m-1}}{dy^{m-1}} \left(\varphi(y) f(y)\right) = -\frac{d^m}{dy^m} \int_y^{\infty} \varphi(x) f(x) dx,$$

where $\frac{d^0}{dy^0} g(y) = g(y)$, and set for $i \geq 1$,

$$\zeta_m(X_i) = (-1)^{m-1} \left\{\chi_m(X_i) - \mathrm{E}\chi_m(X)\right\}.$$

Define

$$Y_i = \zeta(X_i) + \sum_{m=0}^{r} \zeta_m(X_i), \quad i \geq 1. \tag{A.2}$$

A.3 Conditions

For the main result one requires the following assumptions:

I For each $m = 1, \cdots, r$, $f^{(m-1)}$ is bounded.

II For some $0 < \alpha \leq 1$, and each $m = 1, \cdots, r$, $f^{(m-1)}$ is in $G_{2r+\alpha-1,K_m}(H)$, where H is some non-negative measurable function satisfying $\mathrm{E}\left[H^2(X)\right] < \infty$, which may depend on $f^{(m-1)}$. $G_{s,k}(H)$ is the class of all measurable functions g on $\mathbb{R}$ for which there is a positive constant $M_K(g)$ such that for all h sufficiently small

$$\left| \frac{1}{h} \int_R g(u) K\left(\frac{x-u}{h}\right) du - g(x) \right| \leq h^s M_k(g) H(x), \quad \forall x.$$

III Uniformly in $x \in \mathrm{supp}(f)$, the function $\varphi(x, y_0, y_1, \cdots, y_r)$ and its partial derivatives $\partial\varphi/\partial y_i, i = 0, \cdots, r$, satisfy a global Lipschitz condition with respect to the variables

$y_0, \cdots, y_r$ on a bounded open convex subset of $\mathbb{R}^{r+1}$ containing the closure of the range,

$$\left\{ \left(F(x), f(x), f^{(1)}(x), \cdots, f^{(r-1)}(x) \right) \middle| \, x \in \mathbb{R}. \right\}.$$

IV $\varphi_m, m = 0, 1, \cdots, r$, are bounded on the support of f.

V For each $m = 1, \cdots, r$, the function χ_m is Lipschitz of order $0 < \beta \leq 1$ and

$$\int_R |u|^\beta |K_m(u)| du < \infty.$$

VI The kernels $K_m^{(m-1)}, m = 1, \cdots, r$, are in $L_2(\mathbb{R})$ and of bounded variation in $\mathbb{R}$.

VII The kernels K_m are smooth and has bounded support.

VIII With $0 < \alpha \leq 1$ as in condition *II*, for $m = 1, \cdots, r$,

$$\sqrt{n} \frac{h_{m,n}^{2r-1}}{\log(1/h_{m,n})} \to \infty \text{ and } \sqrt{n} h_{m,n}^{2r+\alpha-1} \to 0.$$

A.4 Theorem statement

Theorem A.3 *Under conditions I-VIII, for all* $0 < a < b < \infty$*, equation* (A.1) *holds. Moreover, if* $0 < \mathrm{Var}(Y) < \infty$*, where* Y *is defined as in equation* (A.2)*, then the processes* $\left\{ v_n\left(\vec{\lambda}\right) : \vec{\lambda} \in [a,b]^r \right\}$ *converge in law in* $l_\infty([a,b]^r)$ *(in the sense of Hoffmann-Jorgensen) to the constant Gaussian process* $G\left(\vec{\lambda}\right) = G$ *where* G *is* $N(0, \mathrm{Var}(Y_1))$*, that is,*

$$\left\{ T_n(\vec{\lambda} \circ \mathbf{h_n}) - T(F) \right\} \longrightarrow_d N(0, \mathrm{Var}(Y_1)),$$

uniformly in $\vec{\lambda} \in [a,b]^r$.

Appendix B
Selected R code

Introduction

All the simulations and plots in this thesis were produced using the statistical programming language R. In this appendix we present a selection of the R code used in this thesis. The required packages include, but are not limited to, boot, meta, mvmeta, zoo and fdrtool.

Calculating the test statistics based on η

The following simple R function takes two arguments: `x`- a set of data, and `xcuts`- the number of points to use in the numerical integration (used to estimate the variance). The function returns the standardised test statistics based on $\widehat{\eta}$ and $\widehat{\zeta}$.

```
uTnm_T <- function(x , xcuts = 512){

#1. Preamble functions

fisherz<-function(r){
Z = (1/2)*log( (1+r)/(1-r) )
return(Z)
}

AUC <- function(X,Y){
id <- order(X)
int = 0
if( length(Y) > 1){ int = sum(diff(X[id])*rollmean(Y[id],2)) }
return(int)
}

## Calculate integrals based on X at the points x

If2<-function(X,f2,x){

n = length(x)
if2 = rep(0,n)

for( i in 1:n){
```

```
# pick out all the Xs > x

Xx  =  X[ X > x[i] ]
f2x = f2[ X > x[i] ]

if2[i] = AUC(Xx,f2x)

}
return(if2)
}

# #

IfF<-function(X,f,F,x){

n = length(x)

ifF = rep(0,n)

fF = (F - 0.5)*f

for( i in 1:n){

Xx =   X[X > x[i] ]
fFx = fF[X > x[i] ]

ifF[i] = AUC(Xx,fFx)

}

return(ifF)
}

# #

remNaN<-function(x){

y <- x[!is.na(x)]

return(y)
}


#2. Calls a C function to calculate f^(Xi)
```

```
fden<-function(x){
    x = sort(x)
    nx = length(x)
    v = rep(0, nx)
    h = bw.nrd0(x)
    crun = .C("fden", as.double(x), out = as.double(v), as.integer(nx),
        as.double(h))
    fnh = (1/(nx - 1)) * (1/h) * (crun$out)
    return(fnh)
}


#3. Main function: Calculating eta-hat

x=remNaN(x)

x = sort(x)

n = length(x)

f = fden(x)
F = (ecdf(x))(x)

#
denf = density( x , kernel = c("epanechnikov") , n = xcuts )

xs = denf$x
fs = denf$y
Fs = (ecdf(x))(xs)
#

Tcov = -cov( f , F )

etah = -cor( f , F )

zetah = fisherz(etah)


#4. Estimating the variance

ix  = If2(xs,fs^2,x)
ix2 = IfF(xs,fs,Fs,x)

A = var(f)
B = var(F)
```

```
mu = mean(f)

Y = (   (2/sqrt(A*B))*( f*F - 0.5*f ) + (1/sqrt(A*B))*ix +
etah*( (F - 0.5)^2/(2*B) + (f-mu)^2/(2*A)
           + (1/B)*ix2 + (1/A)*(f - mu)*f )  )

sig2 = var(Y)
zsig2 = var(Y)*(1/( 1 - etah^2 )^2)

# #

T1 = sqrt(n)*etah/sqrt(sig2)
T2 = sqrt(n)*zetah/sqrt(zsig2)

return( list( eta = T1 , zeta = T2  ) )
}


##
```

Smooth meta-analysis method (Chapter 7)

The following function is used to perform meta-analyses of individual patient diagnostic test accuracy data using the smooth distribution function. The function takes a number of arguments, but most crucially `XDXND`- the continuous outcome in the diseased and non-diseased group respectively, and `pos_for`- which specifies whether the test gives a positive outcome for large or small values.

```
logit_Fh_dta<-function(XD , XND , xx = xx0 , nx=32 , tc = 0.5 ,
          cc = "study" , meth = "reml" , pos_for =  "large" , Sigma = "unstr",
          bw.scale = 1, max.iter = 100 , study_labels = NULL  ,
          ss_lb = 0 , ss_ub = Inf , scale = "logit"  ){

## xx - the exact thresholds to determine the test accuracy
## nx - the number of equally spaced points at which to calculate the
##        test accuracy

## scale ="raw" carries out the meta-analysis on the raw sens/spec
## scale ="logit" carries out the meta-analysis on the logit scale.

## cc = "all" corrects every 2x2 in every study at every threshold.
## cc = "study" only corrects the 2x2 tables for studies and thresholds where
## there is an  offending 0.
## cc = "single" only corrects the offending observation.

#1. Preamble functions
```

```
remNaN<-function(x){

y <- x[!is.na(x)]

return(y)
}

logit <- function( x ){

log(x/(1-x))
}

ilogit<-function(x){

exp(x)/(exp(x)+1)
}

if( scale != "logit" & scale !="raw" ){

print("Scale must be one of logit or raw")
return()
 }

axb<-function( a , x , b ){

x = max( c( x , a ) )
x = min( c( x , b ) )

return(x)
}

#2. Prepare input data

XDm  = unlist(XD)
XDm  = remNaN(XDm)

XNDm = unlist(XND)
XNDm = remNaN(XNDm)

Xmerge = sort( c( XDm , XNDm ) )

fromm = min(Xmerge)
too = max(Xmerge)

xx0 = seq( fromm , too , length.out = nx )
```

```
if( xx[1] == "all"){ xx = Xmerge  }

nx=length(xx) # no. of meta-analyses

if( length(XD) != length(XND) ){ print("XD and XND have different length")}
k = length(XD)      # no. of studies

###
#3. Prepare outputs

metaA = rep(0,nx)
metaA = as.list( metaA )

Sens = rep(0,nx)
Spec = rep(0,nx)

seDi  = rep(0,nx)
seNDi = rep(0,nx)

Sensl = rep(0,nx)
Sensu = rep(0,nx)

Specl = rep(0,nx)
Specu = rep(0,nx)

Sens.tau = rep(0,nx)
Spec.tau = rep(0,nx)

converge_fail = rep(0,nx)

#4. Calculate smooth distribution function

Fhat<-function(data,xs,bws = 1){

data=sort(data)
n = length(data)
h = bw.nrd0(data)

h = h*bws

Fh = rep(0, length(xs))

for( i in 1:n){

 Fh = Fh + pnorm( (xs - data[i] )/h )
```

```
}

Fh = (1/n)*(Fh)
return(Fh)}

# variance: F(x)(1-F(x))/n
# need to correct for when Fh = 0 and Fh = 1
# Construct 2x2 table

for( j in 1:nx){

theta = xx[j]    #threshold

# tp = rep(0,k)
# fn = rep(0,k)
# tn = rep(0,k)
# fp = rep(0,k)

tpr = rep(0,k) #sens
tnr = rep(0,k) #spec

SS = as.list( rep(0,k))

for( i in 1:k){

Di  = XD[[i]]
NDi = XND[[i]]

Di = remNaN(Di)
NDi = remNaN(NDi)

nDi  = length( Di )
nNDi = length( NDi )


if( pos_for ==  "small" ){

FnXD  = Fhat(Di  ,theta, bws = bw.scale)          #sens
FnXND = Fhat(NDi ,theta, bws = bw.scale)       #1-spec

tp = FnXD*nDi
fn = nDi - FnXD*nDi
tn = nNDi - FnXND*nNDi
fp = FnXND*nNDi

}
```

```
if( pos_for ==  "large" ){

FnXD  = 1-Fhat(Di  ,theta, bws = bw.scale)          #sens
FnXND = 1-Fhat(NDi ,theta, bws = bw.scale)       #1-spec

tp = FnXD*nDi
fn = nDi  - FnXD*nDi
tn = nNDi - FnXND*nNDi
fp = FnXND*nNDi

}

if( cc == "all"){

tp = tp + tc
fn = fn + tc
tn = tn + tc
fp = fp + tc

nnDi  = tp + fn
nnNDi = fp + tn
tpr[i]  = tp/(nnDi)
tnr[i] = tn/(nnNDi)

}

if( cc == "study"){

## check to see if any cell is not greater than 0.
## due to numerical instability check to see whether
## any are >= nDi or nNDi as well

if( tp <= 0 | fn <= 0 | tn <= 0 | fp <= 0 |

tp >= nDi | fn >= nDi | tn >= nNDi | fp >= nNDi ){

tp = tp + tc
fn = fn + tc
tn = tn + tc
fp = fp + tc

}

nnDi  = tp + fn
nnNDi = fp + tn
```

```
tpr[i]  = tp/(nnDi)
tnr[i] = tn/(nnNDi)

}

if( cc == "single"){

##  check to see if any cell is not greater than 0.

 if( tp <= 0 | fn >= nDi  ){ tp = tp + tc }
 if( fn <= 0 | tp >= nDi  ){ fn = fn + tc }
 if( tn <= 0 | fp >= nNDi ){ tn = tn + tc }
 if( fp <= 0 | tn >= nNDi ){ fp = fp + tc }

nnDi  = tp + fn
nnNDi = fp + tn
tpr[i]  = tp/(nnDi)
tnr[i] = tn/(nnNDi)

}

if( scale == "raw" ){

S11 = tpr[i]*(1-tpr[i])/nDi
S22 = tnr[i]*(1-tnr[i])/nNDi

SS[[i]] = matrix(  c(  S11 ,  0 , 0  ,  S22 )  , 2 , 2   )

 }

## scale the variance

if( scale == "logit" ){

S11 = 1/( tpr[i]*(1-tpr[i])*nDi  )
S22 = 1/( tnr[i]*(1-tnr[i])*nNDi )

SS[[i]] = matrix(  c(  S11 ,  0 , 0  ,  S22 )  , 2 , 2   )

 }

## Additional check for computational issues in SS[[i]]]

SS[[i]][1,1] = axb( ss_lb , S11 , ss_ub)
SS[[i]][2,2] = axb( ss_lb , S22 , ss_ub)
```

```
}

## We now have a vector of length k for the estimate of the true
##  positive/negative rate for threshold xx[j]

## Compute the variance

sensSE = sqrt( tpr*(1-tpr)/nDi  )
specSE = sqrt( tnr*(1-tnr)/nNDi )

## Bivar set up

if( scale =="raw" ){

YY =   cbind(tpr,tnr)   }

## Appropriately scale sens and spec

if( scale =="logit" ){

YY = logit( cbind(tpr,tnr)   )

 }

#5. Perform the meta-analysis

metsAj = mvmeta( YY , SS , method = meth, bscov = Sigma ,
control = list( maxiter = max.iter ) )

metaA[[j]]  = metsAj

Sens[j] = (summary( metsAj )$coefficients)[1,1]
Spec[j] = (summary( metsAj )$coefficients)[2,1]

#
# Report the 95% upper and lower bounds
#

Sensl[j] = (summary( metsAj )$coefficients)[1,5]
Sensu[j] = (summary( metsAj )$coefficients)[1,6]

Specl[j] = (summary( metsAj )$coefficients)[2,5]
Specu[j] =  (summary( metsAj )$coefficients)[2,6]

# tau for raw/logit Sens and Spec
```

```
Sens.tau[j] = sqrt(summary( metsAj )$Psi[1,1])
Spec.tau[j] = sqrt(summary( metsAj )$Psi[2,2])

# Note whether convergence failed

converge_fail[j] = 1 - as.numeric( metsAj$converge )

}

## logit transform results

if( scale == "logit" ){

Sens = ilogit( Sens  )
Spec = ilogit( Spec  )

Sensl = ilogit( Sensl  )
Sensu = ilogit( Sensu  )

Specl = ilogit( Specl  )
Specu = ilogit( Specu  )

}

return(list( thresh = xx , Sens = Sens , Spec = Spec  , mets = metaA ,
Sensl = Sensl ,Sensu = Sensu ,Specl = Specl , Specu = Specu ,
Senstau =  Sens.tau, Spectau = Spec.tau , conv = converge_fail ) )
}

##
```

List of References

[1] G. Abo-Zaid, B. Guo, J. J. Deeks, T. P. Debray, E. W. Steyerberg, K. G. Moons, and R. D. Riley. Individual participant data meta-analyses should not ignore clustering. *Journal of Clinical Epidemiology*, 66(8):865–873, 2013.

[2] A. Amery, P. Brixko, D. Clement, A. De Schaepdryver, R. Fagard, J. Forte, J. Henry, G. Leonetti, K. O'Malley, T. Strasser, et al. Mortality and morbidity results from the european working party on high blood pressure in the elderly trial. *The Lancet*, 325 (8442):1349–1354, 1985.

[3] A. Antille, G. Kersting, and W. Zucchini. Testing symmetry. *Journal of the American Statistical Association*, 77(379):639–646, 1982.

[4] A. Azzalini. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12(2):171–178, 1985.

[5] L. M. Bachmann, M. A. Puhan, G. t. Riet, and P. M. Bossuyt. Sample sizes of studies on diagnostic accuracy: literature survey. *BMJ*, 332(7550):1127–1129, 2006.

[6] R. Baker and D. Jackson. A new approach to outliers in meta-analysis. *Health Care Management Science*, 11(2):121–131, 2008.

[7] M. S. Bartlett. Statistical estimation of density functions. *Sankhya: The Indian Journal of Statistics, Series A*, 25(3):245–254, 1963.

[8] F. Bochmann, Z. Johnson, and A. Azuara-Blanco. Sample size in studies on diagnostic accuracy in ophthalmology: A literature survey. *British journal of ophthalmology*, 91(7): 898–900, 2007.

[9] C. A. Boneau. The effects of violations of assumptions underlying the t test. *Psychological bulletin*, 57(1):49, 1960.

[10] D. D. Boos. A test for asymmetry associated with the Hodges-Lehmann estimator. *Journal of the American Statistical Association*, 77(379):647–651, 1982.

[11] G. N. Boshnakov. Some measures for asymmetry of distributions. *Statistics & Probability Letters*, 77(11):1111–1116, 2007.

[12] G. E. Box and G. C. Tiao. *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons, 2011.

[13] G. E. Box and G. S. Watson. Robustness to non-normality of regression tests. *Biometrika*, 49(1-2):93–106, 1962.

[14] S. E. Brockwell and I. R. Gordon. A simple method for inference on an overall effect in meta-analysis. *Statistics in Medicine*, 26(25):4531–4543, 2007.

[15] L. Broemeling. *Bayesian analysis of linear models*. Statistics, textbooks and monographs. M. Dekker, 1985.

[16] C. C. Butler. A test for symmetry using the sample distribution function. *The Annals of Mathematical Statistics*, 40(6):2209–2210, 1969.

[17] P. Cabilio and J. Masaro. A simple test of symmetry about an unknown median. *The Canadian Journal of Statistics*, 24(3):349–361, 1996.

[18] R. J. Carroll, K. Roeder, and L. Wasserman. Flexible parametric measurement error models. *Biometrics*, 55(1):44–54, 1999.

[19] H. Chu and S. R. Cole. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *Journal of Clinical Epidemiology*, 59 (12):1331–1332, 2006.

[20] J. E. Cornell, C. D. Mulrow, R. Localio, C. B. Stack, A. R. Meibohm, E. Guallar, and S. N. Goodman. Random-effects meta-analysis of inconsistent effects: a time for change. *Annals of Internal Medicine*, 160(4):267–270, 2014.

[21] F. Cribari-Neto and G. M. Cordeiro. On Bartlett and Bartlett-type corrections. *Econometric Reviews*, 15(4):339–367, 1996.

[22] F. David and N. Johnson. Some tests of significance with ordered variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 18(1):1–31, 1956.

[23] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and their Applications, Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press Cambridge, 1997.

[24] J. J. Deeks. Systematic reviews of evaluations of diagnostic and screening tests. *BMJ*, 323 (7305):157–162, 2001.

[25] R. DerSimonian and N. Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188, 1986.

[26] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *The Annals of Statistics*, 24(2):508–539, 1996.

[27] B. Efron. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7 (1):1–26, 1979.

[28] B. Efron. Nonparametric standard errors and confidence intervals. *The Canadian Journal of Statistics*, 9(2):139–158, 1981.

[29] B. Efron. Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association*, 82(397):171–185, 1987.

[30] B. Efron and R. Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):54–75, 1986.

[31] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1994.

[32] V. A. Epanechnikov. Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, 14(1):153–158, 1969.

[33] C. Fernandez and M. F. J. Steel. On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93(441):359–371, 1998.

[34] A. P. Field. Meta-analysis of correlation coefficients: a Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods*, 6(2):161–180, 2001.

[35] R. A. Fisher. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521, 1915.

[36] R. A. Fisher. On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32, 1921.

[37] R. C. Geary. Testing for normality. *Biometrika*, 34(3-4):209–242, 1947.

[38] E. Giné and D. Mason. Uniform in bandwidth estimation of integral functionals of the density function. *Scandinavian Journal of Statistics*, 35(4):739–761, 2008.

[39] G. Glass and J. Stanley. *Statistical methods in education and psychology*. Prentice-Hall series in educational measurement, research, and statistics. Prentice-Hall, 1970.

[40] G. V. Glass. Primary, Secondary, and Meta-Analysis of Research. *Educational Researcher*, 5(10):3–8, 1976.

[41] L. Grilli and C. Rampichini. Specification of random effects in multilevel models: a review. *Quality & Quantity (published online ahead of print on 29th July 2014)*, 2014. doi: 10.1007/s11135-014-0060-5.

[42] C. Gu and C. Qiu. Smoothing spline density estimation: Theory. *The Annals of Statistics*, 21(1):217–234, 1993.

[43] M. K. Gupta. An asymptotically nonparametric test of symmetry. *The Annals of Mathematical Statistics*, 38(3):849–866, 1967.

[44] P. Hall. *The bootstrap and Edgeworth expansion*. Springer series in statistics. Springer-Verlag, 1992.

[45] P. Hall and J. S. Marron. Estimation of integrated squared density derivatives. *Statistics & Probability Letters*, 6(2):109–115, 1987.

[46] P. Hall, J. Marron, and B. U. Park. Smoothed cross-validation. *Probability Theory and Related Fields*, 92(1):1–20, 1992.

[47] P. Hall, R. J. Hyndman, and Y. Fan. Nonparametric confidence intervals for receiver operating characteristic curves. *Biometrika*, 91(3):743–750, 2004.

[48] T. Hamza, L. Arends, H. van Houwelingen, and T. Stijnen. Multivariate random effects meta-analysis of diagnostic tests with multiple thresholds. *BMC Medical Research Methodology*, 9(1):73, 2009.

[49] R. J. Hardy and S. G. Thompson. A likelihood approach to meta-analysis with random effects. *Statistics in Medicine*, 15(6):619–629, 1996.

[50] J. Hartung and G. Knapp. On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Statistics in Medicine*, 20(12):1771–1782, 2001.

[51] L. Hedges and I. Olkin. *Statistical Methods for Meta-analysis*. Academic Press, 1985.

[52] S. J. Hershkorn and R. J. Chapman. Symmetric random variables. *The American Mathematical Monthly*, 105(7):670, 1998.

[53] J. P. T. Higgins and S. G. Thompson. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11):1539–1558, 2002.

[54] J. P. T. Higgins, I. R. White, and J. Anzures-Cabrera. Meta-analysis of skewed data: Combining results reported on log-transformed or raw scales. *Statistics in Medicine*, 27 (29):6072–6092, 2008.

[55] J. P. T. Higgins, S. G. Thompson, and D. J. Spiegelhalter. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):137–159, 2009.

[56] M. Hollander. *Testing for Symmetry*. John Wiley & Sons, Inc., 2004.

[57] P. J. Huber. Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, 1(5):799–821, 1973.

[58] J. Hunter and F. Schmidt. *Methods of meta-analysis: correcting error and bias in research findings*. Sage Publications, 1990.

[59] D. Jackson and J. Bowden. A re-evaluation of the quantile approximation method for random effects meta-analysis. *Statistics in Medicine*, 28(2):338–348, 2009.

[60] D. Jackson, I. R. White, and S. G. Thompson. Extending DerSimonian and Laird's methodology to perform multivariate random effects meta-analyses. *Statistics in Medicine*, 29(12):1282–1297, 2010.

[61] D. Jackson, R. Riley, and I. R. White. Multivariate meta-analysis: Potential and promise. *Statistics in Medicine*, 30(20):2481–2498, 2011.

[62] D. Jackson, I. R. White, and R. D. Riley. Quantifying the impact of between-study heterogeneity in multivariate meta-analyses. *Statistics in Medicine*, 31(29):3805–3820, 2012.

[63] M. Jones. Simple boundary correction for kernel density estimation. *Statistics and Computing*, 3(3):135–146, 1993.

[64] M. C. Jones and A. Pewsey. Sinh-arcsinh distributions. *Biometrika*, 96(4):761–780, 2009.

[65] G. Karabatsos, E. Talbott, and S. G. Walker. A Bayesian nonparametric meta-analysis model. *Research Synthesis Methods*, 6(1):28–44, 2015.

[66] M. G. Kenward and J. H. Roger. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, pages 983–997, 1997.

[67] G. Knapp and J. Hartung. Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, 22(17):2693–2710, 2003.

[68] E. Kontopantelis and D. Reeves. Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: A simulation study. *Statistical Methods in Medical Research*, 21(4):409–426, 2012.

[69] E. Kontopantelis and D. Reeves. Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: A comparison between DerSimonian–Laird and restricted maximum likelihood. *Statistical Methods in Medical Research*, 21(6): 657–659, 2012.

[70] O. Kuss, A. Hoyer, and A. Solms. Meta-analysis for diagnostic accuracy studies: a new statistical model using beta-binomial distributions and bivariate copulas. *Statistics in Medicine*, 33(1):17–30, 2014.

[71] A. Lam and P. D. Kerr. Parathyroid hormone: An early predictor of postthyroidectomy hypocalcemia. *The Laryngoscope*, 113(12):2196–2200, 2003.

[72] K. J. Lee and S. G. Thompson. Flexible parametric models for random-effects distributions. *Statistics in Medicine*, 27(3):418–434, 2008.

[73] X. Li and J. M. Morris. On measuring asymmetry and the reliability of the skewness measure. *Statistics & Probability Letters*, 12(3):267–271, 1991.

[74] C. P. Lombardi, M. Raffaelli, P. Princi, S. Santini, M. Boscherini, C. De Crea, E. Traini, A. M. D'Amore, C. Carrozza, C. Zuppi, et al. Early prediction of postthyroidectomy hypocalcemia by one single ipth measurement. *Surgery*, 136(6):1236–1241, 2004.

[75] T. Lumley, P. Diehr, S. Emerson, and L. Chen. The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, 23(1):151–169, 2002.

[76] H. MacGillivray. Skewness and asymmetry: measures and orderings. *The Annals of Statistics*, 14(3):994–1011, 1986.

[77] V. C. Marinho, J. Higgins, S. Logan, and A. Sheiham. Fluoride toothpastes for preventing dental caries in children and adolescents. *The Cochrane Library*, 2003.

[78] P. Martínez-Camblor. Fully non-parametric receiver operating characteristic curve estimation for random-effects meta-analysis. *Statistical Methods in Medical Research*, 2014. doi: 10.1177/0962280214537047.

[79] D. Mavridis and G. Salanti. A practical introduction to multivariate meta-analysis. *Statistical Methods in Medical Research*, 22(2):133–158, 2013.

[80] C. E. McCulloch, J. M. Neuhaus, et al. Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. *Statistical Science*, 26(3):388–402, 2011.

[81] T. P. McWilliams. A distribution-free test for symmetry based on a runs statistic. *Journal of the American Statistical Association*, 85(412):1130–1133, 1990.

[82] D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to Linear Regression Analysis (4th ed.)*. Wiley & Sons, Hoboken, July 2006.

[83] L. E. Moses, D. Shapiro, and B. Littenberg. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Statistics in Medicine*, 12(14):1293–1316, 1993.

[84] H. Noma. Confidence intervals for a random-effects meta-analysis based on bartlett-type corrections. *Statistics in Medicine*, 30(28):3304–3312, 2011.

[85] J. P. Noordzij, S. L. Lee, V. J. Bernet, R. J. Payne, S. M. Cohen, I. K. McLeod, M. P. Hier, M. J. Black, P. D. Kerr, M. L. Richards, C. Y. Lo, M. Raffaelli, R. Bellantone, C. P. Lombardi, J. I. Cohen, and M. S. Dietrich. Early prediction of hypocalcemia after thyroidectomy using parathyroid hormone: An analysis of pooled individual patient data from nine observational studies. *Journal of the American College of Surgeons*, 205(6): 748–754, 2007.

[86] G. W. Oehlert. A note on the delta method. *The American Statistician*, 46(1):27–29, 1992.

[87] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[88] P. N. Patil, P. Patil, and D. Bagkavos. A measure of asymmetry. *Statistical Papers*, 53 (4):971–985, 2012.

[89] P. N. Patil, D. Bagkavos, and A. T. Wood. A measure of asymmetry based on a new necessary and sufficient condition for symmetry. *Sankhya: The Indian Journal of Statistics, Series A*, 76(1):123–145, 2014.

[90] R. J. Payne, M. P. Hier, M. Tamilia, J. Young, E. MacNamara, and M. J. Black. Postoperative parathyroid hormone level as a predictor of post-thyroidectomy hypocalcemia. *The Journal of Otolaryngology*, 32(6):362–367, 2003.

[91] K. Pearson. Contributions to the mathematical theory of evolution. II. skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London A*, 186: 343–414, 1895.

[92] T. Pigott, R. Williams, and J. Polanin. Combining individual participant and aggregated data in a meta-analysis with correlational studies. *Research Synthesis Methods*, 3(4):257–268, 2012.

[93] H. O. Posten. The robustness of the two-sample t-test over the pearson system. *Journal of Statistical Computation and Simulation*, 6(3-4):295–311, 1978.

[94] H. Putter, M. Fiocco, and T. Stijnen. Meta-analysis of diagnostic test accuracy studies with multiple thresholds using survival methods. *Biometrical Journal*, 52(1):95–110, 2010.

[95] R. H. Randles, M. A. Fligner, G. E. Policello, and D. A. Wolfe. An asymptotically distribution-free test for symmetry versus asymmetry. *Journal of the American Statistical Association*, 75(369):168–172, 1980.

[96] R. D. Reiss. Nonparametric estimation of smooth distribution functions. *Scandinavian Journal of Statistics*, 8(2):116–119, 1981.

[97] J. B. Reitsma, A. S. Glas, A. W. Rutjes, R. J. Scholten, P. M. Bossuyt, and A. H. Zwinderman. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology*, 58(10):982–990, 2005.

[98] J. Rice. Boundary modification for kernel regression. *Communications in Statistics-Theory and Methods*, 13(7):893–900, 1984.

[99] R. D. Riley. Multivariate meta-analysis: the effect of ignoring within-study correlation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(4):789–811, 2009.

[100] R. D. Riley, J. Higgins, and J. J. Deeks. Interpretation of random effects meta-analyses. *BMJ*, 342(7305):549, 2011.

[101] R. D. Riley, M. J. Price, D. Jackson, M. Wardle, F. Gueyffier, J. Wang, J. A. Staessen, and I. R. White. Multivariate meta-analysis using individual participant data. *Research Synthesis Methods*, 2014.

[102] R. D. Riley, Y. Takwoingi, T. Trikalinos, A. Guha, A. Biswas, J. Ensor, R. K. Morris, and J. J. Deeks. Meta-analysis of test accuracy studies with multiple and missing thresholds: A multivariate-normal model. *Journal of Biometrics & Biostatistics*, 5(3):1–12, 2014.

[103] R. D. Riley, I. Ahmed, J. Ensor, Y. Takwoingi, A. Kirkham, R. K. Morris, J. P. Noordzij, and J. J. Deeks. Meta-analysis of test accuracy studies: an exploratory method for investigating the impact of missing thresholds. *Systematic Reviews*, 4(1):1–13, 2015.

[104] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956.

[105] E. Rothman and M. Woodroofe. A Cramér von-Mises type statistic for testing symmetry. *The Annals of Mathematical Statistics*, 43(6):2035–2038, 1972.

[106] D. Ruppert. Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *Journal of the American Statistical Association*, 92(439):1049–1062, 1997.

[107] C. M. Rutter and C. A. Gatsonis. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine*, 20(19):2865–2884, 2001.

[108] H. Scheffé. *The Analysis of Variance*. Wiley, New York, 1959.

[109] S. Senn. Testing for baseline balance in clinical trials. *Statistics in Medicine*, 13(17): 1715–1726, 1994.

[110] R. J. Serfling. Multivariate symmetry and asymmetry. *Encyclopedia of Statistical Sciences*, 2006.

[111] S. J. Sheather and M. C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 53(3):683–690, 1991.

[112] K. Sidik and J. N. Jonkman. A simple confidence interval for meta-analysis. *Statistics in Medicine*, 21(21):3153–3159, 2002.

[113] K. Sidik and J. N. Jonkman. Robust variance estimation for random effects meta-analysis. *Computational Statistics & Data Analysis*, 50(12):3681–3701, 2006.

[114] B. W. Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

[115] E. Slutsky. On the law of large numbers. *Statistics Bulletin*, 7(9):1–55, 1925.

[116] H. E. Soper, A. W. Young, B. M. Cave, A. Lee, and K. Pearson. On the distribution of the correlation coefficient in small samples. appendix II to the papers of 'Student' and R. A. Fisher. *Biometrika*, 11(4):328–413, 1917.

[117] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

[118] T. D. Stanley and S. B. Jarrell. Meta-regression analysis: A quantitative method of literature surveys. *Journal of Economic Surveys*, 3(2):161–170, 1989.

[119] Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908.

[120] D. Tjostheim. Measures of dependence and tests of independence. *Statistics: A Journal of Theoretical and Applied Statistics*, 28(3):249–284, 1996.

[121] T. A. Trikalinos, D. C. Hoaglin, and C. H. Schmid. Empirical and simulation-based comparison of univariate and multivariate meta-analysis for binary outcomes. *Agency for Healthcare Research and Quality (US)*, 2013.

[122] W. R. van Zwet. *Convex transformations of random variables.* Mathematisch Centrum, 1964.

[123] J. Voraprateep. Robustness of Wilcoxon signed-rank test against the assumption of symmetry. MPhil thesis, University of Birmingham, December 2013.

[124] M. P. Wand and M. C. Jones. *Kernel smoothing.* CRC Press, 1994.

[125] J.-G. Wang, J. A. Staessen, S. S. Franklin, R. Fagard, and F. Gueyffier. Systolic and diastolic blood pressure lowering as determinants of cardiovascular outcome. *Hypertension*, 45(5):907–913, 2005.

[126] G. S. Watson and M. R. Leadbetter. Hazard analysis II. *Sankhya: The Indian Journal of Statistics, Series A*, 26(1):101–116, 1964.

[127] A. Whitehead and J. Whitehead. A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine*, 10(11):1665–1677, 1991.

[128] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80–83, 1945.

[129] M. B. Wilk and R. Gnanadesikan. Probability plotting methods for the analysis of data. *Biometrika*, 55(1):1–17, 1968.