Robust Models for Accommodating Outliers in Random-Effects Meta-Analysis: A Simulation and Empirical Study

ROBUST MODELS FOR ACCOMMODATING OUTLIERS IN RANDOM-EFFECTS META-ANALYSIS: A SIMULATION AND EMPIRICAL STUDY

BY

MELANIE STACEY, B.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF MATHEMATICS & STATISTICS

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Master of Science (2015) McMaster University

(Mathematics & Statistics) Hamilton, Ontario, Canada

TITLE:	Robust Models for Accommodating Outliers in Random-Effects Meta-Analysis: A Simulation and Empirical Study
AUTHOR:	Melanie Stacey B.Sc., (Statistics) McMaster University, Canada
SUPERVISOR:	Dr. Joseph Beyene
NUMBER OF PAGES:	xi, 99

Abstract

In traditional meta-analysis, a random-effects model is used to deal with heterogeneity and the random-effect is assumed to be normally distributed. However, this can be problematic in the presence of outliers. One solution involves using a heavy tailed distribution for the random-effect to more adequately model the excess variation due to the outliers. Failure to consider an alternative approach to the standard in the presence of unusual or outlying points can lead to inaccurate inference. A heavy tailed distribution is favoured because it has the ability to down-weight outlying studies appropriately, therefore the removal of a study does not need to be considered.

In this thesis, the performance of the t-distribution and a finite mixture model are assessed as alternatives to the normal distribution through a comprehensive simulation study. The parameters varied are the average mean of the non-outlier studies, the number of studies, the proportion of outliers, the heterogeneity and the outlier shift distance from the average mean. The performance of the distributions is measured using bias, mean squared error, coverage probability, coverage width, Type I error and power. The methods are also compared through an empirical study of

meta-analyses from The Cochrane Library (2008).

The simulation showed that the performance of the alternative distributions is better than the normal distribution for a number of scenarios, particularly for extreme outliers and high heterogeneity. Generally, the mixture model performed quite well.

The empirical study reveals that both alternative distributions are able to reduce the influence of the outlying studies on the overall mean estimate and thus produce more conservative p-values than the normal distribution.

It is recommended that a practitioner consider the use of an alternative random-effects distribution in the presence of outliers because they are more likely to provide robust results.

Acknowledgements

I would like to express my thanks and gratitude to my supervisor Dr. Joseph Beyene who offered valued guidance and knowledge throughout my experience with this thesis. I am thankful I was able to work with Dr. Beyene and his extremely resourceful research group (SIGMA).

I must especially thank Dr. Mateen Shaikh who has been a reliable and patient problem solver, through my countless "re-runs" and questions. His expertise with statistics, high performance computing, as well as his ability to write elegant code has made a big difference in the work I was able to produce.

I would like to mention some of my professors who have inspired me during my undergraduate and graduate degree at McMaster: Dr. Peter Macdonald, Dr. Angelo Canty, Dr. Román Viveros-Aguilera, and Dr. Narayanaswamy Balakrishnan; I sincerely enjoyed your courses and guidance over the years.

Thank you to my committee members Dr. Román Viveros-Aguilera, Dr. Greg Pond and Dr. Joseph Beyene.

Thank you to the SIGMA team and other graduate students for being friendly and always happy to help.

To my friends and family, thank you for not forgetting about me during my time as reclusive graduate student.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Evidence-based medicine requires that the best available evidence is used when making decisions about clinical research and individual patient care. This involves combining clinical expertise with external evidence through systematic reviews (Sackett, 1997). Meta-analysis is a tool which allows for the synthesis of studies which aim to address the same scientific question. In a clinical setting, these studies may aim to measure the effect of one intervention versus another, or versus a control group. Combining the best available and most relevant clinical evidence through a meta-analysis allows a researcher to practice evidence-based medicine.

The first distinction between results within a study and results summarized from multiple studies was made during the 18th and 19th centuries in the fields of mathematics and astronomy. In the 20th century these ideas were revisited in an attempt to summarize the results of various clinical trials. Karl Pearson is credited as the first to synthesize clinical study results in 1904 with a set of data which counted

infections in soldiers who had either volunteered or not volunteered for typhoid inoculation (Simpson and Pearson, 1904). However, Pearson questioned whether the group of soldiers who volunteered was homogeneous to the group of soldiers who did not volunteer. In 1935, Ronald Fisher published a textbook which identified the issue that the effects might vary by year or location, in other words, that the effects might not be identical from study to study (O'Rourke, 2007).

In current meta-analysis the results from various clinical studies can be aggregated to assess the strength of the treatments. However, in the presence of outliers, the estimate of the overall mean can be misinformative due to an inflated estimate of the heterogeneity.

Fioravanti and Yanagi (2005) present a meta-analysis of 10 studies which investigate the effect of CDP-choline as a treatment for memory and behavior in an elderly population. Figure 1.1 illustrates that study 8 (Senin 2003) is a potential outlier. It would be worth investigating if the overall treatment effect estimate of 0.39 is influenced by the observed treatment effect of study 8.

1.1 Background and Motivation

Marinho and Higgins (2003) present 70 studies which investigate the effect of fluoride toothpaste on children for the prevention of dental caries. The effect is the difference between the treatment group and the control group, and the negative values indicate that the fluoride is beneficial. Using the standard method the overall mean estimate is -0.30 with a p-value < 0.0001. The forest plot is shown in Figure 1.2.

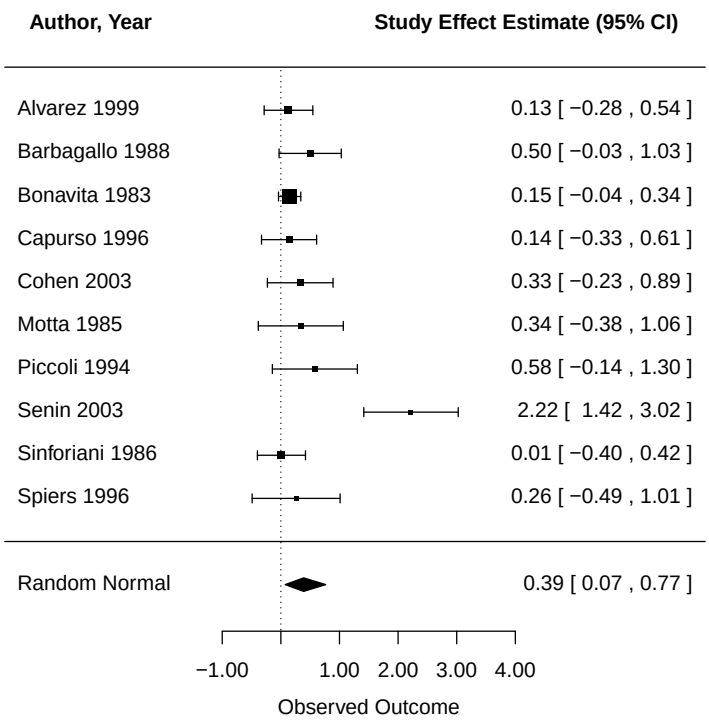

Figure 1.1: CDP-choline data from Fioravanti and Yanagi (2005)

This meta-analysis is atypically large, and even though some outliers are suspected through visual inspection, the overall treatment effect is not in doubt due to the large number of studies included (Gumedze and Jackson, 2011). In this case, it may not be imperative to perform a meta-analysis using a robust random-effects distribution. In fact, using the alternative methods that are discussed in Chapter 2 does not yield a drastically different result as compared to the standard normal approach. Typically, meta-analyses contain a small number of studies and one outlier may substantially influence the mean estimate. Figure 1.3 is a forest plot of 10 studies which explore exercise as a treatment for depression (Lawlor and Hopker, 2001). The standard normal approach yields an overall treatment effect of -1.06 with a p-value < 0.0001. It would be difficult to make a decision about the outliers in this study based only on this forest plot, however, the Mutrie study is the furthest from the mean effect estimate and may have some influence on the estimate. It is also possible that there are no outliers here but instead high heterogeneity. One of the proposed robust distributions will down-weight the outer points in favour of the more precise studies, such as the Martinsen, Singh and Veale studies.

Outlier analysis is commonly discussed in the field of regression and model fitting and there are many accepted standards for outlier detection. Cook and Weisberg (1982) is an excellent resource for the diagnostics of outliers and the assessment of influence. However, outlier analysis is seldom discussed within the context of meta-analysis even though outliers can influence the overall mean estimate in a meta-analysis just as they can in usual regression. Viechtbauer and Cheung (2010) have applied some of the standard regression methods found in Cook and Weisberg (1982)

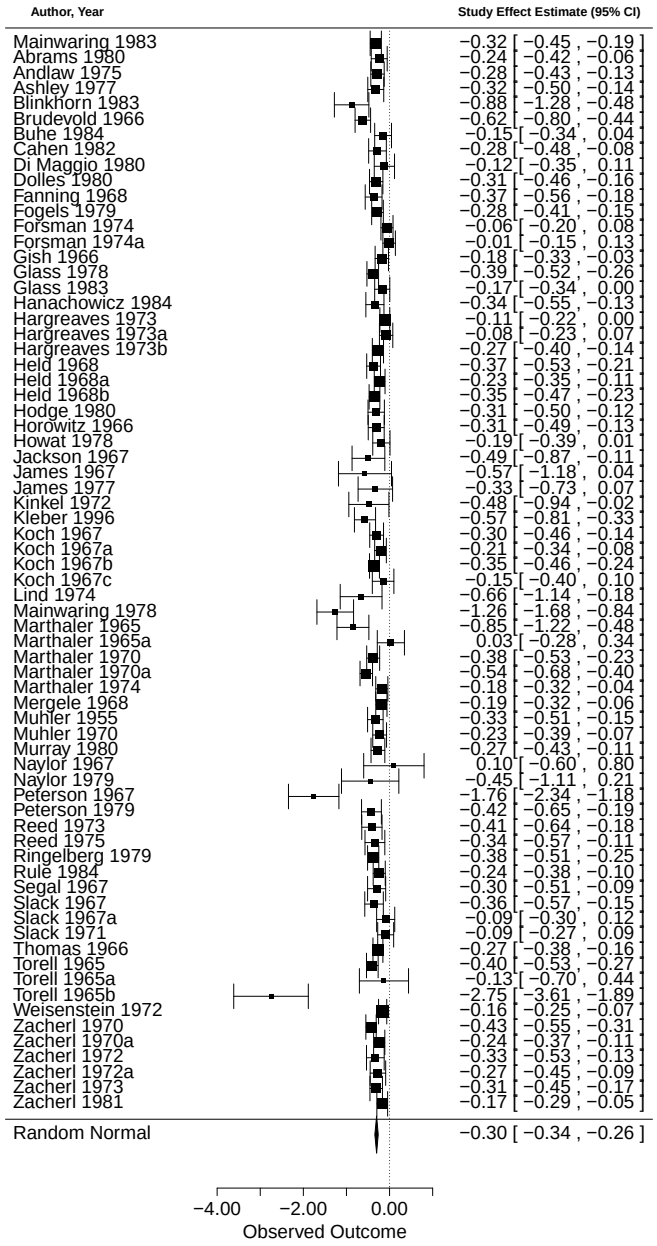

Figure 1.2: Forest plot for Fluoride Data from Marinho and Higgins (2003)

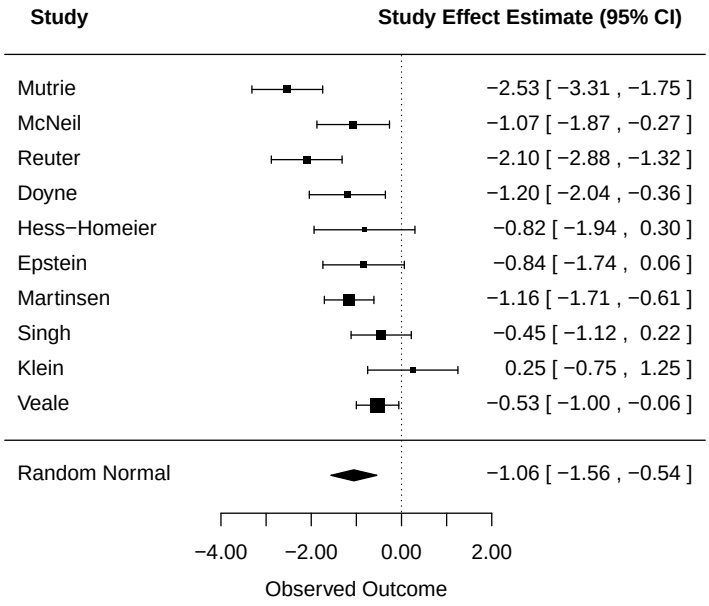

Figure 1.3: Forest plot for Exercise Data from Lawlor and Hopker (2001)

as well as Belsley et al. (1980) to meta-analysis. Some familiar tools such as studentized residuals and Cook's distance are easily applied to meta-analyses and are further discussed in Chapter 2.

When a random-effects model is used, the common assumption is that the distribution of the random-effect is normal. However, Baker and Jackson (2007) suggest that this assumption is not adequate when the data contain unusual values. Since most meta-analyses contain a small number of studies, omitting suspected outliers as though erroneous can be particularly detrimental to analysis since the outlying studies have occurred by natural random chance. It would be preferable to keep the outlying studies but to down weight their influence on the mean estimate. This overcomes the uncertainty and controversial decision to remove the outlier.

To address the problem of outliers in meta-analysis, Baker and Jackson (2007) proposed the use of a heavy-tailed distribution for the random-effect as this method is capable of down-weighting the outlying observations and does not give an inflated estimate of heterogeneity. Using real data applications, the paper compared the standard method's estimates against the estimates for a model using a t-, sinh-, beta- and Subbotin-distributions for the random-effects. The authors conclude that each proposed distribution successfully down-weighted the outliers, leading to a better estimate for the overall treatment mean.

Lee and Thompson (2008) argued that assuming normality for the random-effects is often a restrictive assumption and proposed a flexible model for the random-effects. Using real data applications, the paper compared the estimates from the normal method with the t-distribution, skewed-t and skewed-normal for the random-effects.

Data sets simulated using both a normal distribution and a skewed t-distribution for the random-effects were used to assess the methods when the true random-effects distribution is known. This study focused on the values of the parameter estimates and did not directly measure any performance measures such as bias, coverage, power, etc. Model comparisons were made using the deviance information criterion (DIC). The authors concluded that fitting a skewed distribution to data with a skewed random-effects distribution is more accurate and leads to a better model fit. More generally, using a more flexible distribution is very beneficial when there is suspicion that the normal assumption for the random-effects may be violated (Lee and Thompson, 2008).

Another method which is also capable of identifying potential outliers was proposed by Gumedze and Jackson (2011). The random variance shift outlier model in the paper fits a model for each observation and determines the probability that it is an outlier. This method also down-weights the outliers and produces a more sensible mean estimate. The results are similar to the outcomes achieved by Lee and Thompson (2008), and Baker and Jackson (2007), with the added benefit of identifying the outliers.

Beath (2014) incorporates all of the above methods and verifies them using real data applications with outliers present. Beath (2014) agrees that a heavier tailed distribution adequately down-weights outliers and proposes the use of a finite mixture distribution as an attempt to identify the outliers and model them using a second heterogeneity parameter. This type of mixture model provides heavier tails when outliers are identified and allows for them to be down-weighted during the calculation

of the overall mean effect. This finite mixture model is an extension of the random variance shift outlier model proposed by Gumedze and Jackson (2011). It is found that the finite mixture model provides more robustness over the standard method, and it is noted that the identification of the outlying studies is a useful tool.

While all of this work determines that an alternative random-effects distribution is needed in the presence of outliers, none of the above studies used simulated data with artificially inserted outliers in order to systematically assess the robustness of the methods in comparison to one another.

This thesis project includes a simulation study which varies the proportion of outliers present as well as the magnitude of the outliers, among other parameters. Chapter 2 outlines the methods and assumptions which are necessary to perform a standard meta-analysis as well as perform an analysis using a non-normally distributed random-effect. Chapter 3 describes the design and results of the simulation study. Performance measures such as bias, mean squared error, coverage probability, confidence width, power and Type I error are compared. Chapter 4 demonstrates the effectiveness of the different methods when applied to real data sets. Finally, Chapter 5 summarizes the results of the project and outlines the potential directions for future work.

Chapter 2

Methods

2.1 Standard Meta-Analysis Overview

The purpose of a meta-analysis, in a medical context, is to calculate statistics about a group of studies with greater precision than the individual studies of which the meta-analysis is comprised. For a single study of continuous measurements, such as blood pressure, one might summarize the study using a mean. For a single study which records count data, such as how many patients tested positive for a disease, one might calculate an odds ratio (see Table 2.1 for an example of the discrete data structure using a treatment and control group, where a_i, b_i, c_i, d_i are counts of events in each group). These are known as the "treatment effects" or "effect sizes" (Borenstein et al., 2009).

The effect size represents the impact of a treatment versus a placebo or another treatment. There are different choices of effect sizes, such as the raw mean difference,

	Outcome 1	Outcome 2	Total
Treatment	a_i	b_i	$k_{Ti} = a_i + b_i$
Control	c_i	d_i	$k_{Ci} = c_i + d_i$

Table 2.1: Table of count data for one study (Viechtbauer, 2010)

the standardized mean difference, the odds ratio or the relative risk, etc (see Table 2.2). In this table, y_i is the observed treatment effect, or effect size, for study i where $i = 1, ..., k$ and the meta-analysis is comprised of k studies.

Effect size	Type	Formula	Note
Mean Difference (MD)	Continuous	$y_i = \overline{x}_T - \overline{x}_C$	$\overline{x}_T, \overline{x}_C$ are the means from the treatment and control groups, respectively
Standardized Mean Difference (SMD)	Continuous	$y_i = \frac{\overline{x}_T - \overline{x}_C}{S_{pooled}}$	S_{pooled} pooled standard deviation of the two groups
Odds Ratio (OR)	Discrete	$y_i = a_i d_i / b_i c_i$	a_i, b_i, c_i, d_i come from Table 2.1
Relative Risk (RR)	Discrete	$y_i = (a_i / k_{Ti}) / (c_i / k_{Ci})$	a_i, c_i, k_{Ti}, k_{Ci} come from Table 2.1

Table 2.2: Example of possible effect size calculations

Some choices of effect size may be more appropriate than others. For example, using the log odds ratio would be appropriate to express the doubling and the halving of a value to have the same magnitude. Not all studies use the same effect size, but in order to perform a meta-analysis on a collection of studies it is necessary to have the effect sizes measured the same way. If a study provides the full data set then any effect size can be calculated, but if only the effect size is known it is not always

possible to convert it to another outcome measure. Some conversions are available in Borenstein et al. (2009).

The simulation study in Chapter 3 simulates generic continuous outcomes while the real data analysis in Chapter 4 uses MD, SMD and another measure called a ratio of means proposed by Friedrich et al. (2011).

A key motivation in this project is the heterogeneity of the study effect sizes within a meta-analysis. It is possible that there is no heterogeneity across the true effect sizes for each study; in this case each study would have an identical true effect size. This is called the fixed-effects model. However, due to variations such as patient mixes, geographical location or implementation inconsistencies, there may be dispersion in the true underlying effect sizes from study to study (Borenstein et al., 2010). In this case, the random-effects model would be more appropriate as it is designed to address this concern of heterogeneity.

The goal of a meta-analysis is to calculate an estimate of the overall mean effect. A common approach when summarizing a data set would be to take an average of the data points. In meta-analysis, each study effect size has a different measure of precision which is analogous to the inverse of the within-study-variance called the weight. We would like to give more weight to the studies which present more precise effect size estimates (Borenstein et al., 2010). The weights will be different when considering a fixed-effects model or a random-effects model.

A useful tool in meta-analysis is a forest plot. The forest plot simultaneously displays the observed effect size for each study, the confidence interval and the corresponding within-study-variance. The CDP forest plot (Figure 1.1) contains 10

studies, the effect size is represented by a square, and the value is printed in the same line. The confidence interval is represented by the horizontal line through the square, the upper and lower bound values are printed next to the effect size. Study 3 (Bonavita 1983) has the smallest within-study-variance indicated by a larger square. This study will have the largest weight. The diamond indicates the estimate of the overall mean. The diamond is stretched horizontally to represent the confidence interval of the estimate. Visually, study 8 is suspected to be an outlier, or at the very least it indicates potential heterogeneity in the meta-analysis. It is not always clear whether the meta-analysis follows a fixed-effects model or a random-effects model, Figure 1.1 has been constructed using a random-effects meta-analysis and maximum likelihood estimates in following with other authors who have used these data. This decision is up to the discretion of the analyst, and may be based on the context of the studies. The next sections will describe the differences between the fixed-effects and random-effects models.

2.1.1 Fixed Effects Model

The fixed-effects model is the most basic way of combining the outcome measures. The estimate of the overall mean will be a weighted average of the study estimates, where the weights are determined by the inverse of the study variances. This model assumes that each study shares a common underlying true effect size ($\theta_i = \theta$ for all i). In other words, it is only random sampling error at the subject level that causes the observed effect sizes to differ from study to study. The weights for each study are given by $w_i = 1/v_i$ where v_i is the known within-study-variance. The estimate

of the true mean is

$$\hat{\theta} = \frac{\sum_{i=1}^{k} y_i w_i}{\sum_{i=1}^{k} w_i}$$

where y_i is the effect estimate for study i. Borenstein et al. (2010) suggest that there are two main conditions that should be met before choosing the fixed-effects model. Firstly, if one can make the argument that all of the studies come from a narrowly defined and functionally identical population. Secondly, the estimates are to be used only to describe the particular population, and not to be extrapolated beyond the defined population. These assumptions are not always valid, leading to the discussion of the random-effects model.

2.1.2 Random Effects Model

The most important difference between the fixed-effects model and the random-effects model is that the true underlying effect sizes (θ_i) are stochastic. The calculation of the overall mean estimate for the random-effects model differs from the fixed-effects model in the calculation of the weights. The weights use the estimate of the heterogeneity as well as the within-study-variances since the analysis is designed to recognize two sources of variation. This means that the overall mean estimate can be drastically different than the fixed-effects estimate for studies with high heterogeneity. Borenstein et al. (2010) urges that even if the estimates are identical this does not imply that the methods are interchangeable. As mentioned above, the fixed-effects model is used to describe a narrower population whereas the random-effects model can account for diverse populations.

For any observed value y_i the weight is given by

$$\tilde{w}_i = \frac{1}{v_i + \hat{\tau}^2} \tag{2.1}$$

where $\tilde{w}_i = w_i$ for a fixed-effects model with an absence of heterogeneity ($\hat{\tau}^2 = 0$). $\hat{\tau}^2$ is the estimate of the heterogeneity parameter. This parameter is needed to calculate the weights and the estimate of the overall mean effect. The estimate of the overall mean is given by

$$\hat{\mu} = \frac{\sum_{i=1}^{k} y_i \tilde{w}_i}{\sum_{i=1}^{k} \tilde{w}_i}. \tag{2.2}$$

The observed effect y_i is collectively determined by the true mean, the deviation of the study's true effect size from the true mean and the sampling variance (Borenstein et al., 2010).

To think about this model in terms of statistical distributions we use the same ideas. The observed effect size (y_i) will differ from the study mean (θ_i) by a random amount, determined by the within-study-variance (v_i). The study means (θ_i) will differ from the overall true mean (μ) by a random amount, determined by the between-study-variance (τ^2).

This is viewed as the hierarchical model

$$y_i = \theta_i + \epsilon_i \tag{2.3}$$

$$\theta_i = \mu + \mu_i \tag{2.4}$$

where $\epsilon_i \sim$Normal(0,v_i) and $\mu_i \sim$Normal$(0, \tau^2)$.

The random-effects model is a common choice in many meta-analyses. Even if the fixed-effects model is an appropriate choice, inferences are limited to the specific population. The random-effects model allows for inferences from the meta-analysis to be generalized across more populations. Even though the random-effects model will more accurately describe data which has heterogeneity in the true effects (Borenstein et al., 2009), it has the potential to inaccurately model data which contains outliers.

The heterogeneity parameter τ^2 is often nuisance as we are not directly interested in τ^2 but require it to estimate μ. There are many different methods for estimating τ^2, the most well known is called the "DL" method, named after DerSimonian and Laird (DerSimonian and Laird, 1986). The method is based on a method of moments framework. Two other popular methods for estimating τ^2 are the maximum likelihood (ML) method and the restricted maximum likelihood (REML) method. These are common statistical approaches for determining estimators for parameters in a probability density function and will be used extensively in Chapter 4. Kontopantelis and Reeves (2012) evaluate the performance of some of these methods in a simulation study for non-normally distributed random-effects. The details of the estimation methods will be discussed in the proceeding random-effects distribution sections.

2.2 Outliers in Meta-Analysis

Before exploring how a robust random-effects distribution might be beneficial in dealing with outliers, it is important to understand how to detect an outlier and

how it might affect the result of a meta-analysis. Hedges and Olkin (1985) include a chapter which describes various diagnostic techniques for fixed-effects meta-analyses. Viechtbauer and Cheung (2010) uses the methods of Hedges and Olkin (1985) and adapts them for random-effects meta-analysis. The outlier detection tools discussed are extensions of the standard outlier diagnostics in statistical literature.

Firstly, the internally studentized residual takes the form

$$r_i = \frac{y_i - \hat{\mu}}{\sqrt{\text{Var}[y_i - \hat{\mu}]}} = \frac{y_i - \hat{\mu}}{\sqrt{(1 - h_i)(v_i + \hat{\tau}^2)}}$$

where y_i is the observed effect estimate, $\hat{\mu}$ is the estimate of the overall mean effect and h_i is the i^{th} diagonal entry of the hat matrix, also called the leverage. If y_i were an outlier or an extreme observation then it would have influence over the mean estimate, which is included in the calculation of this residual, therefore it can be useful to use the externally studentized residual, which takes the form

$$r_{(-i)} = \frac{y_i - \hat{\mu}_{(-i)}}{\sqrt{\text{Var}[y_i - \hat{\mu}_{(-i)}]}} = \frac{y_i - \hat{\mu}_{(-i)}}{\sqrt{v_i + \hat{\tau}^2_{(-i)} + \text{Var}[\hat{\mu}_{(-i)}]}}$$

where $\hat{\mu}_{(-i)}$ and $\hat{\tau}^2_{(-i)}$ are the estimate of the overall mean and heterogeneity excluding the i^{th} study, respectively.

There are no universally accepted cut-off points which determine if a particular residual is too extreme, therefore we can examine the residuals in relation to each other. If the studies agree with the model assumption, then the externally studentized residuals will follow a normal distribution, this can be checked using a Q-Q plot (Viechtbauer and Cheung, 2010). Figure 2.1 shows this plot for the CDP-choline

data, study 8 falls outside of the confidence limits.

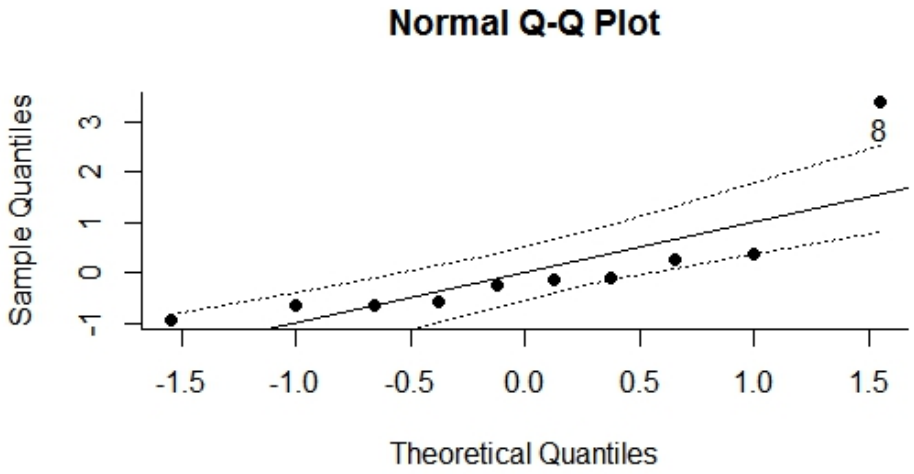

Figure 2.1: Q-Q plot of the externally studentized residuals from CDP data (Fig. 1.1)

The next diagnostic measure is called DFFITS (difference in fit statistics). It is a measure of the difference between the predicted mean effect for study i with and without study i included in the model. DFFITS measures the influence on the results of the meta-analysis by calculating the change in standard deviations for the mean effect after the i^{th} study is excluded (Viechtbauer and Cheung, 2010). The formula is as follows

$$DFFITS_i = \frac{\hat{\mu} - \hat{\mu}_{(-i)}}{\sqrt{h_i(v_i + \hat{\tau}^2_{(-i)})}}.$$

Another useful measure of influence comes from exploring the change in fitted values when the i^{th} study is excluded, this measure is known as Cook's distance and

can be calculated for a random-effect meta-analysis as follows

$$D_i = \sum_{i=1}^{k} \frac{(\hat{\mu} - \hat{\mu}_{(-i)})^2}{v_i + \hat{\tau}^2}.$$

When D_i is larger than $\chi^2_{p+1,0.5}$, the i^{th} study is a potential outlier and should be examined, where p is the number of betas included in a mixed effects model. A D_i value larger than $\chi^2_{p+1,0.5}$ will move the estimate to the $100\times(1-\alpha)^{th}$ confidence boundary which is denoted by $\chi^2_{p+1,1-\alpha}$. This is determined by the idea that D_i can be interpreted as the Mahalanobis distance between the two sets of predicted values when the i^{th} study is included and excluded, respectively (Viechtbauer and Cheung, 2010; Cook and Weisberg, 1982).

The change in the parameter estimates can also be measured as each study is deleted in turn. This measure is called DFBETAS (difference in betas) and is calculated using

$$DFBETAS_i = (\hat{\mu} - \hat{\mu}_{(-i)}) \sqrt{\sum_{l=1}^{k} \tilde{w}_{l(-i)}}$$

where $\tilde{w}_{l(-i)} = 1/(v_l + \hat{\tau}^2_{(-i)})$. For small to medium data sets, a value of $DFBETAS_i$ larger than 1 could indicate an influential point. Viechtbauer and Cheung (2010) suggest that this threshold is still valid for meta-analysis.

A change in the variance-covariance matrix of the overall effect size estimates can be measured using COVRATIO as follows

$$COVRATIO_i = \frac{\text{Var}[\hat{\mu}_{(-i)}]}{\text{Var}[\hat{\mu}]}$$

where $\text{Var}[\hat{\mu}] = 1/\sum_{i=1}^{k} \tilde{w}_i$. Removing study i will yield a more precise effect size estimate when $COVRATIO_i$ is less than 1 (Viechtbauer and Cheung, 2010).

Finally, the change in the estimate of the heterogeneity can be measured with the exclusion of each study, R_i is calculated as a percent change:

$$R_i = 100 \times (\hat{\tau}^2 - \hat{\tau}^2_{(-i)})/\hat{\tau}^2.$$

If a study is influential then its removal will cause a decrease in the estimated heterogeneity and a large positive R_i (Viechtbauer and Cheung, 2010).

All of these methods are available in the *metafor* package using the function **influence()** (Viechtbauer, 2010).

The following is R output of the influence diagnostics for the CDP-choline data from Figure 1.1. Study 8 is identified as influential by an asterisk (*).

	rstudent	dffits	cook.d	cov.r	tau2.del	QE.del	hat	weight	dfb	inf
1	-0.5977	-0.2755	0.0912	1.3316	0.1851	27.3796	0.1178	11.7753	-0.2782	
2	0.2071	0.0058	0.0000	1.3214	0.1868	26.7369	0.1026	10.2597	0.0058	
3	-0.6030	-0.2984	0.1142	1.3926	0.1908	25.9223	0.1445	14.4488	-0.3081	
4	-0.5585	-0.2568	0.0784	1.3224	0.1851	27.4964	0.1104	11.0445	-0.2580	
5	-0.1430	-0.1170	0.0163	1.3354	0.1905	27.6000	0.0988	9.8772	-0.1165	
6	-0.1089	-0.0929	0.0097	1.2697	0.1824	27.6268	0.0795	7.9542	-0.0913	
7	0.3431	0.0566	0.0035	1.2245	0.1736	26.8406	0.0795	7.9542	0.0559	
8	4.8785	1.8244	1.7803	0.2050	0.0000	3.8982	0.0715	7.1506	2.9872	*
9	-0.8932	-0.3594	0.1429	1.2385	0.1668	26.3117	0.1184	11.8417	-0.3615	

```
10  -0.2526 -0.1335 0.0199 1.2586   0.1809 27.6956 0.0769  7.6937 -0.1310
```

These diagnostic tools will help identify studies as warranting further investigation. However, identification of these studies is not the only important task. If the outlier study is not an error then it should be included in the analysis. The next sections describe different random-effects distributions which allow for the inclusion of the outlier or influential point but reduce its influence on the overall mean estimate.

2.3 Random Effects Distributions

The distribution of the random-effect is used to explain the variability in effect sizes when it is believed that the effect sizes do not share a common grand mean. This means that the variability observed in a given meta-analysis is due to random sampling variance (the within-study-variance) and population effect variance (the between-study-variance). The parameter of interest is the overall mean, which can only be estimated by taking into account the heterogeneity that exists across the population effect sizes (Viechtbauer, 2005).

The random-effects hierarchical model is

$$\begin{aligned} y_i &= \theta_i + \epsilon_i \\ \theta_i &= \mu + \mu_i \end{aligned} \tag{2.5}$$

where ϵ_i $\sim$Normal(0,v_i) as before, but now μ_i is the random-effect from some distribution $g(\mu_i|\tau, \phi)$ where τ and ϕ are scale and shape parameters, respectively. Equation 2.6 is the probability density function (pdf) of observing y_i (Baker and

Jackson, 2007).

$$f(y_i|\mu,\tau,\phi) = \frac{1}{\sqrt{2\pi v_i}}\int_{-\infty}^{\infty}\exp\{-(y_i-\mu-\mu_i)^2/(2v_i)\} \times g(\mu_i|\tau,\phi)\,\mathrm{d}\mu_i \tag{2.6}$$

2.3.1 Normal Distribution for Random-Effects

Using similar notation to Baker and Jackson (2007), the pdf of observing y_i is equation 2.6 where $g(\mu_i|\tau,\phi)$ is the pdf of a normal, this gives the following formula

$$f(y_i|\mu,\tau) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{v_i+\tau^2}}\exp\left(-\frac{1}{2}\frac{(y_i-\mu)^2}{v_i+\tau^2}\right). \tag{2.7}$$

Estimating τ^2 using DL gives the estimate

$$\hat{\tau}^2 = \frac{Q-(k-1)}{\sum_{i=1}^k w_i - \frac{\sum_{i=1}^k w_i^2}{\sum_{i=1}^k w_i}} \tag{2.8}$$

where Q is called Cochran's Q-statistic and follows a χ^2_{k-1} distribution under the fixed-effects null hypothesis. The Q-statistic is

$$Q = \sum_{i=1}^k w_i(y_i-\hat{\mu}_F)^2 = \sum_{i=1}^k w_i y_i^2 - \frac{\left(\sum_{i=1}^k w_i y_i\right)^2}{\sum_{i=1}^k w_i} \tag{2.9}$$

where $\hat{\mu}_F$ is the estimated true effect under the fixed effects model. Q is the observed weighted sums of squares and $(k-1)$ is the expected weighted sums of squares under the fixed-effects model, therefore $Q-(k-1)$ is the excess variation. If the fixed-effects model assumptions do not hold then there may be excess variation which can

be attributed to heterogeneity in the true effect sizes (Borenstein et al., 2009).

Another way to estimate the heterogeneity is using the maximum likelihood (ML). This approach gives estimates $\hat{\mu}$ and $\hat{\tau}^2$ which are the values that maximizes the following log-likelihood

$$\log(L(\mu, \tau^2)) = -\frac{1}{2}\left[\sum_{i=1}^{k} \log(2\pi(v_i + \tau^2)) + \sum_{i=1}^{k} \frac{(y_i - \mu)^2}{v_i + \tau^2}\right] \tag{2.10}$$

where $\hat{\mu} \in R$ and $\tau^2 \geq 0$. This is done by taking partial derivatives with respect to μ and τ^2 and setting them equal to zero and solving for $\hat{\mu}$ and $\hat{\tau}^2$. The resulting estimate for τ^2 is

$$\hat{\tau}^2 = \frac{\sum_{i=1}^{k} \tilde{w}_i^2 \left[(y_i - \hat{\mu})^2 - v_i\right]}{\sum_{i=1}^{k} \tilde{w}_i^2}. \tag{2.11}$$

The equation for $\hat{\mu}$ (2.2) contains $\tilde{w}_i$ which must be calculated using $\hat{\tau}^2$ therefore a closed form solution does not exist for both $\hat{\mu}$ and $\hat{\tau}^2$ simultaneously. The estimates are typically computed numerically in an iterative procedure with some starting value for one of the parameters.

The third method of estimation that will be used in this thesis is the restricted maximum likelihood (REML), the log-likelihood for which includes a penalty term and it is maximized for $\hat{\tau}^2$, the log-likelihood is

$$\log(\tilde{L}(\mu, \tau^2)) = \log(L(\mu, \tau^2)) - \frac{1}{2}\log \sum_{i=1}^{k} \frac{1}{v_i + \tau^2} \tag{2.12}$$

again where $\hat{\mu} \in R$ and $\tau^2 \geq 0$. Estimating the parameters here is done in a similar iterative manner as in the ML method.

Recently, abiding by this generally accepted assumption of normally distributed random-effects has received some criticism. Lee and Thompson (2008) outline some of the concerns around this assumption, including the argument that the normal distribution is not appropriate for any situation where there are departures from normality. Specifically, the normal assumption can lead to poor inference when there are outliers involved as it will tend to give an inflated variance estimate to account for observations which lay further out in the tails.

2.3.2 t-Distribution for Random-Effects

The t-distribution has been proposed as natural choice for a robust random-effects distribution because it has heavier tails than the normal distribution (Baker and Jackson, 2007; Lee and Thompson, 2008). Figure 2.2 is a generic example of how the t-distribution can have heavier tails than the normal distribution. However, with higher degrees of freedom, the t- converges toward the normal.

The usual assumption of a normally distributed random-effect comes from the desire for simplicity and asymptotic approximation by the Central Limit Theorem (CLT). However, the approximations due to the CLT may be poor when applied to the random-effect since meta-analyses typically have a low number of studies (Baker and Jackson, 2007).

Referring again to the notation in Baker and Jackson (2007), the pdf of observing y_i is equation 2.6 where

$$g(\mu_i|\tau,\nu) = \frac{\Gamma((\nu+1)/2)}{\tau\sqrt{\pi\nu}\Gamma(\nu/2)}(1+\mu_i^2/(\nu\tau^2))^{-((\nu+1)/2)} \tag{2.13}$$

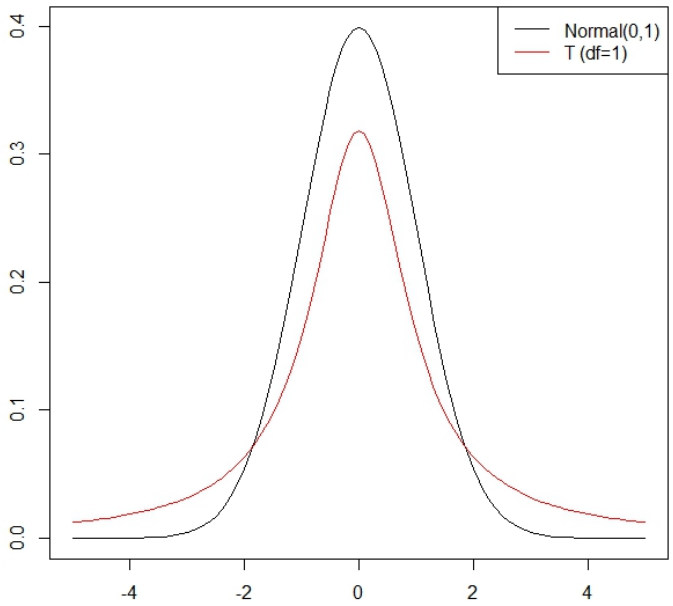

Figure 2.2: Normal Distribution vs. t-Distribution

and the shape parameter ϕ is the inverse of the degrees of freedom ν. The pdf of $g(\mu_i|\tau,\nu)$ is formulated by taking the product of τ and a random variable with a t-distribution (Baker and Jackson, 2007). Both Baker and Jackson (2007) and Lee and Thompson (2008) are referenced by Beath (2015) for the methods used in the *metaplus* package for the calculations using the t-distribution.

$$\begin{aligned}\log(L(\mu,\tau^2,\nu)) = n\log\left(\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\pi\tau\sqrt{2\nu}\Gamma\left(\frac{\nu}{2}\right)}\right) + \sum_{i=1}^{k}\log\Bigg(\frac{1}{\sqrt{v_i}} \\ \times\int_{-\infty}^{\infty}\exp\left(-\frac{(y_i-\mu-\mu_i)^2}{2v_i}\right)\left(1+\frac{\mu_i^2}{\nu\tau^2}\right)^{-\frac{\nu+1}{2}}\mathrm{d}\mu_i\Bigg)\end{aligned} \tag{2.14}$$

The likelihood (equation 2.14) contains an integral over μ_i which is intractable and requires more computationally intensive methods to estimate μ. Baker and Jackson (2007) used a Fortran program as well as NAG library routines to complete the estimation. Lee and Thompson (2008) used WinBUGS to carry out Bayesian MCMC methods for the estimation process. Beath (2015) uses functions in R to compute the numerical integrals as well as a quasi-Newton method to find the $\hat{\mu}, \hat{\tau}^2$ and $\hat{\nu}$ which maximizes the likelihood.

Referring back to the CDP data in Figure 1.1, using the t-distribution for the random-effect yields an overall mean estimate of 0.195 with a 95% confidence interval of [0.053,0.361]. This new estimate is much smaller than the previous estimate of 0.39, thus it is clear that study 8 was too far out in the tails for the normal distribution to handle properly. Using the t-distribution yields a new estimate of -0.27, a smaller confidence interval of [-0.313,-0.247] and a p-value$<$ 0.0001 for the fluoride data in Figure 1.2. This change is not drastic due to the large number of studies, as expected. The exercise data in Figure 1.3 produces almost identical estimates indicating that there is no benefit of using the t-distribution in this case.

2.3.3 Finite Mixture Model for Random-Effects

The finite mixture model attempts to identify outlying studies and model them using a larger variance for the random-effects distribution. The two-component model is a mixture distribution of a non-outlier distribution and an outlier distribution. Similar to the t-distribution, the finite mixture model includes the outliers in the calculation of the overall mean estimate but gives them less weight (Beath, 2014).

The distribution of this mixture model will be heavier tailed, conforming to the general approach suggested by Lee and Thompson (2008).

The detection of the outliers is based on a method proposed by Gumedze and Jackson (2011) which uses a random variance shift outlier model (RVSOM) to fit a mixture model with the current observation having a different random-effect variance than the rest of the observations. Each identified outlier exists in its own group of the mixture model. This approach is favoured over the method of removing potential outlying observations from a meta-analysis. However, Gumedze and Jackson (2011) do not suggest that this method should replace the standard random-effects methodologies for meta-analyses; it should be used as a supplementary tool. More details on this method can be found in Appendix A.

The methods in Beath (2014) differ from the approach developed by Gumedze and Jackson (2011) in that all identified outliers are modeled using a single distribution with a larger random-effects variance (ie. there are only two-components in the model), whereas Gumedze and Jackson (2011) estimates a different variance for each outlying study using a separate model for each outlying study. Also, Beath (2014) allows for any number of outliers, instead of Gumedze and Jackson (2011) allowing only up to three due to the method of using third order statistics.

The mixture model proposed by Beath (2014) assumes that each study belongs to one of two groups, both of which have different random-effects variance. The hierarchical model is

$$\begin{aligned} y_{i|m} &= \theta_{i|m} + \epsilon_i \\ \theta_{i|m} &= \mu + \mu_{i|m} \end{aligned} \tag{2.15}$$

where ϵ_i is the same random within-study-variance as the standard model and $\mu_{i|m}$ is the random-effect for group m with corresponding random-effect variances τ_m^2 for $m = \{1, 2\}$. Now the pdf of observing y_i is the following mixture distribution

$$\begin{aligned} f(y_i|\mu, \tau_1, \tau_2) &= \sum_{m=1}^{2} \pi_m f_m(y_i|\mu, \tau_m) \\ &= \sum_{m=1}^{2} \pi_m \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{v_i + \tau_m^2}} \exp\left(-\frac{1}{2}\frac{(y_i - \mu)^2}{v_i + \tau_m^2}\right) \end{aligned} \tag{2.16}$$

where $\pi_1 + \pi_2 = 1$ and $\pi_i \geq 0$. This is a weighted sum of the pdfs for each group m where each is a standard pdf as in 2.7 and is weighted by the proportion of studies in each group (π_1 and π_2).

Beath (2014) uses an expectation-maximization algorithm to determine the posterior probabilities of a study being an outlier, the mixture model membership probabilities, as well as the estimates for μ, τ^2 and τ_{out}^2. Some details for this method are included in Appendix A. As a faster alternative, in the *metaplus* package, Beath (2015) uses a greedy search and score algorithm (also known as a best-first search and score algorithm) to determine the mixture model memberships. This method starts by assuming there are no outliers (that is, the mixture distribution has one component) then models all possible mixture models with one single outlier, using each study in turn as a potential outlier. The process chooses the model with the highest likelihood and the corresponding point is determined to be the first outlier. Next, all remaining studies are added to the single outlier model one by one to create possible two-outlier models and the one with the highest likelihood is kept. This is repeated until the addition of more studies to the outlier model does not increase

the likelihood (Beath, 2015). Model parameters are estimated using a quasi-Newton method. The estimates $\hat{\mu}$, $\hat{\tau}_1^2$ and $\hat{\tau}_2^2$ are the values which are found to maximize the log-likelihood equation 2.17.

$$\log(L(\mu, \tau_1, \tau_2)) = \sum_{i=1}^{m} \log \sum_{m=1}^{2} \pi_m \frac{1}{\sqrt{(2\pi)(v_i + \tau_m^2)}} \exp\left(-\frac{(y_i - \mu)^2}{2(v_i + \tau_m^2)}\right) \tag{2.17}$$

Using the mixture model for the random-effects of the CDP meta-analysis (Figure 1.1) yields a new estimate of 0.191 with a 95% confidence interval of [0.056,0.348]. Both the t-distribution and the mixture model were able to down weight the effects of study 8 and produce a smaller overall mean estimate of the data. Under the mixture model, the fluoride data in Figure 1.2 has an overall treatment effect estimate of -0.28 and a slightly shifted confidence interval of [-0.315,-0.248]. The associated p-value is < 0.0001. The results have accounted for the outlier but the overall conclusion is the same. The exercise data from Figure 1.3 gives almost identical results under the mixture model and the normal distribution.

2.3.4 Confidence Intervals

For a fixed-effects meta-analysis with heterogeneity present, the standard method of computing confidence intervals using the standard error of the estimate will result in an overestimated (narrower) interval, and thus, low coverage. The DerSimonian and Laird method for random-effects was proposed to fix this by incorporating the estimate for the random-effects variance. However, this method has weaknesses

the confidence interval for the treatment effect does not account for the fact that the heterogeneity is an estimate using observed data (Hardy and Thompson, 1996). REML is often used to improve on this issue. However, the profile likelihood method proposed by Hardy and Thompson (1996) can be used for all of the distributions discussed in Beath (2014), therefore it is the favoured choice for computing the confidence intervals. It is ultimately concluded that the confidence intervals produced are wider than those produced by the standard method, and that the proposed method is preferred to the previously standard approaches (Jackson et al., 2010; Hardy and Thompson, 1996).

Beath (2014) calculates the profile likelihood confidence intervals by using a grid search approach and a step-halving method. The corresponding p-values are computed using a likelihood ratio test. The methods of Beath (2014) have been emphasized throughout this chapter because the R package *metaplus* (Beath, 2015) implements these specific methods and is used in the following chapter.

Chapter 3

Simulation Study

The main goal of the following simulation study is to investigate the effect of outliers on various optimality measures and hypothesis test measures when the random-effects distribution is heavier-tailed and provides different estimates than the standard method. Based on the literature, it was expected that a heavy-tailed distribution would outperform the normal distribution. This simulation study explores differences in three models when 5 input parameters are manipulated.

3.1 Design

There are 5 parameters with a total of 216 scenarios. Each scenario was replicated 1,000 times for a total of 216,000 meta-analyses. Table 3.1 displays the parameter values and descriptions. Our algorithm follows applicable aspects of previously published simulations.

Parameter	Description	Values
μ	True overall μ	$0, 0.5, 1$
k	Number of studies in each meta-analysis	$10, 20, 30, 40$
p	Probability each study being an outlier	$0.1, 0.2$
c	Multiple of non-outlier study standard deviation which is added to μ for the distribution of the outliers	$2.5, 4, 6$
τ^2	Heterogeneity parameter	$0.01, 0.3, 0.5$

Table 3.1: Simulation Parameters Descriptions and Values

The following is a systematic account of how the data were simulated with discussion on the approach following the algorithm:

1. Choose values of μ, τ^2, p, k and c.

2. Generate k observations from a χ^2_1 distribution and divide them by 4. These are the v_i values. If $v_i \in (0.009, 0.6)$ then keep v_i, if not, redraw for a new v_i (Brockwell and Gordon, 2001; Kontopantelis and Reeves, 2012).

3. Generate a single observation from a Binomial(k,p) distribution to obtain the number of studies (k_{out})which will be outliers .

4. Generate $k - k_{out}$ observations from a Normal(μ,τ^2) distribution. These are the non-outlier θ_i values.

5. Generate $k - k_{out}$ observations from a Normal(0,v_i) distribution, for each of the first $k - k_{out}$ v_i's. These are the random study errors ϵ_i for the non-outlier studies.

6. Compute $y_i = \theta_i + \epsilon_i$ for the $k - k_{out}$ non-outlier studies.

7. Generate k_{out} observations from a Normal($\mu + c \times \text{SD}(y_i)$,$\tau^2$) where SD($y_i$) is the sample standard deviation of the $k - k_{out}$ non-outlier y_i values. These are the $\theta_{i_{out}}$ values for the outlier studies. (This method of shifting the mean and utilizing the sample standard deviation is similar to some methods incorporated by Filzmoser (2005), Knight and Wang (2009) and Hardin and Rocke (2004), this is discussed further in this section.)

8. Generate k_{out} observations from a Normal(0,v_i) distribution, for each of the last k_{out} v_i's. These are the random study errors $\epsilon_{i_{out}}$ for the outlier studies.

9. Compute the $y_{i_{out}} = \theta_{i_{out}} + \epsilon_{i_{out}}$ for the k_{out} outlier studies.

10. Together, the $k - k_{out}$ y_i and k_{out} $y_{i_{out}}$ make up the sample of k studies from one meta-analysis with corresponding within-study-variances. v_i. Repeat Steps 2-9 1,000 times to have 1,000 simulated samples under the selected scenario.

11. Return to Step 1 and select the next scenario parameter values.

The method for inserting outlying observations is a combination of some methods found in the literature. Filzmoser (2005) generates "shift outliers" in a multivariate setting, where the "clean data" are realizations of $N_p(\mathbf{0}, I)$ and the n_{out} outlying studies are generated away from the clean data using $N_p(\eta \cdot \mathbf{1}, I)$ where $\eta = \{1.5, 3\}$ and $n_{out}/n \in [0.05, 0.45]$. This is a relatively small shift distance compared to other studies. The values $c = \{2.5, 4, 6\}$ were selected through trial simulations as well as real data sets to ensure that the generated outliers were far enough away from the rest of the data. Rousseeuw and Driessen (1999) uses a similar "shift" method

for outlier generation. Knight and Wang (2009) insert either 0, 1 or 2 outliers into the simulated data. The sizes of the outliers were randomly generated to be either between three and six standard deviations or six and nine standard deviations from the simulated data.

Rather than increasing the variance of the outlier distribution, which will produce some points close to the mean of the non-outlier distribution (Hardin and Rocke, 2004), shifting the mean of the outlier distribution encourages the points to fall in the tails (Filzmoser, 2005). Also, by utilizing the standard deviation of the "clean data" similar to Knight and Wang (2009) we can ensure to some degree that the outlying points will not severely overlap with the rest of the points. For example, if we take study 8 of the CDP data (Figure 1.1) to be an outlier, then the standard deviation of the non-outlier treatment effects is 0.186. A random-effects meta-analysis of the data with study 8 removed gives a treatment effect estimate of 0.189. Thus, study 8 is more than 10 standard deviations away from the estimated mean of the non-outlier data. More examples of similar approaches to outlier generation can be found in Peña and Prieto (2001) and Yong et al. (2008).

An example of simulated data with varying values of c can be found in Figure 3.1. The first panel displays the two mild outliers. These outliers are not always prominent and may be masked by they other data points. The second panel displays the two moderate outliers, which are more visibly detectable, but other randomly generated "clean" points might still be close in value to the outlier points due to a higher parameter of heterogeneity. The third panel displays the two extreme outliers. These are more likely to be detected visually.

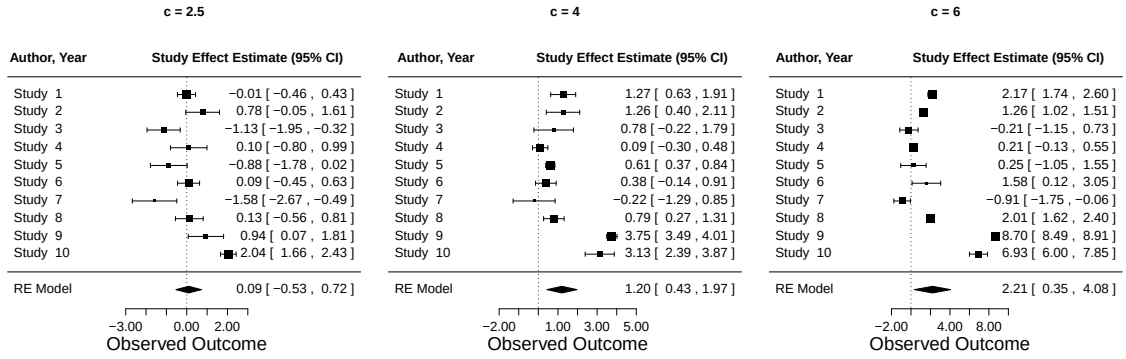

Figure 3.1: Forest plots of 3 sample simulated data sets with different values of c

3.2 Computational Methods

The R package *metaplus* (Beath, 2015) is used extensively, methods for which are outlined in Beath (2014).

Using a sample of y_i and $\sqrt{v_i}$ values, the function **metaplus()** can perform a meta-analysis using either a normal distribution or a t-distribution for the random-effect, or can model the studies using the finite mixture model described in Chapter 2.

For the normally distributed random-effects, the **metaplus()** function is actually using the *metafor* package (Viechtbauer, 2010) with the ML method to produce estimates of the overall mean (μ), the heterogeneity parameter (τ^2) and a 95% confidence interval for μ.

When the random-effects variance is specified to have a t-distribution the **metaplus()** function produces estimates for the overall mean (μ), the heterogeneity parameter (τ^2) and degrees of freedom (ν) using ML, a 95% confidence interval for μ.

For the finite mixture model, the **metaplus()** function produces estimates for the overall mean (μ) and the heterogeneity parameter for the non-outliers and outliers (τ^2 and τ^2_{out}, respectively) using ML, a 95% confidence interval for μ, and the probability that each study is an outlier. These probabilities are not explored in this thesis project.

For consistency, the profile log likelihood (Hardy and Thompson, 1996) is used to compute the confidence intervals for each random-effects model (Beath, 2015). This method takes into account the iterative process of the ML method that results in estimating both parameters simultaneously. The profile likelihood confidence intervals produced are allowed to be asymmetric. Each method uses likelihood ratio tests to compute the corresponding p-values for the mean estimates. All computations and simulations were done using R (R Core Team, 2015).

3.3 Performance Measures

The performance measures used were bias, mean squared error (MSE), coverage probability, confidence width, power and Type I error.

The definition of bias of the estimator $\hat{\theta}$ for the parameter θ is

$$Bias_{\hat{\theta}} = \mathrm{E}[\hat{\theta} - \theta].$$

Thus, we estimate bias as the average

$$bias = \frac{1}{1000} \sum_{j=1}^{1000} \hat{\mu}_j - \mu$$

where $\hat{\mu}_j$ is the estimate of the overall mean for the j^{th} meta-analysis and μ is the true mean.

The definition of mean squared error of the estimator $\hat{\theta}$ for the parameter θ is

$$MSE_{\hat{\theta}} = \mathrm{E}[(\hat{\theta} - \theta)^2].$$

Similarly, the mean squared error was calculated as the average

$$mse = \frac{1}{1000} \sum_{j=1}^{1000} (\hat{\mu}_j - \mu)^2.$$

In the frequentist framework, the confidence intervals are random and the true parameter value is fixed but unknown. The coverage probability is defined as the probability that the random interval contains the true μ or $P(L(X) < \mu < U(X))$ where $L(X)$ and $U(X)$ are the random variables for the lower and upper confidence bounds, respectively, and X is a random sample from the probability distribution determined by μ. Calculating the coverage probability analytically is only possible if the distribution of the random intervals is known. In practice, the distribution is unknown. Therefore, the coverage probability is estimated using simulation by calculating the proportion of simulated intervals which contain the true μ (Brockwell and Gordon, 2001). It is desirable to have the coverage probability closely match the level of confidence assigned to the intervals (Brockwell and Gordon, 2001). The coverage probability is

$$coverage = \frac{1}{1000}\sum_{j=1}^{1000} I_{\mu \in [L_j, U_j]}$$

where L_j and U_j are the lower and upper confidence limits for the j^{th} meta-analysis, respectively, and I_A is the indicator function which equals 1 for the subset A.

The confidence width is simply a measurement of the distance between the upper confidence limit and the lower confidence limit. The average was taken of the 1,000 confidence widths as follows

$$width = \frac{1}{1000}\sum_{j=1}^{1000}(U_j - L_j).$$

Power is defined as the probability of rejecting the null hypothesis when it is false. The p-value given by the **metaplus()** function tests the null hypothesis that the true mean is equal to zero ($H_0 : \mu = 0$) thus, the power can only be calculated for scenarios which have a true mean μ not equal to zero (ie. for which the null hypothesis is actually false). The power for each scenario was calculated to be the proportion of the 1,000 p-values which are smaller than the significance level. Here, the significance level was chosen to be 0.05. Equation 3.1 is used for scenarios with $\mu \neq 0$ only.

$$power = \frac{1}{1000}\sum_{j=1}^{1000} I_{p_j \leq 0.05} \tag{3.1}$$

where p_j is the p-value for the j^{th} meta-analysis.

Type I error is defined as the probability of rejecting the null hypothesis when it

is true. In order to measure this, only the scenarios with $\mu = 0$ can be used since this means the null hypothesis is actually true. Therefore, equation 3.1 can still be used but for scenarios with $\mu = 0$.

3.4 Results

The performance measure results for $\mu = 0.5$ and $\tau^2 = 0.5$ can be seen in Table 3.2. The main trend that can be seen in this table is that as c increases, the measures get worse for the normal and t-distribution, but often improve with c for the mixture model. The hypothesis testing measures can be seen in Table 3.3 for $\tau^2 = 0.5$. More trends are also explored using graphical methods.

Number of	Proportion of	Outlier	Bias			MSE			Cov. Prob.			Conf. Width		
Studies (k)	Outliers (p)	Shift (c)	Norm	t	Mix	Norm	t	Mix	Norm	t	Mix	Norm	t	Mix
10	0.1	2.50	0.19	0.16	0.14	0.13	0.13	0.14	0.91	0.90	0.87	1.25	1.22	1.15
		4.00	0.32	0.22	0.14	0.29	0.24	0.20	0.90	0.90	0.88	1.51	1.41	1.28
		6.00	0.45	0.15	0.05	0.49	0.24	0.14	0.93	0.89	0.90	1.91	1.41	1.22
	0.2	2.50	0.39	0.36	0.31	0.29	0.28	0.26	0.84	0.84	0.82	1.43	1.41	1.34
		4.00	0.63	0.51	0.38	0.68	0.62	0.53	0.80	0.80	0.81	1.87	1.79	1.66
		6.00	0.98	0.67	0.38	1.55	1.28	1.10	0.80	0.78	0.87	2.59	2.18	1.88
20	0.1	2.50	0.19	0.15	0.13	0.09	0.08	0.08	0.89	0.89	0.87	0.89	0.88	0.84
		4.00	0.32	0.15	0.10	0.19	0.11	0.09	0.85	0.89	0.89	1.11	0.99	0.89
		6.00	0.48	0.11	0.03	0.39	0.12	0.06	0.83	0.84	0.91	1.43	0.86	0.80
	0.2	2.50	0.41	0.38	0.32	0.24	0.23	0.21	0.68	0.70	0.71	1.02	1.02	0.99
		4.00	0.65	0.51	0.34	0.55	0.46	0.32	0.55	0.64	0.78	1.35	1.31	1.23
		6.00	0.99	0.61	0.17	1.23	0.87	0.22	0.44	0.59	0.87	1.85	1.45	1.13
30	0.1	2.50	0.20	0.16	0.14	0.08	0.07	0.06	0.83	0.85	0.83	0.72	0.72	0.69
		4.00	0.31	0.13	0.09	0.15	0.07	0.06	0.78	0.85	0.88	0.90	0.77	0.72
		6.00	0.48	0.11	0.02	0.33	0.10	0.03	0.73	0.72	0.93	1.20	0.62	0.65
	0.2	2.50	0.40	0.37	0.32	0.20	0.19	0.17	0.55	0.59	0.64	0.82	0.83	0.81
		4.00	0.64	0.49	0.31	0.49	0.40	0.24	0.38	0.55	0.73	1.10	1.09	1.01
		6.00	0.96	0.52	0.11	1.09	0.67	0.12	0.26	0.48	0.90	1.52	1.08	0.84
40	0.1	2.50	0.20	0.16	0.14	0.07	0.05	0.05	0.78	0.83	0.82	0.63	0.62	0.60
		4.00	0.32	0.13	0.08	0.14	0.06	0.04	0.67	0.83	0.87	0.78	0.65	0.62
		6.00	0.50	0.12	0.03	0.32	0.09	0.02	0.58	0.65	0.93	1.03	0.48	0.56
	0.2	2.50	0.40	0.38	0.33	0.20	0.18	0.16	0.40	0.46	0.53	0.71	0.72	0.72
		4.00	0.64	0.50	0.29	0.49	0.39	0.21	0.21	0.47	0.71	0.94	0.96	0.88
		6.00	1.00	0.60	0.10	1.14	0.77	0.08	0.12	0.37	0.89	1.32	0.91	0.70

Table 3.2: Table of Optimality Measures for $\mu = 0.5$ and $\tau^2 = 0.5$

Number of Studies (k)	Proportion of Outliers (p)	Outlier Shift (c)	Power for $\mu = 0.5$			Power for $\mu = 1$			Type I Error			Average p-value for $\mu = 0.5$		
			Norm	t	Mix	Norm	t	Mix	Norm	t	Mix	Norm	t	Mix
10	0.1	2.50	0.60	0.61	0.61	0.96	0.96	0.94	0.10	0.10	0.13	0.10	0.11	0.11
		4.00	0.60	0.60	0.56	0.96	0.96	0.94	0.10	0.11	0.12	0.09	0.11	0.12
		6.00	0.49	0.55	0.50	0.93	0.96	0.95	0.08	0.15	0.12	0.10	0.13	0.15
	0.2	2.50	0.70	0.69	0.67	0.98	0.98	0.97	0.17	0.17	0.18	0.06	0.07	0.09
		4.00	0.69	0.68	0.62	0.98	0.97	0.94	0.20	0.20	0.17	0.06	0.09	0.11
		6.00	0.64	0.64	0.52	0.94	0.95	0.90	0.21	0.27	0.15	0.06	0.10	0.15
20	0.1	2.50	0.87	0.84	0.83	1.00	1.00	1.00	0.14	0.13	0.14	0.02	0.03	0.04
		4.00	0.88	0.81	0.78	1.00	1.00	1.00	0.15	0.13	0.13	0.02	0.04	0.05
		6.00	0.84	0.81	0.75	1.00	1.00	1.00	0.17	0.23	0.10	0.03	0.05	0.07
	0.2	2.50	0.96	0.94	0.89	1.00	1.00	0.99	0.32	0.31	0.30	0.01	0.02	0.03
		4.00	0.96	0.91	0.83	1.00	1.00	1.00	0.43	0.33	0.21	0.01	0.02	0.04
		6.00	0.97	0.89	0.75	1.00	1.00	0.99	0.49	0.41	0.10	0.01	0.03	0.06
30	0.1	2.50	0.98	0.96	0.95	1.00	1.00	1.00	0.16	0.14	0.14	0.01	0.01	0.01
		4.00	0.98	0.94	0.91	1.00	1.00	1.00	0.24	0.17	0.11	0.01	0.01	0.02
		6.00	0.97	0.92	0.90	1.00	1.00	1.00	0.31	0.31	0.06	0.01	0.02	0.02
	0.2	2.50	1.00	0.99	0.98	1.00	1.00	1.00	0.50	0.45	0.39	0.00	0.00	0.01
		4.00	1.00	0.96	0.92	1.00	1.00	1.00	0.64	0.47	0.24	0.00	0.01	0.02
		6.00	1.00	0.95	0.87	1.00	1.00	1.00	0.77	0.55	0.12	0.00	0.01	0.03
40	0.1	2.50	0.99	0.99	0.98	1.00	1.00	1.00	0.24	0.20	0.19	0.00	0.00	0.00
		4.00	1.00	0.98	0.97	1.00	1.00	1.00	0.34	0.20	0.12	0.00	0.00	0.01
		6.00	0.99	0.97	0.96	1.00	1.00	1.00	0.44	0.43	0.06	0.00	0.01	0.01
	0.2	2.50	1.00	1.00	1.00	1.00	1.00	1.00	0.61	0.56	0.47	0.00	0.00	0.00
		4.00	1.00	0.99	0.97	1.00	1.00	1.00	0.80	0.55	0.30	0.00	0.00	0.01
		6.00	1.00	0.98	0.94	1.00	1.00	1.00	0.88	0.65	0.10	0.00	0.00	0.01

Table 3.3: Table of Hypothesis Measures for $\tau^2 = 0.5$

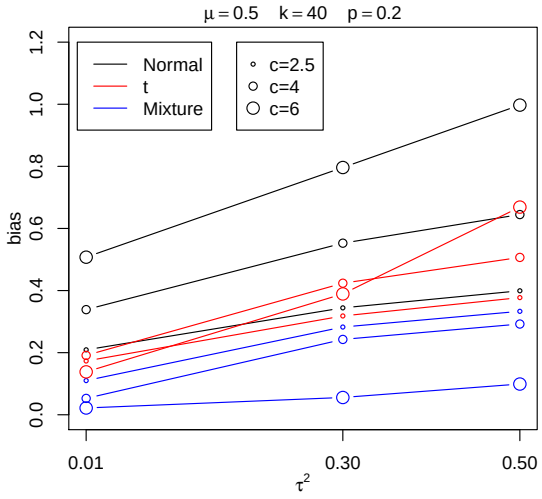

Figure 3.2: Bias for simulation scenario $\mu = 0.5$, $k = 40$ and $p = 0.2$

The normally distributed random-effects model proved to have the most weaknesses in the presence of outliers. Figure 3.2 shows that as heterogeneity increases, bias is highest for the normal distribution with the largest outlier shift. Note that the bias is always positive because the outliers are generated to be on the positive side of the true mean. Figure 3.3 illustrates that the same weaknesses are true for the MSE. In general, all of the performance measures for the normal distribution are consistently worse when c increases which illustrates that it is ill-equipped to handle observations which lay much further out in the tails. This is because this method inflates the estimate of τ^2 to account for the distant points, which leads to an inaccurate estimate of μ.

When the t-distribution is used for the random-effect the bias and MSE are improved over the normal for the majority of the scenarios. For extreme outliers the

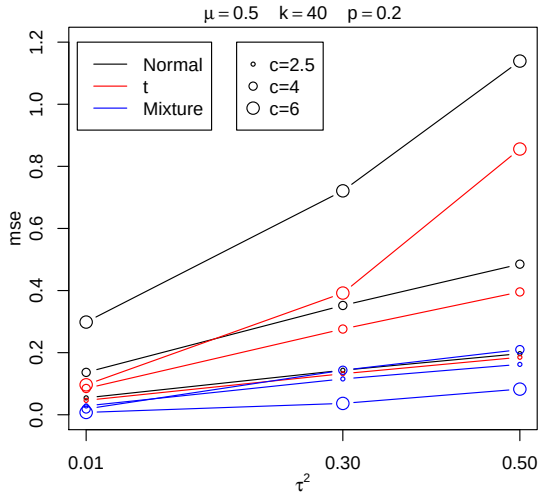

Figure 3.3: MSE for simulation scenario $\mu = 0.5$, $k = 40$ and $p = 0.2$

bias is lower for $p = 0.1$, but is higher for $p = 0.2$ (see Figure 3.6). This suggests that the t-distribution performs poorly when there are a higher number of outlying studies, but still better than the normal distribution. The coverage probability performs well for $\tau^2 = 0.01$ but deteriorates with higher heterogeneity, it is especially low when $p = 0.2$ (see Figure 3.4). In general, the t performs poorly when c is higher and when $p = 0.2$. It is suspected the t-distribution might perform better for higher c and p when outliers are present on both sides of the mean, rather than only one side. This will be discussed further in the next section.

When the mixture model is used for the random-effect there is significant improvement in the bias and MSE as compared to the normal and t-distribution. It is able to maintain the lowest bias and MSE, especially when $p = 0.2$ since it is easier to identify the two groups when there are more outliers present. The coverage

probability deteriorates the least and the Type I error is the smallest for $c = 6$ (see Figure 3.9).

The coverage probability quickly deteriorates as k increases for all scenarios and distributions, as shown in Figure 3.4. As k increases so does the number of outliers (while the proportion stays constant) which causes the estimate of μ to shift dramatically away from the true mean. The bias stays constant (see Figure 3.6) while the confidence widths simultaneously get smaller (see Figure 3.5) due to decreased variance of the estimate, therefore it is more probable that the confidence interval does not capture the true μ. Figure 3.7 demonstrates how the confidence interval for an example scenario are less likely to include the true mean as k increases. Furthermore, a supplementary simulation was done for $\tau^2 = 0.5$, $\mu = 0.5$, $c = 6$ and fixing the number of outliers at 2. It can be seen in Figure 3.8 that the bias, MSE and confidence widths are similar to the previous results. The coverage probability is no longer drastically decreasing, it stays reasonably constant for the normal and mixture models, while the t-distribution has a slight reduction in coverage. This verifies that the decrease in coverage as k increases seen in Figure 3.4 is due to the increasing number of outliers rather than an inherent flaw in the methods.

The confidence widths follow a similar trend for all three distributions. With all scenarios having the largest widths when $k = 10$, and decreasing as k increases due to the decreasing variance of the estimates as k increases. A sample scenario is shown in Figure 3.5. Also, the confidence width increases by approximately 50% when the proportion of outliers increases from $p = 0.1$ to $p = 0.2$.

Figure 3.10 shows that power increases as k increases, Figure 3.11 also shows that

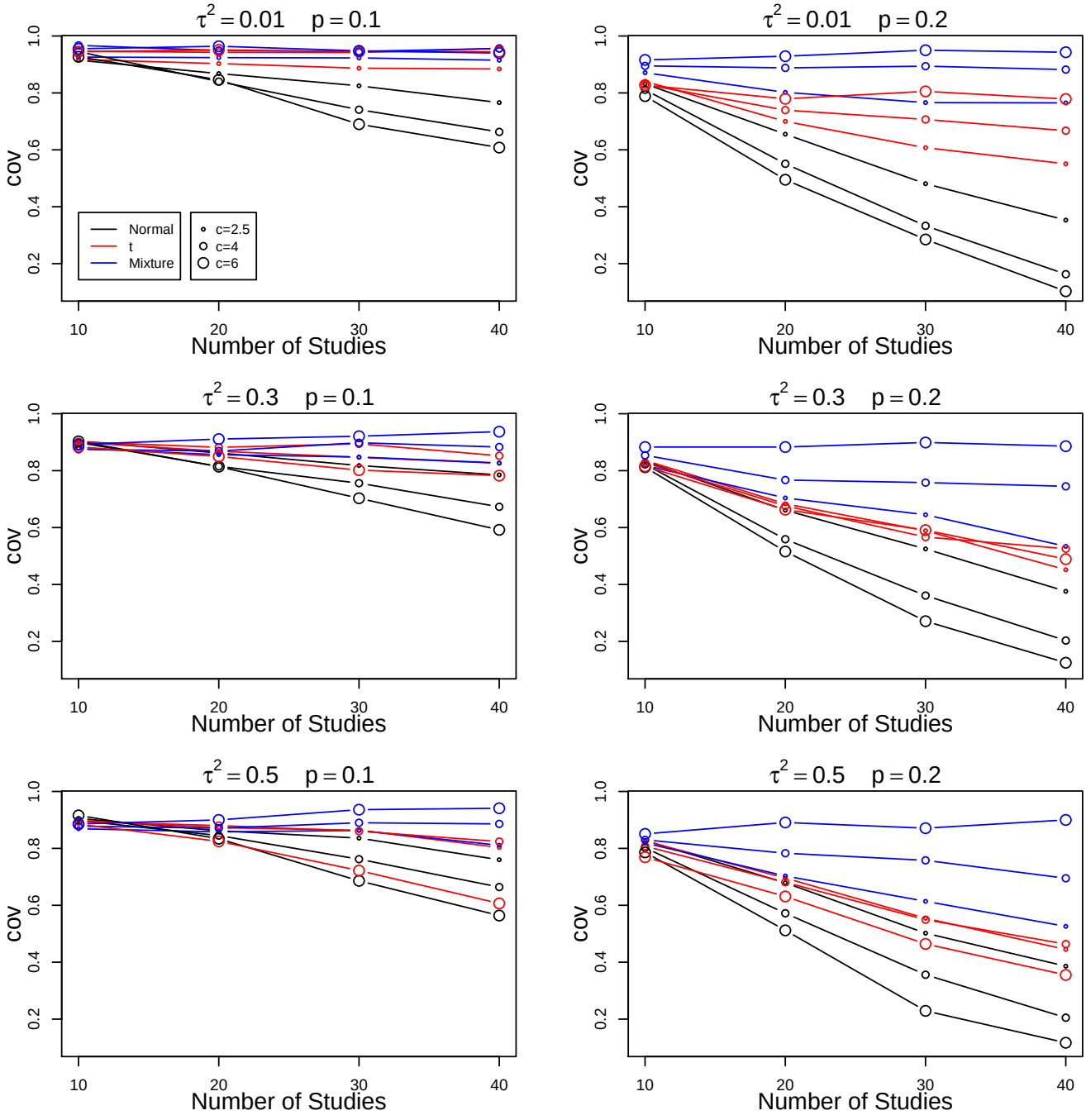

Figure 3.4: Coverage Probability for $\mu = 0.5$

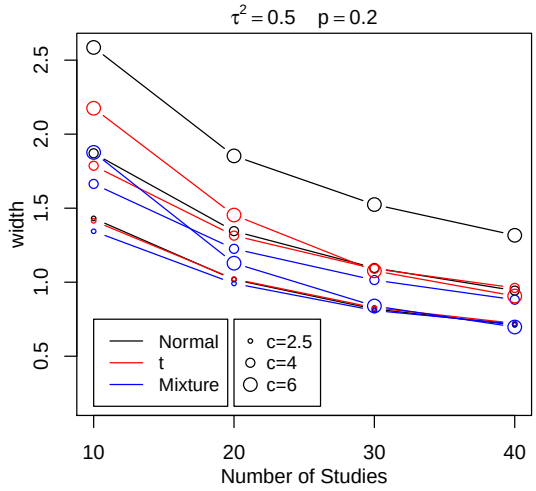

Figure 3.5: Confidence Width for simulation scenario $\mu = 0.5$, $\tau^2 = 0.5$ and $p = 0.2$

power increases as μ increases. This behaviour is expected. The mixture distribution with $c = 6$ has the lowest power while the normal distribution scenarios have the highest. This is a reflection of the normal distribution rejecting the null hypothesis too often due to the presence of outliers on the positive side of μ and far from zero, whereas the robust distributions are able to down weight the outliers and make a more conservative decision about rejection. The low power given by the robust distributions is balanced out by the low Type I error in Figure 3.9. Typically, Type I error is considered the more severe error since a high Type I error leads to incorrectly detecting significance. Though the mixture distribution has the least power in the more extreme scenarios, it has the best Type I error, indicating that it handles outliers well.

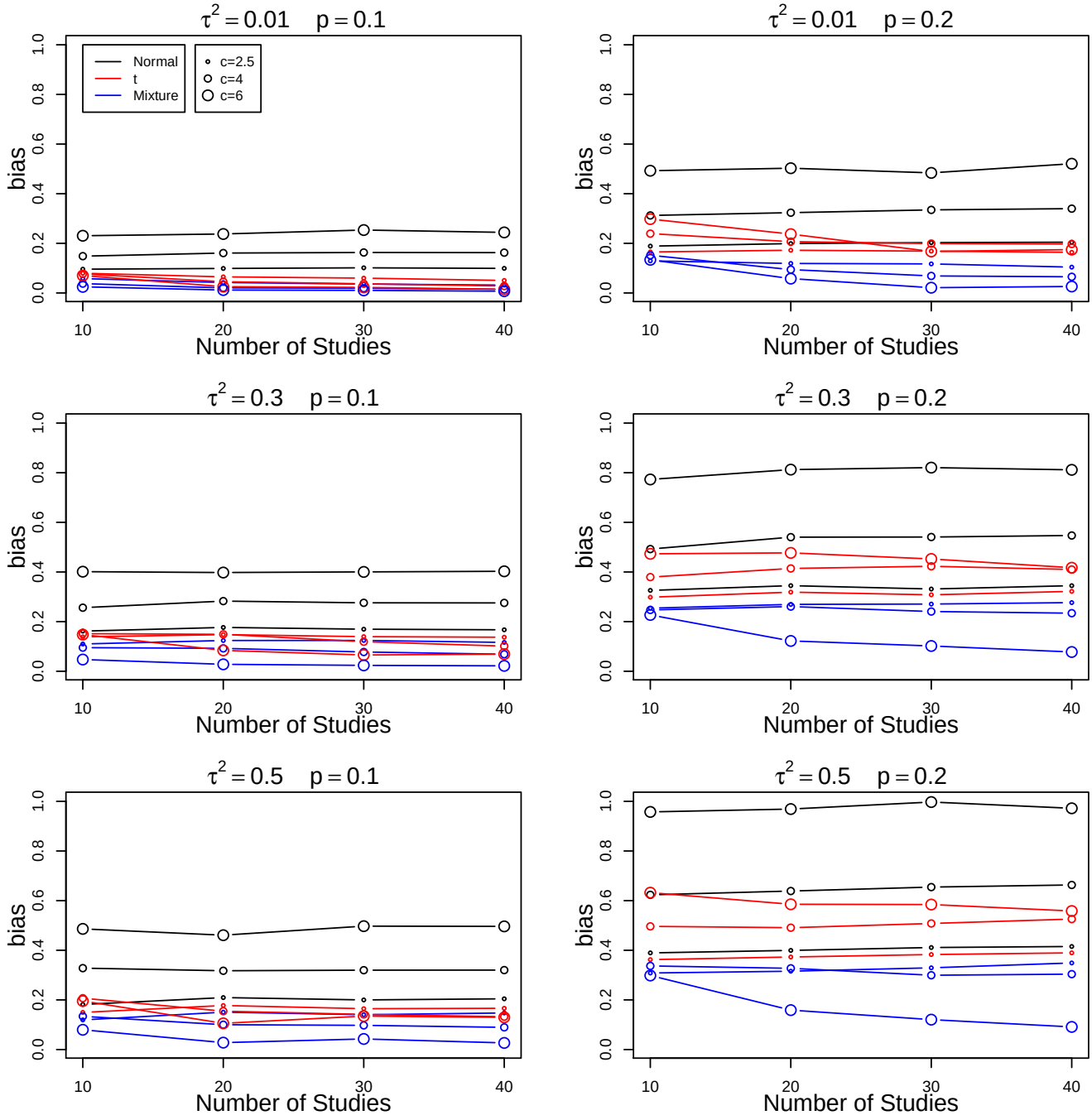

Figure 3.6: Bias for $\mu = 0.5$

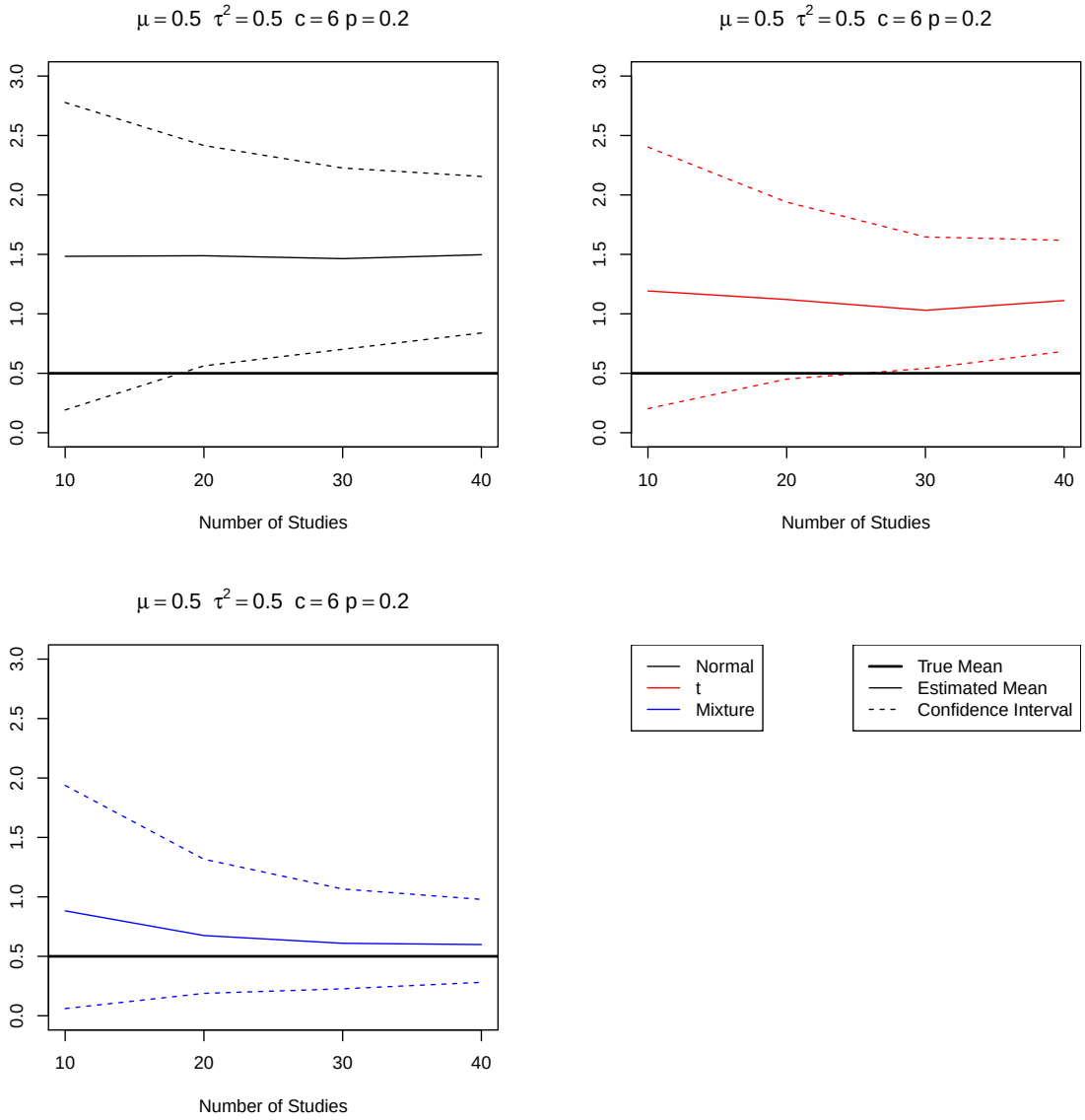

Figure 3.7: Confidence intervals for $\mu = 0.5$, $\tau^2 = 0.5$, $c = 6$ and $p = 0.2$

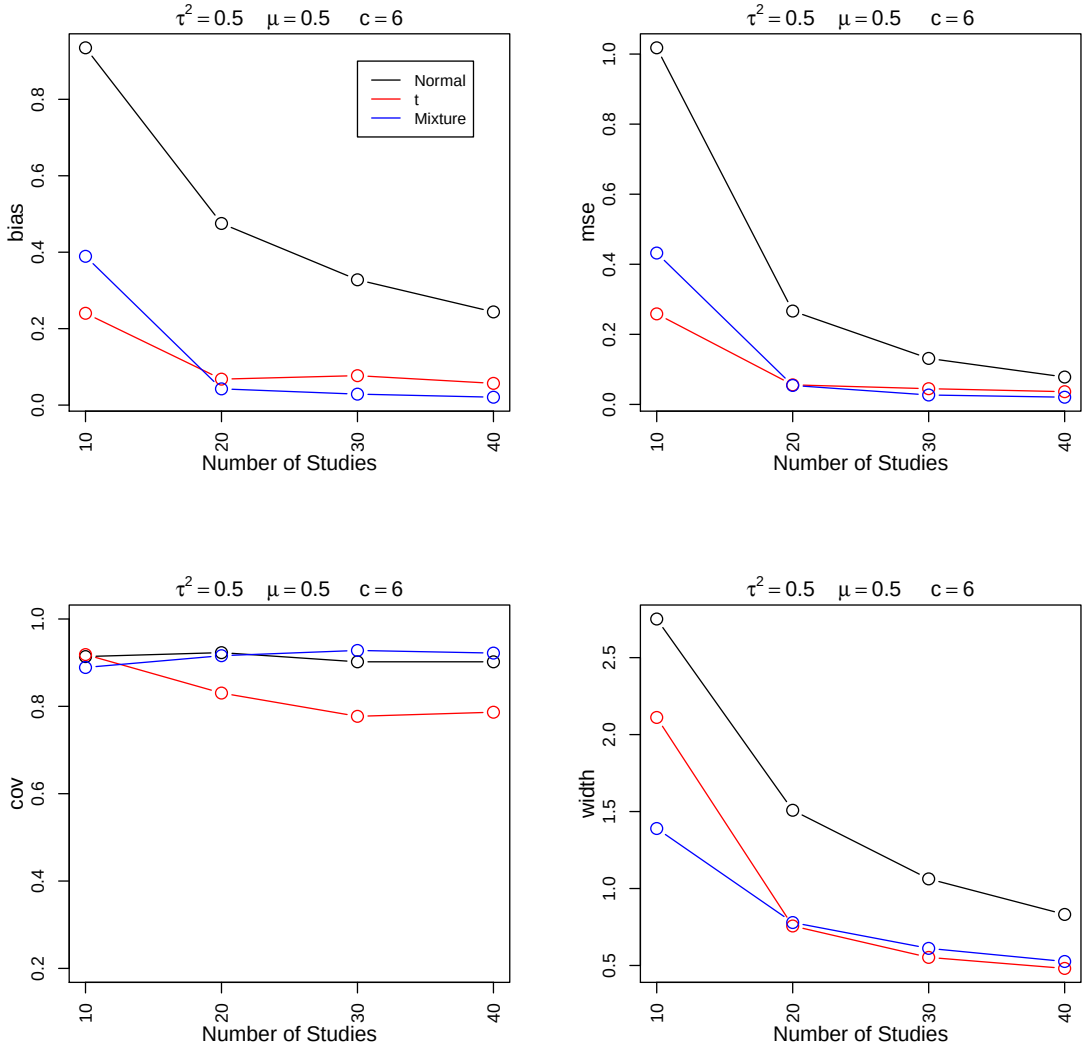

Figure 3.8: Coverage probability for $\mu = 0.5$, $\tau^2 = 0.5$, $c = 6$ and 2 outliers

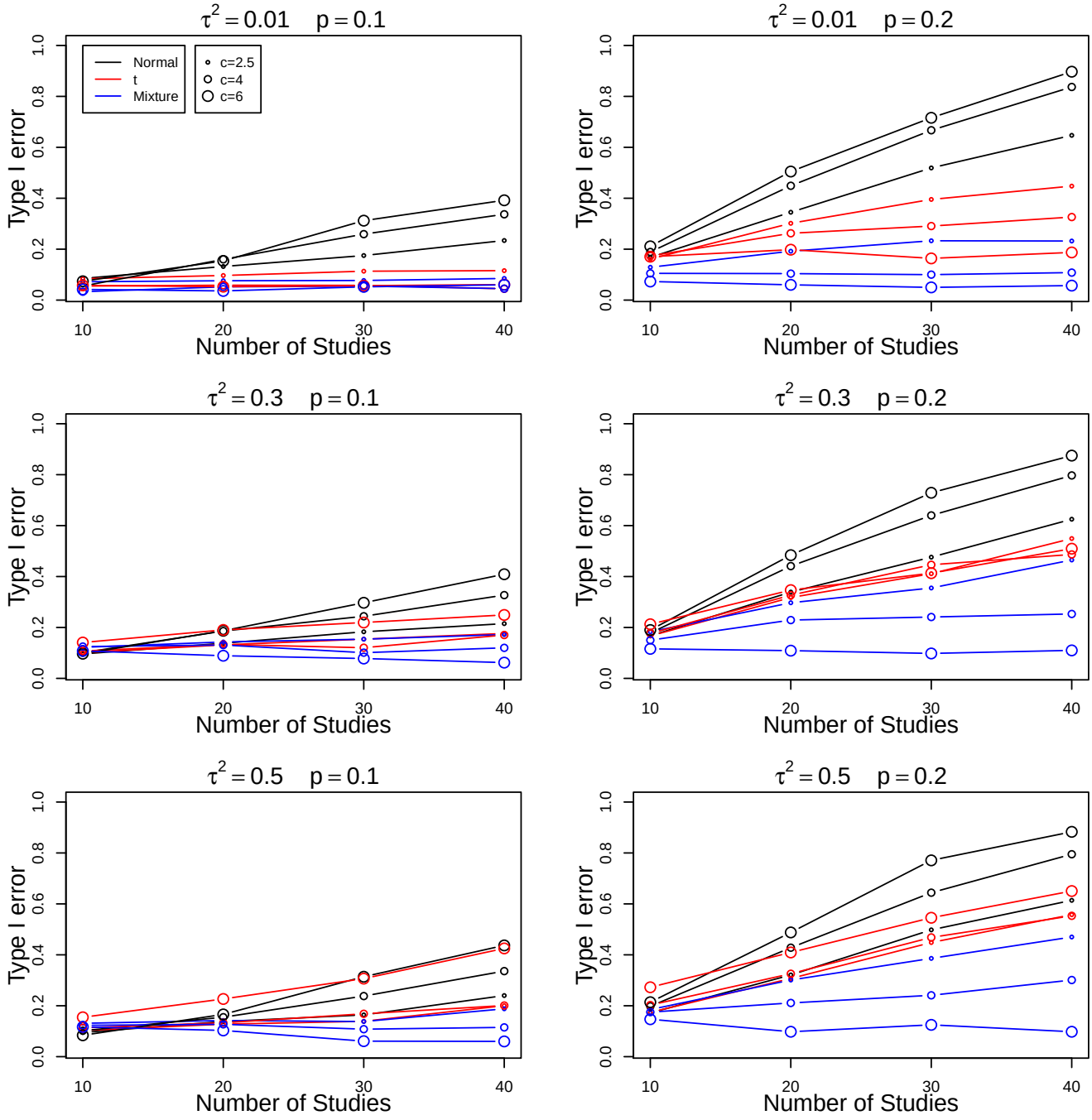

Figure 3.9: Type I Error

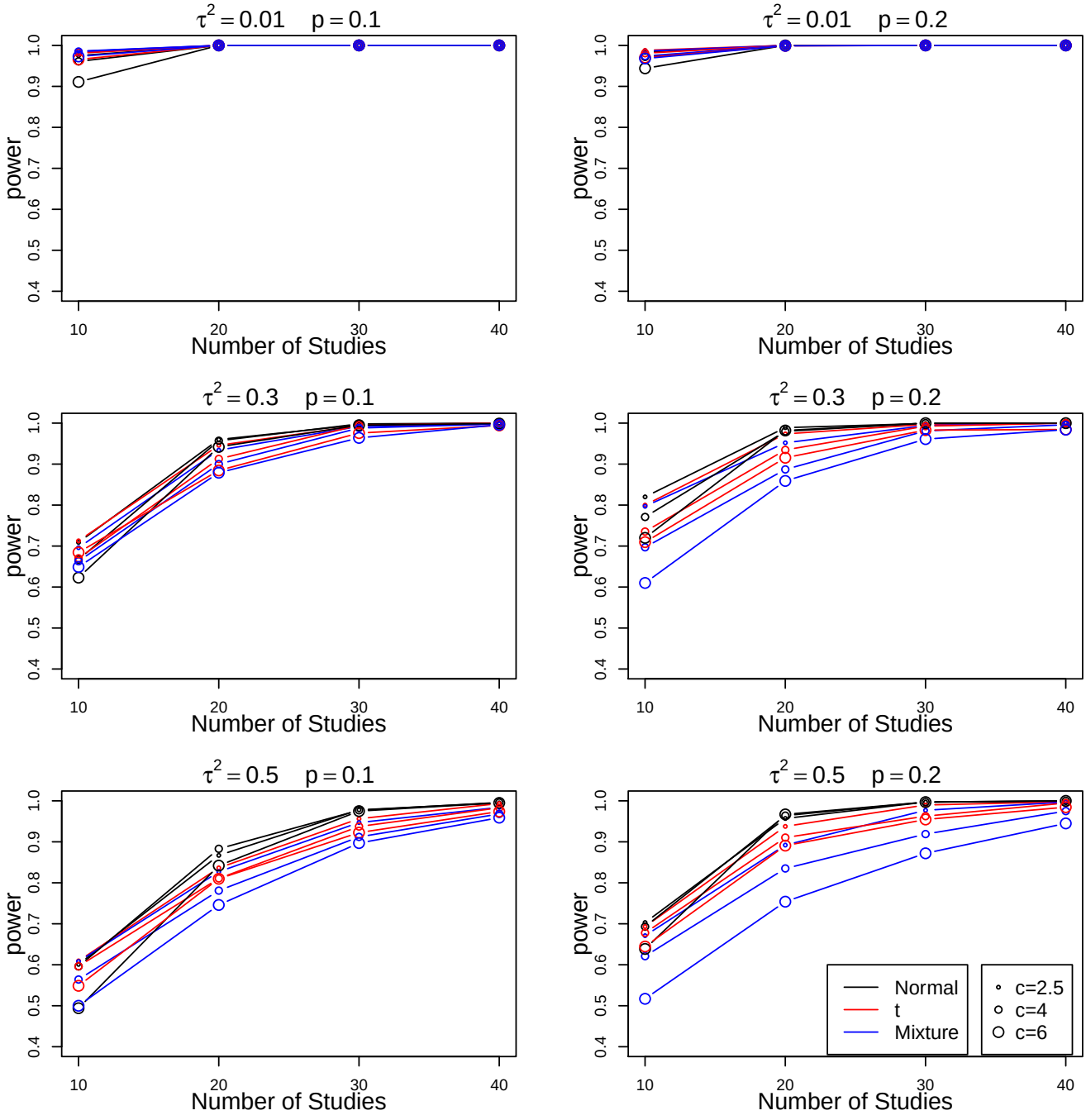

Figure 3.10: Power for $\mu = 0.5$

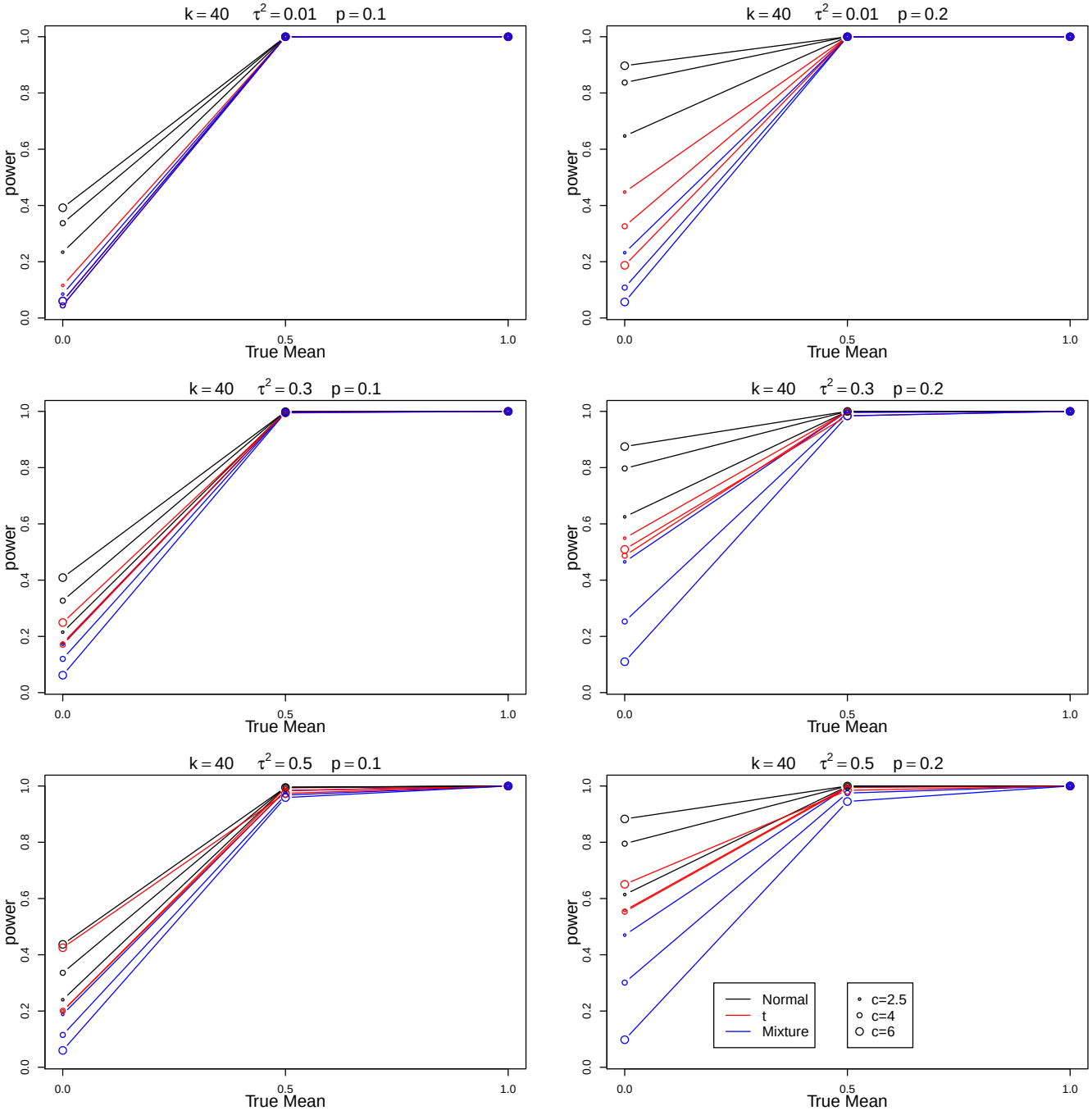

Figure 3.11: Power for $k = 40$

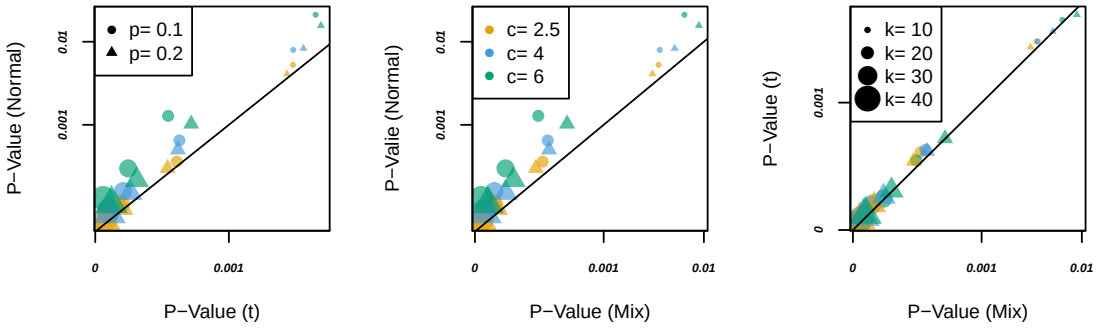

Figure 3.12: P-values by distribution for $\mu = 0.5$, $\tau^2 = 0.01$

3.4.1 p-Values

Figure 3.12 shows that for almost-zero heterogeneity both the t-distribution and the mixture model result in smaller p-values than the normal. The robust distributions are more likely to correctly reject the null hypothesis ($H_0 : \mu = 0$). However, Figure 3.13 indicates that for higher heterogeneity, the normal distribution results in smaller p-values than the other two distributions. In the presence of outliers, the normal distribution estimates μ further from the truth than the t-distribution and mixture model. The consequences are that the conclusions of the meta-analysis using the normal random-effects distribution in the presence of outliers may be misleading since it will use a distant outlier as strong evidence against the null hypothesis.

Friedrich et al. (2011) suggests that the 1/4 power transformation is most useful for displaying p-values and thus was used in Figures 3.12 and 3.13, and will continue to be used throughout this paper. This transformation preserves relative order but helps distribute values near zero to be closer to one, while larger p-values already

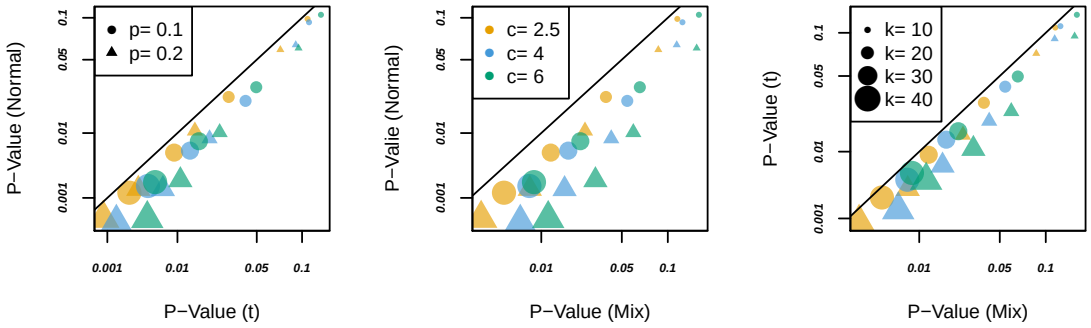

Figure 3.13: P-values by distribution for $\mu = 0.5$, $\tau^2 = 0.5$

close to one are only slightly adjusted.

Figure 3.14 shows the p-value from 4 different scenarios traveling across the $y = x$ line as τ^2 increases incrementally from 0.06 to 0.26, the numeric label on each point indicates the value of τ^2. The heavy-tailed distributions results in larger p-values than the normal distribution as the heterogeneity of the meta-analysis model is increased.

3.5 Investigation into Symmetrically Distributed Outliers

In the main simulation, the artificially inserted outliers were only simulated on the positive side of the true mean. This was done to mimic the "outlier shift" methods of Filzmoser (2005) and Rousseeuw and Driessen (1999). It also exaggerated the effect

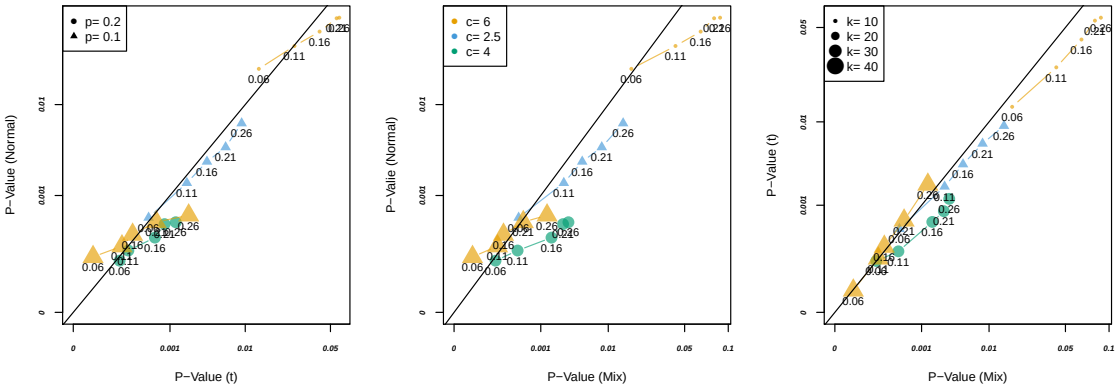

Figure 3.14: P-values by distribution for $\mu = 0.5$, $\tau^2 = [0.06, 0.26]$

of the outliers on point estimates. However, it is possible that the selected random-effects distribution are performing specifically better or worse due to the design of the simulated outliers. A supplementary investigation was done on 24,000 of the previously simulated meta-analyses by randomly permuting the generated outliers about the true mean μ.

Figure 3.15 shows that the bias is nearly zero in all cases. This is because the estimate $\hat{\mu}$ is no longer pulled away from the true mean in one direction and is instead centered around the true mean due to the outliers on both sides. On average, the bias is small, however, the symmetrically distributed outliers cause the estimates to be biased on either side of the true mean in individual cases. Figure 3.16 plots the standard deviation of the biases. We see that the bias for $c = 6$ is most variable, and the bias for the mixture model is the least variable for each c.

The MSE follows a similar trend but with smaller values than the previous results, for the same reason as for the bias. The mixture model still outperforms the other

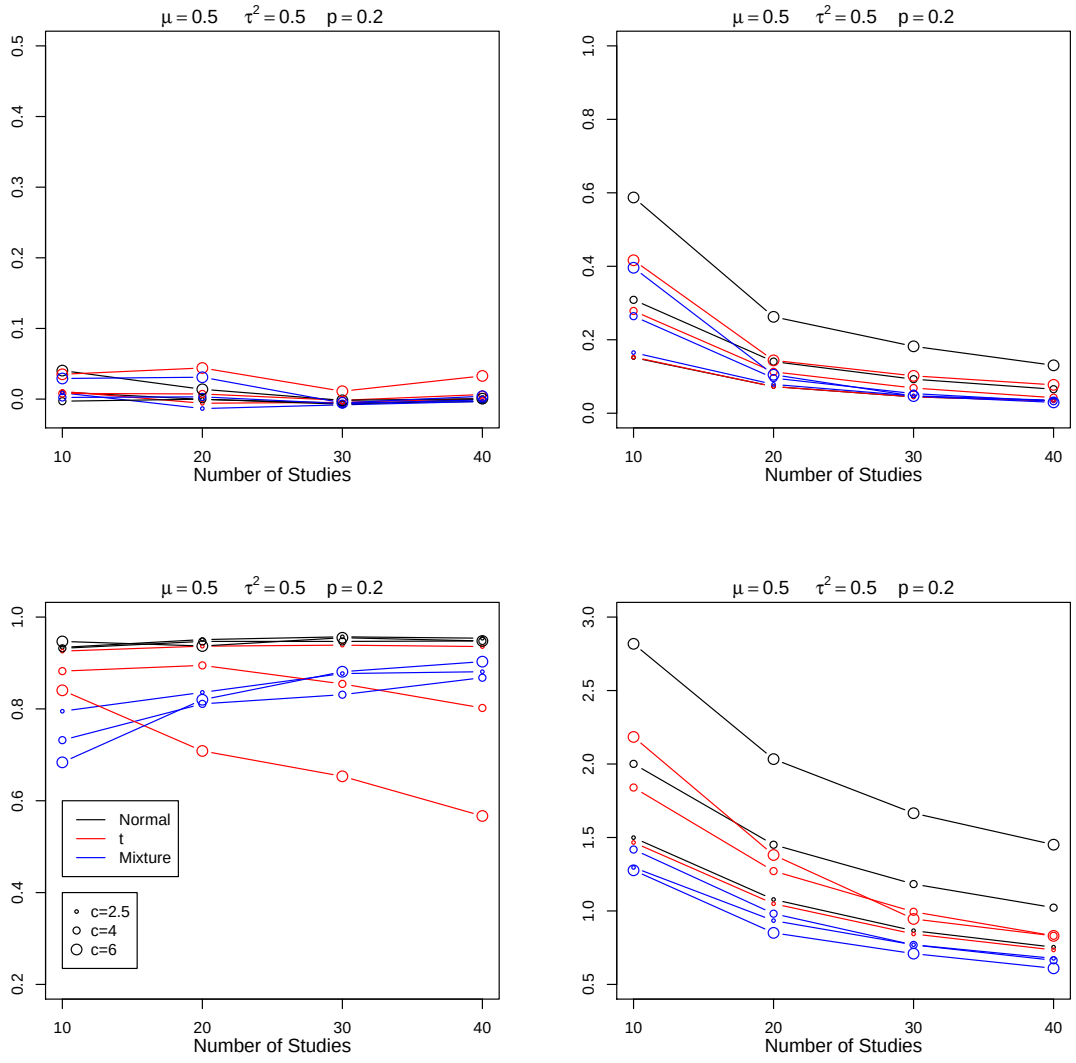

Figure 3.15: Performance measures for $\tau^2 = 0.5$ and $p = 0.2$ for symmetric outliers

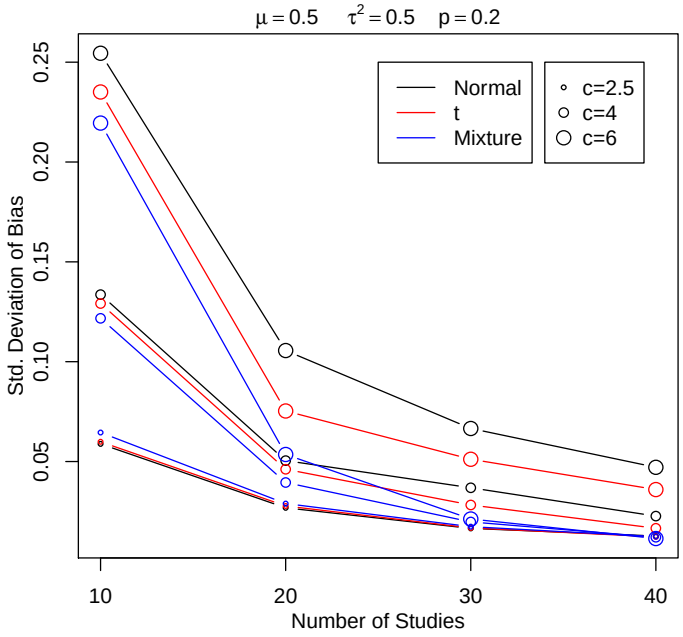

Figure 3.16: Standard deviation of the bias for $\tau^2 = 0.5$ and $p = 0.2$ with symmetric outliers

two distributions. The most notable difference between the one-sided and two-sided outliers is that the coverage probability is no longer drastically decreasing. The coverage is better in this situation for the same reason that the bias is better: the estimate is not being pulled in only one direction anymore, therefore, the confidence interval is more likely to capture the true mean.

For moderate and extreme outliers the t-distribution has decreasing coverage with k, this is a function of the absolute bias and the confidence width. For extreme outliers, the t-distribution has lower coverage then the normal because it has smaller confidence widths but still relatively high absolute bias. The t-distribution has lower coverage than the mixture model because it has relatively small confidence widths but a higher variation in bias meaning it is more likely have confidence intervals which do not include the true mean.

Figure 3.17 and 3.18 illustrate two examples of simulated data with symmetric outliers using the t-distribution. The forrest plot in Figure 3.17 contains five outliers on each side of the true mean $\mu = 0.5$, two of which have high precision on the left and one of which has high precision on the right. The overall estimate is slightly smaller than the truth (0.38) but the confidence interval contains 0.5. The forest plot in Figure 3.18 has six unbalanced outliers, four of which are influencing the estimate toward the negative side. The estimate using the t-distribution is 0.15 and the confidence interval does not contain the true mean. This is an example of high absolute bias and a small confidence width that contributes to a smaller coverage probability for the corresponding scenario.

The power in Figure 3.19 illustrates that symmetric outliers allows the mixture

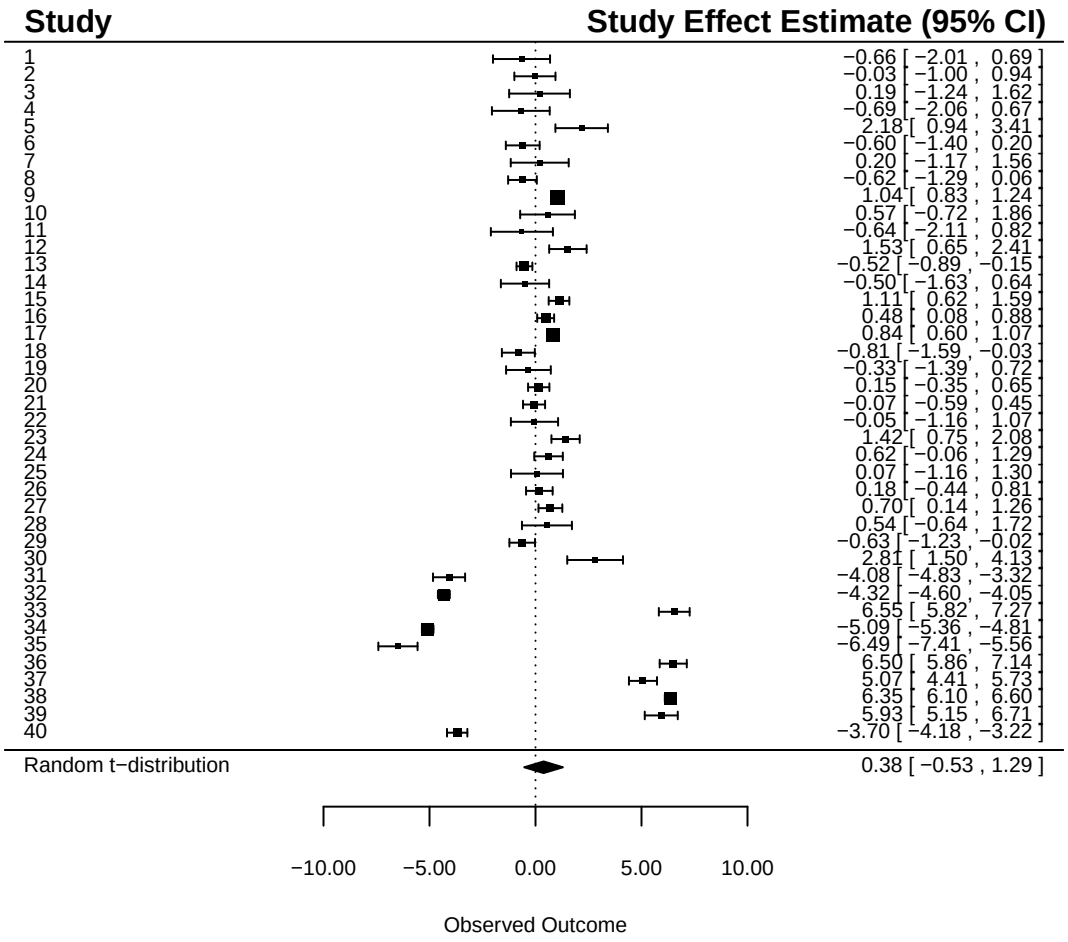

Figure 3.17: Example 1 of symmetric outliers for the t-distribution

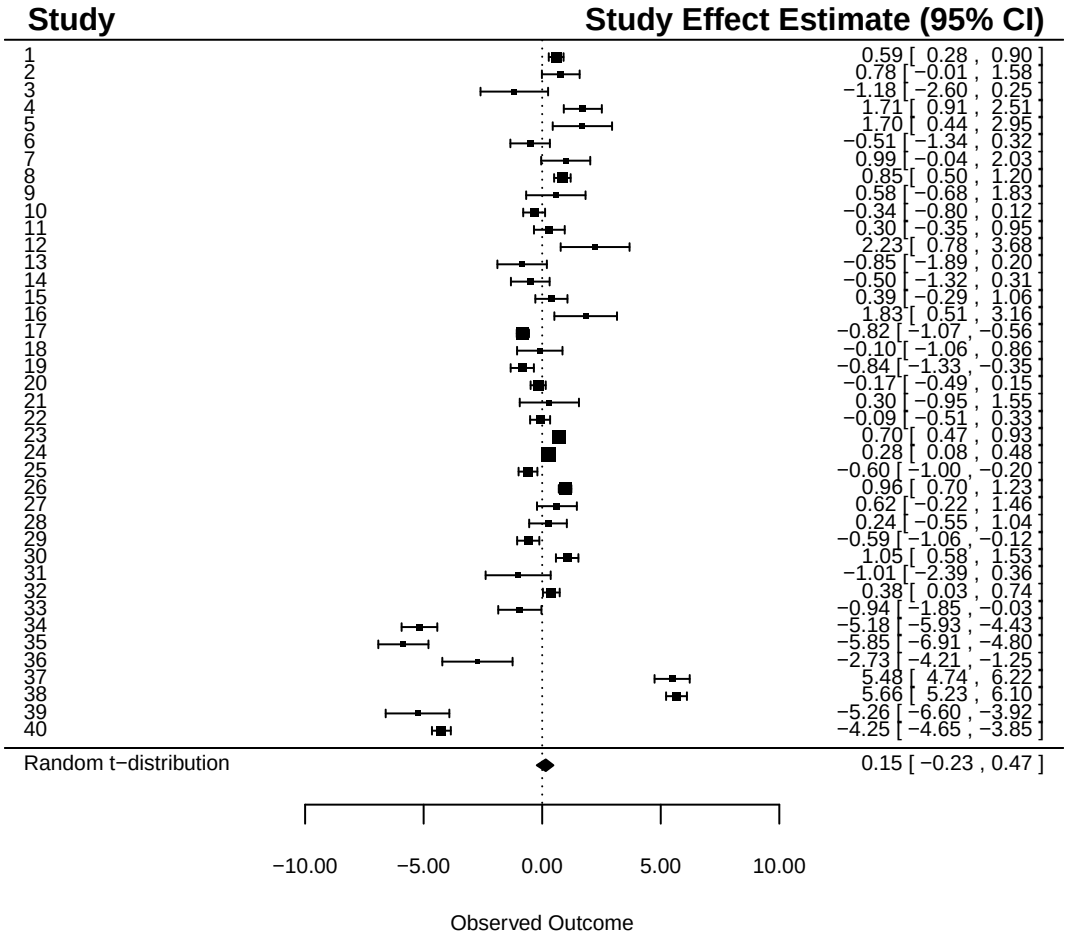

Figure 3.18: Example 2 of symmetric outliers for the t-distribution

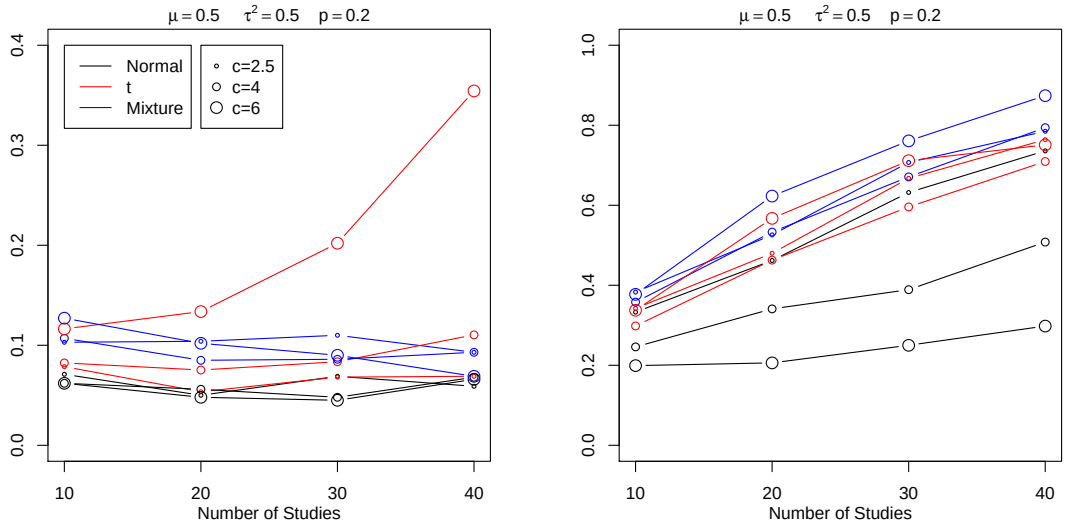

Figure 3.19: Type I Error and Power for $\tau^2 = 0.3$ and $p = 0.2$ for symmetric outliers

model to attain the highest power while the normal has the lowest power. The normal distribution is less likely to detect significance because the estimate of the heterogeneity is inflated while the bias is very small. This gives a small test statistic and a large p-value. The mixture model gives a smaller estimate of the heterogeneity than the normal and smaller p-values and is, therefore, correctly rejecting the null hypothesis more often.

For extreme outliers, the Type I error for the t-distribution deteriorates with k, though at its maximum it is still an improvement over the Type I error in the corresponding scenario from the main simulation (see Figure 3.9). It is falsely detecting significance due to the high absolute bias which also causes low coverage.

The purpose of this investigation was to check that the results of the main simulation apply when there are outliers present on both sides of the mean. This has been verified for the sample scenario of $\tau^2 = 0.5, p = 0.2$, the parameter values for which where chosen to reflect a more extreme case where we would expect the greatest disparity in results, if they exist. As with the results found with the main simulation, it can be expected that the methods will perform better overall with a smaller proportion of outliers. Performance measures and hypothesis measures for this investigation can be found in Appendix Tables B.1 and B.2.

Regardless of the position of the outliers relative to the true mean, the normal distribution under performs as compared to the alternative distributions. The high coverage probability of the normal is a reflection of inflated confidence widths rather than accuracy, and the estimates produced using the normal are highly biased with more extreme outliers present. The t-distribution offers some improvement over the normal with average bias and confidence width, however, due to the variability of the bias in individual cases, the t-distribution has poor absolute bias and decreasing coverage, especially with extreme outliers. Overall, the mixture model proves to be the most advantageous in the case of extreme and abundant outliers.

3.6 Limitations

This simulation study was limited by computational power and time. Each of the 216,000 meta-analyses needed to be evaluated 3 separate times using the **metaplus()** function, one for each random-effects distribution. The computation using normally

distributed random-effects was not very intensive, however, the t-distribution required the approximation of an integral for each meta-analysis, and both the finite mixture model and the t-distribution required a quasi-Newton method for maximizing the likelihoods. In total, the main simulation and analysis took approximately 5 weeks to complete in parallel on 30 nodes. Ideally, more comprehensive scenarios with greater granularity in parameters would have been run.

Another limitation was the heavy reliance on the *metaplus* package. There were meta-analyses which produced errors and warnings. It was expected that some estimates would not converge, or some matrices requiring inversion were near singular and failed to compute. With more time and experience it may have been possible to solve the problems.

Although errors are not desirable, they are not synonymous with mistakes. Often, computational limitations will lead to errors which cannot be easily rectified, this does not necessarily mean there is a problem with the method, it may mean that no solution has been implemented. In this simulation study, the computational limitations arose mainly from using the t-distribution to estimate the parameters of the random-effects model. As seen in equation 2.14, the likelihood contains an intractable integral which must be solved numerically, and then the quasi-Newton method requires the inversion of a matrix to maximizes the likelihood. If these methods do not work for some combinations of values then we can say it is a computational limitation. However, if there is a relationship between the values which cause problems then by excluding those instances we may have a misleading result within the simulation scenario.

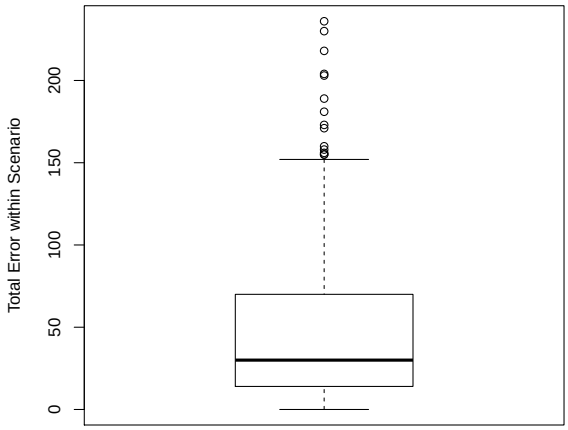

Figure 3.20: Boxplot of error counts per scenario of 1,000 for the t-distribution

There were a total of 11,034 errors out of the 216,000 (5.1%) simulated data sets which were analyzed using the t-distribution. An error instance did not produce a result, thus the total number of results are lessened by the number of instances which produced errors. Using the normal distribution produced relatively few errors ($< 1e^{-4}\%$) and using the mixture model produced none. Some data sets produced reliable parameter estimates with unreliable confidence intervals due to multi-modal profile likelihoods. These results were kept but their respective confidence limits were not included in the coverage or confidence width calculations.

The t-distribution errors are due to a combination of non-invertible matrices and difficult-to-estimate likelihoods. Interpreting error messages was aided through correspondence with the author of *metaplus*. Figure 3.20 displays the frequencies of errors within each of the 216 scenarios, the median is 30 errors per scenario of 1,000

Parameter	Value	Error Count	Total Simulations	Percentage
μ	0	3,832	72,000	5.32%
	0.5	3,560	72,000	4.94%
	1	3,642	72,000	5.06%
τ^2	0.01	2,679	572,000	3.72%
	0.3	3,984	72,000	5.53%
	0.5	4,371	72,000	6.07%
k	10	1,105	54,000	2.04%
	20	2,595	54,000	4.81%
	30	3,209	54,000	5.94%
	40	4,125	54,000	7.63%
p	0.1	4,869	108,000	4.51%
	0.2	6,165	108,000	5.71%
c	2.5	1,160	72,000	1.61%
	4	2,595	72,000	3.60%
	6	7279	72,000	10.1%
Total		11,034	216,000	5.11%

Table 3.4: Error Counts by Parameter Value

replicates. The value of the extreme upper whisker is 152 and there are 14 scenarios which contain more than 152 errors. Table 3.4 displays all of the error counts when broken down by parameter values. It is especially clear that for $c = 6$ the function encounters more errors.

The more frequent errors for $c = 6$ reflects a limitation in the **metaplus()** function. There was no pattern in the simulated data which caused errors in the $c = 6$ scenarios and the simulated data which did not cause errors when $c = 6$. Figure 3.21

illustrates the means and variances of the simulated effect sizes (y_i and $y_{i_{out}}$), the means and variances of the simulated within-study-variances (v_i and $v_{i_{out}}$), as well as the number of outliers produced (k_{out}). Finally, Figure 3.22 compares the means of the simulated outlier values and their associated within-study-variances as well as the distance between the means of the "clean data" and the outlier studies. With no obvious difference between the data which did and did not produced errors within $c = 6$, it would be reasonable to assume that the errors are not occurring for any particular type of data within $c = 6$. Given the chance to re-run this simulation, we would re-simulate enough data to ensure that each scenario had 1,000 useable results.

Ideally, it would be important to attempt to rectify the problem by adjusting the function. Due to the high number of error and warning messages, as well as the complexity of the methods included in *metaplus* it was not possible to individually explore each error instance and attempt to resolve the problem. Instead, it was sufficient to report on the errors and present them as limitations.

It is worth noting that although the t-distribution is less computationally reliable for extreme outliers, it is also questionable whether a researcher would always allow for an extremely distant point to be included in a meta-analysis. We discussed that it is disadvantageous to remove potential outliers, in reality however, a point which is *so* distant from the rest of the points may be irrelevant to include in the meta-analysis, especially if the alternative distribution encounters computational problems. Furthermore, the design of our simulation allowed for a probability of outliers of up to 0.2, which means on average there are 6 and 8 outliers for $k = \{30, 40\}$, respectively.

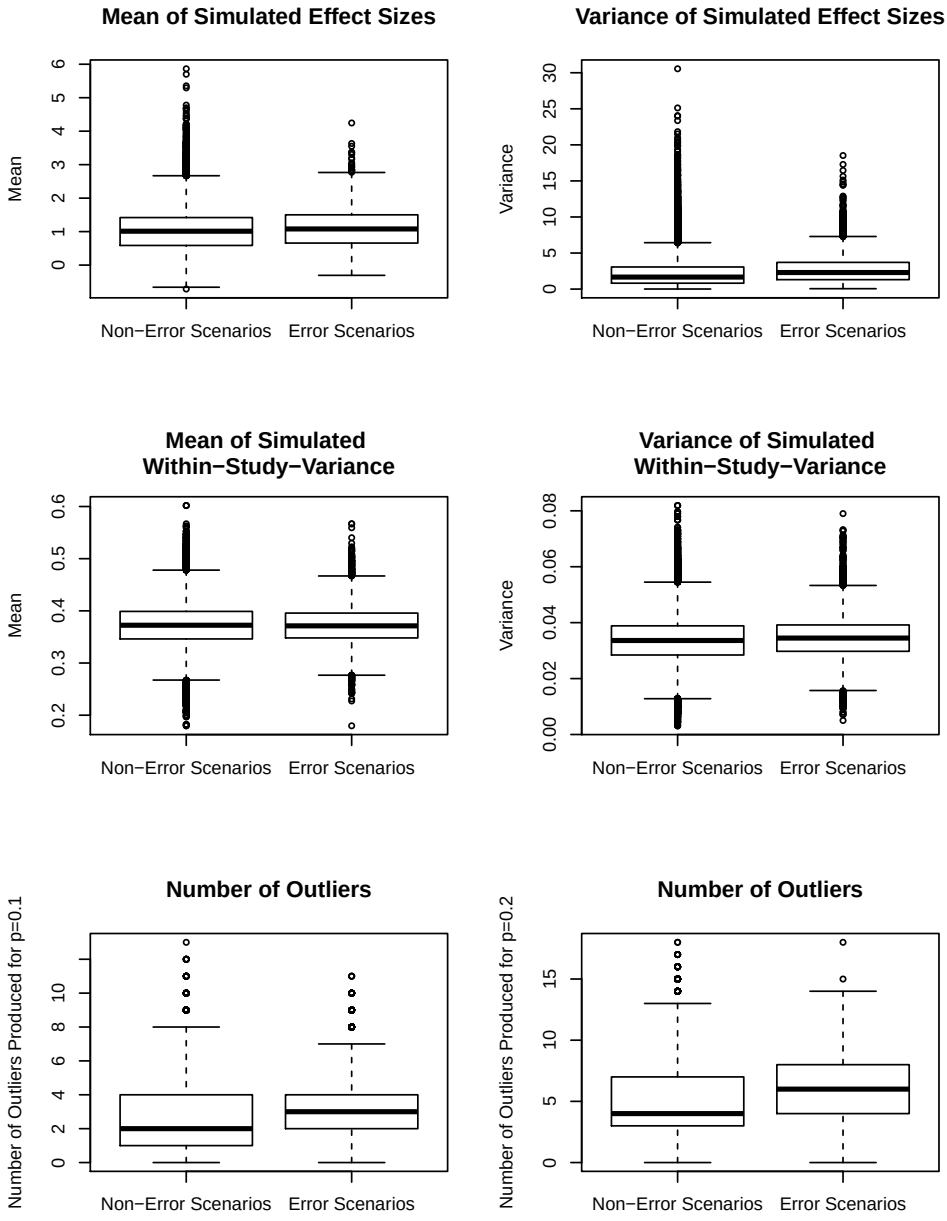

Figure 3.21: Boxplot of simulated data summaries for $c = 6$ scenarios using the t-distribution

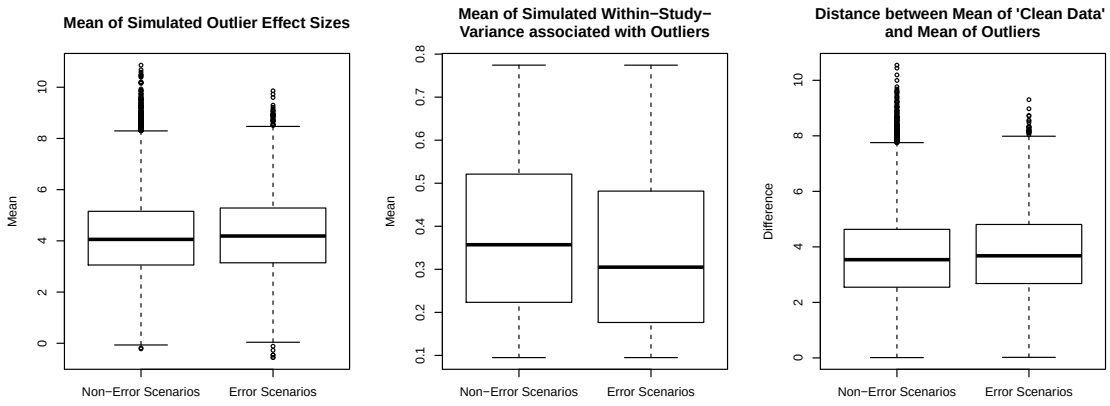

Figure 3.22: Boxplot of outlier summaries for $c = 6$ scenarios using the t-distribution

Having so many outliers indicates that the data should be modeled using two groups, which means a t-distribution would inherently perform poorly in this case and a mixture model would perform best.

Chapter 4

Real Data Application

While a simulation study allows us to understand the behaviour of certain performance measures across the three random-effects distributions, application to real data provides more verification for the results of the study as a whole.

4.1 Data and Methods

We used the data set assembled by Friedrich et al. (2011) using The Cochrane Library (2008). The included meta-analyses had to have at least 5 trials and used either the mean difference (MD) or standardized mean difference (SMD) as the effect measure. There were some other exclusion criteria which ensured consistency in the values. The resulting data set contained 232 meta-analysis. We were only able to calculate the mean difference (MD) for 142 of these, due to the use of subjective measurement scales in the remainder of the meta-anlyses. Calculations of standardized mean difference (SMD) and ration of means (RoM) were possible for all 232 meta-analyses.

Friedrich et al. (2011) compared the MD and SMD for these meta-analyses against the proposed ratio of means measure (RoM). The ratio of means is calculated by dividing the mean of the treatment group by the mean of the control, or second treatment group (Friedrich et al., 2011). In the following real data application we will compare these three effect measures using the t-distribution and mixture model for the random-effects against the normally distributed random-effects model, for which τ^2 has been estimated using either DerSimonian and Laird (DL), maximum likelihood (ML) or restricted maximum likelihood (REML).

4.2 Results

Figure 4.1 and 4.2 compare the p-values from the meta-analyses using the normally distributed random-effect model (y-axis) and robust distributions for the random-effects (x-axis). The majority of the points are below the $y = x$ line, which indicates that the proposed robust distribution is producing larger p-values and is being more conservative in general. The most disparity in p-values between the two methods can be found in the lower left quadrant. The meta-analyses with fewer studies are getting consistently higher p-values.

There are a number of studies which fall into the lower right quadrant which are found to be significant at the 5% level under the normal random-effects model, but are not significant under the robust model. We can test whether the proportion of meta-analyses which switch significance from one method to another is itself significant using a McNemar test and a sign test. The McNemar test is used to test the null

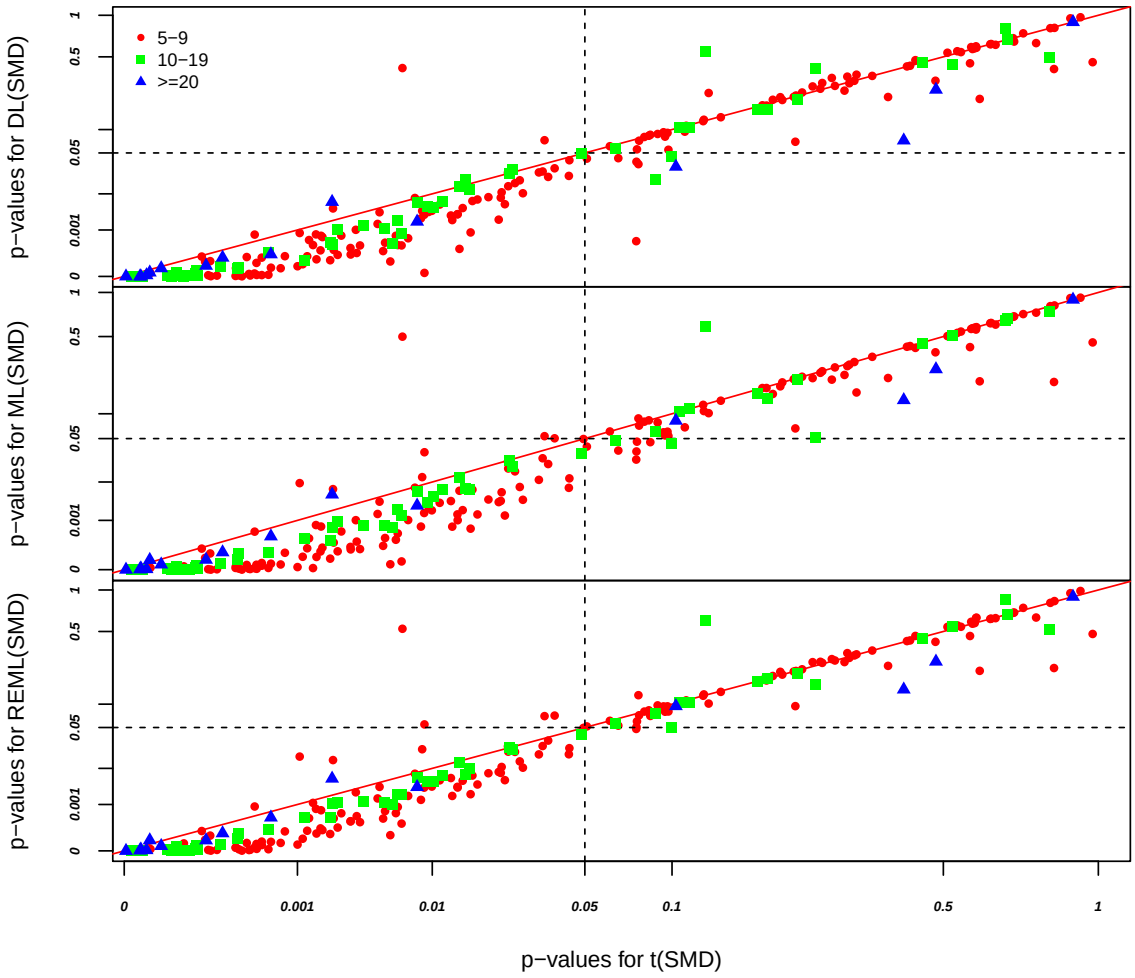

Figure 4.1: P-Values for Normal vs. t using SMD

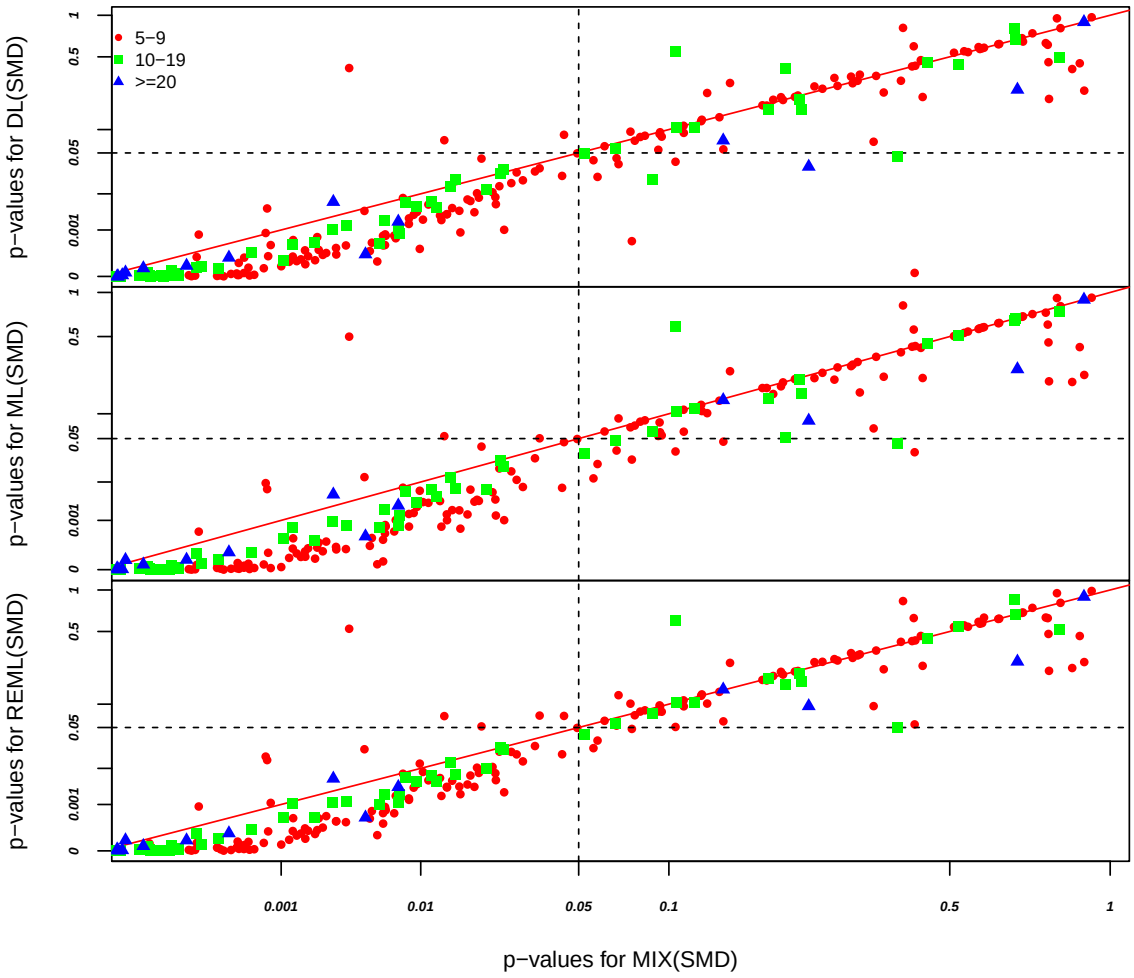

Figure 4.2: P-Values for Normal vs. Mixture using SMD

		Robust Distribution	
		$p < 0.05$	$p \geq 0.05$
Normal Distribution	$p < 0.05$	a	b
	$p \geq 0.05$	c	d

Table 4.1: Example 2x2 Contingency table used in McNemar tests

hypothesis of marginal homogeneity ($H_0 : p_b = p_c$) in paired nominal data. An example of a 2x2 contingency table is Table 4.1, where a, b, c, d are the counts of the meta-analyses and p_b, p_c are the probabilities of falling into the off-diagonal categories in which the p-values disagree. The McNemar tests performed on the real data using a 5% significance level for the meta-analysis did not detect significant differences in the off diagonal entries (see an example in Table 4.2). However, using a more conservative significance level of 1% yielded a significant result in the majority of the tests (see Table 4.3). The associated sign test is a non-parametric test which tests if one group of the paired observations is equally likely to be larger or smaller than the other group. The sign test in Table 4.2 does not reject the null hypothesis, while the sign test in Table 4.3 rejects the null hypothesis in favour of one group having a higher probability of being larger than the other group. According to these tests, under the 1% significance level, if the normal method and the alternative method do not agree in regards to the significance of the meta-analysis, then it is more probable that the alternative method has resulted in non-significant results while the normal method has claimed that there are significant results.

The limits of agreement plot was proposed by Bland and Altman (1999) to investigate where the differences between two methods are expected to lie based on the

		Mixture Model	
		$p < 0.05$	$p \geq 0.05$
Normal Distribution (ML on SMD)	$p < 0.05$	138	10
	$p \geq 0.05$	3	81

McNermar p-value = 0.0961 Sign test p-value = 0.0923

Table 4.2: 2x2 Contingency table used in McNemar tests with $p = 0.05$ cut-off

		Mixture Model	
		$p < 0.01$	$p \geq 0.01$
Normal Distribution (ML on SMD)	$p < 0.01$	102	24
	$p \geq 0.01$	2	104

McNemar p-value=$3.814e^{-5}$ Sign test p-value = $1.049e^{-5}$

Table 4.3: 2x2 Contingency table used in McNemar tests with $p = 0.01$ cut-off

assumption that the differences are normally distributed. Since there are more points above the upper 95% agreement boundary in Figure 4.3 the p-values produced by the robust method are often larger than the p-values produced by the normal method. This indicates that the robust methods are acting more conservatively and would be less likely to reject the null hypothesis. This trend is true for all combinations of effect measures and estimation methods, for both the t-distribution and the mixture model.

Figure 4.4 plots the difference between p-values produced using the t-distribution and p-values produced using the normal distribution against the number of studies. It is clear that the differences are most pronounced for smaller meta-analyses. This suggests that for smaller meta-analyses, outlier analysis should be performed and a robust distribution considered if appropriate. A practitioner should exercise caution

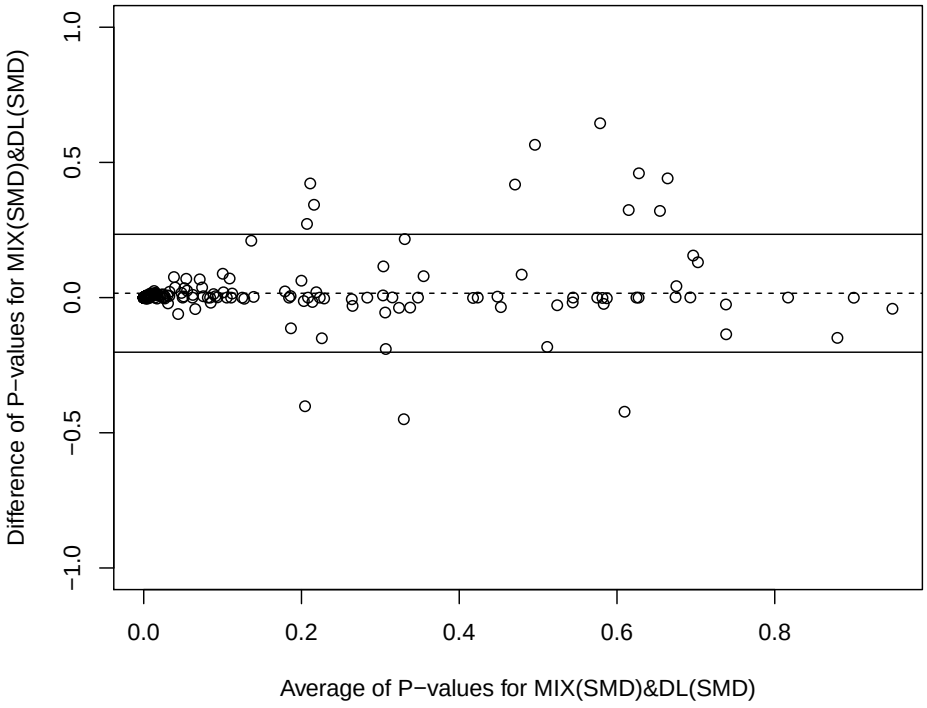

Figure 4.3: Limits of Agreement for Normal vs. Mixture using DL and SMD

since some distributions may not be practical for a meta-analysis with very few studies.

Meta-analyses 78 and 110 are large meta-analyses which are found to have positive outliers. Figure 4.5 shows the forest plot for meta-analysis 78, the potential outlier is study 25. This outlier pulls the treatment effect estimate away from zero when the normal distribution is used. When the heavy-tailed alternative is used, the estimate relaxes toward zero. The p-value is smaller for the normal since it is erroneously detect more evidence against the null hypothesis. This is why ΔP is positive in Figure 4.4. The same holds for meta-analysis 110, there is a negative outlier, but there are more positive outliers with higher precision. Meta-analyses 65 and 193 are displaying the opposite effect; the outlying study is negative while the non-outliers are positive, see Figure 4.6. The negative outlier pulls the treatment effect estimate from positive toward zero, therefore, using the normal distribution will produce a larger p-value than using the robust distribution, this is why ΔP is negative.

In general, the differences between the standard method and the robust methods are most pronounced for small number of studies and for meta-analyses with identifiable outliers.

While the differences in the p-values are informative, the purpose of the study is to investigate the effects on meta-analyses which contain outliers. To classify studies as outliers, the following method of simulated envelopes proposed by Julious and Whitehead (2012) was used:

1. Simulate k treatment effects from Normal$(0, \tilde{w}_i^{-1})$ where $\tilde{w}_i$ is the random-effects study weight (equation 2.1) for study i using the known study variance

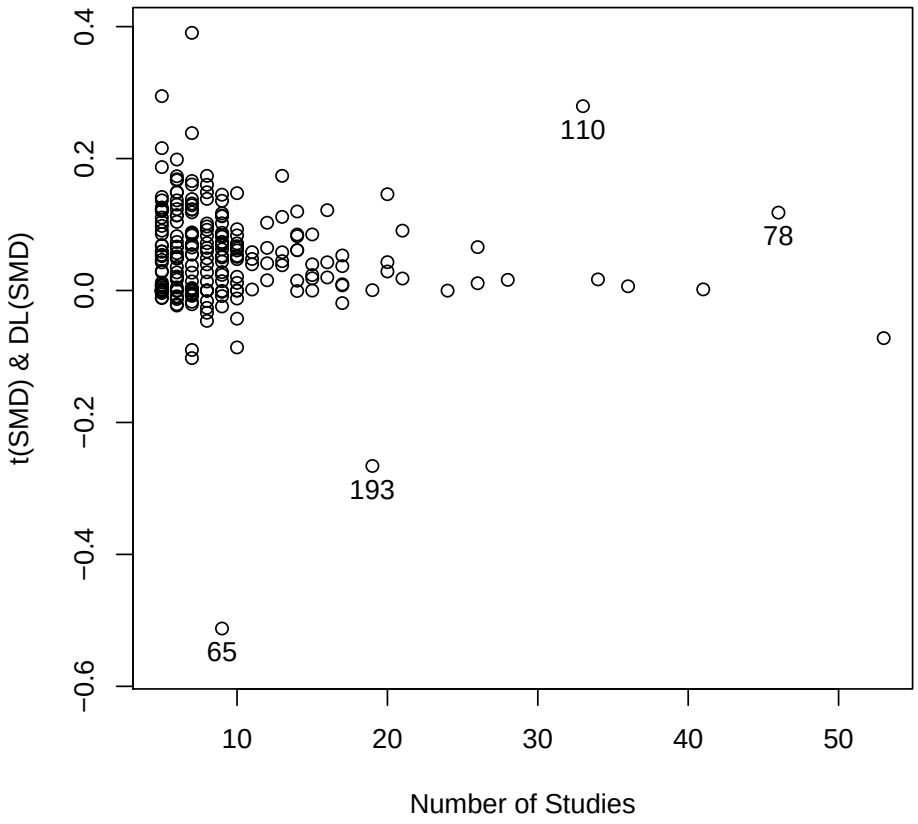

Figure 4.4: ΔP vs. Number of Studies for T-Normal using DL and SMD

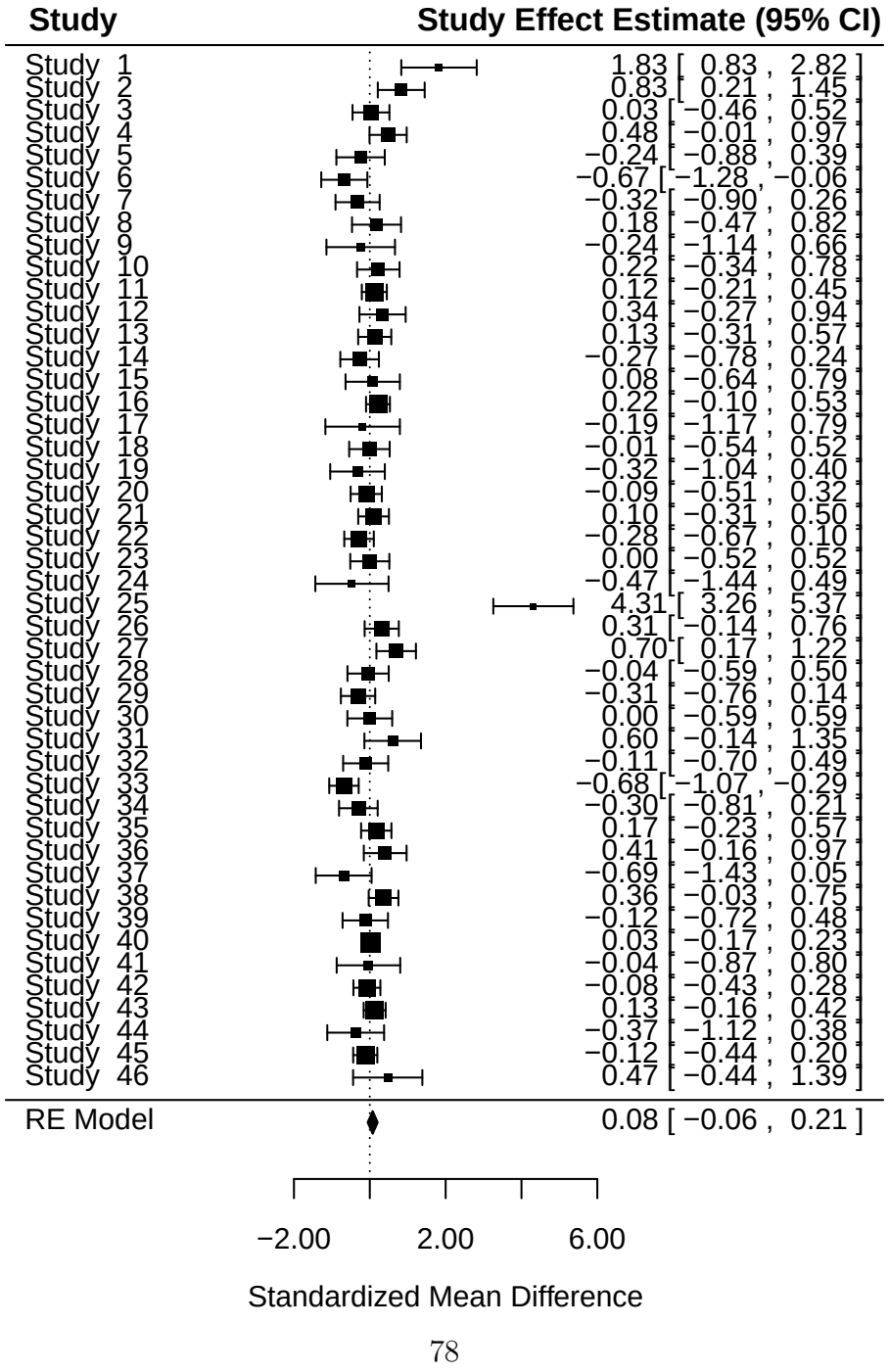

Figure 4.5: Forest plot of meta-analysis 78 using DL and SMD

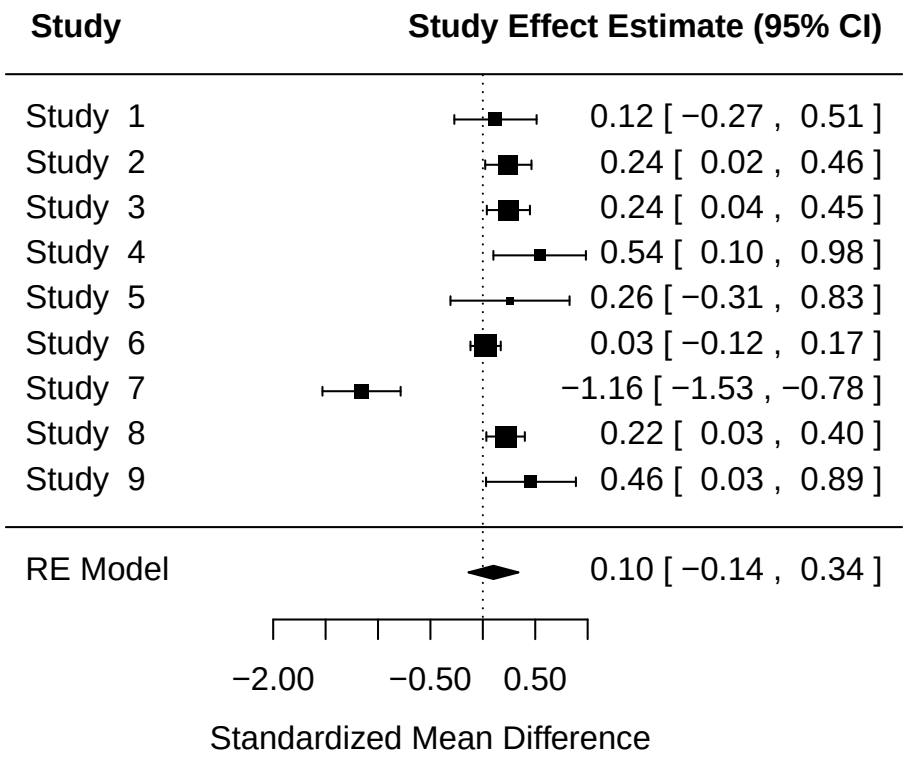

Figure 4.6: Forest plot of meta-analysis 65 using DL and SMD

v_i, and $\hat{\tau}^2$ is the estimate from the random-effects meta-analysis using the observed data.

2. Perform a random-effects meta-analysis using the simulated treatment effects and the study variances v_i, calculate the standardized weighted residuals q_i for each study i where $q_i = \frac{(y_i - \hat{\mu})\sqrt{\tilde{w}_i}}{\sqrt{1 - \tilde{w}_i / \sum_{i=1}^{k} \tilde{w}_i}}$.

3. Order the q_i from smallest to largest: $q_{(1)} \leq q_{(2)} \leq ... \leq q_{(k)}$.

4. Repeat the simulation 1,000 times to obtain 1,000 vectors of ordered standardized weighted residuals.

5. Order all of the 1,000 $q_{(i)}$ for each i and take the 25^{th} and 975^{th} of each to obtain the 2.5% and 97.5% interval for the i^{th} ordered residual.

A study whose residual falls outside of the corresponding interval is considered an outlier. Figure 4.7 and 4.8 compare the p-values of the normal method against the robust distributions with the meta-analyses containing outlier studies differentiated from those which do not. For meta-analyses with at least 10 studies we can see the majority of the outlier points fall below or on the $y = x$ line. This indicates that the robust method is less likely to detect significance than the normal, since the outliers are down-weighted and do not pull the estimate away from the null hypothesis value as much. The non-outlier points below the line are expected, since the robust distributions are more conservative.

This real data application verifies that there are indeed differences between the p-values produced by the standard normal approach and by the alternative distributions. This is consistent across all three different effect sizes (MD, SMD and RoM).

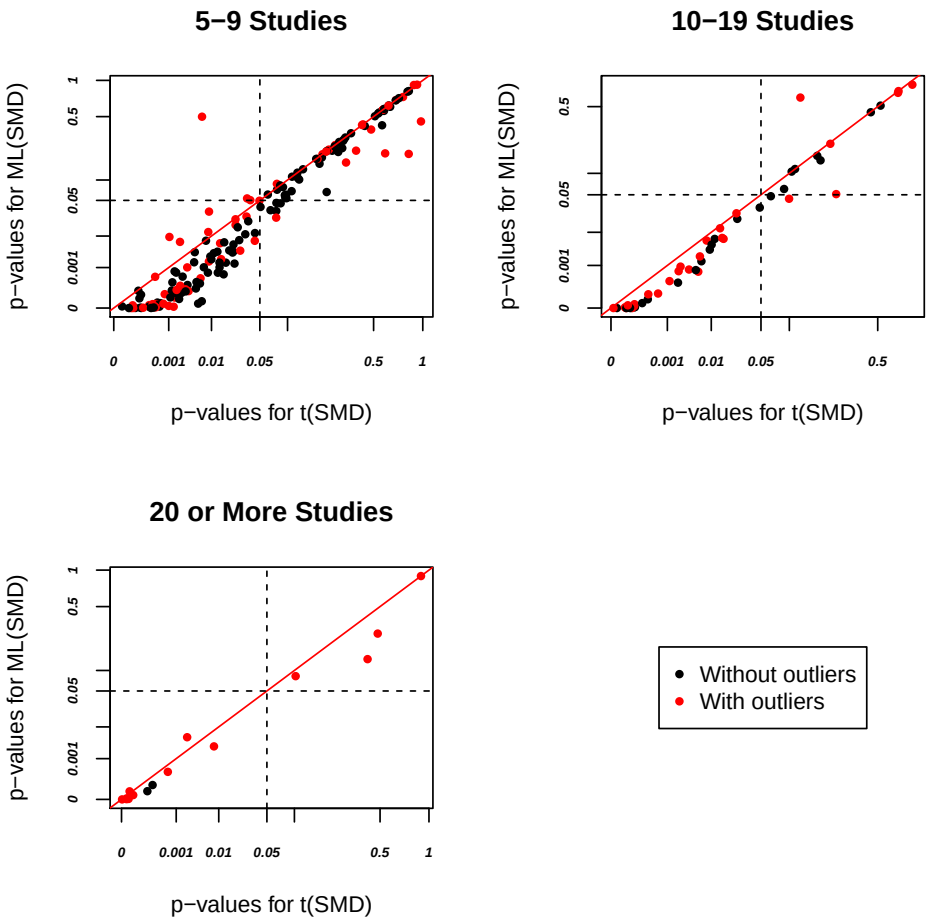

Figure 4.7: P-values for t vs Normal using ML and SMD

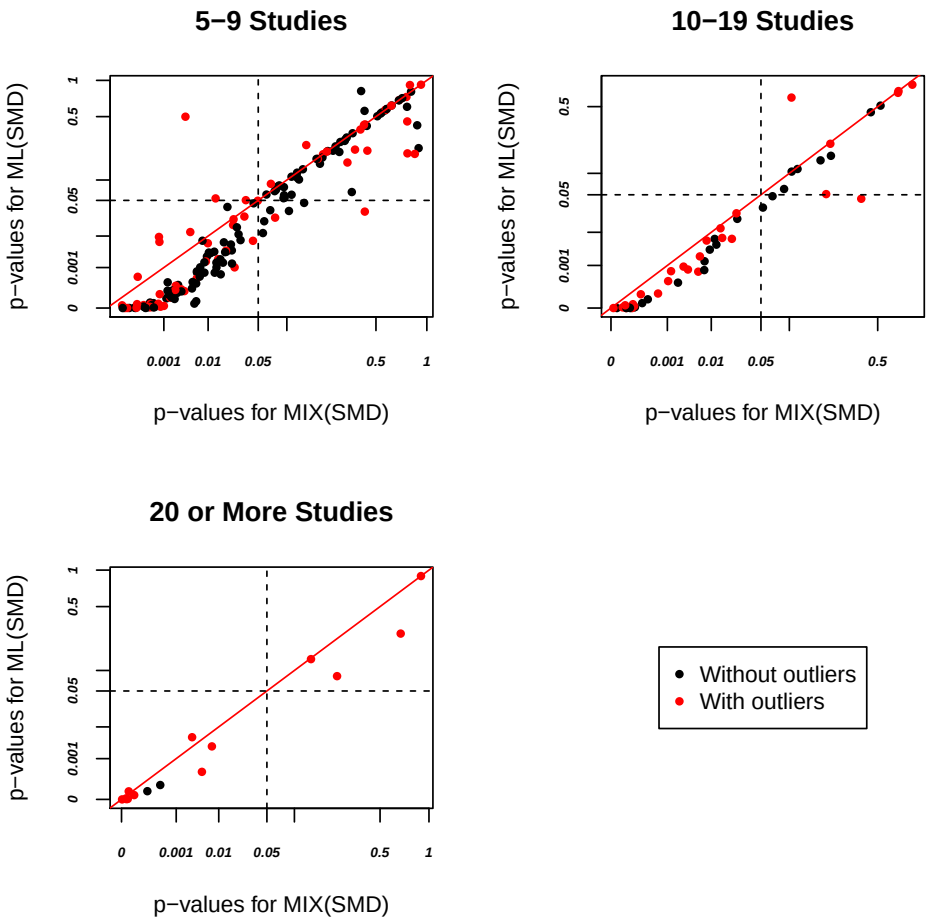

Figure 4.8: P-values for Mixture vs Normal using ML and SMD

These differences are often more severe in the presence of outliers, as well as in meta-analyses with a low number of studies. The t-distribution and mixture model perform very similarly to each other in this real data application.

Chapter 5

Discussion and Future Directions

5.1 Discussion

While meta-analysis is a highly utilized and useful tool for synthesizing clinical study results, the quality of a meta-analysis can be compromised in the presence of outliers. Outlier analysis is important in any data application and should be considered a mandatory step in meta-analysis. Recently, this problem has attracted attention, and many researchers believe that the random-effects should not be automatically assumed to be normally distributed. It is advantageous to consider using a heavier tailed distribution for the random-effects because a suspected outlier can still be included in the analysis without heavily influencing the outcome.

In Chapter 2, we presented the methods which can be used to perform a random-effects meta-analysis using the normal distribution, the t-distribution and the mixture model for the random-effects during the estimation of the model parameters.

Chapter 3 included a novel simulation study which explored the performance of each method in the presence of outliers with varying degrees of severity. Both alternative distributions proved to outperform the standard method in the presence of outliers and high heterogeneity. The conclusions hold true in the presence of outliers on both sides of the true mean.

The main advantage of using one of the alternative distributions is that there is no need to consider removing a suspected outlier. Meta-analyses typically have a small number of studies and it would be unfavourable to further reduce the size. A specific advantage of the mixture model is that each study is assigned a probability of being an outlier, which can be very useful in an analysis. The mixture model was also much more computationally reliable using *metaplus* as compared to the t-distribution which incurred more computational issues. Additionally, the mixture model supplies two estimates for the heterogeneity. This can provide some extra insight about how the "non-outlier" studies are different from the "outlier" studies. The t-distribution and the mixture model both proved to be an improvement over the normal, and they both behaved similar with respect to p-values. However, the mixture model out-performed the t- in terms of bias, mean squared error, coverage probability and Type I error, especially in the presence of the most extreme outliers. In the presence of symmetrically distributed outliers, the mixture model had the highest power and the best coverage for larger meta-analyses. With all of this in mind, a researcher may find the mixture model proposed by Beath (2014) to be the most favourable.

In Chapter 4, we verified the advantages of the robust distributions using a collection of meta-analyses from the Cochrane Collaboration. It was generally observed that the alternative distributions acted more conservatively than the normal distribution.

With careful consideration, the accuracy of a meta-analysis can be greatly improved in the presence of outliers by using a robust random-effects distribution such as the t-distribution or the proposed finite mixture model.

5.2 Future Directions

The simulation study in this paper has revealed that the t-distribution and mixture model proposed give more accurate analysis in the presence of outliers when the interest is to down-weight the outlying points. The mixture model has shown a particular strength in the presence of multiple outliers as well as very large outliers. A next step in this research would be to determine at what threshold the mixture becomes more favourable than the t-, and when the t- offers an improvement over the normal. This could be done by incrementally adding outliers to fixed data sets.

It would also be informative to study how the performance of the methods changes for more granular values of τ^2 and how this interacts with the number and size of outliers. As the true heterogeneity in the non-outlier data increases it can mask the effect of the outlier.

Learning from this simulation study, a future researcher may be interested in inserting a less variable random outlier. This can be achieved by shrinking the

variance of the distribution from which the outlier will be generated. These outliers will be more prominent, which reflects the type of real situation that leads to the consideration of outlier removal or robust distributions.

Investigating the sensitivity of the proposed methods in relation to the size, precision and distance of the outlier study would help in the understanding of when these particular distributions are no longer beneficial. In which case another alternative distribution, such as the sinh or Subbotin may warrant investigation.

Furthermore, as it is common to have a small number of studies in a meta-analysis, an investigation into the lower threshold of studies for which these methods still perform well would be advantageous.

Appendix A

Finite Mixture Model Supplementary Details

The following are supplementary details about the finite mixture model which are not used in the *metaplus* package (Beath, 2015) but are still valid methodologies.

The method of detecting outliers is based on a random variance shift outlier model (RVSOM) proposed by Gumedze and Jackson (2011) to fit a mixture model with the current observation having a different random-effect variance than the rest of the observations. The RVSOM for the j^{th} study is

$$y_j = \mu_j + \delta I_{i=j} + \mu + \epsilon_j \tag{A.1}$$

where μ_j and ϵ_j are the same as in equation 2.5, $\delta \sim N(0, \alpha)$ is an unknown random coefficient for $\alpha \geq 0$, and δ is only included for the j^{th} study. It is clear that the j^{th} study will have inflated variance as compared to all other studies.

A parametric bootstrap is used to generate test statistics under the null hypothesis and the likelihood ratio test determines if the observation should in fact be modeled with a higher random-effect variance. This is repeated for each observation in turn. The RVSOM assesses whether or not the point is an outlier depending on how shifted the new variance parameter is. The threshold for this is determined by ordering the bootstrap likelihood ratio tests and getting 100(1-α)th percentiles for up to third order statistics (Gumedze and Jackson, 2011). The overall effect estimate can be calculated by down-weighting the observations which were determined to be outliers according to how distant the shifted variance is (Beath, 2014; Gumedze and Jackson, 2011).

To identify the outliers Gumedze and Jackson (2011) use order statistics on the ordered set of likelihood ratio test results to determine which set of studies cross the determined threshold. Once the outliers are identified they can be down-weighted during the process of estimating the overall mean.

Beath (2014) uses posterior probabilities from standard literature such as McLachlan and Peel (2000) to determine the probability of a study being an outlier. The posterior probability of study i belonging to the m^{th} class is

$$p_{im} = \frac{\pi_m f_m(y_i|\mu, \tau_m)}{\sum_{j=1}^{2} \pi_j f_j(y_i|\mu, \tau_j)}. \tag{A.2}$$

The estimates for p_{im} and π_m are found simultaneously using the standard expectation-maximization algorithm, with a quasi-Newton method used for the maximization step, to iterate between the estimates.

The study is classified as an outlier when $p_m > 0.9$ where m corresponds to the

outlier group; this is considered strong evidence. The study is indeterminate for $0.1 \leq p_m \leq 0.9$. Finally, the study belongs to the non-outlier group for $p_m < 0.1$. The outliers and non-outliers will be weighted fully based on their classification, whereas the indeterminate studies will be weighted proportionally to their posterior probability (Beath, 2014).

It must be determined if the mixture model is an appropriate model for the given data, thus the fit of the standard model (equation 2.5) is tested against the fit of the mixture model (equation 2.15) (Beath, 2014). Since this requires testing a boundary condition, Beath (2014) refers to McLachlan (1987) in order to assess a test of a single normal distribution against a mixture of two normals. The purpose of this test is to indicate whether the studies should be modeled using a single standard distribution with a normal random-effect, or if a proportion of the studies should be modeled using a secondary standard distribution with a normal random-effect but with a larger random-effects variance.

In McLachlan (1987), the null hypothesis is the simple case of one element in the mixture distribution (ie. a single univariate density), and the alternative hypothesis has two elements in the mixture distribution (ie. a mixture of two normal densities). This is equivalent to saying that the null distribution assumes that all studies share one random-effects variance (Beath, 2014). This test is carried out using bootstrapping techniques for the log-likelihood ratio statistic (McLachlan, 1987). The bootstrap sample is generated under the null hypothesis from the original data and then both the standard model and mixture model are fitted to the data. The likelihood ratio test is computed using these two fits, and the comparison between the

observed value and the simulated value yields the p-value (Beath, 2014; McLachlan, 1987).

The bootstrap simulation of size k used by Beath (2014) is as follows:

1. Simulate a sample of size k of μ_i random-effects from Normal(0,$\hat{\tau}^2$) where $\hat{\tau}^2$ is the estimate of the heterogeneity from the standard (one-component) model
2. Sample n times with replacement from the observed within-study-variances (v_i)
3. Simulate the random study errors (ϵ_i) from Normal(0,v_i) and then determine the n bootstrap values of y_i using equation 2.5 ($y_i = \hat{\mu} + \mu_i + \epsilon_i$, where $\hat{\mu}$ is the estimated grand mean from the standard model)
4. Fit both the standard (null) model and the robust model to the meta-analysis of size n and perform the likelihood ratio test

Repeating K times will provide K values of $-2\log\,\lambda$ where λ is the likelihood ratio statistic, thus the distribution of $-2\log\,\lambda$ can be evaluated and p-value for the test can be found using the ordered bootstrap replicates of the likelihood ratio tests (McLachlan, 1987). Thus, it can then be determined if the one or two component mixture model should be used.

Appendix B

Tables for Symmetric Outliers

True μ	Number of	Outlier	Bias			MSE			Cov. Prob.			Conf. Width		
	Studies (k)	Shift (c)	Norm	t	Mix	Norm	t	Mix	Norm	t	Mix	Norm	t	Mix
0	10	2.50	0.00	0.00	0.00	0.15	0.16	0.17	0.93	0.92	0.90	1.51	1.47	1.37
		4.00	-0.02	-0.03	-0.02	0.28	0.24	0.26	0.94	0.92	0.89	2.03	1.82	1.49
		6.00	0.03	0.00	0.01	0.59	0.43	0.40	0.94	0.91	0.88	2.79	2.12	1.49
	20	2.50	0.00	0.00	-0.01	0.07	0.07	0.08	0.95	0.95	0.89	1.07	1.03	0.95
		4.00	0.02	0.00	0.01	0.14	0.10	0.09	0.94	0.94	0.91	1.44	1.21	0.99
		6.00	0.00	0.00	-0.01	0.26	0.17	0.11	0.95	0.91	0.91	2.05	1.31	0.91
	30	2.50	0.01	0.01	0.01	0.05	0.05	0.05	0.93	0.93	0.89	0.88	0.85	0.78
		4.00	0.01	0.02	0.01	0.09	0.06	0.06	0.95	0.94	0.92	1.20	0.97	0.79
		6.00	0.02	0.03	0.01	0.18	0.09	0.04	0.95	0.82	0.92	1.68	0.96	0.70
	40	2.50	0.00	0.00	0.00	0.04	0.04	0.04	0.94	0.93	0.91	0.76	0.73	0.68
		4.00	0.01	0.01	0.00	0.08	0.06	0.04	0.93	0.91	0.91	1.04	0.83	0.67
		6.00	0.00	0.02	0.00	0.15	0.10	0.03	0.93	0.66	0.94	1.44	0.75	0.61
0.5	10	2.50	0.01	0.01	0.01	0.15	0.15	0.17	0.93	0.93	0.90	1.50	1.42	1.34
		4.00	0.00	0.00	0.00	0.31	0.26	0.26	0.93	0.93	0.88	2.00	1.83	1.53
		6.00	0.04	0.04	0.03	0.59	0.42	0.40	0.95	0.92	0.91	2.82	2.13	1.50
	20	2.50	0.00	-0.01	-0.01	0.07	0.07	0.08	0.95	0.95	0.90	1.08	1.04	0.95
		4.00	0.00	0.01	0.00	0.14	0.11	0.10	0.95	0.94	0.90	1.45	1.24	1.01
		6.00	0.01	0.05	0.03	0.26	0.13	0.10	0.94	0.91	0.92	2.03	1.26	0.88
	30	2.50	-0.01	0.00	-0.01	0.04	0.04	0.05	0.96	0.95	0.92	0.87	0.84	0.78
		4.00	-0.01	0.01	-0.01	0.09	0.06	0.05	0.95	0.94	0.90	1.18	0.97	0.78
		6.00	0.00	0.01	-0.01	0.18	0.10	0.05	0.95	0.81	0.94	1.67	0.89	0.71
	40	2.50	0.00	0.00	0.00	0.03	0.03	0.04	0.95	0.95	0.91	0.75	0.73	0.68
		4.00	0.00	0.00	0.00	0.07	0.04	0.03	0.95	0.94	0.93	1.02	0.81	0.67
		6.00	0.00	0.03	0.00	0.13	0.08	0.03	0.95	0.73	0.93	1.45	0.79	0.61

Table B.1: Performance measures for $\tau^2 = 0.5$ and $p = 0.2$ with symmetric outliers

Number of	Outlier	Power for $\mu = 0.5$			Type I Error			Average p-value for $\mu = 0.5$		
Studies (k)	Shift (c)	Norm	t	Mix	Norm	t	Mix	Norm	t	Mix
10	2.50	0.33	0.35	0.38	0.07	0.08	0.10	0.26	0.25	0.23
	4.00	0.25	0.30	0.36	0.06	0.08	0.11	0.32	0.28	0.24
	6.00	0.20	0.35	0.38	0.06	0.11	0.13	0.36	0.24	0.21
20	2.50	0.46	0.49	0.53	0.05	0.06	0.10	0.17	0.16	0.15
	4.00	0.34	0.48	0.53	0.06	0.07	0.08	0.28	0.20	0.15
	6.00	0.21	0.58	0.62	0.05	0.13	0.10	0.35	0.15	0.10
30	2.50	0.63	0.68	0.71	0.07	0.07	0.11	0.10	0.09	0.09
	4.00	0.39	0.62	0.67	0.05	0.09	0.09	0.21	0.12	0.09
	6.00	0.25	0.71	0.76	0.04	0.21	0.09	0.32	0.10	0.06
40	2.50	0.74	0.76	0.78	0.06	0.07	0.09	0.06	0.06	0.05
	4.00	0.51	0.72	0.79	0.07	0.11	0.09	0.16	0.09	0.05
	6.00	0.30	0.73	0.87	0.07	0.36	0.07	0.27	0.10	0.03

Table B.2: Hypothesis measures for $\tau^2 = 0.5$ and $p = 0.2$ with symmetric outliers

Bibliography

The Cochrane Library, Issue 1. Chichester: Wiley (2008).

Baker, R. and Jackson, D. (2007). A new approach to outliers in meta-analysis. *Health Care Management Science*, 11(2):121–131.

Beath, K. (2014). A finite mixture method for outlier detection and robustness in meta-analysis. *Research Synthesis Methods*, 5:285–293.

Beath, K. (2015). *metaplus: Robust Meta-Analysis and Meta-Regression*. R package version 0.7-4.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics*. John Wiley & Sons, New York.

Bland, J. M. and Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, 8(2):135–160.

Borenstein, M., Hedges, L. V., Higgins, J., and Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. Wiley, Hoboken.

Borenstein, M., Hedges, L. V., Higgins, J., and Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, 1(2):97–111.

Brockwell, S. and Gordon, I. (2001). A comparison of statistical methods for meta-analysis. *Statistics in Medicine*, 20(6):825–840.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, New York.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188.

Filzmoser, P. (2005). Identification of multivariate outliers: A performance study. *Austrian Journal of Statistics*, 34(2):127–138.

Fioravanti, M. and Yanagi, M. (2005). Cytidinediphosphocholine (CDP-choline) for cognitive and behavioural disturbances associated with chronic cerebral disorders in the elderly. *The Cochrane Library*, (2).

Friedrich, J. O., Adhikari, N. K. J., and Beyene, J. (2011). Ratio of means for analyzing continuous outcomes in meta-analysis performed as well as mean difference methods. *Journal of Clinical Epidemiology*, 64(5):556–564.

Gumedze, F. N. and Jackson, D. (2011). A random effects variance shift model for detecting and accommodating outliers in meta-analysis. *BMC Medical Research Methodology*, 11(1):19.

Hardin, J. and Rocke, D. M. (2004). Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. *Computational Statistics and Data Analysis*, 44(4):625–638.

Hardy, R. J. and Thompson, S. G. (1996). A likelihood approach to meta-analysis. *Statistics in Medicine*, 15(June 1995):619–629.

Hedges, L. V. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press, New York.

Jackson, D., Bowden, J., and Baker, R. (2010). How does the DerSimonian and Laird procedure for random effects meta-analysis compare with its more efficient but harder to compute counterparts? *Journal of Statistical Planning and Inference*, 140(4):961–970.

Julious, S. and Whitehead, A. (2012). Investigating the assumption of homogeneity of treatment effects in clinical studies with application to meta-analysis. *Pharmaceutical Statistics*, 11(1):49–56.

Knight, N. L. and Wang, J. (2009). A Comparison of Outlier Detection Procedures and Robust Estimation Methods in GPS Positioning. *The Journal of Navigation*, 62(04):699.

Kontopantelis, E. and Reeves, D. (2012). 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):409–26.

Lawlor, D. a. and Hopker, S. W. (2001). The effectiveness of exercise as an intervention in the management of depression: systematic review and meta-regression analysis of randomised controlled trials. *BMJ (Clinical Research ed.)*, 322(7289):763–7.

Lee, K. and Thompson, S. G. (2008). Flexible parametric models for random-effects distributions. *Statistics in Medicine*, 27:418–434.

Marinho, V. and Higgins, J. (2003). Fluoride toothpastes for preventing dental caries in children and adolescents. *The Cochrane Collaboration*, (1).

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

McLachlan, G. J. (1987). On the likelihood ratio test statistic for the number of components in a normal mixture of unequal variances. *Journal of the Royal Statistical Society*, 36(3):318–324.

O'Rourke, K. (2007). An historical perspective on meta-analysis: dealing quantitatively with varying study results. *Journal of the Royal Society of Medicine*, 100(12):579–582.

Peña, D. and Prieto, F. J. (2001). Multivariate Outlier Detection and Robust Covariance Matrix Estimation. *Technometrics*, 43(3):286–310.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rousseeuw, P. J. and Driessen, K. V. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, 41(3):212–223.

Sackett, D. L. (1997). Evidence-based medicine. *Seminars in Perinatology*, 21(1):3–5.

Simpson, R. J. S. and Pearson, K. (1904). Report on Certain Enteric Fever Inoculation Statistics. *British Medical Journal*, 2(2288):1243–1246.

Viechtbauer, W. (2005). Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model. *Journal of Educational and Behavioral Statistics*, 30(3):261–293.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3):1–48.

Viechtbauer, W. and Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, 1(2):112–125.

Yong, X., Ward, R. K., and Birch, G. E. (2008). Robust Common Spatial Patterns for EEG signal preprocessing. *Conference proceedings : 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society*, 2008:2087–2090.