

University of Warwick institutional repository: <http://go.warwick.ac.uk/wrap>

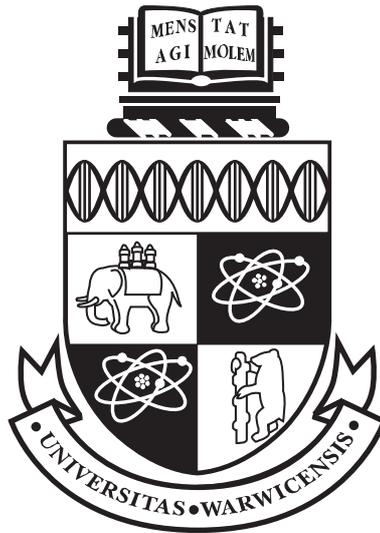
**A Thesis Submitted for the Degree of PhD at the University of Warwick**

<http://go.warwick.ac.uk/wrap/3767>

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.



# Sensitivity Methods for Publication Bias in a Meta-Analysis

by

Paul Malley

Thesis

Submitted to the University of Warwick

for the Degree of

Doctor of Philosophy

Department of Statistics

October 2009

THE UNIVERSITY OF  
**WARWICK**

# Contents

List of Figures . . . . .	ii
List of Tables . . . . .	ii
Acknowledgements . . . . .	iii
Declarations . . . . .	iv
Abstract . . . . .	v
<b>1 Introduction</b>	<b>1</b>
<b>2 A Literature Review</b>	<b>4</b>
2.1 Systematic Reviews . . . . .	4
2.2 Meta-Analysis . . . . .	7
2.3 Issues in Meta-Analysis . . . . .	8
2.3.1 Heterogeneity . . . . .	8
2.3.2 Publication bias . . . . .	8
2.3.3 Heterogeneity and its interaction with publication bias . . . . .	9
2.4 A Brief History of Publication Bias . . . . .	11
2.4.1 The use of selection functions in meta-analysis . . . . .	15
2.5 A Review of Recent Research Investigating Publication Bias . . . . .	18
2.5.1 Research by Duval and Tweedie . . . . .	18
2.5.2 Research by Copas and Jackson . . . . .	22
2.5.3 Research by Henmi <i>et al.</i> . . . . .	24
<b>3 Case Study: A Meta-Analysis Investigating the Effects of Environmental Tobacco Smoke</b>	<b>27</b>
3.1 Environmental Tobacco Smoke and Lung Cancer . . . . .	27
3.2 The Hackshaw <i>et al.</i> Paper . . . . .	29
3.3 Existing Analyses . . . . .	33
3.4 Promoting a Routine Investigation of Publication Bias in Meta-Analysis	38
3.4.1 An updated analysis by Taylor <i>et al.</i> . . . . .	38
3.4.2 A recommended approach in meta-analysis . . . . .	40
3.4.3 Methods for investigating sources of heterogeneity . . . . .	54

3.4.4	Conclusions and further remarks . . . . .	59
<b>4</b>	<b>A Robust P-value in Meta-Analysis with Publication Bias</b>	<b>62</b>
4.1	Motivation . . . . .	62
4.2	Using the Permutation Test in a Meta-Analysis . . . . .	63
4.3	A P-value Using a Normal Approximation . . . . .	65
4.4	Numerical Examples . . . . .	67
4.4.1	Cholesterol lowering dataset . . . . .	67
4.4.2	Passive smoking dataset . . . . .	74
4.5	A Comparison Between the Permutation Test and the Linear Regression Test . . . . .	76
4.6	Concluding Comments . . . . .	78
<b>5</b>	<b>Applications of Parametric Selection Functions in Meta-Analysis</b>	<b>80</b>
5.1	Introduction . . . . .	80
5.2	Using Parametric Selection Functions . . . . .	82
5.3	Numerical Examples . . . . .	85
5.3.1	Passive smoking dataset . . . . .	85
5.3.2	Prophylactic corticosteroids dataset . . . . .	85
5.3.3	The selection functions . . . . .	86
5.3.4	Example 1: passive smoking dataset . . . . .	88
5.3.5	Example 2: corticosteroids dataset . . . . .	91
5.4	Generalising the Parametric Selection Functions Approach . . . . .	93
5.4.1	Description of the model . . . . .	94
5.4.2	Hypothesis tests and confidence intervals . . . . .	96
5.4.3	Example: passive smoking dataset with a fixed effects model . . . . .	97
5.4.4	Alternative model . . . . .	100
5.4.5	Example continued: passive smoking dataset . . . . .	101
5.4.6	Further examples . . . . .	103
5.5	Assessing the Effectiveness of the Bounds for Confidence Intervals . . . . .	106
5.5.1	Examples . . . . .	106
5.5.2	Discussion . . . . .	111

5.6	Concluding Comments . . . . .	113
<b>6</b>	<b>A New Likelihood Method for Monotonic Selection Functions</b>	<b>118</b>
6.1	Setting Up the Step Function . . . . .	120
6.2	A Description of the Algorithm . . . . .	124
6.3	An Example: Passive Smoking . . . . .	126
6.4	A Comparison with Parametric Selection Functions . . . . .	129
6.5	Concluding Comments . . . . .	132
<b>7</b>	<b>Summary and Conclusions</b>	<b>134</b>
	<b>Appendix A1 - Statistics in Medicine (2008) Paper</b>	<b>139</b>
	<b>Appendix A2 - S-Plus Code for the Bounds Method</b>	<b>140</b>
	<b>Appendix A3 - S-Plus Code for the Step Selection Function</b>	<b>144</b>

## List of Figures

3.1	Funnel plot for the 2007 Taylor meta-analysis. The horizontal lines represent each study's confidence interval. . . . .	45
3.2	Forest plot for the 2007 Taylor meta-analysis. The horizontal lines represent each study's confidence interval. . . . .	47
3.3	Radial plot for the 2007 Taylor meta-analysis. . . . .	48
3.4	Confidence limits using the Bounds method for the 2007 Taylor meta-analysis. . . . .	51
3.5	Funnel plot once the Trim and Fill Method is applied to the Taylor dataset. . . . .	53
3.6	Pooled odds ratios with 95% confidence intervals for meta-analyses between 1986 and 2007. . . . .	60
4.1	Cholesterol lowering dataset: funnel plot. . . . .	68
4.2	Cholesterol lowering dataset: permutation distribution of $\hat{\theta}$ . . . . .	69
4.3	Cholesterol lowering dataset: radial plot. . . . .	70
4.4	Cholesterol lowering dataset: radial plot when assuming a random effects model. . . . .	71
4.5	Cholesterol lowering dataset: permutation distribution of $\hat{\theta}$ when assuming a random effects model. . . . .	72
4.6	Passive smoking dataset: permutation distribution of $\hat{\theta}$ . . . . .	74
4.7	Passive smoking dataset: radial plot. . . . .	75
5.1	Corticosteroids dataset: funnel plot. . . . .	86
5.2	Passive smoking dataset: $\hat{\theta}$ and 95% confidence intervals of $\theta$ when assuming different selection functions $a_1, \dots, a_6$ for a range of $p$ . . . . .	89
5.3	Corticosteroids dataset: $\hat{\theta}$ and 95% confidence intervals of $\theta$ when assuming different selection functions $a_1, \dots, a_6$ for a range of $p$ . . . . .	92
5.4	Passive smoking dataset: profile log-likelihood for $p$ when assuming selection function $a_7$ . . . . .	98
5.5	Passive smoking dataset: contour plot of the profile log-likelihood for $p$ and $\theta$ when assuming selection function $a_7$ . . . . .	99

5.6	Passive smoking dataset: $\hat{\theta}$ and 95% confidence intervals against $p$ when assuming selection function $a_7$ . . . . .	99
5.7	Passive smoking dataset: contour plot of the profile log-likelihood for $p$ and $\theta$ when assuming the alternative version of selection function $a_7$ . . . . .	102
5.8	Passive smoking dataset: $\hat{\theta}$ against the $P$ -value when assuming the alternative version of selection function $a_7$ . . . . .	102
5.9	Passive smoking dataset: grid of different plots when a random effects model is used. . . . .	104
5.10	Corticosteroids dataset: grid of different plots when a fixed effects model is used. . . . .	105
5.11	Passive smoking dataset: $\hat{\theta}$ , 95% confidence intervals when assuming different selection functions $a_1, \dots, a_6$ and the bounds for a range of $p$ . . . . .	108
5.12	Passive smoking dataset: $\hat{\theta}$ , 95% confidence intervals when assuming selection function $a_7$ and the bounds for a range of $p$ . . . . .	109
5.13	Corticosteroids dataset: $\hat{\theta}$ , 95% confidence intervals when assuming different selection functions $a_1, \dots, a_6$ and the bounds for a range of $p$ . . . . .	110
5.14	Corticosteroids dataset: $\hat{\theta}$ , 95% confidence intervals when assuming selection function $a_7$ and the bounds for a range of $p$ . . . . .	111
6.1	Selection function $a(t)$ . . . . .	121
6.2	Selection function $a(t)$ with step function $a^*(t)$ . . . . .	123
6.3	Profile likelihood for different intervals of $p$ when assuming selection function $a^*$ . . . . .	126
6.4	95% confidence intervals of $\theta$ given a range of $p$ when assuming selection function $a^*$ . The bounds from the Bounds method have also been added. . . . .	127
6.5	Examples of $a^*(t)$ for a selection of different values of $p$ . . . . .	129
6.6	An example of $a^*(t)$ plotted against other selection functions $a_1, a_3$ and $a_5$ . . . . .	131

# List of Tables

3.1	Hackshaw dataset, 1982-1997: epidemiological studies of the risk of lung cancer in lifelong non-smokers whose spouses smoked relative to the risk in those whose spouses do not smoke. . . . .	36
3.2	Calculations for the fixed effects model and random effects model for the Hackshaw and Taylor meta-analyses. . . . .	43
3.3	Summary statistics for the Hackshaw and Taylor meta-analyses. . . .	44
3.4	S-Plus regression output for the Taylor dataset. . . . .	49
3.5	Taylor dataset, 1982-2006: epidemiological studies of the risk of lung cancer in female lifelong non-smokers whose spouses smoked relative to the risk in those whose spouses do not smoke. . . . .	56
4.1	Cholesterol lowering dataset, where $y_i$ is the log odds ratio. . . . .	73
5.1	Notation for the selection functions used. . . . .	87
5.2	Passive smoking dataset: table of approximate number of missing studies for the confidence intervals to include zero when assuming different selection functions $a_1, \dots, a_6$ . . . . .	90
5.3	Corticosteroids data: table of approximate number of missing studies for the confidence intervals to include zero when assuming different selection functions $a_1, \dots, a_6$ . . . . .	93
6.1	Average values of $p_{a^*}$ and corresponding 95% confidence intervals of $\theta$ . . . . .	128
6.2	Summary of the results for the parametric selection functions $a_1, \dots, a_7$ , $a^*$ and the Bounds method for comparison. . . . .	130

## **Acknowledgements**

I would like to first and foremost thank my supervisor Professor John Copas. He has given me endless support and advice throughout my studies, and for that I give to him my limitless gratitude. I offer my thanks also to all the staff within the Department of Statistics at the University of Warwick, with special acknowledgement to Paula Matthews for her help throughout my time at Warwick.

I thank my loving parents, Pauline and Allan, to whom I dedicate this thesis. Mum, Dad, thank you so much for all the love, support and encouragement that you have given me. I simply would not have achieved my goals without you both.

Finally, thank you to all my family and friends, new and old, who continue to be that joyful counterpoint to all those hours of studying.

## Declarations

I declare that this thesis is my own work, except where explicitly stated, and that it has not been submitted elsewhere.

## Abstract

Meta-analysis is the statistical part of a systematic review. Many researchers have used selection functions to model publication bias in a meta-analysis. The main problem with this approach is that it is impossible to verify that the selection function truly represents the selection process, and so the use of selection functions can only be seen as part of a sensitivity analysis. In this thesis we present new methods that involve selection functions that aim to make as few strong assumptions about selection as possible, including the use of a non-parametric permutation test, and the use of a step selection function. We also investigate the use of parametric selection functions and suggest how researchers could use these as part of a sensitivity analysis, by looking at a range of plausible values for the overall selection probability. As part of this sensitivity analysis, we assess the effectiveness of the Bounds method as presented by Henmi *et al.* Throughout the thesis we illustrate all methods with numerical examples, including a meta-analysis investigating the effects of environmental tobacco smoke on the risk of lung cancer in non-smokers.

# 1 Introduction

Meta-analysis is the statistical part of a systematic review, and due to the increased awareness in the importance of evidence based research, it is a very active area for researchers. Publication bias is one of the most prominent issues in meta-analysis, and there currently exist numerous methods to try to model and account for it. The structure of the thesis is as follows.

Chapter 2 will include a literature review of meta-analysis. This chapter will begin by introducing the broader concepts and issues relevant to meta-analysis. Also discussed will be an overview of a variety of methods that have been developed to model and adjust for publication bias in a meta-analysis. Special attention will be given to the Trim and Fill method by Duval and Tweedie [27], and the Bounds method by Henmi *et al.* [47], as these methods will be used in subsequent chapters.

Chapter 3 will include a case study of a meta-analysis investigating the effects of environmental tobacco smoke. The effect of smoking on a person's health has rightfully received a lot of medical attention and continues to do so. This chapter will provide an overview of the meta-analysis conducted by Hackshaw *et al.* in 1997 [40] that aimed to synthesize together research relevant to this topic. Subsequent analyses that other researchers have carried out to demonstrate techniques of handling publication bias in a meta-analysis relevant to the Hackshaw dataset are discussed. In 2007, Taylor *et al.* [85] carried out an updated meta-analysis investigating environmental tobacco smoke and lung cancer. In this chapter, we will use their updated dataset for two reasons: the first being to promote routine investigations of publication bias

in a meta-analysis, with the aid of simple recommended techniques, both graphical and statistical. The second reason for analysing the Taylor dataset is that it will be interesting to discuss how the main conclusions of the 2007 Taylor meta-analysis may have changed compared to those of the 1997 Hackshaw meta-analysis. S-Plus code relevant to the calculation of the Bounds method of Henmi *et al.* will also be presented in the Appendix. While the theory for the Bounds method might appear highly statistical, it is hoped that the S-Plus code provided can be easily adapted for others with their own research under similar settings.

One standard approach for modelling publication bias in a meta-analysis is to make assumptions about the selection process. These assumptions are unfortunately very difficult if not impossible to fully verify. Chapter 4 presents a robust P-value using the idea of a permutation test. An alternative approximation to this P-value is also presented. The aim of this new method is to avoid making strong assumptions about the selection process. Numerical examples are discussed to demonstrate the method, including the aforementioned passive smoking example, as well as a different example concerning the effectiveness of cholesterol lowering interventions. Limitations concerning the use of this robust P-value will also be discussed. It should be noted that the content in Chapter 4 forms the basis for a paper that was co-authored with J.B. Copas and subsequently accepted for publication in the journal *Statistics in Medicine* in 2008 [20], a copy of which is included in the Appendix.

Chapter 5 will present a general method for using parametric selection functions in meta-analysis. Selection functions (or weight functions as they are also known) describe the probability of a study being selected in a meta-analysis, often conditional upon the study outcome, study size and some adjustable parameter,  $\beta$ . The choice of selection function is entirely arbitrary, and we know little about the value of  $\beta$  since the selection process is unknown. Chapter 5 will describe the maximum likelihood approach and will propose a sensitivity analysis, where we re-calibrate the various selection functions into an interpretable quantity,  $p$ , representing the overall probability of selection, and investigate a plausible range of values of  $p$ . Chapter 5

will include various examples of selection functions where  $\beta$  is a scalar parameter, and this chapter will also generalise the theory to an example where  $\beta$  is a vector of parameters, namely the Copas and Shi selection function [21]. We conclude this chapter of work by assessing the effectiveness of the Bounds method by Henmi *et al.* by comparing the confidence intervals derived from the use of the parametric selection functions with the bounds when the Bounds method is used. Various numerical illustrations will be used throughout this chapter, including the case study example from Chapter 3 and an example concerning the use of prophylactic corticosteroids in cases of premature birth.

Chapter 6 will present a new maximum likelihood method for monotonic selection functions, aiming to make as few assumptions about the selection process as possible. Previous researchers have used the idea of a step selection function when attempting to model publication bias in a meta-analysis, for which Lane and Dunlap [53], and Vevea and Hedges [90] are early examples. One of the main criticisms is that some of these methods require making very strong assumptions about the selection function, for example, where to place the steps in the function, which can not be easily verified. The new method proposed here in Chapter 6 will use a step function in its solution, under few assumptions. An algorithm on how to implement this method in practice will be discussed, and will be illustrated with the aid of the case study concerning environmental tobacco smoke and lung cancer. In addition, the S-Plus code that will be used to implement this method will be given in the Appendix.

## 2 A Literature Review

This chapter starts our journey into publication bias by first introducing the wider definitions and concepts regarding systematic reviews and meta-analysis. A brief history of publication bias is also included, which will provide an interesting account of the major landmarks in the development of this branch of statistics. Finally, this chapter will focus upon the use of parametric selection functions in meta-analysis, including techniques that will be used often in subsequent chapters. One final point before we begin is that Sutton *et al.* [80] was most useful in providing an insightful introduction to systematic reviews and meta-analysis. While the following attempts to highlight the main points, see their recommended text for more details. Another recommendation is the second edition text edited by Cooper *et al.* [14] in 2009 providing a relevant update into the numerous areas of literature reviews and meta-analyses.

### 2.1 Systematic Reviews

The increase in demand for scientific knowledge relevant to health care over the last century has been considerable, spread across a variety of reports and journals. When one searches for the available evidence for a particular area of interest, one normally finds a variety of studies which have used different methods, are of varying quality, and quite possibly will have contradictory conclusions. It can therefore be difficult for the researcher to make sense of the research that they find.

Further to this, there is an increasing need to ensure medical procedures and health policies are based on evidence that is reliable and relevant. Archie Cochrane, a

British epidemiologist, in 1972 spoke of the need for “rigorous evaluations to inform choices made by policy makers” which eventually led to the Cochrane Collaboration in 1993 - a general international initiative that summarises the results of health care experimental evaluations. Briefly, the Cochrane Collaboration is a group of volunteers responsible for ensuring that well conducted reviews are readily available to researchers or practitioners to learn from them.

Evidence based medicine (EBM) refers to the explicit use of current best evidence in making decisions about care of individual patients. It is critical for EBM to have the structure in place to gather together all the evidence in such a way that it is in a usable form by the practitioners. This is the role of the systematic review. Systematic reviews use well-defined and rigorous methods to “identify, critically appraise, include and synthesize relevant research studies”.

A brief summary, as given by Sutton *et al.* [80], of the structure of a systematic review is given below.

1. Specification of the objectives, hypotheses and methods of the systematic review before the study is undertaken.
2. Compilation of relevant primary studies, having searched for all potentially relevant data, documenting all search methods and sources, based on clearly stated *a priori* specifications.
3. Assessment of the methodological quality of the set of studies.
4. Identification of definitions of outcome, explanatory and confounding variables compatible as far as possible with all primary studies.
5. Extraction of estimates of outcome measures and of study and subject characteristics in a standardized way from primary study documentation.
6. **Meta-analysis** using appropriate methods and models, exploring and allowing for all important sources of variation. Confidence intervals around pooled point estimates should be included.

7. When statistical aggregation is inappropriate (for example, if the data is too sparse, or of low quality, or heterogeneous - a concept introduced later), a qualitative summary should be performed, and the formal meta-analysis omitted.
8. Exploration of the robustness of the results of the systematic review, including the impact of study quality; likelihood and impact of **publication bias**; implications of the effect of different model selection strategies; exploration of a reasonable range of values of missing data from studies with uncertain results.
9. Clear presentation of key aspects of all the above stages in the study report, to enable critical appraisal and replication of the systematic review, including a table of key elements of each primary study. Graphical displays can assist interpretation where appropriate.
10. Limitations of the primary studies and the systematic review should be appraised. Any clinical or policy recommendations should be practical and explicit, making clear the research evidence on which they are based. Proposal of a future research agenda should include clinical and methodological requirements as appropriate.

By following these guidelines on systematic reviews, as demonstrated by the Cochrane Collaboration, researchers have an arena to collaborate work, avoid duplication, keep up-to-date and allow everyone to access the combined results. A wide variety of publications are available guiding researchers on how best to perform literature searches of systematic reviews. A good reference by Egger *et al.* provides an insight into comprehensive literature searches, discussing the importance of such searches and the need to assess the quality of the reviews [33]. Another recommendation would be the text on systematic reviews by Egger *et al.* [32] which offers an approachable introduction.

## 2.2 Meta-Analysis

Meta-analysis is the application of statistical methods to systematic reviews. To quote Sutton *et al.* [80],

“meta-analysis is the part of the review process that concerns itself with the analysis of the data extracted from the primary research included, uses quantitative methods to explore the heterogeneity of study results, estimates overall measures of association or effect and assess the sensitivity of the results to possible threats to validity such as publication bias and study quality.”

The concepts upon which meta-analysis is based are not necessarily new ideas. A known example of meta-analysis dates as far back as 1904 when Karl Pearson investigated divergent results from small studies of the effectiveness of inoculation against typhoid fever [59]. The aim of the meta-analysis was to overcome the problem of reduced statistical power in studies with small sample sizes. Meta-analysis has been widely used in the social sciences for over sixty years, for example in the fields of education, sociology and psychology. Researchers such as Glass, Schmidt and Hunter are names associated with advancements in meta-analysis. Glass was in fact the first to use the term meta-analysis in a statistical setting [36]. From the 1980's onwards, it has been used frequently in the health care field.

In spite of the obvious advantages that meta-analysis can provide, there are a few drawbacks. Some statisticians and researchers have criticised the use of systematic reviews. This could be the result of poor practice or inappropriate use of statistics when the assumptions, on which the methods are based, are not satisfied. Another limitation is that meta-analysis of qualitative data is poorly developed. When arriving at decisions, for example with a particular health care policy, it sometimes may be necessary to incorporate informal observations from clinical and patients' experiences.

## 2.3 Issues in Meta-Analysis

The use of meta-analysis and systematic reviews has increased considerably over the last few decades. It is a very active area of research, especially related to medical applications, which has seen substantial development in the last fifteen to twenty years. The predominant reason why the area has received so much attention recently in the health care field is most likely because of the growing awareness of the need for evidence based medicine in policy making. There are however many issues in meta-analysis. Two of the main themes which will be discussed here are **heterogeneity** and **publication bias**.

### 2.3.1 Heterogeneity

Given a research question of interest, results from single studies are collected together and used to estimate an overall effect. Those estimates will almost certainly differ amongst the various studies. If sampling error alone is responsible for this variation (since individual estimates will inevitably vary by chance), then this means that the true effect is the same in each study. In this scenario, the effect estimates would be considered *homogeneous*. If, on the other hand, there are kinds of systematic differences between studies which causes variations that chance alone can not explain, then the effect estimates are called *heterogeneous*. Take as an example a meta-analysis including studies from all parts of the world, spread over a significant period of time, say decades. It seems unlikely that the individual studies would be estimating the same effect. Heterogeneity is therefore an important problem in meta-analysis. It might not be appropriate to combine the study results if the data are “too heterogeneous”, resulting in an overall estimate of little use.

### 2.3.2 Publication bias

Another important problem in meta-analysis is publication bias. This is the bias caused by the generally accepted belief that research with statistically significant results is more likely to be submitted for publication than those studies with non-

significant results [29]. The non-random sampling that is taking place will therefore create bias and in turn pose a serious threat to the validity of the results of the meta-analysis. The sample of results will almost certainly misrepresent the research findings, usually creating over-optimistic conclusions. Clearly publication bias is an issue that needs much consideration. Since a meta-analysis is the statistical part of a systematic review which is in turn critical to EBM, it is important to take into account any biases that may occur in an unrepresentative sample of studies.

Publication bias is not the only type of bias that one encounters when carrying out a systematic review. A non-exhaustive list of examples of other types of publishing and reporting related biases are given below.

1. Retrieval bias - the bias incurred through the process of obtaining unpublished studies.
2. Pipeline effects - the effects of waiting for unpublished studies to become published.
3. The subjective reporting of results may be a consequence of the opinions of the investigator.
4. Duplication of reporting results when, for example, authors submit their results to different journals.
5. Language bias - possible exclusion of studies from non-English speaking countries.
6. MSc dissertations and PhD theses might not get published.
7. Suppression of studies due to conflicts in personal or political interests.

### **2.3.3 Heterogeneity and its interaction with publication bias**

It is widely known that the two issues of heterogeneity and publication bias are not mutually exclusive. Publication bias is one of the more important issues with meta-analysis, however there is a danger that publication bias may be misdiagnosed or

over-estimated within a meta-analysis when it may in fact be an issue of heterogeneity. Funnel plots (discussed in more depth in Section 3.4.2) are routinely used to detect the presence of possible publication bias by investigating asymmetrical patterns within the plot. It is widely acknowledged that heterogeneity is an alternative cause of asymmetry, and so when heterogeneity exists within the data, it becomes difficult to determine whether this is the cause of the asymmetry or if publication bias is the cause, or both [80]. This issue can make the use of funnel plots unreliable.

The importance of investigating heterogeneity within any meta-analysis is paramount, not just how to potentially adjust for it, but to consider the underlying causes of the between-study variation. Research by Peters *et al.* [62] highlighted the need to consider heterogeneity, publication bias and their interaction. Results showed that ignoring heterogeneity when assessing for publication bias can be misleading, and it becomes difficult to disentangle the effects of the two issues when the number of studies within the meta-analysis is small.

There are many causes of heterogeneity, summarised as the following (see [80] and [2] for a good discussion):

1. The underlying cause may be due to chance.
2. The scale used to measure the treatment effect.
3. Treatment characteristics, such as dose levels of the intervention under investigation.
4. Patient-level covariates may provide insight to the cause.
5. Characteristics of the design and conduct of the study, including the quality of the study.
6. The length of follow-up of a trial may influence the treatment effect size.

If there still exists significant amount of heterogeneity that remains unexplained even after considering various possible causes for it, then investigators must ask whether

it is appropriate or not to pool together the various studies. There are numerous approaches for investigating and dealing with sources of heterogeneity. These will be discussed in Section 3.4.3 later on in this thesis.

## 2.4 A Brief History of Publication Bias

The generally agreed notion of publication bias that we are familiar with today dates back over a period of about fifty years with one of the earliest examples by Sterling in 1959 [77]. Sterling referred to research yielding non-significant results not being published. Twenty years later, the term *file drawer problem* was first coined by Rosenthal [70]. It literally refers to researchers filing away studies with negative outcomes, and was one of the earliest methods of assessing for publication bias using a fail safe approach. Another notable reference in this area was Orwin in 1983 [58]. A fail safe approach involved estimating the number of unpublished studies that would threaten the validity of a significant overall estimate from a meta-analysis. Specifically, this fail safe estimate was the number of null results necessary to average the overall estimate to some specified level of significance, say a P-value of greater than 0.05. Around the same time as the file drawer problem did the term *publication bias* first get used in a statistical context [75] in 1980.

Throughout the 1980's and 1990's the issue of publication bias became increasingly well known, and measures to identify and take into account the effects of publication bias were being explored. One such approach involved the use of selection functions to model the probability of a study being published. Pioneering pieces of research included those by Iyengar and Greenhouse in 1988 [50] and Hedges in 1992 [44]. Both offered a more sophisticated statistical approach to taking publication bias into account compared to the fail safe approach, by using a maximum likelihood approach to model the selection process. Statistical tests were presented that could be used to see the effect of assuming no selection compared to assuming selection bias was present.

Continuing this maximum likelihood based approach were landmark research papers by Hedges and Vevea, [90] and [46]. Both are highly statistical papers, the first of which in 1995 presented a general linear model for estimating the effect size when assuming selection is modelled with one-tailed P-values. Hedges and Vevea presented a test for the presence of publication bias and also suggested how to add a correction to the effect estimate if publication bias exists. Related to this was a paper in 1996, presenting a method on how to deal with publication bias focusing on a random effects model, again with the aid of one-tailed P-values. Since the use of selection functions is central to the subsequent chapters, a more in depth discussion is given in Section 2.4.1.

In 1994, Begg and Mazumdar [3] presented the first rank correlation test for assessing the presence of publication bias. Briefly, they proposed testing the independence of study variance and effect size using the non-parametric Kendall's method [52]. Begg's method was praised for its simplicity, and generally was considered to be quite a powerful test given the number of studies in the meta-analysis was large. However, the test does have low power if the number of studies is small (considered to be less than 25) [78].

Another important landmark in the history of publication bias was a paper published in 1997 by Egger *et al.* who presented the Egger test [31]. The Egger test is a regression based test used to assess funnel plot asymmetry. The funnel plot is arguably one of the simplest and most commonly used graphical plots in meta-analysis, plotting (as an example, since there are variations) treatment effect against study precision (defined as the reciprocal of the standard error). Therefore the Egger test was one of the earliest examples of trying to formally assess the output from a funnel plot, rather than just basing it on subjective judgement. More details and an example of the Egger test are discussed in the case study in Chapter 3. This method makes more assumptions than, say, Begg's method, but has been criticised (including by the authors of the paper themselves) of perhaps not being reliable for assessing funnel plot asymmetry when there are only a small number of studies in a meta-analysis.

In spite of this, one could argue that the Egger test is one of the most widely used statistical tests in meta-analysis.

Duval and Tweedie [27]-[28] in 2000 presented their Trim and Fill method. This method, similar to the Egger test, aimed to make use of the funnel plot as a method of testing for publication bias, as well as attempting to adjust for any potential publication bias. This non-parametric method provided a more objective approach to evaluating for bias in a funnel plot. The basic description is as follows [80]. The number of “asymmetric” studies on, say, the right hand side of the funnel is estimated. The “asymmetric” studies can broadly be thought of as studies having no counterparts on the left hand side of the funnel plot. These studies are trimmed from the funnel leaving the symmetric remainder from which the “true” average treatment effect is estimated. The trimmed studies are then replaced, and their missing counterparts filled, mirrored around the axis placed at the calculated average estimate. More details are given in Section 2.5.1.

Even though the Trim and Fill method has been criticised for depending too strongly on assumptions about funnel plot asymmetry, and that adjusting the results of a meta-analysis by inputting “fictional” missing studies is considered controversial, the method has been used numerous times by different researchers. One such example is a BMJ article by Sutton *et al.* [81]. Their paper analysed 48 systematic reviews in the Cochrane database, and with the aid of the Trim and Fill method they concluded that publication and related biases were found to be present in approximately 50% of reviews. In this paper Sutton *et al.* found that the overall conclusions were not reversed in most studies, except for 4 studies. Another good example of where the Trim and Fill method has been applied is a paper by Jennions and Møller [51]. Focusing on systematic reviews relevant to ecology and evolution, the authors concluded that one in five meta-analyses were affected by publication bias, with the aid of the Trim and Fill method.

In response to the increase in variety of approaches to detecting and handling publi-

cation bias, the number of research papers and publications critically appraising these methods also steadily grew in 2000 and onwards. Two notable publications collating together and summarising the more important concepts and methods were *Methods for Meta-Analysis in Medical Research* by Sutton *et al.* in 2000 [80] and *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*, edited by Rothstein *et al.* in 2005 [71]. These publications (amongst others) were fundamental in promoting good techniques in meta-analysis to cater for an audience of both technical and non-technical statisticians and health practitioners.

Other notable research critically appraising methods to handling publication bias include a paper by Sterne *et al.* in 2000 [78] and a paper by Pham *et al.* in 2001 [64]. The paper by Sterne *et al.* investigated the difference in performance of the rank correlation test (Begg's method) and a weighted regression method (Egger's test). The paper by Pham *et al.* compared a variety of methods, including Begg's method, the Egger test, the Trim and Fill method, Rosenthal's file drawer approach, and the use of weighted functions. In both papers, the authors concluded that different methods reach different conclusions, and so there was still ongoing debate about the degree of usefulness of these methods concerning the detection and handling of publication bias.

This section has presented just a selection of the research that has taken place over the last fifty years or so, specific to publication bias in meta-analysis. The constant additions of research in this field illustrates how vibrant the area of interest is. To name just a few recent examples, Trikalinos *et al.* in 2004 [87] investigated effect sizes in cumulative meta-analyses of mental health randomised trials, and they found numerous examples of meta-analyses with small numbers of studies that revealed significant overall results, only to become non-significant as the number of studies in the meta-analysis grew over time. This paper demonstrated the importance of maintaining caution with reviews with small numbers of studies.

Also in 2004 was a paper by Bennett *et al.* [4] which compared the method of capture-recapture against the Egger test, the Trim and Fill method and other techniques used

to assess for publication bias. Capture-recapture, a concept well known in epidemiology, is concerned with trying to detect all individuals within a population of interest. The method that Bennett *et al.* present very much falls in line with the philosophy of the other techniques that they investigated, and put forward this method to rival these techniques.

We conclude this brief overview of the history of publication bias by mentioning two final sources of information summarising the most recent developments in the study of publication bias in meta-analysis. The first is a paper by Sutton and Higgins [83] in 2007 who summarised the most important advances within the topic broken down into more specialist areas, with a wealth of references and directions to good sources of software for meta-analysis. The other notable reference is *The Handbook of Research Synthesis and Meta-Analysis: Second Edition*, a book edited by Cooper *et al.* [14] in 2009. The book appeals to both a technical and non-technical audience, with the inclusion of numerous examples and discussion throughout.

#### 2.4.1 The use of selection functions in meta-analysis

Section 2.4 first discussed the use of selection functions (also known as weight functions) in meta-analysis, but since this is a central concept to the subsequent chapters, a more in depth discussion is given here. As a reminder, the chapter by Hedges and Vevea in [71] is an excellent review of numerous and current selection function approaches. One of the fundamental points to note is that these types of approaches usually involve two parts: modelling the effect size and modelling the selection. Modelling the effect size essentially involves modelling the data before any kind of selection has occurred, so for example, we could assume that  $n$  studies in a meta-analysis have estimates  $y_i$  such that  $y_i \sim N(\theta_i, \sigma_i^2)$ . Here  $\theta_i$  are the parameters of interest and the  $\sigma_i^2$ s are assumed known. The second part, modelling the selection, usually involves some kind of parametric model which is used to describe the varying level of probability of selection assumed to be due to publication bias. Weighted distribution theory is clearly a core idea here used to model publication bias. The basic

idea behind the use of weight functions is that studies included in a meta-analysis constitute a sample from a weighted distribution - the bigger the weight, the bigger the probability of a the study being published. Throughout the thesis, we will usually denote these selection functions as  $a(y_i)$  or  $a(y, \sigma)$  as examples. Therefore,  $a(y, \sigma)$  models the probability that a study is selected given the values of, say,  $y$  and  $\sigma$ .

One of the inherent problems with any kind of selection function approach is that it is very difficult, if not impossible, to fully determine if such a function  $a(y, \sigma)$  is adequately modelling the selection process. Therefore some researchers have instead chosen to assume different selection functions and then perform a sensitivity analysis to investigate the effects of differing degrees of publication bias.  $\beta$  has been frequently used to denote a parameter measuring the degree of publication bias, which is a convention that shall be used throughout this thesis. Preston *et al.* [68] provide examples of different selection functions, and uses a real data set (a systematic review concerning oral rehydration solution in the treatment of dehydration) to see the effect of modelling the selection process on the overall results. Preston *et al.* looked at selection functions such as  $a(y, \sigma) = e^{-\beta V}$  where  $V$  denotes the study's P-value. This is one such example of a selection function that will be used in later chapters of this thesis.

The various selection function approaches can be categorised into two groups: the first is those where the selection function depends on the ratio  $y/\sigma$ , or equivalently the P-value for each study. The second group is those where the selection function depends on both  $y$  and  $\sigma$ , the effect size and its standard error, separately. This first group of selection functions include all examples whereby there is a belief that the probability of a study being published is dependent upon the statistical significance of the overall result. An example of early research is that by Lane and Dunlap [53] in 1978 and Hedges [43] in 1984 who used the extreme selection function  $a(y, \sigma) = 1$  if the P-value  $< 0.05$  and  $a(y, \sigma) = 0$  otherwise.

Another example of a selection function takes the form of a step selection function. A

good early example is that of Dear and Begg [24] in 1992. Hedges and Vevea provide a good illustration of the usage of a step selection function [45]. Briefly, assume that selection depends on a one-tailed P-value, and if we assume that the selection function has  $k$  intervals on which  $a(y, \sigma)$ , the probability of selection, is constant. A step function could therefore take the form

$$a(y, \sigma) = \begin{cases} w_1, & \text{if } 0 < V \leq u_1 \\ w_2, & \text{if } u_1 < V \leq u_2 \\ \vdots & \vdots \\ w_k, & \text{if } u_{k-1} < V \leq u_k \end{cases}$$

where the  $i^{\text{th}}$  interval has end points  $u_{i-1}$  and  $u_i$ , and that if a study's P-value  $V$  falls within interval  $i$ , then that particular study will have weight  $w_i$ . Hedges and Vevea go into a lot more depth into how to use step selection functions of this type, such as how it fits into a likelihood function for the data. They also offer a possible solution to how to estimate all the necessary model parameters.

Copas [15] and Copas and Shi [21] developed a selection function which falls into the more complicated category of a function that depends on both  $y$  and  $\sigma$ , not just the ratio  $y/\sigma$ . We shall discuss this particular selection function in more depth in Chapter 5. Copas *et al.* promote the idea of a sensitivity analysis which essentially involves testing their selection function's fit to the funnel plot. In spite of their research being highly statistical (and therefore arguably too technical and out of reach for those with little statistical knowledge to implement), their method is credited for making a selection function that has more realistic assumptions than those which depend solely upon a study's P-value. So for two studies with the same P-value, the larger of the two will have a higher probability of publication.

There are a few notable criticisms concerning the use of selection functions. The first is that selection functions do not perform well for meta-analyses with small numbers of studies. This intuitively makes sense since smaller numbers implies less information upon which to estimate the various quantities necessary for the model. Another criticism of the use of functions such as the aforementioned step selection functions

is that it is essentially arbitrary as to where to place these steps. This is something that will be discussed further in Chapter 6. One could place these steps at the conventional critical values of say 0.01, 0.05, and so on, but it remains to be seen if this is a valid approach. On the other hand, psychological research has been carried out that suggests people perceive a result to be more conclusive if its P-value is less than 0.05 or 0.01 [69].

As we shall see in subsequent chapters, another issue to be aware of when using selection functions is that clearly there could be a wide variability in the estimates of the meta-analysis dependent upon the choice of the selection function. This is another reason why Copas *et al.* promote the use of a sensitivity analysis approach. In spite of the potentially horrendous computations necessary to implement some of the above mentioned selection functions, it is undeniable how useful these selection functions can be to model selection and how they can reveal possible consequences to how much overall estimates in a meta-analysis may change when modelling publication bias in a meta-analysis.

## **2.5 A Review of Recent Research Investigating Publication Bias**

In this section we focus on a selected few pieces of recent research that have been carried out to investigate methods of handling publication bias in meta-analysis. The main concepts from each piece of research are summarised here, and examples of the methods will be included in subsequent chapters. For further reading, references are given throughout.

### **2.5.1 Research by Duval and Tweedie**

The Trim and Fill method, as first discussed in Section 2.4, was put forward by Duval and Tweedie to provide a more objective assessment of the funnel plot. We present

here a more detailed look into the method, for which full details can be obtained in the original papers [27]-[28].

1. *Estimating  $k_0$ , the number of unobserved studies due to publication bias.*

For  $i = 1, \dots, n$ , we have effect size  $y_i$  estimating some global effect size  $\theta$  and an estimated within-study variance  $\sigma_i^2$ . We assume that there are an additional number,  $k_0$ , of studies not observed due to publication bias. The value of  $k_0$  is unknown and therefore must be estimated. As described by Duval and Tweedie [27], the key assumption behind this non-parametric method is that:

“the suppression has taken place in such a way that it is the  $k_0$  values of the  $y_i$  with the most extreme left-most values that have been suppressed.”

The ranks of the absolute values of the observed effect sizes and the signs of those effect sizes around  $\theta$  are used to form estimators of  $k_0$ . For  $i = 1, \dots, n$ , define

$x_i$  as the observed values of  $y_i - \theta$ ,

$r_i$  as the ranks of the absolute values  $|x_i|$ ,

$T_n$  as the sum of  $r_i$  for positive  $x_i$  only, and

$\gamma^* \geq 0$  as the length of the right-most run of ranks associated with positive values of the observed  $x_i$ .

Define two estimators of  $k_0$  as

$$R_0 = \gamma^* - 1, \tag{1}$$

$$L_0 = \frac{4T_n - n(n+1)}{2n-1}. \tag{2}$$

Both of these estimators have good statistical properties. Full details of these properties, and a description of another estimator of  $k_0$  (namely,  $Q_0$ ), are given by Duval and Tweedie [27]. In practice, we round  $L_0$  and  $R_0$  to the nearest integer since we need to trim whole studies. Duval and Tweedie recommend using both estimators before making a judgement on the number of missing studies.

## 2. The iterative Trim and Fill algorithm.

The iterative Trim and Fill algorithm is as follows. The random effects approach is assumed, but the fixed effects model is also applicable (page 41 for further details of the fixed and random effects model).

### *Step One*

Define  $\hat{\theta}^{(1)}$  as the first estimate of  $\theta$ , using the random effects estimator.

Construct the first set of centred values  $y_i^{(1)} = y_i - \hat{\theta}^{(1)}$ ,  $i = 1, \dots, n$ .

Define  $\hat{k}_0^{(1)}$  as our first estimate of  $k_0$ , for example  $L_0$  as given in equation (2), applied to the set  $y_i^{(1)}$ ,  $i = 1, \dots, n$ .

### *Step Two*

Remove  $\hat{k}_0^{(1)}$  values from the right end of the set of initial values  $y_i$   $i = 1, \dots, n$ .

Define  $\hat{\theta}^{(2)}$  as our second estimate of  $\theta$ , based on the trimmed symmetric set of  $n - \hat{k}_0^{(1)}$  values.

Construct the next set of centred values  $y_i^{(2)} = y_i - \hat{\theta}^{(2)}$ ,  $i = 1, \dots, n$ .

Define  $\hat{k}_0^{(2)}$  as our second estimate of  $k_0$  based on the set  $y_i^{(2)}$ ,  $i = 1, \dots, n$ .

If  $\hat{k}_0^{(2)} = \hat{k}_0^{(1)}$ , then proceed to *Step Four*. Otherwise, continue with *Step Three*.

### *Step Three*

Remove  $\hat{k}_0^{(2)}$  values from the right end of the set of initial values  $y_i$   $i = 1, \dots, n$ .

Define  $\hat{\theta}^{(3)}$  as our third estimate of  $\theta$ , based on the trimmed symmetric set of  $n - \hat{k}_0^{(2)}$  values.

Construct the next set of centred values  $y_i^{(3)} = y_i - \hat{\theta}^{(3)}$ ,  $i = 1, \dots, n$ .

Define  $\hat{k}_0^{(3)}$  as our third estimate of  $k_0$  based on the the set  $y_i^{(3)}$ ,  $i = 1, \dots, n$ .

### *Step Four*

Continue iterating *Step Three* in a similar manner until an iteration,  $J$ , where  $\hat{k}_0^{(J)} = \hat{k}_0^{(J-1)} = \hat{k}_0$ , at which point  $\hat{\theta}^{(J)} = \hat{\theta}^{(J-1)} = \hat{\theta}$ . Fill the funnel plot with the trimmed

$\hat{k}_0$  right hand studies, and input the “missing” counterpart studies

$$y_i^* = 2\hat{\theta} - y_{n-j+1}, \text{ for } j = 1, \dots, \hat{k}_0,$$

with standard errors

$$\sigma_j^* = \sigma_{n-j+1}, \text{ for } j = 1, \dots, \hat{k}_0.$$

An adjusted value of  $\theta$  can be calculated as

$$\hat{\theta} = \frac{\sum_{i=1}^n y_i w_i + \sum_{j=1}^{\hat{k}_0} y_j^* w_j^*}{\sum_{i=1}^n w_i + \sum_{j=1}^{\hat{k}_0} w_j^*}, \quad (3)$$

with a corresponding 95% confidence interval given by

$$\left( \hat{\theta} - 1.96 \left\{ \sum_{i=1}^n w_i + \sum_{j=1}^{\hat{k}_0} w_j^* \right\}^{-1/2}, \hat{\theta} + 1.96 \left\{ \sum_{i=1}^n w_i + \sum_{j=1}^{\hat{k}_0} w_j^* \right\}^{-1/2} \right), \quad (4)$$

where the  $w_i$ s are the usual weights for a random effects model, namely

$$\begin{aligned} w_i &= (\sigma_i^2 + \tau_F^2)^{-1}, \\ w_j^* &= (\sigma_j^2 + \tau_F^2)^{-1}. \end{aligned}$$

$\tau_F^2$  is estimated, based on the entire data set  $\{y_1, \dots, y_n, y_1^*, \dots, y_{\hat{k}_0}^*\}$ .

In principle, this Trim and Fill method is actually quite straightforward to implement, as the example in Section 3.4.2 will hopefully demonstrate. It is perhaps why this method has been frequently used and investigated by other researchers. Due to the aforementioned concerns over its usage (the strong assumptions about funnel plot asymmetry for example), Duval and Tweedie recommend the use of the Trim and Fill method as a means of providing a sensitivity analysis, suggesting to look at how the value of  $\theta$  changes depending upon the number of missing studies. Recent research by Terrin *et al.* suggests that this method may wrongly adjust for publication bias when there is none when studies are heterogeneous, further suggesting the use of methods other than the Trim and Fill method to handling publication bias under these conditions [86].

### 2.5.2 Research by Copas and Jackson

As previously discussed, many approaches to modelling publication bias in a meta-analysis have used selection (weight) functions. These approaches all experience the same limitation, in that it is virtually impossible to estimate the selection mechanism from the observed studies alone. Also the choice of the selection function is entirely arbitrary. We summarise the main concepts from the research by Copas and Jackson [16], who looked at the bias of the “worst case” scenario across a plausible range of values for the number of unpublished studies.

Suppose there are  $n$  studies, each with their own values of the outcome  $y$  and  $\sigma^2$ , the variance of  $y$ . We assume  $y|\sigma$  to be normally distributed,  $y|\sigma \sim N(\theta, \sigma^2)$  with density

$$g(y; \theta, \sigma^2) = \frac{1}{\sigma} \phi\left(\frac{y - \theta}{\sigma}\right),$$

where  $\theta$  is the quantity of interest. Define the selection function  $a(y, \sigma)$  as

$$a(y, \sigma) = P(\text{selection}|y, \sigma).$$

We suppose that the variation of  $\sigma$  to be random with a distribution  $f(\sigma)$ . Note that in a fixed effects model,  $\theta$  denotes the common treatment effect over all studies and  $\sigma^2$  is the study’s variance. For the random effects model,  $\theta$  is the average treatment effect and  $\sigma^2$  denotes the sum of within-study variance, say  $s_i^2$ , and the between-study variance  $\tau^2$ . Concerning the selection procedure, the only assumption made is that the conditional probability  $a(\sigma)$  which is defined as

$$\begin{aligned} a(\sigma) &= P(\text{selection}|\sigma) = \mathbb{E}[a(y, \sigma)|\sigma] \\ &= \int_{-\infty}^{\infty} a(y, \sigma)g(y; \theta, \sigma)dy \end{aligned}$$

is a decreasing function of  $\sigma$ , meaning larger studies are more likely to be selected than smaller studies. This seems like a reasonable and valid assumption to make. Using Bayes’ rule, the joint distribution of  $(y, \sigma)$  for a selected study, denoted  $g_o(y, \sigma)$ , is given as

$$g_o(y, \sigma) = P(y, \sigma|\text{selection}) = \frac{a(y, \sigma)g(y; \theta, \sigma)f(\sigma)}{p}, \quad (5)$$

where  $p$  is the overall selection probability

$$\begin{aligned} p &= P(\text{selection}) = \mathbb{E}[a(\sigma)] = \int_0^\infty a(\sigma)f(\sigma)d\sigma \\ &= \int_0^\infty \int_{-\infty}^\infty \frac{1}{\sigma}\phi\left(\frac{y-\theta}{\sigma}\right)a(y,\sigma)f(\sigma)dyd\sigma. \end{aligned}$$

The distribution of  $\sigma$  for a selected study is

$$f_o(\sigma) = \frac{a(\sigma)f(\sigma)}{p}. \quad (6)$$

The data observed is a random sample  $\{(y_i, \sigma_i^2)\}$ ,  $i = 1, \dots, n$ , from  $g_o(y, \sigma)$  as given in (5). The convention of using the inverse variance estimator of  $\theta$  is followed, where

$$\hat{\theta} = \frac{\sum_{i=1}^n w_i y_i}{\sum w_i},$$

with  $w_i = 1/\sigma_i^2$ . Note that  $\hat{\theta}$  is the standard maximum likelihood estimator of  $\theta$  when  $y_i \sim N(\theta, \sigma_i^2)$ .

It can be shown that the asymptotic bias  $b$  in estimating  $\theta$  with  $\hat{\theta}$  is

$$b = \frac{\int_0^\infty \int_{-\infty}^\infty \sigma^{-1}\phi(z)a(\theta + \sigma z, \sigma)f(\sigma)dzd\sigma}{\int_0^\infty \sigma^{-2}a(\sigma)f(\sigma)d\sigma}.$$

**THEOREM:** For given overall selection probability  $p$ ,

$$|b| \leq \frac{\bar{\sigma}}{p}\phi\left(\Phi^{-1}(p)\right), \quad (7)$$

where

$$\bar{\sigma} = \frac{\int_0^\infty \sigma^{-1}f_o(\sigma)d\sigma}{\int_0^\infty \sigma^{-2}f_o(\sigma)d\sigma} = \frac{\mathbb{E}_o(\sigma^{-1})}{\mathbb{E}_o(\sigma^{-2})}, \quad (8)$$

and  $\mathbb{E}_o$  denotes expectation over  $\sigma$  with respect to the distribution  $f_o(\sigma)$ . The upper bound is attained when [the weight function  $a(y, \sigma)$  is the step function

$$a(y, \sigma) = a(y) = \begin{cases} 1 & \text{if } y \geq \theta - \sigma\Phi^{-1}(p) \\ 0 & \text{if } y < \theta - \sigma\Phi^{-1}(p) \end{cases}$$

for all values of  $\sigma$ ]. The lower bound is attained when  $a(y, \sigma)$  equals the step function with 1 and 0 interchanged and with the minus sign before  $\sigma$  changed to plus.

In practice,  $\mathbb{E}_o$  denotes an average over the observed values  $\sigma_1, \dots, \sigma_n$ . Also,  $p$  can be considered to be the ratio  $\frac{n}{n+m}$ , where  $m$  represents the number of unpublished studies. Therefore, for any given values of  $m$ , the bounds in (7) can be estimated as

$$|b| \leq \frac{n+m}{n} \phi \left\{ \Phi^{-1} \left( \frac{n}{n+m} \right) \right\} \frac{\sum_{i=1}^n \sigma_i^{-1}}{\sum_{i=1}^n \sigma_i^{-2}}. \quad (9)$$

For a sensitivity analysis, we take  $m = 0, 1, 2, \dots$  and plot (9) against  $m$ .

### 2.5.3 Research by Henmi *et al.*

The approach of Copas and Jackson [16] makes no assumptions about the selection process except one - all else equivalent, larger studies are more likely to be published than smaller studies. Their approach asks the question “how bad could the bias be?” Their worst case scenario approach has the obvious advantage that there is no dependence on any untestable assumptions with regards to selection. There is the concern of how useful is this bound. If the bound gives overly cautious values, then its usefulness in practice will be called into question. However, if the limits of the confidence intervals are close to these bounds, when using all the various selection functions to model the selection process, then the bound could be viewed as a very useful tool.

Copas and Jackson’s research was extended to not only examine the bound on the bias of the estimate of  $\theta$ , but looking at the bounds on confidence intervals and P-values - arguably more relevant in practice. Henmi, Copas and Eguchi [47] proposed a sensitivity analysis for publication bias looking at the bounds on confidence intervals and P-values. Again, very few assumptions are made about selection. The main result in Henmi *et al.* is briefly given below, for full details refer to [47]. First, note that  $\hat{\theta}$  remains defined as the conventional weighted average, and  $\bar{w} = \frac{\sum w_i}{n}$ .

**THEOREM:** *The confidence region is an interval with lower and upper limits*

$$\hat{\theta} + \frac{1}{\bar{w}} L(\alpha, p, f_o) \quad \text{and} \quad \hat{\theta} + \frac{1}{\bar{w}} U(\alpha, p, f_o) \quad (10)$$

where

$$L(\alpha, p, f_o) = \min_{\lambda} C_{-}^*(\lambda, \alpha, p, f_o), \quad U(\alpha, p, f_o) = \max_{\lambda} C_{+}^*(\lambda, \alpha, p, f_o)$$

with

$$\begin{aligned} C_{\pm}^*(\lambda, \alpha, p, f_o) &= -B_1^*(\lambda, p, f_o) \pm n^{-1/2} z_{\alpha} \sqrt{B_2^*(\lambda, p, f_o) - \{B_1^*(\lambda, p, f_o)\}^2}, \\ B_1^*(\lambda, p, f_o) &= \frac{1}{p} \mathbb{E}_o \left[ \frac{1}{\sigma} \{ \phi(\lambda\sigma + e) - \phi(\lambda\sigma - e) \} \right], \\ B_2^*(\lambda, p, f_o) &= \mathbb{E}_o \left[ \frac{1}{\sigma^2} \left( 1 + \frac{1}{p} \{ (\lambda\sigma + e)\phi(\lambda\sigma + e) - (\lambda\sigma - e)\phi(\lambda\sigma - e) \} \right) \right] \end{aligned}$$

and where  $e = e(\lambda, \sigma, p)$  is defined by

$$\Phi(\lambda\sigma - e) + \Phi(-\lambda\sigma - e) = p. \quad (11)$$

The proof is not considered here. For full details of the proof, see Henmi *et al* [47].

In practice, we evaluate the confidence interval as

$$\left[ \hat{\theta} + \frac{1}{\hat{w}} \hat{L}(m), \quad \hat{\theta} + \frac{1}{\hat{w}} \hat{U}(m) \right] \quad (12)$$

where

$$\hat{L} = L(\alpha, \hat{p}, \hat{f}_o) = \min_{\lambda} C_{-}^*(\lambda, \alpha, \hat{p}, \hat{f}_o), \quad (13)$$

$$\hat{U} = U(\alpha, \hat{p}, \hat{f}_o) = \max_{\lambda} C_{+}^*(\lambda, \alpha, \hat{p}, \hat{f}_o) \quad (14)$$

and

$$C_{\pm}^*(\lambda, \alpha, \hat{p}, \hat{f}_o) = -B_1^*(\lambda, \hat{p}, \hat{f}_o) \pm n^{-1/2} z_{\alpha} \sqrt{B_2^*(\lambda, \hat{p}, \hat{f}_o) - \{B_1^*(\lambda, \hat{p}, \hat{f}_o)\}^2}. \quad (15)$$

$B_1^*$  and  $B_2^*$  are calculated as follows.

$$B_1^* = \frac{(n+m)}{n^2} \sum_{i=1}^n \frac{1}{\sigma_i} \{ \phi(\lambda\sigma_i + e_i) - \phi(\lambda\sigma_i - e_i) \}, \quad (16)$$

$$B_2^* = \frac{1}{n} \sum_{i=1}^n \frac{1}{\sigma_i^2} \left[ 1 + \frac{(n+m)}{n} \{ (\lambda\sigma_i + e_i)\phi(\lambda\sigma_i + e_i) - (\lambda\sigma_i - e_i)\phi(\lambda\sigma_i - e_i) \} \right] \quad (17)$$

where  $e_i = e(\lambda, \sigma_i, \hat{p})$ , for  $i = 1, \dots, n$ , is defined by

$$\Phi(\lambda\sigma_i - e_i) + \Phi(-\lambda\sigma_i - e_i) = \frac{n}{n+m}. \quad (18)$$

Note that  $z_\alpha = \Phi^{-1}(1 - \alpha/2)$  is the standard normal percentage point with coverage  $1 - \alpha$ . For all examples to follow,  $z_\alpha = 1.96$  so that we consider 95% confidence intervals. Equations (12) - (18) are those that we shall use to estimate the bounds.

An example demonstrating the use of what we shall refer to as the Bounds method is given in Chapter 3: Case study: A Meta-Analysis Investigating the Effects of Environmental Tobacco Smoke. Furthermore, the S-Plus codes that can be used to compute the bounds will be given in the Appendix.

This concludes Chapter 2: A Literature Review, where we introduced the general settings for a meta-analysis, and discussed what the main issues are, for example publication bias. We focused more closely on the various approaches used to modelling publication bias, including specific references to techniques such as the Trim and Fill method and the Bounds method that we shall use in subsequent chapters of this thesis.

## 3 Case Study: A Meta-Analysis Investigating the Effects of Environmental Tobacco Smoke

### 3.1 Environmental Tobacco Smoke and Lung Cancer

There is a vast amount of scientific evidence to suggest tobacco smoking is detrimental to a person's health, such as increasing the risk of lung cancer, heart disease, cardiovascular disease, bronchitis and asthma. Studies as early as 1950 (Peto *et al.* [63] provide an interesting discussion) and even before then have made the association between tobacco smoking and lung cancer. An equally important issue in public health is environmental tobacco smoke.

Environmental tobacco smoke is defined as the combination of "sidestream" smoke (from the burning tip of a cigarette) and "mainstream" smoke (exhaled by the smoker). Environmental tobacco smoke, henceforth ETS, is also known as passive smoking, or secondhand smoking. Similar to the association of tobacco smoking and lung cancer, there is extensive evidence claiming ETS increases the risk of lung cancer. For example, a search of the Medline database returns a list of literally thousands of related articles.

The effects of smoking on a person's health have rightfully been the subject of much research for the past sixty years. A report by the British Medical Association in 2004 provided alarming mortality estimates for the UK. It was estimated smoking related illnesses kills approximately 113,000 people in the UK per year, and passive smoking kills approximately 1,000 people in the UK per year [6]. The same report

also discusses how smoking related illnesses cost the NHS over £1.7bn per year. Considerations of ETS were essential in the decision to ban smoking in public places in the UK, which came into effect by 2007 (2006 in Scotland), with the British Medical Association and other respected bodies believing such a ban would make a significant contribution to public health.

In this chapter, we provide a meta-analysis of environmental tobacco smoke and lung cancer. The data is taken from a paper originally written by Hackshaw *et al.* in 1997 [40]. Although their analysis is over 10 years old, the implications of ETS are as relevant today as they have ever been. The Department of Health's Scientific Committee on Tobacco and Health (SCOTH) in 2004 released a report [74], which supported a smoking ban in public places, in part based upon the analysis by Hackshaw *et al.*

Section 3.2 discusses the findings of the paper written by Hackshaw *et al.* The data will be the primary example of a meta-analysis for the remaining chapters, and so the original paper will be discussed here in quite some detail. Section 3.3 will summarise the subsequent analyses undertaken by other researchers in response to the Hackshaw paper. This will serve two purposes: demonstrating a selection of some of the different approaches one can take when undertaking a meta-analysis, and also highlighting important issues commonly faced within a meta-analysis. The works of Copas and Shi [22] and Henmi *et al.* [47] are discussed in this section. In Chapter 2, a variety of approaches to modelling publication bias were discussed. Most of these methods, while valid in their own right, may be out of reach for those who have limited and relevant statistical knowledge. Since many systematic reviews show no attempts to consider publication bias [81], simple graphical or numerical summaries could be used routinely. With the aid of an updated meta-analysis carried out by Taylor *et al.* [85], Section 3.4 will include a summary of suggestions of established methods concerning possible publication bias in a meta-analysis.

## 3.2 The Hackshaw *et al.* Paper

In 1986 Wald *et al.* [92] reviewed 13 epidemiological studies to investigate the effects of ETS, specifically to the risk of lung cancer. Their paper supported the conclusion that breathing in other people's tobacco smoke can cause lung cancer. Eleven years later the BMJ published a paper by Hackshaw *et al.* [40] which extended the research from the original paper. During this time, a substantial amount more of data was now being considered, with three times as many studies in this latest analysis. What follows is a detailed discussion of the main points in the Hackshaw *et al.* paper. For the full paper, see [40].

A total of 39 relevant studies were included, including five cohort studies and 34 case-control studies. The studies were published between 1981 and 1997, with geographical regions Asia (44%), USA (36%) and Europe (20%). Since the studies originate across a wide global area, this intuitively suggests a random effects analysis may be more appropriate (discussed later). Variation between study effects sizes may be explained by cultural differences in different geographical regions. There is however a noticeable under representation of developing countries from the list of epidemiological studies.

For each of the 39 studies, the relative risk outcome measure was given with a corresponding 95% confidence interval. In 28 of the studies, the risk of lung cancer was calculated only for lifelong non-smokers who were women. In nine studies, two relative risks were reported - one for when the lifelong non-smokers were female, and a separate estimate for when the lifelong non-smokers were male. In the remaining two studies, the estimates used data for male and female lifelong non-smokers combined.

Hackshaw *et al.* decided to base most of their analysis only on the data for female lifelong non-smokers, a total of 37 different studies. This seems like a reasonable thing to do since the majority of reported lung cancer cases (91%) are from the female data. 31 studies reported an increase in risk of lung cancer, of which seven studies were statistically significant (the confidence intervals for the relative risk did not include

the null value of 1). The remaining six studies reported negative results suggesting the passive smoking prevented lung cancer. However, all six negative results were not statistically significant. One of the larger studies that reported a negative result commented that their effect estimate was most likely affected by another cause of lung cancer, namely using open coal fires to cook with little ventilation [95].

The main conclusion of the paper was that the evidence from combining the epidemiological studies corroborated the belief that environmental tobacco smoke causes lung cancer. They used the following points to support their claim.

1. First, the pooled estimate of the relative risk was 1.24, meaning there was a 24% excess risk of lung cancer in female lifelong non-smokers whose spouses smoked compared to those whose spouses did not smoke. The corresponding 95% confidence interval was (1.13, 1.36) so that the observed effect size was statistically significant with a P-value of less than 0.05. Therefore they had strong evidence to reject the null hypothesis (that the relative risk = 1). Hackshaw *et al.* made the conscious decision to only use the female data in their analysis. This ignored the separate data on male lifelong non-smokers in nine studies, as well as the two studies with combined female/male data. When Hackshaw *et al.* included this additional data, there was little difference to the pooled estimate of the relative risk, 1.23, with 95% confidence interval (1.13, 1.34).
2. Also supporting their claim was the dose-response relation they found between the risk of lung cancer and environmental tobacco smoke. 16 studies reported data that could investigate a relationship between the number of cigarettes smoked by the spouse and the risk of lung cancer, see for example Garfinkel [35] and Akiba *et al.* [1]. A positive relationship between risk and number of cigarettes was found, providing supporting evidence of causality. Hackshaw *et al.* also describe how 11 studies reported data that could investigate a relationship between the number of years a woman lived with a smoker and the risk of

lung cancer (Humble *et al.* [49], Stockwell *et al.* [79] and Cardenas *et al.* [9] are good examples). These studies suggested risk increases for women who have lived with a smoker for a longer period of time. A dose-response relationship is arguably one of the stronger criteria from Bradford Hill's criteria for causation [48]. Therefore, the evidence examining the duration of exposure to ETS provides good evidence to support the claim that ETS causes lung cancer.

3. The third area of reasoning by Hackshaw *et al.* supporting their claims that environmental tobacco smoke causes lung cancer was the inability for bias and confounding to explain the apparent association between the two variables. They had identified two possible sources of bias and one possible confounder that could have affected the observed outcome from the data. The first type of bias is misclassification bias. This occurs when current or former smokers are misclassified as lifelong non-smokers. Using existing methods (Wald *et al.* [92]), the relative risks for each study was adjusted to take into account possible misclassification bias. Relevant empirical evidence and national data was used to estimate quantities that Hackshaw *et al.* deemed necessary to in turn estimate the misclassification bias.

The second type of bias was due to the exposure to ETS in the reference group (non-smokers whose spouses did not smoke). It is somewhat inevitable that these people will have been exposed to environmental tobacco smoke from other sources, for example, at the workplace or social venues. It has been shown that the average levels of urinary cotinine (a chemical product only of tobacco smoke) of non-smokers with spouses who do not smoke is not zero, see for example Wald *et al.* [93]. Using empirical evidence, Hackshaw *et al.* adjusted the relative risk to take this type of bias into account.

One possible confounder that was considered to explain the observed association was the diet of the lifelong non-smoker. Previous research suggests the risk

of lung cancer may increase for those who have low levels of fruit and vegetable consumption, see for example Candelora *et al.* [8], and also non-smokers whose spouses smoke are less likely to eat fruit and vegetables than non-smokers whose spouses do not smoke (see for example, Cardenas *et al.* [9]). In spite of the fact that only a few of the 39 studies included in their paper recorded data about diet, Hackshaw *et al.* made adjustments to the estimate of the relative risk to take into account dietary confounding.

After adjusting the relative risk for the confounding and two types of bias, the pooled estimate of the relative risk became 1.26 with 95% confidence interval (1.06, 1.47). Comparing this to the original pooled estimate, 1.24 (1.13, 1.36), the conclusion was the two estimates were similar and therefore bias and confounding could not account for the association between environmental tobacco smoke and lung cancer.

4. The final strand in the arguments of Hackshaw *et al.* to support their claim that environmental tobacco smoke causes lung cancer was the discussion of the existing biological and experimental evidence. Much previous research has shown that tobacco related carcinogens (known substances capable of causing cancer) are found in the blood and urine of non-smokers. Two such examples include the research of Maclure *et al.* [56] and Hammond *et al.* [41]. Very briefly both examples investigated levels of hemoglobin adducts, specifically 4-aminobiphenyl (4-ABP) which is a known carcinogen, within non-smokers. Mean levels of adducts were significantly higher in those non-smokers who were exposed to ETS compared to those non-smokers who were not (defined by having undetectable levels of cotinine). In general, studies of this kind that are investigating specific tobacco carcinogens are very convincing, because they satisfy many of the Bradford Hill criteria, such as the coherence, experimental and biological plausibility criteria [48].

In summary, the four different strands discussed by Hackshaw *et al.* led them to conclude that exposure to ETS causes lung cancer. This includes calculating the pooled estimate of the relative risk (1.24) with a corresponding confidence interval, investigating a dose-response relationship, considering sources of bias and confounding, and discussing experimental evidence of tobacco carcinogens.

### 3.3 Existing Analyses

Since the publication of the Hackshaw paper in the BMJ in 1997, numerous researchers have not only scrutinised their findings, but also used the data that they had collected to demonstrate methods used to model and adjust for publication bias in a meta-analysis. Arguably two important reasons why the research of Hackshaw *et al.* has received so much subsequent attention from others is that the issue of passive smoking has been, and remains still, a serious issue in public health. The other reason is that from an academic point of view, the data is a good example to illustrate researchers' new methods of handling publication bias in a meta-analysis due to the evidence of the presence of potential publication bias within the data. This following section will include some of these existing analyses based on the methods that were first discussed in Chapter 2. The data is given in Table 3.1 (page 36) and throughout the thesis will be referred to as the passive smoking dataset.

In 2000, Copas and Shi used the passive smoking dataset in conjunction with a parametric selection function model that depends on both the effect size  $y$  and its standard error  $\sigma$  [22]. Their research was first discussed in Section 2.4.1 (page 15). Technical details are given later in Chapter 5. Briefly, Copas and Shi used the Hackshaw data to demonstrate their method of proposing a sensitivity analysis in which a range of different assumptions to their selection function can be tested against the fit to the funnel plot. Their paper suggests that, when a likelihood-based confidence interval is calculated and subsequently the number of unpublished studies is estimated, the once statistically significant result becomes non-significant when there are about  $m = 28$  unpublished studies. (Note also that Hedges and Vevea include a practical example

of this Copas selection function in [45].)

In 2005, Hedges and Vevea contributed to an excellent text edited by Rothstein *et al.* [71]. In their chapter, Hedges and Vevea used the passive smoking dataset as one of their prominent examples to illustrate several different selection functions, first discussed in Section 2.4.1 (page 15). The first example they discussed involved a selection function depending only on study P-values, with the selection function estimated from the data. This non-parametric approach used a step function, denoted as  $w(p)$ , with steps at  $p = 0.05, 0.1$  and  $0.5$ , namely

$$w(p) = \begin{cases} 1 & \text{if } 0 \leq p < 0.05 \\ w_2 & \text{if } 0.05 \leq p < 0.1 \\ w_3 & \text{if } 0.1 \leq p < 0.5 \\ w_4 & \text{if } 0.5 \leq p < 1 \end{cases}$$

Hedges and Vevea present the results of their analysis first under the assumption of no selection ( $w_2 = w_3 = w_4 = 1$ ) and then assuming selection via  $w(p)$  with weights ( $w_2 = 2.48, w_3 = 1.01, w_4 = 0.42$ ). They argued that the maximum likelihood estimate of  $\theta$  reduced from 0.22 to 0.13. Incidentally, when looking at the  $100(1 - \alpha)\%$  confidence intervals for  $\theta$  when assuming selection, for any value of  $\alpha$  (where  $\hat{\theta} = 0.13$ ), these confidence intervals always contained the value 0.22 corresponding to the estimate of  $\theta$  when assuming no selection. For more details, see [45].

Hedges and Vevea also discussed a numerical example of selection depending on study P-values, with the selection function specified *a priori*. The passive smoking dataset was used to demonstrate a sensitivity analysis, by assuming four different *a priori* specifications for the selection function, namely a weak and a strong one-tailed and two-tailed P-value selection function. The impact was considered on the overall estimate of  $\theta$  under the different levels of severity of the four assumed selection functions. Hedges and Vevea concluded with their sensitivity analysis that if you assume strong selection, then the overall estimate of  $\theta$  would “be of minimal clinical interest”. For more details, again see [45].

In 2007, Henmi *et al.* used the passive smoking dataset to demonstrate their research on bounds for confidence intervals and P-values for meta-analysis [47]. Their research and main results were first discussed in Section 2.5.3 (page 24). Henmi *et al.* found that a random effects analysis was appropriate, and that there was 23% added risk of lung cancer from exposure to passive smoking, with 95% confidence interval (13%, 35%). When applying their worst-case sensitivity analysis method, the limits of the confidence intervals did not widen large enough to reverse the significance of the overall result (by including the log odds ratio value of zero) until  $m = 19$  unpublished studies. This translates to an overall selection probability of 66%, which the authors believe is rather extreme. If one is willing to accept that such an overall selection probability is not feasible, then the evidence from the Hackshaw meta-analysis still stands.

Table 3.1: Hackshaw dataset, 1982-1997: epidemiological studies of the risk of lung cancer in lifelong non-smokers whose spouses smoked relative to the risk in those whose spouses do not smoke.

Study	Year	Country	Relative risk	95% confidence interval	Data $y_i$	$\sigma_i$
<i>case control studies</i>						
Chan	1982	Hong Kong	0.75	(0.43,1.30)	-0.29	0.28
Correa	1983	USA	2.07	(0.81,5.25)	0.72	0.48
Trichopolous	1983	Greece	2.13	(1.19,3.83)	0.76	0.30
Buffler	1984	USA	0.80	(0.34,1.90)	-0.22	0.44
Kabat	1984	USA	0.79	(0.25,2.45)	-0.25	0.59
Lam	1985	Hong Kong	2.01	(1.09,3.72)	0.70	0.31
Garfinkel	1985	USA	1.23	(0.81,1.87)	0.21	0.21
Wu	1985	USA	1.20	(0.50,3.30)	0.25	0.45
Akiba	1986	Japan	1.52	(0.87,2.63)	0.41	0.28
Lee	1986	UK	1.03	(0.41,2.55)	0.02	0.47
Koo	1987	Hong Kong	1.55	(0.90,2.67)	0.44	0.28
Pershagen	1987	Sweden	1.03	(0.61,1.74)	0.03	0.27
Humble	1987	USA	2.34	(0.81,6.75)	0.85	0.54
Lam	1987	Hong Kong	1.65	(1.16,2.35)	0.50	0.18
Gao	1987	China	1.19	(0.82,1.73)	0.17	0.19
Brownson	1987	USA	1.52	(0.39,5.96)	0.42	0.69
Geng	1988	China	2.16	(1.08,4.29)	0.77	0.35
Shimizu	1988	Japan	1.08	(0.64,1.82)	0.08	0.27
Inoue	1988	Japan	2.55	(0.74,8.78)	0.94	0.63
Kalandidi	1990	Greece	1.62	(0.90,2.91)	0.48	0.30
Sobue	1990	Japan	1.06	(0.74,1.52)	0.06	0.18
Wu-Williams	1990	China	0.79	(0.62,1.02)	-0.23	0.12

*continued on next page*

*continued from previous page*

Study	Year	Country	Relative risk	95% confidence interval	Data	
					$y_i$	$\sigma_i$
Liu	1991	China	0.74	(0.32,1.69)	-0.31	0.43
Jockel	1991	Germany	2.27	(0.75,6.82)	0.82	0.57
Brownson	1992	USA	0.97	(0.78,1.21)	-0.03	0.11
Stockwell	1992	USA	1.60	(0.80,3.00)	0.44	0.35
Du	1993	China	1.19	(0.66,2.13)	0.17	0.30
Liu	1993	China	1.66	(0.73,3.78)	0.51	0.42
Fontham	1994	USA	1.26	(1.04,1.54)	0.24	0.10
Kabat	1995	USA	1.10	(0.62,1.96)	0.10	0.29
Zaridze	1995	Russia	1.66	(1.12,2.45)	0.50	0.20
Sun	1996	China	1.16	(0.80,1.69)	0.15	0.19
Wang	1996	China	1.11	(0.67,1.84)	0.10	0.26
<i>cohort studies</i>						
Garfinkel	1981	USA	1.18	(0.90,1.54)	0.16	0.14
Hirayama	1984	Japan	1.45	(1.02,2.08)	0.38	0.18
Butler	1988	USA	2.02	(0.48,8.56)	0.71	0.73
Cardenas	1997	USA	1.20	(0.80,1.60)	0.12	0.21

### 3.4 Promoting a Routine Investigation of Publication Bias in Meta-Analysis

In 2007, Taylor *et al.* [85] presented an updated meta-analysis concerning the risks of environmental tobacco smoke (ETS) and lung cancer. Their data will be used in this section for two purposes: the first is that it will be interesting to discuss how the main conclusions of their meta-analysis may have changed since 1997, with the addition of more recent studies since those included by Hackshaw *et al.*; the second purpose for including this updated meta-analysis is to provide a simple analysis as a means of promoting routine investigations of publication bias in meta-analysis.

#### 3.4.1 An updated analysis by Taylor *et al.*

Hackshaw *et al.* concluded from their analysis in 1997 that there was convincing evidence to support the on-going debate on the risk of lung cancer due to passive smoking. It may come as little surprise that the tobacco companies believe the risks are over-estimated, but there has been a noticeable amount of disagreement by some believing that the observed excess risk of lung cancer in non-smokers who live with smokers is entirely due to bias. One such critic, Lee [55], in 1992 wrote a book about the available evidence on ETS and the associated health risks. Lee assessed over fifty studies and concluded that, while the majority of studies examining ETS and lung cancer reported a statistically significant excess risk, there were persistent problems such as the presence of biases. Lee described the associations as weak at best and believed there was still no convincing evidence.

Many other researchers have performed meta-analyses over the last twenty years or so to attempt to synthesise the relevant published studies. To list just a few, there was one meta-analysis in 1992 by Tweedie & Mengersen [88]; in 1997, there was another by Wang [94], whose results interestingly contrasted against the majority of reviews by seemingly showing a beneficial exposure to ETS; and in 2002 a meta-analysis by Boffetta [7], again illustrating the numerous reviews undertaken.

Taylor *et al.* set out to contribute to this field of research. Here we discuss how this new meta-analysis compares with that of Hackshaw *et al.*, and more specifically what practical decisions have been made about the data to allow us to implement our own subsequent meta-analysis.

The first steps towards gaining data for our own meta-analysis involves comparing which of Hackshaw's 37 studies are amongst the 55 studies in the meta-analysis by Taylor *et al.* Upon close inspection of the studies included in Hackshaw's meta-analysis, 30 studies were identical to those included in Taylor's meta-analysis (some with slightly different reported estimates), 5 studies were updated studies, for example a cohort study with a longer follow up period, and 2 studies were missing without clear explanation (although one was a PhD thesis, the other was part of a dissertation). There were 20 studies that were included within Taylor's meta-analysis that were not in Hackshaw's. 17 of these were published in 1998 and onwards, but 3 were published before 1998. These 3 studies could be seen as signs of possible publication bias, since these studies were published before Hackshaw performed their analysis, but for one reason or another, these 3 studies were not included. Examples include a study by de Waard *et al.* [91] who are researchers from a university in the Netherlands.

Taylor *et al.* includes a summary of results of all 55 studies included in their analysis. Specifically they provide estimates of  $\theta$  given as the relative risk or odds ratio with corresponding 95% confidence intervals. Adjusted estimates (relative risk or odds ratio) and/or unadjusted estimates of risk were provided for each study. Refer to Tables 1-3 in the 2007 paper [85]. To conduct our own analysis based upon the Taylor data, careful efforts were made to be consistent with the Hackshaw data  $(y_i, \sigma_i)$ . This meant that if a study gave only one estimate (and corresponding confidence interval), clearly that value would have to be used. 36 studies provided only one estimate. If both adjusted and unadjusted estimates were presented (as in 15 studies), the values corresponding to those included in the Hackshaw meta-analysis were used. If a study presented two estimates and it was not included in the 1997 meta-analysis (as in 4

studies), then similar studies would indicate if an adjusted or unadjusted estimate was used. This process of deriving the data resulted in 55 pairs of data  $(y_i, \sigma_i)$  which can be found in the final two columns of Table 3.5 starting on page 56.

Similar to the Hackshaw meta-analysis, the Taylor meta-analysis is made up of a mixture of cohort studies and case-control studies. There are 7 cohort studies and the majority of studies (48) are case-control studies in the Taylor dataset, compared to 4 cohorts and 33 case-control studies in the Hackshaw dataset. A comparison of the overall analyses and conclusions between the 1997 and 2007 study will be discussed further in the following section.

### 3.4.2 A recommended approach in meta-analysis

In this section we use this example to illustrate our recommended approach to meta-analysis. There are of course many various methods and approaches to synthesising evidence together. For a comprehensive view of such methods see the works of Sutton *et al.* [80] and [82], and more recently the text edited by Rothstein *et al.* [71].

As previously discussed in Chapter 2, meta-analysis is the statistical part of systematic reviews. Much debate has occurred, and continues to do so, discussing guidelines for good meta-analytic practice, see for example Deeks *et al.* [25], and Cook *et al.* [13]. The following analysis will not go into as much detail as contained in these guidelines of good practice. Instead the analysis will be roughly divided into three parts: the first is stating the model assumptions and exploring for heterogeneity; the second part will include several of the many different graphical ways of exploring the data; the final part will examine the robustness of the results of the meta-analysis with special consideration of possible effects of publication bias.

We begin this analysis under the assumption that we have collected the data from the various studies included in the systematic review. Table 3.5 starting on page 56 gives the results of 55 studies concerning the risk of lung cancer when exposed to

environmental tobacco smoke. An estimate of the relative risk for cohort studies, or odds ratio for case-control studies was reported along with a 95% confidence interval (columns four and five of Table 3.5). From this we calculate the data  $(y_i, \sigma_i)$ , where  $y_i$  is the  $i^{th}$  log relative risk and  $\sigma_i^2$  is the variance of  $y_i$  (columns six and seven of Table 3.5).

### Step 1: Choice of model and summary statistics

We first assume that the conventional *fixed effects model* is appropriate, namely we have  $n$  independent studies where

$$y_i \sim N(\theta, \sigma_i^2), \quad i = 1, \dots, n. \quad (19)$$

The maximum likelihood estimate of  $\theta$ , with study weights equal to the inverse study variance  $w_i = 1/\sigma_i^2$ , is given as

$$\hat{\theta} = \frac{\sum_{i=1}^n y_i w_i}{\sum_{i=1}^n w_i}. \quad (20)$$

It is also easily shown that

$$v = \text{var}(\hat{\theta}) = \frac{1}{\sum_{i=1}^n w_i}. \quad (21)$$

We are then in a position to calculate statistical quantities such as confidence intervals and P-values. For example,

$$\left( \hat{\theta} - 1.96\sqrt{v} \quad , \quad \hat{\theta} + 1.96\sqrt{v} \right), \quad \text{and} \quad (22)$$

$$P = 2\Phi\left(-|\hat{\theta}|/\sqrt{v}\right) \quad (23)$$

are respectively the 95% confidence interval for  $\theta$  and a two-tailed P-value for the null hypothesis  $H_0 : \theta = 0$ . Since for this example the study estimates  $y_i$  are log values of the relative risk, it is recommended to transform to the relative risk scale by calculating  $RR_i = \exp(y_i)$  when summarising the results in a meta-analysis.

A simple assessment to assess whether or not the fixed effects model is appropriate is the  $\chi^2$  test. We have the null hypothesis (where  $\theta_i$  corresponds to study  $i$ ):

$$H_0 : \theta_1 = \theta_2 = \dots = \theta_n = \theta, \text{ versus}$$

$H_1$ : at least one  $\theta_i$  different.

The statistic,  $Q$ , is defined as

$$Q = \sum_{i=1}^n w_i (y_i - \hat{\theta})^2. \quad (24)$$

Under the null hypothesis,  $Q$  has a  $\chi_{n-1}^2$  distribution. A computationally convenient form of (24) is

$$Q = \sum_{i=1}^n w_i y_i^2 - \frac{(\sum_{i=1}^n w_i y_i)^2}{\sum_{i=1}^n w_i} \quad (25)$$

Briefly, the fixed effects model given in (19) assumes that all studies included in the meta-analysis are estimating the same quantity,  $\theta$ .  $Q$  is simply the sums of squares of the study outcome  $y_i$  around the pooled estimate  $\hat{\theta}$ . If there is considerable variation, more so than what can be reasonably observed with sampling variation alone, then  $Q$  will take large values. So for example, we would reject the null hypothesis  $H_0$  at the 5% level if  $Q \geq \chi_{n-1,0.95}^2$ .

If there is reasonable doubt from the  $\chi^2$  test that the fixed effects model may not be appropriate, we recommend considering the alternative model, namely the *random effects model*. Here we have  $n$  independent studies where

$$y_i \sim N(\theta, \sigma_i^2 + \tau^2), \quad i = 1, \dots, n, \quad (26)$$

where  $\tau^2$  is the random effects variance. This model is more appropriate when (possible) heterogeneity is present between studies. The maximum likelihood estimate of  $\theta$  takes the same form as in (20), except the study weights are now

$$w_i = \frac{1}{\sigma_i^2 + \tau^2}. \quad (27)$$

Once we have estimated  $\tau^2$  (since it is usually unknown) we can proceed just like the fixed effects model with calculating quantities such as P-values and confidence intervals under the random effects model assumptions. We recommend following the method of DerSimonian and Laird [26] to estimate  $\tau^2$ . Essentially, this is calculated as follows.

$$\hat{\tau} = \begin{cases} 0 & \text{if } Q \leq n - 1 \\ \frac{Q - (n-1)}{U} & \text{if } Q > n - 1, \end{cases} \quad (28)$$

where  $Q$  is defined in (25),  $U$  is defined as

$$U = (n - 1) \left( \bar{w} - \frac{s_w^2}{n\bar{w}} \right). \quad (29)$$

$\bar{w}$  and  $s_w^2$  are respectively the sample average and unbiased sample variance of the  $w_i$ s calculated in the conventional way.

Table 3.2 summarises the calculations as carried out with the Taylor dataset. A summary of the same calculations for the Hackshaw dataset have been included to allow for a comparison between the two meta-analyses. This includes the calculation of  $\hat{\theta}$  and corresponding 95% confidence interval for both the fixed effects and random effects model, the observed values of the  $\chi^2$  test and the corresponding P-value (using equations (20)-(29)).

Table 3.2: Calculations for the fixed effects model and random effects model for the Hackshaw and Taylor meta-analyses.

Model	Between variance estimate $\hat{\tau}^2$	$\hat{\theta}$	95% confidence interval	$Q$	$P$ -value ( $\chi^2$ test)
<i>Hackshaw dataset, 37 studies</i>					
Fixed	0	0.183	(0.110, 0.256)	47.7	0.092
Random	0.0174	0.213	(0.122, 0.305)	-	-
<i>Taylor dataset, 55 studies</i>					
Fixed	0	0.197	(0.137, 0.258)	67.9	0.097
Random	0.0138	0.225	(0.151, 0.299)	-	-

For both datasets and both models, all estimates of  $\theta$  are positive and statistically significant, because all 95% confidence intervals exclude the null value, which is zero on the log relative risk scale. This suggests exposure to passive smoking is harmful to a person's health. When applying the  $\chi^2$  test to both data sets (columns five and six in Table 3.2), we see there is some evidence at the 10% level of significance

to suggest heterogeneity is present between studies, suggesting the use of random effects models. What appears to be most striking about the data in Table 3.2 is the similarities between the Hackshaw and Taylor datasets. Ten years later and 50% more studies than the Hackshaw meta-analysis, the main conclusions derived from the fixed effects and random effects models with the Taylor dataset appear consistent.

Even though all the calculations summarised in Table 3.2 are necessary, a typical health practitioner reading through the summary statistics in that table may find some of it confusing, especially with the use of the logarithmic scale. We therefore recommend in any meta-analysis that summary statistics be expressed in a suitable format so that appropriate conclusions can be inferred. Table 3.3 has transformed the summary statistics for both the Hackshaw and Taylor datasets from the log scale so that the outcome of interest is the relative risk. This will give the reader a clearer interpretation of the results.

Table 3.3: Summary statistics for the Hackshaw and Taylor meta-analyses.

Model	Relative risk	95% confidence interval	$P$ -value ( $H_0 : RR = 1$ )
<i>Hackshaw dataset, 37 studies</i>			
Fixed	1.20	(1.12,1.29)	$9.12 \times 10^{-7}$
Random	1.24	(1.13,1.36)	$5.03 \times 10^{-6}$
<i>Taylor dataset, 55 studies</i>			
Fixed	1.22	(1.15,1.29)	$1.50 \times 10^{-10}$
Random	1.25	(1.16,1.35)	$2.50 \times 10^{-9}$

For example, with the fixed effects model for the Taylor data, the relative risk is 1.22, which means the risk of lung cancer for female non-smokers whose spouses smoke is 22% greater compared to those female non-smokers whose spouses do not smoke. Notice how all the estimates of the relative risk and confidence intervals look quite

similar, and that the conclusions from the 2007 Taylor meta-analysis are consistent with the 1997 Hackshaw meta-analysis. Table 3.3 also includes the corresponding  $P$ -values for each of the models, testing the hypothesis  $H_0 : RR = 1$ . Clearly all the  $P$ -values are very small, suggesting strong evidence to reject the null hypothesis.

### Step 2: Graphical displays of the data

Graphical displays of the data are highly recommended within any meta-analysis. The funnel plot first mentioned in Chapter 2 is one of the most commonly used plots, mainly due to its simple graphical display. Figure 3.1 shows a funnel plot for the 2007 Taylor dataset. Study outcome  $y_i$  is plotted against study precision  $1/\sigma_i$ . The 95% confidence interval of each study is also plotted, represented by horizontal lines.

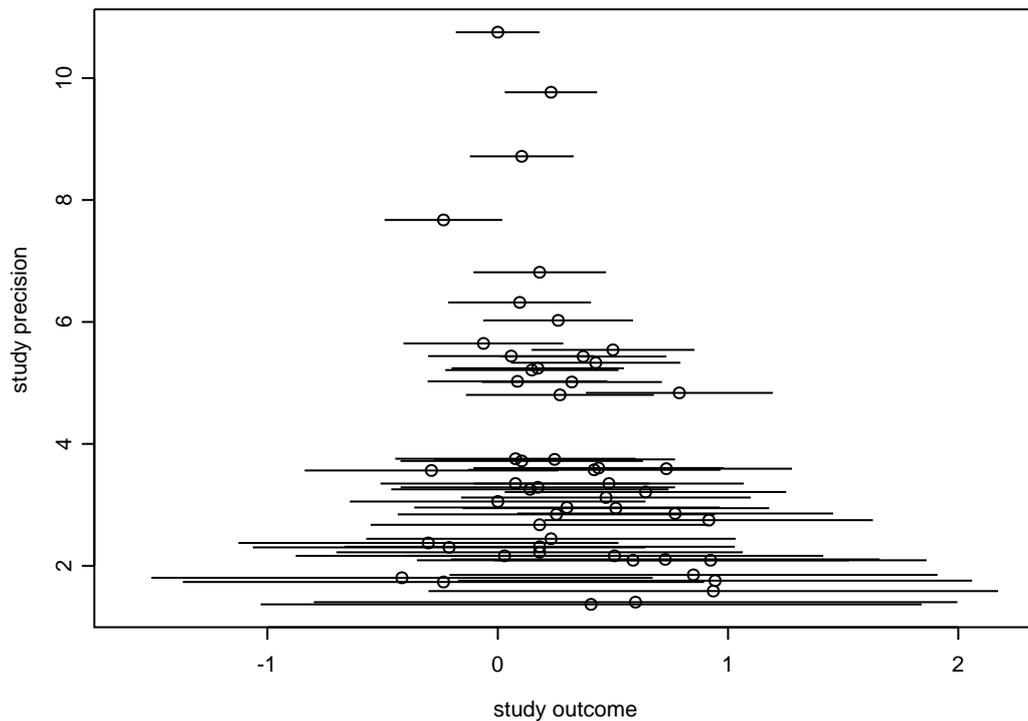


Figure 3.1: Funnel plot for the 2007 Taylor meta-analysis. The horizontal lines represent each study's confidence interval.

Clearly those studies with smaller study precision (smaller studies) have wider confidence intervals compared to studies with larger study precision (larger studies). There appears to be a possible drift towards the right of the plot in Figure 3.1 as the studies become smaller. This suggests that there may be missing smaller studies whose study outcome is nearer zero or even negative. A trend such as this is a classic sign for the presence of possible publication bias.

It is fair to say that inspection of any funnel plot is entirely subjective, where different people may reach contrasting conclusions. Also differences in the choice of outcome measure may change a person's opinion about the conclusions of a meta-analysis. Nonetheless, we recommend the routine use of an appropriate funnel plot in any meta-analysis, being cautious of potentially inferring too much when, say, there are too few studies within the meta-analysis.

Another commonly used graphical display within a meta-analysis is the forest plot. The plot elegantly displays each study's outcome and corresponding confidence interval all on one set of axes. Figure 3.2 shows the forest plot for the Taylor dataset, with 95% study confidence intervals. The size of the plotting symbol is proportional to the study weight  $w_i$  (the reciprocal of the study variance). This means the most influential studies will have the bigger symbols and therefore will stand out visually. The smaller studies will have the widest confidence intervals and are less influential than the bigger studies.

Note that in Figure 3.2 a vertical line has been added to represent the value of  $\log(RR) = 0$ , or in other words when the relative risk is 1. The addition of a line representing the null value always aids interpretation, for example, to the left of the vertical line, the study estimates suggest exposure to environmental tobacco smoke is beneficial to a person's health. If the study estimate is to the right of the line, this suggests exposure is harmful to a person's health. From the plot we see that only 7 studies show a negative value suggesting exposure to ETS is beneficial. However, all of these studies' confidence intervals overlap with the vertical line at  $\log(RR) = 0$

meaning the results are not statistically significant.

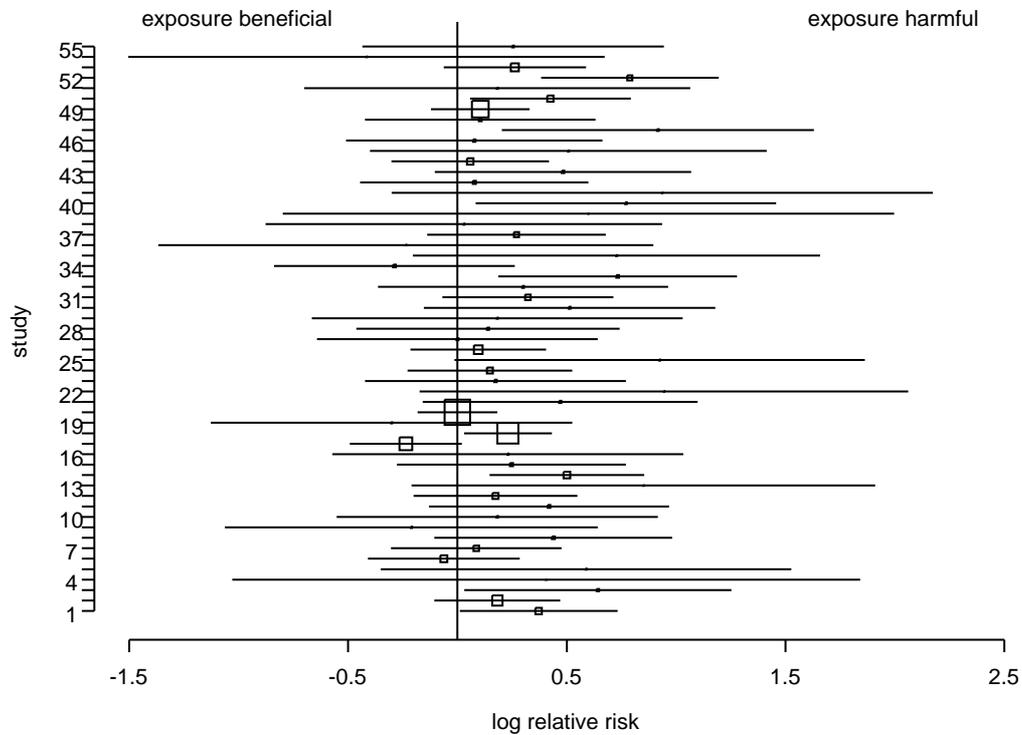


Figure 3.2: Forest plot for the 2007 Taylor meta-analysis. The horizontal lines represent each study's confidence interval.

Another use of the forest plot is to explore for between-study heterogeneity. Non-overlapping confidence intervals would highlight the variability between study estimates, and indicate that the studies are possibly not estimating the same quantity,  $\theta$ . In such an example, the use of the fixed effects model would not be considered an appropriate assumption. Figure 3.2 shows some evidence of non-overlapping intervals, such as study 49 and 52. Just from examining this plot there does appear to be some heterogeneity present within the data, however we would recommend further investigation such as the tests mentioned previously in Step 1 (page 41 onwards).

The radial plot, first put forward by Galbraith [34], is the third of the recommended graphical displays of the data to be mentioned. Also referred to as a Galbraith plot, the essential idea is to examine the degree of funnel plot asymmetry, for which the basic idea follows. Each study's  $z$  statistic,  $z_i = y_i/\sigma_i$ , is plotted against the study precision (reciprocal of the standard error)  $w_i^{1/2} = 1/\sigma_i$ . Studies in a meta-analysis that do not have between-study heterogeneity should be scattered homoscedastically around a line through the origin whose gradient represents the pooled estimate for  $\theta$ . Figure 3.3 shows a radial plot of the Taylor dataset, and plots like this can be used to visually inspect a meta-analysis for the presence of heterogeneity between studies.

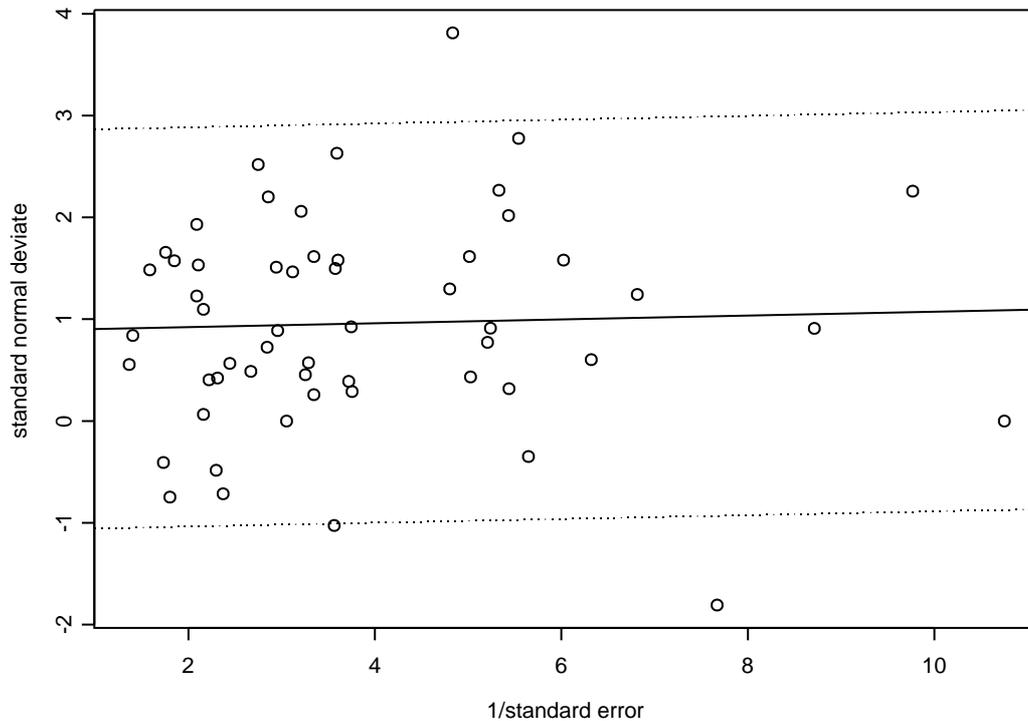


Figure 3.3: Radial plot for the 2007 Taylor meta-analysis.

We could not discuss the radial plot without first discussing its companion, the Egger test [31]. Briefly, this is a regression test where the values of  $z_i$  are fitted against

$w_i^{1/2}$ , study precision, using a standard weighted linear regression with weights  $w_i$ , with equation

$$z = \alpha + \beta \frac{1}{\sigma}. \quad (30)$$

The slope parameter,  $\beta$ , indicates the size and direction of the pooled estimate for  $\theta$  (provided  $\alpha = 0$ ). The intercept parameter,  $\alpha$ , provides a measure of asymmetry: the larger its deviation from zero, the more pronounced the asymmetry. If smaller studies show effects differing systematically from larger studies, then the regression line will not run through the origin, and therefore there will be a non-zero  $\alpha$  value. The Egger test is applied to the Taylor dataset (which can be found in Table 3.5 starting on page 56), and a summary of this weighted regression test, as produced by S-Plus, is given in Table 3.4.

Table 3.4: S-Plus regression output for the Taylor dataset.

Coefficients	Value	Std error	<i>t</i> -value	<i>P</i> -value
Intercept	0.8855	0.3013	2.9386	0.0049
$1/\sigma$	0.0187	0.0689	0.2717	0.7869
Multiple $R^2$	0.0014			
F-statistic	0.0738			

The fitted regression line is given by  $z = 0.89 + 0.02/\sigma$ . This line has been included in Figure 3.3. Parallel lines indicating the limits for 95% confidence intervals have also been added to the plot, denoted by dotted lines. These limits are simply a distance of 1.96 away from the fitted line. Briefly, this is because the original fixed effects model (as discussed in Step 1, page 41) assumes that

$$y_i \sim N(\theta, \sigma_i^2),$$

and therefore transforming the data to  $z_i = y_i/\sigma_i$ , we have

$$z_i \sim N\left(\frac{\theta}{\sigma_i}, 1\right). \quad (31)$$

The  $z_i$  values therefore have constant variance, and so we use this whilst inspecting the radial plot in Figure 3.3. About 95% of the studies should lie within the dotted lines, which is the case with our dataset. Only three studies (approximately 5%) lie outside of the dotted lines, which possibly suggests there is little between-study heterogeneity present within the meta-analysis. Referring to Table 3.4, the intercept of the regression line is 0.89 with corresponding  $P$ -value of 0.0049 which is clearly statistically significant. Therefore, according to the Egger test, we have very strong evidence to reject the null hypothesis  $H_0 : \alpha = 0$ , or in other words, there is strong evidence of publication bias.

We recommend any investigation of a meta-analysis to include graphical displays of the data, and here we have discussed just three different types of plots. See Sutton *et al.* [80] for a good introduction to other graphical displays of meta-analysis data. It should be noted that there is constant on-going debate about the effectiveness of such methods, and how much we should infer from them. For example, examination of radial plots and the Egger test has come under some scrutiny. Recent research by Schwarzer *et al.* [72], Peters *et al.* [60], and Harbord *et al.* [42] suggests the Egger test gives over-inflated significance levels with regards to the test of publication bias  $H_0 : \alpha = 0$ . These biases may account for the very small  $P$ -value noted above with the Egger test. Note that these researchers, amongst others such as Copas and Lozada [17]-[18], have presented alternatives to the Egger test.

### **Step 3: Robustness and modelling publication bias**

This final section discusses two possible ways of modelling and adjusting for publication bias in a meta-analysis. There are many approaches, but the two that will be discussed here are the Bounds method by Henmi *et al.* and the Trim and Fill method by Duval and Tweedie. For the technical details, see Section 2.5.1 and Section 2.5.3 (starting from page 18). The reason why these two particular methods have been chosen is that they are relatively straightforward methods to implement, and that

makes them accessible to both qualified statisticians and health practitioners who may not have as much statistical experience. It is important to remember the issue concerning these methods' incorrect usage before proceeding, namely we are not saying that the adjusted estimates resulting from these methods provide the true estimate, but simply if we are to entertain the possibility that publication bias exists within the meta-analysis, then we can examine the impact upon the overall results.

First consider the Bounds method by Henmi *et al.*. Note that the S-Plus code used to implement the Bounds method in practice is given in Appendix A2 (page 140). We assume here that the random effects model is appropriate from earlier investigations for heterogeneity. Figure 3.4 shows the 95% confidence limits for the Taylor dataset plotted for a range of plausible values for the overall selection probability  $p$ , produced by using the aforementioned S-Plus code.

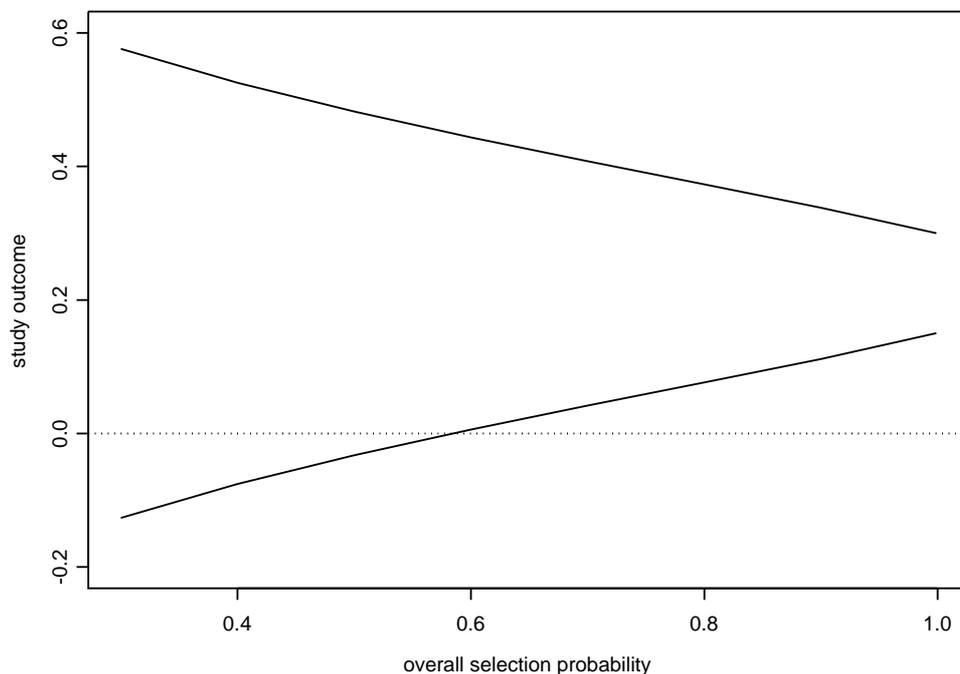


Figure 3.4: Confidence limits using the Bounds method for the 2007 Taylor meta-analysis.

Of particular interest is at what point does the lower confidence limit cross the null line  $\theta = 0$ . This occurs when the overall selection probability is 59% or equivalently when there are 39 unpublished studies. Therefore if there are 39 or more unpublished studies, then the original significance of the overall result will be overturned.

Incidentally, if we compare back to the Hackshaw meta-analysis (10 years previously) when applying this Bounds method, 19 unpublished studies or equivalently an overall selection probability of 66% would result in a reversal in the significance of their overall result. The interpretation here is that, since the point at which the significance becomes overturned has decreased from 66% to 59% for the 1997 and 2007 meta-analyses respectively, the level of publication bias plays less of a significant role within the latter meta-analysis. Even though we do have evidence that the overall estimate may be inflated, we would have to assume a more severe selection process was present to cast doubt on the validity of the overall results of the meta-analysis.

The second method that we would recommend within a meta-analysis investigation is the Trim and Fill method by Duval and Tweedie. We apply this method to the Taylor dataset following the details of the Trim and Fill method as given in Section 2.5.1 (page 18). Recall that we are assuming a lack of studies towards the left hand side of the funnel plot (Figure 3.1 presented on page 45). A random effects model has been used resulting in an initial estimate of  $\theta$  to be  $\hat{\theta}^{(1)} = 0.22$ , assuming all studies have been included. The first set of estimates for  $k_0$  (the number of missing studies on the left hand side of the funnel plot) were 3 and 5 by calculating  $R_0$  and  $L_0$  respectively, and so the average of the two was taken as a rule of thumb, namely  $\hat{k}_0^{(1)} = 4$ .

Repeating the process gave the second set of estimates of  $k_0$  to be 4 and 7, again the average of the two taken to give  $\hat{k}_0^{(2)} = 6$ . After the third iteration of trimming the funnel plot, the estimated number of missing studies was again 6, and so this is our final estimate (corresponding to a high overall selection probability of 90%). Figure 3.5 shows the filled funnel plot. The circle symbols represent the observed studies and the cross symbols represent the inputted studies.

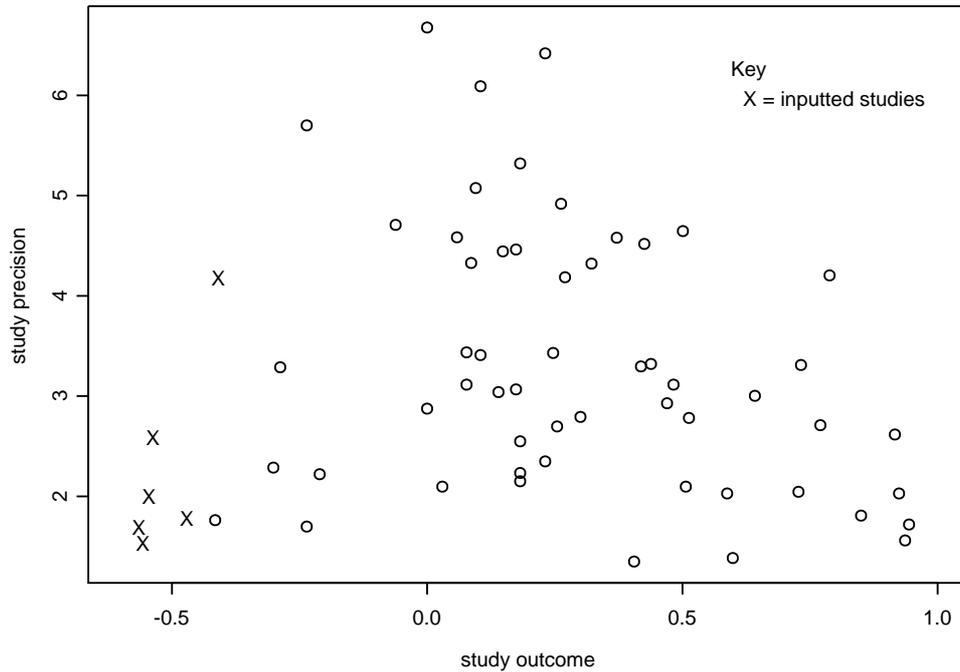


Figure 3.5: Funnel plot once the Trim and Fill Method is applied to the Taylor dataset.

Notice from Figure 3.5 how the filled funnel plot looks more visually symmetric with the additional six studies. Also, after filling, we obtain an overall estimate of  $\theta$  (based now on 61 studies) as 0.19 with corresponding 95% confidence interval (0.12, 0.26). With the additional six studies, the statistical result is still significant. Therefore based on this particular assumption of missing studies in the left tail of the funnel plot, the overall results still stand up to potential publication bias. An important note is that  $k_0$  was taken to be the average of the estimators  $L_0$  and  $R_0$ . In this example  $L_0$  and  $R_0$  were quite different, with higher estimates when using  $L_0$  compared to  $R_0$ . Assuming the higher number of missing studies, say with  $L_0$ , would imply a more severe selection process which would clearly affect the adjusted estimates of  $\theta$  and their confidence intervals. Therefore a sensitivity analysis approach is strongly encouraged, remembering one must not rely on the results of inputted studies when forming a final conclusion about  $\theta$ , but merely the usage of the Trim and Fill should give an indication of a meta-analysis that may require more careful evaluation.

### 3.4.3 Methods for investigating sources of heterogeneity

In Chapter 2 the issue of heterogeneity was first introduced. Here we discuss possible methods for investigating sources of heterogeneity. Identifying the cause of possible heterogeneity is an important part of any analysis, rather than simply combining the study results when it may not be appropriate to do so. Sutton *et al.* gives an excellent discussion about various approaches [80], for which we briefly include some here.

The use of graphical displays of the data is a very useful approach. Some have already been discussed in Section 3.4.2, such as the forest plot and the radial plot. Other plots include the L'Abbé plot, and the plot of normalized Z-scores. The L'Abbé plot involves the event risk in the treatment group plotted against the event risk in the control group for each study, where the outcome is a binary variable. If no heterogeneity is present, the points should form a cloud around a line whose gradient corresponds to the overall treatment effect. Large deviations would suggest possible heterogeneity within the data. The plot of normalized Z-scores, defined as  $z_i = (y_i - \hat{\theta})/se(y_i)$ , would suggest the fixed effects model may be inappropriate if the distribution does not follow an approximate normal distribution with mean zero and variance one.

Alternative approaches that we discuss here include subgroup analysis and meta-regression (Chapter 6 in [80] provides a good discussion). It may be appropriate to conduct a subgroup analysis, which involves investigating subsets of studies defined by either study or patient characteristics. One would use this, say, if it was believed participants within different subsets would have systematic differences. The example discussed by Sutton *et al.* is a meta-analysis of cholesterol lowering interventions and their effect on mortality. The types of treatment amongst the studies within the meta-analysis were classified into three groups: drugs, diets and surgery. Since the type of treatment is a fundamental difference amongst the studies, it was appropriate to conduct a subgroup analysis on these three individual subsets of studies to investigate whether this was the cause of heterogeneity within the meta-analysis. This approach must be carried out with caution, and the way the subsets are created

should be clearly defined prior to the analysis.

Meta-regression is a technique that provides a method of exploring and potentially explaining heterogeneity between studies. A simple definition is that meta-regression is a generalization of subgroup analysis, and it is an extension of either the fixed effects model or the random effects model in which study-level covariates are added to the models in an attempt to explain for heterogeneity [71]. Examples of covariates could include aspects of the interventions, geographical location, dose amount, and so on. The extension of the fixed effects model is called a meta-regression model, and the extension of the random effects model is called a mixed model. One would use a meta-regression model when the variation between study outcomes can be considered accountable by the covariates included. A mixed model would be more appropriate when the covariates do not explain a significant part of the heterogeneity.

A very brief description would be as follows [80]: we have  $n$  independent effect size estimates  $\hat{\theta}_1, \dots, \hat{\theta}_n$  with estimated sampling variances  $v_1, \dots, v_n$  with corresponding parameters  $\theta_1, \dots, \theta_n$ . We suppose there are  $k$  known predictor variables  $X_1, \dots, X_k$  which are related to  $\theta_i$  in the following linear form:

$$\theta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik},$$

where  $x_{i1}, \dots, x_{ik}$  are the values of  $X_1, \dots, X_k$  for study  $i$ , and  $\beta_0, \beta_1, \dots, \beta_k$  are the unknown regression coefficients to be estimated. Standard weighted multiple regression theory can then be applied.

In summary, there are a variety of approaches one should take when considering heterogeneity within a meta-analysis. Whilst Section 3.4.2 has focused more upon publication bias, heterogeneity is an equally important issue to consider. The two issues are very much entangled, and so a meta-analyst must proceed with caution when investigating both.

Table 3.5: Taylor dataset, 1982-2006: epidemiological studies of the risk of lung cancer in female lifelong non-smokers whose spouses smoked relative to the risk in those whose spouses do not smoke.

Study	Year	Country	Relative risk	95% confidence interval	Data $y_i$	$\sigma_i$
<i>case control studies pre-1998</i>						
Chan	1982	Hong Kong	0.75	(0.43,1.30)	-0.29	0.28
Correa	1983	USA	2.07	(0.81,5.25)	0.73	0.47
Trichopolous*	1983	Greece	2.08	(1.20,3.59)	0.73	0.28
Buffler	1984	USA	0.81	(0.34,1.90)	-0.21	0.43
Kabat	1984	USA	0.79	(0.25,2.45)	-0.24	0.58
Garfinkel	1985	USA	1.31	(0.87,1.97)	0.27	0.21
Wu	1985	USA	1.20	(0.60,2.50)	0.18	0.37
Akiba	1986	Japan	1.52	(0.88,2.63)	0.42	0.28
Lee	1986	UK	1.03	(0.41,2.55)	0.03	0.46
Brownson	1987	USA	1.82	(0.45,7.36)	0.60	0.71
Gao	1987	China	1.19	(0.82,1.73)	0.17	0.19
Humble	1987	USA	2.34	(0.81,6.75)	0.85	0.54
Koo	1987	Hong Kong	1.55	(0.90,2.67)	0.44	0.28
Lam	1987	Hong Kong	1.65	(1.16,2.35)	0.50	0.18
Pershagen	1987	Sweden	1.28	(0.76,2.16)	0.25	0.27
Geng	1988	China	2.16	(1.08,4.29)	0.77	0.35
Inoue	1988	Japan	2.55	(0.74,8.78)	0.94	0.63
Shimizu	1988	Japan	1.08	(0.64,1.82)	0.08	0.27
Svensson**	1989	Sweden	1.26	(0.57,2.81)	0.23	0.41
Kalandidi	1990	Greece	1.62	(0.90,2.91)	0.48	0.30
Sobue	1990	Japan	1.06	(0.74,1.52)	0.06	0.18
Wu-Williams	1990	China	0.79	(0.62,1.02)	-0.24	0.13

*continued on next page*

*continued from previous page*

Study	Year	Country	Relative risk	95% confidence interval	Data	
					$y_i$	$\sigma_i$
Liu	1991	China	0.74	(0.32,1.69)	-0.30	0.42
Brownson	1992	USA	1.00	(0.80,1.20)	0.00	0.09
Stockwell	1992	USA	1.60	(0.80,3.00)	0.47	0.32
Liu	1993	China	1.66	(0.74,1.52)	0.51	0.46
Fontham	1994	USA	1.26	(1.04,1.54)	0.23	0.10
De Waard	1995	Netherlands	2.57	(0.83,7.85)	0.94	0.57
Kabat	1995	USA	1.08	(0.60,1.94)	0.08	0.30
Du*	1996	China	1.19	(0.66,2.16)	0.17	0.30
Sun	1996	China	1.16	(0.80,1.69)	0.15	0.19
Wang**	1996	China	2.50	(1.30,5.10)	0.92	0.36
Wang	1996	China	1.11	(0.65,1.88)	0.10	0.27
Zheng**	1997	China	2.52	(1.03,6.44)	0.92	0.48
<i>case control studies 1998 onwards</i>						
Jockel*	1998	Europe	1.11	(0.88,1.39)	0.10	0.11
Zaridze*	1998	Russia	1.53	(1.06,2.21)	0.43	0.19
Boffeta	1999	Europe	1.00	(0.50,1.90)	0.00	0.33
Rapiti	1999	India	1.20	(0.50,2.90)	0.18	0.45
Zhong	1999	China	1.10	(0.80,1.50)	0.10	0.16
Lee	2000	Taiwan	2.20	(1.50,3.30)	0.79	0.21
Wang	2000	China	1.15	(0.60,2.10)	0.14	0.31
Johnson	2001	Canada	1.20	(0.50,2.80)	0.18	0.43
Kreuzer	2002	Germany	1.67	(0.86,3.25)	0.51	0.34
Seow	2002	Singapore	1.30	(0.90,1.80)	0.26	0.17
Zatloukal	2003	Prague	0.66	(0.22,1.96)	-0.42	0.56
McGhee	2005	Hong Kong	1.38	(0.94,2.04)	0.32	0.20
Gorlova	2006	USA	1.29	(0.65,2.57)	0.25	0.35
Yu	2006	China	1.35	(0.69,2.62)	0.30	0.34

*continued on next page*

*continued from previous page*

Study	Year	Country	Relative risk	95% confidence interval	Data $y_i$	$\sigma_i$
<i>cohort studies pre-1998</i>						
Hirayama	1984	Japan	1.45	(1.02,2.08)	0.37	0.18
Cardenas	1997	USA	1.20	(0.80,1.60)	0.18	0.15
<i>cohort studies 1998 onwards</i>						
Jee	1999	Korea	1.90	(1.00,3.50)	0.64	0.31
Speizer	1999	USA	1.50	(0.30,6.30)	0.41	0.73
Nishino	2001	Japan	1.80	(0.67,4.60)	0.59	0.48
Garfinkel*	2003	USA	0.94	(0.66,1.33)	-0.06	0.18
Wen	2006	China	1.09	(0.74,1.61)	0.09	0.20

\* Studies included in Hackshaw's review but have been updated/combined.

\*\* Studies published before 1998 but not included in Hackshaw's review.

#### 3.4.4 Conclusions and further remarks

Having carried out our own analysis in the previous section, we return to the 2007 meta-analysis to provide a basic summary of their findings including any points of interest not yet covered in the preceding sections. The meta-analysis by Taylor *et al.* is an update of a previous analysis carried out in 2001 [84]. The more recent analysis included more studies and considered categorising studies according to continent, study design and year of publication. A computerised literature search of Medline and Embase was undertaken (as well as approaching experts in the field of ETS) to find relevant studies. The reviewers originally found 101 studies published between 1981 and 2006. Of these studies, 46 studies were excluded from the meta-analysis for one of several reasons. Possible reasons of exclusion included studies which reported male/female results combined, studies including fewer than 7 participants, and studies where it was unclear that the risk was due to spousal exposure.

Taylor *et al.* followed a conventional statistical analysis, including familiar meta-analysis techniques such as the use of the fixed and random effects model, the Trim and Fill method by Duval and Tweedie [27], and a modification of Macaskill's test [60] to check for publication bias. A summary of the main results are as follows. 82% of the studies reported an increased risk of lung cancer (with a point estimate of the relative risk or odds ratio greater than one). Adopting a random effects model ( $\chi^2 = 67.9$ ,  $df = 54$  with a  $P$ -value of 0.1) the pooled estimate of the relative risk of lung cancer for non-smoking women spouses was 1.25 with 95% confidence interval (1.16, 1.35). There appeared to be no difference in significance of pooled estimates when categorising by study design or continent. There also was no evidence for any trend over time, with the pooled estimate remaining stable for over twenty years. Data about dose response was also considered within the analysis. Of the 36 studies to include such data, 25 gave evidence that a dose relationship existed between level of exposure and risk of cancer.

A different strand to the 2007 analysis was to gather together all other meta-analyses that have been carried concerning environmental tobacco smoke and lung cancer.

With this topic being such a controversial and far reaching issue, it is not surprising many researchers over the last twenty years have attempted to synthesise the relevant published studies. Figure 3.6 shows a pooled estimate of the odds ratio along with 95% confidence intervals for each of the 21 meta-analyses that were published between 1986 and 2007. These values are taken from [85]. Note that these meta-analyses are often using (though not exclusively) the same set of primary studies.

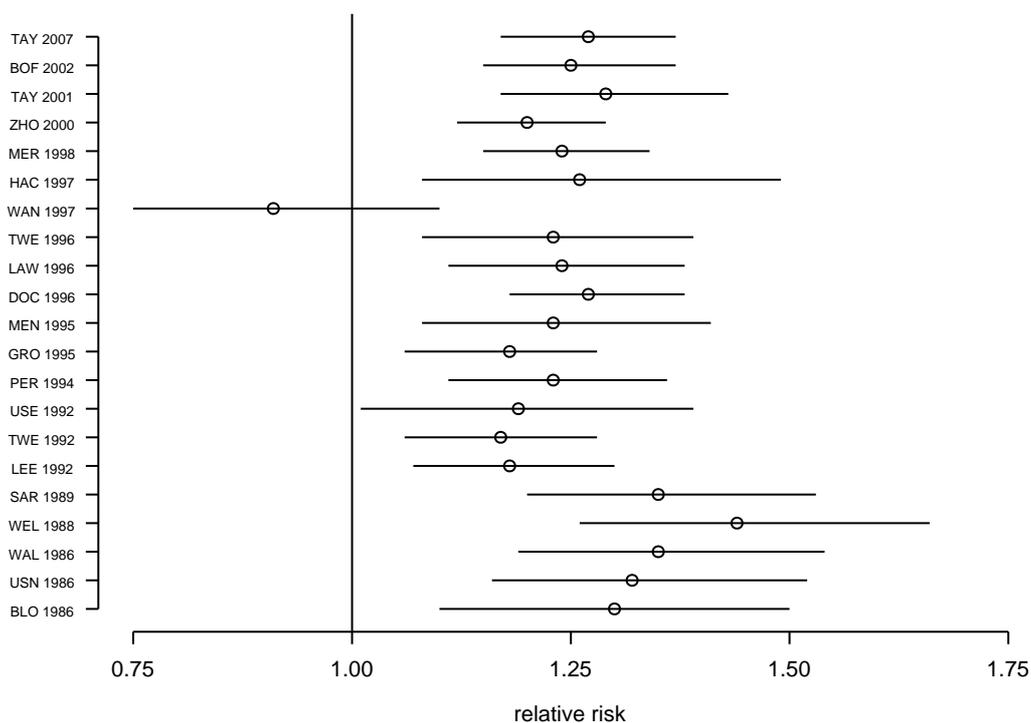


Figure 3.6: Pooled odds ratios with 95% confidence intervals for meta-analyses between 1986 and 2007.

The pooled estimates range between 0.91 and 1.44. Figure 3.6 shows that, with the exception of one meta-analysis (the Chinese meta-analysis by Wang [94] previously mentioned, including only six studies), all meta-analyses conclude there is a statistically significant harmful effect of ETS. This is shown by all but one of the 21

confidence intervals being located above 1 and not including this null value. More recent meta-analyses generally include more studies which explains why the confidence intervals are shorter for those meta-analyses towards the top of the figure.

Misclassification bias and publication bias were both considered to explain the results. Possible explanatory variables contributing to an increase risk of lung cancer include indoor air pollution, lifestyle or diet. The authors of the meta-analysis also noted that many of the studies were case-control studies which are naturally retrospective, relying upon subjective responses in questionnaires to assess the level of exposure. Based upon their analysis, Taylor *et al.* believed the observed excess risk of lung cancer was unlikely to have been caused by publication bias. They also refer to a paper by Bero *et al.* [5] which provided some evidence that there are only a small number of unpublished studies in this specific field of research. Briefly, Bero *et al.* compared peer-reviewed journal articles and symposium articles to determine the proportion of articles reporting statistical significant results. As a result of this, they concluded there was no evidence of publication bias within peer-reviewed literature concerning non-significant results.

In summary, based upon their analysis in 2007, Taylor *et al.* supports the belief that it is preferable that public health policy should introduce or maintain a total ban of smoking in public places. It is hoped that this chapter highlights the need for a routine investigation for publication bias in any meta-analysis, and that there are many tools at a meta-analyst's disposal to do so. Many are straightforward, and for these tools to be accessible to many they should be simple to implement.

## 4 A Robust P-value in Meta-Analysis with Publication Bias

Note that a second version of this chapter was co-authored with J.B. Copas, and subsequently edited for the submission to *Statistics in Medicine* and published in 2008 [20]. A copy of the paper is included in Appendix A1 (page 139).

### 4.1 Motivation

It is often necessary to make assumptions about the selection process when studying publication bias in a meta-analysis. The selection process may be modelled by some parametric function. However, the choice of parametric function would be entirely arbitrary and result in making unverifiable assumptions about the selection process. For this reason it would be desirable to use methods that make as few assumptions as possible. Permutation tests achieve this, and may be utilised to make statistical inferences about publication bias in meta-analysis.

The concept of permutation tests is widely known within the literature, first discussed by Fisher [39] and Pitman [65] - [67]. Permutation tests are also known as randomization tests, re-randomization tests and exact tests. Regardless of their various names, they all essentially work on the following steps [37]. A quantity of interest is to be investigated, with the null and alternative hypotheses stated. A test statistic is chosen and a rejection rule is established to distinguish the null hypothesis with the alternative. Using the data, the test statistic is calculated with the original observations. The main step of the permutation test is to produce a permutation distribution of

the test statistic by calculating all possible values of the statistic under the permutations of the labels of the original observations. The observed value of the statistic is then compared to the distribution to obtain a P-value. This is achieved by finding the proportion of values of the permutation distribution which are as extreme as the observed value of the test statistic.

There are many advantages for the use of permutation tests. First, they are non-parametric statistics, which means the parametric form of the underlying population distribution is not specified explicitly, and therefore would remove substantial assumptions about the selection process in the meta-analysis. Also, by permuting the data, any statistical test (parametric or non-parametric) can be transformed into a distribution-free test. This is a considerable advantage, as the permutation tests do not require specific assumptions such as normality. The P-value itself is very easy to understand, and its computation depends only on calculating a proportion rather than referring to any statistical tables. Permutation tests can be and usually are heavily computational, but with the development of fast and powerful computers over the last thirty years, permutation tests can be easily applied to a range of problems. For detailed discussions about permutation tests and implementing them in practice, see Edgington [30] and Good [37] - [38].

The remainder of this chapter will present the robust P-value for a permutation test in a meta-analysis, as well as providing an approximation to this P-value. Two examples will be discussed to demonstrate the methodology, including a cholesterol lowering dataset, not previously discussed.

## **4.2 Using the Permutation Test in a Meta-Analysis**

The main assumption concerning the use of a permutation test is that observations are exchangeable under the null hypothesis. As an example, suppose we were comparing two treatments in a randomized controlled trial. Under the null hypothesis, this assumption implies that the distribution, from which the data about patients is

drawn, is the same for both treatment arms. Or in other words, this means every patient is the same before sampling and random allocation to treatment groups began. This assumption is central to the following permutation test argument.

Suppose we have  $n$  studies in a meta-analysis each reporting  $(y_i, \sigma_i)$  for  $i = 1, \dots, n$ , where  $y_i$  is the outcome for the  $i^{\text{th}}$  study and  $\sigma_i^2$  is the variance of  $y_i$ . We assume that

$$y_i \sim N(\theta, \sigma_i^2),$$

where  $\theta$  is the outcome of interest. The only assumption we make about selection is that selection can be modelled via its P-value. Assume that the null hypothesis is  $H_0 : \theta = 0$ . Define  $z_i = y_i/\sigma_i$  and the one-tailed P-value for the  $i^{\text{th}}$  study as

$$P_i = \Phi(-z_i).$$

If  $z$  is the vector of observed values of  $z_i$ , then the usual fixed effects meta-analysis estimate of  $\theta$  is

$$\hat{\theta} = \hat{\theta}(z) = \frac{\sum_{i=1}^n v_i z_i}{\sum_{i=1}^n v_i^2},$$

where  $v_i = 1/\sigma_i$ . The meta-analysis P-value is the probability under  $H_0$  that  $\hat{\theta}$  exceeds its observed value.

We assume that the studies included in the meta-analysis constitute a non-random sample from a population of similar studies  $(y, \sigma)$ , where the probability of selection

$$P(\text{selection}|z) = a(z)$$

is modelled by some selection function  $a(z)$ , depending on the study's reported P-value. Under  $H_0$ , the observed values of  $z_i$  are therefore i.i.d. with density

$$f(z) = \frac{a(z)\phi(z)}{\int a(z)\phi(z)dz}. \quad (32)$$

Hence, under  $H_0$ , each member  $Z = (Z_1, \dots, Z_n)$  of the set

$$\mathcal{S} = \{Z | Z \text{ is a permutation of } z\}$$

is equally likely. However, each rearrangement of  $Z_1, \dots, Z_n$  with  $v_1, \dots, v_n$  fixed will produce a different treatment effect  $\hat{\theta}$ . This gives the permutation P-value  $P\{\hat{\theta}(Z) \geq \hat{\theta}(z)|H_0, Z \in \mathcal{S}\}$ , which can easily be shown is equivalent to

$$P\{\hat{\theta}(Z) \geq \hat{\theta}(z)|H_0, Z \in \mathcal{S}\} = P\{\sum \alpha_i Z_i \geq \sum \alpha_i z_i | H_0, Z \in \mathcal{S}\}, \quad (33)$$

where  $\alpha_i = v_i - \bar{v}$ . The values of  $\alpha_i$  are known and fixed, and so (33) can be calculated directed by evaluating  $\sum \alpha_i Z_i$  for all  $n!$  permutations of the observed values of  $z_i$ . According to the concept of permutation tests, the P-value is the proportion of these permutations for which  $\sum \alpha_i Z_i$  exceeds the value of  $\sum \alpha_i z_i$ , when using the observed vector  $z$ .

The key argument is that the observed values of  $z_i$  under  $H_0$  are randomly sampled from the same distribution  $f(z)$  as given in (32). This means that the observed value for each study will be the same under one assignment to  $\alpha_i$  (or equivalently study precision/sample size) compared to any other assignment that could have resulted from the random assignment procedure. Clearly for even moderate values of  $n$  the number of permutations will be incredibly large. Therefore if  $n$  is large, complete enumeration of all permutations can be replaced by sampling random permutations of  $z$ .

### 4.3 A P-value Using a Normal Approximation

It is possible to approximate the permutation P-value as shown in (33) by using a normal approximation. Let  $Z$  be a randomly chosen element of  $\mathcal{S}$ . Then for any fixed  $i$  and  $j$  ( $i \neq j$ ), we have

$$\mathbb{E}[Z_i] = \bar{z} \text{ and } Var(Z_i) = \frac{1}{n} \sum_{i=1}^n (z_i - \bar{z})^2 = s_z^2.$$

Also, it has been easily shown [12] that for  $i \neq j$

$$Cov(Z_i, Z_j) = \frac{1}{n(n-1)} \sum_{a \neq b} (z_a - \bar{z})(z_b - \bar{z}) = -\frac{s_z^2}{n-1}.$$

Since  $\alpha_i = v_i - \bar{v}$ , and so  $\sum \alpha_i = 0$ , it is clear that

$$\mathbb{E}\left[\sum \alpha_i Z_i\right] = 0.$$

Next consider  $Var\left(\sum \alpha_i Z_i\right)$ . First define  $s_v^2$  as

$$s_v^2 = \frac{1}{n} \sum (v_i - \bar{v})^2 = \frac{1}{n} \sum \alpha_i^2.$$

Then

$$\begin{aligned} Var\left(\sum \alpha_i Z_i\right) &= \sum_{i=1}^n \alpha_i^2 Var(Z_i) + \sum_{i \neq j} \alpha_i \alpha_j Cov(Z_i, Z_j) \\ &= s_z^2 \sum_{i=1}^n \alpha_i^2 - \frac{s_z^2}{n-1} \sum_{i \neq j} \alpha_i \alpha_j \\ &= \frac{s_z^2}{n-1} \left\{ (n-1) \sum_{i=1}^n \alpha_i^2 - \sum_{i \neq j} \alpha_i \alpha_j \right\} \\ &= \frac{s_z^2}{n-1} \left\{ n^2 s_v^2 - \sum_{i=1}^n \alpha_i^2 - \sum_{i \neq j} \alpha_i \alpha_j \right\}, \end{aligned}$$

and since

$$\sum_{i,j} \alpha_i \alpha_j = \sum \alpha_i \sum \alpha_j = 0,$$

we have

$$Var\left(\sum \alpha_i Z_i\right) = \frac{n^2 s_z^2 s_v^2}{n-1}.$$

Therefore the asymptotic normal approximation for  $\sum \alpha_i Z_i$  under  $H_0$  is

$$\sum \alpha_i Z_i \sim N\left(0, \frac{n^2 s_z^2 s_v^2}{n-1}\right).$$

The (one-sided) permutation P-value (33) is therefore

$$\Phi\left(-\frac{(n-1)^{1/2} \sum \alpha_i z_i}{n s_z s_v}\right).$$

Since  $a(z)$  is an unspecified selection function, we say that  $z$  is an i.i.d. sample from some distribution, and therefore we have a non-parametric conditional P-value. Furthermore, by observing that  $\sum \alpha_i z_i = \sum (v_i - \bar{v})(z_i - \bar{z})$ , the permutation P-value can be simplified to

$$\Phi(-(n-1)^{1/2} r), \tag{34}$$

where

$$r = \frac{\frac{1}{n} \sum (v_i - \bar{v})(z_i - \bar{z})}{s_v s_z}$$

is the correlation of the observed points  $(z_i, v_i)$  of the radial plot.

Recall that radial plots, also known as Galbraith diagrams [34], plot a study's standardized effect against its precision, or in the notation used here, plot  $z_i$  against  $v_i$ . Briefly, the gradient obtained from drawing a line through the origin to a study corresponds to the study estimate  $y_i$ . Also the gradient of the line constrained through the origin, corresponding to an unweighted regression line, can be interpreted as the conventional fixed effects meta-analysis estimate  $\hat{\theta}$ . Radial plots can be useful in representing the data graphically, exploring possible heterogeneity and identifying possible outliers. This is achieved by forming an approximate 95% confidence region around the regression line.

Recall the P-value for the  $i^{\text{th}}$  study is  $P_i = \Phi(-z_i)$ , implying studies with positive outcomes are more likely to be included in the meta-analysis than those with near zero or negative outcomes. Therefore the form of the approximate P-value in (34) is easy to interpret. A statistically significant P-value will be obtained if the sample correlation  $r$  (between  $z$  and  $v$ ) is large and positive. Also, provided  $n$  is large enough, moderate positive values of  $r$  may be sufficient.

## 4.4 Numerical Examples

### 4.4.1 Cholesterol lowering dataset

Smith *et al.* [76] reviewed 34 randomized controlled trials in a meta-analysis to investigate the effect of cholesterol lowering interventions. First note that the text on systematic reviews by Sutton *et al.* includes the data upon which the following analysis is based [80], and the data is given in Table 4.1 on page 73. Each study reported mortality data for both treatment and control groups. The log(odds ratio) outcome was calculated in each study,  $y_i$ , and the sample variance of  $y_i$ ,  $\sigma_i^2$ , was calculated in

the conventional way. A negative  $\log(\text{OR})$  value suggested treatment was beneficial to lowering cholesterol. Under the assumption of a fixed effects model, a conventional meta-analysis estimate of  $\theta$  suggested that  $\hat{\theta} = -0.166$  with a 95% confidence interval  $(-0.232, -0.105)$ . The corresponding P-value with the null hypothesis  $H_0 : \theta = 0$  is  $1.74 \times 10^{-7}$  suggesting there is very strong evidence to suggest there is a non-zero treatment effect.

A funnel plot for the data is presented in Figure 4.1. Study precision,  $1/\sigma_i$ , is plotted against study outcome  $y_i$ . The dotted line represents the fixed effects meta-analysis estimate of  $\theta$ , which is  $\hat{\theta} = -0.166$ . From the funnel plot, it is clear that there is one study in particular that is much larger (with large precision) showing a beneficial treatment effect.

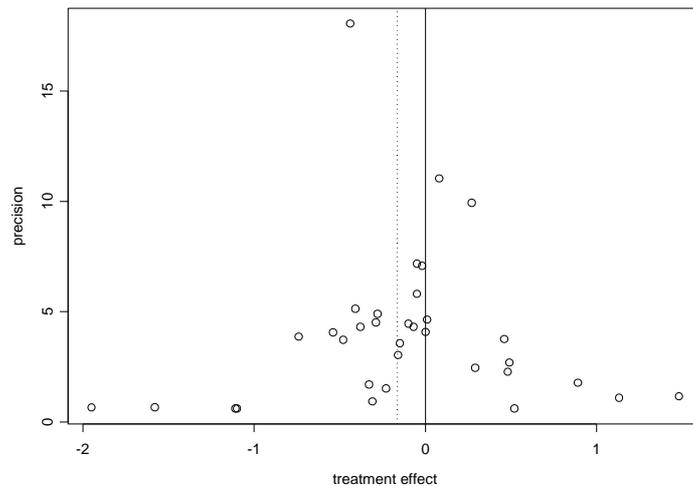


Figure 4.1: Cholesterol lowering dataset: funnel plot.

Clearly the complete enumeration of all  $34!$  permutations would be computationally intensive. The permutation test described in Section 4.2 is implemented by sampling random permutations of  $z$  (the vector of observed  $z_i = y_i/\sigma_i$ ). 100,000 permutations of  $z$  were randomly chosen in S-Plus. The quantity  $\sum \alpha_i Z_i$  was then calculated and compared to the observed value of  $\sum \alpha_i z_i = -98.03$ . In this example we assume

studies with negative outcomes are being selected more frequently than those with positive outcomes. Our null hypothesis was  $H_0 : \theta = 0$  versus the one-sided alternative hypothesis,  $H_1 : \theta < 0$ . The inequality in (33) is therefore reversed, and instead we are interested in the proportion of permutations such that  $\sum \alpha_i Z_i \leq \sum \alpha_i z_i$ . The resulting proportion was **0.02284**, and this is our P-value. We have evidence to reject the null hypothesis at the 5% level suggesting there is a beneficial (negative) treatment effect in lowering cholesterol. The permutation distribution of  $\hat{\theta}$  is presented as a histogram in Figure 4.2. The line represents the observed  $\hat{\theta}(z) = \hat{\theta} = -0.166$ . Clearly the distribution is asymmetrical, with a negative skewed tail, most likely caused by the large, influential study that reported a negative treatment effect.

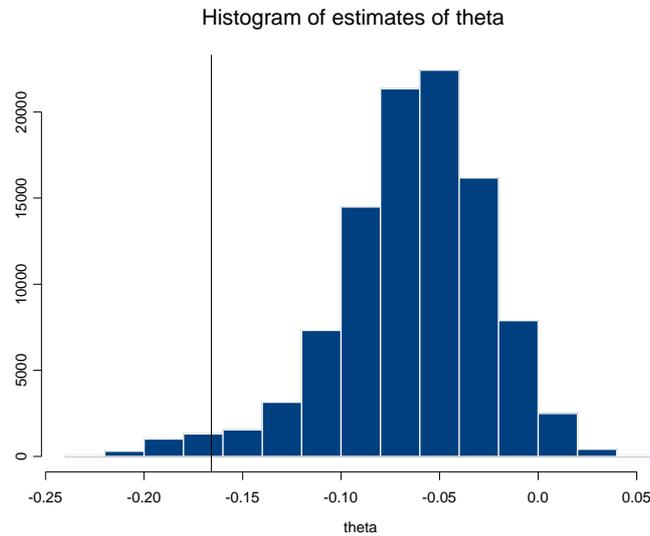


Figure 4.2: Cholesterol lowering dataset: permutation distribution of  $\hat{\theta}$ .

The alternative method to calculate an approximate P-value was presented in section 4.3. The correlation between the observed points  $(z_i, v_i)$  was  $r = -0.459$ . Using (34) we calculate the approximate P-value for the permutation test to be **0.00421**. Using this method, we have very strong evidence to reject the null hypothesis  $H_0 : \theta = 0$ . Figure 4.3 presents the radial plot for the dataset. The solid line represents the meta-analysis estimate of  $\theta$  using a fixed effects model. The negative gradient suggests that treatment is beneficial. The dot-dash lines represent an ap-

proximate 95% confidence region at a distance of two standard errors away from  $\hat{\theta}$ .

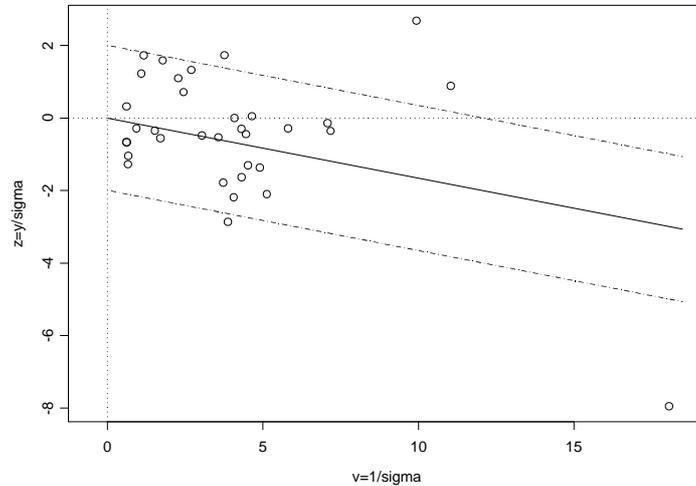


Figure 4.3: Cholesterol lowering dataset: radial plot.

In this example, we do get a highly statistically significant P-value, but the radial plot clearly shows that the fitted regression line through the origin is highly influenced by the very large study in the bottom right of the plot. This point, along with four other studies, lies outside of the approximate 95% confidence region. The presence of heterogeneity appears to be present in this data, and the large study can be considered as an outlier. This could explain why the approximate P-value (0.02284) is noticeably different from the P-value calculated from the permutation distribution (0.00421). The approximation relies upon  $\sum \alpha_i Z_i$  having a symmetrical normal distribution, which in this example (Figure 4.2), it does not.

Since a graphical exploration of the radial plot suggests that there may be heterogeneity present within the dataset, a random effects model is investigated. Applying the standard method of DerSimonian and Laird [26] gives  $\hat{\tau} = 0.0679$ . We assume that  $\tau^2 = \hat{\tau}^2$  is fixed and known. The corresponding random effects analysis gives  $\hat{\theta} = -0.100$  with 95% confidence interval  $(-0.24, 0.04)$ . The corresponding P-value with the null hypothesis  $H_0 : \theta = 0$  is 0.090 suggesting there is no evidence to reject

$H_0$ . The choice of model between fixed and random effects is clearly important since the two models produce contrasting conclusions about the null hypothesis.

Figure 4.4 shows the radial plot when  $z_i = y_i(\sigma_i^2 + \tau^2)^{-1/2}$  is plotted against  $v_i = (\sigma_i^2 + \tau^2)^{-1/2}$ . All studies now lie within the 95% confidence band when taking into account the heterogeneity. As expected, the non-zero  $\tau$  results in the values of  $v_i$  being brought together along the x-axis with less spread. The solid line represents  $\hat{\theta} = -0.1$  using the random effects model. The correlation between the observed points  $(z_i, v_i)$  is now  $r = -0.146$ . Using (34) we calculate the approximate P-value for the permutation test to be **0.4457**. This method suggests that we have no evidence to reject the null hypothesis  $H_0 : \theta = 0$ .

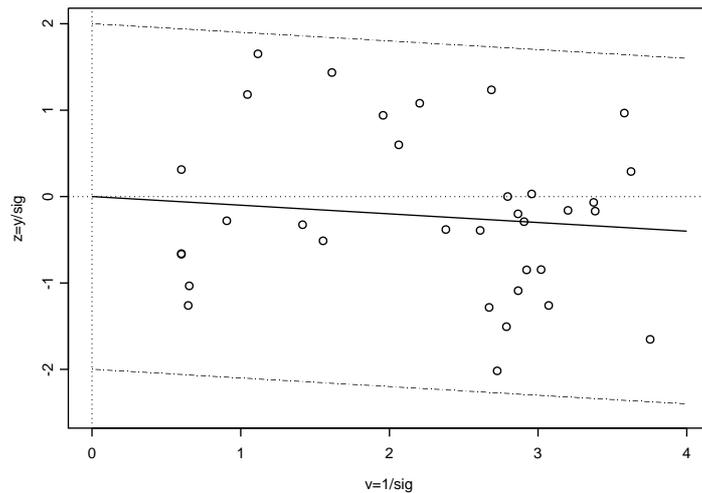


Figure 4.4: Cholesterol lowering dataset: radial plot when assuming a random effects model.

When the permutation test was implemented again by sampling 100,000 random permutations of  $z$  (the vector of observed  $z_i$ ), the quantity  $\sum \alpha_i Z_i$  was calculated and compared to the observed value of  $\sum \alpha_i z_i = -0.72$ . The resulting proportion was **0.447**, and this is our P-value. We have no evidence to reject the null hypothesis. The permutation distribution of  $\hat{\theta}$  is presented as a histogram in Figure 4.5. The line

represents the observed  $\hat{\theta}(z) = \hat{\theta} = -0.1$ . Compared to the histogram under the fixed effects model, this distribution is much more symmetrical. This is because the effect of the original outlier has been reduced by taking the between-study variance into consideration. Notice that the permutation P-value and the normal approximation P-value are very similar. A key assumption to this section of work concerning the random effects model is that we can model selection via the quantity  $z_i = y_i(\sigma_i^2 + \tau^2)^{-1/2}$ .

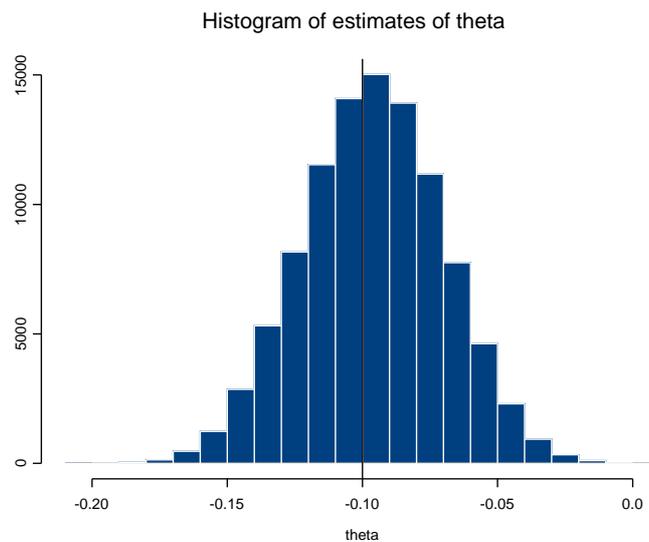


Figure 4.5: Cholesterol lowering dataset: permutation distribution of  $\hat{\theta}$  when assuming a random effects model.

Table 4.1: Cholesterol lowering dataset, where  $y_i$  is the log odds ratio.

Study	Data		95% confidence interval	Study	Data		95% confidence interval
	$y_i$	$\sigma_i$			$y_i$	$\sigma_i$	
1	-0.74	0.265	(-1.25,-0.24)	18	-0.10	0.224	(-0.54,0.34)
2	-0.07	0.224	(-0.53,0.38)	19	-0.23	0.656	(-1.51,1.06)
3	-0.48	0.265	(-1.01,0.04)	20	-0.29	0.224	(-0.72,0.15)
4	-1.48	0.854	(-3.16,0.20)	21	0.46	0.265	(-0.06,0.99)
5	-1.95	1.533	(-4.95,1.06)	22	1.13	0.922	(-0.67,2.94)
6	-0.41	0.200	(-0.79,-0.03)	23	-0.33	0.592	(-1.48,0.83)
7	-0.38	0.224	(-0.84,0.08)	24	0.08	0.100	(-0.10,0.26)
8	-0.16	0.332	(-0.81,0.49)	25	-0.28	0.200	(-0.68,0.12)
9	-0.02	0.141	(-0.30,0.25)	26	-1.11	1.640	(-4.32,2.10)
10	0.00	0.245	(-0.48,0.48)	27	0.49	0.374	(-0.24,1.22)
11	-0.54	0.245	(-1.03,-0.06)	28	-0.05	0.173	(-0.39,0.29)
12	0.48	0.436	(-0.39,1.34)	29	0.01	0.224	(-0.41,0.44)
13	0.29	0.412	(-0.51,1.09)	30	0.89	0.566	(-0.22,1.99)
14	-1.58	1.507	(-4.54,1.37)	31	0.27	0.100	(0.07,0.47)
15	-0.44	0.043	(-0.54,-0.33)	32	-1.10	1.646	(-4.32,2.13)
16	-0.05	0.141	(-0.32,0.23)	33	0.52	1.643	(-2.70,3.74)
17	-0.15	0.283	(-0.70,0.40)	34	-0.31	1.072	(-2.41,1.79)

#### 4.4.2 Passive smoking dataset

The second example returns to the passive smoking dataset, as first reviewed by Hackshaw *et al.* [40]. Refer to Chapter 3 for the dataset. Here, a random effects model was assumed. Note that this assumption means the  $z_i$  are now calculated as  $y_i(\sigma_i^2 + \tau^2)^{-1/2}$ , which is no longer a simple transformation of the study P-values. If  $\tau^2$  is large, we lose the original intuition of the original method.

Recall from the summary table in Chapter 3 (page 43) that  $\hat{\theta} = 0.21$  with a corresponding 95% confidence interval (0.12, 0.31). The permutation test is carried out by sampling 100,000 random permutations of  $z$  in S-Plus. The permutation distribution for the quantity  $\sum \alpha_i Z_i$  was then produced to find the proportion of values for which  $\sum \alpha_i Z_i \geq \sum \alpha_i z_i = -2.22$ . Our null hypothesis is  $H_0 : \theta = 0$  versus the one-sided alternative hypothesis  $H_1 : \theta > 0$ . The resulting proportion was **0.6252**, and this is our P-value. We therefore have no evidence to reject the null hypothesis.

The permutation distribution of  $\hat{\theta}$  is presented in Figure 4.6. The line represents the observed  $\hat{\theta}(z) = \hat{\theta} = 0.21$ . Note that the distribution is clearly symmetric, unlike the cholesterol lowering example in Section 4.4.1.

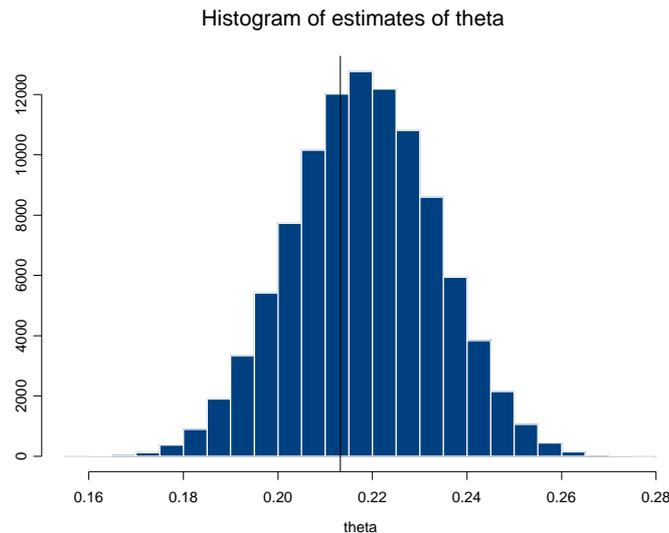


Figure 4.6: Passive smoking dataset: permutation distribution of  $\hat{\theta}$ .

Next consider the approximate P-value. The correlation between  $(z_i, v_i)$  was  $r = -0.0528$ . The approximate P-value for the permutation test was **0.6240**, again suggesting no evidence to reject the null hypothesis. The corresponding radial plot for this dataset is given in Figure 4.7. Note how both the results of obtaining a P-value from both methods are similar. We can relate this robust P-value to the radial plot as discussed in Chapter 3 as follows. The hypothesis test here,  $\theta = 0$ , corresponds to the slope for the fitted line in the radial plot having a gradient of zero. The robust P-value method essentially depends entirely upon whether it is reasonable to constrain the fitted line of the radial plot to have a flat, horizontal slope through the origin. It is clear from Figure 4.7 that this constraint would not fit the data well and are therefore not surprised that the result is non-statistically significant. If we were to remove the anchor through the origin, one can see from Figure 4.7 that the gradient could be anything for this set of data points. Modelling with  $a(z)$  removes this constraint through the origin.

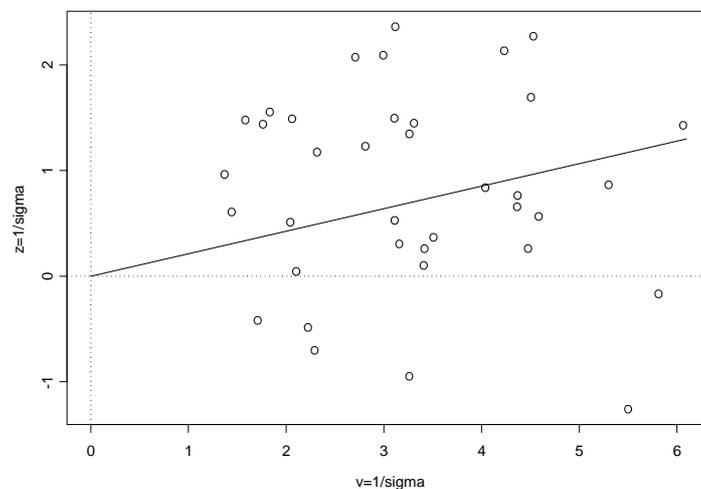


Figure 4.7: Passive smoking dataset: radial plot.

## 4.5 A Comparison Between the Permutation Test and the Linear Regression Test

Egger *et al.* [31] proposed a method of detecting publication bias in a meta-analysis by measuring the level of asymmetry in the funnel plot. A linear regression approach regressed standard normal deviates  $z_i = y_i/\sigma_i$  against study precision  $v_i = 1/\sigma_i$ . The proposed model is

$$\mathbb{E}[z_i] = \alpha + \theta v_i.$$

Recall that Egger *et al.* proposed that the intercept,  $\alpha$ , could provide a measure of asymmetry, namely the larger its deviation from zero, the more pronounced the asymmetry. The slope parameter,  $\theta$ , indicates the size and direction of treatment effect. Therefore, following the conventional methods of standard linear regression, we can perform a statistical test with a null hypothesis  $H_0 : \theta = 0$ . Standard regression results give

$$\hat{\theta} \sim N\left(\theta, \frac{\sigma^2}{n s_v^2}\right), \quad (35)$$

where  $\sigma^2$  is estimated as

$$\hat{\sigma}^2 = \frac{n}{n-2} \left( s_z^2 - \hat{\theta}^2 s_v^2 \right) \quad (36)$$

where  $\hat{\theta}$  is the usual least squares estimate

$$\hat{\theta} = \frac{s_{vz}}{s_v^2},$$

with  $s_v^2 = \sum (v_i - \bar{v})^2$ ,  $s_z^2 = \sum (z_i - \bar{z})^2$  and  $s_{vz} = \sum (v_i - \bar{v})(z_i - \bar{z})$ . Since the correlation between  $(z_i, v_i)$  is given as

$$r = \frac{s_{vz}}{s_v s_z} = \frac{\hat{\theta} s_v}{s_z}$$

we can re-write (36) as

$$\hat{\sigma}^2 = \frac{n s_z^2 (1 - r^2)}{n - 2}. \quad (37)$$

Using (35) and (37) we see that

$$\text{Var}(\hat{\theta}) = \frac{s_z^2 (1 - r^2)}{s_v^2 (n - 2)}.$$

The test statistic  $\hat{\theta}/se(\hat{\theta})$  which we denote here as  $t_2$ , under the null hypothesis is thus calculated as follows:

$$t_2 = \frac{\hat{\theta}}{se(\hat{\theta})} = \frac{s_{vz}s_v(n-2)^{1/2}}{s_v^2s_z(1-r^2)^{1/2}} = \frac{r(n-2)^{1/2}}{(1-r^2)^{1/2}}.$$

Recall from Section 4.3 that the test statistic, which we denote here as  $t_1$ , for the approximation to the permutation test under  $H_0$  is

$$t_1 = (n-1)^{1/2}r.$$

We wish to compare the two test statistics  $t_1$  and  $t_2$ . This is achieved by considering  $t_1/t_2$ , given as

$$\frac{t_1}{t_2} = \sqrt{\frac{n-1}{n-2}}(1-r^2)^{1/2}.$$

Note that

$$\frac{t_1}{t_2} \equiv 1 \Leftrightarrow r = \pm \sqrt{1 - \frac{n-2}{n-1}}.$$

This means the approximation method will provide a smaller value for its test statistics compared to the linear regression method if and only if

$$\sqrt{1 - \frac{n-2}{n-1}} < |r| \leq 1.$$

A meta-analysis with a large number of studies will result in the approximation method providing a smaller test statistic compared to the regression method almost always, unless  $r$  is near zero.

We return to the two numerical examples to compare the permutation test with the regression based method. First consider the cholesterol lowering example. Recall from Section 4.4.1 that the conventional meta-analysis estimate of  $\theta$  was  $\hat{\theta} = -0.166$  with corresponding P-value ( $H_0 : \theta = 0$ ) of  $1.74 \times 10^{-7}$  when using the fixed effects model. The permutation P-value, which we denote here as  $\tilde{P}$ , was **0.02284**, and the approximate P-value, which we denote here as  $\hat{P}$ , was **0.00421**. So with  $\hat{P}$  we have strong evidence to reject the null hypothesis that  $H_0 : \theta = 0$ . We now compare this to the linear regression test. We denote  $P_E$  to be the P-value corresponding to the Egger test ( $H_0 : \alpha = 0$ ), and  $P_{reg}$  to be the P-value corresponding to the significance of the least squares slope of the radial plot ( $H_0 : \theta = 0$ ). Note that the notation for

the various different P-values has followed the notation as set out by the Statistics in Medicine paper as given in Appendix A1. Both  $P_E$  and  $P_{reg}$  are routinely calculated by any regression software, such as S-Plus, following standard linear regression theory. For the cholesterol lowering dataset,  $P_E = 0.2591$ , suggesting there is no evidence of selection bias. Also,  $P_{reg} = \mathbf{0.0064}$ , which is relatively similar to  $\hat{P}$ .

Next consider the passive smoking example. Recall from Section 4.4.2 that the conventional meta-analysis estimate of  $\theta$  was  $\hat{\theta} = 0.213$  with corresponding P-value ( $H_0 : \theta = 0$ ) of  $5.03 \times 10^{-6}$ . The permutation P-value  $\tilde{P} = \mathbf{0.6252}$ , and the approximate P-value  $\hat{P} = \mathbf{0.6240}$ . So for both  $\tilde{P}$  and  $\hat{P}$  we have no evidence to reject the null hypothesis of  $H_0 : \theta = 0$ . We now compare this to the linear regression test. Again, both  $P_E$  and  $P_{reg}$  are routinely calculated within S-Plus, following standard linear regression theory. Here we have  $P_E = 0.0338$ , suggesting there is some doubt about the selection of these studies. Also,  $P_{reg} = \mathbf{0.7561}$ , which is similar to  $\hat{P}$  and clearly both non-significant. Refer to the Statistics in Medicine in Appendix A1 for two further examples comparing the permutation P-value and the regression based method.

## 4.6 Concluding Comments

This chapter forms a basis for a paper published in Statistics in Medicine in 2008, in collaboration with J.B. Copas [20]. A copy is given in Appendix A1 (page 139 onwards), for which some additional technical details and another numerical example for the purposes of illustration can be found. Modelling publication bias in a meta-analysis requires making some assumptions about the selection process, otherwise inference is impossible. The downside to making such assumptions is that it is not possible to verify their validity. In this chapter, we presented a robust non-parametric method which aimed to relax the assumptions about the selection process. The surely plausible and generally widely accepted idea that we adopt here is that selection depends in some unspecified way on a study's P-value.

Two approaches to providing a P-value were presented: the first being the permutation P-value, which was based on standard permutation theory. The basic concept was to permute the  $y_i$  values such that different values of  $\hat{\theta}$  were generated. In practice, taking a sufficiently large sample, the proportion of these different  $\hat{\theta}$  values that were greater than or equal to the observed value of  $\hat{\theta}$  was our P-value. The second approach involved an approximation P-value, which resulted in having quite a simple form, depending on the number of studies in the meta-analysis and the sample correlation of the radial plot.

The concept behind these two approaches is quite elegantly simple. However this trade-off for simplicity and avoiding making strong assumptions about selection comes in the form of loss of power. The published paper as shown in Appendix A1 develops theory about the power functions relating to the approximate robust P-value method, and compares this to the power function from the conventional fixed effects model (which clearly makes more assumptions about selection). There is an inevitable loss of power with the robust P-value, for which the severity of this loss depends mainly on  $\gamma$ , the coefficient of variation of the observed  $v_i$ , defined as  $\gamma = s_v/\bar{v}$ . A small value of  $\gamma$  implies there is a small spread of values along the x-axis of a radial plot, and in these scenarios, the loss of precision is very large. For larger values of  $\gamma$ , the loss of power is still present but less significant.

We conclude this chapter by mentioning that the paper in Appendix A1 includes a different numerical example - namely a meta-analysis of randomised controlled trials of intravenous streptokinase in the prevention of death following myocardial infarction. This example is discussed in the text edited by Egger *et al.* [32] and the data can be found there. Whilst providing an opportunity to present the robust P-value method on a different data set, this meta-analysis is a good example of a dataset where we have a larger value of  $\gamma$  (because we mainly have small studies and a couple of large studies), resulting in the robust P-value still having a loss of power, but not as much as the passive smoking dataset (which had a smaller  $\gamma$ ). Again, full details can be found in the Appendix.

## 5 Applications of Parametric Selection Functions in Meta-Analysis

### 5.1 Introduction

It is becoming increasingly recognised that standard methods in meta-analysis can produce potentially misleading results if certain issues are not addressed. Two such issues are *heterogeneity* and *publication bias*. Heterogeneity refers to variation between studies within a meta-analysis that can not be fully explained by just sampling error alone. It may be inappropriate to combine study results if the studies are not estimating the same quantity of interest.

The issue of publication bias is discussed here. Publication bias is caused by simply assuming that studies within a meta-analysis constitute a random sample of studies from some population of interest. The shared belief is that studies with statistically significant reviews are more likely to be submitted for publication than those with non-significant results [29]. This non-random sampling that is taking place will therefore create bias and in turn pose a serious threat to the validity of the results of the meta-analysis.

As reviewed in Chapter 2, various approaches have attempted to model publication bias in meta-analysis, as reviewed by Sutton *et al.* [80]. One such approach is the use of selection functions. Hedges [43] first introduced selection functions into meta-analysis. Essentially, selection functions model the probability that a study is selected for publication, usually determined by the study's P-value. There are many exam-

ples in the literature of the selection functions taking some kind of parametric form. There usually is an adjustable parameter,  $\beta$ , that models the selection. Since the selection process is unknown, and therefore we know nothing about the value of  $\beta$ , we consider a sensible range of different values of  $\beta$  as part of a sensitivity analysis to investigate how our inferences change.

Following on from Copas and Jackson [16], the selection function takes the form

$$P(\text{selection}|y, \sigma) = a(y, \sigma), \quad (38)$$

for some function  $a(y, \sigma)$ , where it will be assumed that  $y \sim N(\theta, \sigma^2)$ . We let  $p$ , the (unknown) overall selection probability, be defined as  $p = \mathbb{E}[a(y, \sigma)]$ , expectation being over a population  $(y, \sigma)$  of studies. With this definition of  $p$ , and by assuming that  $a(y, \sigma)$  has a parametric form, we will be able to directly assess the likelihood function, and make inferences about a bias-corrected  $\theta$  using a maximum likelihood approach. From this, we will be able to conduct hypothesis tests, and explore datasets by considering likelihood contours. The crucial argument here is that we want a sensitivity analysis for different fixed values of  $p$  since it is impossible to estimate  $a(y, \sigma)$ .

In Section 5.2, a description will be given of the maximum likelihood approach with parametric selection functions. Section 5.3 will focus on selection functions where the adjustable parameter is scalar. Section 5.4 will briefly review the Heckman-type selection model, and re-examine the methods used by Copas and Shi [21]-[23]. Finally, Section 5.5 analyses the effectiveness of the bound for confidence intervals proposed by Henmi *et al.* [47] by comparing the confidence intervals for  $\theta$ , when the selection functions are assumed, and when the Bounds method is used.

Throughout, two examples will be discussed. The first example will be the passive smoking dataset used by Hackshaw *et al.* [40] concerning the relationship between passive smoking and lung cancer. We include a second example to demonstrate the methods used - a dataset that has not yet been discussed in this thesis. The data relates to the effectiveness of prophylactic corticosteroids, an example which was first

discussed in Copas and Jackson [16].

## 5.2 Using Parametric Selection Functions

Our basic model is as follows:

$$\begin{cases} y|\sigma \sim N(\theta, \sigma^2) \\ \sigma \sim f(\sigma) \end{cases}$$

where  $\theta$  is the outcome of interest and  $\sigma^2$  is the variance of  $y$ . We suppose  $\sigma$  to be random with (unknown) distribution  $f(\sigma)$ . We model the selection process with the selection function  $a(y, \sigma)$ , where

$$a(y, \sigma) = P(\text{selection}|y, \sigma).$$

Some of the following definitions here were first mentioned in Section 2.5.2 (page 22) which we include here to introduce the subsequent theory. Define  $a(\sigma)$  as

$$a(\sigma) = P(\text{selection}|\sigma) = \mathbb{E}[a(y, \sigma)|\sigma] = \int_{-\infty}^{\infty} a(y, \sigma) \frac{1}{\sigma} \phi\left(\frac{y - \theta}{\sigma}\right) dy,$$

where  $\phi$  is the density of the standard normal distribution.

The joint distribution of  $(y, \sigma)$  for a selected study,  $g_o(y, \sigma)$ , is given as

$$g_o(y, \sigma) = P(y, \sigma|\text{selection}) = \frac{1}{p\sigma} a(y, \sigma) \phi\left(\frac{y - \theta}{\sigma}\right) f(\sigma), \quad (39)$$

where  $p$  is the overall selection probability

$$\begin{aligned} p = P(\text{selection}) &= \mathbb{E}[a(\sigma)] = \int_0^{\infty} a(\sigma) f(\sigma) d\sigma & (40) \\ &= \mathbb{E}[a(y, \sigma)] = \int_0^{\infty} \int_{-\infty}^{\infty} a(y, \sigma) \frac{1}{\sigma} \phi\left(\frac{y - \theta}{\sigma}\right) f(\sigma) dy d\sigma & (41) \end{aligned}$$

The distribution of  $\sigma$  for a selected study,  $f_o(\sigma)$ , is given as

$$f_o(\sigma) = \frac{a(\sigma) f(\sigma)}{p}. \quad (42)$$

Re-arranging (42) for  $f(\sigma)$ , it is possible to eliminate  $p$  in equation (39). Note also, by doing so,  $g_o(y, \sigma)$  will be written in terms of  $f_o(\sigma)$  rather than  $f(\sigma)$ . We therefore

have

$$g_o(y, \sigma) = \frac{a(y, \sigma) \frac{1}{\sigma} \phi\left(\frac{y-\theta}{\sigma}\right) f_o(\sigma)}{a(\sigma)}. \quad (43)$$

In our meta-analysis, we have data  $\{(y_i, \sigma_i) : i = 1, \dots, n\}$ . Note that for an assumed fixed effects model,  $\sigma_i^2$  is simply  $s_i^2$ , the observed within study variance of  $y_i$ . If a random effects model is used, we take  $\sigma_i^2 = s_i^2 + \tau^2$ , where  $\tau^2$  is the between-study variance. We suppose the selection function is parametric, namely,  $a(y, \sigma; \beta)$ . So the likelihood function under model (43) with a parametric selection function is

$$\begin{aligned} L(\theta, \beta) &= \prod_{i=1}^n g_o(y_i, \sigma_i; \beta, \theta) \\ &= \prod_{i=1}^n \frac{a(y_i, \sigma_i; \beta) \frac{1}{\sigma_i} \phi\left(\frac{y_i-\theta}{\sigma_i}\right) f_o(\sigma_i)}{a(\sigma_i; \beta, \theta)}. \end{aligned}$$

Looking at the log-likelihood function, we have

$$l(\theta, \beta) = \sum_{i=1}^n \left\{ \log a(y_i, \sigma_i; \beta) + \log \phi\left(\frac{y_i - \theta}{\sigma_i}\right) + \log f_o(\sigma_i) - \log \sigma_i - \log a(\sigma_i; \beta, \theta) \right\}. \quad (44)$$

The maximum likelihood estimate of  $f_o(\sigma)$  is the discrete distribution putting probability  $\frac{1}{n}$  on each of the observed  $\sigma_1, \dots, \sigma_n$ . Therefore it is easy to show that

$$\sum_{i=1}^n \log f_o(\sigma_i) \leq -n \log n.$$

Using this fact, to maximise the log-likelihood function in (44), it is sufficient to maximise

$$\sum_{i=1}^n \left\{ \log a(y_i, \sigma_i; \beta) + \log \phi\left(\frac{y_i - \theta}{\sigma_i}\right) - \log \sigma_i - \log a(\sigma_i; \beta, \theta) \right\}, \quad (45)$$

with the constraint that  $p = \mathbb{E}[a(y, \sigma)]$ . From a practical point of view, the constraint in this form is not so helpful, since it relies upon the unknown distribution of  $\sigma$ ,  $f(\sigma)$ , as shown in equation (41). However, we can re-write this constraint as the following.

RESULT: In terms of  $f_o(\sigma)$  and  $a(y, \sigma)$ ,

$$p = \left\{ \mathbb{E}_o \left[ \frac{1}{a(\sigma)} \right] \right\}^{-1}. \quad (46)$$

This results from the fact that for any function  $h(\sigma)$ ,

$$\mathbb{E}_o[h(\sigma)] = \frac{1}{p}\mathbb{E}[a(\sigma)h(\sigma)].$$

The proof of this is straightforward.

$$\begin{aligned}\mathbb{E}_o[h(\sigma)] &= \int_0^\infty h(\sigma)f_o(\sigma)d\sigma = \int_0^\infty h(\sigma)\frac{a(\sigma)f(\sigma)}{p}d\sigma \\ &= \frac{1}{p}\int_0^\infty (h(\sigma)a(\sigma))f(\sigma)d\sigma = \frac{1}{p}\mathbb{E}[a(\sigma)h(\sigma)]\end{aligned}$$

So we simply take  $h(\sigma) = \frac{1}{a(\sigma)}$  and the result in equation (46) follows. In practice, the constraint in (46) becomes

$$p = \left\{ \frac{1}{n} \sum_{i=1}^n \frac{1}{a(\sigma_i; \theta, \beta)} \right\}^{-1}. \quad (47)$$

For a specified parametric selection function,  $a(y, \sigma, \beta)$ , and for a sensible range of values of  $p$ , we calculate the profile likelihood by using (45). The general method that we shall use is as follows. We fix the value of  $p$ . Then for each value of  $\theta$ , we find the set of values of  $\beta$ ,  $\mathcal{B}_{\theta,p}$  say, such that (47) is satisfied. We then have the log-likelihood  $l(\theta, \beta : \beta \in \mathcal{B}_{\theta,p}) = l(\theta, \beta)$ . Define  $l^*(\theta, p)$  as

$$l^*(\theta, p) = \max_{\beta \in \mathcal{B}_{\theta,p}} l(\theta, \beta).$$

We compute numerically to find the maximum likelihood estimate of  $\theta$ , which is the value  $\hat{\theta}$ , such that

$$l(\hat{\theta}) = \max_{\theta} l^*(\theta, p).$$

We repeat this process for different fixed values of  $p$ . This enables us to see what effect assuming a specific parametric selection function will have on the estimate of  $\theta$  for a range of different values of the overall selection probability,  $p$ . In addition to this, we can use this method to find an approximate confidence interval for  $\theta$ . This is achieved by equating

$$2\left(\max_{\theta} l^*(\theta, p) - l^*(\theta, p)\right)$$

to a percentage point of the  $\chi_1^2$  distribution. Two meta-analyses will be used throughout the following sections to demonstrate this method.

## 5.3 Numerical Examples

In this section we implement the method as described previously with the aid of two meta-analyses. The passive smoking dataset was extensively discussed in Chapter 3. We present the second example briefly here, before introducing some examples of selection functions that have a scalar  $\beta$  parameter, and then use these to perform a sensitivity analysis.

### 5.3.1 Passive smoking dataset

Recall from Chapter 3 that Hackshaw *et al.* found a significant increased risk of lung cancer for those exposed to passive smoking compared to those that did not. Since there was evidence of heterogeneity within the meta-analysis, by using the method of DerSimonian and Laird [26], the overall result was that  $\hat{\theta} = 0.21$  with a 95% confidence interval  $(0.12, 0.30)$ . For the remainder of this section we shall use the random effects model, where we use the estimate of  $\tau^2 = 0.017$  and fix it as known.

### 5.3.2 Prophylactic corticosteroids dataset

Copas and Jackson [16] looked at the results of 14 randomised clinical trials concerning the use of prophylactic corticosteroids in cases of premature birth. The data was taken from the Cochrane database. Treatment is administered to the mother if a premature birth is anticipated. The events are the deaths of the infants. The quantity of interest here is  $\theta$ , the log-odds ratio comparing the probability of death in the treatment group with the probability of death in the control group.

For each of the 14 studies, an estimate of the log-odds ratio,  $y_i$ , was calculated, along with  $s_i$ , the standard error of  $y_i$ . Using this data, we calculate  $\hat{\theta}$  and a corresponding 95% confidence interval for  $\theta$  in the usual way. Assuming a fixed effects model,  $\hat{\theta} = -0.48$  with 95% confidence interval  $(-0.71, -0.25)$ . The interpretation of this is that treatment is effective in reducing mortality, remembering that we are comparing

probability of infant deaths.

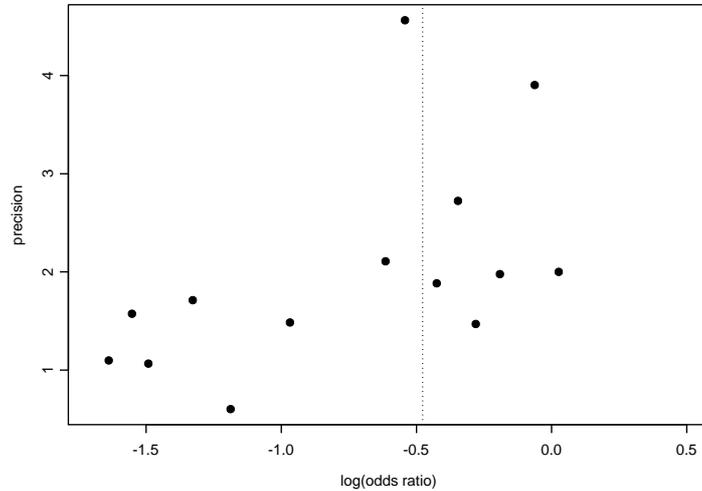


Figure 5.1: Corticosteroids dataset: funnel plot.

A funnel plot of the data is presented in Figure 5.1. The dotted line represents the estimate  $\hat{\theta} = -0.48$ . The plot exhibits the classic asymmetry, where the data is skewed to the left of the plot, ie. the smaller studies with negative values are being reported more often than those with positive values. It is possible that there are missing studies to the right of the plot. Therefore, we suspect publication bias may be present. For the remainder of this section we continue to use the fixed effects model. This is because the method given in DerSimonian and Laird [26] calculated  $\hat{\tau}^2 = 0$ .

### 5.3.3 The selection functions

We suppose that the selection of a study for inclusion in a meta-analysis is biased through some function  $a(y_i, \sigma_i; \beta)$ .  $\beta$  is the parameter measuring the strength of selection. The choice of parametric functions is entirely arbitrary since it is impossible to know the shape of the underlying selection process. We model selection using the

one-tailed and two-tailed P-values, respectively

$$v(y, \sigma) = \Phi(-y/\sigma), \text{ and} \quad (48)$$

$$v(y, \sigma) = 2\Phi(-|y|/\sigma). \quad (49)$$

The following three selection functions are considered.

$$\text{Exponential} \quad a(y, \sigma; \beta) = e^{-\beta v(y, \sigma)}, \quad (50)$$

$$\text{Half-normal} \quad a(y, \sigma; \beta) = e^{-\beta v^2(y, \sigma)}, \quad (51)$$

$$\text{Logistic} \quad a(y, \sigma; \beta) = \frac{2e^{-\beta v(y, \sigma)}}{(1 + e^{-\beta v(y, \sigma)})}. \quad (52)$$

Note that the meaning of  $\beta$  is different in all three selection functions. Notice also as the P-value increases, the probability of publication, ie. the weight of the study, decreases. This intuitively makes sense, since studies reporting highly significant results ( $v \leq 0.01$ , say) will almost certainly be published. Studies reporting non-significant results are less likely to be published, and less weight will be given to it.

A few points should be mentioned before discussing the results. Refer back to Section 5.2 for the description of the method to calculate  $\hat{\theta}$  and the confidence intervals for  $\theta$ . The notation  $a_1, \dots, a_6$  will be used for the six selection functions as shown in Table 5.1.

Selection function	Description
$a_1$	Exponential, One-tail
$a_2$	Exponential, Two-tail
$a_3$	Half-normal, One-tail
$a_4$	Half-normal, Two-tail
$a_5$	Logistic, One-tail
$a_6$	Logistic, Two-tail

Table 5.1: Notation for the selection functions used.

The sign of  $y/\sigma$  in the expression for the one-tailed P-value in (48) needs a moment of consideration. Suppose there is suspicion that studies with positive  $y$  are not being selected (as in the case of the corticosteroids example). To model this, the one-tailed P-value would need to be  $\Phi(y/\sigma)$ . So now, large positive  $y/\sigma$  results in a high P-value. Note that, as it is written, the P-value in (48) assumes the suppression occurs for negative values of  $y/\sigma$ .

When calculating  $\hat{\theta}$  and the 95% confidence intervals of  $\theta$ , it was decided, for the overall selection probability  $p$ , to only consider  $p \in [0.3, 1]$ . Arguably, it seems unlikely in a real world setting that the overall proportion of studies selected in a particular area of interest would be less than 0.5, even more so for  $p < 0.3$ . Therefore, in the following plots, only the range  $p \in [0.3, 1]$  is considered.

### 5.3.4 Example 1: passive smoking dataset

We consider the passive smoking dataset as our first example. Figure 5.2 on page 89 shows 6 plots, corresponding to the six selection functions considered. Each graph plots  $\hat{\theta}$  against  $p$ , with corresponding 95% confidence intervals. The first thing to point out is that in all cases, when  $p = 1$ ,  $\hat{\theta} = 0.21$  with 95% confidence interval (0.12, 0.30). As we would expect, this matches perfectly to the analysis as given in Chapter 3.  $p = 1$  means no studies are being excluded, and so the standard meta-analysis estimates apply.

Figure 5.2 shows that, as  $p$  decreases from 1, the confidence intervals consistently widen faster when a one-tailed P-value is used compared to a two-tailed P-value. It is also clear that the choice of selection function is important in making inferences about  $\theta$ , and can yield quite contrasting results. For example, the gradient of the line representing  $\hat{\theta}$  in Figure 5.2(ii) falls slowly from  $\hat{\theta} = 0.21$  as  $p$  decreases from 1. Compare this with Figure 5.2(iii) where the gradient of the line representing  $\hat{\theta}$  descends at a faster rate as  $p$  decreases from 1. If we consider the case when  $p = 0.6$ , the 95% confidence intervals for  $a_2$  and  $a_3$  are  $(-0.02, 0.24)$  and  $(0.05, 0.25)$  respec-

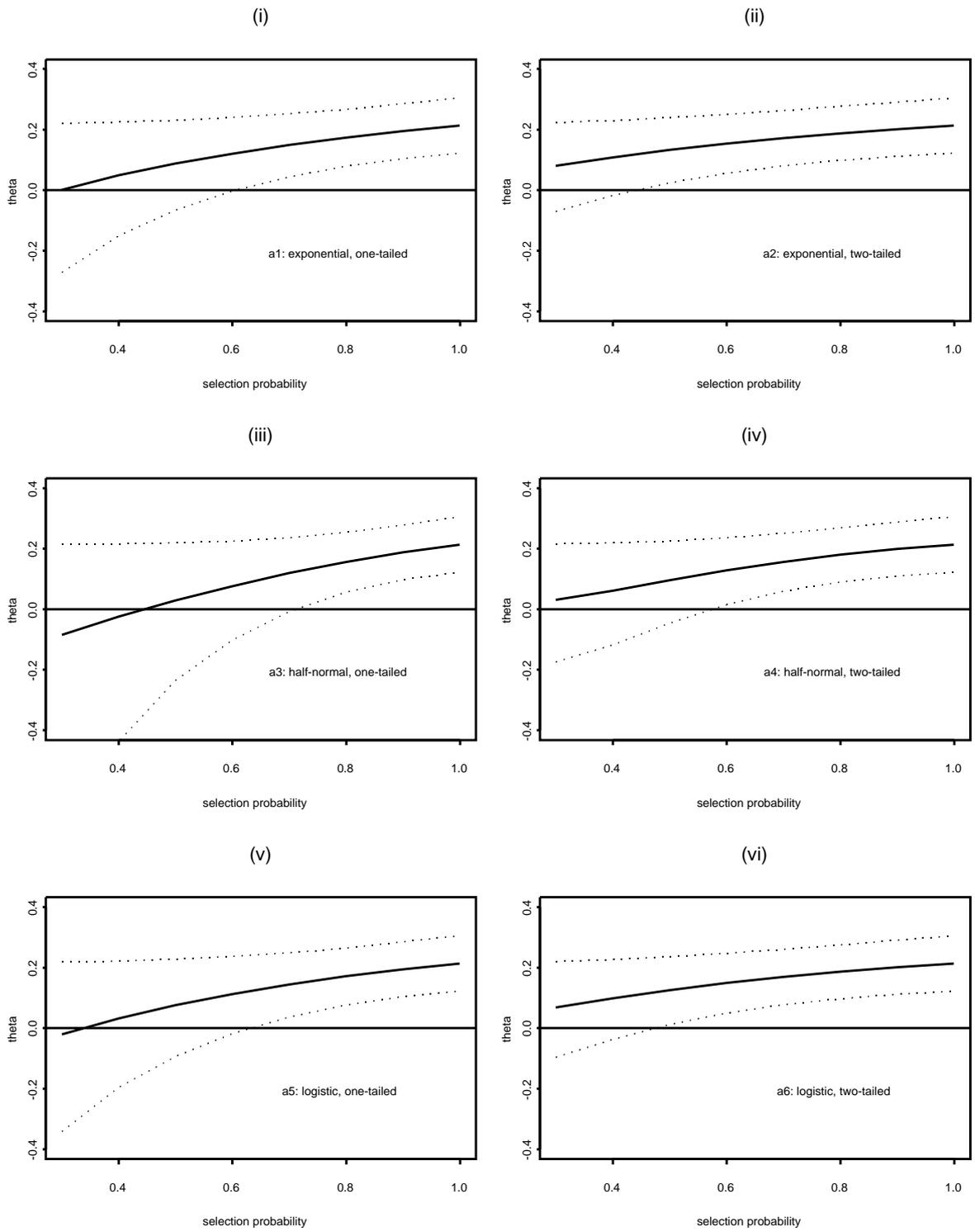


Figure 5.2: Passive smoking dataset:  $\hat{\theta}$  and 95% confidence intervals of  $\theta$  when assuming different selection functions  $a_1, \dots, a_6$  for a range of  $p$ .

tively. Imagine we were to entertain the assumption that only 60% of all passive smoking studies were published in this meta-analysis. Then using these two different selection functions to model the selection procedure, we would reach two contradicting conclusions about the existence of a relationship between passive smoking and lung cancer.

Of particular interest is when the confidence intervals widen enough to include the value zero. From Figure 5.2, we note at what value of  $p$  does the lower limit of the confidence interval cross the line  $\theta = 0$ . These values are given in Table 5.2.

Selection function	Approximate overall selection probability	Approximate number of missing studies
$a_1$	61%	24
$a_2$	44%	46
$a_3$	71%	15
$a_4$	58%	27
$a_5$	63%	21
$a_6$	48%	41

Table 5.2: Passive smoking dataset: table of approximate number of missing studies for the confidence intervals to include zero when assuming different selection functions  $a_1, \dots, a_6$ .

We can approximate  $p$  with  $\hat{p} = \frac{n}{n+m}$ , where  $m$  can be interpreted as the (approximate) number of missing studies and, for this data set,  $n = 37$ . The percentages included in the table represent the upper most value of the overall selection probability such that the 95% confidence interval of  $\theta$  would include zero. Likewise, the values given for the missing studies in the table represent the smallest number of studies that would be necessary to cast doubt on the validity of the results of the meta-analysis. Most noticeable in Table 5.2 is the wide range of values of missing studies. The smallest number of missing studies necessary is 15, whereas the largest

number is 46. It is important at this point to stress that this is a sensitivity analysis. We are taking a plausible range of values for  $p$ , and examining what effect various selection functions will have on the inference of  $\theta$ . Since the percentages of the overall selection vary between 44% and 71%, we conclude that the inferences for  $\theta$  for this dataset is sensitive to the choice of parametric selection function. In other meta-analyses, it may be such that the percentages of the overall selection may be much more consistent, and one could then argue in that scenario that the choice of selection function would be less critical.

### 5.3.5 Example 2: corticosteroids dataset

We now consider the corticosteroids data as our second example. Figure 5.3 on page 92 shows six plots, corresponding to the six selection functions considered. Each graph plots  $\hat{\theta}$  against  $p$ , with corresponding 95% confidence intervals. Again, we notice that in all cases, when  $p = 1$ ,  $\hat{\theta} = -0.48$  with 95% confidence interval  $(-0.71, -0.25)$ . This matches the standard meta-analysis estimates as given in Section 5.3.2.

Figure 5.3 shows the estimate of  $\theta$  in all six cases increases slowly as  $p$  decreases from 1. The confidence intervals remain approximately consistent in width as  $p$  varies, with the slight exception when  $a_5$  is used for small values of  $p$  in Figure 5.3(v). The effect of using a one-tailed or two-tailed  $P$ -value does not appear to change the inferences about  $\theta$  in this example. Comparing the pairs of plots (one-tailed versus two-tailed versions) for the exponential selection function and so on, the shapes of the confidence intervals look similar and so will produce similar inferences. Again, we pay close attention to when the confidence intervals widen enough to include the value zero. From Figure 5.3 we note at what value of  $p$  does the upper limit of the confidence interval cross the line  $\theta = 0$ . These values are given in Table 5.3 on page 93.

The range of percentages in Table 5.3 are more consistent than in the passive smoking example. This agrees with the similar looking plots in Figure 5.3. As discussed earlier, it seems unlikely that the overall selection proportion would be as low as 20%

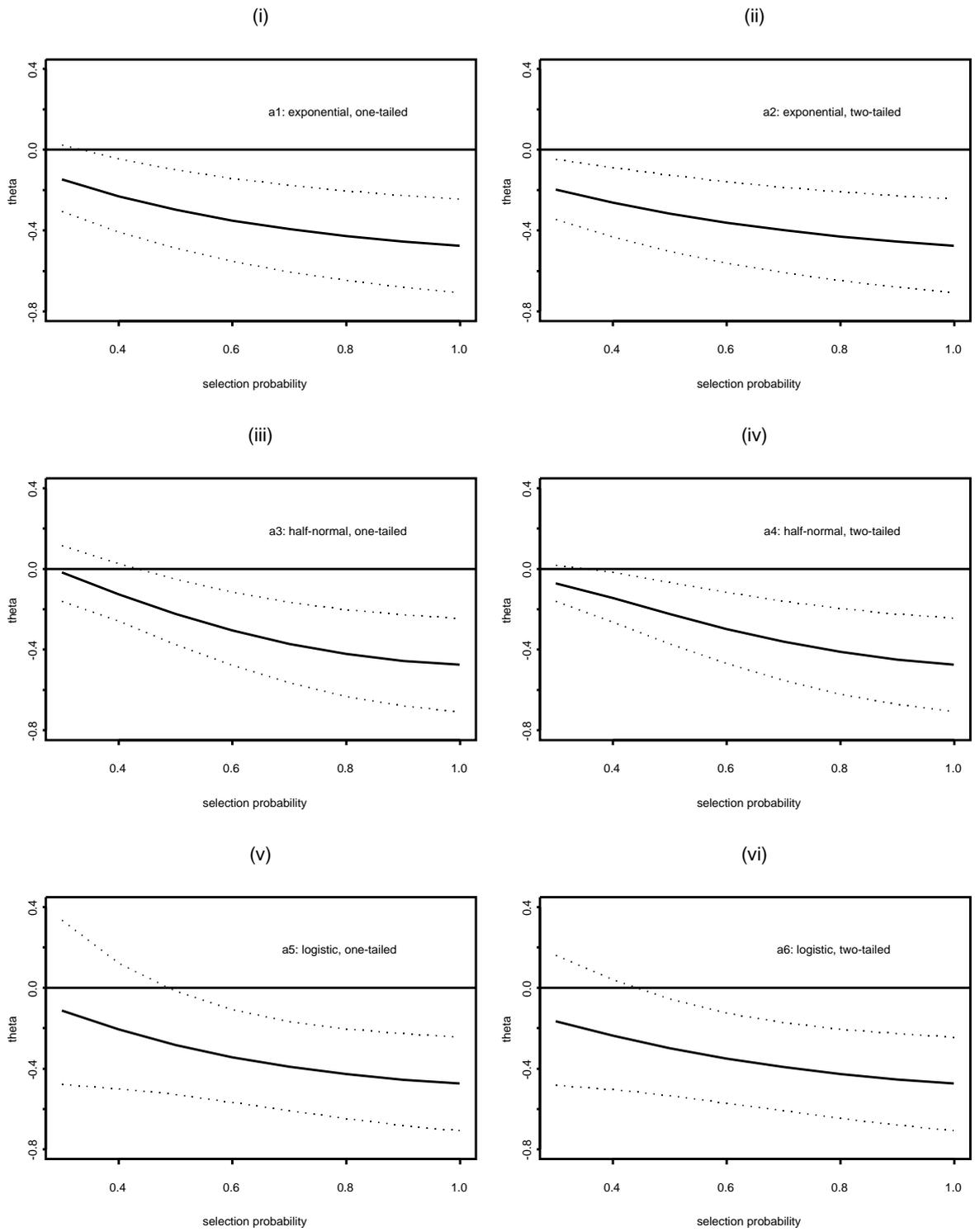


Figure 5.3: Corticosteroids dataset:  $\hat{\theta}$  and 95% confidence intervals of  $\theta$  when assuming different selection functions  $a_1, \dots, a_6$  for a range of  $p$ .

or 40%. Out of the six selection functions, the lowest number of missing studies necessary to reverse the significance of the analysis is 16. In some sense, it seems unlikely that there would be so many unpublished studies of this kind in existence, when only 14 have been published. In the uppermost value of missing studies, it seems surely implausible that there would be 68 studies that have not for some reason been included within this meta-analysis.

Selection function	Approximate overall selection probability	Approximate number of missing studies
$a_1$	33%	28
$a_2$	17%	68
$a_3$	43%	19
$a_4$	34%	27
$a_5$	47%	16
$a_6$	43%	19

Table 5.3: Corticosteroids data: table of approximate number of missing studies for the confidence intervals to include zero when assuming different selection functions  $a_1, \dots, a_6$ .

## 5.4 Generalising the Parametric Selection Functions Approach

So far the selection functions discussed have only included a scalar adjustable parameter,  $\beta$ . This approach of maximising the likelihood function numerically can be extended to include a much more flexible family of selection functions. We now consider when  $\beta$  is a vector of parameters.

The same method applies as before. Assuming the selection process can be modelled with a specific selection function, we simply aim to find the mle of  $\theta$ , maximising over all components of  $\beta$ , for a given value of  $p$ .

The following section has two aims. First, this section aims to illustrate how the general approach of maximum likelihood and parametric selection functions can be extended to include  $\beta$  as a vector of parameters. Second, and more importantly, this section re-evaluates the themes and approaches that were discussed by Copas and Shi [21]. The selection function they consider has been discussed in other works, [15] and [22], and within this thesis the following selection function was first discussed in Section 2.4.1. A review of the main points are given below. Note that the notation has been slightly modified to agree with the notation used in this thesis.

#### 5.4.1 Description of the model

We suppose there are  $n$  studies to be reviewed.  $y_i$  is the estimated treatment effect, and  $s_i$  is the reported standard error for the  $i^{\text{th}}$  study.  $\theta$  is the overall mean effect,  $\tau^2$  is the between-study variance and  $\sigma_i^2$  is the within-study variance. The model is as follows.

$$y_i = \theta_i + \sigma_i \epsilon_i, \quad (53)$$

with  $\epsilon_i \sim N(0, 1)$  and  $\theta_i \sim N(\theta, \sigma_i^2)$  for  $i = 1, \dots, n$ . Note that it is assumed that  $\epsilon_i$  and  $\theta_i$  are independent. Also, assume that  $y_i$  and  $s_i$  are independent.

Following on from Copas and Li [19] and Copas [15], a selection equation is introduced. A correlation parameter,  $\rho$ , aids in modelling the selection.

$$z_i = a + b/s_i + \delta_i, \quad (54)$$

where  $\delta_i \sim N(0, 1)$  and  $\text{corr}(\epsilon_i, \delta_i) = \rho$ . The residuals  $(\epsilon_i, \delta_i)$  are assumed jointly normal.

The purpose of equation (54) is that  $y_i$  will only be observed if the latent variable  $z_i > 0$ . The observed treatment effects within the meta-analysis are modelled by the conditional distribution of  $y_i | z_i > 0$ . The parameters,  $a$  and  $b$ , in equation (54) control the probability that a specific study, with reported standard error  $s_i$ , is published. The parameter  $a$  controls the overall proportion of studies published, while

$b$  controls how publication depends upon study size. We expect  $b > 0$ , so that this ties together the surely plausible assumption that large studies (small values of  $s_i$ ) are more likely to be published than smaller studies. Note also that  $\rho > 0$  implies that smaller studies that are accepted are more likely to be those with large values of  $y_i$ .

We re-write equations (53) and (54) slightly as follows.

$$y_i = \theta + (\sigma_i^2 + \tau^2)^{1/2} \epsilon_i^* \quad (55)$$

$$z_i = a + b/s_i + \delta_i \quad (56)$$

We assume  $\epsilon_i^* \sim N(0, 1)$  and that

$$\text{corr}(\epsilon_i^*, \delta_i) = \tilde{\rho}_i = \frac{\sigma_i}{(\tau^2 + \sigma_i^2)^{1/2}} \rho.$$

We are now in a position to write down the log-likelihood.

$$\begin{aligned} L(\theta, \rho, \tau, a, b) &= \sum_{i=1}^n \log p(y_i | z_i > 0, s_i) \\ &= \sum_{i=1}^n \left[ -\frac{1}{2} \log(\tau^2 + \sigma_i^2) - \frac{(y_i - \theta)^2}{2(\tau^2 + \sigma_i^2)} - \log \Phi(u_i) \right. \\ &\quad \left. + \log \Phi(v_i) \right], \end{aligned} \quad (57)$$

where  $u_i = a + b/s_i$  and

$$v_i = \frac{u_i + \tilde{\rho}_i \frac{y_i - \theta}{(\tau^2 + \sigma_i^2)^{1/2}}}{(1 - \tilde{\rho}_i^2)^{1/2}}.$$

Note that we replace  $\sigma_i^2$  with  $s_i^2$ , assuming sufficiently large sample sizes in each study.

The above model is an example of a parametric selection function, where  $\beta$  is a vector. For the remainder of the remainder of the thesis, denote this selection function as  $a_7$ .

$$a_7(y_i, \sigma_i) = \Phi(v_i) = \Phi \left( \frac{a + b/s_i + \frac{\sigma_i \rho (y_i - \theta)}{(\tau^2 + \sigma_i^2)}}{\sqrt{1 - \frac{\sigma_i^2 \rho^2}{(\tau^2 + \sigma_i^2)}}} \right) \quad (58)$$

To make clear,  $\beta = (a, b, \rho, \tau)$ . For this specific selection function,

$$a_7(\sigma_i) = \Phi(u_i) = \Phi(a + b/s_i). \quad (59)$$

We now proceed as in the previous section. We aim to maximise the likelihood given in equation (57) for a given overall selection probability,  $p$ , subject to the constraint

$$\frac{1}{p} = \mathbb{E}_o \left[ \frac{1}{a_7(\sigma)} \right] = \frac{1}{n} \sum_{i=1}^n \frac{1}{a_7(\sigma_i)}.$$

It is important to comment that the approach that we adopt here is different to the approach by Copas and Shi [21], largely because they perform a sensitivity analysis fixing the pair of values  $(a, b)$  whereas here we fix only the value of  $p$ . One could argue that this approach is more straightforward and makes the sensitivity analysis easier to interpret.

#### 5.4.2 Hypothesis tests and confidence intervals

Before the inclusion of any examples, we first describe what hypothesis tests and plots will be carried out in the analysis.

- The profile likelihood can be plotted against  $p$ . Recall from Section 5.2 that the profile likelihood is calculated as

$$\max_{\theta} l^*(\theta, p),$$

where  $l^*(\theta, p)$  was defined as

$$l^*(\theta, p) = \max_{\beta \in \mathcal{B}_{\theta, p}} l(\theta, \beta).$$

- The following hypothesis test can be performed.  $H_0 : p = 1$  versus  $H_1 : p < 1$ . A value of  $p = 1$  implies that all studies are being included within the meta-analysis and that there is no selection occurring. The following likelihood ratio test statistic is used.

$$X^2 = 2 \left( \max_{\theta, p} l^*(\theta, p) - \max_{\theta} l^*(\theta, 1) \right)$$

Compare  $X^2$  to a  $\chi_1^2$  distribution. At the 5% level of significance, reject  $H_0$  if  $X^2 > 3.84$ .

- A contour plot of  $l^*(\theta, p)$  can be produced to look at the profile likelihood for different values of  $\theta$  and  $p$ .

- With the aid of normalising the plot of  $l^*(\theta, p)$ ,  $\hat{\theta}$  and 95% confidence intervals can be plotted against  $p$ . Of particular interest will be for what values of  $p$  do the confidence intervals include zero.
- Related to the confidence intervals plot, a hypothesis test can be performed concerning  $\theta$ . For a fixed value of  $p$ , we have  $H_0 : \theta = 0$  versus  $H_1 : \theta \neq 0$ . This tests, as an example, whether there is evidence of a significance treatment effect. The following likelihood ratio test statistic is used.

$$X^2 = 2 \left( \max_{\theta} l^*(\theta, p) - l^*(0, p) \right)$$

Compare  $X^2$  to a  $\chi_1^2$  distribution.

We maximise all these functions numerically. The vector of parameters,  $\beta = (a, b, \rho, \tau)$ , has some restrictions. When we search for the maximum over the parameter space,  $a \in (-\infty, \infty)$ ,  $b > 0$ ,  $\rho \in [-0.999, 0.999]$  to avoid the singularities at  $\pm 1$ , and  $\tau \in (0, \infty)$ .

### 5.4.3 Example: passive smoking dataset with a fixed effects model

We return to the passive smoking dataset. First assume that the fixed effects model is appropriate, so that  $\tau^2 = 0$ . Note that  $\sigma_i = s_i$  in (45) as given in Section 5.2. For a range of values of  $p$ ,  $\theta$  can be estimated by maximum likelihood subject to the constraint  $\mathbb{E}[a(y, \sigma)]$ . Figure 5.4 shows the profile log-likelihood for  $p$ . When  $p$  is large and approaches 1, the likelihood falls sharply. This suggests these values of  $p$  are not acceptable. Instead, the likelihood reaches its maximum when  $p$  becomes small. The data fits the model when we assume overall selection probability is small. There is a noticeable dip in the graph when  $p$  is approximately 0.7 to 0.9. Whilst calculating the likelihood,  $\hat{\rho} \approx -0.9$ , which lead to numerical difficulties. The somewhat erratic behaviour seen in the likelihood plot was also noted by Copas and Shi [21].

In Figure 5.4, of particular interest is the case when  $p = 1$ . Using a likelihood ratio test, we test the null hypothesis,  $H_0 : p = 1$  versus  $H_1 : p < 1$ . We compare the

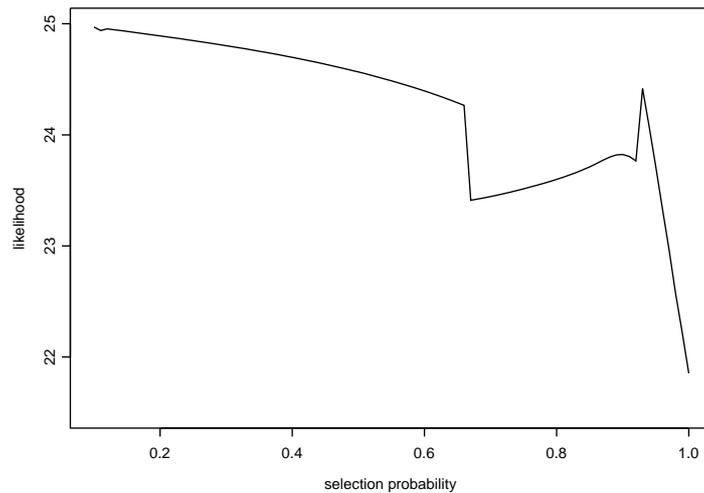


Figure 5.4: Passive smoking dataset: profile log-likelihood for  $p$  when assuming selection function  $a_7$ .

statistic to a  $\chi_1^2$  distribution. The P-value=0.0125. Therefore we have strong evidence to suggest that overall selection probability is not equal to 1, and that there are missing studies.

Figure 5.5 on page 99 shows the profile log-likelihood for given  $p$  and  $\theta$ . The contours show a long band of area that is quite flat. The maximum value of the profile log-likelihood for different values of  $p$  slowly decreases as  $p$  decreases from 1. Using this contour plot, we can plot  $\hat{\theta}$  and its 95% confidence interval for different values of  $p$ . This is shown in Figure 5.6 on page 99. As we allow for more missing studies, we see that  $\hat{\theta}$  does not change that much, and remains somewhere about 0.1 to 0.15. The confidence intervals are of particular interest as, for this example, we are interested to see when the lower limit crosses zero. It is only when  $p < 0.23$  that the 95% confidence interval includes the value 0. That would imply that if there were 124 unpublished studies, then the significance of the result would be overturned. It seems unlikely such a large number of unpublished studies would exist.

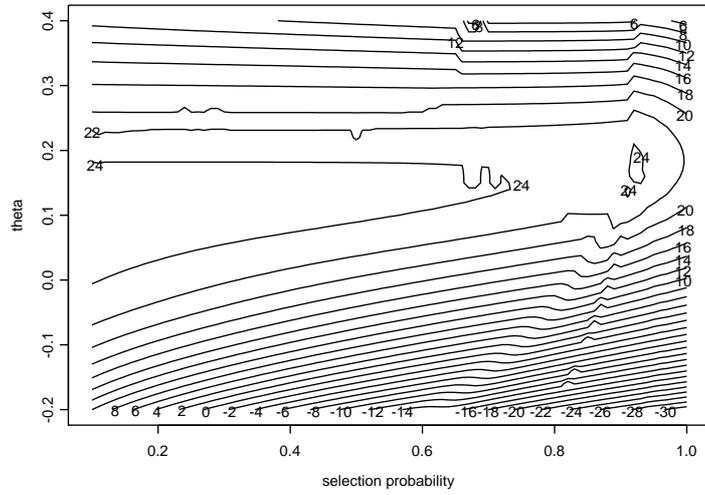


Figure 5.5: Passive smoking dataset: contour plot of the profile log-likelihood for  $p$  and  $\theta$  when assuming selection function  $a_7$ .

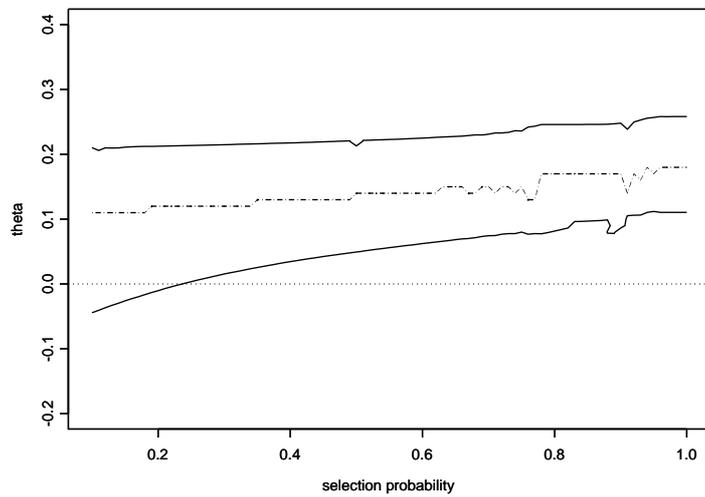


Figure 5.6: Passive smoking dataset:  $\hat{\theta}$  and 95% confidence intervals against  $p$  when assuming selection function  $a_7$ .

#### 5.4.4 Alternative model

We explore the model's adequacy by introducing an alternative model. In this section, we test the fit of the model to the funnel plot. Conventionally, the assertion is that publication bias exists if there exists a trend in the funnel plot. For example, if small studies (large  $s_i$ ) are reporting more frequently large positive values of  $y_i$  than negative values, this is one possible indication of publication bias. This suggests a possible linear relationship of  $y_i$  against  $s_i$  may appear in the funnel plot. We propose the following model.

$$y_i = \theta + \alpha s_i + (\sigma_i^2 + \tau^2)^{1/2} \epsilon_i^* \quad (60)$$

$$z_i = a + b/s_i + \delta_i \quad (61)$$

Note that the choice of the alternative model is entirely arbitrary, for example, the additional term could have been  $\alpha \frac{1}{s_i}$ . We continue in the usual way by considering the extended likelihood, as given in equation (57) in Section 5.4.1, but with the term  $\alpha s_i$  added to  $\theta$ .

$$\begin{aligned} L(\theta, \rho, \tau, a, b, \alpha) &= \sum_{i=1}^n \log p(y_i | z_i > 0, s_i) \\ &= \sum_{i=1}^n \left[ -\frac{1}{2} \log(\tau^2 + \sigma_i^2) - \frac{(y_i - \theta - \alpha s_i)^2}{2(\tau^2 + \sigma_i^2)} \right. \\ &\quad \left. - \log \Phi(u_i) + \log \Phi(v_i^*) \right], \end{aligned} \quad (62)$$

where  $u_i = a + b/s_i$  and

$$v_i^* = \frac{u_i + \tilde{\rho}_i \frac{y_i - \theta - \alpha s_i}{(\tau^2 + \sigma_i^2)^{1/2}}}{(1 - \tilde{\rho}_i^2)^{1/2}}.$$

We compare models to see if the additional  $\alpha s_i$  term helps in explaining the data. If  $\alpha$  is not significantly different from zero, then the earlier model (without the  $\alpha$  term) is giving an adequate explanation of any trend in the funnel plot. For consistency,  $\boldsymbol{\beta}$  will represent the vector of parameters  $(a, b, \rho, \tau)$  and keep  $\alpha$  as a separate parameter (even though it should be considered as a component of  $\boldsymbol{\beta}$ ).

There are various plots and tests that can be carried out with regards to investigating the fit to the funnel plot. The example of the passive smoking dataset (with  $\tau^2$  assumed zero) will be used to illustrate these methods.

- Define  $F(\alpha, p)$  as the following:

$$F(\alpha, p) = \max_{\theta, \beta \in \mathcal{B}_{\theta, p}} l(\theta, \beta, \alpha).$$

A hypothesis test can be performed concerning  $\alpha$ . For a fixed value of  $p$ , we have  $H_0 : \alpha = 0$  versus  $H_1 : \alpha \neq 0$ . This tests the adequacy of the original model. The following likelihood ratio test statistic is used.

$$X^2 = 2 \left( \max_{\alpha} F(\alpha, p) - F(0, p) \right)$$

Compare  $X^2$  to a  $\chi_1^2$  distribution. If the null hypothesis is rejected, for a specific value of  $p$ , then the null model is not adequate in explaining the data.

- The P-values corresponding to the test  $H_0 : \alpha = 0$  can be plotted against  $p$ . This essentially will show what affect the choice of  $p$  has upon the fit of the model to the data.
- Another graph will plot  $\hat{\theta}$  against the P-values corresponding to the test  $H_0 : \alpha = 0$ . This will show how  $\hat{\theta}$  will change as the quality of the fit improves. Of particular interest would be when the P-value is greater than 0.05, ie. for an acceptable fit to the funnel plot.

#### 5.4.5 Example continued: passive smoking dataset

We continue on from Section 5.4.3 with the passive smoking dataset, assuming a fixed effects model. Figure 5.7 on page 102 shows a plot of the P-values relating to the test  $H_0 : \alpha = 0$  against a range of values of  $p$ , the overall selection probability. We reject  $H_0$  when  $p > 0.67$ . This means the original model does not explain the relationship between lung cancer and passive smoking well enough if we were to assume at least two thirds of all studies have been published. Figure 5.8 on page 102 plots  $\hat{\theta}$  against the P-value for  $H_0 : \alpha = 0$ . Using the conventional 5% threshold, and with an acceptable fit to the funnel plot (P-value > 0.05), the estimate of  $\theta$  is at most 0.15. A small technical comment about Figure 5.8 is that the graph looks like a series of dots simply because we calculated  $\hat{\theta}$  for a finite grid of values of  $p$ .

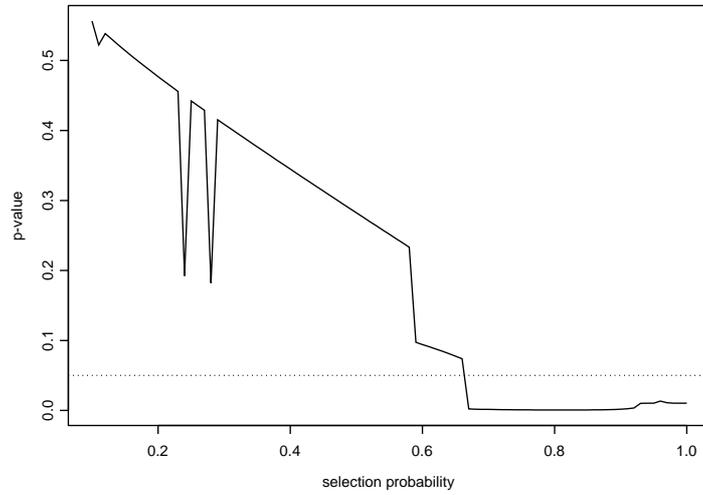


Figure 5.7: Passive smoking dataset: contour plot of the profile log-likelihood for  $p$  and  $\theta$  when assuming the alternative version of selection function  $a_7$ .

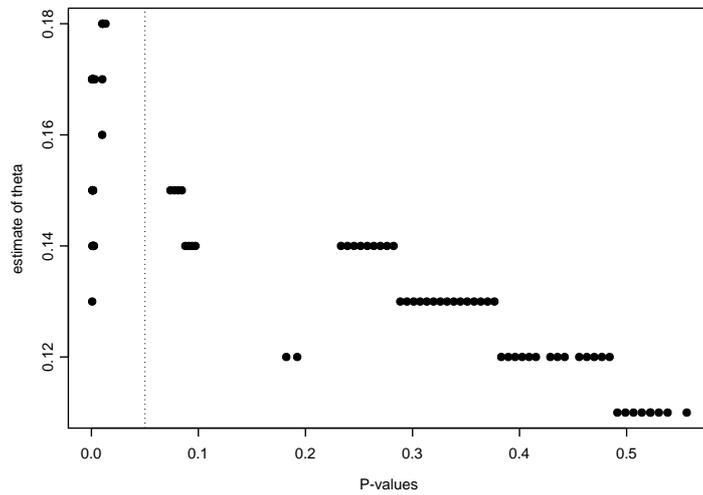


Figure 5.8: Passive smoking dataset:  $\hat{\theta}$  against the  $P$ -value when assuming the alternative version of selection function  $a_7$ .

### 5.4.6 Further examples

In this section further examples, of a suggested approach in analysing a dataset using the Copas and Shi selection function, are presented. The approach combines the hypothesis tests described in Sections 5.4.2 and 5.4.4.

The first example is the passive smoking dataset, this time with a random effects model.  $\hat{\tau}^2$  was estimated as 0.017, and then assumed fixed and known. Figure 5.9(i) on page 104 plots the profile likelihood for given  $\theta$  and  $p$ . As  $p$  decreases from 1, the likelihood rises sharply, then a slight dip, and for values of  $p < 0.6$ , the profile likelihood jumps to a high value and remains constant. The maximum value occurs when  $p = 0.16$ . The test  $H_0 : p = 1$  reports a P-value of 0.0203, hence we reject  $H_0$  at the 5% level and conclude selection of studies is present here. Figure 5.9(ii) plots  $\hat{\theta}$  and 95% confidence intervals for different  $p$ . Compared to the fixed effects model in Section 5.4.3, the values of  $\hat{\theta}$  approach zero at a faster rate as  $p$  decreases from 1. Evidence for the test  $H_0 : \theta = 0$  suggests  $\theta \neq 0$  when  $p > 0.49$  (compared to  $p > 0.23$  for the fixed effects model). Figure 5.9(iii) plots the P-values corresponding to the test  $H_0 : \alpha = 0$  for different  $p$ . There is a surprising “spike” in P-values when  $p \approx 0.9$ , otherwise the P-values are less than  $< 0.05$  for  $p > 0.67$ . Figure 5.9(iv) suggests, for an adequate fit to the funnel plot,  $\hat{\theta}$  could reasonably be as high as 0.22. Compare this to the fixed effects model that suggests a value only as high as 0.15.

The final example uses the corticosteroids dataset, assuming a fixed effects model. Figure 5.10(i) on page 105 plots the profile likelihood. The likelihood increases sharply as  $p$  decreases from 1 to 0.8. The maximum occurs when  $p = 0.66$ . For  $p < 0.66$ , the likelihood drops and flattens off. The P-value for the test  $H_0 : p = 1$  is 0.028, hence we have evidence to suggest selection is present here. Figure 5.10(ii) plots  $\hat{\theta}$  and 95% confidence intervals for  $p$ . There is a clear spike in values of  $\hat{\theta}$  when  $p = 0.55$ . The confidence interval crosses  $\theta = 0$  when  $p = 0.43$ . Figure 5.10(iii) shows that for all values of  $p$ , there is no evidence to reject  $H_0 : \alpha = 0$ , ie. the data fits well to the original model. Figure 5.10(iv) suggests  $\hat{\theta}$  could be as large as  $\theta = -0.48$  when the fit to the funnel plot is reasonable.

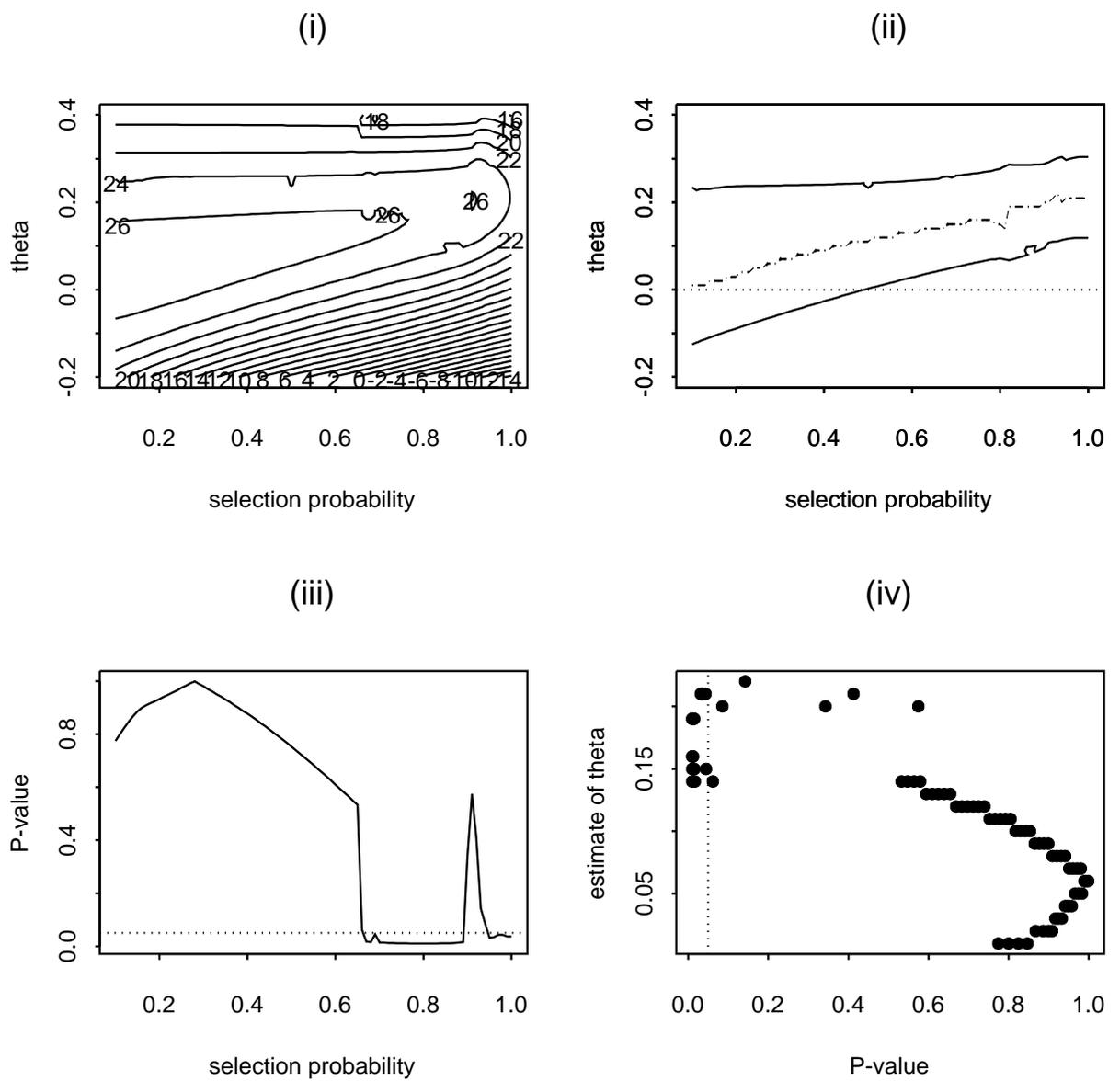


Figure 5.9: Passive smoking dataset: grid of different plots when a random effects model is used.

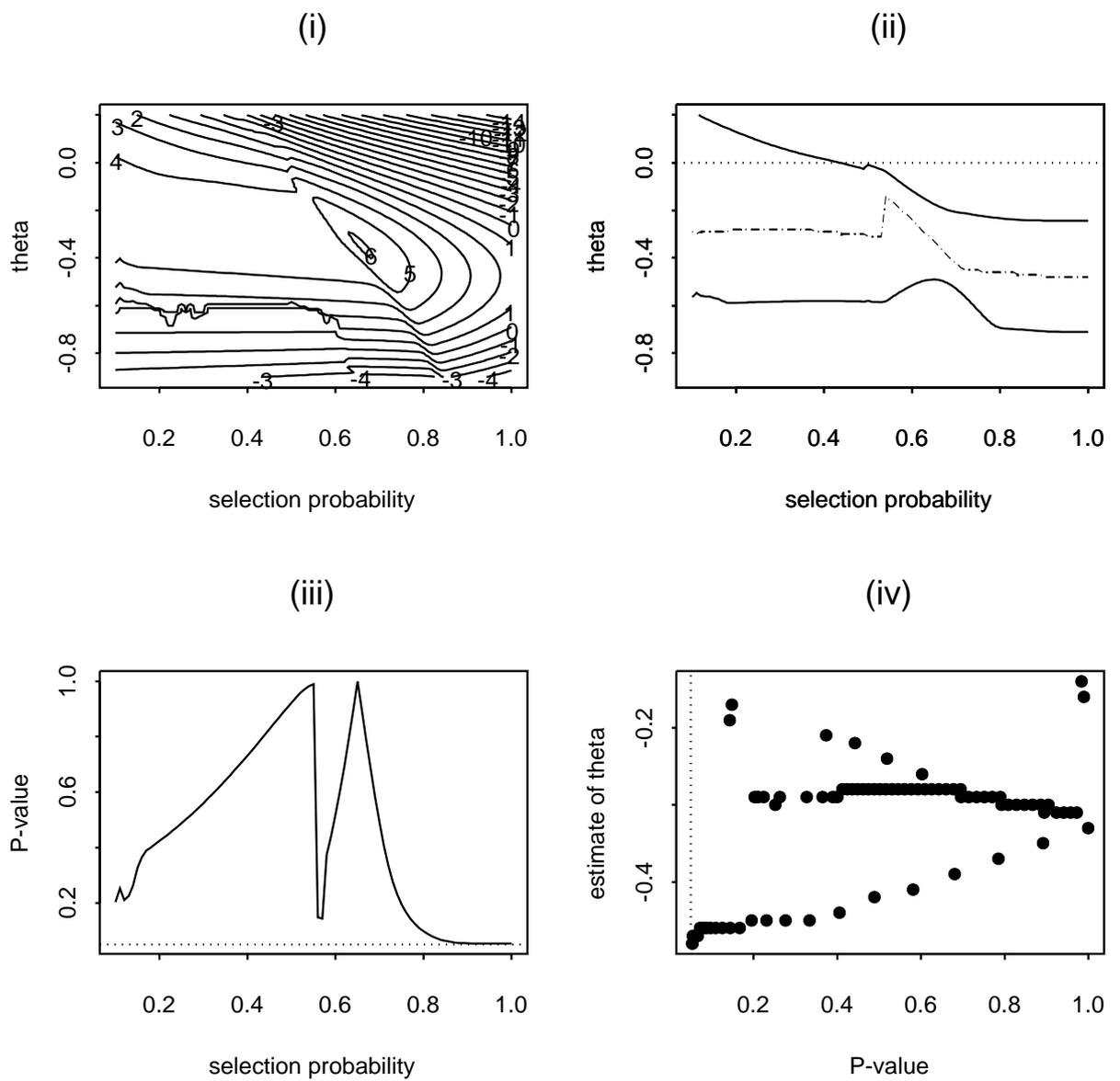


Figure 5.10: Corticosteroids dataset: grid of different plots when a fixed effects model is used.

## 5.5 Assessing the Effectiveness of the Bounds for Confidence Intervals

The problem of publication bias in meta-analysis has been attempted by many researchers. The main argument against a parametric approach to modelling publication bias is that it is impossible to verify if the selection process can be modelled in such a way. Recall from Section 2.5 “the worst case scenario” approaches of Copas and Jackson [16] and Henmi *et al.* [47]. They aimed to make as few assumptions about selection as possible by giving a bound to the bias.

The question of interest in this section is how good is this bound. The maximum likelihood estimates for  $\theta$ , calculated using the parametric approach, will be compared to the bounds suggested by Henmi *et al.* The two approaches are very different, which will be discussed later, nonetheless we aim to compare the bounds with the use of practical examples. Note that the S-Plus code that was used to calculate the bounds corresponding to the Bounds method can be found in Appendix A2 (page 140).

### 5.5.1 Examples

First, we consider the passive smoking dataset. In section 5.3, we considered selection functions  $a_1, \dots, a_6$  that had a scalar  $\beta$  parameter. Figure 5.2 on page 89 plotted  $\hat{\theta}$  and 95% confidence intervals against  $p$ . (Remember for these particular examples, we assumed a random effects model was appropriate, and we estimated  $\hat{\tau}^2$  as 0.017.) We add onto these plots the bounds for the 95% confidence intervals shown in Figure 5.11 on page 108.

The first thing to note is that in all but two graphs, the 95% confidence intervals go outside of the bounds when  $p$  is small. This is true for all selection functions with a one-tailed P-value. Figure 5.11(iii) shows the most extreme case in difference between the 95% confidence interval and bound. The lower bound cross the line  $\theta = 0$  when  $p = 0.66$ . This corresponds to a minimum of approximately 19 unpublished studies

to reverse the significance of the results. Recall from Table 5.2 on page 90 that the approximate number of missing studies for the confidence intervals to include zero ranged between 15 (for selection function  $a_3$ ) to 46 (for selection function  $a_2$ ). With the exception of  $a_3$ , the lower ends of the confidence intervals are relatively close to the lower bound. Furthermore, for the more realistic range of values for  $p$ , say greater than 0.5, the lower limits are contained within the bound.

In Section 5.4.3, Figure 5.6 plotted  $\hat{\theta}$  and 95% confidence intervals against  $p$  using the Copas and Shi selection function  $a_7$ . Figure 5.12 (page 109) shows the same plot with the added bounds. Recall that, for this example, the lower limit of the 95% confidence interval crosses zero when  $p < 0.49$ , or in other words, if there were approximately 39 unpublished studies, then the significance of the result would be overturned. Comparing this with the Bounds method, here the bound crosses  $\theta = 0$  when  $p = 0.66$  (19 missing studies) - an obvious difference. Figure 5.12 shows that for any value of  $p$ , the 95% confidence interval sits within the bounds. For values of  $p > 0.8$ , we see that the two lines are quite close together, but the distance between the lower limit and lower bound of the Bounds method widens as  $p$  decreases from 0.8.

Now consider the corticosteroids dataset. Recall we assumed the fixed effects model was appropriate. Figure 5.3 on page 92 plotted  $\hat{\theta}$  and 95% confidence intervals against  $p$  using selection functions  $a_1, \dots, a_6$ . We add onto these plots the bounds for the 95% confidence intervals. This is shown in Figure 5.13 on page 110.

The first thing to note is that in Figure 5.13(v), the 95% confidence interval goes outside the bounds when  $p$  is small. In all other graphs, the confidence intervals sit within the bounds. The upper bound crosses the line  $\theta = 0$  when  $p = 0.53$ . This corresponds to a minimum of approximately 12 unpublished studies to reverse the significance of the results. Table 5.3 on page 93 showed that the approximate number of missing studies for the confidence intervals to include zero ranged between 16 (for selection function  $a_5$ ) to 68 (for selection function  $a_2$ ). The “distance” of the upper limit of the confidence interval to the bound does vary somewhat dependent

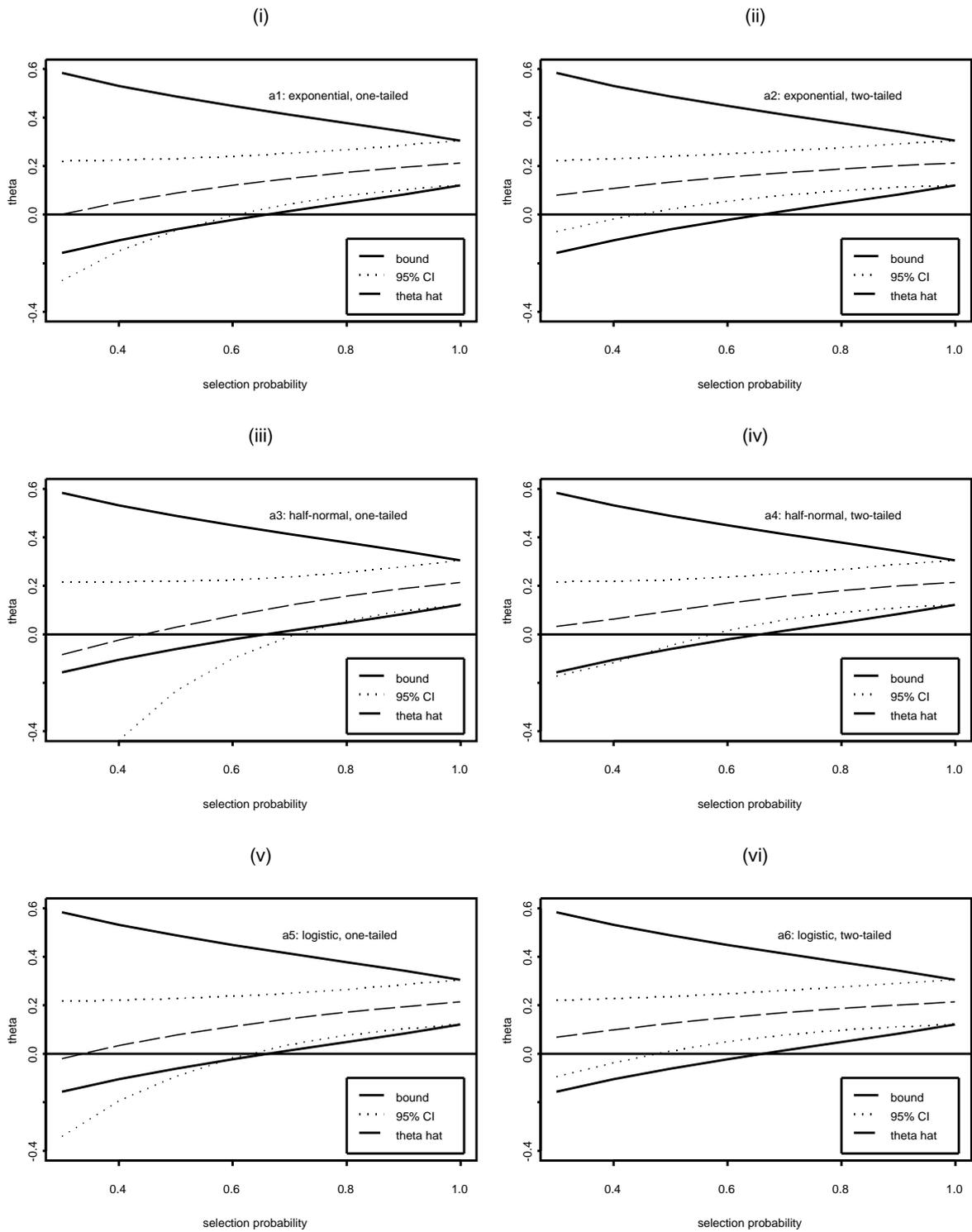


Figure 5.11: Passive smoking dataset:  $\hat{\theta}$ , 95% confidence intervals when assuming different selection functions  $a_1, \dots, a_6$  and the bounds for a range of  $p$ .

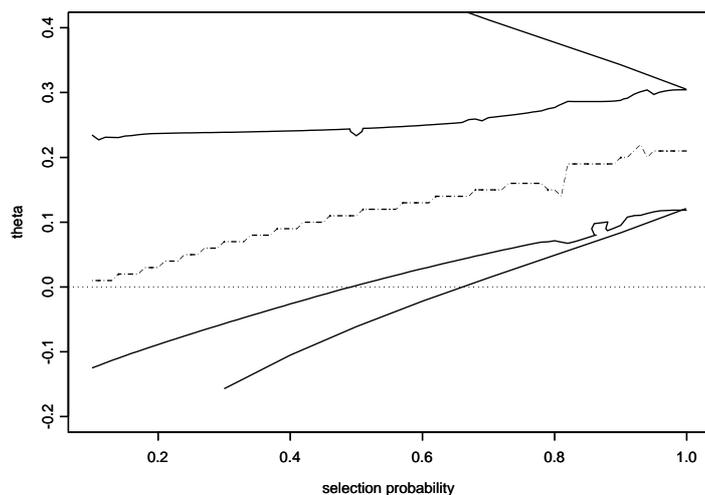


Figure 5.12: Passive smoking dataset:  $\hat{\theta}$ , 95% confidence intervals when assuming selection function  $a_7$  and the bounds for a range of  $p$ .

on the selection function used. For example, the upper limit in Figure 5.13(ii) is noticeably different from the bound, whereas in Figure 5.13(iii), the two give similar values. This implies that if you were to use the bound to examine the worst case scenario, and if  $a_2$  was considered to adequately model the selection process, then the bound here would be too overly cautious. However, if  $a_3$  was an adequate model, then the bound would be considered a useful bound on the bias in the confidence intervals.

Now consider the use of  $a_7$  with the corticosteroids dataset, as discussed in Section 5.4.6. Figure 5.10(ii) plotted  $\hat{\theta}$  and 95% confidence intervals against  $p$ . Figure 5.14 shows the same plot with the added bounds. Recall that, for this example, the lower limit of the 95% confidence interval crosses zero when  $p < 0.44$ , or in other words, if there were approximately 18 unpublished studies, then the significance of the result would be overturned. Comparing this with the lower bound, as discussed earlier, here the bound crosses  $\theta = 0$  when  $p = 0.53$  (12 missing studies). Figure 5.14 shows that for any value of  $p$ , the 95% confidence interval sits within the bounds. For values of  $p \approx 0.5$ , we see that the upper limit comes quite close to the upper bound, otherwise there is a noticeable distance between the two lines.

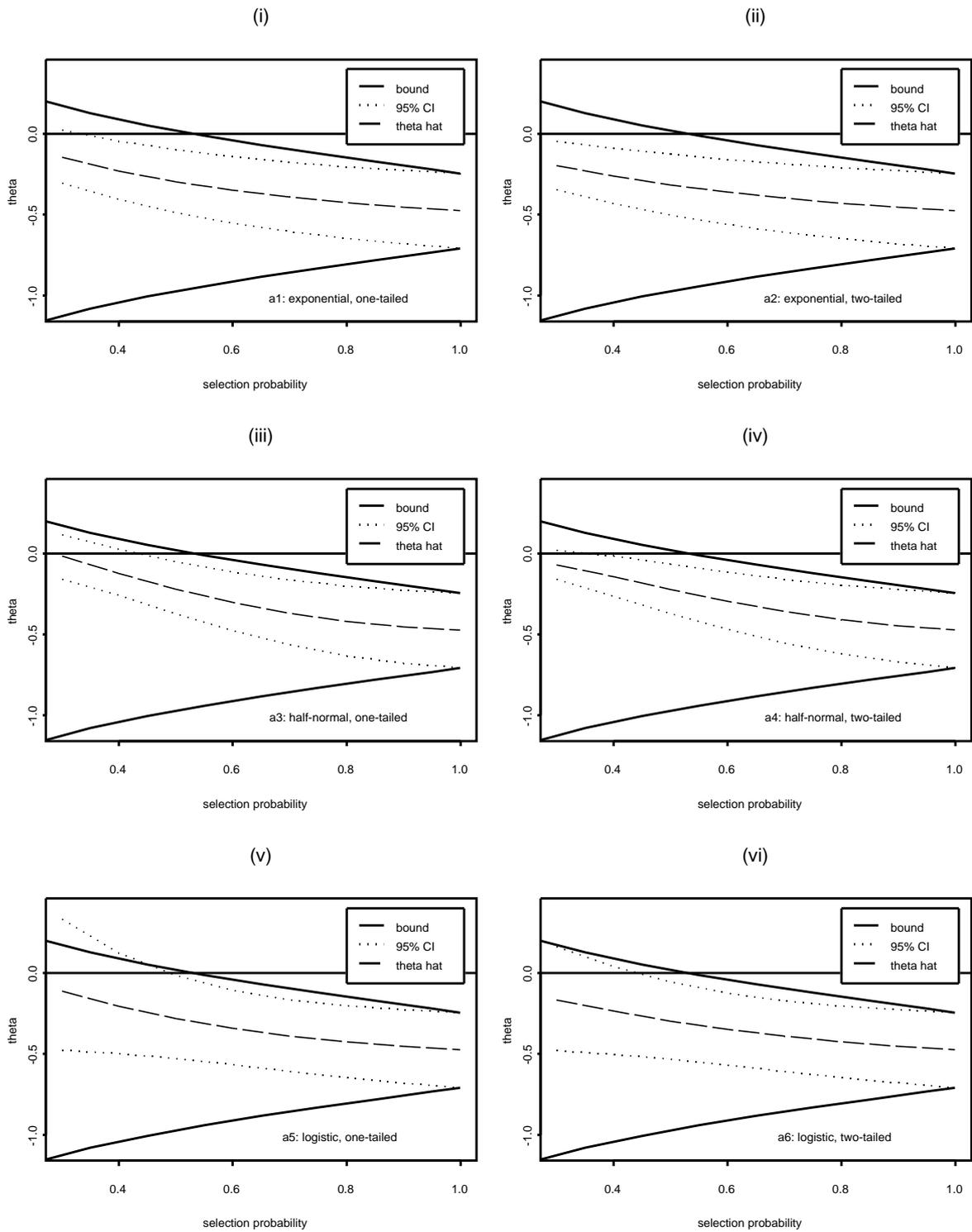


Figure 5.13: Corticosteroids dataset:  $\hat{\theta}$ , 95% confidence intervals when assuming different selection functions  $a_1, \dots, a_6$  and the bounds for a range of  $p$ .

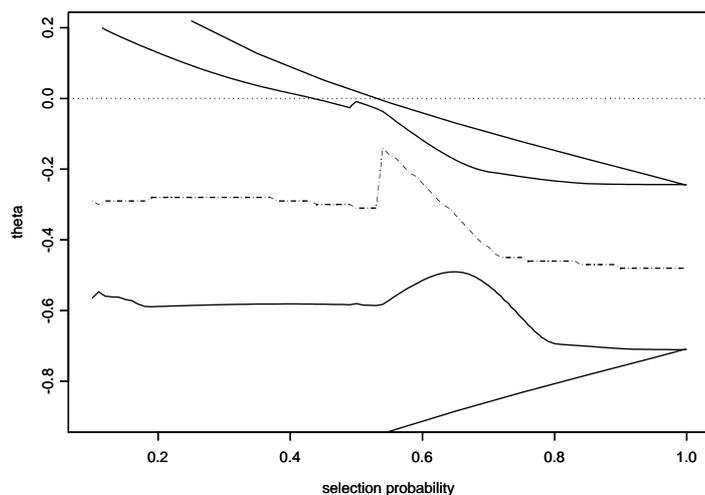


Figure 5.14: Corticosteroids dataset:  $\hat{\theta}$ , 95% confidence intervals when assuming selection function  $a_7$  and the bounds for a range of  $p$ .

An important observation from all examples given in Figures 5.11 to 5.14 is how the bounds are much wider for smaller values of  $p$ , and how there always appear to be a large “gap” between one end of the confidence intervals and the bound. The reason for this is that the Bounds method is based on a much weaker assumption than the likelihood method namely, that  $a(\sigma)$  is a decreasing function of  $\sigma$ . The Bounds method does not use the full information like the likelihood method does.

### 5.5.2 Discussion

The work in this section attempted to compare the analyses of the likelihood approach with the confidence intervals bound method. It should first be noted that the two methods use completely different approaches. The Bounds method is based on the asymptotic distribution of  $\hat{\theta}$ , whereas the main ideas in this chapter are based upon a maximum likelihood approach. It could be argued that the direct comparison of the two approaches should be treated with some caution. Not to take any comparisons too literally is the philosophy we adopt here.

A common observation to both the passive smoking and corticosteroids examples is that in nearly all cases when a one-tailed P-value was used, the relevant end (upper/lower) of the 95% confidence intervals stretched outside of the bounds when the overall selection probability,  $p$ , was small. One could argue therefore that the bounds perform badly in these examples. The reasoning behind why this occurs is to do with the selection function that actually attains the bounds. The bound is attained when the selection function  $a(y, \sigma)$  is of the form

$$a(y, \sigma) = \begin{cases} 1 & \text{if } y \leq \theta + \sigma\{\lambda\sigma - e(\sigma)\} \\ 1 & \text{if } y \geq \theta + \sigma\{\lambda\sigma + e(\sigma)\}, \\ 0 & \text{otherwise} \end{cases}$$

where  $\lambda$  and  $e(\sigma)$  were previously defined in Section 2.5.3. This is a two-sided step function, so that for 5% level of significance, we accept 2.5% of the studies in the large negative direction and 2.5% in the large positive direction. Compare this with a one-tailed selection function that accepts 5% of the studies in say, the large positive direction. Intuitively, there would be less bias with a two-tailed selection function if you are selecting some studies in the opposite direction of interest compared with a one-tailed selection function. This could explain why the width of the confidence intervals are much larger than the bounds when a one-tailed selection function is used.

The bounds approach is based upon asymptotic sampling theory, where we consider the asymptotic distribution of  $\hat{\theta}$  and how this depends on the selection function,  $a$ . The Bounds method gives you the bounds for all different  $a$ . That method is philosophically completely different to the likelihood method that we have used in this Chapter. The likelihood approach uses the funnel plot and the data, and once we fix the selection function,  $a$ , we can obtain confidence intervals. Since these two approaches are fundamentally different, it is entirely possible to witness an example of a confidence interval that falls outside of the bounds. This report only considered two meta-analyses. Using randomly generated data  $(y, \sigma)$ , we could simulate the outcome of a large number of meta-analyses with the intention of further investigating the effectiveness of the bounds.

We have entertained the unverifiable assumptions about modelling the selection process with parametric functions. For each selection function, we have based our inferences on the observed likelihood and have directly used the information available to us in the funnel plot. As the analysis of this section has shown, the bounds work quite well since the absolute distance between the bounds and the limit of the confidence interval of interest is in most cases relatively small. The comparison of the likelihood and asymptotic approaches serves as a way of justifying both methods. For example, the bounds are essentially a “worst case” scenario, asking how bad can the bias be. Since the proximity to the confidence intervals is quite small in most cases, we argue that the worst case is not as “bad” as we think. This means the bounds could be considered as a useful tool in attempting to model publication bias in meta-analysis since its values are relatively close to values that would be seen in practice with real datasets.

On the other hand, with special reference to the Copas and Shi selection function  $a_7$ , the bias given by this model is close to the bias in the bounds. This means that by imposing additional assumptions about selection, the bias is not increasing much more than the case when very few assumptions were made, ie. the bounds approach. In practice therefore, there will not be that much more harm in assuming a particular selection function, say  $a_7$ . The advantage of this is that the likelihood approach allows us to make more detailed inferences with the aid of likelihood contour plots, likelihood ratio tests, etc just as we have done in Section 5.4.

## 5.6 Concluding Comments

This chapter presented a general method for using parametric selection functions,  $a(y, \sigma, \beta)$  to model publication bias in a meta-analysis. The only restraint is that the overall selection probability,  $p$ , is such that  $p = \mathbb{E}[a(y, \sigma, \beta)]$ . The theory generalises to the case when  $\beta$  is a vector of parameters, as illustrated in Section 5.4 with  $a_7$ , the Copas and Shi selection function. Many different examples were investigated as a means of demonstrating how a sensitivity analysis can be carried out.

Section 5.3 in particular has shown that the inference about  $\theta$  can vary greatly according to the choice of selection function. This variation in inference, as well as the fact that it essentially is impossible to verify the assumptions about the selection function, makes the bounds approach appear like an attractive alternative with its few and relatively weak assumptions. Section 5.5 showed that by imposing additional assumptions about selection, the alteration in inference about  $\theta$  was not too great when comparing the bounds method to the parametric approach.

As discussed in Section 5.5.2, further work could involve simulating a large number of meta-analyses as a means of assessing the usefulness of the bound. We should expect approximately 95% of the simulated confidence intervals to lie within the bounds. Conducting simulations is one suggestion for further study. The approach taken would be as follows. Using the passive smoking dataset as a basis, the Trim and Fill method would be applied (as it was done in Section 3.4.2). Recall from Section 3.4.2 that the Trim and Fill method estimated that there were 6 missing studies. We would suppose that the true total number of published and unpublished studies was therefore 43. We would then use these 43 studies to generate different meta-analyses, with potentially different subsets each time, to try and produce a realistic scenario where publication bias may be present.

Certain quantities would remain fixed, such as  $\theta$  and the values of  $\sigma_i$ . For each simulated meta-analysis, we would randomly generate the values of  $y_i$ , say,  $y_i^* = \theta + \sigma_i \epsilon_i$  where  $\epsilon_i \sim N(0, 1)$ . We could calculate the maximum likelihood estimate of  $\theta$  each time, say,  $\hat{\theta}_m$ , and could think of  $\hat{\theta}_m$  as the estimate of  $\theta$  when no publication bias occurs, so that all 43 studies are included within the meta-analysis. The next step would be to model the publication bias in the selection process using selection functions  $a(y^*, \sigma; \beta)$  and then whether or not the study is included within the meta-analysis would be decided by randomly simulating from a uniform distribution  $\epsilon_i^* \sim N(0, 1)$ . If  $\epsilon_i^* \leq a(y_i^*, \sigma_i)$ , then we would include study  $i$  to the meta-analysis.

Once we have our subset of  $s \leq 43$  studies, we calculate the conventional ‘crude’

estimate of  $\theta$ , say  $\hat{\theta}_c$ , which would typically be different from  $\hat{\theta}_m$ . It would then be possible to use the theory as given in Section 5.2 to calculate quantities such as 95% confidence intervals for different selection functions,  $a(y, \sigma)$ . The bounds as given by Henmi *et al.* would also be calculated each time, and we could then compare the confidence bounds using the Bounds method with the confidence intervals for  $\theta$  using the likelihood based approach.

Throughout the simulations certain quantities would remain fixed, such as the 5% level of significance, the true value of  $\theta$  and the 43 values of  $\sigma$ . Certain parameters of interest would have different settings. One proposal could include three different values for the overall selection probability, each representing a different realistic level of selection. For example  $p = 0.9$  would represent a high level of selection,  $p = 0.7$  would represent a moderate level of selection, and  $p = 0.5$  would represent low overall selection. Having any lower values of  $p$  would seem unreasonable, especially with this particular example of passive smoking. Another component that would change during the simulations would be the choice of parametric selection function. The three selection functions would be:

$$\begin{aligned} a(y, \sigma; \beta) &= e^{-\beta v(y, \sigma)}, \\ a(y, \sigma; \beta) &= e^{-\beta v^2(y, \sigma)}, \\ a(y, \sigma; \beta) &= \frac{2e^{-\beta v(y, \sigma)}}{1 + e^{-\beta v(y, \sigma)}}, \end{aligned}$$

where  $v$  would be both one-sided and two-sided P-values. This would result in a total of 18 different settings within our simulation study. Once all settings would be in place, a sufficient number of iterations would be necessary to ensure the accuracy of the findings, say, 20,000 simulations for each setting. Research by Moreno *et al.* [57] and Peters *et al.* [61] provide good examples of simulation studies, and their methodology would be adopted here.

Another suggestion is further investigation of the Copas and Shi model in Section 5.4. For this model, the  $\rho$  parameter was restricted such that  $\rho \in [-0.999, 0.999]$ . The

interpretation of  $\rho$  is as follows: if  $\rho = 0$ , then there is no publication bias, ie.  $y_i$  and  $z_i$  are independent. If  $\rho > 0$ , then selected studies have  $z_i > 0$ , and smaller studies will tend to have larger values of  $y_i$ . It would be interesting to see what effect could there be in restricting  $\rho \in [0, 0.999]$ . At the points where there were unusual jumps in the likelihood plots,  $\rho$  usually had the value of -0.999, ie. exactly on the boundary of  $\rho$ 's parameter space. This in practice is not a sensible value of  $\rho$ . Further study is needed here.

There is potentially an underlying issue about the appropriateness of the theory as given in Section 5.4.2 (page 96) when applying the Copas selection function  $a_7$ . Maximum likelihood estimators possess good asymptotic properties, such as consistency and asymptotic normality. Any good text on asymptotic theory of maximum likelihood estimated is recommended, see for example [54]. There are several regularity conditions for asymptotic maximum likelihood, but in particular a relevant regularity condition here involves the boundary of the parameter space. In general, sometimes the maximum likelihood estimate may lie on the boundary of the parameter space. Standard theory requires the true parameter value to lie away from the boundary. As previously mentioned, there are instances where the  $\rho$  parameter lies exactly on its boundary. Therefore the application of the theory, as given within Section 5.4.2, may be called into question. This is an issue that Copas and Shi recognised in their original paper [22].

Carpenter *et al.* conducted extensive research into the use of the Copas selection function for two reasons: first to develop reliable software so that researchers could use the selection function, and second to provide an empirical evaluation of the method, see for example [10] and [73]. They too experienced estimation problems in 20% of meta-analyses that they investigated “despite considerable programming work”. Throughout Chapter 5 there are seemingly erratic behaviours of the likelihoods and estimates of  $\theta$ , for example where there are sudden spikes within the likelihood plots. Carpenter *et al.* also experienced similar irregular plots, which they too believed were caused by at least one of the underlying maximum likelihood estimates being close to

a boundary. Interestingly, Carpenter *et al.* wrote an R package called ‘copas’ to fit the Copas selection model to adjust for bias in meta-analysis [11]. Their paper further acknowledges irregularities may be observed in contour plots, likelihood plots and so on.

The main criticism about the Copas and Shi approach, [21] and [22], is that is a complicated model. For people with little statistical knowledge and those analysing data for systematic reviews, they will most likely struggle with applying this model. A main achievement of this report is to reconsider the approach given by Copas and Shi, and discuss the sensitivity analysis in terms of the easily understood quantity,  $p$ , rather than the confusing pair of parameters  $(a, b)$ . For example, health practitioners would much more likely understand  $p = 0.5$  compared to  $(a = -0.5, b = 1)$ . Despite this much more accessible and sensible approach to marking inferences in terms of  $p$ , the model still remains complex. A further area of study could possibly involve creating a software program, say, compatible in SAS or STATA, packages familiar to many medical statisticians. These programs would hopefully provide a useful and user-friendly interface such that the application of this model within a sensitivity analysis would be relatively pain-free. It is hoped that the various tests, plots and summaries could then be routinely used by those carrying out systematic reviews to take into consideration the dangers of publication bias in meta-analysis.

## 6 A New Likelihood Method for Monotonic Selection Functions

As reviewed in Chapter 2, the use of selection functions is a long running approach to modelling publication bias in meta-analysis. Also known as weight functions in the literature, selection functions model the selection process by assigning a probability to a study being published based on its effect size estimate. Iyengar and Greenhouse [50] was one of the first to suggest the use of weighted distributions to model selection in meta-analysis, and many other researchers have adopted this approach. A good review of selection method approaches is by Hedges and Vevea [45]. Recall from the Literature Review in Chapter 2 that there are two main classes of selection functions. The first class involves selection functions that depend on the effect size estimate through the study's  $P$ -value, or equivalently, the ratio  $y/\sigma$ . The second class of selection functions depend on the effect size estimate  $y$  and standard error  $\sigma$  separately. In this chapter, we will consider the first class of selection functions.

There have been many attempts to model selection via weight functions that depend only on study  $P$ -values. The motivation for these approaches is that it is a widely accepted assumption that research is more likely to be accepted for publication if it reports statistically significant results. Lane and Dunlap [53] and Hedges [43] modelled selection by giving weight 1 to those studies with statistically significant results (say,  $P$ -value  $< 0.05$  one-tailed) and weight 0 otherwise. This is an interesting model, but perhaps too extreme to satisfactorily model the selection process in real life. This extreme case rejects all research with non-significant results, which just simply does not happen. Nonetheless, the model provides a starting point for more

realistic models.

Consider the work of Hedges [44], Vevea *et al.* [89] and Vevea and Hedges [90]. Briefly, these attempts use a step function as a selection function, where the weights are calculated using the data, for example a maximum likelihood approach. The main assumption is concerned with the location of the steps, which are assumed known and fixed. Of course the location of these steps (say at the P-values of 0.05, then 0.1, and so on) are entirely arbitrary. The intuition is there though. It seems plausible that the journal editor will almost certainly publish if the P-value lies between 0 and 0.05, the editor will be likely to publish if the P-value lies between 0.05 and 0.1, less likely to publish with a P-value of between 0.1 and 0.5, and so on. The familiar problem faced with these attempts are the heavy assumptions about selection that we have to make. A more advanced method by Dear and Begg [24] estimated the selection function as a step function, and assumed that the relative weights and location of the steps were unknown. This semi-parametric method involved making strong assumptions about knowing the distributional form for the summary estimates for individual studies within the meta-analysis. Dear and Begg themselves promote their method only as an exploratory technique prior to conducting a meta-analysis.

The many approaches that have used selection functions to model publication bias in a meta-analysis all experience the same limitation, in that it is virtually impossible to estimate the selection mechanism from the observed studies alone. A common approach is to assume that only a proportion of the total number of studies have been included in a meta-analysis. Call this quantity,  $p$ , the overall selection probability. So for example, if no publication bias exists, we expect all studies to be included in the meta-analysis, and hence  $p = 1$ .  $p$  is an easily interpretable quantity, which is an incredible advantage. This chapter aims to find a selection function  $a^*$  which makes as few assumptions as possible, and allows us to use a likelihood based approach to find the maximum likelihood estimate of  $\theta$  conditional on an interpretable quantity  $p$ .

## 6.1 Setting Up the Step Function

We start with our usual framework of imagining that studies being included within a meta-analysis are selected via a selection function

$$a(y, \sigma) = P(\text{selection}|y, \sigma),$$

where each study reports a study outcome  $y$ , with  $y \sim N(\theta, \sigma^2)$ , estimating overall treatment effect  $\theta$ . We consider a transformation  $t = \frac{y}{\sigma}$  so that selection is now modelled via the selection function

$$a(t) = P(\text{selection}|t).$$

This follows with the familiar approach in meta-analysis by essentially modelling selection via its P-value. The main assumption we make is that  $a(t)$  is an increasing function of  $t$ , so that as  $t$  increases, the probability of selection increases. This intuitively makes sense since, if  $y$  remains constant and  $\sigma$  gets smaller,  $t$  increases, and we would expect to see studies with small  $\sigma$  more frequently than those with large  $\sigma$ . In the examples to come we will imagine positive  $y$  are more likely to be included in the meta-analysis than those with near zero  $y$  values or negative values of  $y$ , and hence our assumption that  $a(t)$  is an increasing function of  $t$ .

For the  $n$  studies in the meta-analysis, we have independent  $t_i \sim N(\frac{\theta}{\sigma_i}, 1)$ . The distribution of the *observed*  $t$  is

$$\begin{aligned} P(t|\text{selection}) &= \frac{P(\text{selection}|t)P(t)}{P(\text{selection})} \\ &= \frac{a(t)\phi(t - \frac{\theta}{\sigma})}{g(\sigma, a)}, \end{aligned}$$

where  $\phi$  is the density function of the standard normal distribution, and  $g(\sigma, a)$  is defined as  $g(\sigma, a) = \mathbb{E}[a(t)] = \int a(t)\phi(t - \frac{\theta}{\sigma})dt$ . The likelihood function is thus

$$L = \prod_{i=1}^n \frac{a(t_i)\phi(t_i - \frac{\theta}{\sigma_i})}{g(\sigma_i, a)}.$$

It can be easily shown that  $\sum_{i=1}^n \log \phi(t_i - \frac{\theta}{\sigma_i}) = A(\theta - \hat{\theta})^2 + B$ , where  $A, B$  are constants and given as

$$A = -\frac{1}{2} \sum_{i=1}^n \frac{1}{\sigma_i^2}$$

and

$$\hat{\theta} = \frac{\sum_{i=1}^n \frac{t_i}{\sigma_i}}{\sum_{i=1}^n \frac{1}{\sigma_i^2}}.$$

To find the maximum likelihood estimate of  $\theta$ , we maximise the log-likelihood function  $l(\theta, a) = \log L$ , or by using the result above, it is sufficient to maximise the following:

$$l(\theta, a) = \sum_{i=1}^n \log a(t_i) - \frac{1}{2} \left( \sum_{i=1}^n \frac{1}{\sigma_i^2} \right) (\theta - \hat{\theta})^2 - \sum_{i=1}^n \log g(\sigma_i, a). \quad (63)$$

An illustration of an arbitrary selection function  $a(t)$  is presented in Figure 6.1, with points  $t_1, t_2, \dots$  highlighted.

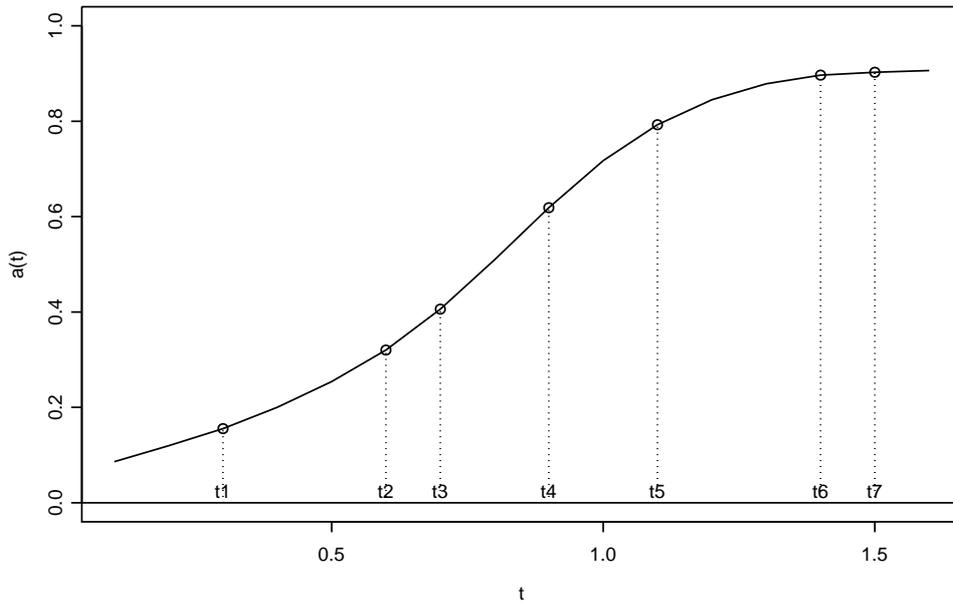


Figure 6.1: Selection function  $a(t)$ .

The question of interest is, given a selection function  $a(t)$ , can we find another selection function  $a^*(t)$  that will provide a larger likelihood function? Define  $a^*(t)$  as the following step function:

$$a^*(t) = \begin{cases} a(t_1) & t \in (-\infty, t_1) \\ a(t_1) & t \in [t_1, t_2) \\ a(t_2) & t \in [t_2, t_3) \\ \vdots & \vdots \\ a(t_{n-1}) & t \in [t_{n-1}, t_n) \\ a(t_n) & t \in [t_n, \infty) \end{cases} \quad (64)$$

The following aims to prove the step function given in (64) is indeed a selection function such that  $l(\theta, a^*) \geq l(\theta, a)$ . Recall that

$$l(\theta, a) = \sum_{i=1}^n \log a(t_i) - \frac{1}{2} \left( \sum_{i=1}^n \frac{1}{\sigma_i^2} \right) (\theta - \hat{\theta})^2 - \sum_{i=1}^n \log g(\sigma_i, a), \quad (65)$$

and  $g(\sigma_i, a) = \int a(t) \phi(t - \frac{\theta}{\sigma_i}) dt$ . The first term in (65) is identical for both  $a(t)$  and  $a^*(t)$  due to the definition of  $a^*(t)$ . The second term in (65) is clearly identical for both  $a(t)$  and  $a^*(t)$  as it depends only on the data. This means for us to show that  $l(\theta, a^*) \geq l(\theta, a)$ , it is sufficient to show that

$$\sum_{i=1}^n \log g(\sigma_i, a) - \sum_{i=1}^n \log g(\sigma_i, a^*) \geq 0.$$

Let  $t_0 = -\infty$  and  $t_{n+1} = \infty$ . We assume that  $a(t_1) > 0$ . Recall that our main assumption about  $a(t)$  is that  $a(t)$  is a non-decreasing function of  $t$ .

$$\begin{aligned} \sum_{i=1}^n \left( \log g(\sigma_i, a) - \log g(\sigma_i, a^*) \right) &= \sum_{i=1}^n \left( \log \int a(t) \phi\left(t - \frac{\theta}{\sigma_i}\right) dt - \log \int a^*(t) \phi\left(t - \frac{\theta}{\sigma_i}\right) dt \right) \\ &= \sum_{i=1}^n \left\{ \sum_{j=0}^n \left[ \log \int_{t_j}^{t_{j+1}} a(t) \phi\left(t - \frac{\theta}{\sigma_i}\right) dt \right. \right. \\ &\quad \left. \left. - \log \int_{t_j}^{t_{j+1}} a(t_j) \phi\left(t - \frac{\theta}{\sigma_i}\right) dt \right] \right\}. \end{aligned}$$

Since we assume that  $a(t) \geq a(t_j) \quad \forall t \in [t_j, t_{j+1})$ , and also note that  $\phi(x) > 0 \quad \forall x$ , then it is true that

$$\begin{aligned} \int_{t_j}^{t_{j+1}} a(t) \phi\left(t - \frac{\theta}{\sigma_i}\right) dt &\geq \int_{t_j}^{t_{j+1}} a(t_j) \phi\left(t - \frac{\theta}{\sigma_i}\right) dt \quad \forall i, j. \\ \Rightarrow \log \int_{t_j}^{t_{j+1}} a(t) \phi\left(t - \frac{\theta}{\sigma_i}\right) dt &- \log \int_{t_j}^{t_{j+1}} a(t_j) \phi\left(t - \frac{\theta}{\sigma_i}\right) dt \geq 0 \quad \forall i, j, \end{aligned}$$

and the result follows.

$a^*(t)$  is a step function, which has minimised the  $\mathbb{E}[a(t)]$  quantity in the likelihood according to the non-decreasing assumption. Figure 6.2 shows an illustration of a selection function  $a(t)$  and the step function  $a^*(t)$ . We re-write  $a^*(t)$  as follows:

$$a^*(t) = \begin{cases} d_1 & t \in (-\infty, t_1) \\ \sum_{i=1}^j d_i & t \in [t_j, t_{j+1}) \text{ for } j = 1, \dots, n. \end{cases} \quad (66)$$

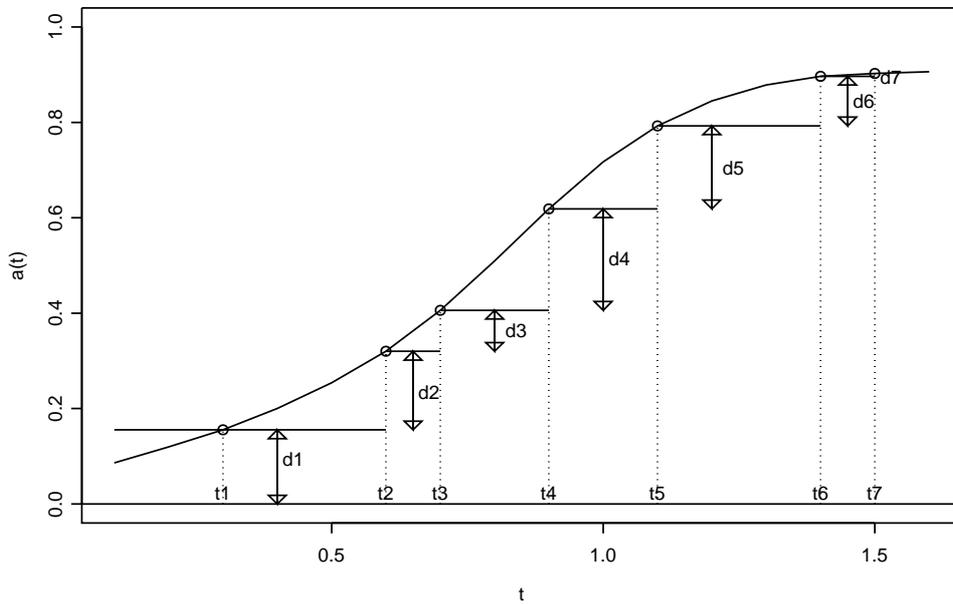


Figure 6.2: Selection function  $a(t)$  with step function  $a^*(t)$ .

Note that we have had to make an assumption about the shape of  $a^*(t)$  for  $t < t_1$ . The chosen option was to suggest that the probability of selecting a study was  $a^*(t) = a(t_1) = d_1$  for  $t < t_1$ . The selection probability of a study is assumed equal to the probability corresponding to the least positive observed  $t_i$  since we have no information about the frequency of observing any studies with  $t < t_1$ . It would be equally arbitrary to assume that  $a^*(t) = \frac{d_1}{2}$  for  $t < t_1$  but the chosen option seems sensible.

Instead of being interested in the likelihood  $l(\theta, a)$ , we need only consider  $l(\theta, a^*)$ . This provides us with an incredible advantage since  $l(\theta, a^*)$  can be expressed as a function of  $n + 1$  parameters, namely,  $l(\theta, d_1, \dots, d_n)$ . For the parameters in  $a^*(t)$ , we impose the constraints that  $d_i \geq 0$  so that  $a^*$  is a non-decreasing function, and  $\sum_{i=1}^n d_i \leq 1$ .

We now have a framework to work with a very rich family of selection functions. For a given data set  $\{t_i : i = 1, \dots, n\}$  and a given value of  $\theta$ , we can calculate  $l(\theta, a^*)$  and  $p_{a^*} = p(d_1, \dots, d_n, \theta)$ , where

$$p_{a^*} = \left\{ \frac{1}{n} \sum_{i=1}^n \frac{1}{\mathbb{E}[g(a^*, \sigma_i)]} \right\}^{-1}. \quad (67)$$

The origins of (67) comes from an earlier result that proves the overall selection probability,  $p$  is given as

$$p = \left\{ \mathbb{E}_O \left[ \frac{1}{a(\sigma)} \right] \right\}^{-1}, \quad (68)$$

where expectation is calculated over the distribution of observed studies (see Chapter 5).

In principle, if we specify a value for  $p$ , we can find a selection function  $a^*(t)$  such that  $p(d_1, \dots, d_n, \theta) = p$ . Conditioning on this  $p$ , we can calculate the profile likelihood  $l(\theta, a^*)$ . Clearly there could be many different values of the vector  $\mathbf{d} = (d_1, \dots, d_n)$  that satisfy  $p_{a^*} = p$  which in turn could cause a variation in the values for  $l(\theta, a^*)$ . We therefore calculate the profile likelihood by finding the vector  $\mathbf{d}^*$  that maximises  $l(\theta, a^*)$  over all  $\mathbf{d}$  such that  $p = p(\mathbf{d})$ . Once the profile likelihood is obtained, it is possible to proceed with calculating statistical quantities such as  $\hat{\theta}_p$  and corresponding 95% confidence intervals.

## 6.2 A Description of the Algorithm

The basic process of the algorithm that produces the profile likelihood is as follows.  $(d_1, \dots, d_n)$  is randomly generated according to the previously mentioned constraints.

The distribution from which the  $d_i$ s are generated is essentially arbitrary. The S-Plus code in Appendix A3 shows that the exponential distribution was used. This method of generating the  $d_i$ s seemed to work well, producing sensible selection functions, however any other sensible method of generating values is acceptable. Along with a given  $\theta$ , the two quantities  $p_{a^*}$  and  $l(\theta, a^*)$  are calculated. This process is repeated thousands of times. If two different randomly generated vectors  $\mathbf{d}$  calculate the same value of  $p_{a^*}$ , the vector  $\mathbf{d}$  resulting in the larger value of the likelihood is retained and used for the profile likelihood.

The values of  $p$  are categorised into bins of interval length 0.05 such that a calculated value of  $p = 0.96$  would fall into the bin  $p \in (0.95, 1)$ ; a value of  $p = 0.92$  would fall into the bin  $p \in (0.9, 0.95)$ , and so on. The reason for partitioning the overall selection probability  $p$  into bins is because searching for  $\mathbf{d}^*$  for a specific value of  $p$  for all  $p \in (0, 1)$  would be computationally expensive making it virtually impractical. If the intervals of  $p$  are sufficiently small, this algorithm essentially tackles the same problem as described previously.

The algorithm involves having a matrix of stored values, where columns of the matrix correspond to the grid of different values of  $\theta$  and the rows correspond to the different intervals of  $p$ . If, for a given value of  $\theta$  and interval of  $p$ , a selection function is found with a greater likelihood, then this current value of the likelihood is stored. Otherwise, the existing value of the likelihood, along with the existing selection function, remains.

From the profile likelihood, it will be possible to calculate 95% confidence intervals for different intervals for  $p$ , for example, 95% confidence intervals for  $p \in (0.95, 1)$ ,  $p \in (0.9, 0.95)$ , and so on. The algorithm retains the different estimated values of  $p_{a^*}$  for the different values of  $\theta$  that are to be examined. This means that rather than having a confidence interval for a band of values of  $p$ , we can very loosely calculate the “average“ value of  $p$ . This allows us to produce plots that show rather loosely what effect varying  $p$  has on the confidence intervals. As we allow for more selection, we ex-

pect the confidence intervals to become nearer the value of  $\theta = 0$ . The algorithm was implemented in S-Plus, using specially written S-Plus code. This code, along with a more detailed description of the algorithm, can be found in Appendix A3 (page 144).

### 6.3 An Example: Passive Smoking

The algorithm, as described in the previous section, was implemented using the passive smoking dataset as means of illustration. Recall that a thorough discussion of the passive smoking example can be found in Chapter 3. Figure 6.3 shows the plots of the profile likelihood for the different intervals of  $p$ .

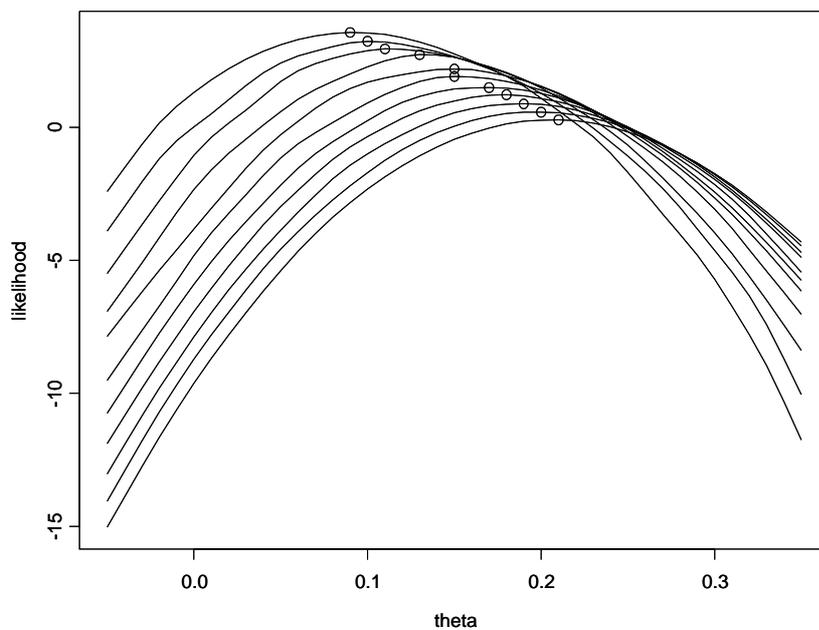


Figure 6.3: Profile likelihood for different intervals of  $p$  when assuming selection function  $a^*$ .

The lowest curve corresponds to the case when  $p \in (0.95, 1)$ . The next highest curve corresponds to the case when  $p \in (0.9, 0.95)$ . The curve third from the bottom corresponds to the case when  $p \in (0.85, 0.9)$ , and so on. The points indicate the maximum likelihood estimate of  $\theta$  for each interval of  $p$ . Note that the slightly uneven pattern in

values of  $\theta$  is a consequence of using a grid of values of  $\theta$  with a step length of 0.01. It is interesting to observe from Figure 6.3 that the likelihood curves attain higher values as the overall selection probability decreases from 1. For example, there is a clear distinction between the likelihood function for  $p \in (0.95, 1)$  and for  $p \in (0.4, 0.45)$ , and so on. This is entirely what we would expect because as  $p$  decreases, we are allowing for a better fit of the data to the model.

From the profile likelihood, it is possible to calculate the 95% confidence intervals. The curves in Figure 6.3 are normalised so that, at the maximum likelihood estimate of  $\theta$ , the likelihood has value zero. Equating  $2\left(l(\hat{\theta}, a^*) - l(\theta, a^*)\right)$ , for each interval of  $p$ , to the relevant percentile of the  $\chi^2$  distribution with one degree of freedom provides an approximate 95% confidence interval for  $\theta$ . These confidence intervals are plotted against  $p$  on the same graph, as shown in Figure 6.4.

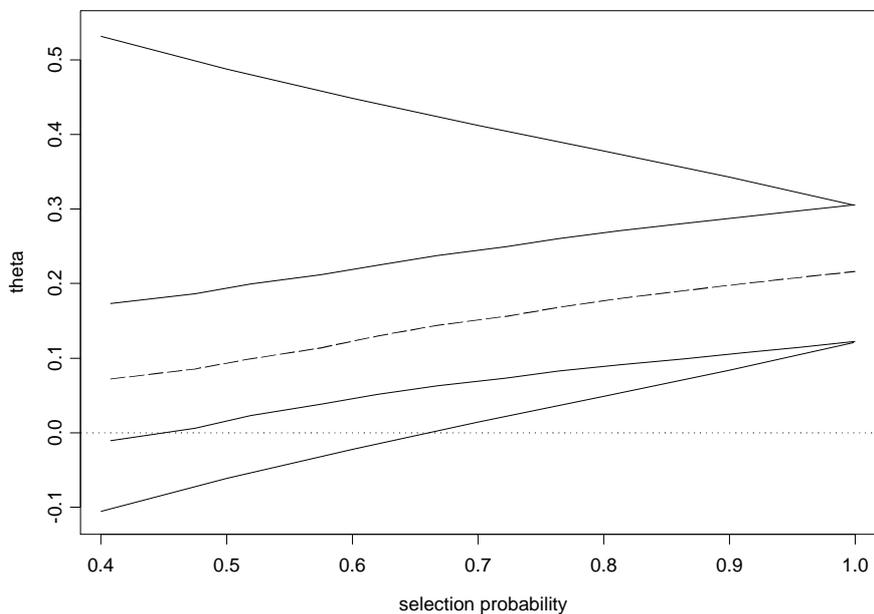


Figure 6.4: 95% confidence intervals of  $\theta$  given a range of  $p$  when assuming selection function  $a^*$ . The bounds from the Bounds method have also been added.

Intervals of $p$	Average value of $p_{a^*}$	$\hat{\theta}$	95% CI
(0.4, 0.45)	0.408	0.07	(-0.010, 0.171)
(0.45, 0.5)	0.475	0.09	(0.007, 0.188)
(0.5, 0.55)	0.519	0.10	(0.025, 0.199)
(0.55, 0.6)	0.575	0.11	(0.037, 0.212)
(0.6, 0.65)	0.620	0.13	(0.055, 0.223)
(0.65, 0.7)	0.668	0.15	(0.061, 0.239)
(0.7, 0.75)	0.722	0.15	(0.075, 0.249)
(0.75, 0.8)	0.764	0.17	(0.083, 0.261)
(0.8, 0.85)	0.812	0.18	(0.091, 0.271)
(0.85, 0.9)	0.864	0.19	(0.099, 0.281)
(0.9, 0.95)	0.911	0.20	(0.108, 0.290)
(0.95, 1)	0.957	0.21	(0.115, 0.297)

Table 6.1: Average values of  $p_{a^*}$  and corresponding 95% confidence intervals of  $\theta$ .

The dashed line in Figure 6.4 represents the values of  $\hat{\theta}$  for given  $p$ , and the dot-dash lines represent the limits of the 95% confidence intervals for given  $p$ . Note that the bounds for the confidence intervals (the solid lines) corresponding to the Bounds method by Henmi *et al.* have also been added to Figure 6.4. The graph shows that, as the overall selection probability decreases from 1, the confidence intervals shift towards the value  $\theta = 0$ . It should be noted that the graph provides a rough but insightful view as to how the confidence intervals for  $\theta$  and  $\hat{\theta}$  change as  $p$  decreases. Rather than plotting a confidence interval against a band of values of  $p$ , the average value of  $p_{a^*}$  was taken over all values of  $\theta$  for each interval of  $p$ . The tabulated values are given in Table 6.1.

The way in which we would interpret Figure 6.4 and Table 6.1 is as follows. When  $p$  is approximately 0.45, the 95% confidence interval includes the value  $\theta = 0$ . This means that if we were to entertain the possibility that only 45% of all studies relating to this research question had been included into this particular meta-analysis, then

this would cast doubt on the validity on the original statistical results. Using a simple transformation,  $p \approx \frac{37}{37+m}$ , we can say that the change to a non-statistically significant result occurs when the number of unpublished studies,  $m$ , is 46. Rather informally, this appears to be a very high number of unpublished studies which, although by no means impossible, sounds somewhat implausible. As previously mentioned, the bounds for the confidence intervals (as proposed by Henmi *et al.* [47]) have been included in Figure 6.4. According to the worst-case sensitivity analysis approach, their method suggests an overall selection probability of approximately 0.66 (or 19 unpublished studies) is sufficient to overturn the original statistical findings.

## 6.4 A Comparison with Parametric Selection Functions

$a^*$  is a non-decreasing function of  $t$ . The location of the steps are situated whenever a study reports a greater value of  $t$ . Figure 6.5 shows a variety of selection functions taken from the stored values in the passive smoking example.

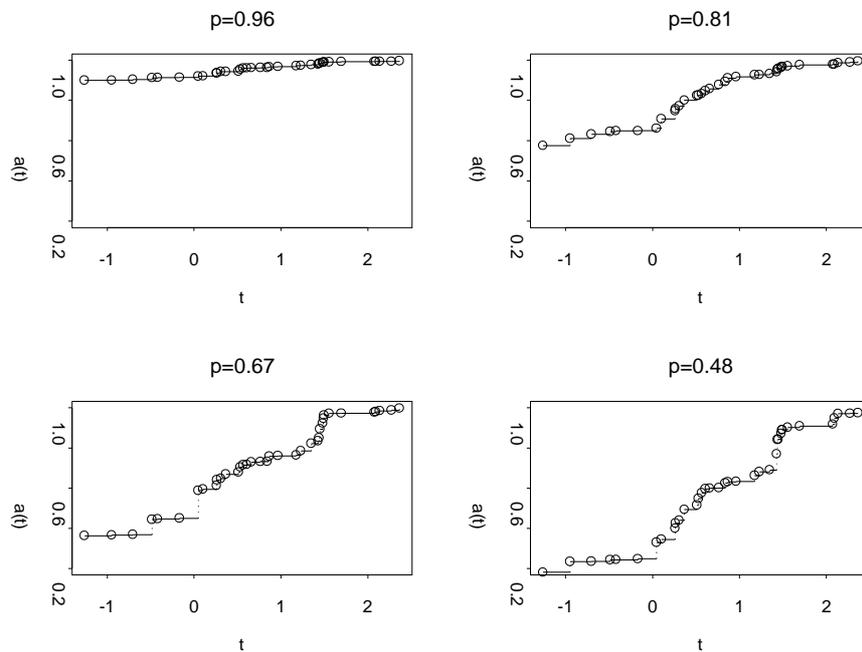


Figure 6.5: Examples of  $a^*(t)$  for a selection of different values of  $p$

Randomly generating  $\mathbf{d}$  repeatedly provides us with a very rich family of selection functions. The shapes of the function are very typical of what  $a(t)$  would look like given assumed values of the overall selection  $p$ . As one would expect, a selection function with overall high probability would have high values of  $a(t)$  for all  $t$ . As  $p$  decreases, the values of  $a(t)$  decreases over a wider range, demonstrating the increasing severity in selection.

In Chapter 5, several parametric selection functions were applied to the passive smoking data. Functions such as  $a(y, \sigma) = e^{-\beta\Phi(-y/\sigma)}$  require us to make much stronger assumptions about selection than those necessary with the  $a^*$  selection function. Table 6.2 summarises the results of the approximate value of  $p$  and equivalently the number of missing studies necessary for the significant result to be overturned for the selection functions  $a_1$  to  $a_7$ .

Selection function	Description	Approximate overall selection probability	Approximate number of missing studies
$a_1$	Exponential, one-tail	61%	24
$a_2$	Exponential, two-tail	44%	46
$a_3$	Half-normal, one-tail	71%	15
$a_4$	Half-normal, two-tail	58%	27
$a_5$	Logistic, one-tail	63%	21
$a_6$	Logistic, two-tail	48%	41
$a_7$	Copas-Shi	23%	124
<b><math>a^*</math></b>	<b>Step function</b>	<b>45%</b>	<b>46</b>
Bounds	Henmi Bounds	66%	19

Table 6.2: Summary of the results for the parametric selection functions  $a_1, \dots, a_7, a^*$  and the Bounds method for comparison.

The estimated step function  $a^*$  has been added for comparison. The values reported roughly lie in the middle of the values reported by  $a_1$  to  $a_7$ , perhaps a little cau-

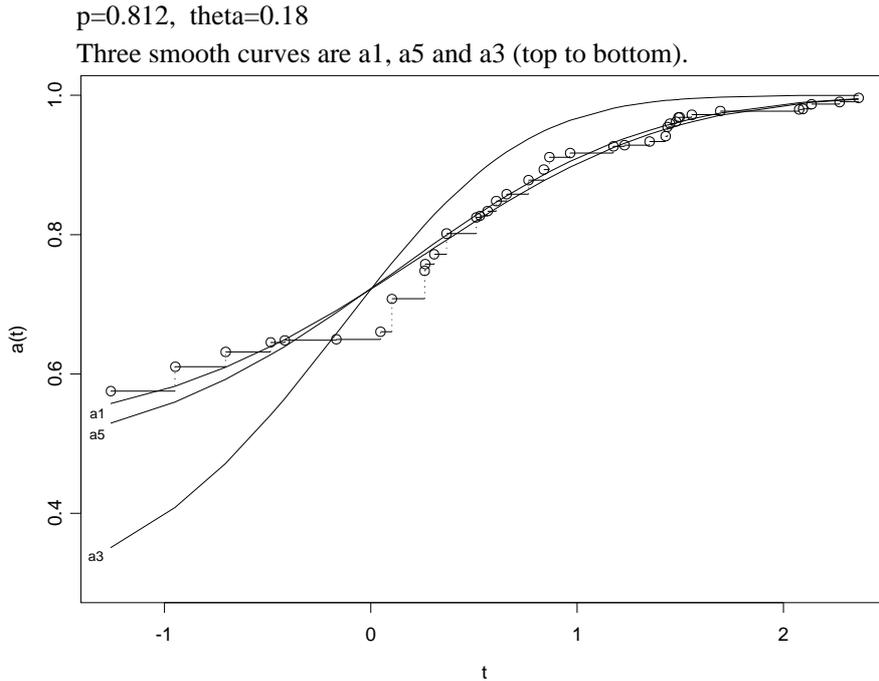


Figure 6.6: An example of  $a^*(t)$  plotted against other selection functions  $a_1$ ,  $a_3$  and  $a_5$ .

tious by comparison. A possible explanation for this is how the confidence intervals were calculated. Within each interval of  $p$  the average value of  $p_{a^*}$  was taken over a range of  $\theta$  (as listed in Table 6.1). This average was usually at the lower end of each bracket. Clearly within each interval of  $p$ , a smaller  $p$  allows for a slightly better fit for the model and so resulting in a larger likelihood which the algorithm will retain.

Figure 6.6 plots  $a^*$  and  $a_1$ ,  $a_3$  and  $a_5$  (one-tailed selection functions). As an example,  $p$  was chosen to be  $p = 0.812$  and  $\theta$  was assumed to be equal to the maximum likelihood estimate as suggested by Table 6.1. Since  $a^*$  is a one-tailed function, no two-tailed versions of the parametric selection functions were considered. Values of  $\beta$  for the parametric functions were calculated in the usual way (as described in Chapter 5).  $a_3$  (the half-normal selection function) is more severe in selection for smaller/negative values of  $t$  but more inclusive for larger positive values of  $t$ . Values of  $a(t)$  are similar for the  $a_1$  and  $a_5$  functions.  $a_1$  (the straightforward exponential

selection function) is slightly more similar in shape to  $a^*$  for this particular example. In fact, in many other numerical examples it appears that  $a_1$  is most similar in shape to that of  $a^*$ , especially when we assume a low overall selection probability  $p$ . The values of  $\hat{\theta}$  are similar for both selection functions for a range of values of  $p$ . In that respect one could argue that we could use  $a_1$  in preference to  $a^*$  to make inference about  $\theta$ . The main reason is that  $a_1$  involves far less computation, and its usage could therefore be recommended when modelling publication with the aid of selection functions.

The method involving  $a^*$  however tends to result in wider confidence intervals resulting in much weaker conclusions compared to the use of, say,  $a_1$ . This can be seen in Table 6.2 (where an overall selection probability of 61% is necessary to overturn the significance of the overall result for  $a_1$  compared to just 45% for  $a^*$ ). This difference in the width of the confidence intervals is intuitive since with  $a^*$ , we are fitting a model with a lot more parameters  $d_1, \dots, d_n$  compared to  $a_1$  which has far fewer parameters.

## 6.5 Concluding Comments

This chapter presented a likelihood based approach to modelling publication bias in a meta-analysis with the aid of a step selection function  $a^*$ . An inherent problem when using selection functions in a meta-analysis is that we are forced to make assumptions about the selection process. The aim of this chapter was to attempt to still adopt a selection function based approach, but trying to make weaker assumptions about selection than, say, methods that were explored in Chapter 5. The only assumption we made is that the selection function was an increasing function of  $t$ , or in other words  $y/\sigma$ .

The corresponding algorithm to  $a^*$  was also included, for which the specific S-Plus codes are given in Appendix A3 (page 144). It is hoped that other researchers could potentially use this selection function - the method is quite computationally intensive, but with the S-Plus code already written, it should not be too big a problem to

adapt the code accordingly to apply the method in other meta-analyses.

The maximum likelihood method that we have presented here uses a very flexible model, however there may be a danger of overfitting due to the large number of parameters in the model. An idea for future research could be to check for this issue with the following method. We could assume that the selection process is known and can be modelled by selection function  $a_1$ . We could simulate data numerous times, calculate the confidence intervals when using  $a_1$  and then when using this new method involving  $a^*$ , and then compare the confidence intervals from these two methods.

## 7 Summary and Conclusions

This thesis discussed several different approaches to modelling publication bias in a meta-analysis, with varying degrees of assumptions about the selection process. Chapter 3 introduced a case study concerning the effects of environmental tobacco smoke (also known as passive smoking) and the risk of lung cancer. The meta-analysis performed by Hackshaw *et al.* in 1997 was reviewed [40]. This particular topic was chosen for two main reasons: the first reason is that, whilst the issue of the effects of passive smoking are widely known, the implications are still very relevant. The Hackshaw analysis was used in part as evidence put forward to reform the legislation concerning smoking in public places [74]. The second reason is that the Hackshaw analysis has been used by other researchers, such as Copas, since it is a good example of a meta-analysis that shows signs of potential publication bias. The corresponding dataset was subsequently used in the later chapters of this thesis to demonstrate the various methods.

In 2007, Taylor *et al.* presented an updated meta-analysis concerning passive smoking [85]. Chapter 3 used the data as given by Taylor *et al.* to demonstrate a personal, recommended approach of performing a good meta-analysis. The literature on good analytic practice is fast growing, for which the texts by Sutton *et al.* [80], and Rothstein *et al.* [71] are recommended examples. This chapter therefore brought together some of the fundamental points in meta-analysis (for example, the choice of model and use of summary statistics, graphical displays of the data, and investigating robustness and modelling for potential publication bias), as a means of promoting merely one possible approach to meta-analysis that hopefully will stand one in good stead. The

Taylor dataset was a relevant update to the analysis carried out by Hackshaw *et al.*, providing more studies with higher numbers of cases of lung cancer. The analysis, as presented in this chapter, can be used to support the ongoing ban of smoking in public places. The comparison between the 2007 Taylor analysis and the 1997 Hackshaw analysis revealed that the two were quite consistent in their conclusions, with the overall estimates of risk and the presence of possible publication bias being quite similar.

Chapter 4 presented a robust P-value in a meta-analysis with publication bias, which is based on the idea of a permutation test. The core idea behind this non-parametric method is relatively straightforward and intentionally makes as few strong assumptions about the selection process. The only strong assumption made was that selection depends in some unspecified way on a study's P-values. Two approaches to providing a P-value were presented: the first was the permutation P-value based upon standard permutation theory. The second approach used an approximation P-value, depending upon only the number of studies in the meta-analysis and the sample correlation associated with the radial plot. Chapter 4 formed the basis for a paper that was co-authored with J.B. Copas and was successfully submitted for publication in the journal *Statistics in Medicine* in 2008 [20].

Without question, for all the approaches presented in this thesis, Chapter 4 included the one with the fewest assumptions about selection. The concepts presented in this chapter are quite elegantly simple, however there is an inevitable trade-off in the form of loss of power. The passive smoking dataset and also the cholesterol lowering dataset were used to demonstrate the methodology of the robust P-value. It is hoped that the methods presented here are simple to implement for both technical statisticians and health practitioners with less statistical knowledge.

Chapter 5 presented a general method for using parametric selection functions in meta-analysis. This chapter is in stark contrast to the previous chapter in the sense that much more stronger assumptions were made about the selection process. The

selection functions investigated included examples such as the exponential selection function,  $a(y, \sigma) = e^{-\beta V}$ , where  $\beta$  is some adjustable parameter indicating the severity of selection and  $V$  represents the study P-value. Since the choice of selection function is entirely arbitrary, a couple of other selection functions were presented, and the methodology recommended using a maximum likelihood approach to calculate the estimate of the overall quantity of interest,  $\theta$ . A sensitivity analysis was recommended, re-calibrating the selection functions  $a(y, \sigma)$  into an interpretable quantity,  $p$ , representing the overall probability of selection, and then investigating a plausible range of values for  $p$ . The recommendation from this section of work would be to investigate the potential impact of publication bias by looking at a few different selection functions and measuring how this impacts upon the estimate of  $\theta$ .

The theory presented in Chapter 5 generalised to the case when  $\beta$  was a vector of parameters, as illustrated by a selection function used by Copas and Shi [21]. A variety of examples were included as means of demonstrating how a sensitivity analysis could be carried out. The results showed that the inference about  $\theta$  varied considerably dependent upon the choice of selection function. Therefore, Chapter 5 also included an analysis investigating the effectiveness of the Bounds method as proposed by Henmi *et al.* [47]. This was achieved by comparing the confidence intervals derived from the use of the parametric selection functions with the bounds when the Bounds method was used. There is an important difference in the two approaches. The Bounds method is based on the asymptotic distribution of  $\hat{\theta}$ , whereas the parametric selection functions approach was based on maximum likelihood estimation.

The overall conclusion about the effectiveness of the Bounds method was that it is a very useful tool to use. The Bounds method does work well, since the absolute distance between the bounds and the limits of the confidence intervals (from the parametric approach) were in most cases relatively small. Since the Bounds method by Henmi *et al.* is essentially looking at the “worst case” scenario, this indicates that using this method, instead of the parametric selections which makes much stronger assumptions about selection, assumes a worst case that is not as “bad” as we think.

When we examined the Copas selection function, which imposes additional assumptions about selection compared to the Bounds method, the bias is not increasing that much more. Chapter 5 showed that the inferences about  $\theta$  were not that different when using the Copas selection function to the Bounds method, which suggests the use of the likelihood approach with the aid of the Copas selection function will be more beneficial as we are able to perform a more detailed analysis with the aid of likelihood contour plots, likelihood ratio tests, etc.

Chapter 6 presented a third approach to modelling publication bias in a meta-analysis again using a maximum likelihood approach but this time with the aid of a step selection function which we denoted as  $a^*$ . The main idea behind Chapter 6 is similar to that of the robust P-value in Chapter 4 in the sense that we are trying to make as few assumptions about the selection process as possible. Specifically, the selection function process makes one assumption about selection, namely the selection function is an increasing function of  $t = y/\sigma$ . Early works using step selection functions (Lane and Dunlap [53], and Vevea and Hedges [90] to name just two) were criticised for making very strong assumptions about the selection function, namely where to place the steps in the function which can not be easily verified.

Therefore Chapter 6 aimed to present a new method using maximum likelihood estimation, and an algorithm was discussed on how to implement this method in practice. The overall conclusions from this chapter was that we have presented a new method of modelling publication bias in a meta-analysis by assuming a step selection function without making strong assumptions about the selection process. One of the interesting observations we found concerning the inferences of  $\theta$  in this chapter was how similar the results were when assuming  $a^*$  as the selection function compared with  $a_1(y, \sigma)$  which was the exponential, one-tailed selection function discussed in Chapter 5. Since the use of  $a_1$  involved far less computation using the maximum likelihood approach than when  $a^*$  was used, one could therefore argue  $a_1$  be used when modelling publication with the aid of selection functions.

Hopefully this thesis demonstrates that there are still plenty of directions for further research to go concerning publication bias in meta-analysis. The theory and algorithm presented in Chapter 6 provides researchers with a new method to modelling publication bias using a maximum likelihood approach but also making as few assumptions about selection as possible. Therefore, as further work to succeed this thesis, we intend to use Chapter 6 as a basis for a paper which will hopefully be submitted to a journal such as *Statistics in Medicine*. The chapter would be modified to include more examples with the hope that others in the field of meta-analysis could potentially use the step selection function in their own research. Related to this further area of work would be to make available the S-Plus codes corresponding to the algorithm used in Chapter 6, or adapt the code to make the method available in R, or perhaps even in SAS to make the method accessible to medical statisticians.

In Chapter 5 we investigated the effectiveness of the Bounds method by Henmi *et al.* Two numerical examples were discussed, namely the passive smoking dataset and the corticosteroids dataset. As means of further work, additional research could involve simulating a large number of meta-analyses as a means of assessing the usefulness of the bounds. We could, say, simulate thousands of meta-analysis datasets and then apply the Bounds method and compare to the confidence intervals when assuming a parametric selection function. This would provide a more thorough analysis of the effectiveness of the Bounds method by conducting simulations.

One of the biggest aims of this thesis was to discuss methods in meta-analysis that are accessible to a wide range of researchers. There are inherent problems with the use of selection functions when modelling publication in meta-analysis, and so the new methods presented here hopefully aim to avoid these problems in such a way that others could implement these methods in practice with relative ease.

## Appendix A1 - Statistics in Medicine (2008) Paper

Refer to Copas, J.B. and Malley, P.F. (2008). A robust P-value for treatment effect in meta-analysis with publication bias. *Statistics in Medicine*, **27**, 4267 – 4278.

## Appendix A2 - S-Plus Code for the Bounds Method

Here we present the S-Plus functions used to calculate the bounds for the confidence intervals as presented by Henmi *et al.*, as first discussed in Chapter 2.5: A Review of Recent Research Investigating Publication Bias. The code presented below is split into two sections: the first contains code that only needs running once; the second contains code that will need re-running for each different example, for which an example is given. A brief commentary follows both sections of code.

### 1. Code that only needs running once.

```
bf1 <- function(x, b, p){
  pnorm(b - x) + pnorm(- b - x) - p}

bf2 <- function(b, p){
  int <- c(0, 1 - p + pnorm(b) - b)
  uniroot(bf1, int, b = b, p = p)$root}

bf3 <- function(b, p){
  sapply(b, bf2, p = p)}

bf4 <- function(lam){
  ee <- bf3(lam * sig, p)
  d1 <- lam * sig - ee
  d2 <- lam * sig + ee
  b1 <- mean((dnorm(d2) - dnorm(d1))/(p * sig))
  b2 <- mean((1 + (d2 * dnorm(d2) - d1 * dnorm(d1))/p)/sig^2)

  - b1 - ga * sqrt(b2 - b1^2)}

bf5 <- function(){
  nlminb(-2/mean(sigpass), bf4, lower = - Inf, upper = 0)$objective}
```

```

bf6 <- function(pgrid){
  igrd <- 1:length(pgrid)
  result <- pgrid
  for(i in igrd) {
    assign("p", pgrid[i], where = 1)
    result[i] <- - bf5()}
  assign("result", result, where = 1)}

```

The above functions correspond to the formulae given in Section 2.5.3 starting on page 24. The function **bf1** is essentially equation (11) for given values of  $(e_i, \lambda\sigma_i, p)$ . Function **bf2** finds the value of  $e_i$  for given values of  $(\lambda\sigma_i, p)$ . Function **bf3** calculates  $e_i$  for a vector of different values of  $\lambda\sigma_i$  for the same given value of  $p$ .

Next, the function **bf4**, for a given  $\lambda$ , calculates the  $e_i$  by solving equation (11). Also the quantities  $B_1^*$  and  $B_2^*$  are calculated, as shown in equations (16)-(17). The final part of function **bf4** calculates the quantity  $C^*$ , as given in equation (15).

The function **bf5** finds the minimum of the  $C^*$ s corresponding to equation (13). Following on from this, function **bf6** simply does all of the above for any specified values of  $p$ . The output of this function is called **result** which we use in the following section of code.

2. *Code that needs running for each new data set. As an example, the passive smoking data is used.*

```

xpass <- c(-0.291, 0.724, 0.758, -0.218, -0.245, 0.700, 0.208, 0.250,
          0.414, 0.022, 0.438, 0.023, 0.849, 0.501, 0.175, 0.422,
          0.767, 0.076, 0.936, 0.481, 0.059, -0.229, -0.307, 0.816,
          -0.029, 0.438, 0.170, 0.508, 0.236, 0.097, 0.505, 0.151,
          0.105, 0.163, 0.376, 0.707, 0.123)

```

```

sigpass <- c(0.307, 0.486, 0.321, 0.450, 0.586, 0.334, 0.248, 0.490,
            0.307, 0.476, 0.302, 0.294, 0.546, 0.221, 0.229, 0.694,
            0.369, 0.293, 0.632, 0.322, 0.223, 0.182, 0.437, 0.568,
            0.172, 0.356, 0.322, 0.432, 0.165, 0.317, 0.236, 0.229,
            0.285, 0.189, 0.222, 0.732, 0.218)

ga <- 1.96/sqrt(37)

pass.theta.hat <- sum(xpass/sigpass^2)/sum(1/sigpass^2)

bf6(c(seq(0.35, 0.95, 0.1), 0.999))

tmp1 <- pass.theta.hat + 1/mean(1/sigpass^2)*result
tmp2 <- pass.theta.hat - 1/mean(1/sigpass^2)*result
tmp3 <- c(0.35, 0.45, 0.55, 0.65, 0.75, 0.85, 0.95, 0.999)

plot(tmp3, tmp1, ylim=c(-0.15,0.6), xlab="selection probability",
     ylab="theta",pch=" ")
lines(tmp3, tmp1)
lines(tmp3, tmp2)
abline(h=0, lty=2)
title("Bound for 95% CI - passive")

```

First, the data  $(y, \sigma^2)$  must be presented in two vectors, called here as **xpass** and **sigpass**. Second, it is necessary to define **ga** which relates to  $\gamma$ , defined as  $n^{-1/2}z_\alpha$ , as seen in equation (15), where  $n$  is the number of studies and  $z_\alpha = \phi^{-1}(1 - \alpha/2)$  is the standard normal percentage point with coverage  $1 - \alpha$ .

**pass.theta.hat** is simply calculating the weighted average of  $\theta$ . Next it is necessary to specify a grid of values for  $p$  to place inside the function **bf6**. Note that there is a convergence issue when you choose  $p = 1$  and so the value 0.999 is used. This

minor detail essentially does not affect the subsequent graph. The final lines of code simply calculate the bounds at each value of  $p$  in the specified grid, and there is some straight forward code to plot these values in a simple graph.

Note that a similar looking graph is presented in Figure 3.4 on page 51, corresponding to the application of the bounds method using the Taylor data set.

## Appendix A3 - S-Plus Code for the Step Selection Function

Here we present the S-Plus functions used to calculate the profile likelihood for Section 6: The Use of Step Selection Functions. For the code to work, you need to specify certain quantities beforehand. First, the data  $(y, \sigma^2)$  must be presented in two vectors. Second, a grid of values of  $\theta$  must be specified, with an appropriate step length. Finally, the number of equal length intervals that  $p \in [0, 1]$  will be divided into must be specified. In the examples presented in this chapter, a grid of  $\theta$  values with step length 0.01, and values of  $p$  divided into intervals of 0.05 was considered appropriate. Examples of these quantities are respectively

```
thetagrid <- seq(-0.1, 0.3, 0.01)
np <- 20.
```

The function **setup** calculates various vectors and matrices according to the pre-defined quantities as described above.

```
setup <- function(y1, var1){
  assign("n", length(y1), where = 1)
  temp <- order(y1/sqrt(var1))
  tee <- (y1/sqrt(var1))[temp]
  assign("tee", tee, where = 1)
  assign("var", var1[temp], where = 1)
  assign("y", y1[temp], where = 1)
  temp <- outer(thetagrid, sqrt(var1), "/")
  temp <- outer(temp, tee, "-")
  wmat <- array(pnorm(temp), c(length(thetagrid), n, n))
  wmat[, , 1] <- 1
  assign("wmat", wmat, where = 1)
  assign("rlik", array(-10, c(np, length(thetagrid))), where = 1)
  assign("rp", rlik + 10, where = 1)
```

```

assign("rn", rlik + 10, where = 1)
assign("rd", array(0, c(np, length(thetagrid), n)), where = 1)}

```

The vector **d** is randomly generated each time, potentially by any method providing they satisfy the necessary constraints. The function **deegen** is one such example.

```

deegen <- function(){
temp <- rexp(n + 1)
temp[1] <- temp[1] * 100
(temp/sum(temp))[1:n]}

```

The function **update** is the main function which is repeated many times.

```

update <- function(j){
dee <- deegen()
rlik1 <- rlik
rp1 <- rp
rd1 <- rd
rn1 <- rn
for(i in 1:length(thetagrid)) {
ay <- wmat[i, , ] %%% dee
pee <- n/sum(1/ay)
lik <- sum(log(cumsum(dee))) - sum(log(ay))
test1 <- ceiling(pee * np)
if(rlik1[test1, i] < lik) {
rlik1[test1, i] <- lik
rd1[test1, i, ] <- dee
rp1[test1, i] <- pee
rn1[test1, i] <- rn1[test1, i] + 1}}
assign("rlik", rlik1, where = 1)
assign("rp", rp1, where = 1)
assign("rd", rd1, where = 1)

```

```
assign("rn", rn1, where = 1)}
```

Once there has been a sufficient number of iterations of **update**, the final function **lika** is used to calculate the profile likelihood. A matrix is produced with the rows corresponding to the intervals of  $p$ , and the columns corresponding to the grid of values of  $\theta$ .

```
lika <- function(){  
  wt <- sum(1/var)  
  thetahat <- sum(y/var)/wt  
  temp <- -0.5 * wt * (thetagrid - thetahat)^2  
  temp <- matrix(temp, ncol = length(thetagrid), nrow = np, byrow = T)  
  rlik + temp * (rlik != -10)}
```

## References

- [1] Akiba S., Kato H. and Blot W.J. (1986). Passive smoking and lung cancer among Japanese women. *Cancer Research*, **46**, 4804-4807.
- [2] Bailey, K.R. (1987). Inter-study differences - how should they influence the interpretation and analysis of results. *Statistics in Medicine*, **6**, 351-360.
- [3] Begg, C.B. and Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, **50**, 1088-1101.
- [4] Bennett, D.A., Latham, N.K., Stretton, C. and Anderson, C.S. (2004). Capture-recapture is a potentially useful method for assessing publication bias. *Journal of Clinical Epidemiology*, **57**, 349-357.
- [5] Bero, L.A., Glantz, S.A. and Rennie, D. (1994). Publication bias and public health policy on environmental tobacco smoke. *Journal of the American Medical Association* **272**, 133-136.
- [6] British Medical Association Tobacco Control Resource Centre (2004). The human cost of tobacco. Passive smoking: doctors speak out on behalf of patients. British Medical Association Report.
- [7] Boffetta, P. (2002). Involuntary smoking and lung cancer. *Scandinavian Journal of Work, Environment & Health*, **28**, 30-40.
- [8] Candelora, E.C., Stockwell H.G., Armstrong A.W. and Pinkham P. (1992). Dietary intake and risk of lung cancer in women who never smoked. *Nutrition and Cancer*, **17**, 263-270.
- [9] Cardenas V.M., Thun M.J., Austin H., Lally C.A., Clark W.S., Greenberg S., et al. (1997). Environmental tobacco smoke and lung cancer mortality in the American Cancer Society's cancer prevention study II. *Cancer Causes & Control*, **8**, 57-64.

- [10] Carpenter, J.R., Schwarzer, G., Rücker, G. and Kunstler R. (2009). Empirical evaluation showed that the Copas selection model provided a useful summary in 80% of meta-analyses. *Journal of Clinical Epidemiology*, **62**, 624-631.
- [11] Carpenter, J.R., Rücker, G. and Schwarzer, G. (2009). copas: An R package for Fitting the Copas Selection Model. *The R Journal*, **1**: 31-36.
- [12] Cochran, W.G. (1963). *Sampling Techniques. Second Edition*. John Wiley & Sons, New York.
- [13] Cook, D.J., Sackett, D.L. and Spitzer, W.O. (1995). Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam Consultation on Meta-Analysis. *Journal of Clinical Epidemiology*, **48**, 167-171.
- [14] Cooper, H., Hedges, L.V. and Valentine, J.C. (eds) (2009). *The Handbook of Research Synthesis and Meta-Analysis: Second Edition*. Russell Sage Foundation: New York.
- [15] Copas J.B. (1999). What works? Selectivity models and meta-analysis. *Journal of the Royal Statistical Society, Series A*, **162**, 95-109.
- [16] Copas, J. and Jackson, D. (2004). A Bound for Publication Bias Based on the Fraction of Unpublished Studies. *Biometrics*, **60**, 146-153.
- [17] Copas, J.B. and Lozada, C. (2007). Asymptotic Approximations for the Radial Plot in Meta Analysis, and a Bias Correction to the Egger Test. Working Paper, No. 07-01, Department of Statistics, University of Warwick.
- [18] Copas, J.B. and Lozada, C. (2007). The radial plot in meta analysis: approximations and applications. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **58**, 329-344.
- [19] Copas J.B. and Li H.G. (1997). Inference for non-random samples. *Journal of the Royal Statistical Society, Series B*, **59**, 55-95.

- [20] Copas, J.B. and Malley, P.F. (2008). A robust P-value for treatment effect in meta-analysis with publication bias. *Statistics in Medicine*, **27**, 4267-4278.
- [21] Copas J.B. and Shi J.Q. (2000a). Reanalysis of epidemiological evidence on lung cancer and passive smoking. *British Medical Journal*, **320**, 417-418.
- [22] Copas J.B. and Shi J.Q. (2000b). Meta-analysis, funnel plots and sensitivity analysis. *Biostatistics*, **1**, 247-262.
- [23] Copas J.B. and Shi J.B. (2001). A sensitivity analysis for publication bias in systematic reviews. *Statistical Methods in Medical Research*, **10**, 251-265.
- [24] Dear, K.B.G. and Begg, C.B. (1992). An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science*, **7**, 237-245.
- [25] Deeks, J., Glanville, J. and Sheldon, T. (1996). Undertaking systematic reviews of research on effectiveness: CRD guidelines for those carrying out of commissioning reviews. Centre for Reviews and Dissemination, York. York Publishing Services Ltd. Report #4.
- [26] DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, **7**, 177-188.
- [27] Duval, S. and Tweedie, R. (2000). Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *The Journal of the American Statistical Association*, **95**,89-98.
- [28] Duval, S. and Tweedie, R. (2000). A non-parametric trim and fill method of assessing publication bias in meta-analysis. *Biostatistics*, **56**,455-463.
- [29] Easterbrook, P.J., Berlin, J.A., Gopalan, R. and Matthews, D.R. (1991). Publication bias in clinical research. *Lancet*, **337**, 867-872.
- [30] Edgington, E.S. (1987). *Randomization Tests. Second Edition*. Marcel Dekker, New York.

- [31] Egger, M., Smith, G.D., Schneider, M. and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, **315**, 629-634.
- [32] Egger, M., Smith, G.D. and Altman, D.G. (2001). *Systematic Reviews in Health Care: Meta-analysis in Context*. BMJ Publishing Group: London.
- [33] Egger, M., Jüni, P., Bartlett, C., Holenstein, F. and Sterne, J. (2003). How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical Study. *Health Technology Assessment*, **7**, 1-76.
- [34] Galbraith, R.F. (1988). A note on graphical presentation of estimated odds ratios from several clinical trials. *Statistics in Medicine*, **7**, 889-894.
- [35] Garfinkel L. (1981). Time trends in lung cancer mortality among nonsmokers and a note on passive smoking. *Journal of the National Cancer Institute*, **66**, 1061-1066.
- [36] Glass, G.V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, **5**, 351-379.
- [37] Good, P. (2000). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer, New York.
- [38] Good, P. (2005). *Introduction to statistics through resampling methods and R/S-PLUS*. John Wiley & Sons, Canada.
- [39] Fisher, R.A. (1935). *Design of Experiments*. Oliver and Boyd, Edinburgh.
- [40] Hackshaw A.K., Law M.R. and Wald N.J. (1997). The accumulated evidence on lung cancer and environmental tobacco smoke. *British Medical Journal*, **315**, 980-988.
- [41] Hammond S.K., Coghlin J., Gann P.H., Paul M., Taghizadeh K., Skipper P.L., et al. (1993). Relationship between environmental tobacco smoke exposure and

- carcinogen-hemoglobin adduct levels in nonsmokers. *Journal of the National Cancer Institute*, **85**, 474-478.
- [42] Harbord, R.M., Egger, M. and Sterne, J.A.C. (2005). A modified test for small study effects in meta analysis of controlled trials with binary endpoints. *Statistics in Medicine*, **25**, 3443-3457.
- [43] Hedges, L.V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Studies*, **9**, 61-85.
- [44] Hedges, L.V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science*, **7**, 246-255.
- [45] Hedges, L.V. and Vevea, J. (2005). *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments. Chapter 9: Selection Method Approaches*. John Wiley & Sons, England.
- [46] Hedges, L.V. and Vevea, J.L. (1996). Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics*, **21**, 299-332.
- [47] Henmi M., Copas J.B. and Eguchi S. (2007). Confidence Intervals and P-Values for Meta-Analysis with Publication Bias. *Biometrics*, **63**, 475-482.
- [48] Hill A.B. (1965). The Environment and Disease: Association or Causation? *Proceedings of the Royal Society of Medicine* **58**, 295-300.
- [49] Humble C.G., Samet J.M. and Pathak D.R. (1987). Marriage to a smoker and lung cancer risk. *American Journal of Public Health*, **40**, 604-609.
- [50] Iyengar, S. and Greenhouse, J.B. (1988). Selection models and the file-drawer problem. *Statistical Science*, **3**, 109-135.
- [51] Jennions, M.D. and Møller, A.P. (2002). Publication bias in ecology and evolution: and empirical assessment using the ‘trim and fill’ method. *Biological Review*, **77**, 211-222.

- [52] Kendall, M. (1938). A New Measure of Rank Correlation. *Biometrika*, **30**, 81-89.
- [53] Lane, D.M. and Dunlap, W.P. (1978). Estimating effect size: Bias resulting from significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology*, **31**, 107-112.
- [54] Le Cam, L. and Yang, G.L. (2000). *Asymptotics in Statistics: Some Basic Concepts (Springer Series in Statistics)* Springer.
- [55] Lee, P.N. (1992). *Environmental tobacco smoke and mortality*. Karger: Basle.
- [56] Maclure M., Ben-Abraham R., Bryant M.S., Skipper P.L. and Tannenbaum S.R. (1989). Elevated blood levels of carcinogens in passive smokers. *American Journal of Public Health*, **79**, 1381-1384.
- [57] Moreno, S.G., Sutton, A.J., Ades, A.E., Stanley, T.D., Abrams, K.R., Peters, J.L. and Cooper, N.J. (2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology*, **9**:2.
- [58] Orwin, R.G. (1983). A fail safe N for effect size in meta-analysis. *Journal of Educational Statistics*, **8**, 157-159.
- [59] Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *British Medical Journal*, **3**, 1243-1246.
- [60] Peters, J.L., Sutton, A.J., Jones, D.R., Abrams, K.R. and Rushton, L. (2006). Comparison of two methods to detect publication bias in meta-analysis. *Journal of the American Medical Association*, **295**, 676-680.
- [61] Peters, J.L., Sutton, A.J., Jones, D.R., Abrams, K.R. and Rushton, L. (2007). Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in Medicine*, **26**, 4544-4562.
- [62] Peters, J.L., Sutton, A.J., Jones, D.R., Abrams, K.R., Rushton, L. and Moreno, S.G. (2010). Assessing publication bias in meta-analyses in the presence of

between-study heterogeneity. *Journal of the Royal Statistical Society: Series A*, awaiting publication.

- [63] Peto R., Darby S., Deo H., Silcocks P., Whitley E. and Doll R. (2000). Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies. *British Medical Journal*, **321**, 323-329.
- [64] Pham, B., Platt, R., McAuley, L., Klassen, T.P. and Moher, D. (2001). Is there a 'best' way to detect and minimize publication bias? An empirical evaluation. *Evaluation and the Health Professions*, **24**, 109-125.
- [65] Pitman, E.J.G. (1937a). Significance tests which may be applied to samples from any populations. *Journal of the Royal Statistical Society, Series B* **4**, 119-130.
- [66] Pitman, E.J.G. (1937b). Significance tests which may be applied to samples from any populations II. *Journal of the Royal Statistical Society, Series B* **4**, 225-232.
- [67] Pitman, E.J.G. (1938). Significance tests which may be applied to samples from any populations III. The analysis of variance test. *Biometrika* **29**, 322-335.
- [68] Preston, C., Ashby, D. and Smyth, R. (2003). Adjusting for publication bias: modelling the selection process. *Journal of Evaluation in Clinical Practice*, **10**, 313-322.
- [69] Rosenthal, R. and Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *Journal of Psychology*, **55**, 33-38.
- [70] Rosenthal, R. (1979). The 'file drawer problem' and tolerance of null results. *Psychological Bulletin*, **86**, 638-461.
- [71] Rothstein, H.R., Sutton, A.J. and Borenstein, M. (eds) (2005). *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. John Wiley & Sons, Ltd: Chichester.

- [72] Schwarzer, G., Antes, G. and Shumacher, M. (2002). Inflation in Type I error rate in two statistical tests for the detection of publication bias in meta analysis with binary outcomes. *Statistics in Medicine*, **21**, 2465-2477.
- [73] Schwarzer, G. Carpenter, J.R. and Rücker. (2009). Empirical evaluation suggests Copas selection model preferable to Trim-and-Fill for selection bias in meta-analysis. *Journal of Clinical Epidemiology*, **63**, 282-288.
- [74] Scientific Committee on Tobacco and Health (SCOTH), Department of Health (2004). Secondhand Smoke: Review of evidence since 1998. SCOTH Report.
- [75] Smith, M.L. (1980). Publication bias in meta-analysis. *Evaluation in Education*, **4**, 22-25.
- [76] Smith, G.D., Song, F. and Sheldon, T.A. (1993). Cholesterol lowering and mortality: The importance of considering initial level of risk. *British Medical Journal* **306**, 1367-1373.
- [77] Sterling, T.D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa. *Journal of the American Statistical Association*, **54**, 30-34.
- [78] Sterne, J.A.C., Gavaghan, D. and Egger, M. (2000). Publication and related bias in meta-analysis: Power of statistical tests and prevalence in literature. *Journal of Clinical Epidemiology*, **53**, 1119-1129.
- [79] Stockwell H.G., Goldman A.L., Lyman G.H., Noss C.I., Armstrong A.W., Pinkham P.A., et al. (1992). Environmental tobacco smoke and lung cancer in nonsmoking women. *Journal of the National Cancer Institute*, **84**, 1417-1422.
- [80] Sutton, A.J., Abrams, K.R., Jones, D.R., Sheldon, T.A. and Song, F. (2000). *Methods for Meta-Analysis in Medical Research*. Chichester: Wiley.
- [81] Sutton, A.J., Duval S.J., Tweedie R.L., Abrams K.R. and Jones D.R. (2000). Empirical assessment of effect of publication bias on meta-analyses. *British Medical Journal*, **320**, 1574-1577.

- [82] Sutton, A.J., Song, F., Gilbody, S.M. and Abrams, K.R. (2000). Modelling publication bias in meta-analysis: a review. *Statistical Methods in Medical Research*, **9**, 421-445.
- [83] Sutton, A.J. and Higgins, J.P.T. (2007). Recent developments in meta-analysis. *Statistics in Medicine*, **27**, 625-650.
- [84] Taylor, R., Cumming, R., Woodward, A. and Black, M. (2001). Passive smoking and lung cancer: a cumulative meta-analysis. *Australian and New Zealand Journal of Public Health*, **25**, 203-211.
- [85] Taylor, R., Najafi, F. and Dobson, A. (2007). Meta-analysis of studies of passive smoking and lung cancer: effects of study type and continent. *International Journal of Epidemiology*, **36**, 1048-1059.
- [86] Terrin, N., Schmid, C.H., Lau, J. and Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine*, **22**, 2113-2126.
- [87] Trikalinos, T.A., Churchill, R., Ferri, M., Leucht, S., Tuunainen, A., Wahlbeck, K. and Ioannidis, J.P.; EU-PSI project (2004). Effect sizes in cumulative meta-analyses of mental health randomized trials evolved over time. *Journal of Clinical Epidemiology*, **57**, 1124-1130.
- [88] Tweedie, R.L. and Mengersen, K.L. (1992). Lung cancer and passive smoking: reconciling the biochemical and epidemiological approaches. *British Journal of Cancer*, **66**, 700-705.
- [89] Vivea, J.L., Clements, N.C. and Hedges, L.V. (1993). Assessing the effects of selection bias on validity data for the General Aptitude Test Battery. *Journal of Applied Psychology*, **78**, 981-987.
- [90] Vevea, J.L. and Hedges, L.V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, **60**, 419-435.
- [91] de Waard, F., Kemmeren, J.M., van Ginkel, L.A. and Stolker, A.A. (1995). Urinary cotinine and lung cancer risk in a female cohort. *British Journal of Cancer*, **72**, 784-787.

- [92] Wald N.J., Nanchahal K., Thompson S.G. and Cuckle H.S. (1986). Does breathing other people's tobacco smoke cause lung cancer? *BMJ*, **293**, 1217-1222.
- [93] Wald N.J., Ritchie C. (1984). Validation of studies on lung cancer in nonsmokers married to smokers. *Lancet* **1**, 1067.
- [94] Wang, T.J. (1997). Meta-analysis of the potential relationship between exposure to environmental tobacco smoke and lung cancer in nonsmoking Chinese women. *Lung Cancer*, **16**, 145-150.
- [95] Wu-Williams, A.H., Dai, X.D., Blot, W., Xu, Z.Y., Sun, X.W., Xiao, H.P., et al (1990). Lung cancer among women in north-east China. *British Journal of Cancer*, **62**, 982-987.