The combining of information: Investigating and synthesizing what is possibly common in clinical observations or studies via likelihood.

Keith O'Rourke

Department of Statistics,

University of Oxford,

January 18, 2008

Abstract

The combining of information: Investigating and synthesizing what is possibly common in clinical observations or studies via likelihood.

A thesis submitted by Keith O'Rourke of Worcester College towards a D.Phil. degree in the Department of Statistics, University of Oxford, Trinity Term, 2003.

The thesis is to develop an analytical framework for a flexible but rigorous model based investigation and synthesis of randomized clinical trials - regardless of outcome measure, probability model assumed or published summary available. This involves the identification of relevant statistical theory, the development and adaptation of necessary techniques and the application of these to a number of examples.

A new strategy for the investigation and synthesis of RCTs regardless of outcome measure, probability model assumed or published summary available was developed to accomplish this. No such general strategy has been explicitly set out before. It provides a quite general method, and with adequate sample information, results in arguably correct and adequate techniques for the assumptions made.

A new method of numerical integration was developed to incorporate flexible random effects models; an importance sampling approach was developed to obtain the needed observed summary likelihoods and a Monte Carlo based diagnostic to assess the adequacy of sample information was produced but remains to be further researched.

Contents

1 Introduction

1.1 Description of thesis

My thesis is to develop a general analysis framework for rigorous model-based investigation and synthesis of randomized clinical trials - regardless of outcome measure, probability model assumed or published summary available - drawing on the relevant statistical theory. This involves the development of a strategy of analysis (i.e. how to discern the relevant pieces of information and then coherently contrast and possibly combine them together), the identification of relevant statistical theory (from both the current and historical literature) to motivate and support the strategy along with the development of necessary techniques to implement the strategy. This first involved the explication of a descriptively appealing and transparent model-based statistical framework[26] for meta-analyses or systematic reviews of (well designed and conducted) randomized clinical trials. The more pervasive but implicit combination of information "activity" in statistics was reviewed in order to gain arguably sufficient insight for such a framework. This review was carried out by recasting some of the theory and techniques of parametric likelihood as being the investigation and synthesis of what is possibly common in a group of individual observations. Here investigation is taken to mean the assessment of what individual observations "suggest" largely on their own and the assessment of conflicts between these individual "suggestions". Here synthesis is taken to mean the explicit rendering of a "common suggestion" usually, but not necessarily, limited to the application of numerical algorithms to individual observations. For instance, the choice of a "best" observation out of n observations is considered a synthesis if "best" is well defined in some sense. Franklin has recently argued for the value of translating algebra into verbal concepts to facilitate the application of the algebra or at least its implications into statistical practice[52] and

the explication of a descriptively appealing and transparent model-based statistical framework was undertaken for just that purpose.

Likelihood is seen to be the always appropriate summary statistic that usually should just be added on the logarithmic scale (i.e. an un-weighted average). A large part of current research in theory in statistics becomes seen as relevant. The history of statistics suggests that this once was more explicit and widely accepted than it is today and was fairly central in Fisher's work. This is apparently somewhat surprising even to some well known scholars of Fisher [private communication AWF Edwards] and this insight may aid those who try to understand Fisher's work.

A novel but perhaps at first distracting aspect of the strategy is to focus on individual observation likelihoods defined as

$$L(\theta_i; y_i) = c(y_i) \Pr(y_i; \theta_i)$$

preferably with the choice of $c(y_i)$ to make

$$L(\theta_i; y_i) = \Pr(y_i; \theta_i) / \Pr(y_i; \hat{\theta}_i).$$

The full sample likelihood is then purposely written out as a multiple of these individual observation likelihoods

$$L(\theta_1, \theta_2; y_1, y_2) = L(\theta_1; y_1) * L(\theta_2; y_2).$$

By doing this, it is made clear which parameters are common, common in distribution or arbitrarily different by observation. Common parameters are identified by the repeated appearance of the same parameter in the likelihood, such as μ in

$$L(\theta_1, \theta_2; y_1, y_2) = L((\mu, \sigma_1); y_1) * L((\mu, \sigma_2); y_2).$$

When there are common in distribution parameters, unobserved random parameters differ by study but are related by being drawn from the same "common" distribution. In such cases, one would likely wish to denote such parameters as random variables χ_i^* drawn from $\Pr(\chi_i^*; \Theta)$. Here for there to be commonness, the Θ must have components that repeat in likelihoods of different (usually groups of) observations. This will be further clarified and more fully discussed later.

No such general strategy for the investigation and synthesis of RCTs, regardless of outcome measure, probability model assumed or published summary available, has been explicitly set out

before. It is quite general, and with adequate sample information, results in arguably correct and adequate techniques for the assumptions made. It facilitates the implementation of either Bayesian or Classical inference, though this thesis largely concentrates on the more challenging to implement Classical inference[39]. It initially entertained an alternative approach for dealing with random effects - using techniques originally developed in robust inference - but found this approach defective for models with unknown scale parameters and more general (e.g. asymmetric) random effects distributions (that have different expected treatment effects than from the fixed effect model)[102][98]. Instead, a new method of numerical integration was developed that provides valid upper and lower bounds to allow the more confident use of a wide range of random effects models.

The thesis also facilitates the thoughtful consideration of possibly very complex models (to deal with publication bias, randomization assignment unblinding, differential loss to follow up, etc.) and directly motivates techniques for the sensitivity analysis of these now more explicit assumptions being made. Limitations of sample information for such complex analyses arise from the "state of the literature" - information cannot simply be created. For a discussion of non-likelihood-based methods should the likelihood approach become prohibitively time-consuming or non-robust, see chapter 8.4 in Cox[27]. On the other hand, prior information might be used to help overcome this limited information[62].

The result could be described as a "generalized meta-analysis model" - analogous to the Generalized Linear Model. But where as the Generalized Linear Model was largely achieved by limiting both the choice of probability models and their parameterizations, the generalized meta-analysis model does not restrict the probability models or their parameterizations but instead implements general estimation by numerical optimization of likelihoods and implements inference for individual parameters by profile likelihood. This extra generality requires a means to obtain the group summary observed likelihoods, better numerical integration techniques for evaluating flexible random effects models that do not have closed form representations, global optimization techniques to avoid local optimizations and ideally a diagnostic for when profile likelihood runs into "Neyman-Scott" type difficulties. Such extra generality is largely available in meta-analysis applications due to there usually being independent groups of independent observations that facilitate the applicability of central limit theorem type results to the group likelihoods though, as it will be shown, not necessarily to the combined likelihood. With the development of the necessary techniques, the

value of the strategy was then demonstrated on a number of meta-analysis examples.

The direct use of the theory of statistics (or at least parametric likelihood-based theory) to motivate and support a general analysis framework for meta-analysis has been largely ignored in the literature. The lack of success in the usual setting of a single study perhaps discouraged such attempts. On the other hand, most researchers in meta-analysis do not see the theory of statistics as a source of concepts and techniques directly applicable to meta-analysis. More specifically, they do not view statistical theory and methods for observations as essentially being meta-analysis of studies having sample size one. Instead, many view meta-analysis as a "two-stage" process involving the calculation of an appropriate summary statistic for each of a set of studies followed by the combination of these statistics into a weighted average[35]. Within this perspective, Generalized Linear Modeling for instance, is perceived only as a technique that uses the "correct" summaries and "correct" weights for the outcome measures - but not as in any way related to a model-based likelihood analysis. Alternatively, while Bayesian researchers use model-based approaches, their emphasis has primarily been on prior distributions and the large role they play[121]. In particular, the role of the likelihood in the combination of observations and studies has been obscured and needs to be re-emphasized as done in O'Rourke and Altman[85]. On the other hand, though many statisticians have worked on meta-analysis, it is usually with respect to a specific technique or topic such as p_value censorship rather than a general theory or approach.

Some statisticians may regard meta-analysis as so similar to usual statistical analysis as to not require special consideration [personal conversation Irwin Guttman]. That claim has been considered in this thesis, and, to a large extent, it is justified for those who are already aware of it - but for a general approach to meta-analysis to be fully developed and widely appreciated this justification needs to be explicitly made. A similar exercise to develop an explicit strategy for analysis of variance has recently been carried out by Andrew Gelman[56]. The strategy he proposes is to view analysis of variance as the batching of coefficients into groups - often considered exchangeable within the groups. The strategy in this thesis, on the other hand, starts with parameters in likelihoods for individual observations and then decides which parameters are common, common in distribution or arbitrary and then which to focus on and which ignore by treating them as nuisance parameters that are somehow to be evaded.

For clinical research practice, the two-stage approach is widely believed to reflect pragmatically important insights and to provide appropriate techniques for the investigation and synthesis of

RCTs in most application areas. This belief is perhaps fostered by it being more or less the case with single studies in many areas of applied statistics (at least when there are independent groups of independent observations). In such cases, complex analyses can often be split into pieces and intelligently and effectively re-assembled using a simple two-stage approach based on means and variances[2]. Such a favorable state of affairs offers important advantages to a clinical research community that is not highly trained in statistics and needs guidance and advice as to if, and when, these simple approaches are appropriate and adequate. In the current context of evidence-based medicine the treatment one will or will not receive is likely to be heavily influenced by simple methods, adequate or otherwise. Adequate ones are arguably preferable and the two-stage approximation should not be undervalued. It was argued in O'Rourke (unpublished) that such simplified approaches to statistical analysis in general were important and perhaps in some wider sense - especially when collaborating with colleagues with limited statistical training - may provide optimal "human information processing" within and between researchers. However, with respect to the combined likelihoods that will be encountered in this thesis that can be very non-quadratic and easily multi-modal - such an approach may not be applicable. A transparent graphical display may be helpful here. The raindrop plot[10] was initially considered in this regard but was found to be somewhat deficient in highlighting the non-quadratic components in some examples. A new plot was developed to make it more transparent as to how the individual log-likelihoods add up to the pooled log-likelihood, in ways that may be very non-quadratic and multi-modal.

In any event, the two-stage approach does need to first "hit on" the "correct" summaries and "correct" weights for the meta-analysis in hand, and although this may seem obvious for many applications, for some it is not. For instance, for dealing with the simple but less common application of combining common proportions from single groups, the "correct" weights are obvious using the strategy of this thesis - n_i - but the common two-stage approach (or at least a misunderstanding of it) has been used to claim the correct weights must be $n_i/(y_i/n_i * (1 - y_i/n_i))$.[43]

It will be argued that the thesis provides a general framework for the solution of statistical questions and difficulties that arise when undertaking meta-analyses in clinical research. For instance, in my MRC funded research fellowship with the Centre for Statistics in Medicine on the meta-analysis of continuous outcomes, a number of identified challenges arose using the popular two-stage approach to meta-analysis of first choosing a "good" summary estimate of something expected to be common from each study and then determining some "optimal" combination of

these summaries. (This approach has been succinctly described in Cox[24] and worked through in numerous practical aspects in the Cochrane Handbook[18].) More specifically, drawing from Cox[24], let $t_1, ..., t_m$be the good summary estimates from m trials that can be assumed to be approximately normally distributed around θ (here a scalar) with known variances $v_1, ..., v_m$, calculated from the internal variability within the separate trials. The "best" combined estimate of θ is $\widetilde{t} = (\sum t_j/v_j)/(\sum 1/v_j)$ where $\widetilde{t}$ will, under the assumptions of normality, be normally distributed around θ with variance $(\sum 1/v_j)^{-1}$. This will be called the two-stage inverse variance weighted approach for meta-analysis or more generally just the two-stage approach (when other weights might be used) and is standard in many clinical research contexts (with various extensions for random effects). It is only appropriate though, if all of the log-likelihoods are essentially quadratic in the region of combination; otherwise the combined log-likelihood will not be approximately quadratic.

The two-stage approach tends to focus on statistical summaries rather than parameters and on procedures and techniques – i.e. linear (weighted) estimating procedures - rather than on underlying probability models that are conceptualized as having generated the observations (random variables) and appropriate inference for those assumed models and observations. Again, without explicitly considering an underlying probability model and appropriate inferences for it, suggested modifications of procedures and techniques can only be on a hit-or-miss basis. There is a temptation to "guesstimate" the needed means and standard deviations and then proceed as if these are known[66]. In the given reference, the example purported to demonstrate the value of such a method that was based on using medians and ranges of within patient change scores, but these were actually unavailable to the authors. Instead the authors "imputed" values using information on group mean scores over various treatment applications on different days, even when considerable drop out occurred [private communication].

Furthermore, the non-parametric conversions they used were based on inequalities between sample medians and ranges and SAMPLE means and standard deviations which are only indirectly relevant - it is population means and standard deviations that are of inferential relevance. In general, such conversions are single imputations of missing summaries, where use of the usual techniques and narrowly defined best estimates can be very flawed. On the other hand, given an appropriate likelihood-based approach, as will be shown, the resolution of the difficulties is largely straightforward except for numerical computing challenges. Fortunately, given the insight from the likelihood-based approach, an adequate quadratic approximation of it (for practical purposes)

will often be achievable and that approximation itself can be carried out as a two-stage approach (which simply provides an arithmetical way to combine the quadratic likelihood functions). This though, does need to be rigorously assessed as being pragmatically valid for the particular case. Given that the likelihood-based approach is not too computationally demanding, it makes good sense to check each particular case, though the development of guidelines would be useful and is future research.

In statistics, parametric likelihood usually does not end with just a presentation of the (combined) likelihood function. This thesis does not accept the suggestion made by certain authors[101] that the likelihood function is all that need be or should be presented . The two most common additions to, or further renderings of, the likelihood function are the construction of confidence intervals or the calculation of posterior probabilities as a function of assumed prior probabilities to get credible intervals. This thesis accepts both as worthwhile additions to the likelihood, though the posterior probabilities less so with contrived rather than "arguable" prior information[57][27]. The likelihood function is being taken in this thesis as simply a device to obtain good confidence or credible intervals and nothing more. Emphatically, if the use of the likelihood function leads to defective confidence or credible intervals in a particular application - it would be rejected for that application. Specifically to avoid defective confidence intervals due to Neyman-Scott problems, a diagnostic was produced to help detect this and remains future research.

In practice, neither the construction of confidence intervals nor that of posterior probabilities has much of an effect on the sample based *investigation and synthesis* of what is possibly common given a model - this is usually all undertaken using likelihood. Priors (though usually just implicit) arguably should have a large role in determining what should be initially entertained as being common between experiments (based on their scientific design), but given this, the empirical investigation is largely likelihood-based except for prior-data conflict[44]. Hence both of these further renderings will not have a large emphasis in this thesis. The Classical versus Bayesian constructions/calculations usually differ in meta-analysis of randomized clinical trials, not with respect to likelihood functions used, but only with respect to how they modify the likelihood functions to get confidence intervals versus posterior probabilities. Additional differences may arise in the ad-hoc choices of how to summarize likelihoods or summarize posterior probabilities such as choosing between the posterior mode and mean. These may result in dramatic differences in the inferences even given identical likelihood functions.

This thesis emphasizes what is usually common in Bayesian and Classical approaches to meta-analysis. Of passing interest, though, empirical Bayes and Bayesian approaches to parametric statistics can be viewed as simply comprising an additional investigation and synthesis of what is common between an estimated or known distribution of a parameter and likelihoods for that parameter. The focus in the Bayesian approach then, will be on both the combination of observations by likelihood and the combination of probabilities by Bayes theorem. This "combining" view has become more prevalent recently[44]. More general combinations have been studied by Dempster[36][37].

1.2 Combination of observations: A parametric likelihood approach

For constructing a general strategy, parametric likelihood arguably provides a better route for the reasons alluded to above. A brief summary of the review of the parametric likelihood-based approach to statistics, conceptualized as the investigation and synthesis of individual observations, is as follows:

1. A descriptively appealing and transparent definition of likelihood is "the probability of re-observing exactly what was observed" under a given probability model – in notation $L(\theta; observed) = f(observed; \theta)$ considered as a function of θ an n-dimensional vector of reals for fixed *observed*, where *observed* can never really be a continuous number but instead some interval and, more often than not, in meta-analysis is a reported summary rather than actual individual observations. For notational convenience, from now on the *observed* will simply be denoted by y.

2. Likelihoods from different observations multiply (after appropriate conditioning if observations are dependent).

3. If there is more than one likelihood and something is common in these likelihoods - i.e. some $\gamma(\theta)$ is repeated in the likelihoods - the multiplication of them provides a combination for that (the $\gamma(\theta)$), and under the probability model, that multiplication provides the "best" combination. A common parameter reparameterization may help make this more apparent. For instance, for $\theta = (\theta_1, ..., \theta_k)$ a common parameter reparameterization is given by $\omega = \omega(\theta_1, ..., \theta_k) = (\gamma, \chi_i)$, where $\gamma = \gamma(\theta_1, ..., \theta_k)$ (the something that is possibly common) and $\chi_i = \chi(\theta_1, ..., \theta_k)_i$ and conversely, $\theta_i = \theta_i(\gamma, \chi_i)$ for each observation i.

4. It does not matter if likelihoods are based on $n = 1$ (a single observation) or $n = k$ (a single study) for 1, 2 & 3 and the order of multiplication also does not matter.

5. There usually is something common - $\gamma(\theta)$ and something non-common $\chi(\theta)_i$ in the probability model entertained -

$$f(y_i; \gamma(\theta), \chi(\theta)_i)$$

The something common may be a particular parameter or a distribution of a particular parameter (often referred to as a random effects model) and that commonness will be with regard to a particular transformation γ. Perhaps a good example of this would be a non-common treatment effect parameter transformable into a common direction parameter and a non-common magnitude parameter.

6. When it is the distribution of a parameter that is common, the multiplication referred to above for combining applies only to the marginal or expected likelihood with respect to the distribution of $\gamma_i^* \sim p(\gamma)$ i.e. $\prod_i E_\gamma L(\gamma_i^*; y_i)$ as $\prod_i L(\gamma_i^*; y_i)$ does not provide a combination.

7. In meta-analysis, usually only a summary of y is available, say $s(y)$, so the required likelihood is $L(\theta; s(y))$ i.e. a marginal likelihood with respect to the distribution of individual observations, which is "forced" upon the meta-analyst. This is in line with Barndorff-Nielsen's definition on page 243[8]

"If u is a statistic we speak of the likelihood function which would have been obtained if only the value of u, and not the basic data y, had been observed as the marginal likelihood function based on u."

It is defined as:

Definition 1 *The marginal likelihood is defined as* $c(s(y)_o) \int_{y^*} \Pr(y; \theta) dy$ *where* $\Pr(y; \theta)$ *is the probability,* y *the set of possible individual observations and* y^* *is a "level set" of* y *such that* $s(y)$ *equals the reported summary* $s(y)_o$*. The* $\int$ *may be a* $\sum$ *depending on the probability model* $\Pr(y; \theta)$ *for the outcomes* y *and it is assumed that the probability model is defined on a partition* $y = \{y\}$ *and that the function* $s(y)$ *is measurable. See also formula* 2 *of Copas and Eguchi along with examples and additional technical reference[20]*

8. Naming conventions are problematic here, marginal in 7 reflecting unobserved (missing) observations and in 6, an unobserved random parameter. Sung and Geyer refer to both situations as missing "Missing data either arise naturally - data that might have been observed are missing - or are intentionally chosen - a model includes random variables that are not observable (called latent variables or random effects)."[115] They also point out that the first marginal likelihood is also called the observed data likelihood. Following this, in this thesis, the marginal over the unobserved data likelihood will be called the observed summary likelihood. For the marginal over the unobserved random effects likelihood, Skrondal and Rabe-Hesketh[107] call the mixing or latent variable distributions higher level distributions, with the level depending on the hierarchy involved. For instance, in meta-analysis, there would usually be only a level 1 and level 2 distribution. Following this, the level 1 likelihood is conditional on the value of unobserved random effect while the level 2 likelihood only involves parameters of the common distribution (from which the unobserved random effect was drawn), obtained by integrating over the unobserved random effect. Their terminology serves to both indicate the hierarchy involved in the likelihood as well as to distinguish it from other methods of dealing with unobserved random parameters, such as h-likelihoods[74]. The terminology clearly and succinctly identifies the likelihood that involves only the parameters of the common distribution and confirms that this likelihood was obtained by integration. It has thus been adopted in this thesis for consistency, clarity and convenience.

9. As the integrals in 6 and 7 will not be tractable in general, numerical methods will be required. For 6, the dimension is usually equal to one (arguments for this are given later) and for this, new numerical integration methods that provide valid lower and upper bounds were developed. For 7, the dimension is usually high (number of observations in $s(y)_o$) and if y^* can be uniformly sampled from, $\sum_{y^*} \Pr(y;\theta)$ will in principle provide a stochastic approximation[100]. Alternatively, as will be shown later $L(\theta; s(y)_o) = \int \frac{d\Pr(y|\theta)}{d\Pr(y|\theta_0)} d\Pr(y|s(y)_o, \theta_0)dy$ and this suggests the use of $\sum \frac{p(y|\theta)}{p(y|\theta_0)}$ as a stochastic approximation where samples are simulated from $p(y|s(y)_o, \theta_0)$ (note here only a single value of θ_0 is required). The EM algorithm and other systematic approximations might also be considered when the exact marginal distribution is not available, but they are not addressed in this thesis.

10. Bayes (Empirical Bayes) is the combination of probabilities of something common by mul-

tiplication: $\Pr(\theta) * L(\theta; y)$ where $\Pr(\theta)$ is the known (or estimated in parametric empirical Bayes) distribution of θ and $L(\theta; y)$ is as in 1.

11. In general, it is not known how to get exact confidence intervals from $\Pi_i L(\theta; y_i)$ - in particular for a common mean with arbitrary variances, e.g. where the combined likelihood is equal to $\Pi_i L(\mu, \sigma_i; y_i)$, they are known not to exist even under assumptions of Normality, (see page 77 of Sprott)[109]. Alternatively, while it is known how to get credible intervals, in general it is unknown as to how to get the meaningful priors required[57], especially if θ is of high dimension, or how the shape of these intervals should be chosen [personal conversation, Mike Evans]. However if $\log \Pi_i L(\theta; y_i)$ is approximately quadratic in the region of its maximum, at least from a practical point of view, confidence intervals and regions based on the likelihood ratio (using first order results relating to the likelihood ratio being distributed approximately as a chi-square random variable [8]) and credible intervals using non-informative or reference priors are non-controversially obtainable and are usually quite similar[99]. Unfortunately, especially with unknown scale parameters, $\log \Pi_i L(\theta; y_i)$ may be far from quadratic. Additionally, as samples are sometime "small", a possible diagnostic for Neyman-Scott type problems has been produced and initially considered .

12. With the combined likelihood $\Pi_{i=1}^n L(\theta; y_i)$ based on a common θ and

$$\Pi_{i=1}^n E_\theta L(\theta_i^*; y_i)$$

based on a common distribution of $\theta_i^* \sim p(\theta)$, some seem to suggest the use of a location estimate assuming a common θ (i.e. from $\Pi_{i=1}^n L(\theta; y_i)$) along with an empirical estimate of scale (sometimes called a robust estimate) instead of both location and scale estimates based on assuming a common distribution of θ (i.e. $\Pi_{i=1}^n E_\theta L(\theta_i^*; y_i)$) - even if the commonness is believed to just be a common distribution of θ – see page 305 of Barndorff-Nielsen and Cox [8] and also Stafford[111]. This particular way of dealing with an arbitrary common distribution of $\theta_i^* \sim \Pr(\theta)$ using an asymptotically motivated adjustment by Stafford[111] was initially appealing but was found deficient for models with unknown scale parameters or general random effects distributions such as those with asymmetric random effects, where the expectation of the fixed effect MLE is no longer relevant[102]. In fact, with some cases of multi-modal likelihoods, Stafford's adjustment lead to an increase rather than decrease in

the concentration of the log-likelihood.

13. The parameters can be grouped into within group parameters (e.g. control rate, baseline hazards, variances, etc.) and between group parameters (e.g. treatment effect, study by treatment interaction, baseline imbalance, etc.) and qualified as being of interest as opposed to of "nuisance" and as being common or non-common (just common in distribution or even arbitrary.) Random effects ("parameter" marginalizing) and profile likelihoods ("parameter" maximizing) are often useful strategies for dealing with the within group and non-common parameters (and perhaps more so profile likelihood for the within group and random effects for the between group as suggested in Bjonstad[13]). In some cases conditioning and "sample" marginalizing will also be helpful, but this is less likely for meta-analysis in clinical research than it was for meta-analysis in astronomy[81] where a large number of very small studies were actually encountered or perhaps more importantly could be anticipated.

14. It may also be useful to think of the various values of the nuisance parameters as generating various likelihoods for the parameter(s) of interest, and the issue being again the investigation and synthesis of what is possibly common in these various likelihood functions which differ for unknown values of the nuisance parameters. One could think of the unknown nuisance parameters as being like unknown sources of measurement error that caused observations of something common to differ from each other. Early astronomers debated about using un-weighted versus weighted averages as well as other methods of combining the differing observations. Integrated likelihoods $\int L(\theta_i; y_i)d\theta_i$ are an obvious "un-weighted" combination of likelihoods over nuisance parameters[12]. "Parameter" marginal likelihoods in random effects approaches weight likelihoods by the probability of the unobserved random nuisance parameter -i.e. $E_\theta L(\theta_i^*, y_i)$ or $\int L(\theta_i^*; y_i)\Pr(\theta_i^*; \theta)d\theta_i^*$. Also, according to Barnard when reviewing Bennett's collection of Fisher's correspondence [5], Fisher's interpretation of the t-likelihood was that of the likelihood integrated with respect to the Fiducial distribution of σ^2 based on s^2. Profile likelihood could also be viewed as a particular combination of differing likelihoods - where the combination is to choose from the possible values of the nuisance parameters those that are most likely ("best") for each value of the common parameter.

This parametric likelihood-based approach was to some extent anticipated in the "likelihood menu" sketched out in O'Rourke[83] and the strategy was originally suggested in the appendix

of L'Abbe, Detsky and O'Rourke[72]. This thesis greatly extends its scope and generality to arbitrary summary statistics and general probability models as well as general random effects models. A simple random effects meta-analysis specification might be helpful here. With $Normal$ assumptions for each study i, with a common within study σ_i, an arbitrary control μ_i and a random treatment effect δ_i^*, the parameters are $\theta_{i1} = (u_i, \sigma_i)$ for the control groups, $\theta_{i2} = (u_i + \delta_i^*, \sigma_i)$ with $\delta_i^* \tilde{} Normal(v, \sigma_b)$ for the treatment groups. With m studies and $j = 2$ observations per study, one each in control and treatment, there would be the following level 2 likelihoods (after integration of the unobserved δ_i^*)

$$L(((v, \sigma_b), \sigma_1, u_1); y_{11}, y_{12}) * ... * L(((v, \sigma_b), \sigma_m, u_m); y_{m1}, y_{m2})$$

with common $\Theta = (v, \sigma_b)$ and arbitrary $u_{i,}\sigma_i$.

1.3 New and adapted techniques

1.3.1 Bounds for level 2 likelihoods

The lack of closed form formulas for level 2 likelihoods for many random effects models can possibly be overcome by the use of numerical integration methods. For instance, there are a number of current proponents for the use of Adaptive Gaussian Quadrature for this[107].

Recall that Gauss rules consist of a set of n points and weights (x_i, w_i) such that $\sum_i^n p(x_i)w_i = \int_A p(x)w(x)dx$ for every polynomial $p(x)$ with degree less than or equal to $2n - 1$. By choosing $w(x)$ to be a probability density with the required number of moments, one could chose to rewrite an integrand of interest, say $f(x)$, as $\frac{f(x)}{w(x)}w(x)$. Choosing $w(x)$ to be a $Normal$ density with its mean set equal to the mode of $f(x)$ and variance set equal to the curvature of $f(x)$ at the mode gives Adaptive Gaussian Quadrature. This observation initially lead to an investigation in this thesis of various generalizations of Adaptive Gaussian Quadrature.

First, the mean and variance of a $Normal$ density were chosen to make $\frac{f(x)}{w(x)}$ a lower order polynomial than the standard choice. Second, various densities other than the $Normal$ were considered along with choices of parameter values for these that made $\frac{f(x)}{w(x)}$ a low order polynomial. The set of n points for a reasonable n can sometimes be calculated using techniques from von Mises[108]. Some initial success was obtained for some random effects models (i.e. the $Beta - Binomial$ and $Binomial - Normal$) but more generally, the approach faltered due to the difficulty

encountered in choosing a good density for $w(x)$ to make $\frac{f(x)}{w(x)}$ approximately a low enough order polynomial such that calculating the needed n points was feasible. The great strength of Gauss rules is that if the integrand is known to be a polynomial of degree less than or equal to $2n-1$, and an n point rule is used, the quadrature will have no error. Unfortunately with efforts so far, it was not possible to achieve this, even approximately, and the errors of integration using the generalizations were both unknown and possibly large.

Unfortunately, errors of integration with numerical integration techniques are usually both unknown and possibly large. For instance, with the current proposals to use Adaptive Gaussian Quadrature this would be the case if the random effects model $f(x)$ resulted in a $\frac{f(x)}{w(x)}$ that is far from a polynomial that can be integrated exactly. Typically the error analysis that accompanies numerical integration is heuristic. The error can be intrinsic to the numerical integration technique, sometimes referred to as algorithmic error, or due to representation of the problem on a specific computer, sometimes referred to as round-off error. In this thesis, the discussions are restricted to the algorithmic error which is usually much more important, though round-off error could possibly be important in some applications. With only heuristic bounds on algorithmic error one is always unsure of the validity of level 2 likelihoods from random effects models obtained via numerical integration. For instance, Geyer offers an importance sampling approach to obtaining level 2 likelihoods and comments that except for toy problems, the true level 2 likelihood always remains unknown[115]. There is, however, a numerical integration method developed by Evans and Swartz[45] that does give valid lower and upper bounds. Drawing on basic results from calculus, they showed that for concave functions $f^{(n)}$, (i.e. the n-th derivative of f),

$$\sum_{k=0}^{n}\frac{f^{(k)}(a)}{(k+1)!}(b-a)^{k+1}+\frac{f^{(n)}(b)-f^{(n)}(a)}{(b-a)}\frac{(b-a)^{n+2}}{(n+2)!}\leq\int_a^b f(x)dx\leq\sum_{k=0}^{n+1}\frac{f^{(k)}(a)}{(k+1)!}(b-a)^{k+1}.$$

For convex $f^{(n)}$ these inequalities are reversed. Given this, one can construct valid upper and lower bounds for $\int_a^b f(x)dx$ - if one can find the regions of concavity for $f^{(n)}$. Examples where the regions of concavity can be analytically determined were given by Evans and Swartz[45]. Now the regions of concavity for $f^{(n)}$ are given by the roots of $f^{(n+2)}$ (more precisely, simple roots that are associated with sign changes) and all of these can be found, with some assurance, numerically using results from Topological Degree Theory[67]. It was shown that the following integral equation counts the

number of simple roots of $f^{(n+2)}$ in the interval (a, b)

$$-\frac{1}{\pi}[\gamma \int_a^b \frac{f^{(n+2)}(x)f^{(n+4)}(x) - f^{(n+3)}(x)^2}{f^{(n+2)}(x)^2 + \gamma^2 f^{(n+3)}(x)^2} dx + \arctan(\frac{\gamma f^{(n+3)}(b)}{f^{(n+2)}(b)}) - \arctan(\frac{\gamma f^{(n+3)}(a)}{f^{(n+2)}(a)})]$$

where γ is an arbitrary small positive constant.

Given that this result allows one to count the number of roots in an interval, numerical methods were developed to implement Evans' and Swartz' numerical integration method for random effects models encountered in a meta-analysis setting. These usually involve only one dimension (i.e. a single random effects on the treatment effect - see later for arguments against additional random effects on control group parameters). An algorithm was designed that recursively splits the interval of integration until $f^{(n+2)}$ has zero roots on all of the intervals and hence $f^{(n)}$ is either concave or convex on all the intervals. Specifically, if a given interval is not easily numerically integrated with default numerical integration routines (an heuristic error warning is generated) or if the numerical integration of the integral equation indicated there was more than 1 root, the interval is split in half. On the other hand, if the default numerical integration indicated there was only 1 root, a numerical search for the root was carried out and the interval split at the root that was found. The recursion stops only when all of the numerical integrations of the integral equation on the recursively constructed intervals equal 0 at a given tolerance. Then, the rule is recursively compounded within these intervals until a given gap is achieved between the upper and lower bounds. One last check that all of these intervals have 0 roots can be made at this point (this may require a fairly long computing time). This was implemented in both R[95] and Mathematica software[122] and was successful on a number of "toy example" random effects models of interest in meta-analysis giving extremely narrow lower and upper bounds. The Mathematica program was faster and more flexible than R. As the algorithm involves many calls of standard numerical integration routines and root finding routines, the upper and lower bounds take much longer to calculate than the default numerical integration result. Further details on the algorithm implemented in the Mathematica program are given in appendix F. Perhaps surprisingly with most results to date, the default numerical integration has usually been close to the bounds (given it successfully runs) and usually somewhere between the lower and upper bound. This suggested that the algorithm could be used to provide valid lower and upper bounds at chosen points on the level 2 likelihood surface of interest rather than used to calculate all the points.

With actual bounds on the level 2 likelihoods, the methods of this thesis can safely be applied.

To apply the methods of this thesis, one needs to evaluate a likelihood surface and reduce that to a likelihood curve for a parameter of interest via profiling. It is much more feasible to first use the default numerical integration techniques to get the likelihood surface, and then check the resulting likelihood curve by obtaining the bounds at selection of parameter points from the curve on the surface traced out by the profiling. Additionally, one may wish to check that the profiling based on the default numerical integration techniques actually gave the correct curve by checking a number of points on the surface (i.e. use a grid of points) and taking the maximum of these over the nuisance parameters. This took only a few minutes of computing time for toy examples, but up to a few hours for real examples. The computational burden will not be modest for some, or perhaps even many, meta-analyses likely to be encountered in the literature - especially given the ideal of profiling the treatment effects estimates from the full combined likelihood (often greater that 2 times the number of studies). Both parallel computations and a more informed and organized computational approach along the lines currently being undertaken by Douglas Bates and colleagues[11] for generalized mixed effects models will likely be required. Despite Evans and Swartz's severe criticism of the inadequacy of so-called "error estimates" provided by numerical integration methods, many users seem fairly unconcerned. Apparently, no other numerical integration method at this time provides valid upper and lower bounds for general integrands of one dimension.

1.3.2 Importance sampling approximations for observed summary likelihoods

The lack of closed form formulas for observed summary likelihoods also presented a challenge for this thesis. Initially motivated by a formula given by Barndorf-Nielsen for the analytical derivation of marginal likelihoods, it was realized that rescaled importance sampling allowed the calculation of a likelihood surface when conditional samples were drawn from an "opportunistically" chosen single point in the parameter space. The result used was from Barndorff-Neilsen[7] and simply given (in different notation) as

$$\frac{f_U(u \mid \theta)}{f_U(u \mid \theta_0)} = \int \frac{f_X(x \mid \theta)}{f_X(x \mid \theta_0)} f_{X|U}(x \mid u, \theta_0)\, dx$$

Now, the marginal distribution is simply

$$f_U(u \mid \theta) = \int_{x^*} f_X(x \mid \theta)\, dx \text{ where } x^* \text{ is the level set given by } u = U(\mathbf{x})$$

(or more formally $x \in \{\mathbf{x} : U(\mathbf{x}) = u\}$) but only the (relative) likelihood $\frac{f_U(u|\theta)}{f_U(u|\theta_0)}$ is needed. Now

$$\frac{f_U(u \mid \theta)}{f_U(u \mid \theta_0)} = \int_{x^*} f_X(x \mid \theta)\, dx \frac{1}{f_U(u \mid \theta_0)}$$

$$\frac{f_U(u \mid \theta)}{f_U(u \mid \theta_0)} = \int_{x^*} \frac{f_X(x \mid \theta)}{f_X(x \mid \theta_0)} \frac{f_X(x \mid \theta_0)}{f_U(u \mid \theta_0)} dx$$

$$\frac{f_U(u \mid \theta)}{f_U(u \mid \theta_0)} = \int \frac{f_X(x \mid \theta)}{f_X(x \mid \theta_0)} \frac{f_X(x \mid \theta_0)}{f_U(u \mid \theta_0)} f_{U|X}(u \mid x, \theta_0)\, dx$$

$$\frac{f_U(u \mid \theta)}{f_U(u \mid \theta_0)} = \int \frac{f_X(x \mid \theta)}{f_X(x \mid \theta_0)} f_{X|U}(x \mid u, \theta_0)\, dx$$

Alternatively, starting out as importance sampling

$$f_U(u \mid \theta) = \int \frac{f_X(x \mid \theta)}{f_{X|u}(x \mid u, \theta_0)} f_{X|u}(x \mid u, \theta_0)\, dx$$

$$\frac{f_U(u \mid \theta)}{f_U(u \mid \theta_0)} = \int \frac{f_X(x \mid \theta)}{f_X(x \mid \theta_0)} f_{X|u}(x \mid u, \theta_0)\, dx$$

Conditional samples were simply generated by rejection sampling. An assumed probability distribution for the unobserved sample values was set to opportunistically chosen parameter values and a sample of the same size drawn and only those matching the reported summaries within a given tolerance were kept. An overly high rejection rate here may be suggestive of the assumed probability distribution for the unobserved sample values being inappropriate. Likelihoods for each of the kept samples were then added (no need to normalize) to get an approximation of the observed summary likelihood. This was done for reported summaries where the observed summary likelihood is available in closed form and the approximation was found to be very close, as can be seen in Figure 1 for 13 studies that reported minimums, medians and maximums on the original scale (a more complete analysis is provided later). The graph shows simulated versus exact observed summary log-likelihoods under $LogNormal$ assumptions for the log mean, with log variance treated as known and equal to 1 (for even sample sizes there are two lines, one for $n-1$ and one for $n+1$). Full likelihood methods and profile likelihood methods have also been applied to the simulated observed summary likelihoods - though the computational challenges increase with the need to deal with many simulated likelihoods.

Of course, for models with sufficiency, one only needs to condition on the sufficient statistics. Without sufficiency, one would possibly need to condition on all reported summaries. If this be-

comes a problem of feasibility there may be some pragmatic justification for just conditioning on as many as possible. More likely, there will be too little to condition on. For instance, if just p-values and sample sizes were reported on, except in exceptional models, the parameters would not be identifiable - perhaps prior information could then be used to overcome this[62]. There would be the additional challenge of less than adequate information to suggest an good θ_0 to sample from. Also, although one needs to make probability assumptions to generate the samples, those same assumptions need not necessarily be made in the analysis. The use of empirical likelihood[88] is worth at least considering and remains to be further researched. Originally, BLUE location-scale estimators that had some applicability for various assumed probability models and particular summaries of the outcomes were considered. But as these were limited to probability models in the location scale class, allowed only the use of summaries that have expectations determined by location scale transformations of the standard parameters and gave only quadratic log-likelihood approximations, they have been abandoned. They did however, turn out to be useful for obtaining good points in the parameter space for importance sampling of the observed summary log-likelihoods. Good points, that is, in the sense of making the conditional sampling efficient (not discarding too many draws) and providing parameter points not too far from the MLE thus avoiding an importance sampling distribution with a highly variable integrand.

1.3.3 Neyman-Scott diagnostics based on simulated modified profile likelihood

Concerns about Neyman-Scott problems are hard to definitively rule out. Wide experience shows that with even a few observations per nuisance parameter, they can be very minor. Theory suggests comparisons with conditional or marginal likelihoods as a gold standard test when such are available[32]. This was originally addressed in this thesis for Odds Ratios with some suggestion that the more generally available integrated likelihood be used as a comparison. Theory also suggests comparison with modified profile likelihood which is more generally available - but still limited in most cases by the need for sample space derivatives that have yet to be forthcoming. Within this theory, approximations to modified profile likelihood have been proposed that both avoid the need for sample space derivatives and can be obtained from Monte-Carlo simulation[106][90]. The required simulations can be quite demanding, but similar in strategy to the bounds for numerical integration, the profile likelihoods can be first used to get approximate confidence intervals and crucial points assessed by simulated modified profile likelihood.

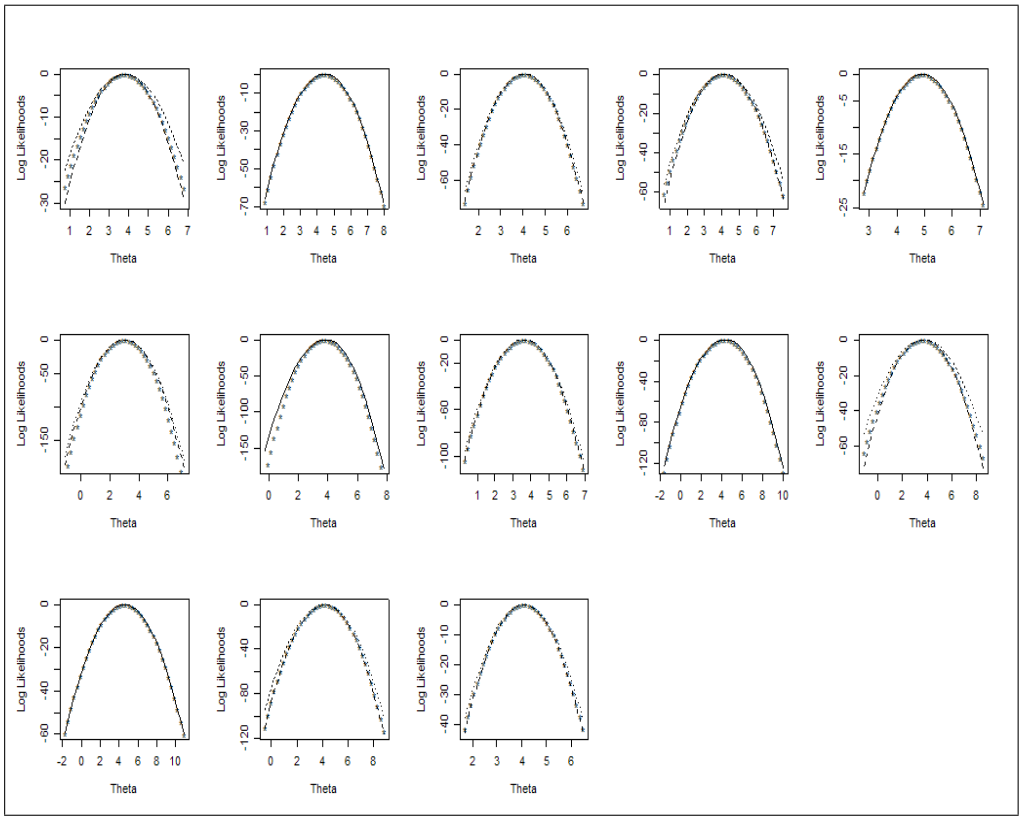

Figure 1: Simulated versus exact observed summary log-likelihoods for 13 studies that reported minimum, median and maximum. Simulated = * , Exact = solid, dotted or dashed line (odd, even +1, even - 1)

1.4 Preview of examples and findings

A number of single-group and two-group examples will be used to demonstrate the approach of this thesis. Binary and continuous outcomes will be covered. An example will be presented of a meta-analysis where some studies reported treatment group means and variances, others treatment group minimums, medians and maximums. Studies in another meta-analysis reported minimums, means and maximums. (The derivation of joint observed summary distributions minimums, means and maximums is difficult to obtain in general - for instance see page 228 Arnold, Balakrishnan and Nagaraja[3].)

When the approach of this thesis was applied even to studies that reported means and variances under the assumption the outcomes were $Normal$ - where the closed form likelihoods are directly available - the two-stage procedure (using a weighted average of study MLEs for the difference in means with weights equal to the inverse variance of the MLEs) provided a very poor approximation compared with the approach of this thesis - see especially Example 5.3. This occurred primarily due to the allowance in uncertainty in the within study scale parameters in the approach of this thesis and the range of individual MLEs. It is a good demonstration that quadratic approximations for all studies need to be close enough in the area of the maximum of the *combined likelihood* for the combined likelihood to be approximately quadratic and not just near the individual study $MLEs$. On the other hand if one entertains a $Normal - Normal$ random effects model, the individual $MLEs$ are more likely "to be local" (the total variance is inflated to help ensure this) - but the approach of this thesis with the initially entertained robust variance adjustment resulted in a very different, almost multi-modal, log-likelihood. In fact, in some cases with multi-modal log-likelihoods the robust variance adjustment from Stafford resulted in a more concentrated multi-modal log-likelihood.

This can easily be seen in the following plot of a fictional example where the $Normal - Normal$ random effects log-likelihood is plotted along with the fixed effect log-likelihood, as shown in Figure 2. The fictional data is comprised of two single group studies with exactly the same sample sizes and observed within study variances. The profile $Normal - Normal$ random effects log-likelihood for the mean is seen to be unimodal with t-distribution like tails, whereas the profile fixed effect log-likelihood for the mean is seen to be bimodal with a "batman" like shape - a simple rescaling of the fixed effect log-likelihood by a constant cannot possibly bring these two likelihoods close together. As the robust approach needs to work for the $Normal - Normal$ model, something

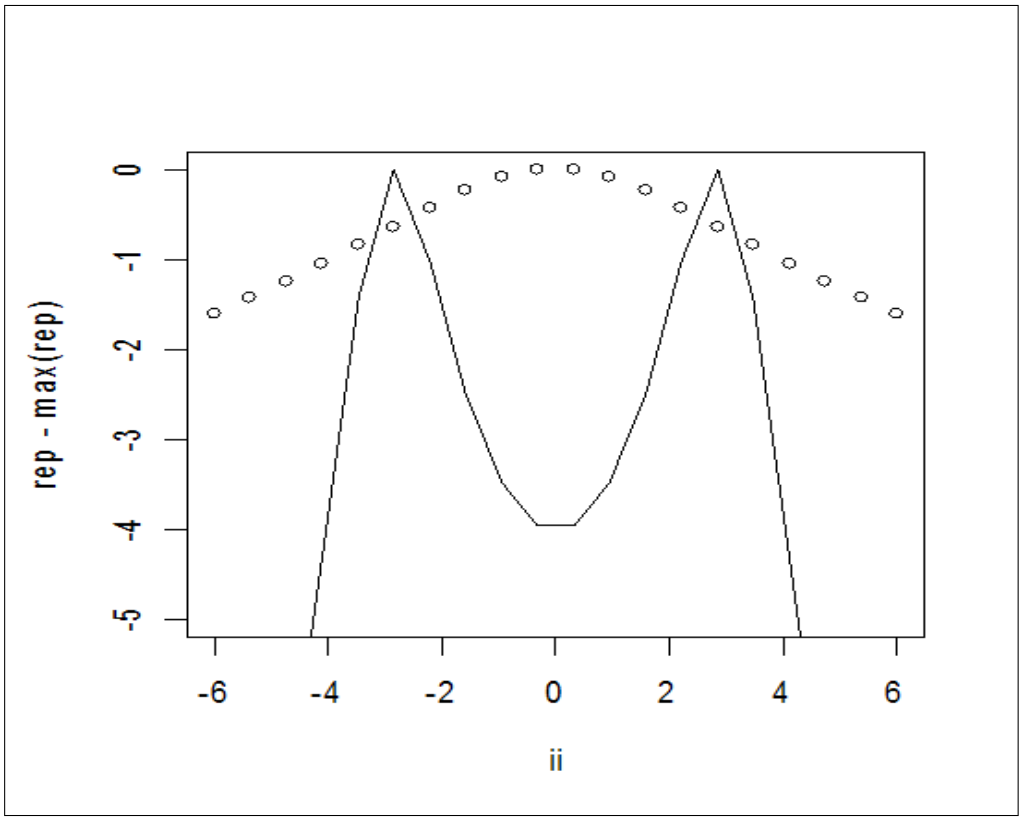

Figure 2: Fictional example to demonstrate potentially very different pooled log- likelihoods: Normal-Normal (circles) versus Robust (solid line)

"wrong" with the profile likelihood-based robust variance adjustment was discovered. On the other hand, David Andrews recently conjectured in his 2006 Statistical Society of Canada Gold Medal address, that if one will settle for just a 95% confidence interval - not a nested set of intervals but just that one in particular - there may be some "hope" for the robust approach of having approximately correct coverage. A simple simulation of multiple studies under $Normal - Normal$ random effects assumptions ruled this out for Stafford's adjustment (the adjustment actually made the coverage worse) but returning to an earlier adjustment referred to in O'Rourke as Fisher's robust adjustment[83], 93% coverage was obtained. This remains as future research. In any event, the widely accepted convention[24] of treating within study variances as known, and estimated without error is seriously misleading for fixed effect meta-analysis and even occasionally misleading for the $Normal - Normal$ random effects model[119].

1.5 Summary

By viewing much of the theory and techniques of parametric likelihood as being the investigation and synthesis of what is possibly common in observations, meta-analysis is seen to be parametric likelihood applied when what is observed are study summaries. Here, commonness of something (replication of the studies) is both of central concern and to some degree always suspect (or at least much more suspect than replication between observations within a given study). With an explicit probability model for the generation of outcomes within studies, with due consideration of what parameterization is most appropriate, common parameters or common distributions of parameters are entertained, which give rise to multivariate observed summary likelihoods for each study, based on exactly whatever summaries are available in those studies. Likelihood regions just for $\gamma(\theta)$ (the something that is possibly common) can be formed that indicate which values of the various parameters made the observations more probable than other of values. Profile likelihood provides a means to focus in on particular parameters of interest. The issue of the acceptability of profiling out all nuisance parameters needs to be addressed. There are actually two issues here - the loss of between study information on common nuisance parameters if the multivariate optimization is not feasible and univariate optimizations are used, and possible Neyman-Scott difficulties. For the first, it is simply a numerical challenge. For the second, a diagnostic was considered to help identify problematic cases and this "Neyman-Scott diagnostic" offers some assurance that this will be satisfactory but remains as future research.

In principle, this observed summary likelihood can be adjusted to take into account possible selection rules for the summaries reported[33]. Exactly what to do with this adjustment is not straightforward, but sensitivity analysis, rather than adjustment, may be more sensible[22]. Alternatively, a more formal Bayesian approach with informative priors on parameters for which there is little to no sample information may provide a more perceptive, or at least quantitative, sensitivity analysis[60].

In summary, perhaps the only real difference between meta-analysis and usual statistical applications involving a single study are the heightened concern about commonness, the real possibility of multi-modal and very non-quadratic combined log-likelihoods given assumptions of commonness or commonness in distribution, the forced marginalization of the likelihoods due to the reporting of summaries rather than raw data, and a possible need to adjust the observed summary likelihood for informative selection of summaries reported. (Especially, if p-value based censorship is

considered an extreme form of marginalization to no reporting at all.) Issues such as the quality of studies[61] that seem more relevant in meta-analysis than in individual studies perhaps simply reflect the lack of commonness of quality of studies versus quality of observations within a study. In this thesis, though, the situation of uniformly high quality randomized clinical trials is being addressed. In principle, extensions to more realistic situations are available by the inclusion of various bias parameters in the likelihood and perhaps even using prior distributions for the bias parameters [60].

The descriptively appealing and transparent model-based statistical basis in this thesis may minimize or even eliminate the differences perceived between meta-analyzing continuous versus discrete outcomes. It provides an obvious method for dealing with mixed discrete and continuous outcomes (i.e. same outcome sometimes reported on as percentage that fell above a given cut point and sometimes by summaries like the mean and variance) by using the observed summary likelihood from continuous outcomes given the discrete summary actually observed. It suggests that summaries (for groups or group differences) are neither necessary nor recommendable in statistics nor meta-analysis. Although some dimensionality reduction may be required to get confidence or credible intervals via Classical or Bayesian inference, any summarization before this implies a loss of information (under model uncertainty) and should be avoided. This becomes very clear once one considers alternative assumptions for generating the summaries that were reported - for some assumptions the summaries are sufficient and the original (and observed summary) likelihood is immediately available - while for other assumptions the original likelihood is unavailable and perhaps the observed summary likelihood intractable and simulation remains the only alternative. Something immediately available from the outcomes (a comparison of different likelihoods under different assumptions) is lost with summarization.

It encourages a more explicit search for, and wider consideration of, what is common between studies and perhaps encourages the choice to be much finer - i.e. just direction of treatment effect rather than both direction and magnitude. Magnitudes can then be treated as arbitrary and profiled out. It offers a straightforward means to investigate the sensitivity of conclusions to differing probability assumptions (i.e. robustness - at least in the sense of Barnard's conditional robustness - for the data set observed, did the assumptions matter?) as well as assumptions about informative choice of differing summaries and/or incomplete summaries in various studies (although what to do when inference is sensitive to these possibilities is not straightforward).

Hopefully, it will help clarify where robust based random effects models can be safely used - essentially quadratic study *combined* log-likelihoods with random effects distributions where the expected value of the fixed effect MLE remains relevant. Here, the true (but unknown to us) random effects log-likelihoods differ from fixed effect only by their maximum and curvature - so only the maximum and curvature are wrong. Now, the expectation of the "wrong" maximum is relevant and correct (because the expected value of the MLE under the random effects model "happens" to be equal to the expected value under the incorrect fixed effect model) albeit it is inefficient. Additionally, the curvature of the fixed effect log-likelihood is too large, but the robust adjustment replaces the wrong curvature with an estimate that, albeit again inefficient, is consistent. If we were certain as to the correct random effects model, there would be little to no reason not to use it (given the envelope numerical integration bounds made available in this thesis). It is this lack of certainty of a given model as being approximately true, or even the presence of very good reasons to believe it is false, that suggests the robust random effects would have been the "least wrong" random effects model[83]. Unfortunately, unknown scale parameters can lead to very non-quadratic profile likelihoods in regions of importance, and a wrong fixed effect log-likelihood cannot be simply remedied by replacing the curvature estimate. In the preceding plot, a possible meta-analysis with two studies contrived to illustrate this under $Normal - Normal$ assumptions was displayed - which shows that the random effects profile log-likelihood for the common mean is very different from the fixed effect profile log-likelihood which is bimodal there.

Additionally, the thesis will highlight clearly the value of randomization (by removing the need to consider and model between group nuisance parameters that likely are rather different for each study) and clarifies the principle of "as randomized" analysis (i.e. not marginalizing over nuisance parameters such as the control rate which may well have varied markedly by study, but estimating, profiling or conditioning them out). It provides a framework for meta-analysis of multiple parameters as well as adjusting for multiple parameters (i.e. regression, change from baseline) and for meta-epidemiology (analysis of meta-analyses) - in fact - practically anything that can be done with likelihood methods.

1.6 A preview of the sections

Chapter 2 surveys parametric likelihood-based statistical theory in terms of single observations as multiple sources of evidence "on their own", the evaluation of their conflict or consistency and their

combination. Here, the claim is not that statistics is meta-analysis, but that it in many (helpful) senses, can be viewed as being so. Here, how the individual observations within a study can stand on their own as statistical elements (though not necessarily independent of the other observations) and how to study, contrast and combine these individual statistical elements via likelihood will be explicated. This focussed look at the investigation and synthesis of individual observations casually defines some familiar likelihood-based statistical techniques. A strategy that then emerges from this exercise is to start with the consideration of parameters for individual observations, one at a time, and discern which components are common, common in distribution or arbitrarily different across individual observations. Given this, it is then decided which parameters are to be to focused on as interest parameters, and which are to be ignored as nuisance parameters that are somehow to be evaded. Likelihood-based methods are then reviewed as a partially successful, but not generally successful, way of doing this. Later, when studies are considered rather than individual observations, it will be argued that for most meta-analysis problems, likelihood-based methods are widely successful. A diagnostic technique will initially be considered to help check that such is the case in particular meta-analyses. Chapter 3 provides a short historical overview on the consideration of multiple sources of evidence, their individual evaluation, the evaluation of their conflict or consistency and their combination starting in 1778 and ending with the author's early experience in meta-analysis of clinical trials in the 1980's. This chapter suggests that statistical theory should be easily relatable to meta-analysis as some of its roots were there - combining observations from different experiments was, in fact, the objective in much of the early development of methods of statistical inference. Chapter 4 addresses the question of what meta-analysis of randomized clinical trials is, or should be. It provides a sketch of the differences (perhaps more apparent than real) between analysis of single and multiple studies and a brief overview of issues that arise in the consideration of reported summaries from studies perhaps selectively and incompletely reported in clinical research. Then it gives a very brief sketch of what is believed to be the current "commonly accepted" statistical approach to meta-analysis in clinical research. Chapter 5 implements the approach of the thesis in the setting of both single-group and two-group randomized studies by using various examples. Some of the examples involve mixed reporting of outcome summaries, where some studies reported treatment group means and variances, and others some combination of treatment group minimums, means, medians, maximums and variances. Chapter 6 summarizes the main findings as well as pointing out areas requiring further work.

2 Investigation and synthesis of observations

2.1 Introduction and background

In this section, a closer look at the investigation and synthesis of individual observations via likelihood, then pairs of individual observations and finally batches of observations will be undertaken. This section suggests statistics is meta-analysis or at least, it is worthwhile to try to view statistics as such. Essentially, individual observations (or pairs or batches of observations) are treated as if they are different studies. The findings for individual observations are admittedly somewhat uninteresting, being perhaps very much like reviewing set operations on null sets. A number of statistical examples (without data) are worked through in appendix C. The main purpose is to carry through parametric likelihood inference, or at least calculations for single observations (as well as pairs and batches of observations), to provide insight into individual evidence assessment, evaluation of conflict and combination for something common rather than develop definitive individual observation inference methods. With only single observations, there certainly are apparent and real limitations and no attempt was made to fully identify and resolve these. These limitations are known to persist even with a single batch of observations (i.e. in the usual statistical setting or context), and have not yet been adequately resolved[25]. Also, Sprott commented that Fisher many times expressed the view that in inductive inference comparatively slight differences in the mathematical specification of a problem may have logically important effects on the inferences possible - e.g. common ratio versus common difference in paired data[110].

In any given single study with repeated observations, acceptance that the individual observations have something in common should not be automatically and uncritically assumed. To suggest that the need to evaluate commonness of observations is not usually addressed in the statistical literature would be misleading - for instance the simple scatter plot does this. However, less explicit thought does seem to have been given to what the observations support "individually" – loosely speaking - individual observation inferences and explicit investigation and synthesis of these. One author who did this, at least with respect to defining individual elements that could be combined to get the usual linear summary based statistical techniques, was Pena[93].

Pena claimed that thinking about statistics used in single studies as a combination of estimates based on individual observations may very well make statistical theories and techniques more transparent. Pena did single observation inferences using generalized least squares and Pena

suggested that this was a useful way to understand common statistical techniques. For this thesis, a parametric likelihood approach seemed much more suitable. Pena did not consider a likelihood approach and has not done further work on this topic since (personal communication, 2002).

Now many, if not most techniques in applied statistics are based on or are closely related to the likelihood[9]. Barndorff-Nielsen[8] states

> "Likelihood is the single most important concept of statistics." and further states it is mainly just the relative likelihood - "We are almost always interested only in relative values of the likelihood at different values of θ."

Why the likelihood is so useful in developing statistical techniques - in fact so useful as to eliminate the need for any other aspect of the observations - has been the subject of a long literature. Essentially, under an assumed model, the likelihood captures all "useful inputs" for statistical procedures. Hald made the following perhaps cryptic comment on the likelihood being sufficient – "which is self-evident in view of the fact that it is a combination of the given (independent) observations and the hypothetical parametric model". Pace and Salvan[89] suggested sufficiency as a basis to restrict one's attention to relative likelihood values as they are sufficient. More formally, the distribution of any statistic conditioned on the observed relative likelihood is independent of the parameters (in the assumed model) - and hence they can provide no further "information".

The relative likelihood is therefore sufficient given the assumed model - in fact, minimal sufficient (i.e. a function of any other sufficient statistics - see page 339 Fraser[53]) . Disagreements do arise though as to what is sufficient in a given problem, for instance that between Fisher and Bartlett, with Fisher considering s^2 a being sufficient for σ^2 in the absence of knowledge about μ and Bartlett formally considering $s^2 + (\overline{y} - \mu)^2$ (see Barnard[5]).

2.2 Groups of single observations

2.2.1 Combination given a common parameter

The likelihood is the probability of the observations for various values of the parameters and will be denoted as $c(y_i)\Pr(y_i;\theta_i)$, where Pr is the assumed probability model that generated the observations, the observations are taken as fixed and the "parameter" θ_i is varied. The generic positive constant function $c(y_i)$ emphasizes that only relative values are of interest in applications, with variable y_i to indicate a possible dependence of this on y_i. Fraser avoids the direct use of a generic positive constant by formally defining the likelihood function as an equivalence class[54].

The θ_i with subscript i in the likelihood emphasizes that the parameters or some components of them may well differ by observation. A more formal definition of likelihood is that it is simply a mathematical function

$$L(\theta_i; y_i) = c(y_i) \Pr(y_i; \theta_i)$$

The formal definition as a mathematical function though, may blur that the likelihood is the probability of re-observing what was actually observed. In particular, one should not condition on something that was not actually observed such as a continuous outcome, but instead some appropriate interval containing that outcome (i.e. see page 52 of Cox & Hinkley[28][6]).

Now, the likelihood for a single observation is an "individual observation statistical element" in that the function is defined with just a single observation as $c(y_i) \Pr(y_i; \theta_i)$. Trivially, from the definition of likelihood as the probability of re-observing the outcome that was observed, for any given allowable combination of values of the parameters, the probability of re-observing the outcome that was observed is defined. Now if the observation itself is not defined in the probability model for some parameter values, the probability of that observation must be zero by the rule of total probability. See also Little and Rubin[75] where this is dealt with in terms of a natural parameter space - the set of parameter values for which $\Pr(y; \theta)$ is a proper density.

Given there is more than one individual observation likelihood and as they are probabilities, they multiply. For a probability model for a single outcome that involves only a scalar parameter denoted $\Pr(y_.; \theta_.)$, recall that

$$\Pr(y_1, y_2; \theta_1, \theta_2) = \Pr(y_1; \theta_1) * \Pr(y_2|y_1; \theta_1, \theta_2) = \Pr(y_2; \theta_2) * \Pr(y_1|y_2; \theta_1, \theta_2)$$

and simply

$$\Pr(y_1; \theta_1) * \Pr(y_2; \theta_2)$$

for independent observations. Re-emphasized as likelihoods

$$L(\theta_1, \theta_2; y_1, y_2) = L(\theta_1; y_1) * L(\theta_1, \theta_2; y_2|y_1) = L(\theta_2; y_2) * L(\theta_1, \theta_2; y_1|y_2)$$

and

$$L(\theta_1; y_1) * L(\theta_2; y_2)$$

If some parameter in the probability models used to represent each of the observations is common, then there is a combination for that parameter simply by that multiplication. With a common parameter reparameterization $\omega = \omega(\theta_1, \theta_2) = (\gamma, \chi)$, where $\gamma = \gamma(\theta_1, \theta_2)$ (the something that is possibly common) and $\chi_i = \chi(\theta_1, \theta_2)_i$ and conversely, $\theta_1 = \theta_1(\gamma, \chi_i)$ and $\theta_2 = \theta_2(\gamma, \chi_i)$, the multiplication for independent observations gives

$$L(\gamma(\theta_1, \theta_2), \chi(\theta_1, \theta_2)_{(1,2)}; y_1, y_2) = L(\gamma(\theta_1, \theta_2), \chi(\theta_1, \theta_2)_1; y_1) * L(\gamma(\theta_1, \theta_2), \chi(\theta_1, \theta_2)_2; y_2)$$

versus

$$L(\theta_1, \theta_2; y_1, y_2) = L(\theta_1; y_1) * L(\theta_2; y_2)$$

with no common $\gamma(\theta_1, \theta_2)$ and for dependent observations gives

$$L(\gamma(\theta_1, \theta_2), \chi(\theta_1, \theta_2)_{(1,2)}; y_1, y_2) = L(\gamma(\theta_1, \theta_2), \chi(\theta_1, \theta_2)_1; y_1) * L(\gamma(\theta_1, \theta_2), \chi(\theta_1, \theta_2)_1, \chi(\theta_1, \theta_2)_2; y_2|y_1)$$

versus

$$L(\theta_1, \theta_2; y_1, y_2) = L(\theta_1; y_1) * L(\theta_1, \theta_2; y_2|y_1).$$

with no common $\gamma(\theta_1, \theta_2)$. In these last two cases the combination is apparent for $\chi(\theta_1, \theta_2)_1$ in first case and θ_1 in the second case, given the dependence between y_1 and y_2 there is information about either $\chi(\theta_1, \theta_2)_1$ or $\theta_1$in y_2. For parameters that are not common, the multiplication results in a product space for them and not a combination of them - with only separate parameters for each observation. Multiplication then is "the" way to combine individual observation likelihoods that have something in common.

The main point here is that the likelihood is defined for single observations (no matter how uninteresting) and the likelihood of all the observations is some multiple of these - so we automatically have individual observation statistical elements for common parameters that, in some sense, stand on their own, so that they can be assessed for commonness and, if common, combined simply by multiplication. For instance, with binary observations when the parameter θ is a scalar - with no covariates or knowledge of order of observations - the likelihood equals either θ or $(1-\theta)$ as the observation was a success or a failure. With these likelihoods, there is no means to assess whether or not θ was common. But as is shown in appendix C, they do "automatically" provide individual observation statistical elements for individual observation inference and a means for their formal

investigation and explicit synthesis. In general, the standing "in some sense on their own" can be very weak for vector parameters θ.

2.2.2 Combination given a common distribution of a parameter

The combination of individual observations that just have the distribution of parameter(s) in common rather than a common parameter is not so straightforward because the unobserved random parameters are not common but differ, the multiplication of the $L(\theta_i; y_i)$ would initially result in a product space for them not a combination of them. But as these now considered random parameters are not observed, perhaps they should be marginalized over and if so, the distribution only involves the parameters of the higher level distribution[16]. That is

$$L^{Mp}(\Theta_i; y_i) = E_{p(\theta_i^*)}[L(\theta_i^*; y_i)] = \int L(\theta_i^*; y_i) * f(\theta_i^*; \Theta_i) d\theta_i^*$$

where Θ_i represents the parameters of the level 2 distribution (recall the definition in introduction), where the subscript indicates that the parameter components may well differ also by observation in this level 2 distribution. The superscript Mp in L^{Mp} indicates that the likelihood is marginal over the parameter where the parameter has a physical probability distribution associated with it. Some authors such as Bjornstad[13] and Berger[12] strongly argue for integrating out the parameter in this case but the acceptance of these arguments may not be universal - for instance see Lee and Nelder[74] and then Little and Rubin for direct criticisms of Lee and Nelder[75]. This will be further discussed in the appendix, but acceptance of $\int L(\theta_i^*; y_i) * f(\theta_i^*; \Theta_i) d\theta_i^*$ simplifies many issues that arise in this thesis. It is necessary to distinguish this from the case where the probability distribution associated with the parameter is simply conceptual or Bayesian, and for this

$$L^{I}(\Theta_i; y_i) = E_{p(\theta_i^*)}[L(\theta_i^*; y_i)] = \int L(\theta_i^*; y_i) * f(\theta_i^*; \Theta_i) d\theta_i^*$$

will be used and for the special case of assuming a (possibly improper) uniform distribution for the parameter,

$$L^{I_u}(\Theta_i; y_i) = \int L(\theta_i^*; y_i) d\theta_i^*$$

will be used and called respectively, the integrated and uniform integrated likelihood. Of course, these all have to be distinguished from the marginal over the sample likelihood, and this will be called the observed summary likelihood.

In the level 2 distribution $f(\theta_i^*;\Theta_i)$ the parameter(s) Θ_i could have common components and for simplicity say this applies to all of them and so Θ_i can be replaced by Θ and the multiple of these marginal likelihoods $L^{M_p}(\Theta;y_i) = \int L(\theta_i^*;y_i) * f(\theta_i^*;\Theta_i)d\theta_i^*$ does provide a combination for Θ. Here the multiplication is after the integration and under assumptions of independence of the "sampling of random" parameters Θ is simply

$$\prod_{i=1}^{k} \int L(\theta_i^*;y_i) * f(\theta_i^*;\Theta)d\theta_i^*$$

Here we are dealing with single observations drawn from the distribution with a given "sampled" parameter - when we move to dealing with pairs or groups of observations - i.e. multiple observations drawn from the distribution with a given "sampled" parameter, the probabilities of the individual observations with a given "sampled" parameter must be multiplied prior to the integration to get the marginal distribution for the pair or group. Furthermore, as pointed out earlier, there is more than one target for estimation – the common parameter Θ of the higher level distribution and the differing unobserved random parameters θ_i^* that generated the individual observations y_i but in most clinical research based on randomized trials, it is the common parameter that is of almost exclusive interest[15]. For completeness, although it will not be of concern in this thesis, for inference on the unobserved random parameters θ_i^* , $L(\theta_i^*;y_i)$ is combined with the estimated higher level distribution $f(\theta_i^*;\hat{\Theta})$ by analogy with the combination of a known distribution with a likelihood by Bayes theorem - i.e. $f(\theta_i^*;\hat{\Theta})L(\theta_i^*;y_i)$ - for instance see page 83 of Casella[16]. Given this need for random effects models with unobserved random parameters, we briefly discuss special likelihood issues that some consider important when there are unobserved random components in the level 1 likelihood under the topic "generalized likelihood" in appendix B.

This focussed look at the investigation and synthesis of individual observations via likelihood casually defines some familiar likelihood-based statistical objects and concepts such as the deviance, deviance residuals, estimated likelihood, profile likelihood, marginal likelihood, etc. A possibly effective strategy that emerges from this exercise is to start with the consideration of parameters for individual observations (or pairs or batches), one at a time, and discern which components are common, common in distribution or arbitrarily different across individual observations - then decide which parameters to focus on, and which to ignore, by treating them as nuisance parameters that are somehow to be evaded. This strategy is then seen as one way to motivate and reconstruct techniques in statistics for a single batch of observations. Here the inference limitations with a single

batch of observations (i.e. the usual statistical setting or context) are unavoidably encountered and explicated using Fisher's three classes of problems in statistics: problems of specification, inference and distribution.

The strategy of considering parameters for individual observations one at a time, and discerning which components are common, common in distribution or arbitrarily different across individual observations is seen to be an integral of part of the problem of specification - identification of an appropriate statistical model for the observations being investigated. Likelihood is one general method of estimation to "solve" the problem of inference and that method is chosen for this thesis. The further elaboration of the problem of inference in deciding which parameters to focus on, and which to ignore by treating them as nuisance parameters, then attempts to align this "general solution" to a more particular inference interest and estimated likelihood, profile likelihood, marginal likelihood, etc. are seen as partially successful but not generally successful techniques for doing this. The problem of inference also includes indications of goodness of fit for the specified model. Perhaps somewhat wider than Fisher anticipated, in a Bayesian approach it should also include the evaluation of conflict between prior and data. Box conflated both goodness of fit and prior data conflict in his approach to model checking. It is here suggested that this should be kept separate as does Evans[44] and it is further suggested that goodness of fit be further separated (when possible) into common parameter conflict and lack of fit given arbitrary (rather than common) parameters (i.e. could the model fit with arbitrary parameters?). The problem of distribution then remains, given the specification and inference, but this is now solvable in principle by Monte-Carlo simulation for most, if not all, meta-analyses of randomized clinical trials[30].

This is model-based statistics, and as some[55] have argued for in the context of investigating a single batch of observations, is the way statisticians should think and act. This thesis argues that statisticians should think about investigating multiple studies or meta-analysis in the same manner. The target output being a "set of intervals the end points of which, taken as a whole, reproduce the likelihood function, and which in repeated samples have the constant coverage frequency property that defines confidence intervals"[110]. Instead, some statisticians restrict their thinking and acting to "safe" subsets of model-based statistics (e.g. linear models or generalized linear models) and thereby both avoid particular inference interest problems (at least for large sample sizes) as well as perhaps the explicit working out of the specification, inference and distribution aspects of their application. The lack of appreciation of the value of an explicit working out of the specification,

inference and distribution aspects is perhaps encouraged by the fact that in wide generality in applied statistics, weighted least squares with empirically estimated weights can be constructed to adequately approximate pretty much any model-based statistical approach, given reasonable sample sizes of groups of independent observations. These are often confused as linear models or generalized linear models, but often violate the underlying assumptions.

On the other hand, some largely avoid the problem of particular inference interests by accepting any convenient prior distribution (perhaps very poorly motivated and even improper) which then allows the other parameters to be integrated out for particular inference interests (still marginality problems can arise[34]). Some even then try to dissuade statisticians of thinking or acting via model-based statistics unless they assume such particular priors[39]. That particular priors lead to joint distributions of parameters and data that allow the application of probability calculus to directly arrive at conditional and marginal probabilities is extremely convenient, but the salience of the inference then largely rests with how salient or helpful those particular priors are believed to be. Sensitivity analysis on a range of priors (and data models) would seem to be required and this may not be as straightforward as some of the current literature suggests[73].

For most meta-analysis problems, the particular inference interest problems encountered with profile likelihood (further discussed later) will likely not be of practical importance for most model specifications currently considered. A diagnostic technique was briefly considered to help check whether such is the case in particular meta-analyses and remains promising future research. The only real challenge then that remains, is with the specifications considered. For instance, which parameters are common and if not common, what is the appropriate specification of the form of the common distribution? Given the specification of the common distribution, the level 2 distribution arises as a problem of distribution and here it is solved by a systematic sample with known errors (i.e. numerical integration with valid lower and upper bounds). Additionally, many, many nuisance parameters may be involved in an adequate specification relating to publication bias, randomization assignment unblinding, differential loss to follow up, etc. It would seem reasonable to worry about this over the particular inference interest problem discussed above and this is simply not resolvable at present. Some recourse is achievable via sensitivity analyses with[60] or without[22] the use of prior information.

2.2.3 Lessons from single observation inference and a suggested graphical display

In appendix C, it has been shown that it is possible to define estimated or profile likelihoods that allow one to "individually" assess and combine single observations for a chosen parameter of focus or interest. A somewhat informal approach was taken in calling them likelihoods - they do not, however, correspond to the definition of likelihoods as being the probability of re-observing what was observed given various probability models. The evaluation of conflicts is also problematic - i.e. the deviance was not always helpful for assessing commonness. Also, the parameters might be highly related in that these estimated or profile likelihoods for a given observation may depend highly on the other observations that were combined to get estimates of the other parameters, and the word "individually" should not be taken as suggesting independence.

An expository example will be provided below involving an actual data set to show how to get individual observations likelihoods so that they may be assessed, compared and combined. This could have simply been accomplished by taking one parameter at a time and replacing all the others by their values in the joint MLE using all the observations, then calculating the likelihood for each observation with the now single unknown parameter. Instead the second route used in the $Normal(\alpha, \sigma^2)$ example in the appendix of profile likelihood was taken. Here again, one parameter at a time is taken, but for a reasonably possible range of values of that parameter the MLE of the other parameters is calculated given each value in that range. To calculate the likelihood for one observation then, there is only one unknown parameter but for each value of that unknown parameter the MLEs of the other parameters change to the joint MLE under that value. This was done numerically rather than in closed form, as it can easily be programmed as a Generalized Linear Model, in S-Plus. With the experience gained in this thesis, it would now be computationally more straightforward to use the full data set to obtain the profile likelihood path (values of all nuisance parameters that give the profile likelihood along the range for the variable of interest) and then simply plot the individual log-likelihoods along that path. A plot to make the addition of the individual log-likelihoods, as used in the later examples, would also be preferable.

The example was inspired by a paper entitled "There are no outliers in the stack-loss data" by J. A. Nelder[80]. In this paper, it was suggested that with the right [probability] model, there were no outliers in the stack-loss data. The stack-loss data set apparently has been used in 90 distinct papers and books on multiple linear regression and there is a loose sense of agreement about observations $1, 3, 4$, and 21 being outliers, but the union of all sets of outliers found by various

authors contains all but five of the 21 observations. Nelder chose a gamma-log link generalized linear model, and with selection and recoding of predictor variable found that the residuals were "no longer outliers, but merely close to the expected extremes". This invited a look at whether the individual likelihoods all supported something in common - i.e. examine what the individual likelihoods suggest and whether these conflict. Informally, individual likelihood plots did confirm Nelder's claim. For a better demonstration though a model where there were clearly outliers and where some conflict could be expected was chosen. For this, an earlier model that Nelder had chosen for his comparison from Draper and Smith which simply involved Normal linear regression with two predictors was used. The result is shown in Figure 3.

This graphical method for assessing the commonality of parameters is an extension of the method of locating several outliers in multiple-regression, using elemental sets[64] to single observations as opposed to sets of observations of size p. (Elemental sets of size p are the smallest set of observations that will allow consistent estimates of p parameters). All possible such sets need to be formed and investigated. It is also distinct from the graphical method for assessing the fit of a logistic regression model by Pardoe[91] which is based on Box's global approach to joint model fit (prior and data models). The method here focuses directly on data model fit and just the commonality of parameters, rather than a mixture of commonality of parameters and shape. Whether parameters appear common or not depends heavily on shape - what appears as common with heavy tailed distributions will appear non common with a light tailed distribution. If the distributional assumptions are fixed and one focuses on likelihoods under those assumptions (recall the likelihood is minimal sufficient) one is focusing on commonness. Hence, it supports the emphasis in this thesis on a strategy of considering parameters for individual observations, one at a time, and discerning which components are common, common in distribution or arbitrarily different across individual observations.

2.2.4 Tibshirani and Efron's "Pre-validation" - a non-meta-analysis example of the value of viewing parametric statistics as the combination of single observations

At the 2003 Joint Statistical Meeting in San Francisco, Robert Tibshirani gave a talk entitled "Pre-Validation and Inference in Microarrays", joint work with Bradley Efron, in which he reported on work from their earlier paper[117]. The abstract (which was the same for both the paper and talk) was as follows -

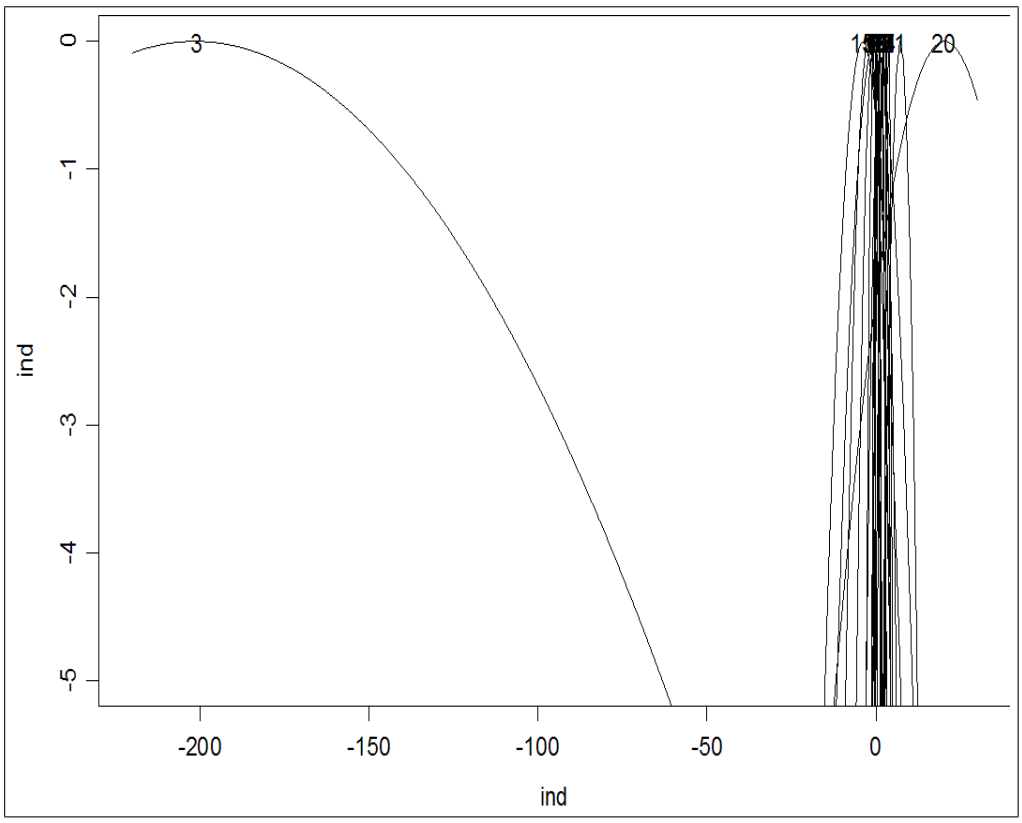

Figure 3: Individual observation log-likelihoods - Stack Loss Data

> "In microarray studies, an important problem is to compare a predictor of disease outcome derived from gene expression levels to standard clinical predictors. Comparing them on the same dataset that was used to derive the microarray predictor can lead to results strongly biased in favor of the microarray predictor. We propose a new technique called "pre-validation" for making a fairer comparison between the two sets of predictors. We study the method analytically and explore its application in a recent study on breast cancer."

The new technique called "pre-validation" had been intuitively motivated and their work involved a more rigorous evaluation of its properties - in particular determining if the degrees of freedom for the microarray predictor using the pre-validation technique were correct. Both in the talk and the paper, an alternative method was first described as the usual k-fold cross-validation approach, where the microarray predictor and clinical predictor were compared on subsets of data that omitted data on which the microarray predictor was developed, and then these comparisons on subsets were averaged (unweighted) in the paper and "somehow to be combined" in the talk. When asked about the "somehow to be combined", Robert Tibshirani made it clear that likelihood combination had not been considered. The approach of this thesis though does suggest the consideration of likelihood combination as outlined below.

First, the details from the paper are convenient to quote and provide a concise summary -

> "The microarray predictor was constructed as follows:
>
> 1. 70 genes were selected, having largest absolute correlation with the 78 class labels
>
> 2. Using these 70 genes, a nearest centroid classifier (described in detail in Section 6) was constructed.
>
> 3. Applying the classifier to the 78 microarrays gave a dichotomous predictor z_j for each case j.
>
> It was of interest to compare this predictor to a number of clinical predictors ...
>
> In order to avoid the overfitting problem ... we might try to use some sort of cross-validation:
>
> 1. Divide the cases up into say K approximately equal-sized parts
>
> 2. Set aside one of parts. Using the other K -1 parts, select the 70 genes having the largest absolute correlation with the class labels, and form a nearest centroid classifier.
>
> 3. Fit a logistic model to the kth part, using the microarray class predictor and clinical predictors
>
> 4. Do steps 2 and 3 for each of the k = 1, 2, . . . K parts, and average the results from the K resulting logistic models.
>
> The main problem with this idea is step 3, where there will typically be too few cases to fit the model. In the above example, with K = 10, the 10th part would consist of only 7 or 8 cases. Using a smaller value of K (say 5) would yield a larger number of cases, but then might make the training sets too small in step 2. Use of multiple random splits can help cross-validation a little in this case.

> Pre-validation is a variation on cross-validation that avoids these problems. It derives a "fairer" version of the microarray predictor, and then this predictor is fit along side the clinical predictors in the usual way. Here is how pre-validation was used in the bottom half of Table 1[not shown]:
>
> 1. Divide the cases up into K = 13 equal-sized parts of 6 cases each.
>
> 2. Set aside one of parts. Using only the data from the other 12 parts, select the genes having absolute correlation at least .3 with the class labels, and form a nearest centroid classification rule.
>
> 3. Use the rule to predict the class labels for the 13th part
>
> 4. Do steps 2 and 3 for each of the 13 parts, yielding a "pre-validated" microarray predictor $\tilde{z}_j$ for each of the 78 cases.
>
> 5. Fit a logistic regression model to the pre-validated microarray predictor and the 6 clinical predictors."

Drawing from this thesis, the main problem identified by Tibshirani and Efron with the first method - "there will typically be too few cases to fit the model" - does not apply to likelihood combination as likelihoods are defined for single observations (i.e. even "leave one out cross-validation" is feasible). The real problem is deciding what (if any) combination is ideal or adequate. The microarray predictors ($\tilde{z}_{g(i)}$) from the k parts are different predictors and so the coefficients for these in the logistic regression almost surely would not be common. This would also make the coefficients for the clinical covariates non-common (given inclusion of non common microarray predictors as covariates). If however, the likelihoods were given common coefficients and multiplied (an incorrect combination) one gets exactly the pre-validation technique. The pre-validation technique, being an incorrect combination, should not be expected to have good properties. First some notation will be useful.

Pre-validation is defined by Tibshirani and Efron, starting with an expression predictor $z = (z_1, z_2, ..., z_n)$ which is adaptively chosen from the data X and y

$$z_j = f_{X,y}(x_j)$$

where their notation indicates that z_j is a function of the data X and y, and is evaluated at x_j. Rather than fit $f_{X,y}$ using all X and y Tibshirani and Efron instead divide the observations into K roughly equal-sized groups, and denote by $g(k)$ the observations composing each part k. For $k = 1, 2, ...K$, Tibshirani and Efron form the pre-validated predictor

$$\tilde{z}_{g(k)} = f_{X^{-g(k)}, y^{-g(k)}}(x_{g(k)}); \text{ for } k = 1, 2, ...K$$

where the notation indicates that cases $g(k)$ have been removed from X and y. Finally, Tibshirani and Efron fit the model to predict y from $\tilde{z}$ and the clinical covariates c, and compare the contributions of $\tilde{z}$ and c in this prediction - i.e. the $\tilde{z} = (\tilde{z}_{g(1)}, \tilde{z}_{g(2)}, ..., \tilde{z}_{g(n)})$ are included with c in a multivariate statistical model. In their particular example Tibshirani and Efron used a linear logistic model

$$f\left(y|\frac{\exp(\alpha+\tilde{z}\beta+c\gamma)}{1+\exp(\alpha+\tilde{z}\beta+c\gamma)}\right).$$

Now the K within group linear logistic models with common parameters α, β and γ are

$$f\left(y_{g(i)}|\frac{\exp(\alpha+\tilde{z}_{g(i)}\beta+c_{g(i)}\gamma)}{1+\exp(\alpha+\tilde{z}_{g(i)}\beta+c_{g(i)}\gamma)}\right)$$

and their multiplication together is

$$\prod_{g(i)} f\left(y_{g(i)}|\frac{\exp(\alpha+\tilde{z}_{g(i)}\beta+c_{g(i)}\gamma)}{1+\exp(\alpha+\tilde{z}_{g(i)}\beta+c_{g(i)}\gamma)}\right).$$

Now as the observations y_j are considered independent in the linear logistic model

$$f\left(y|\frac{\exp(\alpha+\tilde{z}\beta+c\gamma)}{1+\exp(\alpha+\tilde{z}\beta+c\gamma)}\right)=\prod_j f\left(y_j|\frac{\exp(\alpha+\tilde{z}_j\beta+c_j\gamma)}{1+\exp(\alpha+\tilde{z}_j\beta+c_j\gamma)}\right)$$

and

$$f\left(y_{g(i)}|\frac{\exp(\alpha+\tilde{z}_{g(i)}\beta+c_{g(i)}\gamma)}{1+\exp(\alpha+\tilde{z}_{g(i)}\beta+c_{g(i)}\gamma)}\right)=\prod_{j\in g(i)} f\left(y_j|\frac{\exp(\alpha+\tilde{z}_j\beta+c_j\gamma)}{1+\exp(\alpha+\tilde{z}_j\beta+c_j\gamma)}\right)$$

so

$$f\left(y|\frac{\exp(\alpha+\tilde{z}\beta+c\gamma)}{1+\exp(\alpha+\tilde{z}\beta+c\gamma)}\right)=\prod_{g(i)}\prod_{j\in g(i)} f\left(y_j|\frac{\exp(\alpha+\tilde{z}_j\beta+c_j\gamma)}{1+\exp(\alpha+\tilde{z}_j\beta+c_j\gamma)}\right)$$

and it is clear that pre-validation is simply the multiplication of the likelihoods from cross-validation.

Given that pre-validation is an incorrect combination of the cross-validation linear logistic models, problems with its performance should be expected. Tibshirani and Efron did determine that pre-validation did not provide the correct degrees of freedom, but did not report on the performance of the unweighted average of the cross-validations.

2.3 Groups of pairs of observations

The smallest sample size where realistic inference can arise - two observations with at least one parameter in common (or common in distribution) is now addressed. Here real problems with likelihood as a general estimation approach were perhaps first encountered (or at least written about) by Neyman & Scott[81]. These have yet to be fully resolved in the literature and the technical details are reviewed in appendix D. It is suggested that there are important lessons from the Neyman and Scott examples for meta-analysis and, in fact, these examples were actual meta-analyses from Astronomy.

In terms of combination of observations, the common means and arbitrary variance problem arises in that the correct combination for the mean, on its own, is not quite known because it depends on the variance and this is not well estimated. Given the assumptions of Normality if one knew the relative variances, this could be fixed by multiplying the individual likelihoods by the ratio of variances and then combining by multiplication. On the other hand, treating the non-common variances as random variables changes the probability specification to one where this problem no longer remains – combination for the common mean is simply by the multiplication of the marginal (over the unknown variances) likelihoods that involve just common parameters. Of course, the mis-specification of the random distribution of these non-common variances raises additional if not more serious problems - the form of the distribution can be a quite problematic nuisance parameter and sensitivity analysis both in Classical and Bayesian inference is highly recommended.

In terms of combination of observations, the common variance and arbitrary means problem arises in that in the full likelihood, the likelihood component for the variance depends on the unknown value of the means – if the values of the means were known the correct combination would simply be accomplished by the multiplication of the likelihoods given the true means. As the likelihoods from a given pair vary for the unknown value of the mean and as they are believed to have something in common about the variance, a combination of them is desired. As they are likelihoods, a combination of probabilities is required and this requires a prior distribution for the unknown value of the means. But by considering just the marginal observations of pair differences, there is only one marginal (over the sample) likelihood that does not depend on the mean (and there is no loss of information) and a combination of it is immediate. Alternatively, treating the means as random variables changes the specification to one where this problem no longer remains –

combination is simply by the multiplication of the marginal (over the unknown mean) likelihoods.

2.4 A strategy for appropriate analyses in general

The Neyman-Scott examples involved pairing, which essentially reflects a strong belief that pairs of observations have a parameter in common. This leads naturally to specifications of common parameters by pair along with either arbitrarily different parameters across pairs or a common distribution of parameters across pairs. Given this, the single observation inference is extended to a more general setting, and the question "does it set out an appropriate analysis plan?" arises. Gelman's recently published paper[56] addressed the same question of appropriate analyses using a different approach. Quoting from Gelman's abstract

> "Analysis of variance (Anova) is an extremely important method in exploratory and confirmatory data analysis. Unfortunately, in complex problems (for example, splitplot designs), it is not always easy to set up an appropriate Anova. We propose a hierarchical analysis that automatically gives the correct Anova comparisons even in complex scenarios. The inferences for all means and variances are performed under a model with a separate batch of effects for each row of the Anova table."

Essentially, Gelman identified the analysis of variance with the structuring of parameters into batches - a separate batch of effects for each row of the table - and explicated an hierarchical modelling approach that lead to conventionally correct analyses for some analysis of variance examples, including splitplot designs. The same will now be attempted, but instead starting with individual observations, and considering which parameters are common, common in distribution or arbitrarily different, along with ways of focussing on parameters of interest while evading parameters that are not of interest.

Gelman starts with two simple examples. First, a simple experiment with 20 units completely randomized to 2 treatments, with each treatment applied to 10 units. The suggested analysis plan is a regression with 20 data points and 2 predictors: 1 constant and 1 treatment indicator (or no constant and 2 treatment indicators). This has 18 degrees of freedom available to estimate the residual variance, just as in the corresponding one way Anova, that he accepts as being appropriate. The single observation inference approach would specify 20 likelihoods, 10 with parameters (μ_c, σ) and 10 with parameters (μ_t, σ). Note the common σ over both groups and common $\mu_.$ within each

group. Now by independence the likelihood is combined as

$$\prod_{i=1}^{10} L(\mu_c, \sigma; y_{ci}) \prod_{i=1}^{10} L(\mu_t, \sigma; y_{ti})$$

and reparameterized for common $\gamma = \gamma(\mu_c, \mu_t, \sigma) = \mu_t - \mu_c, \mu_c, \sigma$ with $\delta = \mu_t - \mu_c$, is re-written as

$$\prod_{i=1}^{10} L(\mu_c, \sigma; y_{ci}) \prod_{i=1}^{10} L(\delta + \mu_c, \sigma; y_{ti})$$

or without indicating the functional relationships between the parameters δ and μ_c

$$\prod_{i=1}^{10} L(\mu_c, \sigma; y_{ci}) \prod_{i=1}^{10} L(\delta, \mu_c, \sigma; y_{ti})$$

A likelihood for δ presumably would be of interest and the profile likelihood

$$\prod_{i=1}^{10} L(\hat{\mu}_{c_\delta}, \hat{\sigma}_\delta; y_{ci}) \prod_{i=1}^{10} L(\delta + \hat{\mu}_{c_\delta}, \hat{\sigma}_\delta; y_{ti})$$

would be used to focus on δ while evading μ_c, σ. With small sample sizes, modified profile likelihood may seem preferable, but the profile likelihood is just overconcentrated here and Monte-Carlo calibration could correct this and would be needed to calibrate the modified profile likelihood in any event[92]. Gelman's solution, as confirmed in McCullagh's comment that "it is the simplest neoclassical procedure ... i.e. first to compute the variance components using residual maximum likelihood ... and then to compute ... summary statistics", was to replace $\hat{\sigma}_\delta$ by the MLE from the marginal over the sample likelihood (commonly referred to as the restricted MLE). The result is a quadratic log-likelihood with a $\hat{\sigma}_r$ slightly larger than $\hat{\sigma}_{\hat{\delta}}$.

Next, Gelman considers a design with 10 pairs of units, with the 2 treatments randomized within each pair. The corresponding regression analysis he suggests has 20 data points and 11 predictors: 1 constant, 1 indicator for treatment, 9 indicators for pairs, and, if you run the regression, the standard errors for the treatment effect estimates are automatically based on the 9 degrees of freedom for the within-pair variance (by convention, he used the restricted MLE for the estimation of σ). Alternatively, he could have used 1 indicator for treatment and 10 indicators for pairs. The single observation inference approach would again specify 20 likelihoods, 10 with parameters (μ_j, σ) (a different μ_j for each pair j) and 10 with parameters (δ, μ_j, σ) (the same μ_j for within each pair

j). Note within the pair j, by independence the likelihood is simply

$$L(\mu_j, \sigma; y_{cj}) L(\delta + \mu_j, \sigma; y_{tj})$$

and by independence of pairs, and a common δ and σ the combined likelihood over pairs is

$$\prod_{j=1}^{10} L(\mu_j, \sigma; y_{cj}) L(\delta + \mu_j, \sigma; y_{tj}).$$

A likelihood for δ would presumably be of interest and possibly focussed on by the profile likelihood

$$\prod_{j=1}^{10} L(\hat{\mu}_{j_\delta}, \hat{\sigma}_\delta; y_{cj}) L(\delta + \hat{\mu}_{j_\delta}, \hat{\sigma}_\delta; y_{tj}).$$

As in the first example, the same profile likelihood problems and solutions arise. But in this example, one may wish to use the marginal (over the sample) likelihood-based just on the paired differences (for δ not just σ)

$$\prod_{j=1}^{10} L^{Ms}(\delta, \sigma; d) = \prod_{j=1}^{10} \int_{y_{cj}, y_{tj}: y_{cj} - y_{tj} = d} \Pr(y_{cj}, y_{tj}; \delta, \mu_j, \sigma) dy$$

or one might wish to consider the μ_j^* as random and use

$$\begin{aligned} \prod_{j=1}^{10} L^{Mp}(\delta, \mu, \sigma; y_{cj}, y_{tj}) &= \prod_{j=1}^{10} E_{p(\mu_j)}[L(\mu_j^*, \sigma; y_{cj}) L(\delta, \mu_j^*, \sigma; y_{tj})] \\ &= \prod_{j=1}^{10} \int L(\mu_j^*, \sigma; y_{cj}) L(\delta, \mu_j^*, \sigma; y_{tj}) * f(\mu_j^*; \mu) d\mu_j^*. \end{aligned}$$

The latter would likely be "wrong" for treatments randomized within pairs (see arguments later).

Now, for the same experimental setup, if the outcomes were binary, the profile likelihood

$$\prod_{j=1}^{10} L(\hat{\mu}_{j_\delta}; y_{cj}) L(\delta, \hat{\mu}_{j_\delta}; y_{tj})$$

would result in unconditional logistic regression with a known scale parameter but which is known to be highly biased for δ. On the other hand, with continuous outcomes, if the choice is to focus on the common σ rather than the common δ (i.e. let δ be arbitrary) and treat μ_j^* and δ as nuisance parameters to be dealt with by profile likelihood the Neyman-Scott problem arises again.

So the individual observation likelihood approach gives more explicit options and shows some as poor or even unacceptable. Gelman concludes that the different analyses for paired and unpaired designs are confusing for students, but argues in his approach they are clearly determined by the principle of including in the regression all the information used in the design (perhaps using an argument by "authority"). The individual observation likelihood approach suggests that it is whether parameters are considered common, common in distribution or arbitrarily different along with which parameters of interest are focussed on versus which nuisance parameters were evaded and how, that determines an "appropriate" analysis. This also depends highly on the form of the distributions assumed, as the last example clearly highlighted.

2.5 Summary

In this investigation and synthesis of observations, various challenges arose when dealing with multiple parameters, many of which could be nuisance parameters (which here includes the random effects in random effects models). Quoting from Cox[25] page 171 "Numerous variants have been proposed, many based on marginalizing or conditioning with respect to well-chosen statistics or using inefficient estimates of some components while retaining asymptotic efficiency for the components of interest. The incorporation of all these into a systematic theory would be welcome." In this thesis, *profile likelihood* will be used for evading fixed nuisance parameters and the *likelihood curvature adjustment* (details in appendix E) will be considered but rejected for evading random nuisance parameters, except possibly for essentially quadratic combined log-likelihoods where the expectation of the fixed effect MLE is still considered relevant. Profile likelihood is known to have poor properties in certain situations (i.e. the Neyman-Scott examples) and although these are unlikely to arise in many meta-analyses where standard asymptotic results as reviewed on page 243 of Barndorff-Nielsen[8] are likely to apply, a diagnostic to help ensure acceptable performance in particular applications has initially been considered. Further considerations regarding "evasions" of nuisance parameters and other statistical background issues are given in appendix F.

3 Short history of likelihood for meta-analysis

3.1 Pre-Fisher

Meta-analysis, or at least the combination of observations "not necessarily made by the same astronomer and or under different conditions", figured prominently in the initial development of statistical theory in the 18th and 19th centuries[113][63][84]. There was a belief (or at least a hope) that something could be gained from combining the observations however, exactly how the observations should be combined to achieve exactly what gain was far from obvious. This spurred the development of "Bayesian-like" methods that utilized the likelihood as a means of combining observations and offered justifications of this method of combination as providing the most "probable" true value (though originally conceived somewhat less directly as the most probable error of measurement made in the observations). The justifications though, were not as well formalized and understood as current Bayesian justifications of, for instance, the most probable value of an unknown parameter, given an explicitly known prior probability distribution for that unknown parameter[63]. The confusion in the justifications in fact, was wide-spread before and some time after Fisher's thesis of 1912[46], when according to Hald, Fisher vaguely drew attention to some of the difficulties. An exception may have been Keynes' 1911 paper[70] which will be mentioned below where the role of various assumptions was very clearly delineated.

Often, justifications of intuitively reasonable combinations that involved either the mean or various weighted means of multiple observations were argued about and sought. In fact, the early attempts to justify combinations based on "likelihood" were largely abandoned when the mathematical analysis under the "primitive" probability models assumed for the observations was found intractable at the time[63]. One early attempt was by Daniel Bernoulli in a 1778 paper entitled "The most probable choice between several discrepant observations and the formation therefrom of the most likely induction." which was reprinted in Biometrika 1961[69].

In the English translation, most references to what was being combined in the paper were to observations, but the term observers was also used - suggesting that observations were not highly distinguished as coming from the same or different investigators/studies. In this he enunciated a principle of maximum likelihood "of all the innumerable ways of dealing with errors of observations one should choose the one that has the highest degree of probability for the complex of observations as a whole." He believed that the mean was a poor combination using intuitive arguments that

observations further from the "centre" should be given less weight, except in the case where the observations were believed to be Uniformly distributed - where he incorrectly (from a likelihood combination perspective) believed all observations should be equally weighted.

He assumed a semicircular distribution and derived the likelihood function as

$$L = \prod_{i=1}^{n} \sqrt{a^2 - (y_i - \mu)^2}$$

and tried to maximize L^2 with respect to μ but was unable to do this for more than 2 observations, as it lead to an equation of the fifth degree. For just 2 observations, it was maximized by the mean. For some numerical examples with 3 observations he noted that L^2 was maximized by weighted means. The idea of using a probability model to determine the best combination was definitely there, and he did realize that the probability of individual independent observations multiplied to provide the joint probability of the complex of observations. Interestingly, he actually used the smallest observation as the origin i.e.

$$y_i - \mu = \left(y_i - y_{(1)}\right) - \left(\mu - y_{(1)}\right)$$

which emphasizes the correction that needs to be added to $y_{(1)}$ - $u - y_{(1)}$ as the unknown. Unfortunately for him, if instead of $\partial L^2/\partial u = 0$ he had used $\partial \log L/\partial u = 0$ he would have found

$$\begin{aligned} \sum \frac{y_i - \mu}{a^2 - (y_i - \mu)^2} &= 0 \\ \sum \frac{y_i}{a^2 - (y_i - \mu)^2} - \mu \sum \frac{1}{a^2 - (y_i - \mu)^2} &= 0 \end{aligned}$$

$$\begin{aligned} \mu \sum \frac{1}{a^2 - (y_i - \mu)^2} &= \sum \frac{y_i}{a^2 - (y_i - \mu)^2} \\ \mu &= \sum \frac{y_i}{a^2 - (y_i - \mu)^2} \Big/ \sum \frac{1}{a^2 - (y_i - \mu)^2}. \end{aligned}$$

This shows that $\hat{\mu}$ is the weighted average of the observations, the weights being the reciprocal of the squared density, that is, increasing with the distance from μ. It is also unfortunate that he did not consider multiplying individual observation likelihoods assuming $Uniform(u - h, u + h)$ with h known, as the mathematics is simple and the best combination involves only the most extreme observations on each side of the centre – about the most different combination from the one he

intuitively thought best in this case (the equally weighted mean which puts equal weight on all observations).

Somewhat later, circa 1800, Laplace and Gauss fully investigated the multiplying of probabilities of individual observations as the means of combination of observations given a common parameter[84]. Laplace initially concentrated on the probable errors, and often specifically the most probable error, given all the observations (and a more or less explicit assumption of a prior uniform distribution on the possible errors). Gauss moved towards concentrating on the probable values rather than errors and specifically the most probable value given all the observations (and a very explicit assumption of a prior uniform distribution on the possible values). Gauss was also perhaps the first with some real practical success. He reversed the reasoning that Bernoulli had used earlier – rather than trying to establish that the mean is the best combination for some "motivated by first principles" distribution, and he found the distribution for which "likelihood multiplication" would determine that the best combination was the mean. According to Hald, he did not check the distribution empirically[63].

In 1839 Bienayme had remarked that the relative frequency of repeated samples of binary outcomes often show larger variation than indicated by a single underlying proportion and proposed a full probability-based random effects model (suggested earlier by Poisson) to account for this. Here, the concept of a common underlying proportion was replaced by a common distribution of underlying proportions. It is interesting that a random effects model where what is common in observations is not a parameter, but a distribution of a parameter, followed so soon after the development of likelihood methods for combination under the assumption of just a common parameter.

The 1911 paper of Keynes mentioned above, acknowledged and revisited Gauss's derivation of the *Normal* distribution as the only symmetric distribution whose best combination was the mean and also investigated this for the median and the mode. Here, simply for interest in itself, the result of the *Normal* distribution as the only symmetric distribution whose best combination is the mean, is presented in modern form but following that given in Keynes' paper.

The assumption of a symmetric distribution obviously does not imply that

$$f\left(y_i;\mu\right) = Be^{\theta(\mu-y_i)^2}.$$

It is required to show that

$$\sum_{i=1}^{n} \frac{\frac{\partial}{\partial \mu} f(y_i, \mu)}{f(y_i; \mu)} = 0 \text{ being equivalent to } \sum_{i=1}^{n} (\mu - y_i) = 0$$

along with symmetry does imply this.

The most general form of $\sum_{i=1}^{n} (\mu - y_i) = 0$ is $\sum_{i=1}^{n} g(\mu)(\mu - y_i) = 0$, where g is an arbitrary function of μ. Assuming $g(\mu)$ to be twice differentiable, without loss of generality, one may write $g(\mu) = \varphi''(\mu)$. Since y_i is arbitrary, the equivalence requires

$$\frac{\frac{\partial}{\partial \mu} f(y_i, \mu)}{f(y_i; \mu)} = \varphi''(\mu)(\mu - y_i)$$

or

$$\log f(y_i; \mu) = \int \varphi''(\mu)(\mu - y_i)\, d\mu + \psi(y_i),$$

where $\psi(y_i)$ is an arbitrary function of y_i. Integration by parts gives

$$\log f(y_i; \mu) = \varphi'(\mu)(\mu - y_i) - \varphi(\mu) + \psi(y_i).$$

Now, it is required that $f(y_i; \mu)$ be symmetric about μ, *i.e.* invariant under $y \to 2\mu - y$. Thus

$$\varphi'(\mu)(\mu - y_i) - \varphi(\mu) + \psi(y_i) = \varphi'(\mu)(y_i - \mu) - \varphi(\mu) + \psi(2\mu - y_i)$$

or

$$2\varphi'(\mu)(\mu - y_i) = \psi(2\mu - y_i) - \psi(y_i).$$

Taylor expand $\psi(2\mu - y_i)$ about $\mu = y_i$.

$$\psi(2\mu - y_i) = \psi(y_i) + 2\psi'(y_i)(\mu - y_i) + 2\psi''(y_i)(\mu - y_i)^2 + \sum_{j=3}^{\infty} \frac{2^j}{j!} \psi^{(j)}(y_i)(\mu - y_i)^j,$$

which results in

$$\varphi'(\mu)(\mu - y_i) = \psi'(y_i)(\mu - y_i) + \psi''(y_i)(\mu - y_i)^2 + \sum_{j=3}^{\infty} \frac{2^{j-1}}{j!} \psi^{(j)}(y_i)(\mu - y_i)^j,$$

which simplifies to

$$\varphi'(\mu) = \psi'(y_i) + \psi''(y_i)(\mu - y_i) + \sum_{j=3}^{\infty} \frac{2^{j-1}}{j!} \psi^{(j)}(y_i)(\mu - y_i)^{j-1}.$$

But $\varphi'(\mu)$ is a function of μ alone for arbitrary y_i, which implies that $\psi''(y_i) =$ a constant along with $\psi'(y_i) - y_i\psi''(y_i) = 0$, which implies that $\psi(y_i) = ky_i^2$. Then

$$\varphi'(\mu) = 2k\mu \quad \Rightarrow \quad \varphi(\mu) = k\mu^2 + C.$$

Substituting in the equation for $\log f((y_i;\mu))$,

$$\begin{aligned} \log f(y_i;\mu) &= 2k\mu(\mu - y_i) - k\mu^2 - C + ky_i^2 \\ &= k(\mu - y_i)^2 - C, \end{aligned}$$

or

$$f(y_i;\mu) = Ae^{k(\mu - y_i)^2}.$$

Note that

$$\int f(y_i;\mu)\, dy_i = 1 \quad \Rightarrow \quad k < 0.$$

Keynes' paper provides a good indication of the central role played by the combination of observations in statistics prior to Fisher. Apparently though, only Keynes, Gauss, Laplace and perhaps a few others were fully aware of the need for, and arbitrariness of, a prior distribution for the probability justifications for the combination, and both Gauss and Laplace became at some point uncomfortable with this and turned to sampling distribution-based justifications instead[63]. In particular, Gauss developed optimal combinations based on a restriction to unbiased linear combinations of unbiased estimates (i.e. least squares or inverse variance weighted combinations). This approach allowed for varying but known differences in the variances of the estimates and implicitly assumed the estimates and variance were uncorrelated so that weighted averages of unbiased estimates would give unbiased combinations (which is, of course, trivially true for known variances).

Somewhat later, based on Laplace's expositions of his own and Gauss's work, Airy made an extension for the estimation of unknown variances in 1861[1]. He also made a related extension to

Bienayme's "random effects model" by developing methods based on within day and between day variances of observations to allow for imperfect but partial replication of independent estimates. Consideration of the within day sampling errors had shown that in some applications, observations on different days were not in fact replicating in the usual sense – they had larger variations than would be expected from the within day sampling errors. It was conceptualized that there was some unknown day error and that some allowance should be made for this.

The two-stage summary approach to meta-analysis used today is close to this approach, but where the implicit assumption is often violated as, for instance, with effect measures which are slightly correlated with their variances.[65] Fisher though, as we will see below, returned to likelihood (separated from the prior) and again provided arguments for likelihood multiplication as the "best" basis for combining observations in the early 1900s. Pearson wrote an editorial on Airy's book[84] and Fisher, as a graduate student, either studied Airy's book or related ones on the combination of observations[63]. Pearson meta-analysed medical examples in the early 1900s, drawing attention to opportunities suggested by the heterogeneous study outcomes. Fisher and Cochran meta-analysed agricultural trials in the 1930s[49][19]. Fisher drew attention to the need to carefully consider the reasons for less than perfect replications between trials (i.e. whether in fact it was a treatment interaction with place and time or differing measurement errors) and various ways of dealing with it for different inferential purposes. It apparently is one of Fisher's few publications on random effects models (private conversation with D. Sprott and J. A. Nelder). Cochran explicated the full $Normal - Normal$ random effects model with a likelihood-based meta-analysis in 1937. Further details are given in O'Rourke[84].

3.2 Early Fisher

In some ways, perhaps most interesting of all, Fisher in his 1925[47] and 1934[48] papers in which he mainly developed his theory of statistics, thought through the issues of multiple experiments when addressing the loss of information when summarizing data. In the 1925 paper, he points out that if there is no loss of information in a summary (i.e. when there are sufficient statistics) then the summary of two combined samples from the same population must be some function of the two summaries of the individual samples without recourse to the individual observations from either sample. He then concludes the paper with a section on ancillary statistics whose purpose was defined as providing a true, rather than approximate, weight for combining the multiple individual

sample summaries.

In the case of a [small] number of large samples, he shows that the likelihood from all the individual observations collected from all the samples can be recovered from the MLEs of the multiple individual samples via a weighted average of those MLEs with weights equal to the observed information (second derivative of the log-likelihood evaluated at the MLE) of each individual sample. Essentially this is because, for large samples, the log-likelihoods are approximate quadratic polynomials and their addition only involves their maximums (MLEs) and curvatures (observed informations evaluated at the MLE essentially estimated without error and taken as known).

Following Hald[63] and using modern notation

$$l = \log L(\theta, y_{all}) = \sum_k \log L(\theta, y_k) = \sum_k l_k$$

and therefore

$$l^{'}(\widehat{\theta}) = \sum_k l_k^{'}(\widehat{\theta}).$$

By Taylor series expansion about $\widehat{\theta}_k$

$$l_k^{'}(\widehat{\theta}) = l_k^{'}(\widehat{\theta}_k) + (\widehat{\theta} - \widehat{\theta}_k) l_k^{''}(\widehat{\theta}_k) + ...$$

$$\sum_k l_k^{'}(\widehat{\theta}) = \sum_k l_k^{'}(\widehat{\theta}_k) + \sum_k (\widehat{\theta} - \widehat{\theta}_k) l_k^{''}(\widehat{\theta}_k) + ...$$

$$\sum_k l_k^{'}(\widehat{\theta}) = \sum_k 0 + \sum_k (\widehat{\theta} - \widehat{\theta}_k) l_k^{''}(\widehat{\theta}_k) + ...$$

$$\sum_k l_k^{'}(\widehat{\theta}) = \sum_k (\widehat{\theta} - \widehat{\theta}_k) l_k^{''}(\widehat{\theta}_k) + ...$$

but $\sum_k l_k^{'}(\widehat{\theta}) = 0$ so

$$\sum_k (\widehat{\theta} - \widehat{\theta}_k) l_k^{''}(\widehat{\theta}_k) \approx 0$$

$$\widehat{\theta} \sum_k l_k^{''}(\widehat{\theta}_k) - \sum_k \widehat{\theta}_k l_k^{''}(\widehat{\theta}_k) \approx 0$$

$$\widehat{\theta}\sum_k l_k^{''}(\widehat{\theta}_k) \approx \sum_k \widehat{\theta}_k l_k^{''}(\widehat{\theta}_k)$$

$$\widehat{\theta} \approx \frac{\sum_k \widehat{\theta}_k l_k^{''}(\widehat{\theta}_k)}{\sum_k l_k^{''}(\widehat{\theta}_k)}$$

Since each of the estimates $\widehat{\theta}_k$ is asymptotically $Normal(\theta, 1/nI)$, the combination based simply on the unweighted average $\overline{\theta} = \sum_k \widehat{\theta}_k/m$ would have variance $1/mnI$. Note, however, that the above combination recovers the likelihood from the full data and the $\widehat{\theta}$ from this is asymptotically $Normal(\theta, 1/l^{''}(\widehat{\theta}))$.

The advantage is perhaps more easily seen in terms of variances from the finite sample version given by Rao[96] -

> "Suppose that we have two independent samples X and Y, giving information on the same parameter θ, from which estimates $T_1(x)$ and $T_2(y)$ obtained are such that
>
> $$\begin{aligned} E[T_1(X)] &= E[T_2(Y)] = \theta, \\ V[T_1(X)] &= v_{1,} V[T_2(Y)] = v_{2,} \end{aligned}$$
>
> where v_1 and v_2 are independent of θ. Further, suppose that there exist statistics $A_1(X)$ and $A_1(Y)$ such that
>
> $$\begin{aligned} E[T_1|A_1(X) &= A_1(x)] = \theta, \\ E[T_2|A_2(Y) &= A_2(y)] = \theta, \end{aligned}$$
>
> $$\begin{aligned} V[T_1|A_1(X) &= A_1(x)] = v_1(x), \\ V[T_2|A_2(Y) &= A_2(y)] = v_2(y), \end{aligned}$$
>
> where x and y are observed values of X and Y, respectively, and $v_1(x)$ and $v_2(y)$ are independent of θ. Then, we might consider the conditional distributions of T_1 and T_2 given A_1 and A_2 at the observed values and report the variances of T_1 and T_2 as $v_1(x)$ and $v_2(y)$, respectively, as an alternative to v_1 and v_2. What is the right thing to do?
>
> Now, consider the problem of combining the estimates T_1 and T_2 using the reciprocals of v_1, v_2 and $v_1(x)$, $v_2(y)$ as alternative sets of weights:
>
> $$t_1 = (\frac{T_1}{v_1} + \frac{T_2}{v_2})/(\frac{1}{v_1} + \frac{1}{v_2}),$$
>
> $$t_2 = (\frac{T_1}{v_1(x)} + \frac{T_2}{v_2(y)})/(\frac{1}{v_1(x)} + \frac{1}{v_2(y)}).$$
>
> It is easy to see that the unconditional variances of t_1 and t_2 satisfy the relation
>
> $$V(t_1) \geq V(t_2)$$

[by application of the Guass-Markov Theorem, conditional on x and y]."

In the 1934 paper, he addressed the same question for small samples (where the log-likelihoods can be of any form) and concluded that, in general, single estimates will not suffice but that the entire course of the likelihood function would be needed. He then defined the necessary ancillary statistics in addition to the MLE in this case as the second and higher differential coefficients at the MLE (given that these are defined). These would allow one to recover the individual sample log-likelihood functions (although he did not state the conditions under which the Taylor series approximation at a given point recovers the full function - see Bressoud[14]) and with their addition, the log-likelihood from the combined individual observations from all the samples.

The concept of ancillary statistics has changed somewhat since - in fact very soon afterwards, as a year later Fisher treated "ancillary" as a broader term of art not specifically wedded to local behavior of the likelihood function[114]. This was its original conceptualization though – how to "correctly" (without loss of information) combine results from separate sample summaries, given a choice of what the separate sample summaries should be but no access to the individual observations in the separate samples. Here, "correctly" is defined as getting some multiple of the likelihood function from all the observations but with access only to the collection of summaries.

It is perhaps tempting to suggest that Fisher's key ideas in his theory of statistics (the breadth of which is for instance reflected in Efron's claim that modern statistical theory has added only one concept, that of invariance, which is not well accepted[41]) arose from his thinking of statistics as the combination of estimates. Fortunately for us, Fisher as much said so in a 1935 paper read at the Royal Statistical Society[50]. In discussing overcoming the preliminary difficulty of multiple criteria for judging estimates – better for what? – he argued

> "Whatever other purpose our estimate may be wanted for, we may require at least that it shall be fit to use, in conjunction with the results drawn from other samples of a like kind, as a basis for making an improved estimate. On this basis, in fact, our enquiry becomes self contained, and capable of developing its own appropriate criteria, without reference to extraneous or ulterior considerations."

And later in the next paragraph –

> " . . . , where the *real* problem of finite samples is considered, the requirement that our estimates from these samples may be wanted as materials for a subsequent process of estimation [combined somehow with results drawn from samples of a like kind?] is found to supply the unequivocal criteria required." [italics in the original]

3.3 Late Fisher

Fisher continued to exhibit numerous references to multiple estimates or studies in his 1956 book Statistical Methods and Scientific Inference[51]. For instance, on page 75, he states

> "It is usually convenient to tabulate its [the likelihoods] logarithm, since for independent bodies of data such as might be obtained by different investigators, the "combination of observations" requires only that the log-likelihoods be added."

On page 163 he further notes

> "In practical terms, if from samples of 10 two or more different estimates can be calculated, we may compare their values by considering the precision of a large sample of such estimates each derived from a sample of only 10, and calculate for preference that estimate which would at this second stage [meta-analysis stage] give the highest precision."

Finally on page 165 he concludes

> "...it is the Likelihood function that must supply all the material for estimation, and that the ancillary statistics obtained by differentiating this function are inadequate only because they do not specify the function fully."

Given this, it is suggested that Fisher considered the theory of estimation as validly based on the idea of retaining "all" of the likelihood in the estimates "summarized" from studies so that the overall likelihood-based on the individual observations from similar studies could be re-constituted by just using the studies' estimates. This metaphor or model of estimation was continually referred to through many of his publications - though perhaps even few familiar with Fisher's work have noticed that (AWF Edwards, private communication). Fisher was even cited as being the main impetus for one of the earliest papers on p-value censorship[112]. There is some note of it given in Savage [104], which suggested to the author that Fisher's papers should be reviewed for this, and also in Rao[96].

In conclusion, the early development of statistics in the context of combination of observations and Fisher's numerous and continued references to multiple estimates or summaries in his statistical writing suggests that statistical theory should be easily relatable to meta-analysis as some of the roots and elaborations of statistical theory were based on meta-analytical considerations.

3.4 Post-Fisher

The history chapter in this thesis started with the combination of observations made by different astronomers and geodesists in the late 1700's and early 1800's and then concluded with some excerpts from Fisher's 1956 book. Unfortunately, the quantitative combination of estimates from randomized clinical trials was quite rare before about 1980 so there is a need to bridge the gap. Meta-analysis for psychological and educational research started somewhat earlier, and by 1976 Glass highlighted the desirability of the tradition of combining estimates from different studies and apparently first coined the term meta-analysis. Some authors argue that meta-analysis methods for clinical research were initially based on this activity in psychological and educational research. In educational and psychological research however, studies would very often use different outcomes or scales, and to this end, Glass proposed the use of an index of effect magnitude that did not depend on the arbitrary scaling of the outcomes so that combining in some sense, made sense. Presumably, in response to this, Hedges and Olkin wrote a book[65] in 1985 directed (as the authors indicated) at providing different statistical methods from those of Fisher & Cochran that were designed to specifically deal with this new and different kind of meta-analysis - that of combining different outcomes using an index of magnitude. In 1990, Olkin[82], quoting Fisher, again highlighted this arguably different class of meta-analyses (which apparently are more common in psychology and education than clinical research) of determining the combined significance of independent tests on outcomes "that may be of very different kinds" (by combining their p_values.)[84].

Hedges and Olkin's book, although a substantial and now classic book for combining different outcomes using an index of magnitude, is somewhat out of place for the more usual situation encountered in clinical research where a series of randomized clinical trials have identical or very similar outcomes. Here Fisher and Cochran's methods would be arguably more appropriate. (With recent changes in clinical research, specifically the inclusion of Quality of Life measures which are comprised of various scales, this may be less the case for those outcomes.)

DerSimonian and Laird[38], published in 1986 what was perhaps one of the first "modern" papers on statistics for meta-analysis for randomized clinical trails. It drew on and referenced a 1981 paper[97] that W. G. Cochran was the senior author on (published posthumously) that was comprised of simulation studies of various estimators of combined estimates from Cochran's 1937 $Normal - Normal$ random effects model[19]. DerSimonian and Laird chose to adopt one the closed form non-iterative formulas from this paper and adapted it for binary outcomes. Two more

methodological as well as statistical papers appeared in the next year - Sacks et al[103] and L'Abbe, Detsky and O'Rourke[72] (the author of this thesis). The authors of these three papers had been loosely collaborating since 1985. In particular, Chalmers had provided a draft of his quality scoring system and DerSimonian and Laird had provided their draft paper to the author when the L'Abbe group were developing their ideas and paper. There it was suggested that logistic regression be used for conducting meta-analyses of randomized two group experiments with binary outcomes as it provided a likelihood-based approach (the author was the statistician on the paper and wrote the statistical appendix for it). First, the logistic regression is set up to include an indicator term for study, a term for treatment group, and an interaction term (treatment by study). The indicator term for study allows a separate baseline estimate for each study so that each study's treatment effect estimate contribution is relative to its own control group. The treatment group term allows for a common treatment effect estimate and the interaction term allows for a separate treatment effect estimate for each individual study (the same as one would get using each study's data alone). The consistency of study results is then quantitatively analyzed by investigating the variation in the individual study treatment effect estimates and their confidence intervals and, less preferably, the statistical significance of omitting the interaction term in the logistic regression. A warning about the low power of this test was given along with a suggestion that clinical judgement was preferable. With the omission of the interaction term, a common "pooled" treatment effect is constructed along with estimates and likelihood ratio-based confidence intervals and tests. The likelihood for the confidence intervals for the common treatment parameter τ is obtained by profiling out the within study baseline parameters c_i

$$L(y_1, ..., y_n; \tau, \hat{c}_1, ..., \hat{c}_n)$$

which is of course equal to

$$\prod_i L(y_i; \tau, \hat{c}_i)$$

as the $\hat{c}_i, s$ are mutually independent. Thus it was equivalent to the approach in this thesis, but with the marginal likelihood being immediately given by sufficiency and random effects neglected.

Random effects were later allowed for in a technical report[87] using a method from Cox and Snell[29] that Venables and Ripley claim was first suggested by Finney in 1971[120] and is now often referred to as quasi-likelihood - where the scale parameter, rather than being set equal to

one, is estimated by the deviance or Pearson Chi-square statistic divided by the residual degrees of freedom. Quasi-likelihood though, is a much more general approach, not tied to specific estimates of scale. To second order, this scale estimate has the effect of simply increasing the standard error of the MLE as the MLE itself is unaffected. As reviewed in appendix E, Tjur gave reasons for preferring that the MLE be unaffected, which McCullagh was then easily able to set aside. Many authors though, simply reject this allowance for random effects by scale estimation as being *ad hoc*. Stafford's adjustment[111] was adopted earlier in this thesis, as it provides an asymptotic rationale for the allowance for random effects which may overcome such objections to its use. But, unless the likelihoods are essentially quadratic, as is usually the case with binary outcomes, it is unlikely to modify the fixed effect likelihoods to adequately approximate possibly true level 2 likelihoods.

In Statistics in Medicine in 1986[94], Richard Peto provided an explanation for a statistical method he had used in earlier applications. For ruling out the null hypothesis of no effect, he had used a test based on the unweighted sum of observed minus expecteds Oi – Ei, and for combined estimation of an odds ratio, he had used a weighted sum of Oi – Ei with the weights being the inverse variance of Oi – Ei. These quantities could be directly motivated as being quadratic approximations to maximum likelihood estimation under a conditional logistic regression model, as for instance was shown in Cox[23] and referenced by Peto[123] in 1985. Of course there is always more than one way to motivate a quantity – it is just suggesting this is one possible way.

In 1986, Peto emphasized entirely different justification of the use of Oi – Ei by starting with the question "But why use observed minus expecteds rather than some logistic model?" His answer had two parts – one was that observed minus expecteds would be readily understandable to physicians and that it provided a typical estimate of odds ratios that did not depend on assumptions of the sort needed for logistic regression (although this does follow from assuming a conditional logistic regression model and approximating the conditional MLE by the score statistic – see O'Rourke[84]). Unfortunately, he did not define what he meant by "typical" nor the "depend[ence] on assumptions". Perhaps most strikingly, he dismissed the use of random effects models using very similar arguments that Fisher had used for the certain cases where Fisher thought random effects specifically should not be considered – see O'Rourke[84]. It is perhaps more tenuous to relate this Oi – Ei approach back to Fisher and Cochran than the approach of DerSimonian and Laird and L'Abbe, Destky and O'Rourke but more or less indirectly the methods of Fisher and Cochrane became central for the meta-analyses of randomized clinical trials.

The pressure for clinical researchers to actually carry out meta-analysis of randomized controlled trials in their various fields had been building perhaps soon after Archie Cochrane published an essay in 1979, in which he suggested that "It is surely a great criticism of our profession that we have not organized a critical summary, by speciality or subspecialty, adapted periodically, of all relevant randomized controlled trials" . In 1985, an international collaboration to prepare systematic reviews of controlled trials in the field of pregnancy and childbirth, resulting in the publication in 1989 of: "Effective Care in Pregnancy and Childbirth (ECPC): A Guide to Effective Care in Pregnancy and Childbirth (GECPC)", and "The Oxford Database of Perinatal Trials (ODPT)". Encouraged by the reception given to the systematic reviews of care during pregnancy and childbirth, Michael Peckham, first Director of Research & Development in the British National Health Service, approved funding for "a Cochrane Centre" to facilitate the preparation of systematic reviews of randomized controlled trials of health care, in 1992. Later that year, "The Cochrane Centre" opened in Oxford, UK. In 1993, an international and comprehensive concept of the Cochrane Collaboration was presented at a conference ("Doing more Good than Harm") organized by Kenneth Warren and Frederic Mosteller at the New York Academy of Sciences, and in June of that year the development of Cochrane Collaboration's Handbook as a tangible means to facilitate the preparation of systematic reviews of randomized controlled trials of health care began with the arrival of the 1st Cochrane Visiting Fellow at the UK Cochrane Centre.

In 1993, a Cochrane Collaboration Workshop on statistical methods for data synthesis was conducted and a report drafted. The list of participants included D. Altman, P. Armitage, C. Baigent, J. Berlin, M. Bracken, R. Collins, K. Dickersin, D. Elbourne, R. Gray, K. McPherson, A. Oxman, M. Palmer, R. Peto, S. Pocock, K. Schulz and S. Thompson, all of whom were statisticians, epidemiologists or physicians with expertise in statistical methods for data synthesis. The workshop was convened to develop guidelines on statistical methods for data synthesis for the Cochrane Collaboration's eventual handbook and to identify useful research topics in that area.

In the report, the deliberations are outlined and a set of implications for the Cochrane Collaboration are given. It was assumed that only published summary statistics would be available for the foreseeable future, although the preferability of having individual participant data was indicated. Issues of inclusion criteria for systematic reviews were not considered except for those having to do with methodological quality. There was a major discussion on effect measures with greatest emphasis on binary outcomes where the relative merits of odds ratio versus relative risk were dis-

cussed at length. Here the odds ratio was favoured as a default, but it was stated that the relative risk and risk difference should not be ruled out as options. Some felt the choice of effect measure should depend in some part on a test of heterogeneity, while others disagreed. Several participants felt it would be preferable to use different measures for presentation than were used for analyzing the data. Continuous outcome measures received much less attention with weighted mean differences being suggested as appropriate, along with the possible consideration of standardizing by the control group standard deviation (to get an "effect size"). Most felt the area merited deeper study - difficulties being anticipated about choice of effect measure, the issue of data distribution, use of medians rather than means, handling of before and after measurements, weighing of studies and missing data. Further research on these was recommended. The issue of binary and continuous data also arose with some suggestion of automatic transformation of continuous outcome to binary, but further study was recommended. Here some of the issues now resolved by this thesis were being identified and highlighted 20 years ago.

As for approaches to aggregation, many but not all, recommended the use of a test of heterogeneity with the issue of low power being identified as a concern along with a suggested Type I error level of .10 rather than the customary .05. As for aggregation, given the determination that "substantial" heterogeneity is not present, after some discussion and a suggestion that results would be similar, a fixed effect models was decide upon as the default approach. Considerable disagreement ensued, however, when the discussion turned to the preferred approach under conditions of statistically demonstrable heterogeneity. Both random and fixed effect models had strong proponents. The report cautions that characterizing in a few words the differences between fixed and random effects proponents would be challenging.

Some claimed the fixed effect approach was "assumption free" and is not [should not be] directly influenced by heterogeneity while others claimed that it would produce an artificially narrow confidence interval, as it does not reflect between-trial variance. They suggested random effects did not make as stringent an assumption as there being no differences between the underlying true treatment effects in the individual trials and hence was preferable. Common ground under these widely contrasting views was then summarized : the analyst should attempt to explore the reasons for the heterogeneity and explain it, especially with regard to varying methodological quality, that the ruling out of an overall null hypothesis of no effect in all trials need not distinguish the alternative to be fixed or random but "at least one of the trials" has an effect; that whether heterogeneity

was present or not, the fixed effect estimate is an informative average measure of treatment effect; and, finally, that as random effects methods have rather amorphous assumptions, it was an area requiring more research into the importance of the assumptions and robustness to them. Here, the pragmatic concern arose regarding random effects methods giving relatively more weight to smaller studies when these often are of poorer quality and more subject to publication bias.

The entire discussion regarding appropriate approaches for aggregation under conditions of heterogeneity pertained to binary data. The same general principles, however, were thought to apply to continuous data and it was mentioned that the same discussion about fixed versus random effects models had occurred many years ago, relative to continuous data. They felt they should acknowledge that various fixed and random effects approaches were available and that future research should compare DerSimonian and Laird's approach to those based on maximum likelihood methods.

This thesis provides a general approach for both discrete and continuous data, regardless of reported summaries, based on the observed summary likelihood. Additionally, DerSimonian and Laird's approach can be compared to likelihood methods using numerous assumed distributions for random effects. It is a bit surprising that it has taken 20 years for this to be undertaken.

4 Background for randomized clinical trials

4.1 Statistics as the combination of observations

In this chapter, apparent differences between statistical analyses of single studies and multiple studies are sketched out. A brief overview of issues that arise given the ever present possibility of incompletely and/or selectively reported studies that arise in the current clinical research context are then given. Recent scandals of extremely high false positive publication rates in the genetics literature - that lead at least one journal to threaten to refuse to publish new results until they were replicated [comment, John Ioannidis] - underline the importance of this material. A "commonly accepted" statistical approach for dealing with multiple studies is then briefly sketched.

Statistics is often presented as providing a mathematical approach to address variation in observations that arise under identical or similar conditions. Something is believed to be, or pragmatically taken as being common, in all the observations even though they do actually differ from each other. As stated in the introduction, this commonness is made explicit in parametric

likelihood by parameters repeating in the multiple of the observation likelihoods. Observations are conceptualized as being generated from probability models and if a parameter repeats (in some chosen re-parameterization), then there was replication of that parameter (assuming the model is true).

Certainly, the concept of something being common in observations even when they vary is central in statistics - especially if the observations are taken under apparently identical conditions. The appropriate extension of this concept to something being common in slightly or even largely different situations or contexts is admittedly more challenging. Parameters being common in distribution is an even further extension to treating admittedly different things as though they are exchangeable or common in some probabilistic sense. Here the interest could be primarily on what is common in the probability distribution that generated the lower level probability distributions, or something about a particular lower level distribution. Again, it is the first that is usually of primary interest in clinical research[15].

On the other hand, a working assumption of commonness arguably underlies the use of any statistical analysis of observations, and may fail to hold no matter how similar the conditions of observation were contrived to be. The conditions under which replication should be expected and whether it actually happened (both the conditions for it and observed commonness) are far from trivial. Replication of studies should be the norm in scientific studies (recall the current concern in the genetics literature). Studies conducted under identical or similar conditions should be expected to have something in common. The conditions for, and assessment of, the actual achievement of replication of something common here are likely to be much more difficult – especially if what is conceptualized as common is just something about a "postulated" higher level distribution. Perhaps the need to stress both investigation and synthesis, not just synthesis, cannot be overemphasized in any area of statistics. It is important to perhaps keep model fit separate from commonness of a parameter - i.e. a parameter may be common, but the distribution form (a nuisance parameter) may be mis-specified with light tails making the truly common parameters appear non-common.

4.2 Meta-analysis or systematic review

Meta-analysis or *systematic review* are common terms used to refer to the quantitative and qualitative analysis of multiple studies. Perhaps less common terms are research synthesis, overviews, pooling and (periodic) scientific audits. Whatever the term, it is ideally a full investigation and

synthesis of what is possibly common in different studies utilizing appropriate quantitative and qualitative analysis. As an aside, some object to the term meta-analysis as being somewhat pretentious. As Aronson[4] concluded from a linguistic analysis of the prefix meta -

> "meta-analysis is an analysis of analyses, in which sets of previously published (or unpublished) data are themselves subjected as a whole to further analysis. In this statistical sense it [the term meta-analysis] was first used in the 1970s by GV Glass (Educ Res 1976;3(Nov):2)."

The analysis of the results of trials not the analyses actually carried out in the trial publications per se, is the real concern here. The analysis used in a given trial may or may not be informative as to the results, but it is the results that are of primary concern, not the analysis perhaps indiscriminately used. A term more reflective of the refereeing or auditing of the results of a collection of studies may have been preferable[86].

The further role of meta-analysis in encouraging better individual studies and providing feedback was also emphasized in O'Rourke and Detsky[86] where it was argued that meta-analysis was just part of being scientific by critically evaluating all the available supposed evidence on a question, and that a meta-analysis which concluded that there was no evidence was not to be thought as less of a meta-analysis. It was further argued that the desire to "somehow" pool the data, even if it was questionable, needed to be guarded against.

To address this haunting concern about not being able to refrain from somehow combining studies no matter how unlikely it was that something was common, the term systematic review is often taken to mean, especially in Europe, a qualitative investigation and synthesis perhaps without any quantitative investigation (analysis of the replication) but certainly without any quantitative synthesis. Meta-analysis is then taken to be just the quantitative combining of estimates from studies (perhaps conceived by some as being largely mechanical). It is debatable whether quantitative techniques for investigating differences in estimates from studies (i.e. analysis of the replication) fall under the term systematic review or meta-analysis (Iain Chalmers, editor of The James Lind Library, private discussion). The choice to not actually carry out a quantitative combining of estimates may in itself be based on both quantitative and qualitative considerations and as long as this is borne in mind, the interchange of terms systematic review/meta-analysis should not be of great concern.

What is meant by different experiments has yet to be defined and a broad definition of this may be that the experimental observations simply involved different contexts. This definition would

tentatively include, rather than exclude, multi-centre clinical trials as (at least potentially) a systematic review/meta-analysis. Apparently, many think of multi-centered trials as being very different from meta-analysis and especially vice-versa, though not everyone would agree with this[105]. Some might argue that adherence to a common protocol prevented the contexts from being importantly different. Here the motivating concern is that if the context is importantly different, some things may very well have changed, but there may still be something common that can be "estimated" in both contexts and productively pooled. It would be a mistake to not make allowance for what may have changed but also to disregard what may have remained common and in what sense. What exactly is meant by context is unavoidably left somewhat vague.

Some writers have claimed that statisticians usually, or even almost exclusively, deal with "single studies" rather than multiple studies, using terms like "myth of a single study"[17], "cult of the single study"[76] and "single sets of data"[42]. With a more general reading of the literature, exceptions turn out to be so numerous that one wonders how such an impression arose. For instance, one finds the analysis of multiple studies addressed by numerous statisticians in the bibliography of this thesis.

Some statisticians though, may seldom encounter projects that involve more than a single study and may well have some doubts about how to proceed when they first encounter the need to analyze multiple studies. Whether there is a real basis for these doubts depends largely on the question of whether there is anything substantially different from a statistical perspective when the data come from multiple studies. Cox[24] also addressed this question starting off with a concise claim that "the words combination of data encompass the whole of the statistical analysis of data." In elaborating on this, he pointed out that separate sets of data can arise in two distinguishable ways. A single data set may be divided into simpler subsections that are analyzed separately and then combined. Alternatively, the separate sets of data may come from quite distinct investigations. Cox then suggested that the technical statistical problems in these two situations may well be identical with the proviso that assumptions of homogeneity (i.e. questions of whether repeated studies are actually addressing some common entity or underlying process, again the analysis of the replication) may well be more questionable and problematic.

This thesis will attempt to add to Cox's paper, the concept that with respect to any parametric likelihood approach there are always implicit and trivial separate sets of data consisting of the individual observations (which Cox in another paper[25] coined the trivial likelihood factorization)

that have somehow been (perhaps somewhat uncritically) combined in the statistical analysis. Hence statistical analysis is (or should be) always in some sense concerned with the investigation and synthesis of what is common in the individual observations. From the perspective of this thesis, the two papers by Cox referenced above come to be seen as directly related to each other.

4.3 The scientific issues of a haphazard collection of studies

There is a need to be aware of the important scientific and substantive issues of how a haphazard collection of studies - some perhaps censored, some very poorly done and some even mis-reported - arise in practice. Special concerns and issues do arise when there is a (potential) haphazard collection of studies. The need to locate, investigate and synthesize repeated or similar studies surely occurs almost all the time in most areas of clinical research. It would be a rare exception where only one study was ever conducted that addressed a particular question or some aspect of it. Ignoring studies done by others (known or unknown) is a very inadequate way to locate, investigate and synthesize all the available evidence on a given question. Determining that there was indeed only one study is perhaps a necessary step in any complete analysis of any single data set. Selective reporting of studies is not just a problem for meta-analyses or systematic reviews - any particular study in hand may have been through some selection process and hence be biased in some fashion[83].

Adequate ways of locating, investigating and synthesizing all the evidence are surely desirable. This thesis is primarily concerned with the statistical aspects of investigating and synthesizing repeated or similar studies given the summaries available. Admittedly, the statistical aspects may often not be as critical as the qualitative scientific components (i.e. figuring out which studies one should expect to have exactly what in common, figuring out if some have not been published and which were published twice misleadingly as different studies, etc.), but the statistical issues deserve full and proper consideration. For this thesis, the focus is largely on the investigation and synthesis of the extracted summaries in hand - taking them as adequate and correct. In any particular meta-analysis though, these wider issues may be much more important.

Checking the correctness of extracted summaries should be part of a full investigation, that would also involve many other aspects of study design, conduct and publication. A full synthesis may involve qualitative scientific considerations that are not easily dealt with in a statistical framework. An adequate grasp of the actual practice of conducting and publishing clinical studies,

especially the deficiencies, is essential for a full investigation and synthesis of a given set of studies in hand.

The analysis of extracted summaries taken as correct, amongst other things, is comprised of the analysis of the replication of estimates that were expected to be similar in different experiments - i.e. were they consistent? Was the parameter common or just something in a postulated distribution of it? This was indeed stressed in many early papers on meta-analysis of clinical studies (L'Abbe[72], Sacks[103], Greenland[58]) and meta-analysis in general (Cox[24]) and with the combining of prior and sample information inherent in Bayes theorem (Fisher[51], and Sprott[109]). The synthesis of extracted estimates taken as correct, amongst other things, is comprised of the determination of the most appropriate combined estimate, and the quantification of the uncertainty in it, given the accepted commonness of the parameter or the commonness of the distribution of the parameter.

It is perhaps not often enough stressed that the most appropriate synthesis could just be taking what was believed to be the only good estimate available on its own or refusing to accept any of the estimates at all. There should be no commitment to get some combined estimate against good judgement or to get just one combination. Again, the ideal is a full investigation and synthesis of the studies with regard to what was expected to be common among them – was it common in what sense and if so what inference is best - given all available knowledge of the studies whether published or not.

4.4 Additional issues re: common distribution of parameters

Special considerations arise when it is the distribution of parameters that is common rather than the parameters themselves and these special considerations were not fully discussed when we dealt with observations. But with studies, it is important to distinguish what it was that lead to "loss of commonness" of the parameter that was "hoped" to be common. It could very well be that the true treatment effect does vary amongst RCTs due to variation in patient characteristics, for instance. Or there may have been "slight breakage" of the individual RCTs – for instance when some confounding arises that would necessitate inclusion of appropriate confounding parameters. For example, some RCTs may have been in varying degrees somewhat lax on following up patients who were non-compliant on treatment and the biases in this informative loss of patient follow up that varied by study may give the impression that the treatment effect haphazardly varied amongst the RCTs. In the first case, *true treatment effect variation*, these variations in observed treatment

effects corresponds to a biological treatment interaction and there may or may not be an interest in estimating it or ignoring it for the patients studied so far - the question of representativeness of the observed interaction should immediately come to mind[49].

On the other hand, if the variations in observed treatment effects correspond solely to methodological flaws in the RCTs - at least for inferences to patient populations of interest - they are an unfortunate nuisance, for which some allowance has to be made. It would be very unusual to want to know about a specifically confounded effect from a population of similarly flawed studies, at least in the absence of a known true treatment effect. In applications, the two situations are likely to be inseparable and it would seem prudent to treat the haphazard variation as a nuisance before taking it as evidence of something of interest.

Suggested random effects models to deal with variation in treatment effect seldom if ever differentiate these causes - they are simply suggested as a method for dealing with either. For instance, this lead Simon Thompson to propose, at a Methods in Meta-analysis (MIM) meeting at the Royal Statistical Society, that observed variation between trials possibly exceeding that due to chance should be interpreted as real variation in treatment effects and for purposes of the generalization of this to the practice setting, some quantification of the distribution of these varying treatment effects should be strongly encouraged. The author objected on the grounds that it would be unusual for the variation to be largely due to true treatment variation but rather a mix of methodological variation (varying biases) and true treatment variation. Following the MIM meeting, the author drafted the initial writing of a revised entry for the Cochrane Handbook so that observed variation between trials exceeding that due to chance would NOT immediately be interpreted as real and worth trying to generalize -

> "Clinical variation will lead to heterogeneity if the treatment effect is affected by the factors that vary across studies – most obviously, the specific interventions or patient characteristics. In other words, the true treatment effect will be different in different studies. Differences between trials in terms of methodological factors, such as use of blinding and concealment of allocation, or if there are differences between trials in the way the outcomes are defined and measured, may be expected to lead to differences in the observed treatment effects. Significant statistical heterogeneity arising from methodological diversity or differences in outcome assessments suggests that the stud-

ies are not all estimating the same quantity, but does not necessarily suggest that the true treatment effect varies. In particular, heterogeneity associated solely with methodological diversity would indicate the studies suffer from different degrees of bias".

For true treatment effect variation, a distribution of treatment effects of some sort may make sense along with considerations of what are good descriptive summary measures of this. Good description summary measures are only straightforward for symmetric distributions but as symmetric distributions are almost always assumed [J. N. K. Rao, response to suggestion of asymmetric distributions for random effects made by the author at an Edmonton Conference in 2000] discussions about this issue seldom arise. For the case where confounding has displaced a common treatment effect, the random effects model usually represents a common symmetric distribution of an overall confounding parameter that has expectation zero to represent the "hope" that this overall confounding will fortuitously cancel out. The location or centre of the level 2 distribution is then interpreted as a useful estimate of the unconfounded treatment effect. Background considerations would surely suggest an asymmetric distribution for confounding where the confounding is not expected to fortuitously cancel out.

The sensitivity of distributional assumptions about random effects is often investigated using families of symmetric distributions and concentrates on estimation of the location (usually the expectation) of the level 2 distribution - but again non-symmetric distributions are the more realistic threat. These can easily arise if biases from faulty trials are more likely to favour a specific treatment arm, or from there being a mixture of two or more underlying treatment effects. Again, even choosing a summary for a non-symmetric distribution is problematic, let alone the challenge of estimating skewness with a small number of studies. An illuminating example of this was provided by Efron[40].

This haphazard variation - i.e. individual estimates of something hoped to be constant, such as a treatment effect, varying more than expected by chance - could be modelled formally by the random effects model or informally by the "likelihood curvature adjustment" method. For the formal approach, the estimation of parameters of the common distribution is natural. In the informal approach, the combined likelihood is modified so that asymptotically correct variance estimates are available from the usual likelihood quantities. Tjur pointed out the lack of, as he put it, "a meaningful theory of generalized linear models with overdispersion [random effect]"[118]. Unfortunately, in meta-analysis in particular, the number of RCTs is usually fairly small. Because

of this the asymptotics may not be a good guide for the informal approach, and for the formal approach the particular probability distributions for the haphazard variation in treatment effect estimates may give different results, and the appropriateness of the various models would be very difficult to discern amongst - so that complete resolution of the conceptual challenges is quite unlikely, as for instance argued in O'Rourke.[83] Keiding[68] has made similar comments about random effects or frailty modeling in survival analysis.

An important argument in favour of the informal approach is that it avoids a possibly serious bias that would occur when studies are not exchangeable and the change in weights under the formal random effects model versus fixed model is correlated in an unfortunate way with, say, study quality or amount of bias. By avoiding this bias, it may provide an *ad hoc* but very reasonable compromise where the pooled estimate is kept the same as when no random effects were considered and the log-likelihood is widened to allow for the extra uncertainty by the likelihood curvature adjustment. This adjustment is inefficient under fixed effect assumptions but robust under various probability models for random effects and almost efficient when the random effects are small[9]. Arguably, much more important though, it avoids the possibly serious bias that occurs when the random effects weights become correlated with the treatment effect estimates because of study size related quality and p-value censorship.[83]

The informal approach which simply modifies the profile likelihood for treatment effect as for instance suggested by Stafford[111] initially seemed to be the less risky way to proceed and a related technique, based on an approach used by Fisher, was coined the least wrong random effects model for meta-analysis in O'Rourke[83]. Stafford's approach was found deficient for models with unknown scale parameters even when the fixed effect MLE was relevant.

4.5 Comparative experiments with random assignment

It is of great advantage to be able to randomize subjects to treatments when these are to be compared. Randomization prevents very complicated non-commonness from arising between studies – i.e. some between group difference (confounding of comparison groups) that may vary between studies and would need to be represented by between group confounding parameters. These parameters are likely to be very complicated. Comparative trials would have likelihoods such as

$$L\left(study_i; \mu_{ti} = s_i(t) + \tau, \mu_{ci} = s_i(c), \sigma_i\right)$$

where $\mu_{\cdot}$ is a location parameter of the groups, τ a common treatment effect, $s_i(.)$ some process of selecting patients into the comparison groups and σ_i a common within group standard deviation. In randomized studies, subjects are first recruited and then randomized to the groups so $s_i(t) = s_i(c)$, say μ_i and the likelihood really has only three parameters $\{\mu_i, \tau, \sigma_i\}$ (or possibly $\{\mu_i, \tau_i, \sigma_i\}$ if τ is not common). In non-randomized studies, the likelihood would involve $\{s_i(t), \tau, s_i(c), \sigma_i\}$ where $s_i(t) \neq s_i(c)$ and these selection processes most likely depend on an unknown number of parameters, many for which the likelihood would be constant or near constant. In non-randomized studies, it is possible that the differential selection processes between comparison groups is common across studies - i.e. $\{s(t), \tau, s(c), \sigma_i\}$. Here commonness could actually be bad as although $s(t) + \tau - s(c)$ can be better estimated (there is a combination for it), this would need to be corrected for $-s(t) + s(c)$ to get unconfounded estimates of τ and there most likely will be no unconfounded estimates of this.

With randomization there is no need for between group parameters (in addition to the treatment effect parameter and possibly a treatment by study interaction parameter) at least to get unconfounded estimates of τ and this simplifies considerably the investigation and synthesis of what is possibly common about τ. Even with randomization though, one needs to be careful and not use likelihoods or probability distributions for parameters that possibly confound the investigation and synthesis of τ. For instance, using a common probability distribution for the control mean parameters leads to incorrect inference, as the combination of the control means over studies would not in general equal the true control rate even under slight model misspecification. That is, under the usual assumption of a single draw of the random $\mu_{\cdot i} \sim P(\mu)$ for both groups (again patients are first recruited and then randomly assigned to groups) the study likelihoods for known variances would be

$$L\left(study_i; \mu_{ti} = \mu_{ci} + \tau, \mu_{ci} = s_i(c)\right).$$

The proposed strategy is to define a common treatment effect for instance $\mu_{ti} - \mu_{ci}$

$$L\left(study_i; \mu_{ti} - \mu_{ci} = \tau, \mu_{ci}\right).$$

Then the non-common parameters are profiled out within each study

$$L\left(study_i; \mu_{ti} - \widehat{\mu}_{ci(\mu_{ti})}, \widehat{\mu}_{ci(\mu_{ti})}\right).$$

Now $\widehat{\mu}_{ti} - \widehat{\mu}_{ci(\widehat{\mu}_{ti})}$ will equal τ asymptotically and specifically 0 when $\tau = 0$. Alternatively, with μ_{ci} dealt with as a random variable the level 2 likelihood would be

$$= L\left(study_i; \mu_{ti}, \mu_c^* = g(\mu_{c1,\ldots,}\mu_{cn})\right),$$

where $\mu_c^* = g(\mu_{c1,\ldots,}\mu_{cn})$ is some kind of shrinkage estimator appropriate under the particular random effects model. Now here, as discussed earlier, any particular random effects model is highly suspect and μ_c^* unlikely to be even consistent for the true μ_{ci}. Making such suspect assumptions in the hope of gaining some extra efficiencies in estimation - given randomization - is a very poor strategy and is not recommended.

With random assignment, the defining of treatment effects, given likelihoods from two or more groups, just involves transformation of the parameters. Whether or not the treatment effect defined that way is actually common is an empirical question. If it is common, then the other non-common parameters can be profiled out. A closer look at how random assignment facilitates this is now taken.

As the groups are randomized, they are independent and the likelihoods involve parameters relating only to the groups, and not between the groups, and the likelihood for both of them is simply the multiple of the two. This of course, does not imply there is no between group likelihood components. For instance, if we knew the treatment was a placebo, the probability models for both groups would be the same (all parameters would be common) and the groups would be combined by multiplying the likelihoods

$$L\left(group_1; \mu_1 = \mu_2\right) L\left(group_2; \mu_2\right).$$

If there is a treatment effect, then at least some parameter is not common (between the groups) and the multiple of likelihoods will have two terms for that parameter in it, one for each group

$$L\left(group_1; \mu_1\right) L\left(group_2; \mu_2\right).$$

More realistically perhaps

$$L\left(group_1; \mu_1, \sigma\right) L\left(group_2; \mu_2, \sigma\right)$$

or even

$$L\left(group_1; \mu_1, \sigma_1\right) L\left(group_2; \mu_2, \sigma_2\right).$$

These log-likelihoods can be re-expressed by taking any invertible function of them and nothing will be lost – the probability of re-observing exactly the same observations will be the same. Starting with

$$L\left(group_1, group_2; g_1(\mu_1, \mu_2), g_2(\mu_1, \mu_2), \sigma\right),$$

the probability of re-observing the outcome given certain values of the parameters (which defines the log-likelihoods) is not affected by re-expressing those parameters. Certain re-expressions define treatment effects along a single dimension and these may be of interest, and maybe expected to be common from well designed randomized experiments, such as

$$L\left(group_1, group_2; (\mu_1 - \mu_2, \mu_2), \sigma\right)$$

or perhaps

$$L\left(group_1, group_2; (\mu_1/\mu_2, \mu_2), \sigma\right)$$

or even

$$L\left(group_1, group_2; (\log(\mu_1) - \log(\mu_2), \log(\mu_2)), \sigma\right).$$

These are re-expressions of parameters in the probability model and not re-expressions of sample summaries. As before there will be various options in dealing with the non-common parameters, again grouped into non-common treatment effect parameters and non-common within study parameters.

4.6 The popular two-stage approach to meta-analysis

The most popular quantitative methods for dealing with multiple studies in clinical research, as mentioned in the introduction, involve a two-stage process of first determining a "good" summary estimate of something expected to be common from each study, and then determining some "optimal" combination of these summaries, usually based on weighting by inverse variance of the summary estimates. In fact, it was stated in Deeks, Alman and Bradburn that

"Meta-analysis is a two-stage process. In the first stage a summary statistic is calculated

> for each study. ... In the second stage the overall treatment effect is calculated as a weighted average of these summary statistics. ... All commonly used methods of meta-analysis follow these basic principles."[35]

In practice, the results from the more usual two-stage approach are believed to be very similar to those from the explicit likelihood approach suggested here[116] - especially if "informed by" the likelihood approach (i.e. the choice of the "good" summary and "optimal weights", including choice of transformation of scale). However, this experience is largely limited to binary outcomes. The two-stage approach, though, arguably lacks the model base and achievable transparency of the likelihood approach. On theoretical grounds, the two-stage approach can be criticized as being arbitrary, the criticism of arbitrariness perhaps being best put by Fisher's comment with regard to the necessarily arbitrary choice of a "good" summary – "good for what?"[50].

5 Meta-analysis application examples

Here, only well designed and conducted randomized studies will be considered in detail, starting first with random samples from different contexts (single group experiments) and then with random assignment to comparison groups in different contexts -where it is believed something of interest was constant and most likely something of non-interest may have varied. Non-randomized studies involve additional considerations. These additional considerations for non-randomized studies were only discussed briefly. This is not because they are felt to be unimportant or not amenable to likelihood-based approaches, but because they involve considerably more model mis-specification risk that constrains inference largely to sensitivity analysis [21][59] and their use is possibly advisable only with informative Bayesian priors on the model mis-specification[60].

5.1 Computational strategies and tactics

In this section, the computational strategies and tactics used to implement calculations for the examples in this thesis will be briefly outlined. This has involved the writing of tens of thousands of lines of computer code and a hundred plus days of computing. Ideally, one would wish to have a few procedures that would facilitate the meta-analysis of differing examples with the setting of options or minor re-programing. This is now being approached, but much work remains. Each example used similar procedures, but with considerable modifications. For instance, the success of various versions of the global optimizations to obtain profile likelihoods differ by example and can

take hours to run. Originally, all of the programing was done in R (or SPlus) but had to be redone in Mathematica. Mathematica software was a second choice to R - given R's free availability and wide use in the Statistical community - but only Mathematica could handle the global optimizations to get profile log-likelihoods and the computational algebra needed for the envelope numerical integration bounds for level 2 likelihoods. Hence, the Mathematica based approach is outlined here, but again ideally, freely available software would be better. It might be possible to obtain or implement procedures in R to overcome the current limitations. This remains future research.

A list structure was used to represent the reported group summaries such as $g_\cdot = \{mean, sd, n\}$ or $g_\cdot = \{\min, mean, \max, n\}$ within the m multiple studies. For single group studies, the list would be $\{g_1, ..., g_m\}$, for two group studies $\{\{g_{11}, ..., g_{1m}\}, \{g_{21}, ..., g_{2m}\}\}$ and more generally for k group studies $\{\{g_{11}, ..., g_{1m}\}, ..., \{g_{k1}, ..., g_{km}\}\}$. Next, a list of probability models, one for each of these same groups, was defined. For single groups studies, a simple specification could be $\{\Pr[\mu, \sigma_1], ..., \Pr[\mu, \sigma_m]\}$ with Pr being for instance the *Normal* distribution. More generally, the specification could vary much more from study to study as in $\{\Pr_1[\mu, \sigma_1], ..., \Pr_m[\mu, \nu_m, \gamma]\}$. Random effects probability specifications would be nested as in $\{\Pr[\mu_1 \tilde{} \Pr[0, \sigma_b], \sigma_1], ..., \Pr[\mu_m \tilde{} \Pr[0, \sigma_b], \sigma_m]\}$ though one may more easily start with $\{\Pr[\mu_1, \sigma_1], ..., \Pr[\mu_m, \sigma_m]\}$ and leave the specification regarding the commonness in distribution of $\mu_\cdot$ to a later step. Initially, any default or canonical parameterization will suffice. For instance, for two group randomized studies, a common (starting) specification would be $\{\{\Pr[\mu_{11}, \sigma_1], ..., \Pr[\mu_{1m}, \sigma_m]\}, ..., \{\Pr[\mu_{21}, \sigma_1], ..., \Pr[\mu_{2m}, \sigma_m]\}\}$, representing arbitrary control means, arbitrary treatment means and arbitrary but common within study standard deviations.

A list of marginal likelihood approximations for all these groups can then be generated from these two lists, as appropriate. For reported group summaries that are sufficient for their associated probability models, a single sample of the appropriate size n with exactly the same sufficient summaries will provide a fully accurate "recreation" of the original data likelihood - i.e. $\{y_1, ..., y_n\}$. For special closed form marginal likelihoods, a sublist of necessary probability specifications to provide the marginal likelihood directly from the reported summaries can be created - i.e. for reported minimums and maximums $\big\{\{\min, \max, n\}, p(\min)[P(\max) - P(\min)]^{n-2}p(\max)\big\}$. Finally for the general case, a sample of k samples of size n are generated that have approximately the same summaries as the reported summaries and the importance sampling observed summary likelihood approximation formula is applied to these, i.e. $\{\{y_1, ..., y_n\}_1, ..., \{y_1, ..., y_n\}_k\}$, to get the approximate observed

summary likelihood. In order to acquire the needed conditional samples within reasonable computing time, an opportunistic value in the probability models parameter space is "guesstimated" or searched for and then unconditional samples, each of size n, are drawn (using only these parameter values). Only those that are within a given tolerance are kept to get the conditional sample (rejection sampling). This can be done in subsets (randomly or perhaps based on varying tolerances) to give some sense of the accuracies being obtained. Usually tens of thousands of samples are drawn and rejected to meet chosen tolerances but extremely high rejection rates (in the millions) suggests either poorly chosen values in the parameter space or probability model/reported summary conflict. In a given meta-analysis a resulting list such as the following can result, illustrating all three cases of sufficiency, closed form observed summary likelihood and non-sufficient summary $\{\{y_1, ..., y_n\}, .., \left\{\{\min, \max, n\}, p(\min)[P(\max) - P(\min)]^{n-2}p(\max)\right\}, ..., \{\{y_1, ..., y_n\}_1, ..., \{y_1, ..., y_n\}_k\}$.

Reparameterizations are then required to highlight the arbitrariness, commonness or commonness in distribution of the various parameters amongst all the groups. A rewrite of the list of probability models is perhaps most convenient for this. For instance, for the two group randomized study one such rewrite could be

$$\{\{\Pr[\mu_{11}, \sigma_1], ..., \Pr[\mu_{1m}, \sigma_n]\}, .., \{\Pr[\mu_{21}, \sigma_1], ..., \Pr[\mu_{2m}, \sigma_m]\}\}$$

$$\Updownarrow$$

$$\{\{\Pr[\mu_1, \sigma_1], ..., \Pr[\mu_m, \sigma_m]\}, .., \{\Pr[\mu_1 + \delta, \sigma_1], ..., \Pr[\mu_m + \delta, \sigma_m]\}\}$$

for fixed effect and

$$\left\{\{\Pr[\mu_1, \sigma_1], ..., \Pr[\mu_m, \sigma_m]\}, .., \left\{\int \Pr[\mu_1 + \delta_1 \tilde{} \Pr[0, \sigma_b], \sigma_1] d\delta_1, ..., \int \Pr[\mu_m + \delta_m \tilde{} \Pr[0, \sigma_b], \sigma_m] d\delta_m\right\}\right\}$$

for random treatment effects (note only the treatment group has a common in distribution parameter).

Optimizations will then need to be carried out to focus on a common parameter of interest. The needed optimization is succinctly given as another rewriting of the list of probability models.

For instance, for focussing on δ

$$\left\{\left\{\Pr[\widehat{\mu}_{1_{(\delta)}}, \widehat{\sigma}_{1_{(\delta)}}], ..., \Pr[\widehat{\mu}_{m_{(\delta)}}, \widehat{\sigma}_{m_{(\delta)}}]\right\}, .., \left\{\Pr[\widehat{\mu}_{1_{(\delta)}} + \delta, \widehat{\sigma}_{1_{(\delta)}}], ..., \Pr[\widehat{\mu}_{m_{(\delta)}} + \delta, \widehat{\sigma}_{m_{(\delta)}}]\right\}\right\}$$

for fixed effect and similarly for random effects. Fortunately, these optimizations can sometimes be factorized by study. In general, a meta-analysis likelihood from m studies is given as

$$\prod_{i}^{m} L(\gamma(\theta), \gamma(\lambda), \chi(\lambda)_i; y_i)$$

with θ representing the interest parameters, λ the nuisance parameters and $\gamma(.)$, $\chi(.)_i$ isolating the common and non-common parameters. Here there is no parameter based factorization and the full likelihood must be used. If however there are no common nuisance parameters

$$\prod_{i}^{m} L(\gamma(\theta), \chi(\lambda)_i; y_i)$$

the profile likelihood value for a given $\gamma(\theta)^{'}$ becomes factorized as

$$\sup_{\chi(\lambda)_i \in \Omega} \prod_{i}^{m} L(\chi(\lambda)_i; y_i, \gamma(\theta)^{'})$$

and as long as the $\chi(\lambda)_i$ are variation independent components (i.e. $\chi(\lambda)_i \in \Omega_i$ and $\Omega_1 \times \Omega_2 \times ... \times \Omega_n = \Omega$), the profile likelihoods can be obtained separately since

$$\sup_{\chi(\lambda)_i \in \Omega} \prod_{i}^{m} L(\chi(\lambda)_i; y_i, \gamma(\theta)^{'}) = \prod_{i}^{m} \sup_{\chi(\lambda)_i \in \Omega_i} L(\chi(\lambda)_i; y_i, \gamma(\theta)^{'}).$$

This simplifies the required numerical optimization considerably. Unfortunately, the random effects meta-analysis likelihoods have common nuisance parameters, and this simplification is not available - these optimizations need to be carried out jointly over studies or the information lost by ignoring common elements is somehow argued to be unimportant[27]. However, for a given "proposed" profile likelihood path (obtained by a global optimization), re-optimization given the proposed common interest and nuisance parameter estimates will allow a study-wise optimization that should match exactly. This then would provide a less compositionally error-prone check on $\chi(\lambda)_i$, given a certain value of $\gamma(\lambda)$. The study-wise optimization is still error-prone using standard optimization programs and can even fail when the global optimization succeeds. This suggests future work on

identifying the least error-prone approach for low dimensional optimization, perhaps a robust grid based approach[27].

It may always be useful to start with fixed effect assumptions where there are no common nuisance parameters as the sum of the separately optimized log-likelihoods provides a less "optimization error-prone" means of obtaining the pooled log-likelihood. This might also provide good starting values for the joint optimizations required for the random effects model.

Mathematica's minimization routine "FindMinimum" for finding single local minima was soon abandoned for the examples in this thesis, as it frequently failed. Instead, the newer Mathematica function that tries to find global minimums, "NMinimum" was used. This function has a number of implemented methods - "NelderMead", "DifferentialEvolution", "SimmulatedAnnealing" and "RandomSearch". The Mathematica documentation suggested DifferentialEvolution frequently found better minima and this seemed to be the case. Earlier gridding methods written in R will be re-written in Mathematica for future checking of the optimizations (at least when the dimension is moderate). Use of multiple implementations and multiple starting ranges is advisable and was undertaken for the examples. It might be possible to implement similar methods in R (perhaps using "c" routines) and this would remove any reliance on an admittedly largely "blackbox" implementation. In particular, a Differential Evolution Optimization package became available for R in July, 2006.

An intuitive likelihood-based meta-analysis plot was found wanting - raindrop plots had been entertained earlier in the thesis but were found deficient in displaying non-quadratic components in combined likelihoods as well as how the individual log-likelihoods add (or in a sense subtract) to provide the pooled log-likelihood. In some cases, the focus of this will need to shift from a global view to display any possible conflict between the individual log-likelihoods to a local pooled likelihood view, to display how the individual log-likelihoods add and subtract near the pooled likelihood maximum. A plot was developed to do this, taking advantage of the arbitrary additive constant involved in log-likelihoods. First, all of the individual log-likelihoods have constants added to them so that their maximums are 0. These are then added to get the pooled log-likelihood. A constant is then added to this pooled log-likelihood so that its maximum is 2. This provides an approximate 95% pooled likelihood ratio based confidence interval when it intersects with a line drawn at 0. Next, all of the individual log-likelihoods have an equal fraction of the constant that was added to the pooled log-likelihood added to them. With these arbitrary constants the

individual log-likelihoods add exactly to the pooled log-likelihood and, for instance, one can focus on which log-likelihoods add or subtract to give 0 at the two endpoints of the approximate 95% confidence interval. Note that as the likelihood ratio based confidence interval is based on a drop from the maximum, addition of any constants to any of the log-likelihoods would not change this. Once one is satisfied that a pooled confidence is sensible, given how the individual log-likelihoods add and subtract for a given model, the maximum and points at the drop of 2 in the pooled log-likelihood can be extracted. This plot is referred to as an additive support plot. Of course, more formal deviance and other heterogeneity tests remain available.

At first, it seemed impossible to make such a plot for profile log-likelihoods when there was between study information, i.e. where the optimization needed to be undertaken globally over all studies to get the true pooled profile log-likelihood. But if one thinks of profile likelihood as a path or curve through the full multivariate parameter space (running along the crest for given values of the interest parameter), one can simply save these multivariate parameter values from the global optimizations and calculate and plot the individual log-likelihoods along this path. Other possible paths immediately come to mind. For instance, one could set all of the parameter values other than the interest parameter to their joint $MLEs$ and take this as a path. A more interesting path would correspond (at least approximately) to the various meta-analysis methods used in the Cochrane approach. For instance, for the inverse weighted means meta-analysis approach, the path for the common mean would correspond to one with the individual group standard deviations set equal to the sample standard deviations (different by group within a study) with the arbitrary control mean maximized or profiled out. In this way a similar plot can be used to look at the impact of the various approaches. Perhaps almost as important, these other paths can be used as starting values for the multivariate optimizations needed to get the final global profile likelihood path.

Essentially, any probability model can be entertained with these methods. In the examples, distributional assumptions that provided closed form verifications of the various calculations - such as $Beta - Binomial$, $Normal$, $Normal - Normal$ and $LogNormal$ were predominantly used. Numerous other distributional assumptions were tried from time to time, and seemed to provide only housekeeping challenges - i.e. being careful about parameterizations, starting values for optimizations and default methods for numerical integrations. The usual challenges that arise are the loss of sufficiency, the need to simulate the observed summary likelihoods and the loss of closed form level 2 likelihoods. The simulation of the observed summary likelihoods was greatly

facilitated by any contrived means to guess at good parameter values under which to simulate the unconditional samples. These parameter values are arbitrary and the main downside of a poor choice is inefficient simulation (most of the generated samples are discarded) and highly variable integrands. In some cases there may be a conflict between the distributional assumptions and the reported summaries. For instance in one of the examples the reported group summaries from 24 outcomes were a minimum of .4, a median of 6.5 and a maximum of 150. These are in conflict with assumptions of Normally distributed outcomes and sampling from a $Normal$ with mean set equal to 6.5 and SD set equal to $(150 - .4)/4$ failed to result in one sample with summaries within 75% of the reported ones in a sample of $100,000,000$. On the other hand, in this same example, very close samples were quickly generated using $LogNormal$ assumptions. Such conflicts may not be so easy to untangle from a poor choice of parameter values to sample from, for other distributions. The reported maximum was changed to 35 so that the example could be done under $Normal$ assumptions. The programs to generate these samples needed to be written for the particular mix of reported summaries and the particular underlying distributional assumptions. The original ones written in R needed to run overnight while newer ones in written in Mathematica were much more efficient. Given the number of variations required for the examples, only a few were re-written in Mathematica.

The envelope numerical integrations bounds for level 2 likelihoods also provided some numerical challenges. Given that many numerical integrations are needed to discover and verify the intervals of concavity, these challenges were not unexpected. When the computation time started to become excessive on a Pentium 4 desktop computer, other computing platforms were investigated. The usual machines on the Department of Statistics network running Mathematica on Linux actually took longer. The faster machines such as "blackbird" - a dual AMD 250 64bit processor - was about 50% faster. The cluster of Sun computers at the Ottawa Health Research Institute running under Solaris were actually much slower - the system administrator suggested that Mathematica may not have been optimized for Sun computers, as it was only accessing a single processor. Bounding individual rather than pooled level 2 likelihoods considerably decreased the required computing time. As one needs to evaluate conflict amongst the individual log-likelihoods, these need to be bound in any case and can be carefully added together to bound the pooled log-likelihood[45]. Another issue involves the truncation that is currently needed with the envelope numerical integration bounds. One may need to adjust this to correctly bound the level 2 likelihoods. Perhaps more rigorously,

Study	y	**n**
1	3	20
2	0	17
3	3	37
4	1	80
5	2	78
6	8	84
7	1	37
8	1	79
9	2	213
10	8	121
11	2	35
12	4	272

Table 1: Events observed in similar studies

the envelope numerical integration bounds as currently implemented, truly only bound truncated level 2 likelihoods and one needs to check that the truncation is not too severe. Also, one may need to adjust the arbitrary γ parameter as it needs to be "small enough" and so is not entirely arbitrary. Some further implementational details are given in appendix F.

5.2 Single group examples

5.2.1 Example 5.1 - binary outcomes

The example involves a number of studies that recorded events. As it is confidential research only the numbers are given in Table 1.

A perhaps common but somewhat mistaken approach following Cochrane methods (mentioned earlier) would be to enter the estimates and their standard errors and undertake a fixed effect and random effects inverse variance weighted meta-analysis. This is clearly mistaken for fixed effect because if there is a common proportion, there is a common variance, and the true standard errors would be equal for equal sample sizes. Using the Meta library for R software[95] we would get the following standard graph with "diamonds" representing the pooled fixed and random effects 95% confidence intervals, as shown in Figure 4. The fixed effect 95% confidence interval was $(.0201, .0451)$ and the random effects 95% confidence interval was $(.0190, .0614)$.

An analysis was undertaken using methods from this thesis, first for a common proportion (fixed effect) and then for non-common proportions (random effects) using two different formal random effects models. The two formal random effects models used were the $Beta - Binomial$ [29] and

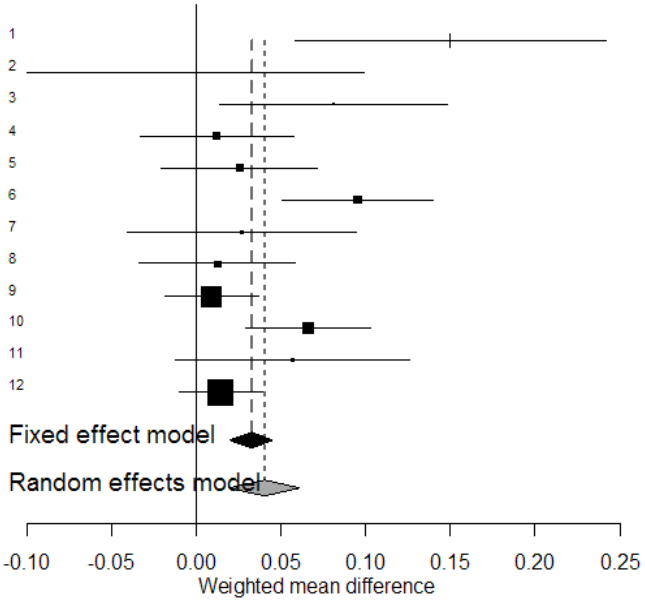

Figure 4: Standard Cochrane Meta-analysis for Example 5.1

$Binomial - Normal$[92]. The $Beta - Binomial$ model is available in closed form and facilitates an exact test of the numerical integration lower and upper bounds. Raindrop plots were used to display the basic features of the individual and combined log-likelihoods in an earlier version of this thesis but were found to not transparently display how individual log-likelihoods added together especially when they were not approximately quadratic. Because of this the additive support plot as described earlier is now used. Again, where the pooled log-likelihood intersects the zero line, provides an approximate 95% confidence interval. A program using a root finding algorithm was written in Mathematica to extract the confidence intervals from a drop of 1.9208 in the log-likelihood. As these intervals are only approximate, directly extracting confidence intervals from these plots, perhaps over a finer range of values, may be adequate in most applications. When confidence intervals are obtained from the root finding algorithm they are given to four decimal places. The individual and pooled log-likelihoods for a common proportion, are given first and shown in Figure 5. The approximate 95% confidence interval would be $(.0231, .0444)$. Note that studies 6 and 1 essentially completely subtract from the interval of support on the left while studies 9 and 12 mostly subtract on the right.

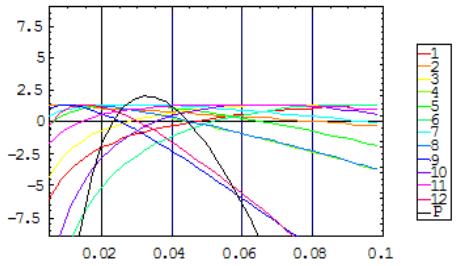

Figure 5: Common Proportion: Individual and Pooled Log-Likelihoods

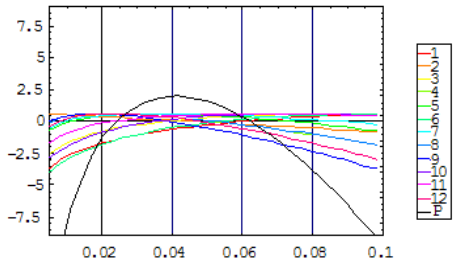

Figure 6: Random Effects Proportion: Individual and Pooled Log-Likelihoods with β set to its value in the joint MLE

When moving to random effects models there are two parameters and the focus is on the unconditional mean of the random effects model. The $Beta - Binomial$ model is dealt with first. The first "path" displayed in Figure 6 takes the nuisance parameter (here β, with the interest parameter being $\frac{\alpha}{\alpha+\beta}$) as set equal to its MLE (found by joint numerical optimization). The approximate 95% confidence interval would be $(.0242, .0616)$.

The next path is the profile path and is shown in Figure 7. The approximate 95% confidence interval would be $(.0237, .0743)$. The points on all the individual profile log-likelihoods shown in the plots were bounded with the envelope numerical integration techniques with a maximum negative gap of -0.000208589 and minimum positive gap of 0.000242802 in approximately 60 minutes.

The example was redone using $Normal-Binomial$ assumptions. The joint profile log-likelihood path was obtained using Adaptive Guassian Quadrature and the individual and pooled profile log-likelihoods plotted in Figure 8. The approximate 95% confidence interval would be $(.0161, .0566)$.

The envelope numerical integration confirmed that these individual log-likelihoods were between the upper and lower bounds. The maximum negative gap was -0.0116158 and minimum positive gap was 0.0121057. Here there are no true known log-likelihoods to bound, but this pro-

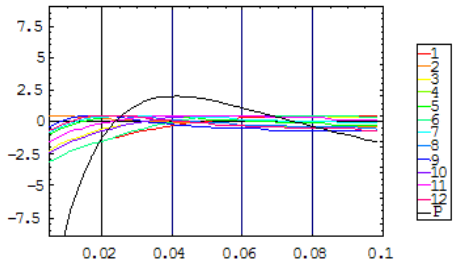

Figure 7: Random Effects Proportion: Beta-Binomial Individual and Pooled Profile Log-Likelihoods

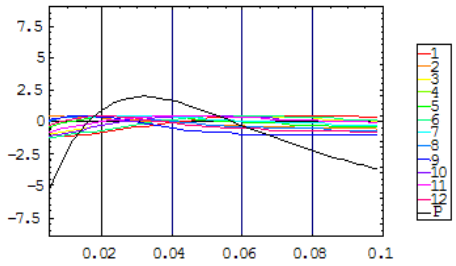

Figure 8: Random Effects Proportion: Normal-Binomial Individual and Pooled Profile Log-Likelihoods

vides numerical assurance that these log-likelihoods are correct - unless the Adaptive Guassian Quadrature and envelope numerical integration have both failed in a manner consistent with this pattern. This was a fairly simple meta-analysis situation - there being only one parameter that is either common or not with two basic choices evaluated for dealing with common in distribution.

In this particular example, the usual Cochrane inverse variance weighted meta-analysis with weights known to be incorrect (as pointed out earlier), provides very similar fixed effect and random effects confidence intervals to the approach of this thesis. Differences do arise in all approaches between the fixed effect and random effects confidence intervals - with respect to the upper limit. The amount, rounded to two decimal places was .04 versus .06 or .07. This may or may not be of practical importance in a given application. If it is, and there is a strong motivation for the fixed effect assumption - such as all studies being a random sample from the same source or population - then the fixed effect confidence interval should be used. More likely, the studies will have been undertaken in slightly different ways and or in different settings and the assumption of one fixed proportion will likely be questionable. Here some random effects assumption would often seem more reasonable. With two different random effects assumptions for this example we obtained

two different upper limits, .06 and .07. Again, this may or may not be of practical importance. If it is, and there is not some background information favoring one of the two assumptions, further analyses using other random effects assumptions or a non-parametric random effects approach may be advisable.

5.2.2 Example 5.2 - continuous outcomes

To investigate whether the parameter that was anticipated to be common was in fact so when there are also non-common parameters, the likelihood for that common parameter needs to be separated to the extent possible from the other parameters. The default strategy is to profile out the non-common parameters – that is for

$$\prod_i L\left(study_i; \alpha, \sigma_i\right) \rightarrow \prod_i L\left(study_i; \alpha, \widehat{\sigma}_{i\alpha}\right) \equiv \prod_i L\left(study_i; \alpha\right)$$

which should be satisfactory if the sample size within the studies is not too small. With small sample sizes problems may arise with this. (And as can be seen from the asymptotics, the difference between methods for separating the common parameter of interest needs to be evaluated with respect to the reasonable number of studies conducted to date and the foreseeable future - small consistent differences can add to a large difference and this will be a practical problem if there are likely to be a large number of studies.) It is therefore desirable to try the simulated modified profile likelihood here. First though, it was done for an example from Pawitan[92] on 11 measurements that were assumed to be Normally distributed with values of $-5.3, -4.5, -1.0, -0.7, 3.7, 3.9, 4.2, 5.5, 6.8, 7.4, 9.3$. Here the modified profile likelihood for both the mean and standard deviation are available in closed form. The following graph in Figure 9 compares two simulated versions to the true modified and simple profile log-likelihoods. Unfortunately, although the approximations were very promising here in this single group single study example, it required $5 - 10,000$ replications. A more informed and comprehensive computational strategy needs to be developed for real meta-analysis examples and is future research.

The example itself involves failure times of air-conditioning units in DC8 jetliners taken from Sprott[109] where the raw data from each "study" is available and various methods can be tried out by summarizing the data in various ways. For instance, the times might be summarized using sufficient statistics (that depend on the choices of distributional assumptions), various order statistics, various linear functions of order statistics, or some mixture of the three. With any of

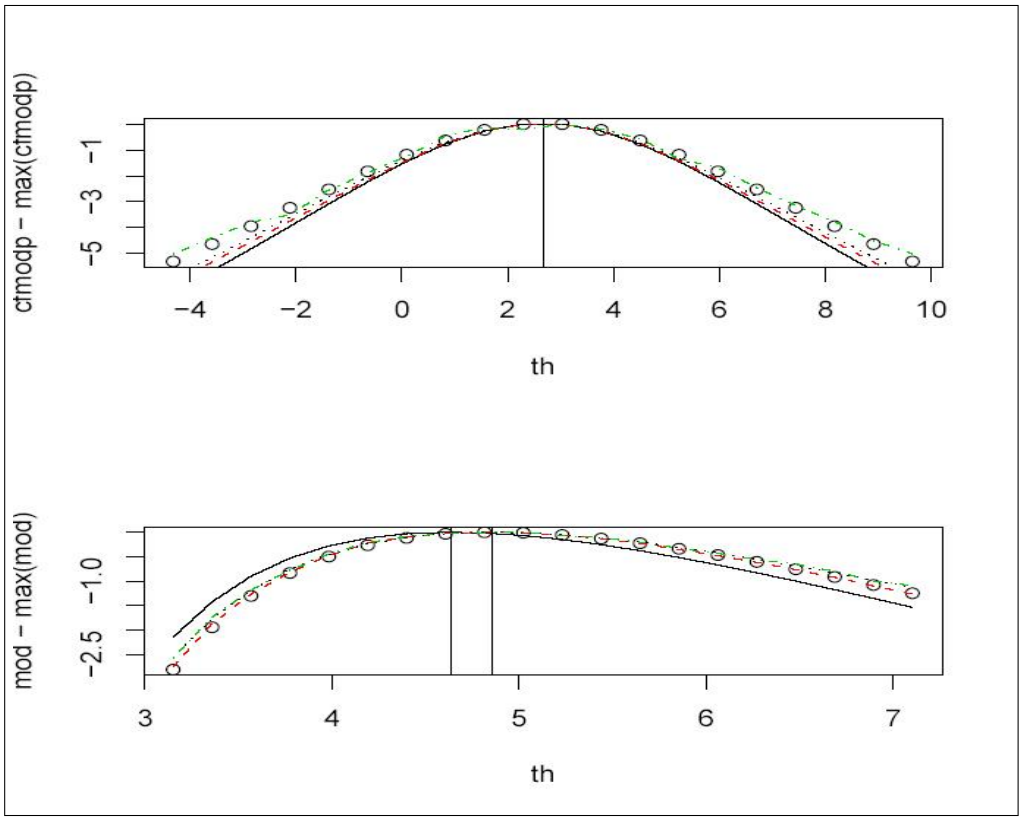

Figure 9: Simulated Modified Profile Log Likelihoods. Modified - "o", Profile - solid line, Simulated - dashed lines.

these summarizations, an appropriate or convenient distribution may be assumed, say *Normal*, *LogNormal*, *Gamma* or *Logistic*. Not all of these 4 * 4 = 16 choices are equally sensible. For instance, with summaries of means and variances, *Normal* assumptions are convenient and perhaps appropriate, with summaries of log means and log variances, *LogNormal* assumptions would be arguably be better, while with order statistics, *Logistic* assumptions would simplify the numerical calculations involving the distribution functions that are available in closed form. Each choice involves a two parameter distribution, and a further choice of parameterization may be required to get commonness of a parameter, focus interest on a parameter or simply for numerical convenience (i.e. in order to avoid constrained optimizations). Differing assumptions involve differing constraints on which parameters can be common while others are arbitrary, for instance with *Gamma* assumptions it is the coefficient of variation not the standard deviation that can be common when the mean is not common over groups/studies[79]. With these specifications there are 7 rather distinct situations, both parameters common, only one parameter common and the other arbitrary or commonly distributed (4) and both parameters arbitrary or commonly distributed (2).

In some of these specifications and summarizations, the raw data likelihood will be directly available in closed form (e.g. *Normal* with mean and variance summaries and common mean and arbitrary variance), for others the observed summary likelihood will be available in closed form (e.g. *Logistic* with order statistic summaries and common scale and arbitrary location), while for others the observed summary likelihood will be available only from an approximation (e.g. *Normal* with order statistic and mean summaries and common mean and arbitrary variance). For specifications with commonly distributed parameters, the level 2 likelihood that is needed will sometimes be available in closed form (e.g. *Normal* with mean and variance summaries and mean parameter Normally distributed and variance parameter arbitrary), while more usually it will only be available via an approximation (e.g. *Normal* with mean and variance summaries and mean parameter Logistically distributed and variance parameter arbitrary).

These choices are opportunities, perhaps sometimes "insurmountable opportunities" and in particular applications they may matter more than one would hope[110]. They represent different views of the data, through different assumptions, obtained more or less directly. In this example, a few that were easy to verify against closed form solutions were chosen. It is anticipated that most, if not all, could eventually be implemented with due care, effort and computational resources. A "package" that easily facilitates the investigation of numerous different views remains future

194	413	90	74	55	23	97	50	359	50	130	487	102
15	14	10	57	320	261	51	44	9	254	493	18	209
41	58	60	48	56	87	11	102	12	5	14	100	14
29	37	186	29	104	7	4	72	270	283		7	57
33	100	61	502	220	120	141	22	603	35		98	54
181	65	49	12	239	14	18	39	3	12		5	32
	9	14	70	47	62	142	3	104			85	67
	169	24	21	246	47	68	15	2			91	59
	447	56	29	176	225	77	197	438			43	134
	184	20	386	182	71	80	188				230	152
	36	79	59	33	246	1	79				3	27
	201	84	27		21	16	88				130	14
	118	44			42	106	46					230
		59			20	206	5					66
		29			5	82	5					61
		118			12	54	36					34
		25			120	31	22					
		156			11	216	139					
		310			3	46	210					
		76			14	111	97					
		26			71	39	30					
		44			11	63	23					
		23			14	18	13					
		62			11	191	14					
					16	18						
					90	163						
					1	24						
					16							
					52							
					95							

Table 2: Failure times of air-conditioning units for 13 DC8 jetliners, one column for each jetliner

research.

The observed failure times for 13 planes are given in Table 2. A standard Cochrane approach given the individual scores would likely be to take the logs (given access to individual outcomes) and then base a meta-analysis on a weighted mean difference analysis of the means and sample standard deviations, as shown in Figure 10. The fixed effect 95% confidence interval was $(3.8196, 4.1297)$ and the random effects 95% confidence interval was $(3.7482, 4.2152)$.

A common choice is to either take logs and use $Normal$ likelihoods or use the outcomes on the original scale and use $LogNormal$ likelihoods. Initially both were done (as a check on the program - as the relative likelihoods should be exactly the same) but $Normal$ likelihoods were chosen so that closed form level 2 $Normal - Normal$ likelihoods would be available as a "test" for the envelope numerical integration lower and upper bounds. In Figures 11 and 12, the Cochrane "path" profile

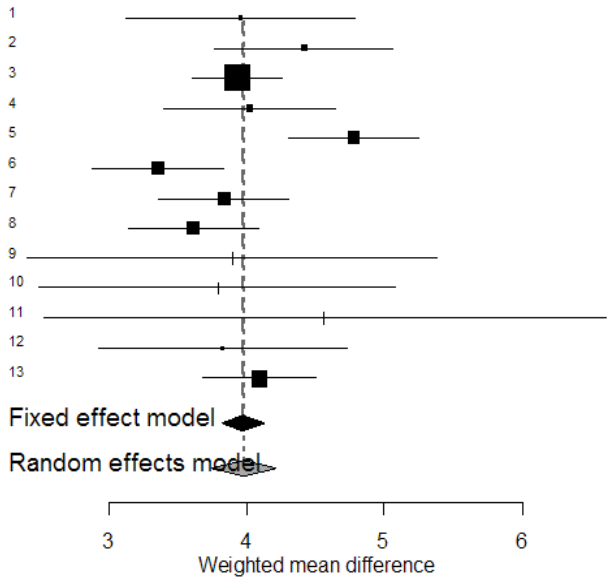

Figure 10: Standard Cochrane Meta-analysis for Example 5.2

likelihoods are given as the first step, where the standard deviations are taken as known and equal to the sample standard deviations. The approximate 95% confidence interval obtained is exactly the same to four decimal places as from the Meta library in R - $(3.8196, 4.1297)$. An alternative approximation to the standard Cochrane approach could also be based on taking the standard deviations as known and equal to their $MLEs$.

Very similar results were seen with studies 5 and 6 being the main "subtracters" from the interval of support. The full profile path, is then given in Figures 13 and 14, where all of the parameters (standard deviations) other than that of interest (mean) are maximized out. The approximate 95% confidence interval would be $(3.7656, 4.1084)$. There is very little change here, except for a much less quadratic log-likelihood for study 11 as seen in the global view.

For the formal random effects model the $Normal - Normal$ random effects model perhaps first used by Cochran in 1937[19] was used. The closed form observed summary likelihood was obtained using algebraic results from C.R. Rao as given in Pawitan[92].

In Figures 15 and 16 the Cochrane path log-likelihoods are given first, where the standard deviations are taken as known and equal to the sample standard deviations and the between study

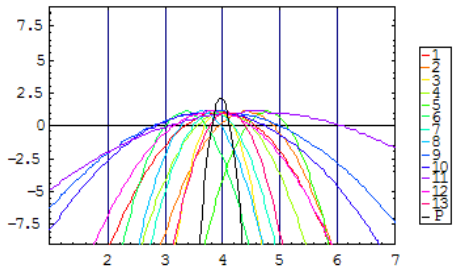

Figure 11: Common Mean: Individual and Pooled "Cochrane Profile" Log-Likelihoods (global view)

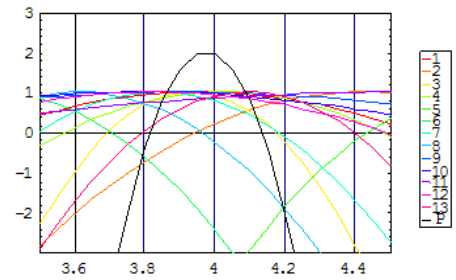

Figure 12: Common Mean: Individual and Pooled "Cochrane Profile" Log-Likelihoods (local view)

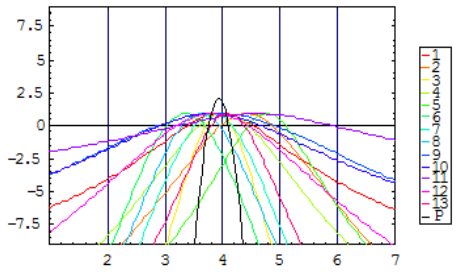

Figure 13: Common Mean: Individual and Pooled Profile Log-Likelihoods (global view)

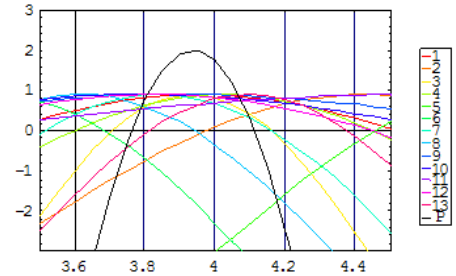

Figure 14: Common Mean: Individual and Pooled Profile Log-Likelihoods (local view)

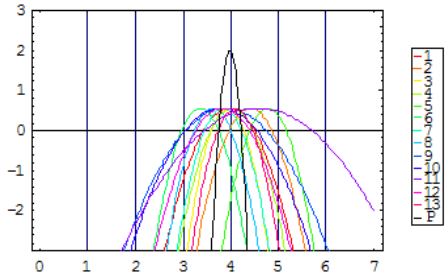

Figure 15: Random Effects Mean: Individual and Pooled Cochrane Profile Log-Likelihoods (global view)

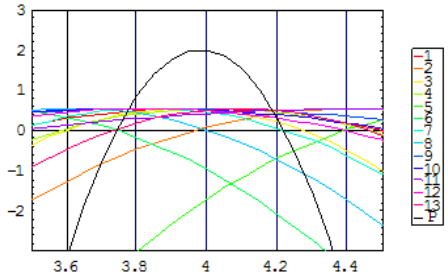

Figure 16: Random Effects Mean: Individual and Pooled Cochrane Profile-Log Likelihoods (local view)

variance taken as known and equal to the DerSimonian-Laird estimate. The approximate 95% confidence interval would be $(3.7481, 4.2153)$.

Then the full profile path is found by optimization over all studies, and parameters and individual and pooled log-likelihoods are plotted along this path, as shown in Figures 17 and 18. The approximate 95% confidence interval was again the same $(3.7328, 4.2312)$.

Here a very different log-likelihood is seen for study 5, but a similar pooled log-likelihood as

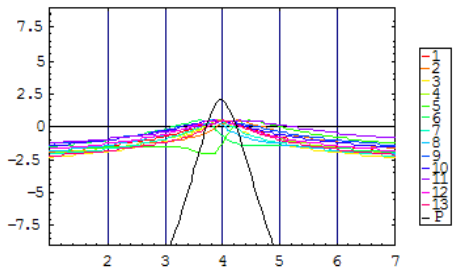

Figure 17: Random Effects Mean: Individual and Pooled Profile Log-Likelihoods (global view)

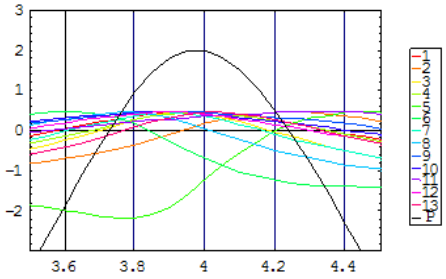

Figure 18: Random Effects Mean: Individual and Pooled Profile Log-Likelihoods (local view)

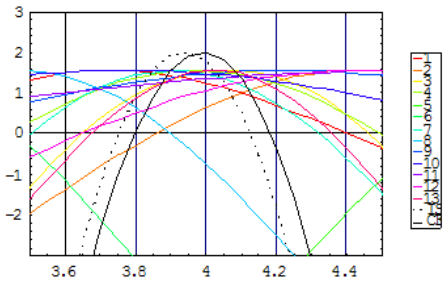

Figure 19: Closed form (cf) versus simulated observed summary (p) pooled log-likelihoods (local view - Common Mean: Individual and Pooled Profile Log-Likelihoods for Min, Med, Max)

in earlier figures. It was then confirmed that the envelope numerical integration lower and upper bounds did bound all the individual (closed form) log-likelihoods along this profile path for a subset of points (52). This took 43 minutes on a Pentium 4 with a maximum negative gap of $-3.37 * 10^{-6}$ and minimum positive gap of $3.85 * 10^{-6}$.

These observations were then summarized by the minimum, median and maximum to investigate methods that would be needed for those summaries. Closed form (involving distribution and density functions) observed summary likelihoods are available for these summaries (exact for odd n, approximate for even n). Monte-Carlo importance samples were then generated to get pseudo data for the simulated observed summary log-likelihoods. Two samples for each study with similar summaries (all within 3%) were obtained in 16 minutes. Study 11 only had 3 observations so these were matched exactly. Accuracy between the simulated and exact observed summary likelihoods was strikingly good for the individual log-likelihoods and the full profile log-likelihood. The pooled profile log-likelihood plots from both along with individual simulated log-likelihoods are shown in Figure 19.

The investigation of the accuracy of the simulated observed summary log-likelihoods in a general

setting remains future research, likely involving importance sampling theory - but here it is very accurate with even just 2 samples. The envelope numerical integration methods were also run on both the closed form observed summary likelihoods and the importance sample observed summary likelihoods. Rather excessive computational time was required for the closed form observed summary likelihoods which involved derivatives of powers of the *Normal* Distribution Function (which grow with the sample size). Currently it is feasible only for small sample sizes. The importance sample observed summary likelihoods did much better (at least for importance samples of size 2), taking about 10 times as long as when the raw data was available. Future research will involve the investigation of more efficient algorithms along with possible parallel computing strategies to make these computations practical for observed summary likelihoods. As for the example itself, at least with knowledge of the individual observations, there were very small differences between fixed effect and random effects confidence intervals regardless of the approach used to construct them.

5.3 Two group randomized examples

5.3.1 Example 5.3 - RCTs with means and standard deviations

Here an example from the Cochrane library on pain relief for neonatal circumcision that had 4 studies, all of which gave group means and sample standard deviations is analyzed. This is the ideal case for the Cochrane approach and considered by many as quite sound even though strong assumptions of approximate Normality and the treating of unknown variances as known underlie the methods. The standard approach is given in Figure 20. The fixed effect 95% confidence interval was $(4.7030, 6.1932)$ and the random effects 95% confidence interval was $(.5616, 8.3243)$.

The starting point is a plot of the profile likelihood along the Cochrane path - separate standard deviations for the groups estimated separately by the sample standard deviations and taken as known. Then the control group means are profiled out. These are shown in Figures 21 and 22. The approximate 95% confidence obtained is exactly the same to four decimal places as from the Meta library in R - $(4.7030, 6.1932)$.

The full profile path is then found by optimizing over all studies and parameters (still allowing separate group standard deviations) and then plotting individual and pooled log-likelihoods along this path, as shown in Figures 23 and 24. The approximate 95% confidence interval would be $(2.0648, 4.0417)$. Note the difference caused by allowing for uncertainty of standard devia-

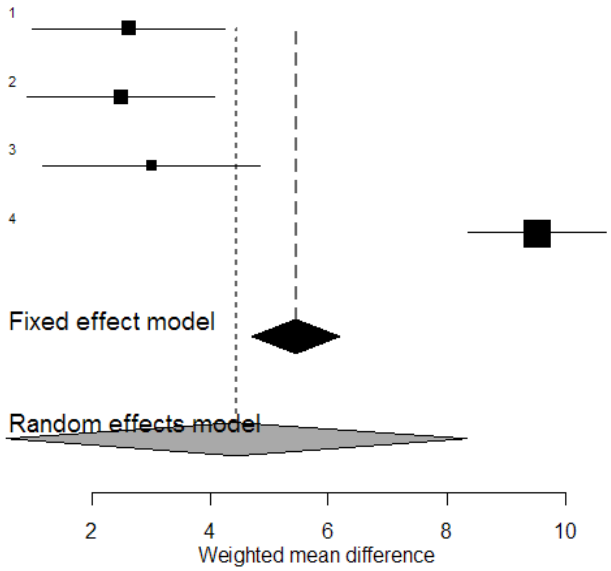

Figure 20: Standard Cochrane Meta-analysis for Example 5.3

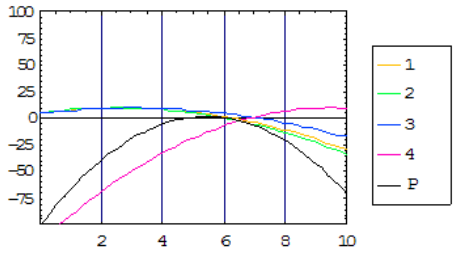

Figure 21: Common Mean: Individual and Pooled Cochrane Profile Log-Likelihoods (global view)

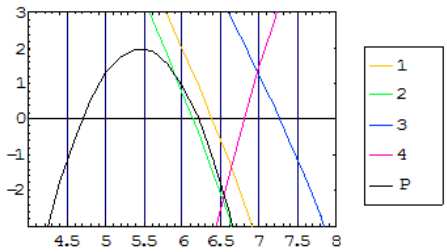

Figure 22: Common Mean: Individual and Pooled Cochrane Profile Log-Likelihoods (local view)

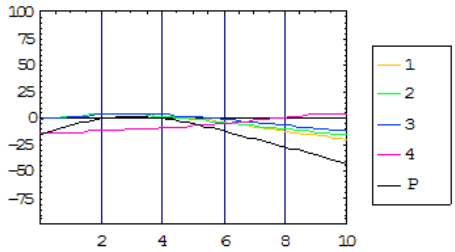

Figure 23: Common Mean: Individual and Pooled Profile Log-Likelihoods (global view)

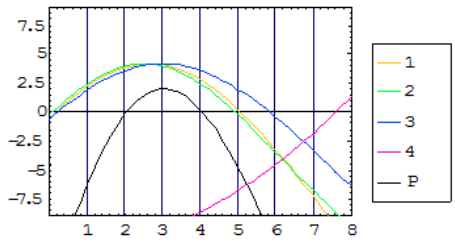

Figure 24: Common Mean: Individual and Pooled Profile Log-Likelihoods (local view)

tions in study 4. This example was also redone using assumptions of equal standard deviations within a study, as shown in Figure 25. The approximate 95% confidence interval would then be $(2.6744, 4.8446)$. $Normal - Normal$ random effects were then done and it is first given along the Cochrane path in Figure 26. The approximate 95% confidence interval would be $(.5984, 8.2593)$. Then it is given along the full profile path (with common standard deviations) in Figure 27. The approximate 95% confidence interval from this would be $(.6999, 8.1749)$. These individual level 2 likelihoods were then bounded by the numerical integration technique for all points (50) along the path with maximum gaps of -0.0561791 and 0.0568516 observed.

In this example, the log-likelihood from study 4 appeared to be quite different from the others,

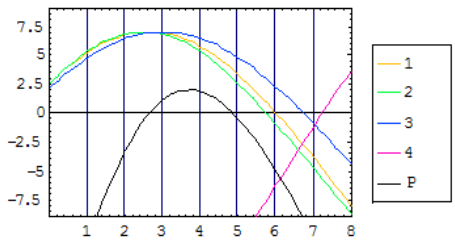

Figure 25: Common Mean and Common Study SD: Individual and Pooled Profile Log-Likelihoods with SDs taken as equal within studies

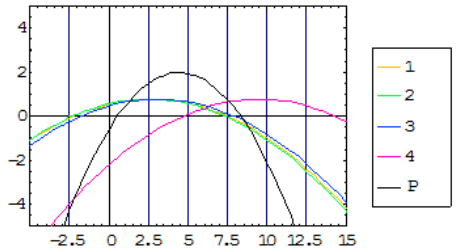

Figure 26: Random Effects Mean: Individual and Pooled Cochrane Profile Log-Likelihoods , non-common SDs within study

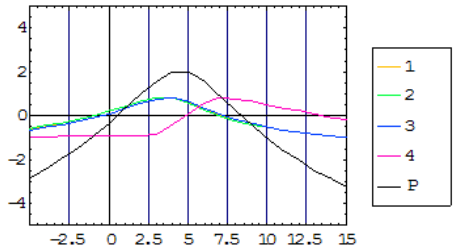

Figure 27: Random Effects Mean: Individual and Pooled Profile Log-Likelihoods, common SD within study

especially under the fixed effect assumption. This lead to a noticeable difference in the fixed effect and random effects confidence intervals, regardless of the approach used to construct them. If there are strong motivations for making the fixed effect assumption or they are simply maintained out of principle, the assumption that the standard deviation is known, will have a large effect. On the other hand, random effects seem more reasonable here, and with that approach, assuming the standard deviation is known has very little effect. A "negative" (expected) effect was ruled out with both fixed and random effects assumptions, though with random effects, the size of the expected effect is much less certain and much more difficult to interpret. As is likely always advisable, anything that appears quite different in a statistical analysis should be subject to further checking and scrutiny. For instance, if study 4 was of higher appraised quality than the others, acceptance of the negative effect as being ruled out is better supported than if it was appraised as being the lowest quality study.

5.3.2 Example 5.4 - RCTs with minimums, medians and maximums

It is not uncommon in the clinical research literature that authors use suspected or observed skewness of outcomes as a factor in their choice of which summary statistics to provide in the

Study	High Dose	Low Dose
Slogoff 1989	N=254, mean=22.8, sd=12.3	N=248, mean=14.7, sd=5.4
Bell 1994	N=19, median=12.96	N=20, median=4.42
Cheng 1996	N=51, mean=18.9, sd=1.4	N=51, mean=4.1, sd=1.1
Myles 1997	N=66, mean=21.5, sd=5.1	N=58, mean=11.4, sd=9.9
Silbert 1998	N=42, median=7.0, range=$(2.1, 19)$	N=38, median=4.0, range=$(0.5, 15.5)$
Michalo. 1998	N=72, mean=11.6, sd=1.3	N=72, mean=7.3, sd=0.7
Sakaida 1998	N=20, mean=14.5, sd=4.5	N=20, mean=5.6, sd=1.6
Berry 1998	N=42, median=12.62, range=$(8.23, 10.67)$	N=43, median=1.83, range=$(0.1, 4.25)$
Myles 2001	N=24, median=9.7, range=$(1.1, 25)$	N=24, median=6.5, range=$(0.4, 35/150)$

Table 3: Reported summaries for studies in Example 5.5

publication of their RCT results, perhaps having been tutored that skewed distributions are more appropriately summarized using median and range values (detailing the minimum and maximum). Unfortunately, the phrase "appropriately summarized" that is often used in statistical textbooks refers to descriptive rather than "inferential" purposes - skewed distributions are poorly described by the mean and variance, but the mean and variance can still provide valuable information about the distribution of outcomes. This can be made fully precise by evaluating the amount of information contained in the observed summary likelihoods.

In this example, some studies reported group means and standard deviations, others just reported group minimums, medians and maximums and one just reported the group medians. The data is given in Table 3.

Using methods suggested to impute means and standard deviations from minimum, median and maximum[66], a standard Cochrane analysis (omitting the study that just reported group medians) can be done, as shown in Figure 28. The fixed effect 95% confidence interval was $(-8.0634, -7.5199)$ and the random effects 95% confidence interval was $(-13.5909, -3.2681)$.

The studies that reported minimums, medians and maximums are dealt with first. The marginal likelihoods for these studies are available in closed form when the number of observations within a group is odd, and by numerical integration when it is even. From David[31] the joint distribution for two or more order statistics is given by

$f(y_1, y_2, ..., y_k) =$

$$cP^{n_1-1}(y_1)p(y_1)[P(y_2) - P(y_1)]^{n_2-n_1-1}p(y_2)...[1 - P(y_1)]^{n-n_k}p(y_k)$$

where $c = \frac{n!}{(n_1-1)!(n_2-n_1-1)!...(n-n_k)!}$, $(1 \leq n_1 < n_2 < ... < n_k \leq n; 1 \leq k \leq n)$ and $y_1 \leq y_2 \leq ... \leq$

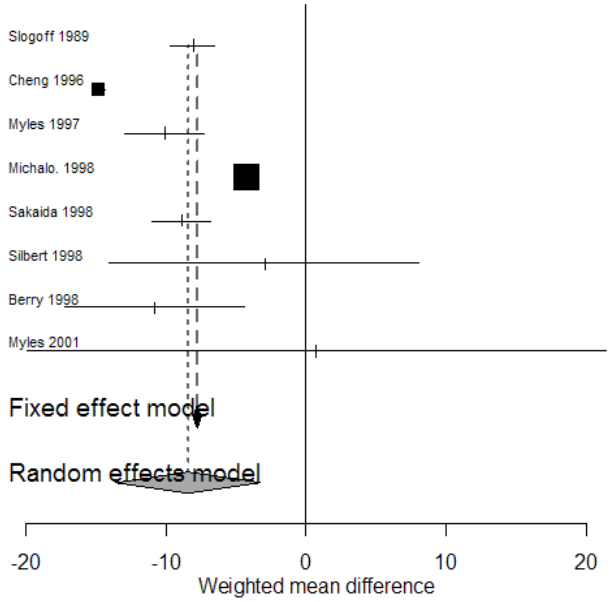

Figure 28: Standard Cochrane Meta-analysis for Example 5.4

y_k. In particular for the minimum, median and maximum with odd n

$f(y_1, y_{n/2+.5}, y_n) =$

$$cp(y_1)[P(y_{n/2+.5}) - P(y_1)]^{n/2-1.5}p(y_{n/2+.5})[P(y_n) - P(y_{n/2+.5})]^{n/2-1.5}p(y_n).$$

For even n median of $y_1 \leq y_2 \leq ... \leq y_k$ is no longer an order statistic but conventionally calculated as $(y_{n/2} + y_{n/2+1})/2$. If the $y_{n/2}$ and $y_{n/2+1}$ had been observed, the joint distribution would be

$f(y_1, y_{n/2}, y_{n/2+1}, y_n) =$

$$cp(y_1)[P(y_{n/2}) - P(y_1)]^{n/2-2}p(y_{n/2})p(y_{n/2+1})[P(y_n) - P(y_{n/2+1})]^{n/2-2}p(y_n).$$

Since $y_{n/2}$ and $y_{n/2+1}$ have not observed, but $(y_{n/2} + y_{n/2+1})/2$ has, one can make the change of variables $y = (y_{n/2} + y_{n/2+1})/2$ and $y_{n/2} = y_{n/2}$ and integrate out $y_{n/2}$. This gives

$f(y_1, y, y_n) =$

$$c \int_{-\infty}^{y} p(y_1)[P(y_{n/2}) - P(y_1)]^{n/2-2} p(y_{n/2}) p(y_{n/2+1})[P(y_n) - P(y_{n/2+1})]^{n/2-2} p(y_n) dy_{n/2}$$

which cannot be integrated in most cases[3] but can be numerically integrated. This also showed that using the same formula with $n-1$ or $n+1$ for even n provided a close approximation. Thus for studies that reported group minimums, medians and maximums essentially exact calculations of the marginal likelihoods are possible. Recall that for simulation of these, the observed maximum for Myles 2001 was changed from 150 to 35 as indicated in the table.

The study that just reported the sample medians also needs to be explicitly dealt with. Here any realistic probability model requires at least three parameters, one for each study mean or location and one for a possibly common variance or scale. (Assumption of a common variance for different groups is standard for RCTs, though perhaps less so for meta-analysis). With *Normal* assumptions and two summaries (one from each group) such as the group medians, it is well known that the likelihood will become unbounded as the MLE for μ_1 and μ_2 equals the respective group medians and the MLE for the variance approaches 0[28]. In such a situation, the likelihood by itself seems incapable of providing information on the variance and hence the "weight" to be given to the study is undetermined, and the study should be set aside in the meta-analysis [unless vague informative priors are used]. It is important to check whether the same obviously wrong scale estimate results from the approach of this thesis. The numerical optimization approach also easily identified an ever-increasing likelihood for decreasing variances at $\hat{\mu}_1$ and $\hat{\mu}_2$ equal to the respective group medians. Here a closed form observed summary likelihood was available. If this is not available, there is an additional problem of finding a good point in the parameter space for the simulations.

For the studies from the same meta-analysis that reported means and standard deviations, under assumption of Normality, the observed summary likelihood given these sufficient summaries is a multiple of the likelihood from any sample that has the same mean and standard deviation. Hence to get the observed summary likelihoods for studies that reported means and standard deviations, observations with the given means and standard deviations are generated, and the usual likelihood results from the use of these observations.

Having obtained likelihoods for each of the studies (except the one that just reported medians) these were then plotted together and combined under the assumption that the difference in means

was common. The plot shown in Figure 29 shows that the log-likelihoods are not quadratic over the range where they are to be combined.

Note that the vertical line in all these plots in Figure 29 is the combined MLE of the quadratic log-likelihoods but that often the profile log-likelihoods are not very quadratic there. This highlights the need for approximations to be good local to the combined maximum and not just the individual maximums - and perhaps, more generally, for a theoretically motivated approach as in this thesis, so that such things do not remain unnoticed in practice. As Copas and Eguchi put it,"it is usual to take these variances as known and to ignore the fact that in practice we use sample estimates".[20] The new plot was designed to help make these issues more apparent. In these "old" plots the direction of treatment effect was defined in the opposite direction to that of the Cochrane software and the new additive support plots.

First the log-likelihoods are plotted for the fixed effect approach on the Cochrane path in Figures 30 and 31. The approximate 95% confidence interval would be $(-8.0352, -7.4919)$. These summary observed log-likelihoods are based on importance samples rather than closed form formulas. The pooled exact and importance sample log-likelihoods are compared using the subset of studies that reported the minimum, median and maximum using a sample size of 2 in Figure 32. Next, in Figures 33 and 34, the log-likelihoods are plotted along the full profile path with common within study standard deviations. The approximate 95% confidence interval would be $(-5.5825, -4.7525)$. Random effects is strongly suggested here and it is first shown on the Cochrane path in Figure 35, where the approximate 95% confidence interval would be $(-13.6490, -3.250)$. It is then shown on the full profile path with common within study standard deviations in Figure 36, where the approximate 95% confidence interval would be $(-11.0569, -4.4922)$. The root finding algorithm took over 50 hours to determine this particular confidence interval.

This example was also done using $LogNormal$ assumptions as these are more "natural" for outcomes constrained to be positive. The full profile path is presented in Figures 37 and 38. The approximate 95% confidence interval (now for the difference in log means), just visually extracted from the graphs, would be $(-.57, -.50)$.

In this example, there is again a noticeable difference in the fixed effect and random effects confidence intervals, regardless of the approach used to construct them. If there are strong motivations for making the fixed effect assumption or they are simply maintained out of principle, the assumption that the standard deviation is known, will again have a large effect in this example -

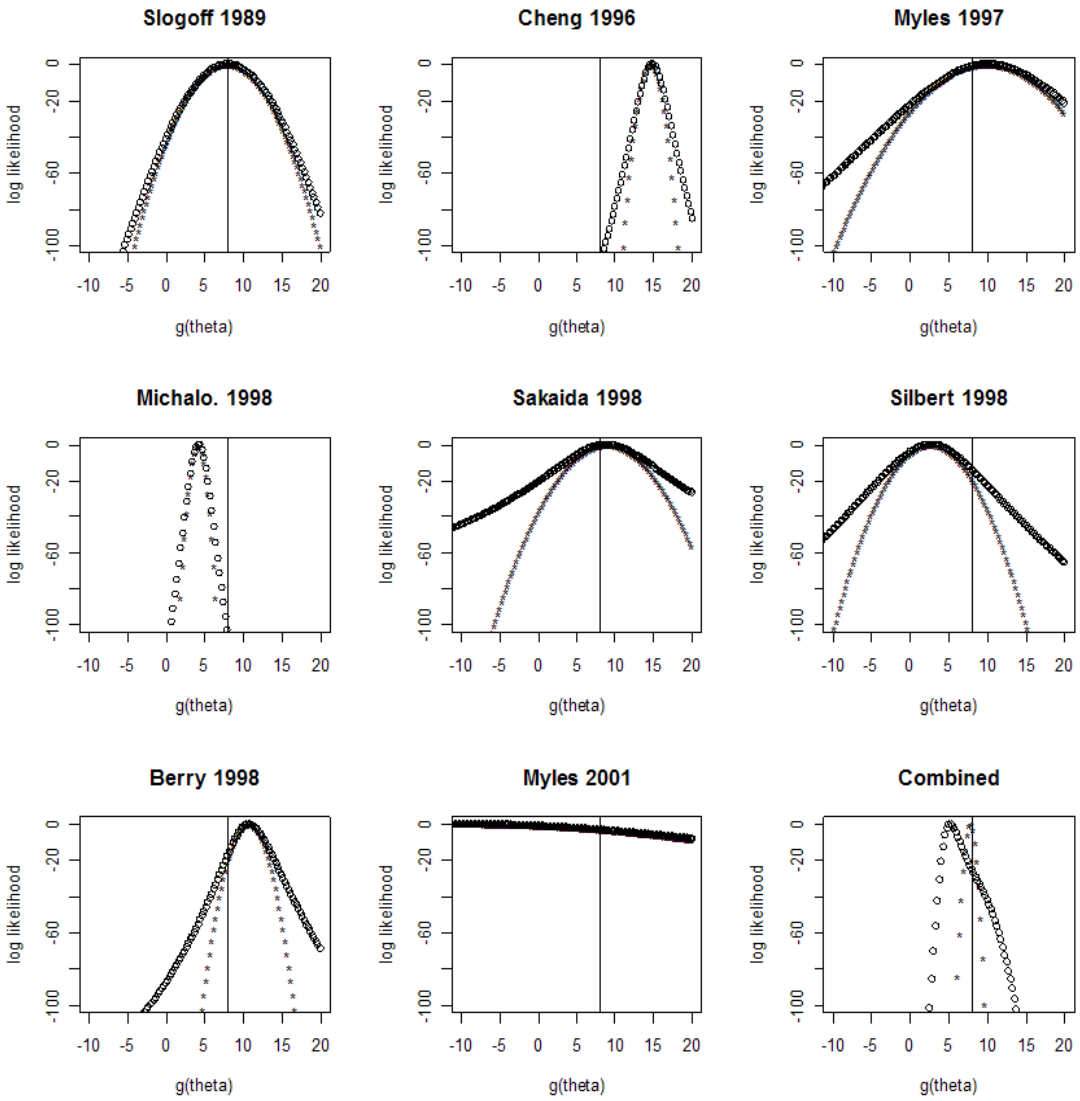

Figure 29: Comparison of observed summary log likelihoods - o - to their quadratic approximations - * - over the full range of $g(\theta)$. Vertical line is the combined quadratic MLE. NOTE - direction of treatment effect is reversed in these plots and observed maximum in Myles 2001 was 150.

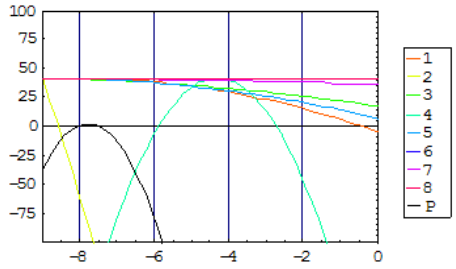

Figure 30: Common Mean: Individual and Pooled Cochrane Profile Log-Likelihoods (global view)

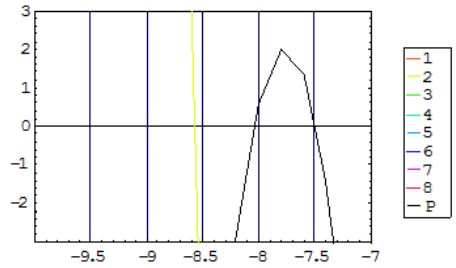

Figure 31: Common Mean: Individual and Pooled Cochrane Profile Log-Likelihoods (local view)

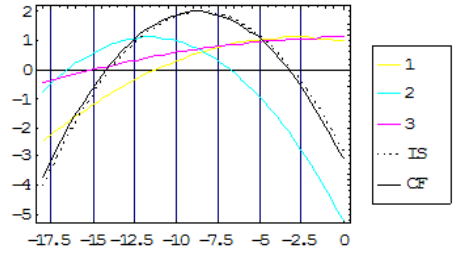

Figure 32: Pooled exact versus importance sample log-likelihoods using sample size 2

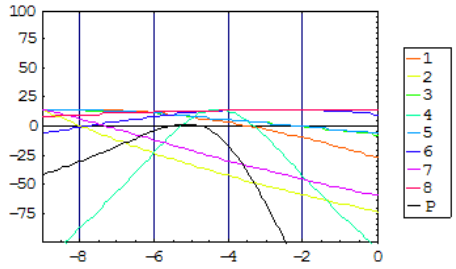

Figure 33: Common Mean and Common Study SD: Individual and Pooled Profile Log-Likelihoods with SDs taken as equal within studies (global view)

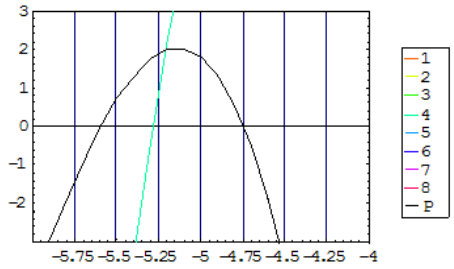

Figure 34: Common Mean and Common Study SD: Individual and Pooled Profile Log-Likelihoods with SDs taken as equal within studies (global view)

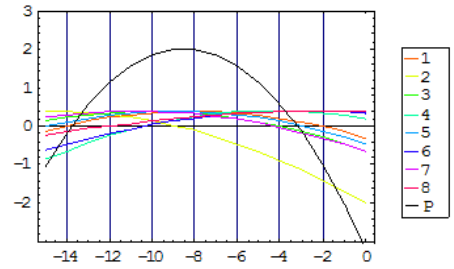

Figure 35: Random Effects Mean: Individual and Pooled Cochrane Profile Log-Likelihoods

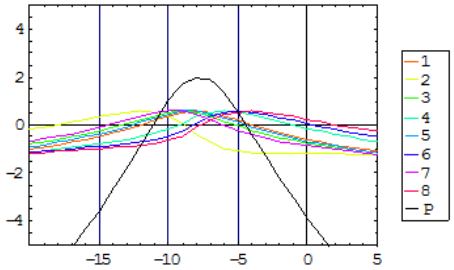

Figure 36: Random Effects Mean: Individual and Pooled Profile Log-Likelihoods

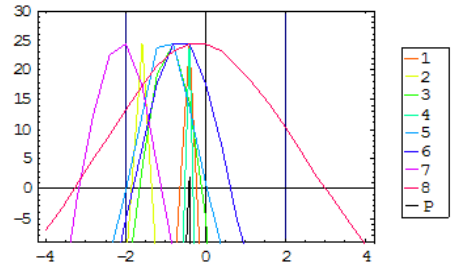

Figure 37: Common LogMean: Individual and Pooled Profile Log-Likelihoods with *LogNormal* Assumptions (global view)

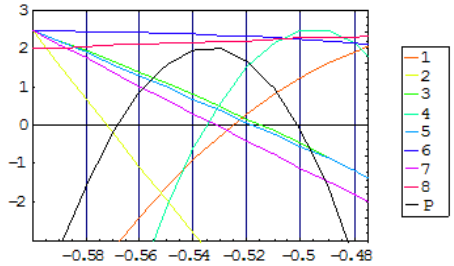

Figure 38: Common LogMean: Individual and Pooled Profile Log-Likelihoods with *LogNormal* Assumptions (local view)

Study	N	Mean	SD	Min	Med	Max	N	Mean	SD	Min	Med	Max
Maekawa 1993	27	1.67	.65	0.9		3.7	25	1.72	.48	1.00		2.80
Meakin 1985	31			1.3	1.7	6.0	14			1.50	1.8	2.40
Splinter 1990	30	1.70	.60	1.3	1.5	4.2	31	1.70	.60	1.20	1.5	4.00
Welborn 1993	41	1.45		1.3		1.6	43	1.41		0.93		1.65
vanderWalt 1986	17			1.5		5.0	13			1.50		7.00

Table 4: Reported summaries for studies in Example 5.5, Rx in right columns, Pl left

the confidence intervals do not even overlap here. If random effects are assumed, the assumption that the standard deviation is known in this example, does still have an effect - it actually narrows the confidence interval. This time a "positive" (expected) effect was ruled out with both fixed and random effects assumptions, though again with random effects, the size of the expected effect is much less certain and much more difficult to interpret. The log-likelihoods under the fixed effect assumption seem to vary considerably, though all but one supported "positive" over "negative" effects. (Recall that Myles 2001 with the actual reported maximum had a negative MLE). Log-likelihoods under *LogNormal* fixed effect assumption were considered as a possible alternative but still seemed to vary considerably. Formal rules for addressing such an apparent lack of replication, as in this example, seem elusive.

5.3.3 Example 5.5 - RCTs with various reported summaries

In another systematic review, some studies reported group minimums, means and maximums, and joint marginal distributions for these summaries are not readily available[3]. To address this, there was no choice but to use the importance sampling observed summary likelihoods. Table 4 gives the summaries of the 5 studies that reported reported various combinations of group means, standard deviations, minimums, medians and maximums.

Simple imputation of means and standard deviations based on available information and rules

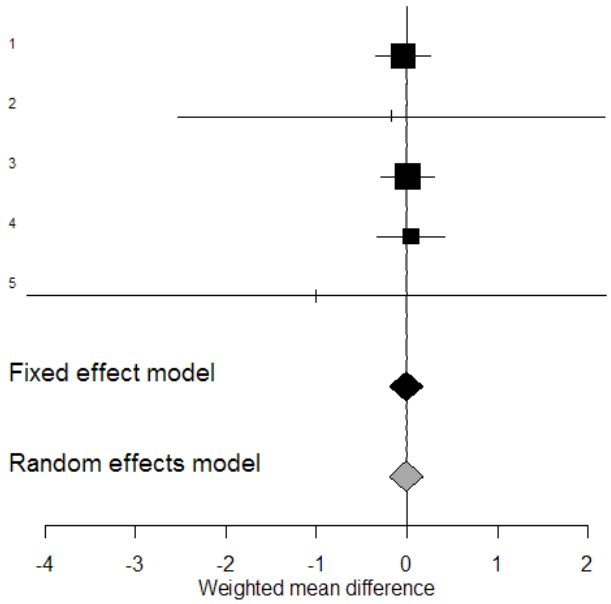

Figure 39: Standard Cochrane Meta-analysis for Example 5.5

from Hozo[66] allows standard meta-analysis methods to be carried out using imputed means and sample standard deviations. This is shown in Figure 39. The fixed effect 95% confidence interval was $(-.2002, .1737)$ and the random effects 95% confidence interval was exactly the same $(-.2002, .1737)$.

First assumptions of Normally distributed outcomes and a fixed effect were used and importance sampling was then used to get the observed summary likelihoods. Under these assumptions, for studies that reported the means and sample standard deviations only a single sample that has exactly that mean and sample standard deviations was needed and there is no need to condition on any other reported summaries.

Again, the log-likelihoods for the fixed effect approach are first plotted on the Cochrane path using an importance samples of size 2 in Figure 40. The approximate 95% confidence interval would be $(-.2027, .1722)$.

Then a plot along a full profile path, with common group standard deviations, is shown in Figure 41. The approximate 95% confidence interval would be $(-.0694, .0939)$.

This example was also done using a $LogNormal$ assumption as this is more "natural" for

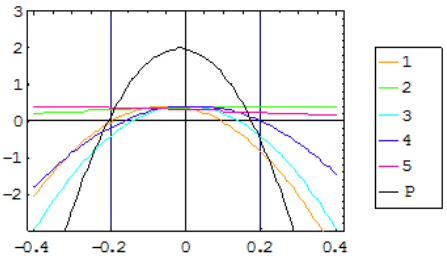

Figure 40: Common Mean: Individual and Pooled Cochrane Profile Log-Likelihoods

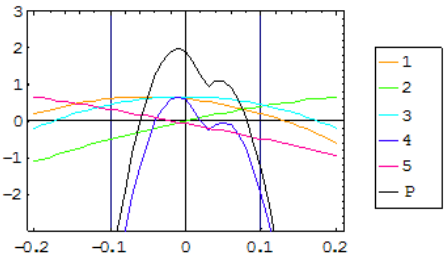

Figure 41: Common Mean: Individual and Pooled Profile Log-Likelihoods, common SDs

outcomes constrained to be positive. Under this assumption, the full profile path is plotted in Figure 42. The approximate 95% confidence interval (now for the difference in log means) would be $(.0176, .1024)$. Note, not only have the log-likelihoods become more quadratic, the pooled log-likelihood provides an approximate 95% confidence interval that excludes the null. Random effects are not suggested from these plots and analyses were only undertaken for $Normal - Normal$ assumptions which confirmed the between study variation was very small, if not zero. Special methods are given in appendix D for obtaining the MLE when the between study variation is very small.

In this example, imputations were required to undertake standard Cochrane analyses. These

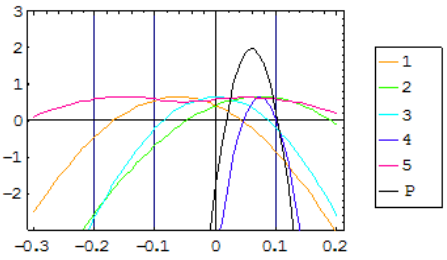

Figure 42: Common LogMean: Individual and Pooled Profile Log Likelihoods

seemed adequate when the standard deviation was taken as known. With the standard Cochrane analyses, there was no noticeable difference in the fixed effect and random effects confidence intervals. Allowance for uncertainty in the study standard deviations lead to a narrower fixed effect confidence interval. The log-likelihood for study 4 was more concentrated than the others and multi-modal. Given this, the log-likelihoods suggested little need for random effects assumptions and were centered about 0. When the example was redone under a $LogNormal$ assumption, study 4's log-likelihood was much more quadratic. The fixed effect confidence interval under the $LogNormal$ assumption now excluded 0. Had the study observations been available, it would likely not have been at all controversial to take a log transformation. This is a well known standard practice in applied statistics when observations are constrained to be positive and likely skewed. With this transformation, one would have likely arrived at very similar conclusions. With the approach of this thesis, the same can be achieved by making a $LogNormal$ assumption and obtaining the observed summary likelihood.

6 Conclusions and future research

In this thesis parametric statistics have been viewed as the investigation and synthesis of individual observation likelihoods. This motivated a generic strategy for the investigation and synthesis of repeated or similar randomized clinical trials, often referred to as meta-analysis or systematic review. The generic strategy relied on the realization that the observed summary likelihood, given what summaries were obtainable from the trials, was the basic ingredient for this investigation and synthesis. Working with this, some commonness of a parameter of interest $\gamma(\theta)$ was sought, where commonness may only mean common in distribution. The investigation and synthesis was then carried out via likelihood with focussed inference via profile likelihood-based approximate confidence intervals. The observed summary likelihood approach was identified earlier for meta-analysis of mixed continuous and dichotomous responses by Dominici and Parmigiani[39], called a latent-variables approach and a fully Bayes approach implemented but a Classical approach abandoned. The reasons cited for this were the difficulty of deriving exact confidence intervals from likelihoods versus the straightforwardness of obtaining credible intervals without resorting to asymptotic approximations. By choosing just the Bayesian approach on examples where observed summary likelihoods are readily available in closed form, considerable computational burden was avoided. This thesis, on the other hand, undertook a Classical approach as well as provided a

general means for obtaining observed summary likelihoods (which had been lacking for the Bayesian approach). MCMC implementations of Bayesian analysis do not require level 2 likelihoods, though some suggest that would still be useful in Bayesian approaches[12] and the envelope lower and upper bounds would be useful there.

The statistical and meta-analysis literature has failed to inform clinical researchers about the perils of summarizing data in their research reports. Both the applied and theoretical literature tend to be misleading here. The applied literature, which suggests that medians and quartiles are appropriate summaries for skewed distributions and means and variances are not, has lead to an uninformative choice of limited summaries being reported. The theoretical literature has suggested that sufficiency and probability models that admit sufficient statistics of dimension less than number of observations are important concepts that have a role to play in what should be reported and what class of models are useful. Fisher was as misleading as any here, suggesting that the log-likelihood (sufficient) be reported in studies so that the studies could later be combined by simply adding up their log-likelihoods. Cox makes a similar suggestion involving profile likelihoods[27]. These suggestions assume complete certainty of the probability model assumed by the investigators - if this should differ between them or at any point come under question (i.e. such as when a consistent mean variance relationship is observed over many studies when assumptions were $Normal$) - the observed summary log-likelihood under the new model given the reported log-likelihood from the assumed model - would have to be simulated or derived for the studies to be properly contrasted and combined. The simulation of general reported summaries, such as a likelihood function, an observed p_value, etc. remains future research. Of particular interest are studies reporting various summaries from survival analysis, especially when stratification was pre-specified but largely only limited strata summaries are provided. A much better solution for all concerned would be the archiving of the data, or the reporting of as many order statistics as confidentiality (and the journal editor) will allow. This has to be by relevant grouping, i.e. simply by treatment and control if unstratified but separately by strata if stratified.

Some are suggesting that more generally, sufficient statistics of dimension less than number of observations are simply a misleading distraction for the theory of statistics[55]. From the perspective of this thesis, they are seen as a trivial consequence of the observed summary likelihood being a multiple of the likelihood from any sample that had the same value for the sufficient statistics. In terms of the Monte-Carlo simulation of observed summary likelihoods (up to a

multiplicative constant), the variance is zero and so the simulation approximation is exact. This may be an important advantage of convenience, but only if the original sample is not available! However, this should not distract from the role of sufficiency in separating evaluation of model fit from that of parameter estimation[27].

Further, with models that involve nuisance scale parameters such as within study variances, the danger of assuming these are known has been underestimated for both fixed and random effects models, for instance as in Cox[24] and the "bad interaction" of profiling out the unknown scale parameter and robust adjustments for random effects models needs to be pointed out - Stafford's adjustment may actually make things worse. This thesis has identified it as a potentially serious issue for meta-analysis. A conjecture recently made by David Andrews in his 2006 Statistical Society of Canada Gold Medal address, along with another means to adjust in particular the 95% pooled log-likelihood-based confidence interval, remains future research.

Profile likelihood has been heavily relied on in this thesis. Profiling out nuisance parameters can result in problematic inference with sparse data and it is hard to define sparseness. To address this, the investigation of simulated modified profile likelihood was suggested and a simple demonstration in a "toy" example was carried out. Further, numerically challenging work remains to be carried out for real examples.

The work in this thesis would suggest that those undertaking meta-analysis initially start with a standard Cochrane approach, perhaps using the convenient Meta package provided in R. Simple imputations of means and standard deviations for studies that do not provide these summaries, would also be a reasonable starting point as would the approximate $Normal - Normal$ random effects model implemented using the DerSimonian-Laird estimate of the between study variance. This approximate initial analysis allows a simple view of the contrast and combination of study results and provides starting values for a more rigorous model-based analysis - but it needs to be checked against a more rigorous model-based analysis. This more rigorous model-based analysis should use observed summary likelihoods and level 2 likelihoods as basic ingredients and (try to) globally profile out all nuisance parameters to focus on a pooled confidence interval for the common parameter of interest. Alternatively, a Bayesian approach could be used along with observed summary likelihoods. A number of distributions for both the outcomes and random parameters should be considered. This may not be computationally feasible for some meta-analysis and suboptimal analyses may need to be accepted (i.e. study-wise profiling out of nuisance parameters or limited

use of distributions). If the practical results of the initial analyses hold up to the more rigorous analyses, then they may be useful simpler views of the contrast and combination of study results. Otherwise, they are importantly misleading and should be abandoned or retracted. On the other hand, it is important to consider that the difference, especially mutli-modality in the likelihoods, may be a sign of poor model fit[27], and the more rigorous analyses done on more reasonable assumptions.

The thesis was directed at ideally conducted randomized clinical trials, where as this is seldom the case in practice. However, it is considered worthwhile to know what to do in an ideal setting prior to dealing with a less ideal setting. Some discussion of sensitivity analysis was given for informative selection by authors, of which summaries to make available and the possible need for sensitivity analysis using informative priors was pointed out. Although beyond the scope of this thesis, a recent example of such work for non-randomized studies and meta-analysis is given in Greenland [59][60] and Copas and Eguchi[20].

The conceptual value of viewing parametric statistics as the investigation and synthesis of single observation likelihoods may be of some value outside meta-analysis applications per se. The recasting of recent work by Tibshirani and Efron was presented as a possible example of this. This is a difficult point to be certain of, as conceptual value is in the eye of the "conceptualizer".

References

[1] Airy, G. B. *On the algebraical and numerical theory of errors of observations and the combination of observations*. MacMillan and Co., Cambridge, 1861.

[2] Andrews, D. F. Comment on : Chatfield C. Avoiding statistical pitfalls. *Statistical Science 6*, 3 (1991), 253.

[3] Arnold, B. C., Balakrishnan, N., and Nagaraja, H. N. *A first course in order statistics*. John Wiley and Sons, New York, 1992.

[4] Aronson, J. When I use a word: Meta-. *BMJ 324*, 7344 (2002), 1022.

[5] Barnard, G. A. Review of statistical inference and analysis: Selected correspondence of R. A. Fisher. Edited by J. H. Bennett. *Statistical Science 7* (1993), 5–12.

[6] Barnard, G. A., and Copas, J. B. Likelihood inference for location, scale and shape. *Journal of Statistical Planning and Inference 108* (2002), 71–83.

[7] Barndorff-Nielsen, O. *Information and exponential families: In statistical theory*. Chichester, New York, 1978.

[8] Barndorff-Nielsen, O. Likelihood theory. In *Statistical Theory and Modelling*, D. V. Hinkley, N. Reid, and E. J. Snell, Eds. Chapman and Hall, London, 1990.

[9] Barndorff-Nielsen, O., and Cox, D. R. *Inference and asymptotics*. Chapman and Hall, London, 1994.

[10] Barrowman, N. J., and Myers, R. A. Raindrop plots: A new way to display collections of likelihoods and distributions. *The American Statistician 57*, 4 (2003), 268–274.

[11] Bates, D. Sparse matrix representation of linear mixed models. Tech. rep., University of Wisconsin - Madison, 2006.

[12] Berger, J. O., Liseo, B., and Wolpert, R. L. Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science 14* (1999), 1–28.

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

[14] Bressoud, D. *A radical approach to real analysis*. The Mathematical Association of America, Washington, 1994.

[15] Carlin, B. P., and Louis, T. A. Empirical Bayes: Past, present and future. In *Statistics in the 21st Century*, A. Raftery, M. Tanner, and M. Wells, Eds. Chapman & Hall, London, 2001.

[16] Casella, G. An introduction to empirical Bayes data analysis. *The American Statistician 39* (1985), 83–87.

[17] Chatfield, C. Model uncertainty, data mining and statistical inference (Disc: p444-466). *Journal of the Royal Statistical Society, Series A, General 158* (1995), 419–444.

[18] Clarke, M., and Oxman, A. D. *Formulating the problem. Cochrane Reviewers' Handbook 4.2.0 [updated March 2003]*. Cochrane Collaboration, 2003. http://www.cochrane.dk/cochrane/handbook/handbook.htm (accessed 30 April 2003).

[19] Cochran, W. G. Problems arising in the analysis of a series of similar experiments. *Journal of Royal Statistical Society Supplement 4*, 1 (1937), 102–118.

[20] Copas, J. B., and Eguchi, S. Local model uncertainty and incomplete-data bias. *Journal of the Royal Statistical Society B 67*, 4 (2005), 459–513.

[21] Copas, J. B., and Li, H. G. Inference for non-random samples (Disc: p77-95). *Journal of the Royal Statistical Society, Series B, Methodological 59* (1997), 55–77.

[22] Copas, J. B., and Shi, J. Q. A sensitivity analysis for publication bias in systematic reviews. *Statistical Methods in Medical Research 10*, 4 (2001), 251–265.

[23] Cox, D. R. *The analysis of binary data*. Methuen, London, 1970.

[24] Cox, D. R. Combination of data. In *Encyclopedia of Statistical Science*, S. Kotz and N. L. Johnson, Eds. Wiley, New York, 1982.

[25] Cox, D. R. Some remarks on likelihood factorization. *IMS Lecture Note Series 136* (2000), 165–172.

[26] Cox, D. R. Comment on: Statistical modeling: The two cultures. *Statistical Science 16* (2001), 216–218.

[27] Cox, D. R. *Principles of statistical inference*. Cambridge University Press, Cambridge, 2006.

[28] Cox, D. R., and Hinkley, D. V. *Theoretical statistics*. Chapman & Hall Ltd, 1974.

[29] Cox, D. R., and Snell, E. J. *Analysis of binary data*. Chapman & Hall Ltd, 1989.

[30] Cox, D. R., and Wermuth, N. *Multivariate dependencies: models, analysis, and interpretation*. Chapman & Hall Ltd, 1998.

[31] David, H. A. *Order statistics*. John Wiley & Sons, 1981.

[32] Davison, A. C. *Statistical Models*. Cambridge University Press, Cambridge, 2003.

[33] Dawid, A. P., and Dickey, J. M. Likelihood and Bayesian inference from selectively reported data. *Journal of the American Statistical Association 72* (1977), 845–850.

[34] Dawid, A. P., Stone, M., and Zidek, J. V. Marginalization paradoxes in Bayesian and structural inference. *Journal of the Royal Statistical Society Series B 35*, 2 (1991), 189–233.

[35] Deeks, J. J., Altman, D. G., and Bradburn, M. J. Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In *Systematic reviews in health care: Meta-analysis in context*, M. Egger, G. Smith, and D. Altman, Eds. BMJ Books, London, 2001, ch. 15, pp. 285–312.

[36] Dempster, A. P. Logicist statistics I. Models and modeling. *Statistical Science 13* (1998), 248–276.

[37] Dempster, A. P. Logicist statistics II. Inference. *Submitted to Statistical Science* (2002).

[38] DerSimonian, R., and Laird, N. Meta-analysis in clinical trials. *Controlled Clinical Trials 7*, 3 (1986), 177–88.

[39] Dominici, F., and Parmigiani, G. Combining studies with continuous and dichotomous responses: A Latent-Variables approach. In *Meta-Analysis in Medicine and Health Policy*, D. Stangl and D. Berry, Eds. Marcel Dekker, 2000, ch. 5, pp. 105–125.

[40] Efron, B. Empirical Bayes methods for combining likelihoods. *Journal of the American Statistical Association 91*, 434 (1996), 538–565.

[41] Efron, B. 1996 R. A. Fisher lecture. *Statistical Science 13*, 2 (1998), 95–122.

[42] Ehrenberg, A. S. C., and Bound, J. A. Predictability and prediction. *Journal of the Royal Statistical Society, Series A, General 156* (1993), 167–206.

[43] Einarson, T. Pharmacoeconomic applications of meta-analysis for single groups using antifungal onychomycosis lacquers as an example. *Clin Ther. 19* (1997), 559–569.

[44] Evans, M., and Moshonov, H. Checking for prior-data conflict. Tech. rep., University of Toronto, 2005. University of Toronto Technical report No 0413.

[45] Evans, M., and Swartz, T. An algorithm for the approximation of integrals with exact error bounds, Oct. 15 1997.

[46] Fisher, R. A. On an absolute criterion for fitting frequency curves. *Messeng. Math 41* (1912), 155–160.

[47] Fisher, R. A. Theory of statistical estimation. *Phil. Trans., A 222* (1925), 309–368.

[48] Fisher, R. A. Two new properties of mathematical likelihood. *Journal of the Royal Statistical Society, Series A, General 144* (1934), 285–307.

[49] Fisher, R. A. *The Design of Experiments.* Oliver and Boyd, Edinburgh, 1935.

[50] Fisher, R. A. The logic of inductive inference. *Journal of the Royal Statistical Society 98* (1935), 39–54.

[51] Fisher, R. A. *Statistical Methods and Scientific Inference.* Oliver and Boyd, Edinburgh, 1959.

[52] Franklin, J. *The Science of conjecture: Evidence and probability before Pascal.* The Johns Hopkins University Press, Baltimore and London, 2001.

[53] Fraser, D. A. S. *Probability and Statistics: Theory and application.* Duxbury Press, North Scituate, 1976.

[54] Fraser, D. A. S. *Inference and linear models.* McGraw Hill, New York, 1979.

[55] Fraser, D. A. S. Fields Institute Lectures: Is the future Bayesian or frequentist? http://www.utstat.utoronto.ca/dfraser/, 2002.

[56] Gelman, A. Analysis of variance - why it is more important than ever. *Annals of Statistics 33*, 1 (2005), 1–53.

[57] Gelman, A. The boxer, the wrestler, and the coin flip: a paradox of robust Bayesian inference and belief functions. Tech. rep., Columbia University, 2006. To appear in The American Statistician.

[58] Greenland, S. Quantitative methods in the review of epidemiologic literature. *Epidemiol Rev. 9* (1987), 1–30.

[59] Greenland, S. The impact of prior distributions for uncontrolled confounding and response bias: A case study of the relation of wire codes and magnetic fields to childhood leukemia. *Journal of the American Statistical Association 98*, 461 (2003), 47–54.

[60] Greenland, S. Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society Series A 168*, 2 (2005), 267–306.

[61] Greenland, S., and O'Rourke, K. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. *Biostatistics 2*, 4 (2001), 463–471.

[62] Gustafson, P., Gelfand, A. E., Sahu, S. K., Johnson, W. O., Hanson, T. E., and Joseph, L. On model expansion, model contraction, identifiability and prior information: Two illustrative scenarios involving mismeasured variables. *Statistical Science 20*, 2 (2005), 111–140.

[63] Hald, A. *A history of mathematical statistics from 1750 to 1930*. John Wiley & Sons, 1998.

[64] Hawkins, D. M., Bradu, D., and Kass, G. V. Location of several outliers in multiple-regression data using elemental sets. *Technometrics 26*, 3 (1984), 189–233.

[65] Hedges, L. V., and Olkin, I. *Statistical methods for meta-analysis*. Academic Press, Orlando, FL, 1985.

[66] Hozo, S. P., Djulbegovic, B., and Hozo, I. Estimating the mean and variance from the median, range, and the size of a sample. *BMC Medical Research Methodology 5*, 1 (2005), 5–13.

[67] Kavvadias, D. J., and Vrahatis, M. N. Locating and computing all the simple roots and extrema of a function. *J-SIAM-J-SCI-COMP 17*, 5 (Sept. 1996), 1232–1248.

[68] Keiding, N. Comments on understanding the shape of the hazard rate: A process point of view. *Statistical Science 16* (1998), 19–20.

[69] Kendall, M. G., Bernoulli, D., Allen, C. G., and Euler, L. Studies in the history of probability and statistics: XI. Daniel Bernoulli on maximum likelihood. *Biometrika 48* (1961), 1–18.

[70] Keynes, J. M. The principal averages and the laws of error which lead to them. *Journal of the Royal Statistical Society 74* (1911), 322–331.

[71] Knight, K. *Mathematical statistics*. Chapman & Hall Ltd, 2000.

[72] L'Abbe, K. A., Detsky, A. S., and O'Rourke, K. Meta-analysis in clinical research. *Annals of Internal Medicine 107*, 2 (1987), 224–233.

[73] Lambert, P., Sutton, A., Burton, P., Abrams, K., and Jones, D. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine 24*, 5 (2005), 2401–2428.

[74] Lee, Y., and Nelder, J. A. Hierarchical generalized linear models. *J. Roy. Statist Soc. (B) 58* (1996), 619–656.

[75] Little, R. J. A., and Rubin, D. B. *Statistical analysis with missing data: Second edition.* John Wiley & Sons, New Jersey, 2002.

[76] Longford, N., and Nelder, J. Statistics versus statistical science in the regulatory process. *Statistics in Medicine 18* (1999), 2311–2320.

[77] McCullagh, P. Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B, Methodological 42* (1980), 109–142.

[78] McCullagh, P. Rejoinder: What is a statistical model? *The Annals of Statistics 30*, 5 (2002), 1300–1310.

[79] McCullagh, P., and Nelder, J. A. *Generalized linear models (Second edition).* Chapman & Hall Ltd, 1989.

[80] Nelder, J. A. There are no outliers in the stack-loss data. *Student 3*, 3 (2000), 211–216.

[81] Neyman, J., and Scott, E. L. Consistent estimates based on partially consistent observations. *Econometrica 16* (1948), 1–32.

[82] Olkin, I. History and goals. In *The future of meta-analysis*, K. W. Wachter and M. L. Straf, Eds. The Belknap Press of Harvard University Press, Cambridge, Massachusettes, 1990.

[83] O'Rourke, K. Meta-analysis: Conceptual issues of addressing apparent failure of individual study replication or "inexplicable" heterogeneity. In *Empirical Bayes and likelihood inference* (2001), pp. 161–183.

[84] O'Rourke, K. Meta-analytical themes in the history of statistics: 1700 to 1938. *Pakistan Journal of Statistics [Split into Series A and B, 1986-1994] 18*, 2 (2002), 285–299.

[85] O'Rourke, K., and Altman, D. G. Bayesian random effects meta-analysis of trials with binary outcomes: Methods for the absolute risk difference and relative risk scales. *Statistics in Medicine 24*, 17 (2005), 2733–2742.

[86] O'Rourke, K., and Detsky, A. S. Meta-analysis in medical research: strong encouragement for higher quality in individual research efforts. *Journal of Clinical Epidemiology 42*, 10 (1989), 1021–1024.

[87] O'Rourke, K., and et al. Incorporating quality appraisals into meta-analyses of randomized clinical trials. Tech. rep., University of Toronto. Dept. of Statistics, 1991.

[88] Owen, A. *Empirical likelihood*. Chapman and Hall, Boca Raton, 2001.

[89] Pace, L., and Salvan, A. *Principles of statistical inference: From a Neo-Fisherian perspective*. World Scientific, London, 1997.

[90] Pace, L., and Salvan, A. Adjustments of the profile likelihood from a new perspective. Tech. rep., University of Padova, 2004. http://www.stat.unipd.it/LIKASY/articles.html.

[91] Pardoe, I., and Cook, R. D. A graphical method for assessing the fit of a logistic regression model. *The American Statistician 56*, 4 (2002), 263–272.

[92] Pawitan, Y. *In all likelihood: Statistical modelling and inference using likelihood*. Clarendon Press, Oxford, 2001.

[93] Peña, D. Combining information in statistical modeling. *The American Statistician 51*, 5 (1997), 326–332.

[94] Peto, R. Discussion. *Statistics in Medicine 6* (1987), 242.

[95] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. ISBN 3-900051-07-0.

[96] Rao, C. R. R. A. Fisher: The founder of modern statistics. *Statistical Science 7*, 1 (1992), 34–48.

[97] Rao, P. S., Kaplan, J., and Cochran, W. G. Estimators for the one-way random effects model with unequal error variances. *Journal of the American Statistical Association 76* (1981), 89–97.

[98] Ritz J, Spiegelman, D. Equivalence of conditional and marginal regression models for clustered and longitudinal data. *Statistical Methods in Medical Research 13*, 4 (2004), 309–323.

[99] Robins, J., and Wasserman, L. The foundations of statistics: A vignette. http://lib.stat.cmu.edu/www/cmu-stats/, 1999.

[100] Ross, S. M. *Simulation*. Academic Press, 2002.

[101] Royall, R. M. *Statistical Evidence: A Likelihood Paradigm*. Chapman & Hall / CRC, London, 1997.

[102] Royall, R. M., and Tsou, T. Interpreting statistical evidence by using imperfect models: Robust adjusted likelihood functions. *Journal of the Royal Statistical Society. Series B: Statistical Methodology 65*, 2 (2003), 391–404.

[103] Sacks, H. S., Berrier, J., Reitman, D., Ancona-Berk, V. A., and Chalmers, T. C. Meta-analyses of randomized controlled trials. *N Engl J Med. 316* (1987), 450–455.

[104] Savage, L. J. On rereading Fisher. *Annals of Statistics 4* (1976), 441–500.

[105] Senn, S. The many modes of meta. *Drug Information Journal 34* (2002), 535–549.

[106] Severini, T. An approximation to the modified profile likelihood function. *Biometrika 85* (1998), 403–411.

[107] Skrondal, A., and Rabe-Hesketh, S. *Generalized latent variable modeling*. Chapman and Hall, Boca Raton, 2004.

[108] Smith, H. Step distribution approximation. In *Encyclopedia of Statistical Sciences*, S. Kotz, N. Johnson, and C. Read, Eds. Wiley, New-York, 1996, pp. 763–764.

[109] Sprott, D. A. *Statistical inference in science*. Springer-Verlag, New York, 2000.

[110] Sprott, D. A. The estimation of ratios from paired data. In *Empirical Bayes and likelihood inference* (2001), pp. 161–183.

[111] Stafford, J. E. A robust adjustment of the profile likelihood. *The Annals of Statistics 24* (1996), 336–352.

[112] Sterling, T. D. Publication decisions and their possible effects on inferences drawn from tests of significance–or vice versa. *Journal of the American Statistical Association 54*, 285 (1959), 30–34.

[113] Stigler, S. M. *The history of statistics: The measurement of uncertainty before 1900.* Harvard University Press, 1986.

[114] Stigler, S. M. Ancillary history. In *State of the Art in Probability and Statistics* (2001), pp. 555–567.

[115] Sung, Y. J., and Geyer, C. J. Monte-Carlo likelihood inference for missing data models. Tech. rep., University of Minnesota, 2006. Submitted (5 Jan 2005) and revised and resubmitted (10 Jan 2006).

[116] Thompson, S. G., Prevost (Nee Smith), T. C., and Sharp, S. J. Reply to comment on "Investigating underlying risk as a source of heterogeneity in meta-analysis". *Statistics in Medicine 18* (1999), 113–115.

[117] Tibshirani R J, Efron, B. Pre-validation and inference in microarrays. *Statistical Applications in Genetics and Molecular Biology 1*, 1 (2002), 1–18.

[118] Tjur, T. Discussion: What is a statistical model? *The Annals of Statistics 30*, 5 (2002), 1297–1300.

[119] Vangel, M. G., and Rukhin, A. L. Maximum likelihood analysis for heteroscedastic one-way random effects ANOVA in interlaboratory studies. *Biometrics 55*, 1 (1999), 129–136.

[120] Venables, W. N., and Ripley, B. D. *Modern Applied Statistics With S (4th ed.).* Springer-Verlag, New York, 2002.

[121] Warn, D. E., Thompson, S. G., and Spiegelhalter, D. J. Bayesian random effects meta-analysis of trials with binary outcomes: Methods for the absolute risk difference and relative risk scales. *Statistics in Medicine 21*, 11 (2002), 1601–1623.

[122] Wolfram Research, Inc. *Mathematica.* Wolfram Research, Inc., Champain, Illinois, 2004.

[123] Yusuf, S., Peto, R., and Lewis, J. Beta blockade during and after myocardial infarction: An overview of the randomized trials. *Progress in Cardiovascular Diseases 27*, 5 (1985), 335–371.

A Evasions of nuisance parameters in general

Here, more generally, some suggested strategies for dealing with nuisance parameters in inference will be reviewed from the statistical literature that has arisen since Neyman-Scott's paper. Intuitively, the heart of the issue of dealing with unknown nuisance parameters is to "evade" them by constructing something like a likelihood - say pseudo-likelihood - that does not explicitly involve the nuisance parameters but captures all of the relevant information "just about" the interest parameter(s) from the full likelihood (that did in fact involve the nuisance parameters) and such that this pseudo-likelihood has the usual properties of a likelihood (where such evasions were not taken.) Asymptotic consistency and efficiency of MLE for the interest parameter and at least consistency of variance of the MLE are perhaps the most important features to insist upon. Such evasions would be a real achievement and would greatly simplify meta-analysis (as well as many other areas of statistics) but this turns out to be possible only in certain situations which have resisted full generalization. "Partial" generalizations, where they have been made, are difficult to analytically implement in common practice[89] and current work on Monte-Carlo approaches (such as the one identified) is in need of further development. Notation from Sprott[109] to look at a select few but important evasions to get a sense of the possibilities and issues will be used.

Conditional evasion

$$\begin{aligned} L(\mu,\lambda,y) \quad \propto \quad f(\mu,\lambda,y) \quad &= \quad f(\mu,\lambda,t)f(\mu,y|t) \\ &\propto \quad L_{res}(\mu,\lambda,t)L_c(\mu,y) \end{aligned}$$

where t is chosen so that $f(\mu,\lambda,t) = f(\lambda,t)$ or $f(\mu,\lambda,t) \approx f(\lambda,t)$ and then one just uses

$$L(\mu,y) \equiv L_c(\mu,y)$$

for the likelihood for μ.

Marginal (over sample) evasion

$$\begin{aligned} L(\mu,\lambda,y) \quad \propto \quad f(\mu,\lambda,y) \quad &= \quad f(\mu,t)f(\mu,\lambda,y|t) \\ &\propto \quad L_m(\mu,t)L_{res}(\mu,\lambda,y) \end{aligned}$$

where t is chosen so that $f(\mu, \lambda, y|t) = f(\mu, y|t)$ or $f(\mu, \lambda, y|t) \approx f(\mu, y|t)$ and then one just uses

$$L(\mu, t) \equiv L_m(\mu, t)$$

for the likelihood for μ.

The success of these two evasions depends on how much information for μ is in L_{res}. For certain classes of probability models - exponential (using certain parameterizations) and location scale, these evasions are considered very successful. As they both arise from probability models (for subsets of the data) they are true likelihoods and hence they have the usual properties.

Maximized (over parameter) evasion - the profile likelihood is

$$L_p(\mu, \lambda, y) = \quad L(\mu, \hat{\lambda}_{(\mu)}, y)$$

where $\hat{\lambda}_{(\mu)}$ is the MLE of λ for a given value of μ. Then one just uses

$$L_p(\mu, y) \equiv L(\mu, \hat{\lambda}_{(\mu)}, y)$$

for the likelihood for μ.

Note in effect the likelihood is conditioned on $\hat{\lambda}_{(\mu)}$ as if it were known when in fact it is not. The profile likelihood is known sometimes to provide asymptotically inconsistent or inefficient MLEs for the parameter of interest. It is suspected that this happens mostly when $\hat{\lambda}_{(\mu)}$ is poorly estimated due to small sample size - the Neyman-Scott examples being the canonical examples of this. On the other hand, this evasion can be very widely applied, being "mechanical" or numerically implemented, and it is known to approximate the two previous successful evasions where those are available. Modifications are available for the profile likelihood to improve this approximation but again are difficult to analytically apply (widely) in practice[89]. Again, we identified approximations that are available via Monte-Carlo simulation and offer some promise to becoming widely applicable.

Marginal (over parameter) evasion

$$L^{M_p}(\mu, y) = \quad \int L(\mu, \lambda^*, y) p(\lambda^*) d\lambda^*$$

where λ is integrated out with respect to $p(\lambda^*)$ which represents physical random variation of λ^*.

This is referred to as a level 2 likelihood in this thesis. Again, this needs to carefully distinguished from a conceptual (Bayesian) representation of the uncertainty of λ (not necessarily a random variable) which gives $L^I(\mu, y)$ in the special case of $p(\lambda)$ being uniform $L^{I_u}(\mu, y)$ (often referred to as simply the integrated likelihood). For λ_i^* being considered a sample from a common distribution of $p(\lambda^*)$ one would have

$$L^{M_p}(\mu, y) = \int \cdots \int L(\mu, \lambda_1^*, y)p(\lambda_1^*)...L(\mu, \lambda_n^*, y)p(\lambda_n^*)d\lambda_1^*...d\lambda_n^*$$

or more conveniently

$$L(\mu, y) = \int L(\mu, \lambda_1^*, y)p(\lambda_1^*)d\lambda_1 ... \int L(\mu, \lambda_n^*, y)p(\lambda_n^*)d\lambda_n^* \ .$$

Note $p(\lambda_1^*, \lambda_2^*) = p(\lambda_1^*)p(\lambda_2^*)$ has been considered as the interest in meta-analysis and it is usual to assume the "random samples" of the treatment effect are independent from study to study.

B Generalized likelihood

Given this need for random effects models with unobserved random parameters, generalized likelihood is now briefly discussed. In Bjornstad[13], a generalization of the likelihood is given where "unobservables" which may consist of random parameters (or variables to be predicted - though these are of little direct interest in this thesis) are explicitly denoted by ψ to distinguish them from fixed unknown parameters θ. Various quantities of interest for statistical inference - λ which are functions of ψ, i.e. $\lambda = f(\psi)$ are also explicitly denoted. With this he defines the complete specified probability model as

$$\Pr = \{f_\theta(y; \theta^*), \theta \in \Theta\}$$

where unobserved random parameters θ^* and fixed unknown parameters θ both enter but "separately". He then defines the generalized likelihood as

$$L_y(\lambda, \theta) = f_\theta(y; \lambda)$$

where the unobserved random parameters θ^* do not enter directly except through the function $\lambda = f(\theta^*)$ - that again denotes the quantities of interest that involve θ^*. Given this definition, he

points out the following:

1. The part of θ^* that is not of inferential interest is integrated out. For example, if no part of θ^* is of interest, then θ^* is integrated out completely and $L_y(\theta) = f_\theta(y)$, the usual parametric likelihood [in our case, usually study specific random effects].

2. We do not condition on the variable, λ, of interest.

3. The fixed unknown θ must be included in the likelihood whether or not θ is of inferential interest.

The force of these claims depends on Bjornstad's generalization providing a likelihood for which Birnbaum's theorem generalizes. Birnbaum's theorem establishes that principles of sufficiency and conditionality imply the (strong) likelihood principle. Intuitively, this can be expressed as the essential idea that statistical inference should be based on all relevant information and not any non-relevant information (sufficiency being the all and conditional being the relevant). Now the likelihood principle is not universally or even widely accepted and hence it is far from ideal or convincing to use it to resolve the controversy of whether unobserved random parameters should be integrated out of the likelihood when they are not of direct interest - i.e. when the interest is primarily in the parameters of the higher level distribution. Having made the controversy clear, we do however use Bjornstad's generalization for the purposes of this thesis.

C Single observation inference examples

Starting with single observations and single unknown parameters it is possible to motivate deviance residuals and deviance and demonstrate single observation "inference". Then with single observations and multiple unknown parameters it is possible to motivate estimated and profile likelihood and demonstrate single observation "inference" for multiple parameters. A quantitative evaluation of commonness succeeded or failed based on the assumed model and or parameter of interest. It is though, demonstrated that it is the likelihood from the assumed probability distribution that generated what was observed, that "tells" one what is the "best" combination of observations, given the parameters were common or common in distribution and provides at least a qualitative assessment of the commonness of the parameters.

C.1 Example 4.1 - Bernoulli distribution

The situation of binary outcomes with a common proportion is perhaps the simplest situation to deal with but also the least satisfying or informative (given only one group and no order of observation information). Likelihood is p for observed success and $(1-p)$ for an observed failure giving $p^y(1-p)^{1-y}$ where $y = 1, 0$ for a success or failure respectively. Plots or functions of these are not likely to facilitate an informative investigation about whether p was actually common. Over all observations one could check how relatively less probable a common p makes the observations than a separate p_i for each observation i.e.

$$\frac{\prod_{i=1}^{n} p^{y_i}(1-p)^{(1-y_i)}}{\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{(1-y_i)}}.$$

Using the maximum likelihood estimates under both assumptions the relative probability is

$$p^{\sum_i^n y_i}(1-p)^{(n-\sum_i^n y_i)}$$

(recall that $p^0 \equiv 1$ for all p in $p^y(1-p)^{1-y}$). Given that this is just a function of n and $\sum y_i$, it is not helpful for investigating commonness.

Rather than directly focusing on how relatively less probable a common p makes the observations than a separate p_i, we can derive some familiar likelihood-based residuals and the so called goodness of fit statistic. First the log-likelihood is used

$$l\left(p; y_i\right) = y_i \log(p) + (1-y_i)\log(1-p)$$

and the simple difference between the maximum of each used

$$\begin{aligned} & \sum_{i=1}^{n} l\left(\widehat{p}_i; y_i\right) - \sum_{i=1}^{n} l\left(\widehat{p}; y_i\right) \\ = & \sum_{i=1}^{n} \{l\left(\widehat{p}_i; y_i\right) - l\left(\widehat{p}; y_i\right)\} \end{aligned}$$

where $(\widehat{p}_i, \widehat{p})$ are the values that make the observations observed most probable and the subtraction is defined to make the result positive.

A simple re-scaling of this is formally called the deviance (in generalized linear models for

instance) and is defined as

$$dev\left(\widehat{p},\widehat{p}_i;y\right)=2\sum_{i=1}^{n}\{l\left(\widehat{p}_i;y_i\right)-l\left(\widehat{p};y_i\right)\}$$

and one may wish to look at the components of this individually, i.e.

$$2\{l\left(\widehat{p}_i;y_i\right)-l\left(\widehat{p};y_i\right)\}$$

and a function of this

$$sign(\widehat{p}_i-\widehat{p})\sqrt{2\{l\left(\widehat{p}_i;y_i\right)-l\left(\widehat{p};y_i\right)\}}$$

is formally called the deviance residual.

In this example

$$\begin{aligned}
& 2\{l\left(\widehat{p}_i;y_i\right)-l\left(\widehat{p};y_i\right)\} \\
= \;& 2\{y_i\log(\widehat{p}_i)+(1-y_i)\log(1-\widehat{p}_i)-y_i\log(\widehat{p})+(1-y_i)\log(1-\widehat{p})\}
\end{aligned}$$

which with $\widehat{p}_i$ replaced with y_i and $\widehat{p}$ replaced by $\bar{y}$ (the MLEs respectively)

$$\begin{aligned}
= \;& 2\{y_i\log(y_i)+(1-y_i)\log(1-y_i)-y_i\log(\bar{y})+(1-y_i)\log(1-\bar{y})\} \\
= \;& 2\{-y_i\log(\bar{y})+(1-y_i)\log(1-\bar{y})\} \\
= \;& -2\log(\bar{y}) \text{ for } y_i=1 \text{ and } 2\log(1-\bar{y}) \text{ for } y_i=0.
\end{aligned}$$

In terms of the deviance

$$\begin{aligned}
& 2\sum_{i=1}^{n}\{l\left(\widehat{p}_i;y_i\right)-l\left(\widehat{p};y_i\right)\} \\
= \;& 2\sum_{i=1}^{n}\{y_i\log(\widehat{p}_i)+(1-y_i)\log(1-\widehat{p}_i)-y_i\log(\widehat{p})+(1-y_i)\log(1-\widehat{p})\}
\end{aligned}$$

which with $\widehat{p}_i$ replaced with y_i and $\widehat{p}$ replaced by $\bar{y}$ (the MLEs respectively)

$$
\begin{aligned}
& 2\sum_{i=1}^{n}\{y_i\log(y_i)+(1-y_i)\log(1-y_i)-y_i\log(\bar{y})+(1-y_i)\log(1-\bar{y})\} \\
= \; & 2\sum_{i=1}^{n}0-2\sum_{i=1}^{n}\{y_i\log(\bar{y})+(1-y_i)\log(1-\bar{y})\} \\
= \; & 2\log(\bar{y})\sum_{i=1}^{n}y_i+2\log(1-\bar{y})\sum_{i=1}^{n}(1-y_i) \\
= \; & 2\log(\bar{y})n\bar{y}+2\log(1-\bar{y})(n-n\bar{y}).
\end{aligned}
$$

Again, given that it is just a function of n and $\bar{y}$, it is not helpful for investigating commonness. Note the attempt was to directly assess the commonness of the parameter, not for instance whether the observed data seems like a reasonable sample from the given probability model. Combination is by multiplication i.e. $p^{\sum_i^n y_i}(1-p)^{\sum_i^n 1-y_i}$ where y_i are the multiple observed outcomes.

When the p arbitrarily varies by observation the multiplication simply gives

$$\prod_{i=1}^{n}p_i^{y_i}(1-p_i)^{(1-y_i)}.$$

But if the p come from a common distribution , say, $f(\theta)$ $\theta\in\Theta$ then for the level 2 likelihood

$$\int_0^1\theta^y(1-\theta)^{1-y}f(\theta)d\theta$$

say, $f^*(\theta)$ a multiplication does provide a combination (for the common parameters in $f^*(\theta)$). As a convenient example if Θ has the beta density,

$$\theta^{\alpha-1}(1-\theta)^{\beta-1}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

we have that

$$
\begin{aligned}
p(y) &= \binom{1}{y}\int_0^1\theta^y(1-\theta)^{1-y}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \\
&= \frac{\Gamma(1+1)}{\Gamma(y+1)\Gamma(1-y+1)}\int_0^1\theta^{y+\alpha-1}(1-\theta)^{(1-y)+\beta-1}d\theta\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \\
&= \frac{\Gamma(1+1)}{\Gamma(y+1)\Gamma(1-y+1)}\frac{\Gamma(y+\alpha)\Gamma(1-y+\beta)}{\Gamma(1+\alpha+\beta)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}.
\end{aligned}
$$

Note that with this marginal likelihood the multiplication results in the combination of α and β

$$\prod_{i=1}^{n} \frac{\Gamma(1+1)\Gamma(y_i+\alpha)\Gamma(1-y_i+\beta)\Gamma(\alpha+\beta)}{\Gamma(y_i+1)\Gamma(1-y_i+1)\Gamma(1+\alpha+\beta)\Gamma(\alpha)\Gamma(\beta)}.$$

Under this distribution the expectation of y_i is $\alpha/(\alpha+\beta)$ but unfortunately as y_i can only take values zero and one, the variance equals $p-p^2$ and similarly all higher moments are functions of p so α and β are unidentifiable (various values of α and β give rise to same distribution).

C.2 Example 4.2 - Gaussian distribution with known scale

Perhaps the next simplest probability model to work with in this manner is the Gaussian distribution with known scale

$$\begin{aligned} f\left(y_i;\alpha,\sigma_0^2\right) &= (2\pi\sigma_0^2)^{-1/2} e^{-(y_i-\alpha)^2/2\sigma_0^2} \\ L\left(\alpha;y_i\right) &= c(\sigma_0^2)^{-1/2} e^{-(y_i-\alpha)^2/2\sigma_0^2} \end{aligned}$$

and for convenience we will take logarithms

$$\begin{aligned} \log L\left(\alpha;y_i\right) &= -\frac{1}{2\sigma_0^2}(y_i-\alpha)^2 + \log(c) \\ l\left(\alpha;y_i\right) &= -\frac{1}{2\sigma_0^2}(y_i-\alpha)^2 + \log(c). \end{aligned}$$

Note the log-likelihood $l\left(\alpha;y_i\right)$ is a second degree polynomial that is straight forward to add and find the maximum of. They can be plotted to graphically investigate if the log-likelihoods support a common α.

In this example the deviance components are

$$\begin{aligned} & 2\{l\left(\hat{\alpha}_i;y_i\right) - l\left(\hat{\alpha};y\right)\} \\ = \; & 2\{-\frac{1}{2\sigma_0^2}(y_i-\hat{\alpha}_i)^2 + \frac{1}{2\sigma_0^2}(y_i-\hat{\alpha})^2\} \end{aligned}$$

which with $\hat{\alpha}_i$ replaced with y_i and $\hat{\alpha}$ replaced by $\bar{y}$ (the MLEs respectively)

$$
\begin{aligned}
&= 2\{-\frac{1}{2\sigma_0^2}(y_i - y_i)^2 + \frac{1}{2\sigma_0^2}(y_i - \bar{y})^2\} \\
&= 2\{\frac{1}{2\sigma_0^2}(y_i - \bar{y})^2\} \\
&= \frac{1}{\sigma_0^2}(y_i - \bar{y})^2.
\end{aligned}
$$

Here the deviance provides an informative quantitative assessment of commonness of α.

$$
\begin{aligned}
dev\left(\widehat{\alpha}, \widehat{\alpha}_i; y\right) &= 2\sum_i \{l\left(y_i; \widehat{\alpha}_i\right) - l\left(y_i; \widehat{\alpha}\right)\} \\
&= \sum_i \frac{1}{\sigma_0^2}(y_i - \bar{y})^2
\end{aligned}
$$

C.3 Example 4.3 - Laplacian distribution with known scale

Another simple probability model is a Laplacian distribution also with known scale

$$
\begin{aligned}
f\left(y_i; \alpha, \sigma_0\right) &= \frac{1}{2\sigma_0} e^{-|y_i - \alpha|/\sigma_0} \\
L\left(\alpha; y_i\right) &= c e^{-|y_i - \alpha|/\sigma_0} \\
l\left(\alpha; y_i\right) &= -|y_i - \alpha|/\sigma_0 + \log(c)
\end{aligned}
$$

Here the log-likelihood is $-|y_i - \alpha|/\sigma_0$ a triangular function with maximum at $\alpha = y_i$. They can also be plotted to graphically investigate if the log-likelihoods support a common α. In this example the deviance components are

$$
\begin{aligned}
& 2\{l\left(\hat{\alpha}_i; y_i\right) - l\left(\hat{\alpha}; y_i\right)\} \\
= \; & 2\{-|y_i - \hat{\alpha}_i|/\sigma_0 + |y_i - \hat{\alpha}|/\sigma_0\}
\end{aligned}
$$

which with $\hat{\alpha}_i$ replaced with y_i and $\hat{\alpha}$ replaced by the median y (the MLEs respectively)

$$
\begin{aligned}
&= 2\{-|y_i - y_i|/\sigma_0 + |y_i - y|/\sigma_0\} \\
&= 2|y_i - y|/\sigma_0.
\end{aligned}
$$

Here again the deviance provides an informative quantitative assessment of commonness of α.

$$\begin{aligned} dev\left(\hat{\alpha},\hat{\alpha}_i;y\right) &= 2\sum_i \{l\left(\hat{\alpha}_i;y_i,\sigma_0\right) - l\left(\hat{\alpha};y_i,\sigma_0\right)\} \\ &= 2\sum_i |y_i - y|/\sigma_0. \end{aligned}$$

C.4 Example 4.4 - Gaussian distribution with unknown mean and unknown scale

The only other probability distribution to be looked at with single outcomes is Gaussian distribution with unknown mean and unknown scale. With more than one parameter multidimensional surfaces have to be dealt with

$$\begin{aligned} f\left(y_i;\alpha,\sigma^2\right) &= (2\pi\sigma^2)^{-1/2} e^{-(y_i-\alpha)^2/2\sigma^2} \\ L\left(\alpha,\sigma^2;y_i\right) &= c(\sigma^2)^{-1/2} e^{-(y_i-\alpha)^2/2\sigma^2} \\ l\left(\alpha,\sigma^2;y_i\right) &= -\frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y_i-\alpha)^2 + \log(c). \end{aligned}$$

They can be plotted to graphically investigate if the log-likelihoods support a common α and σ. The likelihood here is unbounded at $(\alpha = y_i, \sigma = 0)$ and so also would be the deviance which is a function of it. But as y_i has only been observed to some level of accuracy, the observed summary likelihood - $\int_{y_i-\epsilon}^{y_i+\epsilon} L\left(\alpha,\sigma;y_i\right) dy_i$ needs to be used and with this correct likelihood the MLE of σ is a function of ϵ and reaches a finite maximum for σ small relative to ϵ[28].

C.4.1 Example 4.4 - Estimated likelihood

Given the likelihood for the Gaussian distribution with unknown mean and unknown scale involves two parameters and sensible estimates of both are not available from just one observation, it might be sensible to initially forgo the investigation of commonness and just assume it tentatively in order to get initial sensible estimates of both parameters. That is multiply the likelihoods for the observations together and get the joint MLE for the two parameters. Then one could take one parameter as known (in turn) and return to the single observation likelihoods for the other parameter. For this we will want to start with the combination

$$L\left(\alpha,\sigma^2;y\right) = c(\sigma^2)^{-n/2} e^{-\sum_i^n (y_i-\alpha)^2/2\sigma^2}$$

and for convenience we will take logarithms

$$\log L\left(\alpha,\sigma^2;y\right)=-\frac{n}{2}\log\sigma^2-\frac{1}{2\sigma^2}\sum_i^n(y_i-\alpha)^2+c.$$

Taking derivatives of $\log L\left(y;\alpha,\sigma\right)$ we obtain what are referred to as score functions

$$\begin{aligned}S_1\left(\alpha,\sigma^2;y\right) &= \frac{\partial}{\partial\alpha}\log L\left(\alpha,\sigma^2;y\right)=\frac{1}{\sigma^2}\sum_i^n(y_i-\alpha)\\ S_2\left(\alpha,\sigma^2;y\right) &= \frac{\partial}{\partial\sigma^2}\log L\left(\alpha,\sigma^2;y\right)=-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_i^n(y_i-\alpha)^2.\end{aligned}$$

Equating these to zero yields the joint MLEs

$$\begin{aligned}\hat{\alpha} &= \bar{y}\\ \hat{\sigma}^2 &= \frac{1}{n}\sum_i^n(y_i-\bar{y})^2.\end{aligned}$$

The question of commonness of σ^2 can be left aside (for now) by taking it as equal to $\hat{\sigma}^2$ from the joint MLE $\left(\hat{\alpha},\hat{\sigma}^2\right)$ and assess, evaluate and combine single observations just for α.

$$\begin{aligned}l\left(\alpha;y_i\right) &= \frac{1}{2\hat{\sigma}^2}(y_i-\alpha)^2\\ &= -\frac{1}{2\frac{1}{n}\sum_i^n(y_i-\bar{y})^2}(y_i-\alpha)^2.\end{aligned}$$

In this example the deviance components are

$$\begin{aligned}& 2\{l\left(\hat{\alpha}_i;y_i\right)-l\left(\hat{\alpha};y_i\right)\}\\ = \;& 2\{-\frac{1}{2\frac{1}{n}\sum_i^n(y_i-\bar{y})^2}(y_i-\hat{\alpha}_i)^2+\frac{1}{2\frac{1}{n}\sum_i^n(y_i-\bar{y})^2}(y_i-\hat{\alpha})^2\}\end{aligned}$$

which with $\hat{\alpha}_i$ replaced with y_i and $\hat{\alpha}$ replaced by $\bar{y}$ (the MLEs respectively)

$$\begin{aligned}= \;& 2\{-\frac{1}{2\frac{1}{n}\sum_i^n(y_i-\bar{y})^2}(y_i-y_i)^2+\frac{1}{2\frac{1}{n}\sum_i^n(y_i-\bar{y})^2}(y_i-\bar{y})^2\}\\ = \;& 2\{\frac{1}{2\frac{1}{n}\sum_i^n(y_i-\bar{y})^2}(y_i-\bar{y})^2\}.\end{aligned}$$

Here the deviance does not provide an informative quantitative assessment of commonness of α, and of course is based on using $\widehat{\alpha} = \overline{y}$ to estimate σ^2

$$\begin{aligned} dev\left(\widehat{\alpha}, \widehat{\alpha}_i; y\right) &= 2\sum_i^n \{l\left(y_i; \widehat{\alpha}_i\right) - l\left(y_i; \widehat{\alpha}\right)\} \\ &= 2\sum_i^n \frac{1}{2\frac{1}{n}\sum_i^n (y_i - \bar{y})^2}(y_i - \bar{y})^2 \\ &= \frac{n}{\sum_i^n (y_i - \bar{y})^2}\sum_i^n (y_i - \bar{y})^2 \\ &= n. \end{aligned}$$

Alternatively, the question of commonness of α can be left aside (for now) by taking it as equal to $\hat{\alpha}$ and assess, evaluate and combine single observations just for σ^2.

$$\begin{aligned} l\left(\sigma^2; y_i\right) &= -\frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y_i - \hat{\alpha})^2 \\ &= -\frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y_i - \bar{y})^2. \end{aligned}$$

In this example the deviance components are

$$\begin{aligned} & 2\{l\left(\hat{\sigma}_i^2; y_i\right) - l\left(\hat{\sigma}^2; y_i\right)\} \\ = \; & 2\{-\frac{1}{2}\log\hat{\sigma}_i^2 - \frac{1}{2\hat{\sigma}_i^2}(y_i - \bar{y})^2 + \frac{1}{2}\log\hat{\sigma}^2 + \frac{1}{2\hat{\sigma}^2}(y_i - \bar{y})^2\} \end{aligned}$$

which with $\hat{\sigma}_i^2$ replaced with $(y_i - \bar{y})^2$ and $\hat{\sigma}^2$ replaced by $\frac{1}{n}\sum_i^n (y_i - \bar{y})^2$ (the MLEs respectively)

$$\begin{aligned} = \; & 2\{-\frac{1}{2}\log(y_i - \bar{y})^2 - \frac{1}{2(y_i - \bar{y})^2}(y_i - \bar{y})^2 \\ & + \frac{1}{2}\log\frac{1}{n}\sum_i^n (y_i - \bar{y})^2 + \frac{1}{2\frac{1}{n}\sum_i^n (y_i - \bar{y})^2}(y_i - \bar{y})^2\} \\ = \; & 2\{-\frac{1}{2}\log(y_i - \bar{y})^2 - \frac{1}{2} \\ & + \frac{1}{2}\log\frac{1}{n}\sum_i^n (y_i - \bar{y})^2 + \frac{1}{2\frac{1}{n}\sum_i^n (y_i - \bar{y})^2}(y_i - \bar{y})^2\} \\ = \; & \{-\log(y_i - \bar{y})^2 - 1 + \log\frac{1}{n}\sum_i^n (y_i - \bar{y})^2 + \frac{1}{\frac{1}{n}\sum_i^n (y_i - \bar{y})^2}(y_i - \bar{y})^2\}. \end{aligned}$$

Here the deviance does provide an informative quantitative assessment of commonness of σ^2

$$
\begin{aligned}
dev\left(\hat{\sigma}^2, \hat{\sigma}_i^2; y\right) &= 2\sum_i \{l\left(\hat{\sigma}_i^2; y_i\right) - l\left(\hat{\sigma}^2; y_i\right)\} \\
&= 2\sum_i \{-\frac{1}{2}\log(y_i - \bar{y})^2 - \frac{1}{2} + \\
&\quad \frac{1}{2}\log\frac{1}{n}\sum_i^n (y_i - \bar{y})^2 + \frac{1}{2\frac{1}{n}\sum_i^n (y_i - \bar{y})^2}(y_i - \bar{y})^2\} \\
&= \sum_i \{-\log(y_i - \bar{y})^2 - 1 \\
&\quad + \log\frac{1}{n}\sum_i^n (y_i - \bar{y})^2 + \frac{1}{\frac{1}{n}\sum_i^n (y_i - \bar{y})^2}(y_i - \bar{y})^2\} \\
&= n\log\frac{1}{n}\sum_i^n (y_i - \bar{y})^2 - \sum_i \log(y_i - \bar{y})^2 - n + n \\
&= n\log(\frac{1}{n}\sum_i^n (y_i - \bar{y})^2) - \sum_i \log(y_i - \bar{y})^2.
\end{aligned}
$$

This process of taking "all else equal" and investigating one aspect at a time has been a pragmatically useful device in science. Formally in statistics, likelihoods with unknown parameters replaced by estimates (usually MLEs) are called estimated likelihoods, although the motivation to do single observation inference probably played no part in their development.

C.4.2 Example 4.4 - Profile likelihood

For reasons that will be clearer later, a slightly different way to replace an unknown parameter, say σ^2 with an estimate and treat it as known would be to replace it with an estimate based on a given value of α (a "best estimate" given that value of α) in particular with the MLE given that value of α, with suggestive notation for this being $\hat{\sigma}_\alpha^2$. In the score equation for this the α is taken as fixed, and hence it only depends on σ^2

$$
S\left(\sigma^2; y\right) = \frac{\partial}{\partial\sigma^2}\log L\left(\sigma^2; y\right) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_i^n (y_i - \alpha)^2
$$

and setting this equal to zero results in

$$
\hat{\sigma}_\alpha^2 = \frac{1}{n}\sum_i^n (y_i - \alpha)^2.
$$

Replacing σ^2 with this equation in the log-likelihood results in

$$\begin{aligned} l(\alpha; y_i) &= \frac{1}{2\hat{\sigma}^2}(y_i - \alpha)^2 \\ &= -\frac{1}{2\frac{1}{n}\sum_j^n (y_j - \alpha)^2}(y_i - \alpha)^2. \end{aligned}$$

(Note the use of index j for the summation involving all observations in the group.)

In this example the deviance components are

$$\begin{aligned} & 2\{l(\hat{\alpha}_i; y_i) - l(\hat{\alpha}; y_i)\} \\ = \; & 2\{-\frac{1}{2\frac{1}{n}\sum_j^n (y_j - \hat{\alpha}_i)^2}(y_i - \hat{\alpha}_i)^2 + \frac{1}{2\frac{1}{n}\sum_j^n (y_j - \hat{\alpha})^2}(y_i - \hat{\alpha})^2\} \end{aligned}$$

which with $\hat{\alpha}_i$ replaced with y_i and $\hat{\alpha}$ replaced by $\bar{y}$ (the MLEs respectively)

$$\begin{aligned} &= 2\{-\frac{1}{2\frac{1}{n}\sum_j^n (y_j - y_i)^2}(y_i - y_i)^2 + \frac{1}{2\frac{1}{n}\sum_j^n (y_j - \bar{y})^2}(y_i - \bar{y})^2\} \\ &= 0 + \frac{1}{2\frac{1}{n}\sum_i^n (y_i - \bar{y})^2}(y_i - \bar{y})^2\}. \end{aligned}$$

And the deviance then is

$$\begin{aligned} dev(\alpha, \widehat{\alpha}_i; y) &= 2\sum_i \{l(y_i; \widehat{\alpha}_i) - l(y_i; \widehat{\alpha})\} \\ &= 2\sum_i \{0 + \frac{1}{2\frac{1}{n}\sum_i^n (y_i - \bar{y})^2}(y_i - \bar{y})^2\} \\ &= n \end{aligned}$$

which again, as with the estimated likelihood, is still not helpful in assessing the commonness of α.

The same could be considered for α (replacing it with $\widehat{\alpha}_{\sigma^2}$)

$$S(\alpha; y) = \frac{\partial}{\partial \alpha} \log L(y; \alpha) = \frac{1}{\sigma^2}\sum_i^n (y_i - \alpha) = c\sum_i^n (y_i - \alpha)$$

which, regardless of the value of σ^2, has MLE $\bar{y}$ so it is simply the same as the estimated likelihood. This alternative way of getting one dimensional likelihoods is formally called profile likelihood.

D Neyman and Scott examples

Interestingly, meta-analysis problems from astronomy (Neyman & Scott)[81] originally drew attention to the challenge of dealing with common and non-common parameters via parametric likelihood with a relatively small number of observations per non-common parameter. In particular they looked at pairs of observations. It may be important to keep in mind that meta-analyses in clinical research seldom, if ever, face the same degree of challenge as was faced in astronomy where a large number of very small studies were encountered, but the problems are still instructive. The most important lesson is perhaps that it is the number of studies that can be foreseen, not just the number in hand, that needs to be considered when evaluating methods of summarizing for future analysis and eventually undertaking that final analysis.

In two of the three problems Neyman & Scott addressed, there were repeated studies that all had two observations. In the first problem the mean was considered common and the variance non-common, and in the second the variance common and the mean non-common. For both, they assumed the observations were Normally distributed. In the first, the likelihood-based estimate is consistent but its asymptotic variance is not minimum (where the asymptotics fixes the number of observations per study and allows the number of studies to go to infinity), while in the second, the likelihood-based estimate is not even consistent. Various approaches have been offered to address the second situation but no approach is yet fully satisfactory for the first (which is also known as the Fisher-Berhans problem for common mean).

According to Barndorff-Nielsen and Cox[9], essentially the approaches to salvage the likelihood separate into two, one is to find an exact or approximate factorization of the likelihood so that one factor contains all or most of the information about the common parameter, sometimes utilizing conditional or marginal probability models and the second replaces the specification of arbitrary non-commonness of the non-common parameter with a common distribution for that parameter. A common parameter then resides in the marginal (over the non-common parameters) level 2 distribution and difficulties presented by having to (separately) estimate the non-common parameters disappear.

D.1 Example 1 - common mean, arbitrary variance

Quoting from Neyman and Scott

"Let α be some physical constant such as the radial velocity of a star or the velocity

of light. Assume that s series of measurements are to be made and let y_{ij} stand for the result of the j th measurement of the i th series $(i = 1, 2, ..., s; j = 1, 2, ..., n_i)$. We will assume that the measurements follow the normal law with the same mean α and an unknown standard error σ_i which may and probably does vary from one series of observations to another. Thus the probability density function of y_{ij} is

$$f\left(y_{ij}; \alpha, \sigma_i\right) = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-(y_{ij}-\alpha)^2/2\sigma_i^2}$$

This is exactly the case when α stands for the radial velocity of a star and the y_{ij} are its measurements obtained from n_i different spectral lines on the i th plate. ... This is also the situation in all cases where it is desired to combine measurements of physical quantities, made in different laboratories, by different experimenters, etc."

The log-likelihood in general is

$$-\frac{1}{2}\sum_i n_i \ \log(\sigma_i^2) - \frac{1}{2}\sum_i \sum_j (y_{ij} - \alpha)^2/\sigma_i^2$$

and the score function (differentiation of above with respect to α)

$$U_\alpha = \sum_i \sum_j (y_{ij} - \alpha)^2/\sigma_i^2 = \sum_i n_i \overline{y}_{i.} \ /\sigma_i^2 - \alpha \sum_i n_i/\sigma_i^2.$$

As is well known with σ_i^2replaced by $\widehat{\sigma}_{i\widehat{\alpha}}^2$ (the maximum likelihood estimate of σ_i^2 for $\widehat{\alpha}$ not $\widehat{\sigma}_{i\alpha}^2$- see below) the expectation of U_α is zero, so the estimated likelihood of $\alpha; L_p\left(\alpha; y_{ij}\right) = c(y_{ij}) f\left(y_{ij}; \alpha, \widehat{\sigma}_{i\widehat{\alpha}}^2\right)$ is consistent for α.

But if the σ_i^2 were known, $\sigma_i^2 = \sigma_{i0}^2$, say, the inverse variance weighted mean

$$\frac{\sum_i \overline{y}_{i.} \ (\sigma_{i0}^2/n_i)^{-1}}{\sum_i (\sigma_{i0}^2/n_i)^{-1}}$$

would be normally distributed with mean α and variance $1/\sum_i (\sigma_{i0}^2/n_i)^{-1}$. Now the estimated likelihood MLE for α is

$$\widehat{\alpha} = \frac{\sum_i \overline{y}_{i.} \ (\widehat{\sigma}_{i\widehat{\alpha}}^2/n_i)^{-1}}{\sum_i (\widehat{\sigma}_{i\widehat{\alpha}}^2/n_i)^{-1}}$$

where

$$\widehat{\sigma}_{i\widehat{\alpha}}^2 = \sum_j \frac{(y_{ij} - \widehat{\alpha})^2}{n_i} = \frac{\sum_j (y_{ij} - \overline{y}_{i.} \ )^2 + n_i (\overline{y}_{i.} \ - \widehat{\alpha})^2}{n_i}$$

has (Barndorff-Nielsen and Cox*[9]*) asymptotic (for fixed n_i and $s \to \infty$) normal distribution with

mean α and variance

$$\frac{\sum_i n_i/\{(n_i-2)\sigma_i^2\}}{(\sum_i n_i/\sigma_i^2)^2}$$

exceeding that of $1/\sum_i(\sigma_{i0}^2/n_i)^{-1}$ with σ_i^2 known. Additionally, different weights can result in smaller asymptotic variances with σ_i^2 unknown but it is unclear as to the best estimator for all σ_i^2.

D.2 Example 2 - common variance, arbitrary mean

This is the same set up as example 1, but now the precision of measurements does not change from one series to another yet the quantity measured does.

$$f\left(y_{ij};\alpha_i,\sigma^2\right)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(y_{ij}-\alpha_i)^2/2\sigma^2}.$$

With $n_i=2$ for all i the log-likelihood is

$$-\frac{1}{2}2s\ \log(\sigma^2)-\sum_i^s\left\{\frac{(y_{i1}-\alpha_i)^2+(y_{i2}-\alpha_i)^2}{2\sigma^2}\right\}$$

and the score function (differentiation of above with respect to σ^2)

$$U_t=-\frac{2s}{2\sigma^2}+\frac{1}{2(\sigma^2)^2}\sum_i^s\left\{(y_{i1}-\alpha_i)^2+(y_{i2}-\alpha_i)^2\right\}.$$

As is well known with α_i replaced by $\widehat{\alpha}_{i\sigma}$ (the maximum likelihood estimate of α_i for a given σ) the expectation of U_t is not zero but $-s/2\sigma^2$ so the profile likelihood of σ^2, $L_p\left(\sigma;y_{ij}\right)=c(y_{ij})f\left(y_{ij};\widehat{\alpha}_{i\sigma},\sigma\right)$, is not consistent for σ^2 (note that here $\widehat{\alpha}_{i\sigma}$ does not in fact depend on σ so that it is also the estimated likelihood).

D.3 Example 1 recast - common mean, common distribution of variance

With assumptions that σ_i^2 are independently inverse gamma distributed as

$$p_0(\sigma^2)=(\frac{1}{2}d_0\sigma_0^{2'})^{\frac{1}{2}d_0}(\sigma^2)^{-\frac{1}{2}d_0-1}e^{\left(-\frac{1}{2}d_0\sigma_0^{2'}/\sigma^2\right)}/\Gamma(\frac{1}{2}d_0)$$

where d_0 is an effective degrees of freedom and the "prior" mean is $\sigma_0^{2'} d_0/(d_0-2)$. For one sample of size r the likelihood would be

$$\int_0^\infty \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}r}} e^{(-\sum \frac{(y_i-\alpha)^2}{2\sigma^2})} p_0(\sigma^2) d\sigma^2.$$

The full log-likelihood is

$$\frac{1}{2}\sum_i (r_i + d_0) \log\left\{1 + \frac{r_i(\overline{y}_{i.} - \alpha)^2}{(y_{ij} - \overline{y}_{i.})^2 + d_0\sigma_0^{2'}}\right\}.$$

Note here that there are now just two parameters α and $\sigma_0^{2'}$ for all the observations.

D.4 Example 2 recast - common variance, common distribution of mean

With assumptions that α_i are independently normally distributed with mean ν and variance ω. The pairs $(y_{i1}, y_{i2})^T$ are now independently bivariate normal with mean $(\nu, \nu)^T$and covariance matrix

$$\begin{bmatrix} \omega + \sigma^2 & \omega \\ \omega & \omega + \sigma^2 \end{bmatrix}$$

It follows either via the bivariate normal form or by integrating the joint density with respect to the α_i that the log-likelihood is

$$-\frac{1}{2}s\log(\omega + \frac{1}{2}\sigma^2) - \frac{\sum_i(\overline{y}_{i.} - \nu)^2}{2\omega + \sigma^2} - \frac{1}{2}s\log(2\sigma^2) - \frac{\sum_i(y_{i2} - y_{i1})^2}{4\sigma^2}.$$

The maximum likelihood estimate of σ^2 is $\frac{1}{2}\sum_i(y_{i2} - y_{i1})^2/s$ unless

$$\frac{1}{2}\sum_i (y_{i2} - y_{i1})^2/s > 2\sum_i (\overline{y}_{i.} - \overline{y}_{..})^2/(s-1)$$

then it is

$$(\frac{1}{2}\sum_i (y_{i2} - y_{i1})^2 + 2\sum_i (\overline{y}_{i.} - \overline{y}_{..})^2)/(2s-1).$$

The complication arising because the parameter space is $\nu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+$, so that if

$$\frac{1}{2}\sum_i (y_{i2} - y_{i1})^2/s > 2\sum_i (\overline{y}_{i.} - \overline{y}_{..})^2/(s-1)$$

then the maximum likelihood is achieved on the boundary $\omega = 0$. Except for this complication, the usual properties of likelihood-based procedures hold[9] and note that only the parameters ω, σ^2 and ν are involved.

E A perhaps less familiar evasion of nuisance random effects

When the nuisance parameter is a "random sample" from some unspecified $p(\lambda_i^*; \lambda)$, the "likelihood curvature adjustment evasion" is when the likelihood combination is made as if all the $\lambda_i^* = \lambda$ but an adjustment is then made to the log-likelihood to make allowance for the assumption of $\lambda_i^* \sim p(\lambda_i^*; \lambda)$ actually being correct instead of $\lambda_i^* = \lambda$. At least, in the special case where the MLE of λ under the assumption of all $\lambda_i^* = \lambda$ is still considered relevant. This gives

$$\log(\textstyle\prod_i L(\lambda_i, y)) \equiv \quad \log(\prod_i L(\lambda, y))c$$

or more conveniently in terms of log-likelihoods

$$\textstyle\sum_i l(\lambda_i, y) \equiv \quad c\sum_i l(\lambda, y)$$

where c is chosen so that the usual likelihood-based variance estimate

$$-\left[\frac{\partial^2 c\sum_i l(\lambda, y)}{\partial \lambda^2}\bigg|_{\mu=\hat{\mu}}\right]^{-1}$$

provides an asymptotically consistent estimate.

The background details that motivate this adjustment to the log-likelihood are set out here. Assuming that $y_1, y_2, ..., y_n$ are i.i.d. random variables with common density function $f(y;\theta)$ where θ is a real valued parameter, define $l(y;\theta) = \log f(y;\theta)$ and let $l^{'}(y;\theta), l^{''}(y;\theta)$, and $l^{'''}(y;\theta)$ be the first three derivatives of $l(y;\theta)$ with respect to θ. The following assumptions will be made about $f(y;\theta)$:

(A1) The parameter space Θ is an open subset of the real-line.

(A2) The set $A = \{y : f(y;\theta) > 0\}$ does not depend on θ.

(A3) $f(y;\theta)$ is three times continuously differential with respect to θ for all y in A.

(A4) $E_\theta[l^{'}(Y_i;\theta)] = 0$ for all θ and $Var_\theta[l^{'}(Y_i;\theta)] = I(\theta)$ where $0 < I(\theta) < \infty$ for all θ.

(A5) $E_\theta[l^{''}(Y_i;\theta)] = -J(\theta)$ where $0 < J(\theta) < \infty$ for all θ.

(A6) For each θ and for $\delta > 0, |l^{'''}(t;\theta)| \leq M(y)$ for $|\theta - t| \leq \delta$ where $E_\theta[M(Y_i)] < \infty$.

Note that by condition (A2)

$$\int_A f(y;\theta)dy = 1 \text{ for all } \theta \in \Theta$$

and so

$$\frac{d}{d\theta}\int_A f(y;\theta)dy = 0.$$

If the derivative can be taken inside the integral then

$$\begin{aligned} 0 &= \int_A \frac{d}{d\theta} f(y;\theta)dy \\ &= \int_A l^{'}(y;\theta)f(y;\theta)dy \\ &= E_\theta[l^{'}(Y_i;\theta)]. \end{aligned}$$

Moreover, if $\int_A f(y;\theta)dy$ can be differentiated twice inside the integral sign,

$$\begin{aligned} 0 &= \int_A \frac{d}{d\theta}(l^{'}(y;\theta)f(y;\theta))dy \\ &= \int_A l^{''}(y;\theta)f(y;\theta)dy + \int_A (l^{'}(y;\theta))^2 f(y;\theta)dy \\ &= -J(\theta) + I(\theta) \end{aligned}$$

and so $J(\theta) = I(\theta)$.

From standard results, as for instance, on page 248 of Knight[71]

$$\sqrt{n}(\widehat{\theta} - \theta) \rightarrow_d \frac{Z}{J(\theta)} \sim N(0, Var_\theta[l^{'}(Y_i;\theta)]/(E_\theta[l^{''}(Y_i;\theta)])^2)$$

and since given the above assumptions $Var_\theta[l^{'}(Y_i;\theta)] = -E_\theta[l^{''}(Y_i;\theta)]$ this is simply

$$\sqrt{n}(\widehat{\theta} - \theta) \rightarrow_d \frac{Z}{J(\theta)} \sim N(0, 1/-E_\theta[l^{''}(Y_i;\theta)]).$$

But especially when what is common is only common in distribution one will want to know what happens for mis-specified models. For mis-specified models consider the functional parameter $\theta(F)$ defined by $\int_{-\infty}^{+\infty} l^{'}(y;\theta(F))dF(y) = 0$, does $\sqrt{n}(\widehat{\theta} - \theta(F)) \rightarrow_d N(0, \frac{\int_{-\infty}^{+\infty}[l^{'}(y;\theta(F))]^2 dF(y)}{\left(\int_{-\infty}^{+\infty} l^{''}(y;\theta(F))dF(y)\right)^2})$?

Given $\widehat{\theta}$ satisfies the "estimating equation" $\sum_i^n l^{'}(y_i;\widehat{\theta}) = 0$ and

a) $l^{'}(y;\theta)$ is a strictly decreasing (or increasing) function of θ (over the open set Θ) for each y,

b) $\int_{-\infty}^{+\infty} l^{'}(y;\theta)dF(y) = 0$ has a unique solution $\theta = \theta(F)$ where $\theta(F) \in \Theta$,

c) $I(F) = \int_{-\infty}^{+\infty}[l^{'}(y;\theta(F))]^2 dF(y) < \infty$

d) $J(F) = -\int_{-\infty}^{+\infty} l^{''}(y;\theta(F))dF(y) < \infty$

e) $|l^{'''}(y;t)| \leq M(y)$ for $\theta(F) - \delta \leq t \leq \theta(F) + \delta$ and some $\delta > 0$ where $\int_{-\infty}^{+\infty} M(y)dF(y) < \infty$

then $\widehat{\theta} \rightarrow_p \theta(F)$ and

$\sqrt{n}(\widehat{\theta} - \theta(F)) \rightarrow_d Z \sim N(0, I(F)/J^2(F))$

This suggests $\frac{\sum_i^n [l^{'}(y_i;\hat{\theta})]^2}{[-\sum_i^n l^{''}(y_i;\hat{\theta})]^2}$ rather than $\frac{1}{\sum_i^n [l^{'}(y_i;\hat{\theta})]^2}$ or $\frac{1}{-\sum_i^n l^{''}(y_i;\hat{\theta})}$ as an estimate of the variance of $\hat{\theta}$ in mis-specified models.

Now if one multiplied the log-likelihood by $\frac{-\sum_i^n l^{''}(y_i;\hat{\theta})}{\sum_i^n [l^{'}(y_i;\hat{\theta})]^2}$ i.e.

$$l_a(y;\theta) = \frac{-\sum_i^n l^{''}(y_i;\hat{\theta})}{\sum_i^n [l^{'}(y_i;\hat{\theta})]^2} l(y;\theta)$$

and then calculated $\sum_i^n l_a^{''}(y_i;\hat{\theta})$ it would equal

$$\begin{aligned} & \frac{-\sum_i^n l^{''}(y_i;\hat{\theta})}{\sum_i^n [l^{'}(y_i;\hat{\theta})]^2} * (\sum_i^n l^{''}(y_i;\hat{\theta})) \\ = \quad & \frac{-[\sum_i^n l^{''}(y_i;\hat{\theta})]^2}{\sum_i^n [l^{'}(y_i;\hat{\theta})]^2} \end{aligned}$$

so that a correction to the quadratic term has been made and the usual observed information estimate from $1/-l_a^{''}(y_i;\hat{\theta})$ would provide $\frac{\sum_i^n [l^{'}(y_i;\hat{\theta})]^2}{[-\sum_i^n l^{''}(y_i;\hat{\theta})]^2}$ as the variance estimate. Stafford[111] suggests just this but for $l_p(y;\theta) = l(y;\theta,\hat{\lambda}_\theta)$ - the profile log-likelihood - and establishes the asymptotic consistency of it for this, as well as its invariance under interest respecting transformations (i.e. of the form $\{\theta,\lambda\} \rightarrow \{g(\theta), h(\lambda,\theta)\}$).

Originally in this thesis, this was considered a particularly attractive evasion of random effects. It has a long history and many variants and was suggested as the "least wrong" random effects approach for meta-analysis in O'Rourke[83]. In this evasion the MLE of the common parameter is not affected - just the estimate of its variance. Quoting from McCullagh and Nelder[79] about a variant of this (page 126)

> "the mean is unaffected but the variance is inflated by an unknown factor σ^2."

Tjur[118] recently recast the issue as

> "Can we find a class of statistical models that extends the original generalized linear model by some scale parameter, in such a way that:
>
> 1. the original generalized linear model comes out as a special case when the scale parameter is fixed and equal to 1;
>
> 2. the maximum likelihood estimates in the model for the original "link-function-linear" parameters coincide with those of the original model (ignoring the overdispersion).
>
> The answer to this question is no ... [nothing so far] answers the fundamental question whether there is a way of modifying conditions (1) and (2) above in such a way that a meaningful theory of generalized linear models with overdispersion comes out as the unique answer."

McCullagh[78] responded to this suggestion which Tjur had based on an analogy to McCullagh's earlier work on Ordinal Logistic Regression[77], dismissing (2) above by pointing out it is the parameter that should not change when the model is extended, and not its MLE. As was argued earlier, random effects models change the MLEs and in meta-analysis applications they are arguably changed for the worse. In other words, it is not that the MLEs should not change with a random effects model but that they should not be made "worse."

Unfortunately, the adjustment has been found to have very poor properties when there are unknown scale parameters and this would preclude its use more generally. As an alternative, numerical integration techniques have been developed so that at least a range of random effects models can be used instead of just particularly convenient ones. There are also non-parametric approaches to random effects models but they are also problematic for most meta-analysis applications given the usually small number of studies involved.

F Other statistical issues and techniques

F.1 Construction of confidence and credible intervals

Some writers do suggest that likelihoods are all that are necessary as outputs from statistical analyses and some further argue that they should be the only outputs from statistical analyses[101]. These later claims are certainly controversial and also perhaps a distraction from the perspective in this thesis – here multiplied or combined likelihoods are simply used as a means to get confidence intervals for frequency based inference or to be combined with prior information to get posterior probabilities for Bayesian inference. Both Bayesian and classical approaches need to do something

to the likelihood to get credible regions or confidence regions and in turn credible intervals or confidence intervals for a single parameter of interest. A quick discussion on how to first get regions and then intervals, is given now.

Bayesian credible regions could be indirectly characterized as likelihood regions where the probability content is determined by the integral of the posterior probabilities over the region defined by the likelihood region -

$\Pr(\theta \in [\theta_l, \theta_u]) \propto \int_{\theta_l}^{\theta_u} \pi(\theta|obs_i)$, where θ_l and θ_u are determined by relative likelihood values.

Bayes theorem only stipulates that the posterior probabilities are equal to likelihood times prior (times an arbitrary constant so that posterior probabilities add to one)

$$\pi(\theta|obs_i) \propto f(\theta, obs_i) * \pi(\theta) = f(obs_i|\theta) * \pi(\theta),$$

not how particular regions are to be determined. In practice, likelihood regions are not first formed and then probability content determined for them, but this is not an essential difference, as the posterior probability is in principle calculable for any chosen region. The likelihood region's calibration of probability content given the prior would give a credible region with boundaries that may not be the conventional ones that a Bayesian approach would use, but it would be an allowable one. (Care is needed though with parameters that have a common distribution - as these might or might not be integrated out of the likelihood regions before constructing credible regions and this might matter. This issue is addressed below.)

As for confidence regions, most often in meta-analysis, these are simply likelihood regions for the "right" choice of a threshold – i.e. by calibrating the likelihood region by choosing a different threshold value so that under an "appropriate" sampling model this new region has approximately the correct coverage while retaining the same shape.[9][30] In fact, with multiple randomized studies in clinical research, it is almost always the case that profile likelihood ratio based regions have coverage properties consistent with that suggested by the Chi-square approximation of the drop in log-likelihood page 243 of Barndorff-Nielsen[8] and this thesis will only attempt to get confidence regions (and confidence intervals) in such cases (the purpose of the simulated modified profile likelihood is to identify when this is not the case).

To more fully clarify the essential difference between a Bayesian and a Classical parametric

likelihood approach though, it might be helpful to characterize the Classical approach as involving only one combination, and the Bayesian approach as involving two combinations. In the Classical approach the only combination is the combination of data (individual observations and studies) and that is accomplished by likelihood multiplication. In the Bayesian approach, the combination of observations is also accomplished exactly the same way by likelihood multiplication, but there is a second combination that involves combining probabilities which is accomplished by Bayes theorem.

Here the probabilities of possible values of the parameter (prior) are combined with the probability of the data, given various values the parameter (likelihood), resulting in the (posterior) probabilities of possible values of the parameter (given the data and the prior) i.e.

$$\pi(\theta|obs_i) \propto f(\theta, obs_i) * \pi(\theta) = f(obs_i|\theta) * \pi(\theta).$$

Quoting Sprott*[109]* -

> "Bayes' theorem combines prior information about θ in the form of a probability distribution with the current experimental information in the form of the likelihood, to obtain the posterior distribution containing all the information. Bayes' theorem is merely a statement of the multiplication rules for combining probabilities." If the true unknown value of the parameter under which the observations actually in hand were generated, can be thought of as being random sample from a known probability distribution, there is nothing controversial about this "multiplication to achieve the combination of probabilities"

– it is just arithmetic or probability calculus.

There are other ways to characterize and interpret the Bayesian approach. The characterization proposed in this thesis puts Bayes theorem in a combining framework. Sprott for instance, has used this characterization to call attention to the need to check that the information from two sources that is to be combined is mutually consistent – to avoid the combining apples and oranges danger in a naive Bayesian approach where "good" prior information may be inadvertently ruined when combined with "bad" likelihood sample information or vice versa.

In applications, an interval for a single parameter of interest will usually be required. This will necessitate an "extraction" from a higher dimensional region and the Bayesian and Classical approaches differ in how this extraction is accomplished. The Bayesian approach usually integrates out other parameters from the credible region to get credible intervals for individual parameters of interest. Since a prior distribution is required to get a credible region that results in a credible

interval when appropriately integrated – this extraction of a credible interval from a credible region is considered to be simply the result of the combining of likelihood and prior. Note, though that parameters that have a common distribution - i.e. random parameters - might or might not be integrated out of the likelihood regions before constructing credible regions - according to whether or not one accepts that this should be done[12][13]. We now outline the conditions for when this does not matter - i.e. that the same posterior intervals result from either approach.

Recall the uniform integrated likelihood (over nuisance parameter λ) is

$$L^U(\theta) = \int L(\theta, \lambda) d\lambda$$

where θ is the parameter of interest. More generally the integrated likelihood is

$$L(\theta) = \int L(\theta, \lambda)\pi(\lambda|\theta) d\lambda$$

where $\pi(\lambda|\theta)$ is the weight function or conditional prior density of λ given θ, and the general likelihood is

$$f(y, \theta^*, \lambda^*|\theta, \lambda), \text{ where } \theta^*, \lambda^* \text{ are unobserved parameters.}$$

Here it is claimed to be noncontroversial to use $f(y|\theta) = \int f(y, \lambda^*|\theta) d\lambda^*$or more generally noncontroversial to use

$$f(y, \theta^*|\theta, \lambda) = \int f(y, \theta^*, \lambda^*|\theta, \lambda) d\lambda^*.$$

But the subjective Bayesian will base analysis an a full prior

$$\pi^B(\theta, \lambda) = \pi^B(\theta)\pi^B(\lambda|\theta)$$

and they will seek $\pi(\theta|y) \propto L(\theta)\pi^B(\theta)$ and would accept the integrated likelihood $L(\theta)$ if

$$\pi(\theta|y) \propto L(\theta)\pi^B(\theta).$$

It is "easy to see" that the only $L(\theta)$ which satisfies this relationship is given up to a multiplicative constant by

$$L^B(\theta) = \int f(y|\theta, \lambda)\pi^B(\lambda|\theta) d\lambda.$$

Now as

$$\pi^B(\theta \mid y) = \int L(\theta, \lambda) * \pi^B(\theta)\pi^B(\lambda \mid \theta)d\lambda$$

and integrated likelihood is

$$L(\theta) = \int L(\theta, \lambda)\pi(\lambda \mid \theta)d\lambda$$

then

$$\pi^B(\theta \mid y) \propto L(\theta)\pi^B(\theta)$$

iff

$$\pi(\lambda \mid \theta) = \pi^B(\lambda \mid \theta)$$

while almost always in meta-analysis $\pi^B(\lambda|\theta) = 1$. (It would be unusual for anyone under the belief that an unobserved parameter λ^* was randomly sampled from a known distribution - i.e. from $N(\xi, \tau^2)$ given the value of $\theta = \left(\xi, \tau^2\right)$ - to specify any further information about it.)

The Classical approach on the other hand, extracts a likelihood interval from the likelihood region and then calibrates that interval. A general approach that has been adopted in this thesis, is to maximize out the unknown parameters using profile likelihood where, for each value of the parameter of interest, the other unknown parameters are replaced by their maximum likelihood estimates given that value of the interest parameter. In meta-analysis practice, the combined credible intervals are often very similar numerically to the combined confidence intervals, as for instance, was the case in Warn, Thompson and Spiegelhalter [121].

F.2 Sensitivity analyses for possibly informative choice of reported summary

Dawid's[33] approach is presented here for possibly informative choices of a reported summary as distinct from the approach in this thesis which assumed the summary would always be chosen. Using Dawid's notation, let $\{S_n(y)\}$ be a finite set of possible summaries, one of which is reported. Recall, it has been shown that the observed summary likelihood - if that summary is always reported is

$$c(\overline{S_n(y)}) * \int_{y:S_n(y)=\overline{S_n(y)}} f(y|\theta)dy \text{ where } \overline{S_n(y)} \text{ is the value reported.}$$

The alternative formulation of the likelihood-based on the distribution of the summary reported and the fact that the summary was chosen to be reported on the basis of $\overline{S_n(y)}$ and θ is

$$p(R_n|S_n(y) = \overline{S_n(y)}, \theta) * L(S_n(y)|\theta).$$

This can be rewritten as

$$p(R_n|S_n(y) = \overline{S_n(y)}, \theta) * c(\overline{S_n(y)}) * \int_{y:S_n(y)=\overline{S_n(y)}} f(y|\theta)dy.$$

There might be non-data influences of θ on the choice of what is reported. Non-data influences on the choice of reported summary could arise easily from data from previous trials. These would require careful and specific thought - including the possibility of interaction between within and outside trial data.

F.3 Obtaining valid envelope numerical integration bounds

In order to obtain valid bounds on an integral using the envelope rules from Evans and Swartz[45], these rules must be applied separately within each region of concavity for $f^{(n)}$ where f is the desired integrand. Accuracy of the rule is improved with increasing n, but also by simply compounding a given rule within the regions of concavity. In this thesis only one dimensional integrals were dealt with so only intervals of concavity needed to be dealt with. The pragmatic choice of n versus the degree of compounding for the desired accuracy of the bounds may vary by integrand. For the real meta-analysis examples in this thesis we found it was usually best to set $n = 0$ and depend on compounding to get the desired accuracy.

Application of the rules, (i.e. the calculation of $\sum_{k=0}^{n} \frac{f^{(k)}(a)}{(k+1)!}(b-a)^{k+1} + \frac{f^{(n)}(b)-f^{(n)}(a)}{(b-a)} \frac{(b-a)^{n+2}}{(n+2)!}$ and $\sum_{k=0}^{n+1} \frac{f^{(k)}(a)}{(k+1)!}(b-a)^{k+1}$) involves the first $n+1$ derivatives of the integrand and the determination of the intervals of concavity involves sign changes of the $n+2$ derivative. The validity of the bounds from the rules involves the ruling out of simple roots of the $n+2$ derivative within any of the intervals in which the rules are applied. Calculation of symbolical derivatives is straightforward in the Mathematica package.

Now the integral equation that counts the number of simple roots of $f^{(n+2)}$ in a given interval (a, b) is $(-\frac{1}{\pi}[\gamma \int_a^b \frac{f^{(n+2)}(x)f^{(n+4)}(x)-f^{(n+3)}(x)^2}{f^{(n+2)}(x)^2+\gamma^2 f^{(n+3)}(x)^2}dx + \arctan(\frac{\gamma f^{(n+3)}(b)}{f^{(n+2)}(b)}) - \arctan(\frac{\gamma f^{(n+3)}(a)}{f^{(n+2)}(a)})])$ and usu-

ally needs to be evaluated by numerical integration. The intervals can recursively be sub-divided to help enable this. Again there is no need to know any of these roots per se - just the need to ensure no roots persist inside any interval in a given set of disjoint intervals that covers the full interval of integration. Thus the strategy used in this thesis was to recursively partition the interval of integration until a disjoint set of intervals resulted, all of which had no simple roots as determined by numerical integration. Mathematica has a FixedPoint function which facilitates the implementation of recursion until this has been achieved - i.e. no intervals were split in the last iteration. The standard Mathematica NIntegrate function was used to carry out the numerical integrations. For some problems, more specialized methods of integration may be more efficient or even in some cases necessary and remains future research.

The rules for splitting the intervals involved two cases - intervals with more than 1 simple root or that were not numerically integrable (with default settings) were simply split in half, while intervals with 1 simple root were searched for the root using the Mathematica function Root and the interval split at the root found. This needed to be modified so that when a root was found outside the interval, that interval would simply be split in half and the process repeated. In particular, this avoided roots that gave parameter values outside the permitted range for the underlying probability distributions. Numerical separation of intervals needed to be assured around simple roots when they were located and this was achieved by the use of a small constant - usually of order 10^{-6}. Additionally, as the γ in the integral equation is an arbitrary small positive constant, some checking that the chosen one is small enough is required and it needed to be changed in a couple of examples that were undertaken.

Application of the rules to intervals with an infinite endpoint can be challenging but were simply dealt with by truncation in this thesis. Strictly speaking, the envelope bounds are only truly valid for a truncated random effects model. Desired accuracy of the bounds - given an initially constructed set of disjoint intervals without any simple roots - was simply achieved by recursively splitting each interval in half. There are obvious ways in which this could be made adaptive, but remains future work.