Statistical Methods for Combining Results of Independent Studies

Habilitationsschrift

Guido Knapp

Fakultät Statistik
Technische Universität Dortmund

Dortmund, im Juli 2008

Contents

Chapter 1

Introduction

The term *meta-analysis* was coined by Glass (1976) in the social sciences and Glass defined meta-analysis as 'the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings'. Beside the social sciences, meta-analysis is nowadays widely accepted and applied in the life sciences. Following Draper et al. (1992), there are a lot of other fields in which statistical methods for meta-analysis are applied, for instance, archaeology, astronomy, chemistry, engineering, environmental sciences, geosciences, military operations analysis, official statistics, physics, and psychology.

Combining results from independent studies has a long history in statistics, though the term meta-analysis was only coined around thirty years ago. As an early application in biometry, Pearson (1904) used data from five small independent samples and computed a pooled estimate of correlation between mortality and inoculation with a vaccine for enteric fever in order to evaluate the efficacy of the vaccine. In the physical sciences, Birge (1932) combined estimates across experiments at different laboratories to establish reference values for some fundamental constants in physics. Early works of Cochran (1937), Yates and Cochran (1938), Tippett (1931), and Fisher (1932) dealt with combining information in the agricultural sciences in order to derive estimates of treatment effects and test their significance.

As the scope of meta-analysis grew over the years, several terminologies also came into existence, such as combining experiments, combination of information, combination of results, systematic review, quantitative research synthesis, research integration, or pooling evidence. The basic statistical methods behind these various terms, however, are all the same and we will always use the term meta-analysis in the following.

Meta-analysis can be seen as a process which consists of four important stages of research synthesis: problem formulation, data collection, data evaluation, and data analysis and interpretation, see the introduction in Hartung, Knapp, and Sinha (2008) for a detailed description of these stages. The main focus of this thesis is on the data analysis stage, that is, given the results of the independent studies we deal with the problem how to combine these results using sound statistical methods. Several text books on statistical methods of meta-analysis which merely deal with this data analysis stage are nowadays available, notably Hedges and Olkin (1985), the edited volume by Cooper and Hedges (1994), Whitehead (2002), and Hartung, Knapp, and Sinha (2008).

The emphasis of the present thesis is on statistical methods for combining results when only published data from the individual studies are available. This is the scenario Glass (1976) had in mind defining the term meta-analysis and this is still the most common situation in research. Individual data from all the studies could clearly improve the findings from a meta-analysis, but in practice it is usually very difficult, if not impossible, to get all the data from the different experiments.

The experiments or studies we are interested in are comparative studies, that is, studies in which a hypothesis is tested comparing a new intervention or treatment with a standard intervention or control. The difference or the association between the two counterparts can be modelled using a single parameter, we generally will call effect size in the following. Possible effect sizes are difference of normal means, standardized mean difference, risk difference, or odds ratio. The data situation for the meta-analysis is then that estimates of the effect size of interest are available from each study as well as estimates of the precision of each study-specific effect size estimate.

The foundation of the statistical methods in meta-analysis stems from the comparison of several normal populations. Assuming a common mean in all the normal populations, but possibly unequal variances, statistical inference about this common mean is not trivial and has attracted a lot of researchers in the last decades. Chapter 2 contains many results for this common mean problem but the presentation is restricted to those results which can be extended to the meta-analysis for effect sizes. The statistical methods presented in this chapter build the foundation of the so-called *fixed effects model of meta-analysis*.

In case the means of the several populations are possibly unequal, but vary about an overall mean, statistical inference about this overall mean in the one-way random effects model of analysis of variance with possibly heterogeneous error variances is an appropriate tool. Chapter 3 contains many results for the statistical inference in this model, but again the presentation is restricted to those results which can also be used or easily extended for combining results of comparative studies. The statistical methods presented in this chapter build the foundation of the so-called *random effects model of meta-analysis*.

Chapter 4 is devoted to the combination of results from comparative studies with normal outcomes. We discuss the meta-analytical techniques for the effect sizes difference of normal means, standardized difference of normal means, and ratio of normal means when the difference of two populations is of interest.

The meta-analysis of comparative studies with binary outcomes is discussed in Chapter 5. The effect sizes considered are difference of probabilities, also sometimes called risk difference, (logarithmic) relative risk, and (logarithmic) odds ratio. Beside the general meta-analysis methods, meta-analysis methods for sparse data with binary outcomes are stressed that can lead to some additional difficulties.

A crucial decision in meta-analysis is whether ong should use the fixed effects or the random effects meta-analysis model. When using a random effects model, explaining heterogeneity is a further important task in meta-analysis. From a statistical point of view, one can use study-specific covariates in regression models to explore possible sources of heterogeneity. The analysis in this type of regression models, briefly called *meta-regression*, is the topic of Chapter 6.

Most of the presented meta-analysis methods are based on the so-called frequentist approach. Bayesian methods heavily rely on informative prior distributions on the parameters. Using non-informative priors, results of meta-analysis are nearly identical in both approaches, frequentist and Bayesian approach. Moreover, the appropriate choice of prior distributions depends on the actual problem at hand. Thus, we present ideas of Bayesian methods when appropriate, but do not provide details on the Bayesian analysis.

Chapter 2

The Common Mean Problem

Let us consider k independent normal populations, where the ith population follows a normal distribution with mean $\mu \in I\!R$ and variance $\sigma_i^2 > 0$, $i = 1, \ldots, k$. Let $\bar{Y}_i$ denote the sample mean in the ith population, S_i^2 the sample variance, and n_i the sample size, $i = 1, \ldots, k$. Then, we have

$$\bar{Y}_i \sim N\left(\mu\ ,\ \frac{\sigma_i^2}{n_i}\right) \quad \text{and} \quad \frac{(n_i - 1)\ S_i^2}{\sigma_i^2} \sim \chi^2_{n_i - 1}, \qquad i = 1, \ldots, k, \tag{2.1}$$

and the statistics are all mutually independent. Note that $(\bar{Y}_i, S_i^2, i = 1, \ldots, k)$ is minimal sufficient for $(\mu, \sigma_1^2, \ldots, \sigma_k^2)$ even though it is not complete.

If the population variances $\sigma_1^2, \ldots, \sigma_k^2$ are completely known, the maximum likelihood estimator of μ is given by

$$\hat{\mu} = \frac{\sum_{i=1}^k n_i\ \bar{Y}_i/\sigma_i^2}{\sum_{j=1}^k n_j/\sigma_j^2}. \tag{2.2}$$

The estimator (2.2) is also the minimum variance unbiased estimator under normality as well as the best linear unbiased estimator without normality for estimating μ. The variance of $\hat{\mu}$ is given by

$$\mathrm{Var}\,(\hat{\mu}) = \frac{1}{\sum_{i=1}^k n_i/\sigma_i^2}. \tag{2.3}$$

If the population variances $\sigma_1^2, \ldots, \sigma_k^2$ are completely unknown, the log-likelihood func-

tion of the minimal sufficient statistics $(\bar{Y}_i, S_i^2, i = 1, \ldots, k)$ is

$$L^* = \sum_{i=1}^{k} \left[\text{constant} - \frac{n_i}{2} \ln(\sigma_i^2) - \frac{(n_i - 1)\ S_i^2 + n_i \left(\bar{Y}_i - \mu\right)^2}{2\sigma_i^2} \right]. \tag{2.4}$$

Differentiations of L^* w.r.t to μ, $\sigma_1^2, \ldots, \sigma_k^2$ and setting the derivatives equal to zero yield the maximum likelihood estimators $\hat{\mu}_{ML}$ and $\hat{\sigma}_{i(ML)}^2$, $i = 1, \ldots, k$, which must satisfy

$$\hat{\sigma}_{i(ML)}^2 = \frac{(n_i - 1)S_i^2}{n_i} + \left(\bar{Y}_i - \hat{\mu}_{ML}\right)^2 \ , \ i = 1, \ldots, k, \tag{2.5}$$

and

$$\hat{\mu}_{ML} = \frac{\sum_{i=1}^{k} n_i\ \bar{Y}_i / \hat{\sigma}_{i(ML)}^2}{\sum_{j=1}^{k} n_j / \hat{\sigma}_{j(ML)}^2}. \tag{2.6}$$

Clearly, the maximum likelihood estimator (MLE) of μ in Eq. (2.6) does not have a closed form and has to be found numerically.

The literature has not paid much attention to likelihood methods in the common mean problem since Cochran's (1937) seminal paper. Cochran (1937) considered experiments with equal sample sizes and recommended the use of a weighted mean statistic, which is nowadays known as the Graybill-Deal estimator, see Eq. (2.7) below, if at least 15 degrees of freedom are available in S_i^2. With fewer than 15 degrees of freedom, Cochran (1937) preferred the maximum likelihood estimator since its increased precision is well worth the extra labor it involves.

In the Behrens-Fisher problem ($k = 2$ populations), Suguira and Gupta (1987) showed that the likelihood equation for estimating the common mean has either a unique solution with large probability or three solutions with small probability. When it has three solutions, the maximum likelihood estimator of the common mean is given by either minimum or maximum real root of a cubic equation, and when it has a unique solution, it is just the maximum likelihood estimator. This shows that one should be careful in obtaining maximum likelihood estimates by numerical iterations.

Recently, Pal et al. (2007) also considered maximum likelihood estimation of the common mean in case of $k = 2$ populations. They showed that the maximum likelihood estimator of μ is unbiased and, via simulation study, compared the variance of the MLE

of μ with the variance of the Graybill-Deal estimator. The finding of their simulation study is that the MLE of μ has better overall performance than the Graybill-Deal estimator. Maybe, this work will stimulate future research on likelihood methods in the common mean problem.

An estimator of the common mean given in a closed form can be obtained by replacing σ_i^2 by S_i^2 in Eq. (2.2). This yields the already mentioned well-known Graybill-Deal (1959) estimator given as

$$\hat{\mu}_{\mathrm{GD}} = \frac{\sum_{i=1}^{k} n_i \ \bar{Y}_i/S_i^2}{\sum_{j=1}^{k} n_j/S_j^2}. \tag{2.7}$$

Clearly, $\hat{\mu}_{\mathrm{GD}}$ is an unbiased estimator of the common mean μ for the statistics $\bar{Y}_i$ and S_i^2, $i = 1, \ldots, k$, are stochastically independent.

For calculating the variance of $\hat{\mu}_{\mathrm{GD}}$, a standard conditional argument first yields

$$\begin{aligned} \mathrm{Var}\left(\hat{\mu}_{\mathrm{GD}}\right) &= \mathrm{E}\left[\mathrm{Var}\left(\hat{\mu}_{\mathrm{GD}}|S_1, \ldots, S_k\right)\right] + \mathrm{Var}\left[\mathrm{E}\left(\hat{\mu}_{\mathrm{GD}}|S_1, \ldots, S_k\right)\right] \\ &= \mathrm{E}\left[\left(\sum_{i=1}^{k} \frac{n_i \ \sigma_i^2}{S_i^4}\right) \Big/ \left(\sum_{i=1}^{k} \frac{n_i}{S_i^2}\right)^2\right]. \end{aligned} \tag{2.8}$$

Meier (1953) derived a first order approximation of the variance of $\hat{\mu}_{\mathrm{GD}}$ as

$$\mathrm{Var}\left(\hat{\mu}_{\mathrm{GD}}\right) = \frac{1}{\sum_{i=1}^{k} n_i/\sigma_i^2}\left[1 + 2\sum_{i=1}^{k} \frac{1}{n_i - 1}\ c_i\ (1 - c_i) + O\left(\sum_{i=1}^{k} \frac{1}{(n_i - 1)^2}\right)\right] \tag{2.9}$$

with

$$c_i = \frac{n_i \ / \ \sigma_i^2}{\sum_{j=1}^{k} n_j \ / \ \sigma_j^2}, \quad i = 1, \ldots, k.$$

Since $\hat{\mu}_{\mathrm{GD}}$ uses sufficient statistics, the question naturally arises whether $\hat{\mu}_{\mathrm{GD}}$ is a uniformly better unbiased estimator of μ than is each $\bar{Y}_i$, $i = 1, \ldots, k$, that is, $\mathrm{Var}(\hat{\mu}_{\mathrm{GD}}) \leq \sigma_i^2/n_i$, $i = 1, \ldots, k$ for all $\sigma_1^2, \ldots, \sigma_k^2$. In case of $k = 2$ populations, Graybill and Deal (1959) showed that

$$\frac{n_1 \ \bar{Y}_1/S_1^2 + n_2 \ \bar{Y}_2/S_2^2}{n_1/S_1^2 + n_2/S_2^2}$$

is a uniformly better unbiased estimator of μ than is $\bar{Y}_1$ or $\bar{Y}_2$ if and only if n_1 and n_2 are each greater than 10. Norwood and Hinkelmann (1977) extended this result for $k > 2$

populations and showed that $\hat{\mu}_{\mathrm{GD}}$ is a uniformly better estimator of μ than each $\bar{Y}_i$ if and only if each sample size n_i, $i = 1, \ldots, k$, is greater than 10 or $n_i = 10$ for some i and n_j greater than 18 for all $j \neq i$.

For further statistical inference on the common mean, an estimator of the variance of $\hat{\mu}_{\mathrm{GD}}$ should be available. Sinha (1985) derived an unbiased estimator of the variance of $\hat{\mu}_{\mathrm{GD}}$ that is a convergent series. A first order approximation of this estimator is

$$\widehat{\mathrm{Var}}_{(1)}\left(\hat{\mu}_{\mathrm{GD}}\right) = \frac{1}{\sum_{i=1}^{k} n_i/S_i^2} \left[1 + \sum_{i=1}^{k} \frac{4}{n_i + 1} \left(\frac{n_i \;/\; S_i^2}{\sum_{j=1}^{k} n_j \;/\; S_j^2} - \frac{n_i^2 \;/\; S_i^4}{\left(\sum_{j=1}^{k} n_j \;/\; S_j^2\right)^2} \right) \right]. \quad (2.10)$$

This estimator is comparable to the approximate estimator

$$\widehat{\mathrm{Var}}_{(2)}\left(\hat{\mu}_{\mathrm{GD}}\right) = \frac{1}{\sum_{i=1}^{k} n_i/S_i^2} \left[1 + \sum_{i=1}^{k} \frac{4}{n_i - 1} \left(\frac{n_i \;/\; S_i^2}{\sum_{j=1}^{k} n_j \;/\; S_j^2} - \frac{n_i^2 \;/\; S_i^4}{\left(\sum_{j=1}^{k} n_j \;/\; S_j^2\right)^2} \right) \right] \quad (2.11)$$

due to Meier (1953).

In view of generalizing results from this chapter to comparative experiments with possibly non-normal outcomes in later chapters, we present two further estimators of the variance of $\hat{\mu}_{\mathrm{GD}}$ which can be easily adapted for later purposes. One rough estimator of the variance of $\hat{\mu}_{\mathrm{GD}}$ is given by simply replacing σ_i^2 by S_i^2 in Eq. (2.3), that is,

$$\widehat{\mathrm{Var}}_{(3)}\left(\hat{\mu}_{\mathrm{GD}}\right) = \frac{1}{\sum_{i=1}^{k} n_i/S_i^2}. \quad (2.12)$$

Another estimator of the variance of $\hat{\mu}_{\mathrm{GD}}$ is based on a direct estimator of the variance (2.3). An unbiased estimator of the variance (2.3), assuming completely known variances $\sigma_1^2, \ldots, \sigma_k^2$, is given by

$$\widehat{\mathrm{Var}}(\hat{\mu}) = \frac{1}{k-1} \sum_{i=1}^{k} \frac{n_i \;/\; \sigma_i^2}{\sum_{j=1}^{k} n_j \;/\; \sigma_j^2} \left(\bar{Y}_i - \hat{\mu}\right)^2 \quad (2.13)$$

with $\hat{\mu}$ from Eq. (2.2). Using standard linear model arguments, we can show that $\hat{\mu}$ and $\widehat{\mathrm{Var}}(\hat{\mu})$ are stochastically independent and $(k-1)\ \widehat{\mathrm{Var}}(\hat{\mu})/\mathrm{E}[\widehat{\mathrm{Var}}(\hat{\mu})]$ follows a χ^2-distribution with $(k-1)$ degrees of freedom, see Hartung (1999). By replacing σ_i^2 through

S_i^2 in Eq. (2.13), we obtain an approximate variance estimator of $\hat{\mu}_{\mathrm{GD}}$, that is,

$$\widehat{\mathrm{Var}}_{(4)}\left(\hat{\mu}_{\mathrm{GD}}\right) = \frac{1}{k-1}\sum_{i=1}^{k}\frac{n_i \;/\; S_i^2}{\sum_{j=1}^{k} n_j \;/\; S_j^2}\left(\bar{Y}_i - \hat{\mu}_{\mathrm{GD}}\right)^2. \tag{2.14}$$

By applying Meier's general theorem (Meier, 1953), Hartung and Knapp (2005b) derived the unconditional expected value of $\widehat{\mathrm{Var}}_{(4)}\left(\hat{\mu}_{\mathrm{GD}}\right)$ as

$$\mathrm{E}\left[\widehat{\mathrm{Var}}_{(4)}\left(\hat{\mu}_{\mathrm{GD}}\right)\right] = \frac{1}{\sum_{i=1}^{k} n_i/\sigma_i^2}\left[1 + 2\sum_{i=1}^{k}\frac{1}{n_i - 1}\left[\frac{k\; c_i\; (1-c_i)}{k-1} + \frac{(1-c_i)^2}{k-1}\right] + O\left(\sum_{i=1}^{k}\frac{1}{(n_i-1)^2}\right)\right]$$

with

$$c_i = \frac{n_i \;/\; \sigma_i^2}{\sum_{j=1}^{k} n_j \;/\; \sigma_j^2}, \quad i = 1, \ldots, k.$$

Note that the expected value of $\widehat{\mathrm{Var}}_{(4)}\left(\hat{\mu}_{\mathrm{GD}}\right)$ is in close agreement to the first order approximation (2.9) of the variance of the Graybill-Deal estimator.

2.1 Approximate Confidence Intervals

Using the Graybill-Deal estimator (2.7) for the common mean and an appropriate estimator of the variance of $\hat{\mu}_{\mathrm{GD}}$, for instance, an estimator from Eqs. (2.10), (2.11), (2.12), or (2.14), approximate $100(1-\alpha)\%$ confidence intervals for μ can be constructed on the basis of a suitable normalization of $\hat{\mu}_{\mathrm{GD}}$.

A simple large sample $100(1-\alpha)\%$ confidence interval, which is widely used in meta-analysis, is given by

$$\mathrm{CI}_{(1)}(\mu): \;\; \hat{\mu}_{\mathrm{GD}} \mp \sqrt{\widehat{\mathrm{Var}}_{(3)}(\hat{\mu}_{\mathrm{GD}})}\; z_{1-\alpha/2}, \tag{2.15}$$

where $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$-quantile of the standard normal distribution. This interval, however, mostly proves to be too narrow and the actual confidence coefficient of the interval (2.15) can be dramatically less than the nominal one, see Li, Shi, and Roth (1994) and Böckenhoff and Hartung (1998). Based on concavity corrections for the estimates of $1/\sigma_i^2$, $i = 1, \ldots, k$, and following the lines of the interval (2.15), Böckenhoff

and Hartung (1998) worked out improved confidence intervals for μ. A larger coverage probability can also be achieved by using the $(1-\alpha/2)$-quantile of a t-distribution with ν degrees of freedom, say $t_{\nu;1-\alpha/2}$, instead of $z_{1-\alpha/2}$. Follmann and Proschan (1999) suggested the choice of $\nu = k-1$ degrees of freedom.

But it is more appealing to use the more accurate variance estimators (2.10) and (2.11) for constructing approximate confidence intervals on the common mean. By using Patnaik's (1949) approximation of equivalent degrees of freedom, Meier (1953) showed that the distribution of $\widehat{\mathrm{Var}}_{(2)}(\hat{\mu}_{\mathrm{GD}})$ can be approximated by a scaled χ^2-distribution with estimated degrees of freedom $\hat{\nu}$, where

$$\frac{1}{\hat{\nu}} = \sum_{i=1}^{k} \frac{1}{n_i - 1} \left(\frac{n_i \ / \ S_i^2}{\sum_{j=1}^{k} n_j \ / \ S_j^2} \right)^2 .$$

Using the same approximate distribution for $\widehat{\mathrm{Var}}_{(1)}(\hat{\mu}_{\mathrm{GD}})$, two approximate $100(1-\alpha)\%$ confidence intervals on μ are given as

$$\mathrm{CI}_{(2)}(\mu): \ \hat{\mu}_{\mathrm{GD}} \mp \sqrt{\widehat{\mathrm{Var}}_{(1)}(\hat{\mu}_{\mathrm{GD}})} \ t_{\hat{\nu};1-\alpha/2} \tag{2.16}$$

and

$$\mathrm{CI}_{(3)}(\mu): \ \hat{\mu}_{\mathrm{GD}} \mp \sqrt{\widehat{\mathrm{Var}}_{(2)}(\hat{\mu}_{\mathrm{GD}})} \ t_{\hat{\nu};1-\alpha/2}. \tag{2.17}$$

Finally, an approximate $100(1-\alpha)\%$ confidence interval for μ, that does not require the estimation of degrees of freedom, can be constructed using the variance estimator (2.14). Since, suitably scaled, $\widehat{\mathrm{Var}}_{(4)}(\hat{\mu}_{\mathrm{GD}})$ can be well approximated by a χ^2-distribution with $k-1$ degrees of freedom, an approximate $100(1-\alpha)\%$ confidence interval for μ is given as

$$\mathrm{CI}_{(4)}(\mu): \ \hat{\mu}_{\mathrm{GD}} \mp \sqrt{\widehat{\mathrm{Var}}_{(4)}(\hat{\mu}_{\mathrm{GD}})} \ t_{k-1;1-\alpha/2}. \tag{2.18}$$

But in the common mean problem, several exact confidence intervals on μ are available, which will be presented in the next section. The approximate intervals, especially intervals (2.15) and (2.18), however, can be also applied to situations when results of independent studies should be combined and the parameter of interest is not a normal mean or a difference of normal mean. This will be shown in Chapters 3-5.

2.2 Exact Confidence Intervals

Since

$$t_i = \frac{\sqrt{n_i}\ \left(\bar{Y}_i - \mu\right)}{S_i} \sim t_{n_i-1} \tag{2.19}$$

or, equivalently,

$$F_i = \frac{n_i\ \left(\bar{Y}_i - \mu\right)^2}{S_i^2} \sim F_{1,n_i-1} \tag{2.20}$$

are test statistics for testing hypotheses about μ based on the ith sample, suitable linear combinations of these test statistics or other functions thereof can be used as a pivotal quantity to construct exact confidence intervals for μ.

Cohen and Sackrowitz (1984) considered $M_t = \max_{1\le i\le k}\{|t_i|\}$ as test statistic for testing hypotheses about μ. We can use M_t to construct an exact confidence interval for μ after determining the quantile of the distribution of M_t, say $c_{1-\alpha/2}$, which satisfies the following equation

$$1-\alpha = \mathrm{P}\left(M_t \le c_{1-\alpha/2}\right) = \prod_{i=1}^{k} \mathrm{P}\left(|t_i| \le c_{1-\alpha/2}\right).$$

Since the distribution of M_t essentially depends on the degrees of freedom of the t-test statistics t_i, the quantile $c_{1-\alpha/2}$ can be readily found using appropriate statistical software packages. An exact $100(1-\alpha)\%$ confidence interval for μ is then given by

$$\begin{aligned} \mathrm{CI}_{(5)}(\mu): \qquad & \left[\max_{1\le i\le k}\left\{\bar{Y}_i - \frac{c_{1-\alpha/2}\ S_i}{\sqrt{n_i}}\right\}\ ,\ \min_{1\le i\le k}\left\{\bar{Y}_i + \frac{c_{1-\alpha/2}\ S_i}{\sqrt{n_i}}\right\}\right] \\ = & \bigcap_{i=1}^{k}\left[\bar{Y}_i - \frac{c_{1-\alpha/2}\ S_i}{\sqrt{n_i}}\ ,\ \bar{Y}_i + \frac{c_{1-\alpha/2}\ S_i}{\sqrt{n_i}}\right]. \end{aligned} \tag{2.21}$$

An alternative approach is to use the confidence interval

$$\begin{aligned} \mathrm{CI}_{(6)}(\mu): \qquad & \left[\max_{1\le i\le k}\left\{\bar{Y}_i - \frac{c^{(i)}_{1-\alpha/2}\ S_i}{\sqrt{n_i}}\right\}\ ,\ \min_{1\le i\le k}\left\{\bar{Y}_i + \frac{c^{(i)}_{1-\alpha/2}\ S_i}{\sqrt{n_i}}\right\}\right] \\ = & \bigcap_{i=1}^{k}\left[\bar{Y}_i - \frac{c^{(i)}_{1-\alpha/2}\ S_i}{\sqrt{n_i}}\ ,\ \bar{Y}_i + \frac{c^{(i)}_{1-\alpha/2}\ S_i}{\sqrt{n_i}}\right], \end{aligned} \tag{2.22}$$

where $c^{(i)}_{1-\alpha/2}$ satisfies the equation

$$\mathrm{P}\left(|t_i| \le c^{(i)}_{1-\alpha/2}\right) = (1-\alpha)^{1/k}.$$

Clearly, $\mathrm{CI}_{(6)}(\mu)$ is an exact $100(1-\alpha)\%$ confidence interval for μ. Since both intervals $\mathrm{CI}_{(5)}(\mu)$ and $\mathrm{CI}_{(6)}(\mu)$ can be described as intersections of individual confidence intervals, these intersections may be empty. Consequently, both intervals are not necessarily always genuine intervals.

Fairweather (1972) suggested using a weighted linear combination of the t_i's, namely

$$W_t = \sum_{i=1}^{k} u_i \; t_i, \quad u_i = \frac{[\mathrm{Var}(t_i)]^{-1}}{\sum_{j=1}^{k} [\mathrm{Var}(t_j)]^{-1}}, \quad i = 1, \ldots, k. \tag{2.23}$$

Let $b_{1-\alpha/2}$ denote the quantile of the distribution of W_t satisfying the equation

$$1 - \alpha = \mathrm{P}\left(|W_t| \leq b_{1-\alpha/2}\right),$$

then the exact $100(1-\alpha)\%$ confidence interval for μ is given by

$$\mathrm{CI}_{(7)}(\mu): \; \frac{\sum_{i=1}^{k} \sqrt{n_i} \; u_i \; \bar{Y}_i \; / \; S_i}{\sum_{j=1}^{k} \sqrt{n_j} \; u_j \; / \; S_j} \mp \frac{b_{1-\alpha/2}}{\sum_{j=1}^{k} \sqrt{n_j} \; u_j \; / \; S_j}. \tag{2.24}$$

Let t_ν denote a t-distributed random variable with ν degrees of freedom, then it holds $\mathrm{Var}(t_\nu) = \nu/(\nu-2)$, $\nu > 2$, so that the distribution of W_t essentially depends on the degrees of freedom of the t-test statistics. Fairweather (1972) provided an approximation of the distribution of W_t that can also be used to approximate the required quantile $b_{1-\alpha/2}$. Since W_t is a linear combination of t-distributed random variables, the distribution of W_t should resemble a scaled t-distribution, that is, we approximate the distribution of W_t by a $c \; t_\nu$-distribution so that the second and fourth moment of both distributions coincide. The solution is given by $\nu = 4 + 1/\sum_{i=1}^{k}[u_i^2/(n_i-5)]$ and $c = \sqrt{(\nu-2) \; / \; (\nu \; A)}$ with $A = \sum_{i=1}^{k}(n_i-3)/(n_i-1)$, see Fairweather (1972). Note that Fairweather's interval is always a genuine interval for $0 < \alpha < 0.5$.

Jordan and Krishnamoorthy (1996) suggested using a linear combination of the F-test statistics (2.20), namely

$$W_f = \sum_{i=1}^{k} w_i \; F_i, \quad w_i = \frac{[\mathrm{Var}(F_i)]^{-1}}{\sum_{j=1}^{k} [\mathrm{Var}(F_j)]^{-1}}, \quad i = 1, \ldots, k. \tag{2.25}$$

Note that $\mathrm{Var}(F_i) = 2 \; m_i^2 \; (m_i-1)/[(m_i-2)^2 \; (m_i-4)]$ with $m_i = n_i - 1$, $i = 1, \ldots, k$. After determining the quantile $a_{1-\alpha/2}$ satisfying the equation

$$1 - \alpha = \mathrm{P}\left(W_f \leq a_{1-\alpha/2}\right),$$

an exact $100(1-\alpha)\%$ confidence interval for μ is given as

$$\mathrm{CI}_{(8)}(\mu): \sum_{i=1}^{k} p_i \ \bar{Y}_i \mp \Delta, \tag{2.26}$$

where

$$p_i = \frac{w_i \ n_i \ / \ S_i^2}{\sum\limits_{j=1}^{k} w_j \ n_j \ / \ S_j^2}, \quad i = 1, \ldots, k,$$

and

$$\Delta^2 = \frac{a_{1-\alpha/2}}{\sum_{i=1}^{k} w_i \ n_i \ / \ S_i^2} - \left\{ \sum_{i=1}^{k} p_i \ \bar{Y}_i^2 - \left(\sum_{i=1}^{k} p_i \ \bar{Y}_i \right)^2 \right\}.$$

Since Δ^2 is not always positive, the interval (2.26) is not always a genuine interval. Jordan and Krishnamoorthy (1996) suggested approximating the distribution of W_F by a $d \ F_{k,\nu}$-distribution, with numerator degrees of freedom equal to the number of populations, so that the first two moments of both distributions coincide. The solutions for d and ν are given by, recall that $m_i = n_i - 1$, $i = 1, \ldots, k$,

$$\nu = \frac{4 \ k \ M_2 - 2 \ (k+2) \ M_1^2}{k \ M_2 - (k+2) \ M_1^2} \quad \text{and} \quad d = (\nu - 2) \ M_1 \ / \ \nu,$$

where

$$M_1 = \mathrm{E}(W_f) = \sum_{i=1}^{k} \frac{w_i \ m_i}{m_i - 2}$$

and

$$M_2 = \mathrm{E}(W_f)^2 = \sum_{i=1}^{k} \frac{3 \ w_i^2 \ m_i^2}{(m_i - 2)(m_i - 4)} + 2 \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{w_i \ w_j \ m_i \ m_j}{(m_i - 2) \ (m_j - 2)},$$

see Jordan and Krishnamoorthy (1996).

Yu, Sun, and Sinha (1999) derived exact $100(1-\alpha)\%$ confidence intervals for μ using p-values of the F-test statistics F_i from Eq. (2.20). Recall that F_i is a F_{1,n_i-1}-distributed random variable, then the ith p-value P_i is defined as

$$P_i = \int_{F_i}^{\infty} h_i(x) \ \mathrm{d}x \ ,$$

where $h_i(x)$ denotes the probability density function of the F-distribution with 1 and $(n_i - 1)$ degrees of freedom. Note that $P_1, \ldots, P_k$ are independently uniformly distributed random variables on the unit interval.

There are several methods for combining p-values, see Hedges and Olkin (1985), that can be used for constructing exact confidence intervals for μ. We restrict the presentation here to the two most familiar methods, the inverse normal method by Stouffer et al. (1949) and the inverse χ^2-method by Fisher (1932). The general construction principle for the confidence intervals is the inversion of the acceptance region a family of level-α-tests. Note that by using Tippett's minimum p-value method, one obtains the interval $\mathrm{CI}_{(6)}(\mu)$ from Eq. (2.22), see Yu, Sun, and Sinha (1999).

Using the inverse normal method, hypotheses about μ will be rejected if

$$\frac{\sum_{i=1}^{k} \Phi^{-1}(P_i)}{\sqrt{k}} < z_\alpha,$$

where Φ^{-1} denotes the inverse of the cumulative distribution function Φ of the standard normal distribution. Consequently, an exact $100(1-\alpha)\%$ confidence interval for μ is given by inverting the acceptance region, that is,

$$\mathrm{CI}_{(9)}(\mu): \left\{\mu: \frac{\sum_{i=1}^{k} \Phi^{-1}(P_i)}{\sqrt{k}} > z_\alpha\right\}. \tag{2.27}$$

Note that this approach does not necessarily yield a genuine interval.

Using Fisher's inverse χ^2-method, hypotheses about μ will be rejected if

$$-2\sum_{i=1}^{k} \ln(P_i) > \chi^2_{2k;1-\alpha},$$

where $\chi^2_{2k;1-\alpha}$ denotes the $(1-\alpha)$-quantile of a χ^2-distribution with $2\,k$ degrees of freedom. Again, by inverting the acceptance region, we obtain an exact $100(1-\alpha)\%$ confidence interval for μ as

$$\mathrm{CI}_{(10)}(\mu): \left\{\mu: -2\sum_{i=1}^{k} \ln(P_i) < \chi^2_{2k;1-\alpha}\right\}. \tag{2.28}$$

Like the interval (2.27), the interval (2.28) is not necessarily a genuine interval. Yu, Sun, and Sinha (1999) derived sufficient conditions for the inverse χ^2-method and the inverse normal method to produce genuine intervals. Moreover, in a small simulation study for $k = 2$ populations, they showed that the interval with the inverse χ^2-method outperforms

the other p-value based exact confidence intervals for μ in terms of expected length. Compared to the other exact intervals, they recommended the use of Fairweather's interval, when the two population variances are close and small, followed by the interval with inverse χ^2-method and Jordan and Krishnamoorthy's interval. When the two variances are widely apart, they recommended the use of the inverse χ^2-method followed by Jordan and Krishnamoorthy (1996) and Fairweather (1972).

Hartung and Knapp (2005b) used the t-test statistics t_i from Eq. (2.19) and suggested two broad classes of exact $100(1-\alpha)\%$ confidence intervals for μ. Let $F_{t_{n_i-1}}$ be the cumulative distribution function of the t-distribution with $(n_i - 1)$ degrees of freedom. Then it holds

$$F_{t_{n_i-1}}(t_i) =: u_i \sim U(0,1) \quad \text{and} \quad \Phi^{-1}(u_i) \sim N(0,1),$$

where $U(0,1)$ stands for the uniform distribution on the unit interval. Let us consider the weighted inverse normal combination statistic

$$Z(\mu) = \sum_{i=1}^{k} \sqrt{\frac{\gamma_i}{\sum_{j=1}^{k} \gamma_j}}\ \Phi^{-1}\left(F_{t_{n_i-1}}(t_i)\right) \tag{2.29}$$

with some positive weights γ_i, $i = 1, \ldots, k$. Clearly, $Z(\mu)$ is a standard normal random variable. One possible choice of positive weights is $\gamma_i = 1$, $i = 1, \ldots, k$. This means that the precision of each result is only represented through the cumulative distribution function $F_{t_{n_i-1}}$. Since the results of larger experiments are usually more precise, a natural choice of the weights γ_i may be the sample size n_i or the degrees of freedom $n_i - 1$.

The functions $F_{t_{n_i-1}}(\cdot)$ and $\Phi^{-1}(\cdot)$ are monotone increasing functions in their arguments $(\cdot)$, so that $Z(\mu)$ from Eq. (2.29) is a monotone decreasing function in μ. Consequently, an exact $100(1-\alpha)\%$ confidence interval for μ is given by

$$\mathrm{CI}_{(11)}(\mu) : [\ \mu_{L,Z}\ ;\ \mu_{U,Z}\ ], \tag{2.30}$$

where the bounds $\mu_{L,Z}$ and $\mu_{U,Z}$ are the unique solutions for μ of the equations

$$Z(\mu) = \Phi^{-1}(1-\alpha/2) \quad \text{and} \quad Z(\mu) = \Phi^{-1}(\alpha/2)\ .$$

A second class of exact confidence intervals for μ suggested by Hartung and Knapp (2005b) is based on the inverse χ^2-method. Let $G_{\gamma_i}^{-1}$ denote the inverse of the cumulative

distribution function G_{γ_i} of a χ^2-distribution with γ_i degrees of freedom. The general inverse χ^2-combination statistic is then given by

$$S(\mu) = \sum_{i=1}^{k} G_{\gamma_i}^{-1}\left(F_{t_{n_i-1}}(t_i)\right). \tag{2.31}$$

Clearly, $S(\mu)$ is a χ^2-distributed random variable with $\gamma_\Sigma = \sum_{i=1}^{k} \gamma_i$ degrees of freedom. Since $F_{t_{n_i-1}}(\cdot)$ and $G_{\gamma_i}^{-1}(\cdot)$ are monotone increasing functions in their arguments $(\cdot)$, $S(\mu)$ is monotone decreasing in μ. Consequently, an exact $100(1-\alpha)\%$ confidence interval for μ is given by

$$\mathrm{CI}_{(12)}(\mu) : [\ \mu_{L,S}\ ;\ \mu_{U,S}\ ], \tag{2.32}$$

where the bounds $\mu_{L,S}$ and $\mu_{U,S}$ are the unique solutions for μ of the equations

$$S(\mu) = \chi^2_{\gamma_\Sigma;1-\alpha/2} \quad \text{and} \quad S(\mu) = \chi^2_{\gamma_\Sigma;\alpha/2}\ .$$

Table 2.1 contains the simulation results concerning the expected lengths of the exact confidence intervals (2.30) and (2.32) for $k = 2$ populations. For interval (2.30), we considered the weights $\gamma_i = 1$ and $\gamma_i = n_i$, $i = 1, \ldots, k$. For interval (2.32), we considered the weights $\gamma_i = 2$, that is, the weights of Fisher's (1932) method for combining p-values, and again the sample sizes $\gamma_i = n_i$, $i = 1, \ldots, k$. We used the simulation design from Yu, Sun, and Sinha (1999).

We observe from Table 2.1, that the exact intervals $\mathrm{CI}_{(11)}$ based on the inverse normal method are always shorter than the exact intervals $\mathrm{CI}_{(12)}$ based on the inverse χ^2-method. For the intervals $\mathrm{CI}_{(12)}$, the weights equal to the sample sizes always produce on average shorter intervals than the constant weights. For the intervals $\mathrm{CI}_{(11)}$, the intervals using the sample sizes as weights are on average shorter than the intervals with the constant weights if the smaller sample size is associated with the larger variance. If the smaller sample size is associated with the smaller variance, the intervals using constant weights are on average shorter.

Table 2.1. Comparison of expected lengths of four exact confidence intervals for μ given a nominal confidence coefficient of $1 - \alpha = 0.95$

Sample size		Standard deviation		Average length			
				$CI_{(11)}$	$CI_{(11)}$	$CI_{(12)}$	$CI_{(12)}$
n_1	n_2	σ_1	σ_2	$\gamma_i = 1$	$\gamma_i = n_i$	$\gamma_i = 2$	$\gamma_i = n_i$
7	10	1	0.5	0.689	0.664	0.771	0.679
7	10	1	1	1.032	1.028	1.080	1.038
7	10	1	5	2.107	2.279	2.594	2.508
7	10	1	10	2.610	2.952	4.169	3.774
7	10	1	20	3.036	3.601	7.073	5.929
10	7	1	0.5	0.730	0.751	0.773	0.768
10	7	1	1	1.035	1.030	1.081	1.039
10	7	1	5	1.826	1.687	2.706	1.890
10	7	1	10	2.081	1.867	4.482	2.469
10	7	1	20	2.285	2.003	8.049	3.553
10	10	1	0.5	0.640	0.640	0.697	0.654
10	10	1	1	0.936	0.936	0.977	0.944
10	10	1	5	1.732	1.732	2.349	1.910
10	10	1	10	2.021	2.021	3.843	2.623
10	10	1	20	2.233	2.233	6.761	3.847
10	15	1	0.5	0.544	0.523	0.617	0.531
10	15	1	1	0.830	0.825	0.873	0.832
10	15	1	5	1.635	1.780	2.065	1.933
10	15	1	10	1.946	2.213	3.300	2.773
10	15	1	20	2.177	2.582	5.680	4.223
15	10	1	0.5	0.583	0.602	0.620	0.612
15	10	1	1	0.831	0.826	0.874	0.833
15	10	1	5	1.390	1.280	2.151	1.384
15	10	1	10	1.556	1.399	3.623	1.692
15	10	1	20	1.662	1.469	6.563	2.181
21	21	1	0.5	0.421	0.421	0.463	0.426
21	21	1	1	0.622	0.622	0.653	0.625
21	21	1	5	1.090	1.090	1.575	1.153
21	21	1	10	1.221	1.221	2.583	1.413
21	21	1	20	1.303	1.303	4.632	1.768

2.3 Generalized Confidence Intervals

The concept of *generalized p-values* was first introduced by Tsui and Weerahandi (1989) to deal with the statistical testing problem in which nuisance parameters are present, and it is difficult or impossible to obtain a non-trivial test with a fixed level of significance. Weerahandi (1993) then introduced the concept of *generalized confidence intervals* in this setting. Although, a lot of exact confidence intervals for the common mean μ exist, see Section 2.2, the generalized confidence interval approach may be an alternative in the common mean problem as some of the exact confidence intervals do not always yield genuine intervals.

The general setup for constructing a generalized confidence interval is as follows: Let $\boldsymbol{X}$ be a random quantity having a density function $f(\boldsymbol{X}|\zeta)$, where $\zeta = (\theta, \boldsymbol{\eta})$ is a vector of unknown parameters, θ is the parameter of interest, and $\boldsymbol{\eta}$ is a vector of nuisance parameters. Suppose we are interested in a confidence interval for θ. Let $\boldsymbol{x}$ denote the observed value of $\boldsymbol{X}$ and consider the generalized variable $T(\boldsymbol{X}; \boldsymbol{x}, \zeta)$, which depends on the observed value $\boldsymbol{x}$ and the parameters ζ, and satisfies the following requirements:

(A) The distribution of $T(\boldsymbol{X}; \boldsymbol{x}, \theta, \boldsymbol{\eta})$ does not depend on any unknown parameters.

(B) The observed value of $T(\boldsymbol{X}; \boldsymbol{x}, \theta, \boldsymbol{\eta})$ is free of the nuisance parameters.

Then, we say $T(\boldsymbol{X}; \boldsymbol{x}, \theta, \boldsymbol{\eta})$ is generalized pivotal quantity. If t_1 and t_2 are such that

$$\mathrm{P}(t_1 \leq T(\boldsymbol{X}; \boldsymbol{x}, \theta, \boldsymbol{\eta}) \leq t_2) = 1 - \alpha, \tag{2.33}$$

then,

$$\{\theta : t_1 \leq T(\boldsymbol{X}; \boldsymbol{x}, \theta, \boldsymbol{\eta}) \leq t_2\}$$

is a $100(1-\alpha)\%$ generalized confidence interval for θ. For example, if the value of $T(\boldsymbol{X}; \boldsymbol{x}, \theta, \boldsymbol{\eta})$ at $\boldsymbol{X} = \boldsymbol{x}$ is θ, then

$$[T(\boldsymbol{x}; \alpha/2)\ ,\ T(\boldsymbol{x}; 1-\alpha/2)]$$

is a $(1-\alpha)$ confidence interval for θ, where $T(\boldsymbol{x}; \kappa)$ stands for the κth quantile of $T(\boldsymbol{X}; \boldsymbol{x}, \theta, \boldsymbol{\eta})$.

Recall that we have independent samples from k normal populations with common mean μ and possibly unequal variances σ_i^2, $i = 1, \ldots, k$. The sample sizes n_i, $i = 1, \ldots, k$, may differ from sample to sample. Let $\bar{Y}_i$ and S_i^2 be the sample mean and sample variance in the ith population. It is noted that $\bar{Y}_i$ and S_i^2 are stochastically independent with

$$\bar{Y}_i \sim N\left(\mu, \frac{\sigma_i^2}{n_i}\right), \quad U_i = \frac{(n_i - 1)\ S_i^2}{\sigma_i^2} = \frac{V_i}{\sigma_i^2} \sim \chi^2_{n_i - 1}, \quad i = 1, \ldots, k. \tag{2.34}$$

Let $\bar{y}_i$ and s_i^2 denote the observed values of $\bar{Y}_i$ and S_i^2, and v_i stands for the observed value of V_i.

Krishnamoorthy and Lu (2003) considered a weighted linear combination of sample generalized pivotal quantities. Within each sample, a generalized pivotal quantity for μ is given as

$$\begin{aligned} T_i &= \bar{y}_i - \left(\frac{\bar{Y}_i - \mu}{\sigma_i/\sqrt{n_i}}\right)\sqrt{\frac{\sigma_i^2\ v_i}{n_i\ V_i}} \\ &= \bar{y}_i - \frac{Z_i}{\sqrt{U_i}}\frac{\sqrt{v_i}}{\sqrt{n_i}}, \\ &= \bar{y}_i - t_i\frac{s_i}{\sqrt{n_i}}, \end{aligned} \tag{2.35}$$

with $Z_i \sim N(0,1)$ and $t_i = \sqrt{n_i - 1}\ Z_i/\sqrt{U_i} \sim t_{n_i - 1}$, $i = 1, \ldots, k$. A generalized pivotal quantity for σ_i^2 is given as

$$R_i = \frac{\sigma_i^2}{V_i}\ v_i = \frac{v_i}{Q_i}, \quad Q_i = \frac{V_i}{\sigma_i^2} \sim \chi^2_{n_i - 1}, \quad i = 1, \ldots, k. \tag{2.36}$$

Define $\bar{\boldsymbol{Y}} = (\bar{Y}_1, \ldots, \bar{Y}_k)'$ and $\boldsymbol{V} = (V_1, \ldots, V_k)'$ and let be $\bar{\boldsymbol{y}}$ and $\boldsymbol{v}$ the corresponding observed values. Then, the generalized pivotal quantity for the common mean μ is given as

$$T_{KL}\left(\bar{\boldsymbol{Y}}, \boldsymbol{V}; \bar{\boldsymbol{y}}, \boldsymbol{v}\right) = \frac{\sum_{i=1}^k\ W_i\ T_i}{\sum_{j=1}^k\ W_j} \tag{2.37}$$

with

$$W_i = n_i\ Q_i/v_i = n_i\ R_i^{-1}.$$

The generalized pivotal quantity T_{KL} fulfills the two conditions (A) and (B) above and the observed value of T_{KL} is μ. Consequently, $\mathrm{GCI}_1(\mu) : \left(T_{KL;\alpha/2}, T_{KL;1-\alpha/2}\right)$ is a generalized

confidence interval for μ. Note that Krishnamoorthy and Lu (2003) used two different χ^2-random variables U_i and Q_i in the definitions of T_i and R_i even though they are related to the same sample sum of squares. As Krishnamoorthy and Lu (2003) pointed out, the use if the same χ^2-random variable in the generalized pivotal quantity produced confidence limits that are too liberal. Since closed-form expressions for the required quantiles are not available, they may be estimated by simulating the distribution of $T_{KL}\left(\bar{\boldsymbol{Y}}, \boldsymbol{V}; \bar{\boldsymbol{y}}, \boldsymbol{v}\right)$ using the following algorithm:

For given data $(\bar{y}_i, s_i^2, n_i)$, $i = 1, \ldots k$:

For $j = 1, \ldots, m$:

1. Generate $t_{n_1-1}, \ldots, t_{n_-1}$.
2. Generate Q_i $\chi^2_{n_i-1}$, $i = 1, \ldots, k$.
3. Compute $W_1, \ldots, W_k$.
4. Compute $T_{KL,j} = \sum_{i=1}^{k} W_i \left(\bar{y}_i - t_i \; s_i/\sqrt{n_i}\right) / \sum_{j=1}^{k} W_j$.

(end j loop)

Compute the $\alpha/2$- and $(1-\alpha/2)$-quantile of $T_{KL,1}, \ldots, T_{KL,m}$.

Then, $(T_{KL;\alpha/2}, T_{KL;1-\alpha/2})$ is a $100(1-\alpha)\%$ generalized confidence interval on μ.

Lin and Lee (2005) first considered the best linear unbiased estimator for μ assuming that the variances σ_i^2, $i = 1, \ldots, k$, are known. This estimator is given as, see Eq. (2.2),

$$\hat{\mu} = \frac{\sum_{i=1}^{k} n_i \; \bar{Y}_i/\sigma_i^2}{\sum_{j=1}^{k} n_j/\sigma_j^2} \tag{2.38}$$

with

$$\hat{\mu} \sim N\left(\mu, \; \left[\sum_{i=1}^{k}(n_i/\sigma_i^2)\right]^{-1}\right).$$

Consequently,

$$\sqrt{\sum_{i=1}^{k}(n_i/\sigma_i^2)} \; (\hat{\mu} - \mu) = Z \sim N(0,1).$$

The generalized pivotal quantity for μ is then given as

$$\begin{aligned} T_{LL}\left(\bar{\boldsymbol{Y}}, \boldsymbol{V}; \bar{\boldsymbol{y}}, \boldsymbol{v}\right) &= \frac{\sum_{i=1}^{k}(n_i/\sigma_i^2)\ \bar{y}_i\ (V_i/v_i)}{\sum_{j=1}^{k}(n_j/\sigma_j^2)\ (V_j/v_j)} - \frac{\sqrt{\sum_{i=1}^{k} n_i/\sigma_i^2}\ (\hat{\mu}-\mu)}{\sqrt{\sum_{j=1}^{k}(n_j/\sigma_j^2)\ (V_j/v_j)}} \\ &= \frac{\sum_{i=1}^{k} n_i\ U_i\ \bar{y}_i/v_i}{\sum_{j=1}^{k} n_j\ U_j/v_j} - \frac{Z}{\sqrt{\sum_{j=1}^{k} n_j\ U_j/v_j}} \\ &= \frac{\sum_{i=1}^{k} W_i\ \bar{y}_i}{\sum_{j=1}^{k} W_j} - \frac{Z}{\sqrt{\sum_{j=1}^{k} W_j}} \end{aligned} \tag{2.39}$$

with

$$W_i = n_i\ U_i/v_i, \quad i = 1, \ldots, k.$$

The generalized pivotal quantity T_{LL} fulfills the two conditions (A) and (B) and the observed value of T_{LL} is μ. Consequently, $\mathrm{GCI}_2(\mu) : \left(T_{LL;\alpha/2}, T_{LL;1-\alpha/2}\right)$ is a generalized confidence interval for μ. Again, closed-form expressions for the required quantiles are not available, but they may be estimated by simulating the distribution of $T_{LL}\left(\bar{\boldsymbol{Y}}, \boldsymbol{V}; \bar{\boldsymbol{y}}, \boldsymbol{v}\right)$ using the following algorithm:

For given data $(\bar{y}_i, s_i^2, n_i)$, $i = 1, \ldots k$:

For $j = 1, \ldots, m$:

1. Generate $Z \sim N(0, 1)$.
2. Generate $U_i \sim \chi^2_{n_i-1}$, $i = 1, \ldots, k$.
3. Compute $W_1, \ldots, W_k$.
4. Compute $T_{LL,j} = \sum_{i=1}^{k}\ W_i\ \bar{y}_i / \sum_{j=1}^{k} W_j - Z/\sqrt{\sum_{i=1}^{k} W_i}$.

(end j loop)

Compute the $\alpha/2$- and $(1-\alpha/2)$-quantile of $T_{LL,1}, \ldots, T_{LL,m}$.

Then, $(T_{LL;\alpha/2}, T_{LL;1-\alpha/2})$ is a $100(1-\alpha)\%$ generalized confidence interval on μ.

A new third approach also starts with the best linear unbiased estimator $\hat{\mu}$ from Eq. (2.38). Moreover, the statistic

$$\widehat{\mathrm{Var}}(\hat{\mu}) = \frac{1}{k-1}\left(\sum_{i=1}^{k}\frac{n_i}{\sigma_i^2}\right)^{-1}\sum_{i=1}^{k}\frac{n_i}{\sigma_i^2}\left(\bar{Y}_i - \frac{\sum_{j=1}^{k} n_i\ \bar{Y}_i/\sigma_i^2}{\sum_{\ell=1}^{k} n_\ell/\sigma_\ell^2}\right)^2 \tag{2.40}$$

is an unbiased estimator of the variance of $\hat{\mu}$ and stochastically independent of $\hat{\mu}$, see

Hartung (1999). Hartung (1999) also showed that

$$(k-1)\sum_{i=1}^{k}(n_i/\sigma_i^2)\ \widehat{\mathrm{Var}}(\hat{\mu}) \tag{2.41}$$

is a χ^2-distributed random variable with $k-1$ degrees of freedom.
Consequently, $(\hat{\mu}-\mu)/\sqrt{\widehat{\mathrm{Var}}(\hat{\mu})}$ is a t-distributed random variable with $k-1$ degrees of freedom.

A new generalized pivotal quantity is then given by

$$\begin{aligned} T_{new}\left(\bar{\boldsymbol{Y}},\boldsymbol{V};\bar{\boldsymbol{y}},\boldsymbol{v}\right) &= \frac{\sum_{i=1}^{k} n_i\ U_i\ \bar{y}_i/v_i}{\sum_{j=1}^{k} n_j\ U_j/v_j} \\ &\quad - t_{k-1}\ \sqrt{\frac{1}{k-1}\left(\sum_{i=1}^{k}\frac{n_i\ U_i}{v_i}\right)^{-1}\sum_{i=1}^{k}\frac{n_i\ U_i}{v_i}\left(\bar{y}_i-\frac{\sum_{j=1}^{k}(n_j\ U_j/v_j)\ \bar{y}_j}{\sum_{\ell=1}^{k}(n_\ell\ U_\ell/v_\ell)}\right)^2} \\ &= \frac{\sum_{i=1}^{k} W_i\ \bar{y}_i}{\sum_{j=1}^{k} W_j} - t_{k-1}\ \sqrt{\frac{1}{k-1}\left(\sum_{i=1}^{k} W_i\right)^{-1}\sum_{i=1}^{k} W_i\left(\bar{y}_i-\frac{\sum_{j=1}^{k} W_j\ \bar{y}_j}{\sum_{\ell=1}^{k} W_\ell}\right)^2} \end{aligned} \tag{2.42}$$

with

$$W_i = n_i\ U_i/v_i, \quad i=1,\ldots,k.$$

Again, the two conditions (A) and (B) above are fulfilled and the observed value of T_{new} is μ. Consequently, $\mathrm{GCI}_3(\mu)$: $\left(T_{new;\alpha/2}, T_{new;1-\alpha/2}\right)$ is a generalized confidence interval for μ. As closed-form expressions for the required quantiles are not available, they may be estimated by simulating the distribution of $T_{new}\left(\bar{\boldsymbol{Y}},\boldsymbol{V};\bar{\boldsymbol{y}},\boldsymbol{v}\right)$ using the algorithm:
For given data $(\bar{y}_i, s_i^2, n_i)$, $i=1,\ldots k$:

For $j=1,\ldots,m$:

1. Generate t_{k-1}.
2. Generate $U_i \sim \chi^2_{n_i-1}$, $i=1\ldots,k$.
3. Compute $W_1,\ldots,W_k$.
4. Compute $T_{new,j} = \sum_{i=1}^{k} W_i\ \bar{y}_i/\sum_{j=1}^{k} W_j$
   $- t_{k-1}\left[1/(k-1)\left(\sum_{i=1}^{k} W_i\right)^{-1}\sum_{i=1}^{k} W_i\left(\bar{y}_i-\sum_{j=1}^{k} W_j\ \bar{y}_j/\sum_{\ell=1}^{k} W_\ell\right)^2\right]^{1/2}$.

(end j loop)

Compute the $\alpha/2$- and $(1-\alpha/2)$-quantile of $T_{new,1}, \ldots, T_{new,m}$.
Then, $(T_{new;\alpha/2}, T_{new;1-\alpha/2})$ is a $100(1-\alpha)\%$ generalized confidence interval on μ.

In Table 2.2, results for simulated actual confidence coefficients and expected lengths of three generalized confidence intervals $\mathrm{GCI}_1(\mu)$, $\mathrm{GCI}_2(\mu)$, and $\mathrm{GCI}_3(\mu)$ are arranged. We used the same simulation design like in Table 2.1.

This small simulation study shows that the generalized confidence interval $\mathrm{GCI}_1(\mu)$ is either slightly conservative or almost exact as already pointed out by Krishnamoorthy and Lu (2003). The Lin and Lee (2005) generalized confidence interval $\mathrm{GCI}_2(\mu)$, however, is either (slightly or moderately) liberal or almost exact, but never conservative. The actual confidence coefficient of the newly proposed generalized confidence interval $\mathrm{GCI}_3(\mu)$ always lies between the two other confidence coefficients. It is either slightly liberal or almost exact. But the average length of $\mathrm{GCI}_3(\mu)$ is not acceptable. Since for $k = 2$ populations, the t-distribution with one degree of freedom is involved in the calculation, $\mathrm{GCI}_3(\mu)$ is simply too wide. The other two generalized confidence intervals have nearly comparable average length. Since the actual confidence coefficient of $\mathrm{GCI}_2(\mu)$ is always less than or equal to the actual of confidence coefficient of $\mathrm{GCI}_1(\mu)$, $\mathrm{GCI}_2(\mu)$ is on average always shorter than $\mathrm{GCI}_1(\mu)$. Compared to the average lengths of the CI_{11}, see Table 2.1, it is noteworthy that the average length of $\mathrm{GCI}_1(\mu)$ is often smaller than the average length of CI_{11}, when $\mathrm{GCI}_1(\mu)$ almost exactly attains the nominal confidence coefficient.

Table 2.2. Simulated confidence coefficients (in %) and expected lengths of three generalized confidence intervals for μ given a nominal level of $1 - \alpha = 0.95$

Sample size		Standard deviation		Confidence coefficient			Average length		
n_1	n_2	σ_1	σ_2	GCI_1	GCI_2	GCI_3	GCI_1	GCI_2	GCI_3
7	10	1	0.5	95.8	94.4	94.9	0.680	0.619	2.869
7	10	1	1	95.9	93.6	94.6	1.139	1.004	4.372
7	10	1	5	95.6	94.5	94.8	1.799	1.691	7.980
7	10	1	10	95.1	94.7	94.9	1.797	1.752	8.698
7	10	1	20	94.8	94.7	94.8	1.795	1.776	8.975
10	7	1	0.5	95.5	93.5	94.1	0.778	0.691	3.028
10	7	1	1	95.8	93.2	94.2	1.139	1.003	4.367
10	7	1	5	95.2	94.5	94.9	1.429	1.377	7.090
10	7	1	10	95.0	94.7	94.7	1.407	1.391	7.441
10	7	1	20	95.0	95.0	94.5	1.395	1.390	7.349
10	10	1	0.5	95.1	93.7	94.6	0.644	0.593	2.737
10	10	1	1	95.6	93.6	94.3	1.019	0.919	4.146
10	10	1	5	95.4	94.8	94.8	1.394	1.349	6.712
10	10	1	10	95.3	95.0	94.9	1.399	1.384	7.057
10	10	1	20	94.8	94.8	95.1	1.389	1.384	7.182
10	15	1	0.5	95.3	94.2	94.8	0.522	0.493	2.398
10	15	1	1	95.4	93.8	93.9	0.882	0.811	3.752
10	15	1	5	95.1	94.4	94.9	1.381	1.334	6.559
10	15	1	10	95.2	94.9	95.2	1.394	1.377	6.904
10	15	1	20	94.9	94.9	94.8	1.389	1.385	6.977
15	10	1	0.5	95.6	94.1	94.5	0.605	0.560	2.594
15	10	1	1	95.7	93.8	94.5	0.881	0.810	3.718
15	10	1	5	95.0	94.8	95.1	1.086	1.067	5.545
15	10	1	10	94.8	94.7	95.2	1.087	1.082	5.703
15	10	1	20	95.0	95.0	95.0	1.087	1.085	5.772
21	21	1	0.5	95.3	94.7	94.9	0.409	0.395	1.940
21	21	1	1	95.5	94.4	94.9	0.647	0.617	3.007
21	21	1	5	95.0	94.8	95.0	0.887	0.877	4.470
21	21	1	10	95.0	94.9	94.9	0.895	0.892	4.669
21	21	1	20	95.1	95.1	95.0	0.896	0.895	4.641

2.4 Tests of Homogeneity

The crucial assumption in the previous sections is that there is a common mean in all the populations or studies. In this section we present some selected tests of testing homogeneity of normal means which can be extended to testing homogeneity of other effect sizes in later chapters. A more detailed discussion of homogeneity tests in the common mean problem can be found in Hartung, Knapp, and Sinha (2008, Chapter 6).

Let Y_{ij} be the observation on the jth subject of the ith population/study, $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$. Then the standard one-way ANOVA model is given by

$$Y_{ij} = \mu_i + e_{ij} = \mu + \beta_i + e_{ij}, \quad i = 1, \ldots, k, \quad j = 1, \ldots, n_i,$$

where μ is the common mean for all the k populations, β_i is the effect of population i with $\sum_{i=1}^{k} \beta_i = 0$, and e_{ij} are error terms which are assumed to be mutually independent and normally distributed with

$$\mathrm{E}(e_{ij}) = 0, \quad \mathrm{Var}(e_{ij}) = \sigma_i^2, \quad i = 1, \ldots, k, \quad j = 1, \ldots, n_i.$$

Under the above set up, we are interested in testing the hypothesis

$$H_0 : \mu_1 = \cdots = \mu_k$$

or, equivalently,

$$H_0 : \beta_1 = \cdots = \beta_k.$$

Assuming equal error variances, one uses the standard likelihood ratio F-test for testing homogeneity which is also known to be the optimum from an invariance point of view. This test statistic, say F_{an}, is given by

$$F_{\mathrm{an}} = \frac{N-k}{k-1} \, \frac{\sum_{i=1}^{k} n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2}{\sum_{i=1}^{k} (n_i - 1) S_i^2}, \tag{2.43}$$

with $N = \sum_{i=1}^{k} n_i$, $\bar{Y}_{i.} = \sum_{j=1}^{n_i} Y_{ij}/n_i$, $\bar{Y}_{..} = \sum_{i=1}^{k} n_i \bar{Y}_{i.}/N$, and
$S_i^2 = \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2/(n_i - 1)$.

Under the null hypothesis, F_{an} has an F-distribution with $k-1$ and $N-k$ degrees of freedom. The test rejects H_0 at level α if $S_{\mathrm{an}} > F_{k-1,N-k;1-\alpha}$, where $F_{k-1,N-k;1-\alpha}$ denotes

the $(1-\alpha)$-quantile of the F-distribution with $k-1$ and $N-k$ degrees of freedom. This ANOVA F-test has the weakness of not being robust with respect to heterogeneity in the intra-population error variances (Brown and Forsythe, 1974).

Based on the standard ANOVA F-test statistic, several modifications have been proposed for testing equality of means in the case of heteroscedastic error variances, for instance, the Brown-Forsythe (1974) test, a modification of the Brown-Forsythe test proposed by Mehrotra (1997), or an approximate F-test by Asiribo and Gurland (1990).

For testing H_0 in case of heteroscedastic error variances, Cochran (1937) suggested the test statistic

$$Q_{\mathrm{C}} = \sum_{i=1}^{k} \hat{v}_i \left(\bar{Y}_{i.} - \sum_{j=1}^{k} h_j \, \bar{Y}_{j.} \right)^2, \tag{2.44}$$

where $\hat{v}_i = n_i / S_i^2$, $h_i = \hat{v}_i / \sum_{i=1}^{k} \hat{v}_i$. Under H_0, Cochran's statistic is approximately χ^2-distributed with $k-1$ degrees of freedom. The test rejects H_0 at level α if $Q_{\mathrm{C}} > \chi^2_{k-1;1-\alpha}$, where $\chi^2_{k-1;1-\alpha}$ is the $(1-\alpha)$-quantile of the χ^2-distribution with $k-1$ degrees of freedom. Cochran's test is often used as the standard test for testing homogeneity of effect sizes in meta-analysis. The popularity of this test stems from the fact that the test statistic can be easily adapted to other parameters than the normal mean. However, in the common mean problem, Cochran's test can be very liberal for small or moderate sample sizes in the groups and, thus, cannot be recommended for practical use in this situation, see the extensive simulation study by Hartung, Argac, and Makambi (2002).

An improved test based on Q_{C} from Eq. (2.44) in terms of attaining the nominal level was suggested by Welch (1951). The Welch test statistic is given by

$$Q_{\mathrm{W}} = \frac{\sum_{i=1}^{k} \hat{v}_i \left(\bar{Y}_{i.} - \sum_{j=1}^{k} h_j \, \bar{Y}_{j.} \right)^2}{(k-1) + 2\left[(k-2)/(k+1)\right] \sum_{i=1}^{k} (1-h_i)^2 / (n_i - 1)}, \tag{2.45}$$

where $\hat{v}_i = n_i / S_i^2, h_i = \hat{v}_i / \sum_{i=1}^{k} \hat{v}_i$. Under H_0, the statistic Q_W has an approximate F-distribution with $k-1$ and ν_g degrees of freedom, where

$$\nu_g = \frac{(k^2-1)/3}{\sum_{i=1}^{k} (1-h_i)^2 / (n_i - 1)}.$$

This test rejects H_0 at level α if $Q_{\mathrm{W}} > F_{k-1,\nu_g;1-\alpha}$. The basic idea of the Welch test is to

approximate the distribution of Cochran's test statistic through a scaled F-distribution, say $c\, F_{k-1,\nu_g}$, so that the first two moments of both distributions coincide under H_0.

Cochran's test as well as Welch's test use estimated weights $\hat{v}_i = n_i/S_i^2$. Since we know that

$$\mathrm{E}(\hat{v}_i) = \mathrm{E}\left(\frac{n_i}{S_i^2}\right) = c_i\ \frac{n_i}{\sigma_i^2},$$

where $c_i = (n_i - 1)/(n_i - 3)$, an unbiased estimator of n_i/σ_i^2 is $n_i/(c_i\ S_i^2)$. Defining $\hat{v}_i^* = n_i/(c_i\, S_i^2)$, Hartung, Argac, and Makambi (2002) proposed a test they called adjusted Welch test, denoted by $Q_{\text{adj.W}}$, which is given by

$$Q_{\text{adj.W}} = \frac{\sum_{i=1}^k \hat{v}_i^*\ (\bar{Y}_{i.} - \sum_{j=1}^k h_j^*\ \bar{Y}_{j.})^2}{(k-1) + 2\ [(k-2)/(k+1)]\ \sum_{i=1}^k (1-h_i^*)^2\ /(n_i - 1)}, \tag{2.46}$$

where $h_i^* = \hat{v}_i^*/\sum_{j=1}^k \hat{v}_j^*, i = 1, \ldots, k$. Under H_0, the adjusted Welch statistic, $Q_{\text{adj.W}}$, is distributed approximately as an F-variable with $k-1$ and ν_g^* degrees of freedom, where

$$\nu_g^* = \frac{(k^2-1)/3}{\sum_{i=1}^k\ (1-h_i^*)^2\ /(n_i - 1)}.$$

The test rejects H_0 at level α if $Q_{\text{adj.W}} > F_{k-1,\nu_g^*;1-\alpha}$.

Note that the numerator of the test statistic (2.46) can be seen as an adjusted Cochran statistic, that is,

$$Q_{\text{adj.C}} = \sum_{i=1}^k \hat{v}_i^*\ \left(\bar{Y}_{i.} - \sum_{j=1}^k h_j^*\ \bar{Y}_{j.}\right)^2 \tag{2.47}$$

and this test rejects H_0 at level α if $Q_{\text{adj.C}} > \chi^2_{k-1;1-\alpha}$.

Hartung, Argac, and Makambi (2002) reported that the use of the unbiased weights $\hat{v}_i^*$ in test statistic (2.46) leads to a very conservative test. Therefore, they considered general weights, say $\tilde{v}_i^* = n_i/(\varphi_i\ S_i^2)$, $i = 1, \ldots, k$, with $\varphi_i = (n_i + \delta_1)/(n_i + \delta_2)$ and δ_1 and δ_2 are real numbers satisfying $1 \leq \varphi_i \leq c_i = (n_i - 1)/(n_i - 3)$. Replacing $\hat{v}_i^*$ by $\tilde{v}_i^*$ in Eq. (2.46) defines a new class of Welch-type test statistics. The motivation for considering adjustments of the Welch test is based on the observation that the Welch test can be liberal for small sample sizes in the groups and increasing number of groups. Based on their simulation study, Hartung, Argac, and Makambi (2002) recommended the use of $\varphi_i = (n_1 + 2)/(n_i + 1)$ as correction factor for adjusting the weights $\tilde{v}_i^*$.

Using the simulation pattern for $k = 9$ groups from Hartung, Argac, and Makambi (2002), which is reproduced in Table 2.3., we investigated the actual level of Cochran's test, Q_{C} from Eq. (2.44), of the adjusted Cochran test, $Q_{\mathrm{adj.C}}$ from Eq. (2.47), of Welch's test, Q_{W} from Eq. (2.45), of the adjusted Welch test, $Q_{\mathrm{adj.W}}$ from Eq. (2.46), and of the recommended adjusted Welch test with $\varphi_i = (n_1 + 2)/(n_i + 1)$, denoted by $Q_{\mathrm{adj.W}}(\varphi)$, via Monte Carlo simulation. The results of the simulation study are presented in Table 2.4.

As Hartung, Argac, and Makambi (2002) already pointed out, Cochran's test is very liberal and cannot be recommended in this situation. The adjusted Cochran test corrects this shortcoming rather well but is still a bit too liberal. The Welch test is too liberal for small sizes. For increasing sample sizes, the actual level of the Welch test tends to the nominal one, but in the present simulation scenario the test always remains a bit too liberal. The adjusted Welch test, $Q_{\mathrm{adj.W}}$, is clearly too conservative and the other adjusted Welch test, $Q_{\mathrm{adj.W}}(\varphi)$, acts quite well for small sample sizes.

Table 2.3. Sample designs for $k = 9$ groups

Pattern	Samples size and variance in the groups									
	i	1	2	3	4	5	6	7	8	9
1	n_i	5	5	5	5	5	5	5	5	5
	σ_i^2	4	4	4	4	4	4	4	4	4
2	n_i	5	5	5	5	5	5	5	5	5
	σ_i^2	2	6	10	2	6	10	2	6	10
3	n_i	10	10	10	10	10	10	10	10	10
	σ_i^2	4	4	4	4	4	4	4	4	4
4	n_i	10	10	10	10	10	10	10	10	10
	σ_i^2	2	6	10	2	6	10	2	6	10
5	n_i	5	10	15	5	10	15	5	10	15
	σ_i^2	4	4	4	4	4	4	4	4	4
6	n_i	5	10	15	5	10	15	5	10	15
	σ_i^2	2	6	10	2	6	10	2	6	10
7	n_i	5	10	15	5	10	15	5	10	15
	σ_i^2	10	6	2	10	6	2	10	6	2
8	n_i	10	20	30	10	20	30	10	20	30
	σ_i^2	4	4	4	4	4	4	4	4	4
9	n_i	10	20	30	10	20	30	10	20	30
	σ_i^2	2	6	10	2	6	10	2	6	10
10	n_i	10	20	30	10	20	30	10	20	30
	σ_i^2	10	6	2	10	6	2	10	6	2

Table 2.4. Simulated actual significance level (in %) of five homogeneity tests given a nominal level of $\alpha = 0.05$.

Pattern	Q_{C}	$Q_{\mathrm{adj.C}}$	Q_{W}	$Q_{\mathrm{adj.W}}$	$Q_{\mathrm{adj.W}}(\varphi)$
1	29.0	5.7	7.6	0.8	4.9
2	29.6	6.0	7.9	1.0	5.2
3	14.5	6.2	5.8	1.9	4.1
4	14.6	6.3	5.8	2.0	4.1
5	19.1	6.4	7.3	1.7	5.1
6	17.3	5.7	6.4	1.4	4.4
7	20.3	6.9	7.9	1.9	5.6
8	10.6	6.0	5.6	2.8	4.4
9	10.0	5.7	5.4	2.7	4.2
10	11.1	6.1	5.7	2.8	4.4

Chapter 3

The One-Way Random Effects Model

The crucial assumption in Chapter 2 is that the means are all equal in the several populations or studies. In Section 2.4, we discussed some selected tests for testing the equality of means in several normal populations, for a more detailed discussion let us refer to Hartung, Knapp, and Sinha (2008). Practically, these homogeneity tests are often used as pre-tests for the choice of the appropriate model of analysis. In case, one cannot reject the null hypothesis of equality of means, one feels confident in analyzing a common mean. If the null hypothesis is rejected, the model to be analyzed will be the so-called one-way random effects model, which is the topic of this chapter.

The derivation of the one-way random effects model can be seen from different views. Using standard linear model theory, one assumes that there is extra variation additionally to the within-population variability and this extra variation is due to random population-by-subject interaction. This interaction term can be modelled as a random variable with mean 0 and variance, say τ^2.

The second approach uses a normal-normal hierarchial model approach. The observational model assumes that each population has a normal mean, say μ_i, and variance σ_i^2, $i = 1, \ldots, k$, and each mean and variance can differ from population to population. In the structural model, one assumes that the means μ_i are random variables coming from

a super-population with normal mean μ and variance, say τ^2. The parameters μ and τ^2 are also called hyperparameters in this approach.

Both approaches finally lead to the one-way random effects model. Let $\bar{Y}_i$ denote the sample mean in the ith population, S_i^2 the sample variance, and n_i the sample size, $i = 1, \ldots, k$.

Then, we have

$$\bar{Y}_i \sim N\left(\mu \ , \ \tau^2 + \frac{\sigma_i^2}{n_i}\right) \quad \text{and} \quad \frac{(n_i - 1)\ S_i^2}{\sigma_i^2} \sim \chi^2_{n_i - 1}, \ i = 1, \ldots, k, \tag{3.1}$$

where $\tau^2 \geq 0$ stands for the variability between the populations and is also called the heterogeneity parameter. The expected value μ is generally called overall mean. In case $\tau^2 = 0$, we have the common mean problem from Chapter 2. Note that $(\bar{Y}_i, S_i^2, i = 1, \ldots, k)$ are minimally sufficient statistics in model (3.1).

If the variances τ^2 and σ_i^2, $i = 1, \ldots, k$, are completely known, the maximum likelihood estimator for μ in model (3.1) is given as

$$\hat{\mu} = \frac{\sum_{i=1}^k (\tau^2 + \sigma_i^2/n_i)^{-1}\ \bar{Y}_i}{\sum_{j=1}^k (\tau^2 + \sigma_j^2/n_j)^{-1}}. \tag{3.2}$$

The estimator (3.2) is also the minimum variance unbiased estimator under normality as well as the best linear unbiased estimator without normality for estimating μ in model (3.1). The variance of $\hat{\mu}$ is given by

$$\mathrm{Var}(\hat{\mu}) = \left[\sum_{i=1}^k (\tau^2 + \sigma_i^2/n_i)^{-1}\right]^{-1}.$$

In practice, the within-population variances σ_i^2, $i = 1, \ldots, k$, can be unbiasedly estimated using the sample variances S_i^2. The heterogeneity parameter τ^2, however, has to be estimated using the sufficient statistics $(\bar{Y}_i, S_i^2)$, $i = 1, \ldots, k$.

3.1 Estimators of the Heterogeneity Parameter

In the literature, a lot of estimators for τ^2 were proposed, see Rao, Kaplan, Cochran (1981). In this section, we review one class of estimators based on quadratic forms of $\bar{Y}_i$,

$i = 1, \ldots, k$, and the estimators are then deduced by applying the method of moments principle. Cochran (1954) set the sample variance of the $\bar{Y}_i$'s, that is,

$$S_Y^2 = \frac{1}{k-1} \sum_{i=1}^{k} \left(\bar{Y}_i - \bar{Y}\right)^2, \tag{3.3}$$

with $\bar{Y} = \sum_{i=1}^{k} \bar{Y}_i / k$, equal to its expected value and solves for τ^2. Replacing σ_i^2 through the sample variance S_i^2, the method of moments estimator for τ^2, also called ANOVA-type estimator, is given as

$$\hat{\tau}_{AN}^2 = \frac{1}{k-1} \sum_{i=1}^{k} \left(\bar{Y}_i - \bar{Y}\right)^2 - \frac{1}{k} \sum_{i=1}^{k} \frac{S_i^2}{n_i}. \tag{3.4}$$

The estimator τ_{AN}^2 may lead to a negative estimate of τ^2, and hence it is used by enforcing non-negativity in practice, that is, $\max\{0, \hat{\tau}_{AN}^2\}$.

A widely used estimator for τ^2, using a similar approach like Cochran (1954), is the method of moments estimator proposed by DerSimonian and Laird (1986). They use Cochran's (1954) statistic

$$Q_{\mathrm{C}}^2 = \sum_{i=1}^{k} v_i \ \left(\bar{Y}_i - \bar{Y}_v\right)^2, \tag{3.5}$$

where $v_i = n_i / \sigma_i^2$ and $\bar{Y}_v = \sum_{i=1}^{k} v_i \, \bar{Y}_i / \sum_{i=1}^{k} v_i$. By equating Q_{C}^2 to its expected value and solving for τ^2 they find the method of moments estimator for τ^2. Replacing σ_i^2 through the sample variance S_i^2 in practice, the method of moments estimator for τ^2, also called DerSimonian and Laird estimator, is given as

$$\hat{\tau}_{\mathrm{DSL}}^2 = \frac{\hat{Q}_{\mathrm{C}}^2 - (k-1)}{\sum_{i=1}^{k} \hat{v}_i - \sum_{i=1}^{k} \hat{v}_i^2 / \sum_{i=1}^{k} \hat{v}_i}, \tag{3.6}$$

where $\hat{v}_i = n_i / S_i^2$ and $\hat{Q}_{\mathrm{C}}^2$ is obtained by replacing v_i by $\hat{v}_i$ in Q_{C}^2. The estimator $\hat{\tau}_{\mathrm{DSL}}^2$ may also yield a negative estimate for the heterogeneity parameter, and hence the truncated version $\max\{0, \hat{\tau}_{\mathrm{DSL}}^2\}$ is usually used.

Recently, a general method of moments estimator for τ^2 was considered by Kacker (2004) using general weights. Note that Hartung, Böckenhoff, and Knapp (2003) already developed methods for combining results using general weights, see Section 3.4.

Using Kacker's approach, suppose $\bar{Y}_a = \sum_{i=1}^{k} a_i \, \bar{Y}_i / \sum_{i=1}^{k} a_i$, where $a_1, \ldots, a_k$ are any positive constants. Then it holds

$$\begin{aligned} & \mathrm{E}\left[\sum_{i=1}^{k} a_i \left(\bar{Y}_i - \bar{Y}_a\right)^2\right] \\ &= \sum_{i=1}^{k} a_i \left(\tau^2 + \sigma_i^2/n_i\right) - \sum_{i=1}^{k} a_i^2 \left(\tau^2 + \sigma_i^2/n_i\right) \Big/ \sum_{j=1}^{k} a_j \\ &= \tau^2 \left(\sum_{i=1}^{k} a_i - \frac{\sum_{j=1}^{k} a_j^2}{\sum_{\ell=1}^{k} a_\ell}\right) + \left(\sum_{i=1}^{k} a_i \, \frac{\sigma_i^2}{n_i} - \frac{\sum_{j=1}^{k} a_j^2 \, \sigma_j^2/n_j}{\sum_{\ell=1}^{k} a_\ell}\right). \end{aligned} \tag{3.7}$$

By replacing σ_i^2 through S_i^2, a general method of moments estimator of τ^2 can be obtained as

$$\hat{\tau}^2_{\mathrm{GMM}} = \frac{\sum_{i=1}^{k} a_i \left(\bar{Y}_i - \bar{Y}_a\right)^2 - \left(\sum_{i=1}^{k} a_i \, \frac{S_i^2}{n_i} - \frac{\sum_{j=1}^{k} a_j^2 \, S_j^2/n_j}{\sum_{\ell=1}^{k} a_\ell}\right)}{\sum_{i=1}^{k} a_i - \frac{\sum_{j=1}^{k} a_j^2}{\sum_{\ell=1}^{k} a_\ell}}. \tag{3.8}$$

In Eq. (3.8), $a_1, \ldots, a_k$ are any positive values reflecting weights assigned to the k studies. Each set of values for the weights yields an alternative estimator for τ^2. Note that for $a_i = 1/k$, $i = 1, \ldots, k$, the estimator (3.8) is the ANOVA-type estimator (3.4), and for $a_i = n_i/S_i^2$, $i = 1, \ldots, k$, the estimator (3.8) is the DerSimonian-Laird estimator (3.6). Again, the general method of moments estimator $\hat{\tau}^2_{\mathrm{GMM}}$ can yield negative values, and hence the truncated version, $\max\{0, \hat{\tau}^2_{\mathrm{GMM}}\}$, is used in practice.

With $a_i = 1/(\tau^2 + \sigma_i^2/n_i)$, $i = 1, \ldots, k$, equation (3.7) reduces to

$$\mathrm{E}\left[\sum_{i=1}^{k} a_i \left(\bar{Y}_i - \bar{Y}_a\right)^2\right] = k - 1. \tag{3.9}$$

By substituting $S_1^2, \ldots, S_k^2$ for $\sigma_1^2, \ldots, \sigma_k^2$ we get the Mandel-Paule (1970) estimating equation

$$Q(\tau^2) = \sum_{i=1}^{k} \tilde{w}_i \left[\bar{Y}_i - \bar{Y}_{\tilde{w}}(\tau^2)\right]^2 = k - 1, \tag{3.10}$$

where $\bar{Y}_{\tilde{w}}(\tau^2) = \sum_{i=1}^{k} \tilde{w}_i \bar{Y}_i / \sum_{i=1}^{k} \tilde{w}_i$ and $\tilde{w}_i = 1/(\tau^2 + S_i^2/n_i)$, $i = 1, \ldots, k$. The solution of Eq. (3.10), say $\hat{\tau}^2_{\mathrm{MP}}$, is called the Mandel-Paule estimator for τ^2. Since $Q(\tau^2)$ is a

strictly monotone decreasing function in τ^2, see, for instance, Hartung and Knapp (2005a), the solution is unique and exists provided that $Q(0) > k-1$. If $Q(0) < k-1$, the Mandel-Paule estimator is set to zero. Like the general method of moments estimator $\hat{\tau}^2_{\text{GMM}}$, the Mandel-Paule estimator $\hat{\tau}^2_{\text{MP}}$ does not require a normality assumption. Ruhkin, Biggerstaff, and Vangel (2000) investigated the properties of $\hat{\tau}^2_{\text{MP}}$ under normality and showed that $\hat{\tau}^2_{\text{MP}}$ is close to the conditionally restricted maximum likelihood estimator for τ^2; the condition being that the observed sample variance $s_1^2, \ldots, s_k^2$ be regarded as the true within-population variances $\sigma_1^2, \ldots, \sigma_k^2$. Note that the estimating equations for the (conditionally) maximum likelihood and restricted maximum likelihood estimator are presented in the next section.

Since the truncated version of the general method of moments estimator has a positive probability of yielding zero as the estimate, this estimator may not be the appropriate choice especially if heterogeneity is actually present. Following the lines in Hartung and Makambi (2002), we can construct an always non-negative estimator for τ^2 using the basic quadratic form of the general method of moments estimator. For simplifying the notation, let be

$$Q_a = \sum_{i=1}^{k} a_i \left(\bar{Y}_i - \bar{Y}_a\right)^2 \ ,$$

$$A = \sum_{i=1}^{k} a_i - \frac{\sum_{j=1}^{k} a_j^2}{\sum_{\ell=1}^{k} a_\ell} \ ,$$

and

$$B(\boldsymbol{\sigma}^2) = \sum_{i=1}^{k} a_i \ \frac{\sigma_i^2}{n_i} - \frac{\sum_{i=1}^{k} a_i^2 \ \sigma_i^2/n_i}{\sum_{i=1}^{k} a_i}$$

with $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_k^2)'$. Then we can briefly write, see Eq. (3.7),

$$\mathrm{E}(Q_a) = \tau^2 \ A + B(\boldsymbol{\sigma}^2).$$

Interpret $Q_1(a) = Q_a/A$ as a positive estimate of τ^2 and define the estimator

$$\hat{\tau}^2(\delta) = \delta \ Q_1(a), \ \delta > 0,$$

then is holds

$$\begin{aligned} |\text{Bias}(\hat{\tau}^2(\delta))| &= |\text{E}(\delta\ Q_1(a)) - \tau^2| \\ &= \left|(\delta-1)\tau^2 + \delta\ B(\boldsymbol{\sigma}^2)/A\right| \\ &\leq \left\|\begin{pmatrix} \delta-1 \\ \delta \end{pmatrix}\right\| \left\|\begin{pmatrix} \tau^2 \\ B(\boldsymbol{\sigma}^2)/A \end{pmatrix}\right\| \end{aligned}$$

by the Cauchy-Schwarz inequality with $||(\cdot)||$ the Euclidean norm of $(\cdot)$. According to the uniformly minimum bias principle by Hartung (1981) we have to minimize

$$(\delta-1)^2 + \delta^2 \text{ for } \delta > 0$$

giving $\delta = 1/2$.

To adjust for bias, let be $\hat{\tau}^2(\eta) = \eta\ Q_1(a)/2 = \eta\ \hat{\tau}^2(\delta)$ such that

$$\text{E}\left(\eta\ \hat{\tau}^2(\delta) + \eta\ \frac{B(\hat{\boldsymbol{\sigma}}^2)}{A}\right) = \text{E}\left[Q_1(a)\right]$$

with

$$B(\hat{\boldsymbol{\sigma}}^2) = \sum_{i=1}^{k} a_i\ \frac{S_i^2}{n_i} - \frac{\sum_{i=1}^{k} a_i^2\ S_i^2/n_i}{\sum_{i=1}^{k} a_i}.$$

Since $\text{E}[B(\hat{\boldsymbol{\sigma}}^2)] = B(\boldsymbol{\sigma}^2)$, we have to choose

$$\eta = \frac{2\ \text{E}[Q_1(a)]}{\text{E}[Q_1(a)] + 2\ B(\boldsymbol{\sigma}^2)}.$$

For practical purpose, the desired non-negative estimator of τ^2 is given as

$$\hat{\tau}^2_{\text{pos}}(\eta) = \frac{Q_1(a)}{Q_1(a) + 2\ B(\hat{\boldsymbol{\sigma}}^2)}\ Q_1(a). \tag{3.11}$$

Recently, Sidik and Jonkman (2005a) proposed another always non-negative heterogeneity estimator based on considerations from the linear regression model. Let $\bar{\boldsymbol{Y}} = (\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_k)^T$ be the vector of the sample means, then it holds

$$\text{E}(\bar{\boldsymbol{Y}}) = \mu \mathbf{1}_{\boldsymbol{k}}$$

and

$$\text{Var}(\bar{\boldsymbol{Y}}) = \tau^2 \boldsymbol{V},$$

where $\mathbf{1_k}$ is a vector of ones with dimension $k \times 1$ and $\boldsymbol{V}$ is a $(k \times k)$-diagonal matrix with entries $\sigma_i^2/(n_i\ \tau^2) + 1$, $i = 1, \ldots, k$. Assume that the ratios $r_i = \sigma_i^2/(n_i\ \tau^2)$, $i = 1, \ldots, k$, are known, then the best linear unbiased estimator of μ is

$$\hat{\mu}_r = \frac{\sum_{i=1}^{k}(r_i + 1)^{-1}\ \bar{Y}_i}{\sum_{j=1}^{k}(r_j + 1)^{-1}}.$$

An estimate of the variance of $\hat{\mu}$ is readily given as

$$\widehat{\mathrm{Var}}(\hat{\mu}_r) = \frac{\hat{\tau}^2}{\sum_{i=1}^{k}(r_i + 1)^{-1}},$$

where $\hat{\tau}^2$ is an estimate of the heterogeneity variance. Using the weighted residual sum of squares, an estimate of τ^2 is

$$\hat{\tau}^2 = \frac{(\bar{\boldsymbol{Y}} - \hat{\mu}_r \mathbf{1_k})^T\ \boldsymbol{V}^{-1}\ (\bar{\boldsymbol{Y}} - \hat{\mu}_r \mathbf{1_k})}{k - 1} = \frac{1}{k-1}\sum_{i=1}^{k}(r_i + 1)^{-1}\left(\bar{Y}_i - \hat{\mu}_r\right)^2. \tag{3.12}$$

However, the estimate (3.12) depends on the ratios r_i which are usually unknown, and each ratio depends on the heterogeneity parameter itself. To overcome this problem, Sidik and Jonkman (2005a) proposed a two-step procedure. First, compute a crude estimator of τ^2, say $\hat{\tau}_0^2$, and estimate the ratio r_i by $\hat{r}_i = S_i^2/(n_i\ \tau_0^2)$, $i = 1, \ldots, k$, and then replace r_i by $\hat{r}_i$ in (3.12). This results in the final estimate

$$\hat{\tau}^2_{\mathrm{SJ}} = \frac{1}{k-1}\sum_{i=1}^{k}(\hat{r}_i + 1)^{-1}\left(\bar{Y}_i - \hat{\mu}_{\hat{r}}\right)^2. \tag{3.13}$$

As a crude estimate of τ^2, Sidik and Jonkman (2005a) used

$$\hat{\tau}_0^2 = \frac{1}{k}\sum_{i=1}^{k}\left(\bar{Y}_i - \bar{Y}\right)^2$$

with $\bar{Y}$ the arithmetic mean of the $\bar{Y}_i$'s.

Finally, using the general approach of nonnegative minimum biased invariant quadratic estimation of variance components proposed by Hartung (1981), Heine (1993) derived the nonnegative minimum biased invariant quadratic estimator of τ^2 in the present model. Let be $N = \sum_{i=1}^{k} n_i$ and if $N - 2n_i \geq 0$, $i = 1, \ldots, k$, this estimator reads

$$\hat{\tau}^2_{\mathrm{PSD}} = \frac{N^2 \sum_{i=1}^{k} n_i^2 \prod_{\ell' \neq \ell}(N - 2n_{\ell'})(\bar{Y}_i - \sum_{j=1}^{k} n_j \bar{Y}_j/N)^2}{\left(\sum_{\ell=1}^{k} n_\ell^2 + 1\right)\sum_{\ell=1}^{k} n_\ell (N - n_\ell)\prod_{\ell' \neq \ell}(N - 2n_{\ell'})}. \tag{3.14}$$

It is interesting to observe that the estimator (3.14) requires that the sample size in each population or study must be less than or equal to the half of the total sample size. A similar condition occurs when estimating the variance of the overall mean in the random effects model with general weights to ensure the positiveness of the estimator, see Section 3.4 and Hartung, Böckenhoff, and Knapp (2003).

3.2 Confidence Intervals for the Heterogeneity Parameter

In this section we review several confidence intervals for the heterogeneity parameter. Recall that

$$\bar{Y}_i \sim N\left(\mu \ , \ \tau^2 + \frac{\sigma_i^2}{n_i}\right) \ , \ i = 1, \ldots, k,$$

then it holds for the log-likelihood function of μ and τ^2, assuming $\sigma_1^2, \ldots, \sigma_k^2$ are known,

$$l(\mu, \tau^2) = \ -\frac{1}{2}\sum_{i=1}^{k} \ln\left(\tau^2 + \frac{\sigma_i^2}{n_i}\right) - \frac{1}{2}\sum_{i=1}^{k} \frac{(\bar{Y}_i - \mu)^2}{\tau^2 + \sigma_i^2}. \tag{3.15}$$

leaving out the additive constant. The two estimating equations for μ and τ^2 are

$$\hat{\mu} = \frac{\sum_{i=1}^{k} w_i \, \bar{Y}_i}{\sum_{j=1}^{k} w_j} \tag{3.16}$$

and

$$\hat{\tau}^2 = \frac{\sum_{i=1}^{k} w_i^2 \left[(\bar{Y}_i - \hat{\mu})^2 - \sigma_i^2/n_i\right]}{\sum_{j=1}^{k} w_j^2} \tag{3.17}$$

with $w_i = 1/(\tau^2 + \sigma_i^2/n_i)$, $i = 1, \ldots, k$. Let $\hat{\mu}_{\mathrm{ML}}$ and $\hat{\tau}^2_{\mathrm{ML}}$ denote the ML estimators. A confidence interval for τ^2 can then be obtained by profiling the likelihood ratio statistic, see Hardy and Thompson (1996) and Biggerstaff and Tweedie (1997). Denote $\tilde{\mu}$ as that value of Eq. (3.16) with $w_i = 1/(\tilde{\tau}^2 + \sigma_i^2/n_i)$. Then, a $100(1-\alpha)\%$ confidence interval for τ^2 is given by

$$\begin{aligned} \mathrm{CI}_1(\tau^2): \quad & \left\{\tilde{\tau}^2 \mid \ -2\left[l(\tilde{\mu}, \tilde{\tau}^2) - l(\hat{\mu}_{\mathrm{ML}}, \hat{\tau}^2_{\mathrm{ML}})\right] < \chi^2_{1;1-\alpha}\right\} \\ & = \left\{\tilde{\tau}^2 \mid l(\tilde{\mu}, \tilde{\tau}^2) > l(\hat{\mu}_{\mathrm{ML}}, \hat{\tau}^2_{\mathrm{ML}}) - \chi^2_{1;1-\alpha}/2\right\}. \end{aligned} \tag{3.18}$$

Alternatively, one can base the confidence interval on the restricted log-likelihood. Following Viechtbauer (2007), it holds for the restricted log-likelihood for τ^2

$$l_R(\tau^2) = -\frac{1}{2}\sum_{i=1}^{k} \ln(\tau^2 + \sigma_i^2/n_i) - \frac{1}{2}\sum_{i=1}^{k} \frac{1}{\tau^2 + \sigma_i^2/n_i} - \frac{1}{2}\sum_{i=1}^{k} \frac{(\bar{Y}_i - \hat{\mu})^2}{\tau^2 + \sigma_i^2/n_i}. \tag{3.19}$$

leaving out the additive constant. The estimating equation for τ^2 is given by

$$\hat{\tau}^2 = \frac{\sum_{i=1}^{k} w_i^2 \left[(\bar{Y}_i - \hat{\mu})^2 - \sigma_i^2/n_i\right]}{\sum_{j=1}^{k} w_j^2} + \frac{1}{\sum_{i=1}^{k} w_i}, \tag{3.20}$$

and let $\hat{\tau}^2_{\text{REML}}$ denote the REML estimate. Then, a $100(1-\alpha)\%$ confidence interval for τ^2 is given by

$$\begin{aligned} \text{CI}_2(\tau^2): \quad & \left\{\tilde{\tau}^2 \mid -2\left[l_R(\tilde{\tau}^2) - l_R(\hat{\tau}^2_{\text{REML}})\right] < \chi^2_{1;1-\alpha}\right\} \\ & = \left\{\tilde{\tau}^2 \mid l_R(\tilde{\tau}^2) > l_R(\hat{\tau}^2_{\text{REML}}) - \chi^2_{1;1-\alpha}/2\right\}. \end{aligned} \tag{3.21}$$

In practice, the observed sample variances $s_1^2, \ldots, s_k^2$ are substituted for $\sigma_1^2, \ldots, \sigma_k^2$ and then treated as known, true within-population variances in Eqs. (3.18) and (3.21), respectively.

The asymptotic sampling variances of the ML and REML estimators of τ^2 can be obtained by taking the inverse of the Fisher information. Following Viechtbauer (2007), these variances are equal to

$$\text{Var}\left(\hat{\tau}^2_{\text{ML}}\right) = 2\left(\sum_{i=1}^{k} w_i\right)^{-1} \tag{3.22}$$

and

$$\text{Var}\left(\hat{\tau}^2_{\text{REML}}\right) = 2\left(\sum_{i=1}^{k} w_i^2 - 2\,\frac{\sum_{i=1}^{k} w_i^3}{\sum_{j=1}^{k} w_j} + \frac{\left(\sum_{i=1}^{k} w_i^2\right)^2}{\left(\sum_{j=1}^{k} w_j\right)^2}\right)^{-1}, \tag{3.23}$$

respectively. Estimates of the sampling variances are obtained by replacing w_i through $\hat{w}_i = 1/(\hat{\tau}^2_{\text{ML}} + s_i^2/n_i)$ or $\hat{w}_i = 1/(\hat{\tau}^2_{\text{REML}} + s_i^2/n_i)$ in Eqs. (3.22) and (3.23), respectively.

Based on the asymptotic normality of ML and REML estimates, $100(1-\alpha)\%$ Wald-type confidence intervals for τ^2 are given by

$$\text{CI}_3(\tau^2): \hat{\tau}^2_{\text{ML}} \mp \sqrt{\widehat{\text{Var}}\left(\hat{\tau}^2_{\text{ML}}\right)}\, z_{1-\alpha/2} \tag{3.24}$$

and

$$\mathrm{CI}_4(\tau^2) : \hat{\tau}^2_{\mathrm{REML}} \mp \sqrt{\widehat{\mathrm{Var}}\left(\hat{\tau}^2_{\mathrm{REML}}\right)}\; z_{1-\alpha/2}. \tag{3.25}$$

Sidik and Jonkman (2005a) recently suggested a new heterogeneity estimator, see Section 3.1, and, based on this estimator, a method for obtaining a confidence interval for τ^2. The proposed method works as follows. First, a rough estimate of τ^2 is calculated with

$$\tau^2_0 = \frac{1}{k}\sum_{i=1}^{k}\left(\bar{Y}_i - \bar{Y}\right)^2,$$

where $\bar{Y}$ is the sample average of the Y_i's. Next, calculate $\hat{\mu}_0$ with Eq. (3.16), where $w^*_i = 1/(\hat{\tau}^2_0 + S^2_i/n_i)$. The heterogeneity estimator is then given by

$$\hat{\tau}^2_{\mathrm{SJ}} = \frac{\hat{\tau}^2_0}{k-1}\sum_{i=1}^{k} w^*_i\left(\bar{Y}_i - \hat{\mu}_0\right)^2. \tag{3.26}$$

Based on the assumption that $(k-1)\hat{\tau}^2_{\mathrm{SJ}}/\tau^2$ approximately follows a χ^2-distribution with $k-1$ degrees of freedom, an approximative $100(1-\alpha)\%$ confidence interval for τ^2 can be obtained as

$$\mathrm{CI}_5(\tau^2) : \left(\frac{(k-1)\hat{\tau}^2_{\mathrm{SJ}}}{\chi^2_{k-1;1-\alpha/2}}\;,\;\frac{(k-1)\hat{\tau}^2_{\mathrm{SJ}}}{\chi^2_{k-1;\alpha/2}}\right). \tag{3.27}$$

Biggerstaff and Tweedie (1997) proposed a confidence interval for τ^2 based on Cochran's homogeneity test statistic. Recall from Chapter 2 that for known within-study variances this statistic is given as

$$Q_{\mathrm{C}} = \sum_{i=1}^{k} v_i\left(\bar{Y}_{i\cdot} - \sum_{j=1}^{k} h_j\,\bar{Y}_{j\cdot}\right)^2,$$

where $v_i = n_i/\sigma^2_i$, $h_i = v_i/\sum_{i=1}^{k} v_i$. Biggerstaff and Tweedie (1997) approximated the distribution of Q_{C} in the random effects model by a gamma distribution with shape parameter r and scale parameter λ. Setting $\mathrm{E}(Q_{\mathrm{C}}) = r/\lambda$ and $\mathrm{Var}(Q_{\mathrm{C}}) = r/\lambda^2$ and solving for r and λ, we have

$$r = \frac{(\mathrm{E}(Q_{\mathrm{C}}))^2}{\mathrm{Var}(Q_{\mathrm{C}})} \quad \text{and} \quad \lambda = \frac{\mathrm{E}(Q_{\mathrm{C}})}{\mathrm{Var}(Q_{\mathrm{C}})}.$$

Note that in model (3.1) it holds (see Biggerstaff and Tweedie, 1997, and Eq. (3.7) for the expected value)

$$\mathrm{E}(Q_{\mathrm{C}}) = k - 1 + \left(\sum_{i=1}^{k} v_i - \frac{\sum_{i=1}^{k} v_i^2}{\sum_{j=1}^{k} v_j} \right) \tau^2$$

and

$$\mathrm{Var}(Q_{\mathrm{C}}) = 2(k-1) + 4 \left(\sum_{i=1}^{k} v_i - \frac{\sum_{i=1}^{k} v_i^2}{\sum_{j=1}^{k} v_j} \right) \tau^2 + 2 \left(\sum_{i=1}^{k} v_i^2 - 2\frac{\sum_{i=1}^{k} v_i^3}{\sum_{j=1}^{k} v_j} + \frac{\left(\sum_{i=1}^{k} v_i^2 \right)^2}{\left(\sum_{j=1}^{k} v_j \right)^2} \right) \tau^4.$$

Based on this approximation, an approximate distribution of the DerSimonian-Laird estimator $\hat{\tau}^2_{\mathrm{DSL}}$ from Eq. (3.6) is a location-shifted, scaled, gamma distribution. The probability density function $f_{\mathrm{DSL}}(\cdot; \tau^2)$ of $\hat{\tau}^2_{\mathrm{DSL}}$ under this distributional assumption is

$$f_{\mathrm{DSL}}(t; \tau^2) = c \, \frac{\lambda^r}{\Gamma(r)} (c\, t + k - 1)^{r-1} \exp[-\lambda(c\, t + k - 1)] \mathbf{1}_{[-(k-1)/c, \infty)}(t)$$

for $\tau^2 \geq 0$, where $c = (\sum_{i=1}^{k} v_i - \sum_{i=1}^{k} v_i^2 / \sum_{j=1}^{k} v_j)$ and $\mathbf{1}_A(\cdot)$ is the indicator of the set A. Recall that r and λ depend on τ^2.

Biggerstaff and Tweedie (1997) then defined the functions $L(\tau^2)$ and $U(\tau^2)$ by

$$\begin{aligned} L(\tau^2) &= \int_{\hat{\tau}^2_{\mathrm{DSL}}}^{\infty} f_{\mathrm{DSL}}(t; \tau^2) \, \mathrm{d}t \\ U(\tau^2) &= \int_{-(k-1)/c}^{\hat{\tau}^2_{\mathrm{DSL}}} f_{\mathrm{DSL}}(t; \tau^2) \, \mathrm{d}t, \end{aligned}$$

where $\hat{\tau}^2_{\mathrm{DSL}}$ stands here for the observed value of the DerSimonian-Laird estimator. A $100(1-\alpha)\%$ confidence interval for τ^2 is then given as

$$\mathrm{CI}_6(\tau^2) = [\hat{\tau}^2_L, \hat{\tau}^2_U], \tag{3.28}$$

where $\hat{\tau}^2_L$ and $\hat{\tau}^2_U$ are solutions for τ^2 in the equations $L(\tau^2) = \alpha/2$ and $U(\tau^2) = \alpha/2$.

Recently, Hartung and Knapp (2005a) and independently Viechtbauer (2007) proposed a confidence interval using the quadratic form $Q(\tau^2)$ from (3.10) which Mandel and Paule (1970) used for their estimator of τ^2. Hartung and Knapp (2005a) derived the first two moments of $Q(\tau^2)$ and discussed the accuracy of the approximation of the distribution of

$Q(\tau^2)$ to a χ^2-distribution with $k-1$ degrees of freedom. Since $Q(\tau^2)$ is a convex and monotone decreasing function in τ^2 and, thus, proposed a $(1-\alpha)$-confidence region for the among-group variance defined by

$$\mathrm{CI}_7(\tau^2) = \left\{\tau^2 \geq 0 \;\middle|\; \chi^2_{k-1;\alpha/2} \leq Q(\tau^2) \leq \chi^2_{k-1;1-\alpha/2}\right\}. \tag{3.29}$$

Since $Q(\tau^2)$ is a monotone decreasing function in $\tau^2 \geq 0$, the function $Q(\tau^2)$ has its maximal value at $Q(0)$. For $Q(0) < \chi^2_{k-1;\alpha/2}$, we define $C_7(\sigma_a^2) = \{0\}$, otherwise the confidence region $\mathrm{CI}_7(\tau^2)$ is a genuine interval. Note that the validity of the inequality $Q(0) < \chi^2_{k-1;\alpha/2}$ only depends on the choice of the level α. To determine the bounds of the confidence interval one has to solve the two equations for τ^2, namely,

$$\begin{array}{ll} \text{lower bound:} & Q(\tau^2) = \chi^2_{k-1;1-\alpha/2}, \\ \text{upper bound:} & Q(\tau^2) = \chi^2_{k-1;\alpha/2}. \end{array} \tag{3.30}$$

Simulation studies by Hartung and Knapp (2005a), Knapp, Biggerstaff, and Hartung (2006), and Viechtbauer (2007) showed that the interval $\mathrm{CI}_7(\tau^2)$ generally outperforms the other intervals with respect to attaining the nominal confidence coefficient. Biggerstaff and Tweedie's interval based on Cochran's statistic turned out to be rather conservative, especially for large values of heterogeneity. The other intervals are often too liberal, that is, too short. Especially the Wald-type intervals cannot be recommended for practical purposes. The profile restricted maximum likelihood interval $\mathrm{CI}_1(\tau^2)$ behaves well in attaining the nominal confidence coefficient in several scenarios and seems to be the only real competitor to the interval $\mathrm{CI}_7(\tau^2)$.

3.3 Inference on the Overall Mean

In this section, we present some results on estimation, tests and confidence intervals of the overall mean μ. Let us recall $\bar{Y}_i \sim N(\mu, \tau^2 + \sigma_i^2/n_i)$. Then, when the within-study variances are known, the uniformly minimum variance unbiased estimator of μ is given by

$$\hat{\mu} = \bar{Y}_w = \frac{\sum_{i=1}^k w_i\, \bar{Y}_i}{\sum_{j=1}^k w_j},$$

where $w_i = (\tau^2 + \sigma_i^2/n_i)^{-1}$, $i = 1, \ldots, k$. Then it holds for the standardized variable

$$Z = \frac{\bar{Y}_w - \mu}{(\sum_{i=1}^k w_i)^{-1/2}} \sim N(0,1).$$

However, in practice, we have to estimate the usually unknown variances. The within-study variances σ_i^2 are estimated by their sample counterparts, and the between-study variance τ^2 can be estimated using an estimator from the previous two sections. Finally, we obtain an approximate $100(1-\alpha)\%$ confidence interval for μ as

$$\hat{\hat{\mu}} = \hat{\bar{Y}}_w = \frac{\sum_{i=1}^k \hat{w}_i \, \bar{Y}_i}{\sum_{j=1}^k \hat{w}_j} \pm \left(\sum_{i=1}^k \hat{w}_i \right)^{-1/2} z_{1-\alpha/2} \tag{3.31}$$

with $\hat{w}_i = (\hat{\tau}^2 + S_i^2/n_i)^{-1}$.

As is well known, in small to moderate number of studies, which is mostly the case in applications, the confidence interval (3.31) suffers from the same weaknesses as its fixed effects counterpart. Namely, the actual confidence coefficient is below the nominal one. Consequently, the corresponding test on the overall mean yields too many unjustified significant results.

Hartung and Knapp (2001a,b) considered the residual sum of squares

$$Q = \sum_{i=1}^k w_i \, (\bar{Y}_i - \bar{Y}_w)^2, \tag{3.32}$$

which is a chi-square random variable with $k-1$ degrees of freedom and stochastically independent of $\bar{Y}_w$. Moreover,

$$Q^* = \widehat{\mathrm{Var}}(\bar{Y}_w) = \frac{1}{k-1} \frac{\sum_{i=1}^k w_i \, (\bar{Y}_i - \bar{Y}_w)^2}{\sum_{j=1}^k w_j} \tag{3.33}$$

is an unbiased estimator of the variance of $\hat{\mu}$ in model (3.1). Consequently,

$$t = \frac{\bar{Y}_w - \mu}{\sqrt{\widehat{\mathrm{Var}}(\bar{Y}_w)}} \tag{3.34}$$

is a t-distributed random variable with $k-1$ degrees of freedom. The test statistic t depends on the unknown variance components which have to be replaced by appropriate

estimates in practice. By substituting the variance components by their estimates, the resulting test statistic is then approximately t-distributed with $k-1$ degrees of freedom. So, the alternative approximate $100(1-\alpha)\%$-confidence interval for μ reads

$$\hat{\hat{\mu}} = \hat{\bar{Y}}_w = \frac{\sum_{i=1}^k \hat{w}_i \, \bar{Y}_i}{\sum_{i=1}^k \hat{w}_i} \pm \sqrt{\hat{Q}^*} \; t_{k-1,1-\alpha/2} \tag{3.35}$$

with $\hat{Q}^*$ the variance estimator according Eq. (3.33), where w_i is replaced by $\hat{w}_i$.

Hartung and Knapp (2001a,b) conducted an extensive simulation study to compare the attained type I error rates for the commonly used confidence interval (3.31) and the proposed modified confidence interval (3.35). It turns out that the interval (3.35) greatly improves the attained confidence coefficient. Moreover, the good performance of the interval (3.35) does not heavily depend on the estimator of the between-study variance used in the analysis, while the performance of the interval (3.31) can be dramatically affect for different estimators of τ^2.

An exact test for μ in the present model is described in Iyer, Wang, and Mathew (2004), using the notion of the generalized confidence intervals. The general concept of generalized confidence intervals has been already introduced in Section 2.3. Basically, the approach by Iyer, Wang, and Mathew (2004) is similar to the approach by Lin and Lee (2005) in the common mean problem. The important contribution by Iyer, Wang, and Mathew is the generalized pivotal quantity for τ^2 based on the residual sum of squares (3.32).

Consider the set of $k+2$ statistics $(\bar{Y}_w, Q, S_i^2, i = 1, \ldots, k)$. Recall that

$$Z = \frac{\bar{Y}_w - \mu}{\sqrt{1/\sum_{i=1}^k w_i}} \sim N(0,1),$$

$$U_i = \frac{(n_i-1)S_i^2}{\sigma_i^2} = \frac{V_i}{\sigma_i^2} \sim \chi^2_{n_i-1}, \quad i = 1, \ldots, k,$$

and

$$Q = \sum_{i=1}^k w_i \, (\bar{Y}_i - \hat{\mu})^2 \sim \chi^2_{k-1}$$

are pivotal quantities, and let $\bar{y}_i$ and s_i^2 denote the observed values of $\bar{Y}_i$ and S_i^2, and v_i stands for the observed value of V_i.

The generalized pivotal quantity for τ^2 can be obtained through an implicit expression for τ^2 given as the solution to the equation

$$Q = \sum_{i=1}^{k} c_i \left(\bar{Y}_i - \sum_{i=1}^{k} c_i \ \bar{Y}_i \Big/ \sum_{j=1}^{k} c_j \right)^2 = \tilde{Q}(\tau^2)$$

with $c_i = 1/[\tau^2 + (\sigma_i^2/(n_i \ U_i)]$, $i = 1, \ldots, k$. The solution for τ^2 is unique, since $\tilde{Q}(\tau^2)$ is a decreasing function of τ^2, and the maximum value is given at $\tau^2 = 0$. Consequently, given a real number $q \geq 0$, there must exist a unique $\tau_*^2 \geq 0$, such that $\tilde{Q}(\tau_*^2) = q$, provided $q \leq \tilde{Q}(0)$.

Define the function

$$h(q) = \begin{cases} \tau_*^2, & \text{if } 0 \leq q \leq \tilde{Q}(0), \\ 0, & \text{otherwise.} \end{cases}$$

Let $\boldsymbol{D} = (\bar{Y}_1, \ldots, \bar{Y}_k, V_1, \ldots, V_k)'$ be the vector of the sufficient statistics and let $\boldsymbol{d} = (\bar{y}_1, \ldots, \bar{y}_k, v_1, \ldots, v_k)'$ be the vector of corresponding observed values.

Define

$$\boldsymbol{T} = \left(\frac{\sigma_1^2 \ ss_1}{n_1 \ SS_1}, \ldots, \frac{\sigma_k^2 \ ss_k}{n_k \ SS_k} \right)' = [ss_1/(n_1 \ Q_1), \ldots, ss_k/(n_k \ Q_k)]' = (T_1, \ldots, T_k)' \, .$$

Note that $\boldsymbol{T}$ is a random vector whose distribution is free of any model parameters, and the observed value of $\boldsymbol{T}$ is $(\sigma_1^2/n_1, \ldots, \sigma_k^2/n_k)'$.

Define

$$W_i = \left(h(Q) + \frac{v_i}{n_i \ U_i} \right)^{-1}, i = 1, \ldots, k, \tag{3.36}$$

where W_i is a random variable whose distribution are free of any model parameters. Note that when the observed statistics $\boldsymbol{d}$ are substituted in $h(Q)$, it reduces to τ^2. Thus, when the observed values $\boldsymbol{d}$ are substituted for $\boldsymbol{D}$ in W_i, the observed value is $1/(\tau^2 + \sigma_i^2/n_i)$.

Denote $\boldsymbol{\theta} = (\mu, \tau^2, \sigma_1^2, \ldots, \sigma_k^2)'$, then a generalized pivotal quantity for μ is given as

$$\begin{aligned} R = R(\boldsymbol{D}; \boldsymbol{d}, \boldsymbol{\theta}) &= \frac{\sum_{i=1}^{k} W_i \ \bar{y}_i}{\sum_{j=1}^{k} W_j} - \left(\frac{\bar{Y}_W - \mu}{\sqrt{1/\sum_{i=1}^{k} w_i}} \right) \left(\sum_{i=1}^{k} W_i \right)^{-1/2} \\ &= \frac{\sum_{i=1}^{k} W_i \ \bar{y}_i}{\sum_{j=1}^{k} W_j} - Z \left(\sum_{i=1}^{k} \left(h(Q(\tau^2)) + \frac{\sigma_i^2 \ ss_i}{n_i \ SS_i} \right)^{-1} \right)^{-1/2} . \end{aligned} \tag{3.37}$$

Note that the distribution of R is free of any model parameters and $R(\boldsymbol{d};\boldsymbol{d},\boldsymbol{\theta}) = \mu$. Thus, R fulfills the requirements to be a generalized pivotal quantity. In actual applications, when closed-form expressions for the required quantiles are unavailable, they may be estimated by simulating the distribution of $R(\boldsymbol{D};\boldsymbol{d},\boldsymbol{\theta})$ using the following algorithm:
For given data $(\bar{y}_i, s_i^2, n_i)$, $i = 1, \ldots k$:

For $j = 1, \ldots, m$:

1. Generate $Z \sim N(0,1)$.
2. Generate $U_i \sim \chi^2_{n_i-1}$, $i = 1, \ldots, k$.
3. Generate $Q \sim \chi^2_{k-1}$.
4. Calculate T_i, $i = 1, \ldots, k$.
5. Calculate $\tilde{Q}(0)$.
6. If $0 \leq Q \leq \tilde{Q}(0)$, find τ_*^2 such that $\tilde{Q}(\tau_*^2) = Q$, otherwise set $\tau_*^2 = 0$.
7. Calculate $W_i = 1/[\tau_*^2 + ss_i/(n_i\, U_i)]$, $i = 1, \ldots, k$.
8. Calculate $\bar{y}_W = \sum_{i=1}^k W_i\, \bar{Y}_i / \sum_{j=1}^k W_j$.
9. Calculate $R(\boldsymbol{D};\boldsymbol{d},\boldsymbol{\theta})_j = R_j$.

(end j loop)

Compute the $\alpha/2$- and $(1-\alpha/2)$-quantile of $R_1, \ldots, R_m$.
Then, $(R_{\alpha/2}, R_{1-\alpha/2})$ is a $100(1-\alpha)\%$ generalized confidence interval on μ.

Note that the above algorithm until step 6 can be used to simulate the distribution of the generalized pivotal quantity for τ^2, and thus, one can compute a generalized confidence interval for τ^2.

Moreover, using W_i from Eq. (3.36), $i = 1, \ldots, k$, and following the lines of the third generalized pivotal quantity for the common mean in Section 2.3, a further generalized pivotal quantity for the overall mean μ is given in the present model as

$$\begin{aligned} S &= S\left(\boldsymbol{D};\boldsymbol{d},\boldsymbol{\theta}\right) \\ &= \frac{\sum\limits_{i=1}^{k} W_i\ \bar{y}_i}{\sum\limits_{j=1}^{k} W_j} - t_{k-1}\sqrt{\frac{1}{k-1}\left(\sum_{i=1}^{k} W_i\right)^{-1} \sum_{i=1}^{k} W_i \left(\bar{y}_i - \frac{\sum_{j=1}^{k} W_j\ \bar{y}_j}{\sum_{\ell=1}^{k} W_\ell}\right)^2}, \end{aligned} \tag{3.38}$$

where t_{k-1} denotes a t-distributed random variable with $k-1$ degrees of freedom. Note that the distribution of S is free of any model parameters and $S(\boldsymbol{d};\boldsymbol{d},\boldsymbol{\theta}) = \mu$. The above

algorithm can be used for simulating the distribution of S by appropriately changing step 9 into calculate $S(\boldsymbol{D}; \boldsymbol{d}, \boldsymbol{\theta})_j = S_j$. Then, compute the $\alpha/2$- and $(1-\alpha/2)$-quantile of $S_1, \ldots, S_m$. Finally, $(S_{\alpha/2}, S_{1-\alpha/2})$ is a $100(1-\alpha)\%$ generalized confidence interval on μ.

3.4 A General Weighting Scheme

In the previous sections as well as in Chapter 2, the weights have been always chosen as the inverses of the variances of the sample means or the inverses of their estimators for practical purposes. Though this choice is an optimal one in a certain sense, practically, however, it may be possible that the overall conclusion from combining results of independent studies using the inverse variance method may not be reasonable. Recall that the smaller the variance of the sample mean of a study the higher the precision and, thus, the more influential the result of the study in the overall analysis. The magnitude of the variance is determined by the ratio of the population variance σ_i^2 and the sample size n_i. If n_i is large, one will be confident in giving the study a large weight. But if n_i is small or moderate and the population variance, or more exactly the estimate of the variance is close to zero, the study will get a large weight and can possibly dominate the overall analysis irrespective of how large the other studies are. This latter scenario may be a reason for searching for different weighting schemes provided by some external process.

Hartung, Böckenhoff, and Knapp (2003) discussed in detail statistical methods for combining results with an arbitrary but fixed weighting scheme. In the sequel, we summarize some main ideas and results.

Let us consider

$$\bar{Y}_i \sim N(\mu, \alpha_i),\ i = 1, \ldots, k, \tag{3.39}$$

where α_i is a general variance. For $\alpha_i = \sigma_i^2/n_i$, we have the common mean problem, for $\alpha_i = \tau^2 + \sigma_i^2/n_i$, we have the one-way random effects model.

Let $b = (b_1, \ldots, b_k)'$ denote an arbitrary but fixed vector of standardized weights, that is, $b_i \geq 0$, $i = 1, \ldots, k$, and $\sum_{i=1}^{k} b_i^2 = 1$. Then, clearly,

$$\hat{\mu}_b = \sum_{i=1}^{k} b_i^2 \ \bar{Y}_i \tag{3.40}$$

is an unbiased estimator of μ with variance

$$\text{Var}\,(\hat{\mu}_b) = \sum_{i=1}^{k} b_i^4 \ \alpha_i. \tag{3.41}$$

Possible choices of b_i, $i = 1, \ldots, k$, may be

$$b_i = \sqrt{n_i/\sigma_i^2} \ \Big/ \ \sqrt{\sum_{j=1}^{k} n_j/\sigma_j^2}, \tag{3.42}$$

(weights from common mean problem)

$$b_i = \sqrt{1/(\tau^2 + \sigma_i^2/n_i)} \ \Big/ \ \sqrt{\sum_{j=1}^{k} (1/(\tau^2 + \sigma_j^2/n_j)}, \tag{3.43}$$

(weights from one-way random effects model)

$$b_i = \sqrt{1/k}, \tag{3.44}$$

(equal weights)

or

$$b_i = \sqrt{n_i / \sum_{j=1}^{k} n_j}. \tag{3.45}$$

(sample size based weights)

For further statistical inference on μ using the estimator $\hat{\mu}_b$, estimators of $\text{Var}(\hat{\mu}_b)$ and α_i, $i = 1, \ldots, k$, are required. Hartung, Böckenhoff, and Knapp (2003) considered the basic statistics

$$u_{ib}^2 = b_i^2 \Big(\bar{Y}_i - \sum_{j=1}^{k} b_j^2 \ \bar{Y}_j \Big)^2 = b_i^2 \left(\bar{Y}_i - \hat{\mu}_b \right)^2. \tag{3.46}$$

They showed that $\sum_{i=1}^{k} \alpha_i^{-1}\, b_i^2\, \bar{Y}_i$ and u_{ib}^2 are stochastically independent, $i = 1, \ldots, k$, and, using Patnaik's (1949) method of moments matching approach, that, for $d = (d_1, \ldots, d_k)$, $d_i \in I\!R$, $i = 1, \ldots, k$, it holds

$$\nu_d\, \frac{\sum_{i=1}^{k} d_i\, u_{ib}^2}{\sum_{i=1}^{k} d_i\, \mathrm{E}(u_{ib}^2)} \overset{\text{appr.}}{\sim} \chi_{\nu_d}^2 \tag{3.47}$$

with

$$\nu_d = \frac{\Big(\sum_{i=1}^{k} d_i\, \mathrm{E}(u_{ib}^2)\Big)^2}{\sum_{i=1}^{k} d_i^2\, \mathrm{Var}(u_{ib}^2) + 2 \sum_{i=1}^{k} \sum_{j>i}^{k} d_i\, d_j\, \mathrm{Cov}(u_{ib}^2, u_{jb}^2)}, \tag{3.48}$$

$$\mathrm{E}(u_{ib}^2) = (1 - 2\, b_i^2)\, b_i^2\, \alpha_i + b_i^2 \sum_{j=1}^{k} b_j^4\, \alpha_j,$$

$$\mathrm{Var}(u_{ib}^2) = 2\, \Big[\mathrm{E}(u_{ib}^2)\Big]^2,$$

and

$$\mathrm{Cov}(u_{ib}^2, u_{jb}^2) = 2\, b_i^2\, b_j^2 \Big(\sum_{\ell=1}^{k} b_\ell^4\, \alpha_\ell - b_i^2\, \alpha_i - b_j^2\, \alpha_j \Big)^2.$$

Note that the degrees of freedom ν_d still contain the unknown general variances α_i. In practice, appropriate estimates of α_i have to be plugged in.

Furthermore, Hartung, Böckenhoff, and Knapp (2003) showed that

$$\widehat{\mathrm{Var}}(\hat{\mu}_b) = \frac{1}{1 + \sum_{j=1}^{k} b_j^4/(1 - 2\, b_j^2)} \sum_{i=1}^{k} \frac{b_i^2}{1 - 2\, b_i^2}\, u_{ib}^2 \tag{3.49}$$

is an non-negative unbiased estimator of $\mathrm{Var}(\hat{\mu}_b)$, if $b_i^2 < 1/2$, $i = 1, \ldots, k$, and $k \geq 3$. Note that $b_i^2 < 1/2$, $i = 1, \ldots, k$, is sufficient but not necessary for the non-negativity of $\widehat{\mathrm{Var}}(\hat{\mu}_b)$. Consequently, since $\hat{\mu}_b$ and $\widehat{\mathrm{Var}}(\hat{\mu}_b)$ are stochastically independent, it holds

$$\frac{\hat{\mu}_b - \mu}{\sqrt{\widehat{\mathrm{Var}}(\hat{\mu}_b)}} \overset{\text{appr.}}{\sim} t_{\nu(b)}, \tag{3.50}$$

where $\nu(b)$ can be determined according to (3.48) noting that $\widehat{\mathrm{Var}}(\hat{\mu}_b)$ can be expressed as $\sum_{i=1}^{k} d_i\, u_{ib}^2$ with

$$d_i = \frac{b_i^2/(1 - 2\, b_i^2)}{1 + \sum_{j=1}^{k} b_j^4/(1 - 2\, b_j^2)},\ i = 1, \ldots, k.$$

An approximated $100(1-\alpha)\%$ confidence interval for μ is then given as

$$\hat{\mu}_b \pm \sqrt{\widehat{\mathrm{Var}}(\hat{\mu}_b)}\; t_{\hat{\nu}(b);1-\alpha/2}, \tag{3.51}$$

where $\hat{\nu}(b)$ stands for the estimated degrees of freedom.

For more sophisticated methods involving quadratic estimation of α_i using C. R. Rao's (1972) MINQUE principle and Hartung's (1981) concept of nonnegative minimum biased invariant quadratic estimation of variance components, let us refer to Hartung, Böckenhoff, and Knapp (2003). It is worth mentioning that Hartung and Knapp (2003) proposed also confidence regions for the general variance components in the present setting.

Chapter 4

Combining Results of Controlled Studies with Normal Response

The fundamentals for combining results from several independent studies or experiments were extensively discussed in the previous two chapters. The methods presented there heavily rely on the assumptions that we have normal means and variance estimators of the means which are stochastically independent of the sample means and follow exactly independent scaled chi-square distributions. Moreover, the methods were presented for one-sample studies or experiments only.

In this chapter we discuss methods for combining results from comparative studies, say treatment (T) versus control (C), with normal outcomes and show which methods of Chapter 2 and 3 can be applied or extended in the present scenario.

Let us assume that, in general, there are k independent studies comparing a treatment (T) versus a control (C). Let $\bar{Y}_{Ti}$ and S^2_{Ti} denote the sample mean and the sample variance of the treatment group in the ith study, let be n_{Ti} the corresponding sample size. Let $\bar{Y}_{Ci}$ and S^2_{Ci} denote the sample mean and the sample variance of the control group in the ith study, let be n_{Ci} the corresponding sample size. Then it holds for $i = 1, \ldots, k$,

$$\bar{Y}_{Ti} \sim N\left(\mu_{Ti}, \frac{\sigma^2_{Ti}}{n_{Ti}}\right), \quad (n_{Ti} - 1)\ S^2_{Ti} \sim \sigma^2_{Ti}\ \chi^2_{n_{Ti}-1}, \tag{4.1}$$

and

$$\bar{Y}_{Ci} \sim N\left(\mu_{Ci}, \frac{\sigma_{Ci}^2}{n_{Ci}}\right), \quad (n_{Ci}-1)\ S_{Ci}^2 \sim \sigma_{Ci}^2\ \chi_{n_{Ci}-1}^2, \tag{4.2}$$

where μ_{Ti} and μ_{Ci} are the means of the treatment and control group, respectively, and σ_{Ti}^2 and σ_{Ci}^2 are the corresponding variances. Note that the statistics (4.1) and (4.2) are all mutually independent. Assuming that in each study the population variances are identical, that is, $\sigma_i^2 = \sigma_{Ti}^2 = \sigma_{Ci}^2$, $i = 1, \ldots, k$, then the pooled sample variance is given by

$$S_i^{*^2} = \frac{1}{n_{Ti} + n_{Ci} - 2}\left[(n_{Ti}-1)S_{Ti}^2 + (n_{Ci}-1)S_{Ci}^2\right], \tag{4.3}$$

and it follows that

$$(n_{Ti} + n_{Ci} - 2)\ S_i^{*^2} \sim \sigma_i^2\ \chi_{n_{Ti}+n_{Ci}-2}^2. \tag{4.4}$$

First, we have to decide which effect size we use for describing the difference between treatment and control group. The following three effect sizes are widely used:

- Difference of means:

$$\mu_{Di} = \mu_{T_i} - \mu_{Ci}.$$

- Standardized difference of means:

$$\theta_i = \frac{\mu_{Ti} - \mu_{Ci}}{\sigma_i},$$

  where σ_i denotes a suitable standard deviation, for instance, an average of the population standard deviations σ_{Ti} and σ_{Ci}.

- Ratio of means:

$$\rho_i = \frac{\mu_{Ti}}{\mu_{Ci}}, \quad \mu_{Ci} \neq 0.$$

We will discuss methods for combining results from independent studies using the different effect sizes in the following three sections.

4.1 Difference of Means

Let $\mu_{Di} = \mu_{Ti} - \mu_{Ci}$, $i = 1, \ldots, k$, be the parameter of interest in each study, then the difference of the sample means, $D_i = \bar{Y}_{Ti} - \bar{Y}_{Ci}$, is an unbiased estimator of μ_{Di} with

$$D_i \sim N\left(\mu_{Di}, \frac{\sigma_{Ti}^2}{n_{Ti}} + \frac{\sigma_{Ci}^2}{n_{Ci}}\right)$$

in general or

$$D_i \sim N\left(\mu_{Di}, \frac{n_{Ti} + n_{Ci}}{n_{Ti}\ n_{Ci}}\ \sigma_i^2\right)$$

for identical population variances in each study.

The variance of D_i can be unbiasedly estimated either by

$$\widehat{\mathrm{Var}}(D_i) = \frac{S_{Ti}^2}{n_{Ti}} + \frac{S_{Ci}^2}{n_{Ci}} \quad \text{or by} \quad \widehat{\mathrm{Var}}(D_i) = \frac{n_{Ti} + n_{Ci}}{n_{Ti}\ n_{Ci}}\ S_i^{*2}.$$

Note that the latter variance estimator is an exactly scaled chi-square distributed random variable, see (4.4), whereas the distribution of $S_{Ti}^2/n_{Ti} + S_{Ci}^2/n_{Ci}$, which is a linear combination of two independent scaled chi-square variables, can only be approximated, for instance, by Satterthwaite's (1946) approximation if the population variances are different. In this case, the Satterthwaite approximation yields

$$\nu_i \left(\frac{S_{Ti}^2}{n_{Ti}} + \frac{S_{Ci}^2}{n_{Ci}}\right) \overset{\text{approx.}}{\sim} \left(\frac{\sigma_{Ti}^2}{n_{Ti}} + \frac{\sigma_{Ci}^2}{n_{Ci}}\right)\ \chi_{\nu_i}^2$$

with

$$\nu_i = \frac{\left(\sigma_{Ti}^2/n_{Ti} + \sigma_{Ci}^2/n_{Ci}\right)^2}{(\sigma_{Ti}^2/n_{Ti})^2/(n_{Ti}-1) + (\sigma_{Ci}^2/n_{Ci})^2/(n_{Ci}-1)}.$$

Since the degrees of freedom depend on the unknown variances, they must be estimated in practice by

$$\hat{\nu}_i = \frac{\left(S_{Ti}^2/n_{Ti} + S_{Ci}^2/n_{Ci}\right)^2}{(S_{Ti}^2/n_{Ti})^2/(n_{Ti}-1) + (S_{Ci}^2/n_{Ci})^2/(n_{Ci}-1)}.$$

In case the assumption of equal variances in each study is fulfilled, we can directly use all the results from Chapter 2 in a fixed effects model or all the results from Chapter 3 in a random effects model. Under the assumption of equality of differences of means, that is, it holds

$$H_0 : \mu_{D1} = \mu_{D2} = \cdots = \mu_{Dk} =: \mu_D,$$

we have the common mean problem from Chapter 2. The fixed effects model is then given as

$$D_i \sim N\left(\mu_D, \frac{n_{Ti}+n_{Ci}}{n_{Ti}\ n_{Ci}}\ \sigma_i^2\right),\ \frac{(n_{Ti}+n_{Ci}-2)S_i^{*^2}}{\sigma_i^2} \sim \chi^2_{n_{Ti}+n_{Ci}-2}, \quad i=1,\ldots,k. \tag{4.5}$$

By replacing the sample mean $\bar{Y}_i$ through D_i, the variance estimator S_i^2/n_i through $(1/n_{Ti}+1/n_{Ci})S_i^{*^2}$, and the degrees of freedom n_i-1 through $n_{Ti}+n_{Ci}-2$, all the results from the common mean problem can be easily transferred to the analysis in model (4.5), even the exact as well as the generalized confidence intervals for μ_D!

The random effects model is given for $i=1,\ldots,k$, as

$$D_i \sim N\left(\mu_D, \tau^2+\frac{n_{Ti}+n_{Ci}}{n_{Ti}\ n_{Ci}}\ \sigma_i^2\right),\ \frac{(n_{Ti}+n_{Ci}-2)S_i^{*^2}}{\sigma_i^2} \sim \chi^2_{n_{Ti}+n_{Ci}-2}, \tag{4.6}$$

where τ^2 again denotes the heterogeneity parameter. By carrying out the same replacement as above, the results from the one-way random effects model can be transferred to the analysis in model (4.6), even the generalized confidence intervals for μ_D!

When the variances of treatment and control group are not identical in each study we have to use the random effects model

$$D_i \sim N\left(\mu_D, \tau^2+\frac{\sigma_{Ti}^2}{n_{Ti}}+\frac{\sigma_{Ci}^2}{n_{Ci}}\right),\ \nu_i\left(\frac{S_{Ti}^2}{n_{Ti}}+\frac{S_{Ci}^2}{n_{Ci}}\right) \overset{\text{approx.}}{\sim} \left(\frac{\sigma_{Ti}^2}{n_{Ti}}+\frac{\sigma_{Ci}^2}{n_{Ci}}\right)\ \chi^2_{\nu_i}, \tag{4.7}$$

$i=1,\ldots,k$. With $\tau^2=0$, we obtain the corresponding fixed effects model. Recall that the degrees of freedom ν_i of the approximate χ^2-distribution depend on the unknown variances σ_{Ti}^2 and σ_{Ci}^2 and have to be estimated in practice. Consequently, by replacing $\bar{Y}_i$ through D_i, S_i^2/n_i through $S_{Ti}^2/n_{Ti}+S_{Ci}^2/n_{Ci}$, and the degrees of freedom n_i-1 by ν_i or $\hat{\nu}_i$, respectively, the exact methods from Chapter 2 are no longer exact, but still approximately valid.

Let $\hat{\tau}^2$ be an estimator of τ^2, then

$$\hat{w}_i = \left(\hat{\tau}^2+\frac{S_{Ti}^2}{n_{Ti}}+\frac{S_{Ci}^2}{n_{Ci}}\right)^{-1}, i=1,\ldots,k,$$

are the estimated weights in the random effects model (4.7). Note that $\hat{v}_i=(S_{Ti}^2/n_{Ti}+S_{Ci}^2/n_{Ci})^{-1}$, $i=1,\ldots,k$, are the corresponding weights in the fixed effects model.

Then

$$\hat{\mu}_{D,\hat{w}} = \frac{\sum_{i=1}^{k} \hat{w}_i \; D_i}{\sum_{j=1}^{k} \hat{w}_j}$$

is an estimator of μ_D. Note that

$$\hat{\mu}_{D,\hat{v}} = \frac{\sum_{i=1}^{k} \hat{v}_i \; D_i}{\sum_{j=1}^{k} \hat{v}_j}$$

is an unbiased estimator of μ_D, since $\bar{Y}_{Ti}$, $\bar{Y}_{Ci}$, S^2_{Ti}, and S^2_{Ci} are mutually independent.

An approximate $100(1-\alpha)\%$ confidence interval on μ_D is given in analogy to interval (3.31) as

$$\frac{\sum_{i=1}^{k} \hat{w}_i \; D_i}{\sum_{j=1}^{k} \hat{w}_j} \mp \left(\sum_{i=1}^{k} \hat{w}_i \right)^{-1/2} z_{1-\alpha/2} \tag{4.8}$$

and a further approximate $100(1-\alpha)\%$ confidence interval on μ_D is given in analogy to interval (3.35) as

$$\frac{\sum_{i=1}^{k} \hat{w}_i \; D_i}{\sum_{j=1}^{k} \hat{w}_j} \mp \sqrt{\frac{1}{k-1} \frac{\sum_{i=1}^{k} w_i \, (D_i - \hat{\mu}_{D,\hat{w}})^2}{\sum_{j=1}^{k} w_j}} \; t_{k-1;1-\alpha/2}. \tag{4.9}$$

Hartung and Knapp (2001a) conducted a simulation study to compare the actual confidence coefficients of the two approximate confidence intervals (4.8) and (4.9) on μ_D. In their simulation study, Hartung and Knapp used the DerSimonian-Laird estimator of τ^2, which is given here in its truncated form as

$$\hat{\tau}^2_{\text{DSL}} = \max \left\{ 0 \; , \; \frac{\sum_{i=1}^{k} \hat{v}_i \; (D_i - \hat{\mu}_{D,\hat{v}})^2 - (k-1)}{\sum_{i=1}^{k} \hat{v}_i - \sum_{j=1}^{k} \hat{v}_j^2 / \sum_{\ell=1}^{k} \hat{v}_\ell} \right\}.$$

Hartung and Knapp (2001a) showed in their simulation study that the interval (4.8) is very liberal for k up to 12 studies in the fixed effects model, when the samples sizes in both groups are small. With increasing sample sizes in the groups, the actual confidence level of the interval moves towards the nominal one. The interval (4.9) maintains in most cases the nominal level in this model, except for small samples, that is, $n_{Ti} = n_{Ci} = 5$, $i = 1, \ldots, k$, where the interval is also liberal, but still better than the interval (4.8) in terms of having an actual confidence level closer to the nominal one.

In the random effects model, interval (4.8) does not yield acceptable actual coverage probabilities when the amount of heterogeneity is moderate or large. The larger the amount of heterogeneity the more liberal is the interval. The interval (4.9), however, mostly has actual confidence coefficients close to the nominal one and this property holds irrespective of the amount of heterogeneity. Like in the fixed effects model, the interval (4.9) is a little bit liberal only for small sample sizes in the groups. Summarizing, the interval (4.9) can be generally recommended when difference of means of several independent experiments are to be combined and the variances in treatment and control group differ in each study. Even in the fixed effects model and in case of homogeneous group variances within the studies, the interval (4.9) possesses acceptable actual confidence levels compared to the nominal one and can be a serious competitor in practice to the more sophisticated exact methods due to its ease of computation.

4.2 Standardized Difference of Means

Recall that the standardized mean difference as an effect size based on means is given as

$$\theta_i = \frac{\mu_{Ti} - \mu_{Ci}}{\sigma_i}.$$

An natural estimator of θ_i is given by

$$\hat{\theta}_i = \frac{\bar{Y}_{Ti} - \bar{Y}_{Ci}}{\hat{\sigma}_i}$$

with $\hat{\sigma}_i$ an suitable estimator of standard deviation σ_i.

One estimator of θ_i, known as Cohen's d (Cohen, 1969), uses

$$S_i^2 = \frac{1}{n_{Ti} + n_{Ci}} \left[(n_{Ti} - 1) S_{Ti}^2 + (n_{Ci} - 1) S_{Ci}^2 \right]$$

as estimator of σ_i^2, that is,

$$d_i = \frac{\bar{Y}_{Ti} - \bar{Y}_{Ci}}{S_i}, \quad i = 1, \ldots, k.$$

Note that Cohen's d is the maximum likelihood estimator of the standardized mean difference under normality.

A second estimator of θ_i, known as Hedges's g (Hedges, 1981, 1982), is defined as

$$g_i = \frac{\bar{Y}_{Ti} - \bar{Y}_{Ci}}{S_i^*}, \quad i = 1, \ldots, k, \tag{4.10}$$

with $S_i^{*^2}$ from (4.3). Note that $S_i^{*^2}$ is an unbiased estimator of a common variance σ_i^2 in the ith study.

Finally, a third estimator measure of θ_i, known as Glass's Δ (Glass, McGaw, and Smith, 1981), is defined as

$$\Delta_i = \frac{\bar{Y}_{Ti} - \bar{Y}_{Ci}}{S_{Ci}}, \quad i = 1, \ldots, k, \tag{4.11}$$

where the standardized quantity is just the sample standard deviation based on the control group alone. This is typically justified on the ground that the control group is in existence for a longer period than the experimental group, and is likely to provide a more stable estimate of the common variance. Moreover, this estimator is often used when several treatments are compared with one control within a study.

In this section, however, we will exclusively consider Hedges's g_i as the estimator of θ_i, $i = 1, \ldots, k$, for ease of presentation.

It can be shown that (see Hedges and Olkin, 1985)

$$\mathrm{E}(g_i) \approx \theta + \frac{3\,\theta}{4n_i - 9}, \tag{4.12}$$

$$\mathrm{Var}(g_i) \approx \frac{1}{\tilde{n}_i} + \frac{\theta^2}{2(n_i - 3.94)}, \tag{4.13}$$

where

$$n_i = n_{Ti} + n_{Ci}, \;\; \tilde{n}_i = \frac{n_{Ti}\, n_{Ci}}{n_{Ti} + n_{Ci}}.$$

In case the population variances are identical in both groups, under the assumption of normality of the data, Hedges (1981) showed that $\sqrt{\tilde{n}_i}\, g_i$ follows a noncentral t-distribution with noncentrality parameter $\sqrt{\tilde{n}_i}\, \theta_i$ and $(n_{Ti}+n_{Ci}-2)$ degrees of freedom. Consequently, the exact mean and variance of Hedges's g_i are given by

$$\mathrm{E}(g_i) = \sqrt{\frac{n_i - 2}{2}}\; \frac{\Gamma\left(n_i/2 - 3/2\right)}{\Gamma\left(n_i/2 - 1\right)}\; \theta_i, \tag{4.14}$$

$$\mathrm{Var}(g_i) = \frac{n_i - 2}{n_i - 4}\; (1 + \theta_i^2) - \theta_i^2\; \frac{n_i - 2}{2}\; \frac{\left[\Gamma\left(n_i/2 - 3/2\right)\right]^2}{\left[\Gamma\left(n_i/2 - 1\right)\right]^2}, \tag{4.15}$$

and $\Gamma(\cdot)$ denotes the gamma function. Note that the variance $\mathrm{Var}(g_i)$ depends on the effect size θ_i.

Since g_i is biased for θ_i, an approximately unbiased estimate of θ_i is given as

$$g_i^* = \left(1 - \frac{3}{4n_i - 9}\right)\ g_i,$$

see Hedges (1981). For increasing total sample n_i, the correction term approaches one, so that the large sample distributions of g_i and g_i^* are identical.

The variance of g_i^* in large samples is given as

$$\mathrm{Var}(g_i^*) \approx \frac{n_{Ti} + n_{Ci}}{n_{Ti}\ n_{Ci}} + \frac{\theta_i^2}{2\ (n_{Ti} + n_{Ci} - 2)},$$

which can be estimated by

$$\widehat{\mathrm{Var}}(g_i^*) = \frac{n_{Ti} + n_{Ci}}{n_{Ti}\ n_{Ci}} + \frac{g_i^2}{2\ (n_{Ti} + n_{Ci} - 2)}.$$

Note that for $\theta_i = 0$, the large sample variance of g_i^* reduces to $(n_{Ti} + n_{Ci})/(n_{Ti}\ n_{Ci})$ and does not depend on θ_i. Otherwise the large sample variance depends on the unknown standardized mean difference and, generally, g_i^* and $\widehat{\mathrm{Var}}(g_i^*)$ are correlated. Consequently, one may seek for a variance-stabilizing transformation of the estimator g_i^*.

Following Hedges and Olkin (1985), the variance-stabilizing transformation of g_i^* is given by

$$h(g_i^*) = \sqrt{2}\ \sinh^{-1}(g_i^*/a_i) = \sqrt{2}\ \ln\left(\frac{g_i^*}{a_i} + \sqrt{\frac{(g_i^*)^2}{a_i^2} + 1}\ \right)$$

with

$$a_i = \sqrt{4 + 2(n_{Ti}/n_{Ci}) + 2(n_{Ci}/n_{Ti})}\ .$$

Note that the exact form of the transformation of g_i^* depends on the balance n_{Ti}/n_{Ci}. For the balanced case $n_{Ti} = n_{Ci}$, it holds $a_i = \sqrt{8}$.

Let $h(\delta_i)$ denote the transformed parameter, then it holds approximately

$$\sqrt{n_i}\ \ [h(g_i^*) - h(\delta_i)] \sim N(0, 1),$$

or, equivalently,

$$h(g_i^*) \sim N\left(h(\delta_i), \frac{1}{n_i}\right).$$

Note that, using the inverse function $h^{-1}(x) = a\sinh(x/\sqrt{2})$, results for $h(\delta_i)$ can be backtransformed to results for δ_i.

In the following, we describe the combination procedure using estimators g_i^* and variance estimators $\widehat{\text{Var}}(g_i^*)$, $i = 1, \ldots, k$, and combine the results directly on the scale of θ_i. Alternatively, one can first combine the transformed estimators $h(g_i^*)$ and then backtransform the results using the inverse function h^{-1}. For combining $h(g_i^*)$, $i = 1, \ldots, k$, we have to replace g_i^* by $h(g_i^*)$ and $\widehat{\text{Var}}(g_i^*)$ by $1/n_i$ in the following formulas.

The homogeneity hypothesis

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_k,$$

that is, all standardized mean differences are identical, can be tested using Cochran's (general large sample) homogeneity statistic. Defining here $\hat{v}_i = 1/\widehat{\text{Var}}(g_i^*)$ and $u_i = \hat{v}_i / \sum_{j=1}^k \hat{v}_j$, $i = 1, \ldots, k$, the test statistic can be obtained as

$$Q_{\text{C}} = \sum_{i=1}^{k} \hat{v}_i \Big(g_i^* - \sum_{j=1}^{k} u_j \; g_j^* \Big)^2. \tag{4.16}$$

Under H_0, Q_{C} is approximately χ^2-distributed with $k - 1$ degrees of freedom. If the homogeneity assumption holds, the fixed effects model is quite appropriate; otherwise, the combination of the results should be carried out in a random effects model.

Recall that the random effects model is given here as

$$g_i^* \sim N\left[\theta, \tau^2 + \left(\frac{n_{Ti} + n_{Ci}}{n_{Ti} \; n_{Ci}} + \frac{\theta_i^2}{2 \; (n_{Ti} + n_{Ci} - 2)}\right)\right], \tag{4.17}$$

where θ denotes the overall effect size and τ^2 stands for the between-study variability.

Following the DerSimonian-Laird (1986) approach, an estimator of τ^2 can be obtained as

$$\hat{\tau}^2 = \frac{Q_{\text{C}} - (k-1)}{\sum_{i=1}^k \hat{v}_i - \sum_{i=1}^k \hat{v}_i^2 / \sum_{j=1}^k \hat{v}_j} \tag{4.18}$$

with Q_{C} from Eq. (4.16), where negative estimates are set to zero.

Let $\hat{w}_i = 1/[\hat{\tau}^2 + \widehat{\text{Var}}(g_i^*)]$, $i = 1, \ldots, k$, denote the estimate of the inverse of the variance in model (4.17), then the estimate of the overall effect θ is given by

$$\hat{\theta} = \frac{\sum_{i=1}^k \hat{w}_i \; g_i^*}{\sum_{j=1}^k \hat{w}_j}.$$

The large sample variance of $\hat{\theta}$ is given as

$$\widehat{\mathrm{Var}}_{(1)}(\hat{\theta}) = \left(\sum_{i=1}^{k} \hat{w}_i \right)^{-1}.$$

Following Hartung (1999), another estimator of the variance of $\hat{\theta}$ is given as

$$\widehat{\mathrm{Var}}_{(2)}(\hat{\theta}) = \frac{1}{k-1} \frac{\sum_{i=1}^{k} \hat{w}_i (g_i^* - \hat{\theta})^2}{\sum_{j=1}^{k} \hat{w}_j}.$$

Consequently, a large sample $100(1-\alpha)\%$ confidence interval for θ is given as

$$\mathrm{CI}_1(\theta): \quad \hat{\theta} \mp \sqrt{\widehat{\mathrm{Var}}_{(1)}(\hat{\theta})}\ z_{1-\alpha/2} \tag{4.19}$$

which can be improved with respect to the actual coverage probability for a small number of studies through

$$\mathrm{CI}_2(\theta): \quad \hat{\theta} \mp \sqrt{\widehat{\mathrm{Var}}_{(2)}(\hat{\theta})}\ t_{k-1;1-\alpha/2}. \tag{4.20}$$

Hartung and Knapp (2001a) carried out a simulation study comparing the actual confidence coefficients of the approximate confidence intervals (4.19) and (4.20) when there is no difference between the treatment and the control group, that is, under $H_0 : \theta = 0$. In the fixed effects approach, the interval (4.19) mostly proves to be conservative, while the interval (4.20) attains the nominal confidence coefficient quite well, only for small sample sizes this interval is also conservative. In the random effects approach, the interval (4.19) turns out to be conservative for small values of heterogeneity, but with increasing heterogeneity, this interval can be very liberal. The interval (4.20) almost has actual confidence coefficients close to the nominal one, except in small sample sizes when the interval is a bit conservative.

Since the simulation study by Hartung and Knapp (2001a) is restricted to $\theta = 0$ and to the combination on the scale of the standardized mean difference, some additional simulation results are provided studying the performance of the intervals for different values of θ and considering the combinations of the estimators g_i^* as well as the transformed estimators $h(g_i^*)$.

In Table 4.1, a part of the simulation results are reported, which represents the main findings. The estimated standardized mean differences were generated on the original

scale and, thus, τ^2 stands for the variability of the true standardized mean differences. The sample sizes were chosen identical in all studies, two scenarios considered the balanced case, and two other unbalanced situations. In Table 4.1, CI_1 and CI_2 stand for the intervals (4.19) and (4.20) and CI_3 stands for interval when the standard approach is applied on the transformed estimators $h(g_i^*)$ and CI_4 is derived in analogy to CI_2 applied on $h(g_i^*)$.

For $k = 3$ studies, the standard confidence interval CI_1 is conservative when no heterogeneity is present. But when heterogeneity is present, this interval turns out to be liberal and, with increasing heterogeneity, the actual confidence coefficient decreases up to 80% given a nominal one of 95%. The true underlying standardized mean difference does not essentially affect the results.

When the results are combined on the transformed scale using the standard approach and then the combined results are backtransformed to the original scale, the resulting confidence interval CI_3 proves to be very conservative in most cases. For increasing heterogeneity, the actual confidence coefficient declines, but still for moderate heterogeneity, for instance, $\tau^2 = 1$, the interval is still rather conservative. The true underlying standardized mean difference affects the performance of the interval for large heterogeneity. For $\theta = 0.5$, the interval is liberal with actual confidence intervals between 85% and 90%, except in the small sample case. For $\theta = 5$, the actual confidence coefficient, however, is around the nominal one.

The confidence interval CI_2 attains well the nominal coefficient in all scenarios as mostly does the confidence interval CI_4 as well. But the true underlying standardized mean difference affects the performance of the interval CI_4. Whereas for $\theta = 5$, the interval CI_4 consistently attains the nominal level, the interval turns out to be a bit liberal for $\theta = 0.5$ and large heterogeneity.

Doubling the number of the studies basically yields the same performance of the four intervals. The liberal intervals with $k = 3$ studies have now actual confidence coefficients closer to the nominal one, but again it clearly turns out that the intervals CI_2 and CI_4 outperforms the other two intervals.

Since the actual confidence coefficients of CI_2 and CI_4 are often close together, the average lengths of these intervals are reported in Table 4.2. Obviously, the intervals

for $k = 6$ studies are, on average, always shorter than the intervals for $k = 3$ studies. Moreover, for $k = 3$ studies, the average length of interval CI_2 is less than the average length of interval CI_4, and consequently, the interval CI_2 is preferred to CI_4. However, for $k = 6$ studies, both intervals can be recommended similarly.

4.3 Ratio of Means

The response ratio, that is, the ratio of mean outcome in the experimental group to that in the control group, and closely related measures of proportionate change are often used as measures of effect sizes in ecology, see Hedges, Gurevitch, and Curtis (1999). The parameter of interest is the ratio of the population means, that is, $\rho_i = \mu_{Ti}/\mu_{Ci}$. The sample response ratio $R_i = \bar{Y}_{Ti}/\bar{Y}_{Ci}$ is an estimate of ρ_i in the ith study. Usually, the combination of the response ratios R_i is carried out on the metric of the natural logarithm for two reasons. First, the natural logarithm linearizes the metric, that is, deviations in the numerator are treated the same as deviations in the denominator. Second, the sampling distribution of R_i is skewed and the sampling distribution of $\ln(R_i)$ is much more normal in small sample sizes than that of R_i. For further discussion on this topic, we refer to Hedges, Gurevitch, and Curtis (1999).

Let $\zeta_i = \ln(\mu_{Ti}) - \ln(\mu_{Ci})$ be the natural logarithm of the ratio of population means in the ith study. Then, ζ_i can be estimated by

$$\hat{\zeta}_i = \ln(\bar{Y}_{Ti}) - \ln(\bar{Y}_{Ci})$$

with

$$\mathrm{Var}(\hat{\zeta}_i) \approx \frac{\sigma_{Ti}^2}{n_{Ti}\ \mu_{Ti}^2} + \frac{\sigma_{Ci}^2}{n_{Ci}\ \mu_{Ci}^2},$$

or

$$\mathrm{Var}(\hat{\zeta}_i) \approx \sigma_i^2 \left(\frac{1}{n_{Ti}\ \mu_{Ti}^2} + \frac{1}{n_{Ci}\ \mu_{Ci}^2} \right),$$

where the latter holds for $\sigma_i^2 = \sigma_{Ti}^2 = \sigma_{Ci}^2$.

Table 4.1. Estimated actual confidence coefficients (in %) of four intervals for different values of the standardized mean difference given a nominal level of $100(1-\alpha)\% = 95\%$

			$k = 3$ studies				$k = 6$ studies			
θ	(n_T, n_C)	τ^2	CI_1	CI_2	CI_3	CI_4	CI_1	CI_2	CI_3	CI_4
0.5	(5,5)	0	96.59	95.44	100	95.23	96.88	95.92	100	95.28
		0.1	95.28	95.67	99.99	95.42	95.61	95.87	99.99	95.17
		1	88.69	95.67	99.65	95.09	91.60	96.31	99.71	95.12
		10	82.50	95.77	96.43	93.44	87.53	96.32	98.29	94.14
0.5	(10,10)	0	96.42	95.07	100	94.95	96.38	95.37	100	95.04
		0.1	93.55	95.26	99.95	95.12	94.00	95.45	99.96	95.09
		1	84.80	95.36	97.67	94.83	89.58	95.56	98.50	94.88
		10	80.25	95.47	89.97	93.34	85.36	95.69	95.78	94.20
0.5	(15,10)	0	96.39	95.25	100	95.17	96.33	95.32	100	95.06
		0.1	92.87	95.17	99.96	95.04	93.67	95.42	99.94	95.10
		1	84.11	95.36	97.03	94.90	89.52	95.51	98.13	94.89
		10	79.77	95.41	88.68	93.53	84.97	95.55	95.09	94.21
0.5	(30,20)	0	96.33	94.99	100	94.93	96.33	95.11	100	95.02
		0.1	90.19	95.12	99.74	95.03	91.74	95.06	99.78	94.88
		1	82.16	95.15	92.87	94.82	88.78	95.30	95.78	94.91
		10	78.92	95.15	84.58	93.29	84.02	95.36	92.63	94.29
5	(5,5)	0	96.69	95.29	100	95.05	96.30	95.16	100	94.14
		0.1	96.08	95.20	100	94.88	95.82	95.06	100	94.34
		1	92.74	95.22	99.99	95.04	92.35	94.53	100	94.68
		10	82.64	95.12	97.39	95.04	85.55	94.56	96.33	95.04
5	(10,10)	0	96.49	95.17	100	95.01	96.11	95.01	100	94.60
		0.1	95.51	95.23	100	95.11	95.54	95.09	100	94.75
		1	89.89	94.93	99.98	94.94	91.10	94.72	99.98	94.97
		10	80.92	95.11	95.92	95.01	85.81	94.99	94.70	95.09
5	(15,10)	0	96.40	95.16	100	95.06	96.31	95.08	100	94.72
		0.1	95.46	95.28	100	95.24	95.28	94.96	100	94.79
		1	88.86	94.91	99.97	94.90	90.70	94.87	99.97	95.03
		10	81.02	95.20	95.72	95.00	86.33	95.20	94.81	95.23
5	(30,20)	0	96.35	94.96	100	94.92	96.18	95.15	100	94.98
		0.1	94.17	94.97	100	94.95	94.42	94.93	100	94.90
		1	86.04	95.08	99.83	95.12	89.56	94.82	99.87	94.96
		10	79.67	95.08	94.43	95.00	85.82	95.07	93.98	95.08

Table 4.2. Average lengths of two intervals for different values of the standardized mean difference

			$k = 3$ studies		$k = 6$ studies	
θ	(n_T, n_C)	τ^2	CI_2	CI_4	CI_2	CI_4
0.5	(5,5)	0	3.163	3.489	1.407	1.446
		0.1	3.517	3.952	1.561	1.605
		1	5.790	7.539	2.561	2.623
		10	15.533	43.197	6.721	6.886
0.5	(10,10)	0	2.093	2.181	0.944	0.953
		0.1	2.552	2.704	1.149	1.159
		1	5.042	6.097	2.269	2.272
		10	14.673	36.433	6.509	6.427
0.5	(15,10)	0	1.892	1.954	0.855	0.861
		0.1	2.378	2.494	1.073	1.080
		1	4.926	5.861	2.222	2.219
		10	14.518	34.551	6.479	6.362
0.5	(30,20)	0	1.314	1.334	0.594	0.596
		0.1	1.932	1.992	0.874	0.876
		1	4.656	5.427	2.108	2.094
		10	14.194	32.723	6.398	6.227
5	(5,5)	0	6.846	7.990	2.944	3.232
		0.1	6.984	8.182	3.011	3.303
		1	8.256	10.061	3.567	3.909
		10	16.299	34.554	7.122	8.157
5	(10,10)	0	4.363	4.634	1.935	2.007
		0.1	4.578	4.881	2.039	2.114
		1	6.296	6.985	2.803	2.909
		10	14.913	28.340	6.696	7.364
5	(15,10)	0	3.852	4.035	1.716	1.764
		0.1	4.104	4.316	1.828	1.878
		1	5.917	6.471	2.647	2.725
		10	14.813	27.423	6.641	7.241
5	(30,20)	0	2.622	2.678	1.182	1.197
		0.1	2.970	3.046	1.341	1.357
		1	5.179	5.531	2.331	2.371
		10	14.305	25.770	6.472	6.984

Let us assume in the rest of this section that $\sigma_i^2 = \sigma_{Ti}^2 = \sigma_{Ci}^2$. Then, the variance of $\hat{\zeta}_i$ can be estimated as

$$\widehat{\mathrm{Var}}(\hat{\zeta}_i) = S_i^{*2}\left(\frac{1}{n_{Ti}\ \bar{Y}_{Ti}^2} + \frac{1}{n_{Ci}\ \bar{Y}_{Ci}^2}\right),$$

where S_i^{*2} is the pooled sample variance from Eq. (4.3).

The homogeneity hypothesis that all the ratios of population means are equal, that is,

$$H_0 : \rho_1 = \rho_2 = \cdots = \rho_k \quad \text{or equivalently} \quad H_0^* : \zeta_1 = \zeta_2 = \cdots = \zeta_k$$

can be tested using Cochran's large sample homogeneity statistic. Defining now $\hat{v}_i = 1/\widehat{\mathrm{Var}}(\hat{\zeta}_i)$ and $c_i = \hat{v}_i / \sum_{j=1}^k \hat{v}_j$, $i = 1, \ldots, k$, the test statistic can be obtained as

$$Q_{\mathrm{C}} = \sum_{i=1}^{k} \hat{v}_i \Big(\hat{\zeta}_i - \sum_{j=1}^{k} c_j\ \hat{\zeta}_j\Big)^2. \tag{4.21}$$

Under H_0 and H_0^*, respectively, Q_{C} is approximately χ^2 distributed with $k-1$ degrees of freedom. If the homogeneity assumption holds, the fixed effects meta-analysis model is quite appropriate; otherwise, the combination of the results should be carried out in a random effects model.

Recall that the random effects model is given here as

$$\hat{\zeta}_i \sim N\left[\zeta, \tau^2 + \sigma_i^2\left(\frac{1}{n_{Ti}\ \mu_{Ti}^2} + \frac{1}{n_{Ci}\ \mu_{Ci}^2}\right)\right] \tag{4.22}$$

where ζ denotes the overall effect size on the logarithmic scale and τ^2 stands for the between-study variability.

Following the DerSimonian-Laird (1986) approach, an estimate of τ^2 can be obtained as

$$\hat{\tau}^2 = \frac{Q_{\mathrm{C}} - (k-1)}{\sum_{i=1}^k \hat{v}_i - \sum_{i=1}^k \hat{v}_i^2 / \sum_{j=1}^k \hat{v}_j}$$

with Q_{C} obtained from Eq. (4.21). This estimator may yield negative values, which are set to zero in practice.

Let $\hat{w}_i = 1/[\hat{\tau}^2 + \widehat{\mathrm{Var}}(\hat{\zeta}_i)]$, $i = 1, \ldots, k$, denote the estimate of the inverse of the variance in model (4.22), then the estimate of the overall effect ζ is given by

$$\hat{\zeta} = \frac{\sum_{i=1}^{k} \hat{w}_i \, \hat{\zeta}_i}{\sum_{j=1}^{k} \hat{w}_j}.$$

The large sample variance of $\hat{\zeta}$ is given as

$$\widehat{\mathrm{Var}}_{(1)}(\hat{\zeta}) = \left(\sum_{i=1}^{k} \hat{w}_i \right)^{-1}.$$

For a small number of studies, Hedges, Gurevitch, and Curtis (1999) recommended the use of the following variance estimator

$$\widehat{\mathrm{Var}}_{(2)}(\hat{\zeta}) = \left(\sum_{i=1}^{k} \hat{w}_i \right)^{-1} \left(1 + 4 \sum_{i=1}^{k} \frac{1}{n_{Ti} + n_{Ci} - 2} \left(\frac{\hat{w}_i}{\hat{v}_i} \right)^2 \frac{\hat{w}_i \, [\sum_{j=1}^{k} \hat{w}_j - \hat{w}_i]}{(\sum_{j=1}^{k} \hat{w}_j)^2} \right).$$

Following Hartung (1999), another estimator of the variance of $\hat{\zeta}$ is given as

$$\widehat{\mathrm{Var}}_{(3)}(\hat{\zeta}) = \frac{1}{k-1} \frac{\sum_{i=1}^{k} \hat{w}_i (\hat{\zeta}_i - \hat{\zeta})^2}{\sum_{i=1}^{k} \hat{w}_i}.$$

A large sample $100(1-\alpha)\%$ confidence interval for ζ is given as

$$\hat{\zeta} \mp \sqrt{\widehat{\mathrm{Var}}_{(1)}(\hat{\zeta})} \; z_{1-\alpha/2},$$

which can be improved with respect to the actual coverage probability for a small number of studies through

$$\hat{\zeta} \mp \sqrt{\widehat{\mathrm{Var}}_{(2)}(\hat{\zeta})} \; z_{1-\alpha/2}.$$

Following Hartung and Knapp (2001a), an alternative $100(1-\alpha)\%$ confidence interval for ζ can be obtained as

$$\hat{\zeta} \mp \sqrt{\widehat{\mathrm{Var}}_{(3)}(\hat{\zeta})} \; t_{k-1;1-\alpha/2}.$$

After combining the results on the log scale, the results will naturally be transformed to the original scale using antilogs. Backtransforming the mean of logs introduces a bias into the estimate of the mean response ratio due to the convexity of the log transform. This bias also arises, for example, in the averaging of correlation coefficients by backtransforming the average of several Fisher's z transforms, or in the averaging of odds ratios by backtransforming the average of several log odds ratios. However, since the magnitude of the bias depends upon the variance of the weighted mean, this bias is usually expected to be slight.

Chapter 5

Combining Results of Controlled Studies with Binary Outcome

An important application of meta-analysis, especially in biometry and epidemiology, is the combination of results from comparative or controlled studies with binary outcomes. Often, in clinical trials or observational studies, the outcome can be generally described as success or failure or as positive or negative, which can be easily coded as 1 or 0.

Let p_{Ti} denote the probability of success in the treatment (T) group in the ith study, $i = 1, \ldots, k$, and n_{Ti} the sample size, then the number of successes, say n_{T1i} is a binomial variate with parameters n_{Ti} and p_{Ti}. Let us denote by n_{T0i} the number of failures in the treatment group in the ith study. By analogy, let us denote by p_{Ci}, n_{Ci}, n_{C1i}, and n_{C0i} the corresponding values in the control (C) group of the ith study. Then the number of successes in the control group, n_{C1i}, is a binomial variate with parameters n_{Ci} and p_{Ci}.

The results of each study can be arranged in a (2×2)-table as shown in Table 5.1. Here, n_{1i} stands for the total number of successes in the ith study, n_{0i} is the total number of failures, and n_i is the total sample size of the ith study.

Table 5.1. Observed frequencies on two binary characteristics in study i

	Success	Failure	Total
Treatment	n_{T1i}	n_{T0i}	n_{Ti}
Control	n_{C1i}	n_{C0i}	n_{Ci}
Total	n_{1i}	n_{0i}	n_i

There are several effect sizes which can be used to quantify a difference between treatment and control group. In the next section, we will describe the effect sizes: probability difference, also known as risk difference, relative risk, also known as risk ratio, and odds ratio.

Given estimates of the effect size and corresponding standard errors, several results from Chapter 2 and 3 can be used for combining these estimates. In Section 5.2, we will describe the method, which is known as generic inverse variance method, for combining effect size estimates in the fixed and random effects model of meta-analysis.

The generic inverse variance method is based on large sample theory. In case of sparse binary data, this method can lead to inconsistent results. In Section 5.3, Mantel-Haenszel type estimators and appropriate variance estimators are presented which are consistent in large samples as well as sparse data situations in the fixed effects model of meta-analysis.

The one-way random effects model of meta-analysis can be derived as the marginal model of a normal-normal hierarchical model. The assumption that the effect size estimator is (at least approximately) normally distributed may be not fulfilled for binary outcomes, especially for small sample sizes. Thus, one may seek for a model which makes direct use of the binomially distributed number of successes. In Section 5.4, we will present binomial-normal hierarchical models which can be used in meta-analysis.

5.1 Effect Sizes

Probability difference

The probability difference in the ith study is defined as $\theta_{1i} = p_{Ti} - p_{Ci}$, and can be unbiasedly estimated by the difference of the observed success probabilities, namely

$$\hat{\theta}_{1i} = \frac{n_{T1i}}{n_{Ti}} - \frac{n_{C1i}}{n_{Ci}}. \tag{5.1}$$

The unbiased estimate of the variance of (5.1) is

$$\widehat{\mathrm{Var}}(\hat{\theta}_{1i}) = \frac{n_{T1i}\ n_{T0i}}{n_{Ti}^2\ (n_{Ti}-1)} + \frac{n_{C1i}\ n_{C0i}}{n_{Ci}^2\ (n_{Ci}-1)}. \tag{5.2}$$

Note that the inverse of the estimator $\widehat{\mathrm{Var}}(\hat{\theta}_1)$ does not exist when both numerators on the right hand side of Eq. (5.2) are equal to zero. When this situation occurs, the study cannot be incorporated in the meta-analysis using the generic inverse variance method, see Section 5.2.

Relative risk

The relative risk in the ith study is defined as the ratio of the success probabilities, that is, p_{Ti}/p_{Ci}. However, it is more convenient to carry out the analysis on the log scale because of the better normal approximation of the corresponding estimator in small samples. Setting $\theta_2 = \ln(p_{Ti}/p_{Ci})$, the logarithm of the relative risk, an estimate of θ_2 may be defined as

$$\hat{\theta}_{2i}^{*} = \ln\left(\frac{n_{T1i}\ /\ n_{Ti}}{n_{C1i}\ /\ n_{Ci}}\right). \tag{5.3}$$

However, the estimate (5.3) cannot be computed when $n_{i1i} = 0$ or $n_{C1i} = 0$. Moreover, there does not exist an unbiased estimate of the log relative risk. So, different proposals exist in the literature for estimating this parameter. Pettigrew, Gart, and Thomas (1986) discussed the proposed estimators with respect to bias and variance, and concluded that there is no optimal solution. The "optimal" solution always depends on the true, but unknown, success probabilities. One widely used estimate in this context is

$$\hat{\theta}_{2i} = \ln\left[\frac{(n_{T1i}+0.5)\ /\ (n_{Ti}+0.5)}{(n_{C1i}+0.5)\ /\ (n_{Ci}+0.5)}\right]. \tag{5.4}$$

The variance of estimate (5.4) is estimated without bias except for terms of order $O(n^{-3})$ by

$$\widehat{\text{Var}}(\hat{\theta}_2) = \frac{1}{n_{T1i} + 0.5} - \frac{1}{n_{Ti} + 0.5} + \frac{1}{n_{C1i} + 0.5} - \frac{1}{n_{Ci} + 0.5}.$$

This variance estimate is always positive if $n_{T1i} \neq n_{Ti}$ or $n_{C1i} \neq n_{Ci}$. If $n_{T1i} = n_{Ti}$ or $n_{C1i} = n_{Ci}$, then the value 0.5 will not be added to n_{Ti} and n_{Ci} to ensure the positiveness of the variance estimate.

Odds ratio

The odds ratio in the ith study is defined as the ratio of the odds, that is, $p_{Ti}/(1 - p_{Ti})$ divided by $p_{Ci}/(1 - p_{Ci})$. Again, it is more convenient to carry out the analysis on the log scale because of the better normal approximation of the corresponding estimator in small samples. Setting $\theta_{3i} = \ln\{[p_{Ti}/(1-p_{Ti})]/[p_{Ci}/(1-p_{Ci})]\}$, the logarithm of the odds ratio, an estimate of θ_{3i} is obtained as

$$\hat{\theta}^*_{3i} = \ln\left[\frac{n_{T1i} \ / \ n_{T0i}}{n_{C1i} \ / \ n_{C0i}}\right] = \ln\left[\frac{n_{T1i} \ n_{C0i}}{n_{T0i} \ n_{C1i}}\right]. \tag{5.5}$$

As in the case of the log relative risk, the estimate (5.5) cannot be computed when there are no successes or only successes in at least one group. Again, no unbiased estimate of the log odds ratio exists, and Gart and Zweifel (1967) investigated several estimators of this parameter with respect to bias and variance. One estimate, originally proposed by Haldane (1955), is widely used, namely

$$\hat{\theta}_3 = \ln\left[\frac{(n_{T1i} + 0.5) \ / \ (n_{T0i} + 0.5)}{(n_{C1i} + 0.5) \ / \ (n_{C0i} + 0.5)}\right] = \ln\left[\frac{(n_{T1i} + 0.5)\,(n_{C0i} + 0.5)}{(n_{T0i} + 0.5)\,(n_{C1i} + 0.5)}\right]. \tag{5.6}$$

The variance of estimate (5.6) is unbiasedly estimated except terms of order $O(n^{-3})$ by

$$\widehat{\text{Var}}(\hat{\theta}_3) = \frac{1}{n_{T1i} + 0.5} + \frac{1}{n_{T0i} + 0.5} + \frac{1}{n_{C1i} + 0.5} + \frac{1}{n_{C0i} + 0.5}.$$

5.2 Generic Inverse Variance Method

Let θ_i be the parameter of interest in the ith study, for instance, probability difference, log relative risk, or log odds ratio, and let us assume that each independent study provides an estimate of θ_i, say $\hat{\theta}_i$, $i = 1, \ldots, k$, as well as an estimate of $\mathrm{Var}(\hat{\theta}_i) = \sigma_i^2(\theta_i)$, say $\hat{\sigma}_i^2(\theta_i)$. Note that the variance $\sigma_i^2(\theta_i)$ may functionally depend on the parameter of interest, and consequently $\hat{\theta}_i$ and $\hat{\sigma}_i^2(\theta_i)$ are then correlated. Of course, within a meta-analysis, the type of the parameter of interest is identical in all the studies.

In the random effects model of meta-analysis, we have, at least approximatively,

$$\hat{\theta}_i \sim N\left[\theta \ , \ \tau^2 + \sigma_i^2(\theta_i)\right] \ , \ i = 1, \ldots, k. \tag{5.7}$$

Here θ stands for the overall effect size and τ^2 denotes the parameter for the between-study variance, also called heterogeneity parameter. If $\tau^2 = 0$, we have the fixed effects model of meta-analysis and θ is then the common effect size in all the studies, see Chapter 2.

For testing the homogeneity hypothesis, $H_0 : \tau^2 = 0$, we can use Cochran's large sample homogeneity test, see Chapter 2, which is given here as

$$Q_{\mathrm{C}} = \sum_{i=1}^{k} \hat{v}_i \left(\hat{\theta}_i - \tilde{\theta}\right)^2 \tag{5.8}$$

with $\hat{v}_i = 1/\hat{\sigma}_i^2(\theta_i)$, $i = 1, \ldots, k$, and $\tilde{\theta} = \sum_{i=1}^{k} \hat{v}_i \ \hat{\theta}_i / \sum_{j=1}^{k} \hat{v}_j$. Under H_0, the statistic Q_{C} is approximately chi-square distributed with $k-1$ degrees of freedom and H_0 is rejected at level α if $Q_{\mathrm{C}} > \chi^2_{k-1;1-\alpha}$.

For estimating the heterogeneity parameter τ^2, the DerSimonian-Laird (DSL) estimator or the restricted maximum likelihood (REML) estimator, see Chapter 3, are commonly used in the present setting. The DSL estimator of τ^2 is given here as

$$\hat{\tau}^2_{\mathrm{DSL}} = \frac{Q_{\mathrm{C}} - (k-1)}{\sum_{i=1}^{k} \hat{v}_i - \sum_{i=1}^{k} \hat{v}_i^2 / \sum_{j=1}^{k} \hat{v}_j} \tag{5.9}$$

with Q_{C} from Eq. (5.8).

Let $w_i(\tau^2) = 1/[\tau^2 + \hat{\sigma}_i^2(\theta_i)]$, $i = 1, \ldots, k$, and

$$\hat{\theta}(\tau^2) = \frac{\sum_{i=1}^{k} w_i(\tau^2) \ \hat{\theta}_i}{\sum_{i=1}^{k} w_i(\tau^2)}.$$

Then, the REML estimate of τ^2 can be found numerically by iterating

$$\tau^2 = \frac{\sum_{i=1}^k w_i^2(\tau^2)\left\{[\hat{\theta}_i - \hat{\theta}(\tau^2)]^2 - \hat{\sigma}_i^2(\theta_i)\right\}}{\sum_{j=1}^k w_j^2(\tau^2)} + \frac{1}{\sum_{i=1}^k w_i(\tau^2)}, \tag{5.10}$$

starting with an initial guess of τ^2, say τ_0^2, on the right hand side of Eq. (5.10).

By profiling the restricted log-likelihood for τ^2, we can construct a $100(1-\alpha)\%$ confidence interval for τ^2 as follows. Recall that the restricted log-likelihood function can be written as

$$l_R(\tau^2) \propto -\frac{1}{2}\sum_{i=1}^k \ln[\tau^2 + \hat{\sigma}_i^2(\theta_i)] - \frac{1}{2}\sum_{i=1}^k \frac{1}{\tau^2 + \hat{\sigma}_i^2(\theta_i)} - \frac{1}{2}\sum_{i=1}^k \frac{[\hat{\theta}_i - \hat{\theta}(\tau^2)]^2}{\tau^2 + \hat{\sigma}_i^2(\theta_i)}.$$

Let $\hat{\tau}^2_{\text{REML}}$ denote the REML estimate, see Eq. (5.10). Then, a $100(1-\alpha)\%$ confidence interval for τ^2 is given by

$$\begin{aligned} \text{CI}(\tau^2): \quad & \left\{\tilde{\tau}^2 \geq 0 \mid -2\left[l_R(\tilde{\tau}^2) - l_R(\hat{\tau}^2_{\text{REML}})\right] < \chi^2_{1;1-\alpha}\right\} \\ & = \left\{\tilde{\tau}^2 \mid l_R(\tilde{\tau}^2) > l_R(\hat{\tau}^2_{\text{REML}}) - \chi^2_{1;1-\alpha}/2\right\}. \end{aligned} \tag{5.11}$$

Using the quadratic form

$$\tilde{Q}(\tau^2) = \sum_{i=1}^k \tilde{w}_i \left(\hat{\theta}_i - \hat{\theta}_{\tilde{w}}\right)^2$$

with $\hat{\theta}_{\tilde{w}} = \sum_{i=1}^k \tilde{w}_i\ \hat{\theta}_i / \sum_{i=1}^k \tilde{w}_i$ and $\tilde{w}_i = [\tau^2 + \hat{\sigma}_i^2(\theta_i)]^{-1}$, and following Hartung and Knapp (2005a), Knapp, Biggerstaff, and Hartung (2006), or Viechtbauer (2007), a further approximate $100(1-\alpha)\%$ confidence interval for τ^2 can be obtained as

$$\text{CI}(\tau^2) = \left\{\tau^2 \geq 0 \;\middle|\; \chi^2_{k-1;\alpha/2} \leq \tilde{Q}(\tau^2) \leq \chi^2_{k-1;1-\alpha/2}\right\}. \tag{5.12}$$

To determine the bounds of the confidence interval explicitly one has to solve the two equations for τ^2, namely

$$\begin{aligned} &\text{lower bound:} && \tilde{Q}(\tau^2) = \chi^2_{k-1;1-\alpha/2}, \\ &\text{upper bound:} && \tilde{Q}(\tau^2) = \chi^2_{k-1;\alpha/2}. \end{aligned}$$

Let $\hat{w}_i = 1/[\hat{\tau}^2 + \hat{\sigma}_i^2(\hat{\theta}_i)]$ be the inverse of the estimated variance in model (5.7), with $\hat{\tau}^2$ being a suitable estimate of τ^2. Then the estimate of the overall effect size is given as

$$\hat{\theta} = \frac{\sum_{i=1}^{k} \hat{w}_i \, \hat{\theta}_i}{\sum_{j=1}^{k} \hat{w}_j}.$$

The standard approximate $100(1-\alpha)\%$ confidence interval of θ is then given as

$$\mathrm{CI}_1(\theta) : \hat{\theta} \mp \left(\sum_{i=1}^{k} \hat{w}_i \right)^{-1/2} z_{1-\alpha/2}, \tag{5.13}$$

whereas the modified approximate $100(1-\alpha)\%$ confidence interval according to Hartung and Knapp (2001b) is obtained as

$$\mathrm{CI}_2(\theta) : \quad \hat{\theta} \mp \sqrt{\hat{q}} \; t_{k-1,1-\alpha/2} \quad \text{with} \quad \hat{q} = \frac{1}{k-1} \frac{\sum_{i=1}^{k} \hat{w}_i \left(\hat{\theta}_i - \hat{\theta} \right)^2}{\sum_{j=1}^{k} \hat{w}_j}. \tag{5.14}$$

Hartung and Knapp (2001b) carried out an extensive simulation study for all three effect sizes, probability difference, log relative risk, and log odds ratio, to investigate the actual levels of the confidence intervals (5.13) and (5.14). The performance of the interval (5.13) depends on the chosen parameter of interest as well as the amount of heterogeneity. For the probability difference, this confidence interval can become very liberal, especially for a small or moderate number of studies. The larger the amount of heterogeneity the smaller the actual confidence level given a predefined level. For the log relative risk, this interval turns out to be very often conservative; only when the sample sizes in the studies extremely differ and large heterogeneity is present, the interval becomes anticonservative. For the log odds ratio, the interval (5.13) turns out to be mostly liberal, except for small values of τ^2. The interval (5.14) generally shows a better performance than the interval (5.13) in attaining a predefined confidence level. For the probability difference and the log odds ratio, the interval satisfactorily attains the nominal level, irrespective of the pattern of sample sizes chosen and the value of the heterogeneity parameter. For the relative risk, the interval tends to be a little conservative in most cases, but like the interval (5.13) can become liberal when the sample sizes in the studies extremely differ and large heterogeneity is present. But nevertheless, interval (5.14) is always preferable

to interval (5.13). An important additional result of the simulation study of Hartung and Knapp (2001b) is, that even if the meta-analysis is done using the random effects approach though no heterogeneity is present, the interval (5.14) will satisfactorily keep the nominal level. Thus, using this approach, a choice between fixed effects and random effects approach in advance is not necessary.

Knapp, Biggerstaff, and Hartung (2006) and Viechtbauer (2007) evaluated the performance of the confidence intervals for the heterogeneity parameter with the log odds ratio as the parameter of interest. It turned out that their interval (5.12) outperforms the other intervals with respect to attaining a predefined confidence level. Only the profile likelihood based confidence interval (5.11) for τ^2 is a reasonable alternative in most but not all cases.

5.3 Sparse Data and Mantel-Haenszel Type Estimators

The general inverse variance method described in Section 5.2 can be applied in the fixed effects as well as in the random effects model of meta-analysis. In some applications, for instance combining results from safety studies of medicinal products when the number of (serious) adverse events is of interest, it may occur that a lot of entries in the (2×2)-tables are small or 0. This situation is known as "sparse" data. Since the results of the general inverse variance method rely on large sample results, the overall meta-analysis results can be inconsistent in sparse-data situation, even when the correction factor like 0.5 is used in the formulas of Section 5.1.

Mantel and Haenszel (1959) proposed an estimator of a common odds ratio of several (2×2)-tables for case-control studies in epidemiology, which can also be generally used in the fixed effects approach of meta-analysis. The Mantel-Haenszel estimator of a common odds ratio is given as

$$\widehat{\mathrm{OR}}_{\mathrm{MH}} = \frac{\sum_{i=1}^{k} n_{T1i}\ n_{C0i}/n_i}{\sum_{j=1}^{k} n_{T0j}\ n_{C1j}/n_j} \tag{5.15}$$

and can also be expressed as a weighted average of the study-specific odds ratio estimates,

namely

$$\widehat{\mathrm{OR}}_{\mathrm{MH}} = \sum_{i=1}^{k} \frac{w_i}{\sum_{j=1}^{k} w_j} \frac{n_{T1i}\ n_{C0i}}{n_{T0i}\ n_{C1i}}, \quad w_i = \frac{n_{T0i}\ n_{C1i}}{n_i}, \quad i = 1, \ldots, k. \tag{5.16}$$

Breslow (1981) developed a large-sample theory to study odds ratio estimation in sparse data, and demonstrated the consistency of the Mantel-Haenszel estimator within that theory. In Breslow's theory, the number of tables increases but the cell sizes remain bounded.

Define for $i = 1, \ldots, k$,

$$\begin{aligned} R_i &= n_{T1i}\ n_{C0i}/n_i, \\ S_i &= n_{T0i}\ n_{C1i}/n_i, \\ P_i &= (n_{T1i} + n_{C0i})/n_i, \\ Q_i &= (n_{T0i} + n_{C1i})/n_i. \end{aligned}$$

Then an estimate of the variance of the logarithm of $\widehat{\mathrm{OR}}_{\mathrm{MH}}$, see Robins, Breslow, and Greenland (1986), which is consistent in both large-stratum and sparse-data situation, is given by

$$\widehat{\mathrm{Var}}(\ln \widehat{\mathrm{OR}}_{\mathrm{MH}}) = \frac{\sum_{i=1}^{k} P_i\ R_i}{2\big(\sum_{j=1}^{k} R_j\big)^2} + \frac{\sum_{i=1}^{k}(P_i\ S_i + Q_i\ R_i)}{2\sum_{j=1}^{k} R_j \sum_{\ell=1}^{k} S_\ell} + \frac{\sum_{i=1}^{k} Q_i\ S_i}{2\big(\sum_{j=1}^{k} S_j\big)^2}. \tag{5.17}$$

This variance estimator of $\ln \widehat{\mathrm{OR}}_{\mathrm{MH}}$ is now generally accepted; see Silcocks (2005) for a discussion on various estimators of the variance of $\ln \widehat{\mathrm{OR}}_{\mathrm{MH}}$.

An approximate $100(1-\alpha)\%$ confidence interval for log odds ratio is given as

$$\ln \widehat{\mathrm{OR}}_{\mathrm{MH}} \mp \sqrt{\widehat{\mathrm{Var}}(\ln \widehat{\mathrm{OR}}_{\mathrm{MH}})}\ z_{1-\alpha/2}. \tag{5.18}$$

Greenland and Robins (1985) considered Mantel-Haenszel-type estimators of the relative risk and the probability difference and derived estimates of the variance of these estimators, which are consistent in large-stratum and sparse data situations. Again, these estimators can be used in the fixed effects approach of meta-analysis.

The Mantel-Haenszel-type estimator of the relative risk is given as

$$\widehat{\mathrm{RR}}_{\mathrm{MH}} = \frac{\sum_{i=1}^{k} n_{T1i}\ n_{Ci}/n_i}{\sum_{j=1}^{k} n_{C1j}\ n_{Tj}/n_j}. \tag{5.19}$$

This estimator can also be displayed as a weighted average of the study-specific relative risk estimators, namely,

$$\widehat{\mathrm{RR}}_{\mathrm{MH}} = \sum_{i=1}^{k} \frac{w_i}{\sum_{j=1}^{k} w_j}\ \frac{n_{T1i}\ n_{Ci}}{n_{Ti}\ n_{C1i}}, \quad w_i = \frac{n_{Ti}\ n_{C1i}}{n_i}, \quad i = 1, \ldots, k. \tag{5.20}$$

The consistent variance estimator of the logarithm of $\widehat{\mathrm{RR}}_{\mathrm{MH}}$ is given by (see Greenland and Robins, 1985)

$$\widehat{\mathrm{Var}}(\ln \widehat{\mathrm{RR}}_{\mathrm{MH}}) = \frac{\sum_{i=1}^{k} (n_{Ti}\ n_{Ci}\ n_{1i} - n_{T1i}\ n_{C1i}\ n_i)/n_i^2}{\sum_{j=1}^{k} n_{T1j}\ n_{Cj}/n_j\ \sum_{\ell=1}^{k} n_{C1\ell}\ n_{T\ell}/n_\ell}. \tag{5.21}$$

Consequently, an approximative $100(1-\alpha)\%$ confidence interval on log relative risk has the form

$$\ln \widehat{\mathrm{RR}}_{\mathrm{MH}} \mp \sqrt{\widehat{\mathrm{Var}}(\ln \widehat{\mathrm{RR}}_{\mathrm{MH}})}\ z_{1-\alpha/2}. \tag{5.22}$$

The Mantel-Haenszel-type estimator of the probability difference is given by

$$\widehat{\mathrm{PD}}_{\mathrm{MH}} = \frac{\sum_{i=1}^{k} n_{T1i}\ n_{Ci}/n_i - n_{Ti}\ n_{C1i}/n_i}{\sum_{j=1}^{k} n_{Tj}\ n_{Cj}/n_j}, \tag{5.23}$$

which can be displayed as a weighted average of the study-specific probability difference estimators, namely,

$$\widehat{\mathrm{PD}}_{\mathrm{MH}} = \sum_{i=1}^{k} \frac{w_i}{\sum_{j=1}^{k} w_j} \left(\frac{n_{T1i}}{n_{Ti}} - \frac{n_{C1i}}{n_{Ci}} \right), \quad w_i = \frac{n_{Ti}\ n_{Ci}}{n_i}, \quad i = 1, \ldots, k. \tag{5.24}$$

The consistent estimator of the variance of $\widehat{\mathrm{PD}}_{\mathrm{MH}}$ is given by (see Greenland and Robins, 1985)

$$\widehat{\mathrm{Var}}(\widehat{\mathrm{PD}}_{\mathrm{MH}}) = \frac{\sum_{i=1}^{k} (n_{T1i}\ n_{T0i}\ n_{Ci}^3 + n_{C1i}\ n_{C0i}\ n_{Ti}^3)/(n_{Ti}\ n_{Ci} n_i^2)}{\left(\sum_{j=1}^{k} n_{Tj}\ n_{Cj}/n_j \right)^2}. \tag{5.25}$$

Consequently, an approximate $100(1-\alpha)\%$ confidence interval of the common probability difference is given by

$$\widehat{\mathrm{PD}}_{\mathrm{MH}} \mp \sqrt{\widehat{\mathrm{Var}}(\widehat{\mathrm{PD}}_{\mathrm{MH}})}\ z_{1-\alpha/2}. \tag{5.26}$$

The Mantel-Haenszel-type estimators are derived from the unconditional distribution of the number of successes (n_{T1i}, n_{C1i}) and are first order approximations to the unconditional maximum likelihood estimators. As already mentioned, all these estimators for the different effect sizes can only be used in the fixed effects approach of meta-analysis. Greenland (1982) showed that if important heterogeneity is present, Mantel-Haenszel-type estimators will not estimate meaningful parameters, and thus will also be inappropriate.

Peto's method (Yusuf et al., 1985), sometimes also called Yusuf-Peto method, is a further method of combining odds ratios in the fixed effects approach of meta-analysis. This method was developed for use in mega-trials in cancer and heart disease, where small effects are likely, yet very important. Consequently, this method may be appealing in meta-analysis when combining results from studies with sparse data. In each study, the log odds ratio is estimated by

$$\ln \widehat{\mathrm{OR}}_{\mathrm{Peto,i}} = \frac{O_i - E_i}{V_i}, \quad i = 1, \ldots, k, \tag{5.27}$$

with

$$\begin{aligned} O_i &= n_{T1i}, \\ E_i &= \frac{n_{Ti}\ n_{1i}}{n_i}, \\ V_i &= \frac{n_{Ti}\ n_{Ci}\ n_{1i}\ n_{0i}}{(n_i - 1)\ n_i^2}, \end{aligned}$$

and the estimate is based on the conditional distribution of n_{T1i} given the total number of successes. The estimator of the variance of $\ln \widehat{\mathrm{OR}}_{\mathrm{Peto}}$ is

$$\widehat{\mathrm{Var}}(\ln \widehat{\mathrm{OR}}_{\mathrm{Peto,i}}) = \frac{1}{V_i}, \quad i = 1, \ldots, k. \tag{5.28}$$

Consequently, the overall estimate of the common log odds ratio is

$$\ln \widehat{\mathrm{OR}}_{\mathrm{Peto}} = \frac{\sum_{i=1}^{k}(O_i - E_i)}{\sum_{j=1}^{k} V_j} = \sum_{i=1}^{k} \frac{V_i}{\sum_{j=1}^{k} V_j}\ \ln \widehat{\mathrm{OR}}_{\mathrm{Peto,i}} \tag{5.29}$$

with

$$\widehat{\mathrm{Var}}(\ln \widehat{\mathrm{OR}}_{\mathrm{Peto}}) = \left(\sum_{i=1}^{k} V_i \right)^{-1}. \tag{5.30}$$

An approximate $100(1-\alpha)\%$ confidence interval of the log odds ratio is then given by

$$\ln \widehat{\mathrm{OR}}_{\mathrm{Peto}} \mp \sqrt{\widehat{\mathrm{Var}}(\ln \widehat{\mathrm{OR}}_{\mathrm{Peto}})}\; z_{1-\alpha/2}. \tag{5.31}$$

It should be noted that O_i, E_i, and V_i are all equal to zero for studies with no events in either study arm. These studies therefore do not contribute to either the point estimate or variance of the pooled odds ratio.

Recently, Sweeting, Sutton, and Lambert (2004) and Bradburn et al. (2007) investigated the performance of meta-analytical methods in sparse data situations. In both papers, methods for combining odds ratios with rare events were investigated. Though the Mantel-Haenszel estimator can handle zero cells, often a continuity correction factor like of 0.5 is still added to each cell in the (2×2)-table. The effect of the use of continuity corrections was also investigated in both papers.

The findings for the odds ratio are similar in both papers. In sparse data situation, the inverse variance method using the standard interval (5.13) performed consistently badly, irrespective of the continuity correction used. In both papers, the improved interval (5.14) was not considered. Peto's method turns out to be the best method when event rates are below 1 per cent, and provided that there is no substantial imbalance between treatment and control group sizes within studies, and treatment effects are not exceptionally large. In other circumstances, the use of the Mantel-Haenszel estimator of the common odds ratio is to be preferred.

Though the methods described in this section make use of the binomially distributed number of successes n_{T1i} and n_{C1i}, they can only be applied in the fixed effects approach of meta-analysis. In the next section, we show how one can make direct use of the binomially distributed number of successes in the random effects approach of meta-analysis.

5.4 Binomial-Normal Hierarchical Models

A critical assumption in the fixed effects or random effects model may be the assumption that the estimator of the treatment difference is normally distributed, especially in small sample sizes. When the number of successes in the treatment groups are known, that is,

the observed (2×2)-table is given, one can make direct use of the binomially distributed numbers of successes. In the random effects approach this can be done in a binomial-normal hierarchical model that can be analyzed with exact likelihood methods or within the Bayesian framework using Markov chain Monte Carlo (MCMC) methods. Essentially, we will present here the basic ideas of the model formulations.

Smith, Spiegelhalter, and Thomas (1995) first presented the formulation for the log odds ratio that is straightforward. Then Warn, Thompson, and Spiegelhalter (2002) also considered the binomial-normal hierarchical model for the log relative risk and the risk difference risk. All the three models have one common feature, namely, that the number of successes n_{T1i} and n_{C1i} are both binomially distributed with parameters n_{Ti} and p_{Ti}, and n_{Ci} and p_{Ci}, respectively, in each study i, $i = 1, \ldots, k$.

Let $\mu_i = \text{logit}(p_{Ci}) = \ln[p_{Ci}/(1 - p_{Ci})]$ be the logarithmic odds in the control group and assume that the logarithmic odds in the treatment group is $\mu_i + \theta_i$. Consequently, θ_i is the study-specific treatment difference on the log odds ratio scale. Finally, assume that θ_i comes from a normal distribution with mean θ, the overall effect of treatment difference, and variance τ^2, the heterogeneity parameter.

In summary, we may write the binomial-normal hierarchical model for the log odds ratio as

$$\begin{aligned} n_{C1i} &\sim \text{Bin}\left(n_{Ci}, p_{Ci}\right), \\ n_{T1i} &\sim \text{Bin}\left(n_{Ti}, p_{Ti}\right), \\ \mu_i &= \text{logit}(p_{Ci}), \\ \text{logit}(p_{Ti}) &= \mu_i + \theta_i, \\ \theta_i &\sim N(\theta, \tau^2). \end{aligned} \tag{5.32}$$

Note that each value of θ_i from the normal distribution yields admissible values of the success probabilities p_{Ti} and p_{Ci}.

For the log relative risk, we set $\mu_i = \ln(p_{Ci})$, that is, the logarithm of the success probability in the control group. Then the logarithm of the success probability in the treatment group is parameterized as $\ln(p_{Ti}) = \mu_i + \theta_i$, and θ_i is the log relative risk. Again, assume that θ_i comes from a normal distribution with mean θ, the overall effect of

treatment difference, and variance τ^2, the heterogeneity parameter. But now, the value θ_i needs to be constrained so that $p_{Ti} \in [0, 1]$. Following Warn, Thompson, and Spiegelhalter (2002) this is equivalent to constraining $\ln(p_{Ti})$ to the interval $(-\infty, 0]$, which is achieved by confining θ_i to be less than $-\ln(p_{Ci})$. Let θ_i^U be the minimum of θ_i and $-\ln(p_{Ci})$, then θ_i^U can take any value in the range $(-\infty, -\ln(p_{Ci}))$. The full model can then be summarized as

$$\begin{aligned} n_{C1i} &\sim \text{Bin}\left(n_{Ci}, p_{Ci}\right), \\ n_{T1i} &\sim \text{Bin}\left(n_{Ti}, p_{Ti}\right), \\ \mu_i &= \ln(p_{Ci}), \\ \ln(p_{Ti}) &= \mu_i + \min\left\{\theta_i, -\ln(p_{Ci})\right\}, \\ \theta_i &\sim N(\theta, \tau^2). \end{aligned} \tag{5.33}$$

Finally, we consider the third effect measure probability difference. Let $\mu_i = p_{Ci}$ be the success probability in the control group. Then the success probability in the treatment group is parameterized as $p_{Ti} = \mu_i + \theta_i$, and as before assume that θ_i arises from a normal distribution with mean θ, the overall effect of treatment difference, and variance τ^2, the heterogeneity parameter. As in the previous case, the value θ_i needs to be constrained so that $p_{Ti} \in [0, 1]$, that is, $\theta_i \in [-p_{Ci}, 1 - p_{Ci}]$. Define two new parameters θ_i^U and θ_i^L, corresponding to upper and lower bounds for θ_i. Let θ_i^L be the maximum of θ_i and $-p_{Ci}$, then θ_i^L can take any value in the range $[-p_{Ci}, \infty)$. Similarly, let θ_i^U be the minimum of θ_i^L and $1 - p_{Ci}$, then θ_i is confined to the required range $[-p_{Ci}, 1 - p_{Ci}]$. The full model is then given by

$$\begin{aligned} n_{C1i} &\sim \text{Bin}\left(n_{Ci}, p_{Ci}\right), \\ n_{T1i} &\sim \text{Bin}\left(n_{Ti}, p_{Ti}\right), \\ \mu_i &= p_{Ci}, \\ p_{Ti} &= \mu_i + \min\left\{\max\left\{\theta_i, -p_{Ci}\right\}, 1 - p_{Ci}\right\}, \\ \theta_i &\sim N(\theta, \tau^2). \end{aligned} \tag{5.34}$$

For a full Bayesian analysis in the models (5.32), (5.33), and (5.34), appropriate prior distributions have to be determined for the hyperparameters θ and τ^2 as well as for

the success probabilities p_{Ci} in the control groups, which may also be called baseline risk. For instance, in their example using the probability difference, Warn, Thompson, and Spiegelhalter (2002) used the uniform distribution on $[-1, 1]$ as prior distribution of the risk difference parameter and the uniform distribution on $[0, 2]$ as prior distribution of the square root of the between-study variance, say τ. For the prior distributions of p_{Ci}, they considered a uniform distribution on $[0, 1]$ and a beta prior distribution with hyperparameters α and β, with a uniform distribution on $[1, 100]$ as hyperprior on each. For the log relative risk parameter, Warn, Thompson, and Spiegelhalter (2002) used vague $N(0, 10)$-distribution and priors identical to the priors above for τ and p_{Ci}.

Note that the problem of a zero cell can arise in the Bayesian analysis like in the generic inverse variance method, see Section 5.2. The usual way to circumvent this problem is to add 0.5 to the count in each cell of all (2×2)-tables containing zero cells prior to the analysis.

Using an exact binomial likelihood approach in model (5.32) leads to a logistic regression model with a random intercept, and is therefore analogous to the individual patient data method as used by Turner et al. (2000). Recently, Hamza, van Houwelingen, and Stijnen (2008) showed that the use of the exact binomial likelihood approach is preferred to the standard generic inverse variance approach when they considered the logit of sensitivity and specificity in the meta-analysis of diagnostic tests.

Chapter 6

Meta-Regression

In case of substantial heterogeneity between the studies, possible causes of the heterogeneity should be explored. In the context of meta-analysis, that can be done by either covariates on the study level that could explain the differences between the studies or by covariates on the subject level. However, the latter approach is only possible when individual data are available. Since often only information on the study level is available, explaining and investigating heterogeneity by covariates on the study level has drawn much attention in applied sciences. The term meta-regression used to describe such analysis goes back to papers by Bashore et al. (1989), Jones (1992), Greenland (1994), and Berlin and Antman (1994).

Since the number of studies in a meta-analysis is usually quite small, there is a great danger of overfitting. So, there is only room for a few explanatory variables in a meta-regression, whereas a lot of characteristics of the studies may be identified as potential causes of heterogeneity. Higgins and Thompson (2004) remarked that explorations of heterogeneity are noted to be potentially misleading. Investigations of differences between the studies and their results are observational associations and are subject to biases (such as aggregation bias) and confounding (resulting from correlation between study characteristics). Consequently, there is a clear danger of misleading conclusions if p-values from multiple meta-regression analyses are interpreted naïvely.

This chapter is organized as follows. In Section 6.1 we describe in detail the analysis of the fixed and random effects meta-regression with one covariate. Section 6.2 contains the general analysis of meta-regression with more than one covariate. Note that the methods described in this chapter can be seen as an extension of the generic inverse variance method of fixed and random effects meta-analysis. The models and methods can be applied for all effect size measures considered in Chapters 2–5, that is, normal means, difference of normal means, standardized mean differences, ratio of means, risk difference, relative risk, and odds ratio.

6.1 Model with One Covariate

In the fixed effects meta-regression we write

$$Y_i \sim N\left(\theta_i\,,\,\sigma_i^2\right), \quad i = 1, \ldots, k, \tag{6.1}$$

where Y_i is the statistic in the ith study and σ_i^2 the within-study variability of the ith study. The study-specific mean θ_i is parameterized as

$$\theta_i = \theta + \beta x_i, \quad i = 1, \ldots, k, \tag{6.2}$$

where x_i denotes a quantitative covariate or an indicator variable for a factor with only two levels, that is, $x_i = 0$ or $x_i = 1$. In case of a factor with two levels, θ represents the effect size given $x_i = 0$ and β is the difference of the effect size given $x_i = 1$ compared to $x_i = 0$. For a quantitative covariate, β stands for the change in the effects size given a unit change in the covariate. When the quantitative covariate is centered around its mean, then θ represents the effect size given the mean of the quantitative covariate.

Additionally to the parameterization of the mean of the study-specific effect size, we can allow for a parameter of the still unexplained variation between the studies. That is, we can consider, in analogy to the random effects model of meta-analysis, see Chapter 3, the following normal-normal hierarchical model

$$\begin{aligned} Y_i &\sim N\left(\theta_i\,,\,\sigma_i^2\right), \quad i = 1, \ldots, k, \\ \theta_i &\sim N\left(\theta + \beta x_i\,,\,\tau^2\right), \quad i = 1, \ldots, k. \end{aligned}$$

The random effects meta-regression with one covariate is given as the marginal model of the above normal-normal hierarchical model, that is,

$$Y_i \sim N\left(\theta + \beta x_i\,,\; \tau^2 + \sigma_i^2\right), \quad i = 1, \ldots, k. \tag{6.3}$$

In the following, we will present the analysis in the random effects meta-regression. The corresponding analysis in the fixed effects meta-regression can be performed by setting $\tau^2 = 0$.

Let $w_i = 1/\left(\tau^2 + \sigma_i^2\right)$, $i = 1, \ldots, k$, be the true inverse of the variance of Y_i, $w = \sum_{i=1}^{k} w_i$, and $\lambda_i = w_i/w$, $i = 1, \ldots, k$, the normed weights, then the weighted least-squares estimators of θ and β are given by (see Knapp and Hartung, 2003)

$$\tilde{\beta} = \frac{\sum_{i=1}^{k} \lambda_i\, x_i\, Y_i - \sum_{j=1}^{k} \lambda_j\, x_j\; \sum_{\ell=1}^{k} \lambda_\ell\, Y_\ell}{\sum_{i=1}^{k} \lambda_i\, x_i^2 - \left(\sum_{j=1}^{k} \lambda_j x_j\right)^2} \tag{6.4}$$

and

$$\tilde{\theta} = \sum_{i=1}^{k} \lambda_i\; Y_i - \tilde{\beta}\; \sum_{j=1}^{k} \lambda_j\, x_j. \tag{6.5}$$

The variances and the covariance of the estimators $\tilde{\theta}$ and $\tilde{\beta}$ are

$$\mathrm{Var}(\tilde{\theta}) = \left[\sum_{i=1}^{k} w_i - \left(\sum_{j=1}^{k} w_j\, x_j\right)^2 \Bigg/ \sum_{\ell=1}^{k} w_\ell\, x_\ell^2\right]^{-1}, \tag{6.6}$$

$$\mathrm{Var}(\tilde{\beta}) = \left[\sum_{i=1}^{k} w_i\, x_i^2 - \left(\sum_{j=1}^{k} w_j\, x_j\right)^2 \Bigg/ \sum_{\ell=1}^{k} w_\ell\right]^{-1}, \tag{6.7}$$

and

$$\mathrm{Cov}(\tilde{\theta}\,,\; \tilde{\beta}) = \frac{-\sum_{i=1}^{k} w_i\, x_i}{\sum_{i=1}^{k} w_i\; \sum_{j=1}^{k} w_j\, x_j^2 - \left(\sum_{\ell=1}^{k} w_\ell\, x_\ell\right)^2}. \tag{6.8}$$

Usually, every study provides an estimate of the within-study variance σ_i^2, say $\hat{\sigma}_i^2$. The between-study variance τ^2 can be estimated using the different estimation procedures discussed in Chapter 3 adapted for the meta-regression model with one covariate. We present some extensions of the between-study variance estimators from Chapter 3 in the following.

In analogy to the DerSimonian-Laird estimator in Chapter 3, the method of moments (MM) estimator of the between-study variance τ^2 can be derived from the statistic $Q_1 = \sum_{i=1}^{k} w_i^*(Y_i - \hat{\theta}^* - \hat{\beta}^* x_i)^2$ in the present model, where $\hat{\theta}^*$ and $\hat{\beta}^*$ are weighted least-squares estimators of θ and β with known weights $w_i^* = 1/\sigma_i^2$, $i = 1, \ldots, k$, that is, the weighted least-squares estimators in the fixed effects meta regression. So, the quadratic form Q_1 can also be seen as the residual sum of squares in the fixed effects meta-regression model. The method of moments estimator is given in its truncated form as (see Thompson and Sharp, 1999)

$$\hat{\tau}^2_{\text{MM}} = \max\left\{ 0\,;\ \frac{Q_1 - (k-2)}{F(\boldsymbol{w}^*, \boldsymbol{x})} \right\} \tag{6.9}$$

with

$$F(\boldsymbol{w}^*, \boldsymbol{x}) = \sum_{i=1}^{k} w_i^* - \frac{\sum w_i^{*2} \ \sum w_i^* \, x_i^2 - 2 \ \sum w_i^{*2} x_i \ \sum w_i^* x_i + \sum w_i^* \ \sum w_i^{*2} x_i^2}{\sum w_i^* \ \sum w_i^* \, x_i^2 - \left(\sum w_i^* \, x_i\right)^2} \, .$$

In practice, the usually unknown variances σ_i^2 have to be replaced by appropriate estimates in Eq. (6.9).

Using the ordinary least squares estimators of θ and β, say $\bar{\theta}$ and $\bar{\beta}$, in model (6.3), Raudenbush (1994) derived an approximated method of moments (AMM) estimator of τ^2, which is given as

$$\hat{\tau}^2_{\text{AMM}} = \max\left\{ 0, \frac{1}{k-2} \sum_{i=1}^{k} \left[Y_i - (\bar{\theta} + \bar{\beta} x_i) \right]^2 - \frac{1}{k} \sum_{i=1}^{k} \hat{\sigma}_i^2 \right\}. \tag{6.10}$$

Note that the estimator $\hat{\tau}^2_{\text{AMM}}$ is equal to the ANOVA-type estimator of τ^2 in the case of no covariates, see Chapter 3 (note that $1/(k-2)$ is replaced by $1/(k-1)$ with no covariates).

The (approximate) restricted maximum likelihood (REML) estimator for the between-study variance in model (6.3) with one covariate is the solution of the estimating equation (see Berkey et al., 1995)

$$\hat{\tau}^2 = \frac{\sum_{i=1}^{k} \hat{w}_i^2 \left((k/(k-2))(Y_i - \hat{\theta} - \hat{\beta} x_i)^2 - \hat{\sigma}_i^2 \right)}{\sum_{j=1}^{k} \hat{w}_j^2}. \tag{6.11}$$

This equation is iteratively solved using a starting value of τ^2, say $\tau^2 = \tau_0^2$, on the right hand side of Eq. (6.11). With the weights $\hat{w}_i = 1/(\tau_0^2 + \hat{\sigma}_i^2)$, the initial values of $\hat{\theta}$ and $\hat{\beta}$

are given. Then the right hand side of Eq. (6.11) can be evaluated to yield a new value of $\hat{\tau}^2$. This provides new weights $\hat{w}_i$, and leads to new estimates of θ and β and finally to a new value of $\hat{\tau}^2$. The procedure continues until convergence under the restriction that $\hat{\tau}^2$ is non-negative.

Knapp and Hartung (2003) considered the quadratic form

$$Q_2 = \frac{1}{k-2} \sum_{i=1}^{k} w_i \left(Y_i - \tilde{\theta} - \tilde{\beta}\, x_i\right)^2 , \; k > 2 \, . \tag{6.12}$$

This quadratic form can be seen as a mean sum of the weighted least-squares residuals with known variance components. Knapp and Hartung (2003) showed that, under normality of Y_i, the quadratic form Q_2 from Eq. (6.12) is stochastically independent of the weighted least-squares estimators $\tilde{\theta}$ and $\tilde{\beta}$, and that $(k-2)Q_2$ is χ^2-distributed with $k-2$ degrees of freedom. Let $\tilde{w}_i = (\tau^2 + \hat{\sigma}_i^2)^{-1}$, $i = 1, \ldots, k$, and consider the quadratic form

$$\tilde{Q}_2(\tau^2) = \sum_{i=1}^{k} \tilde{w}_i \left(Y_i - \tilde{\tilde{\theta}} - \tilde{\tilde{\beta}}\, x_i\right)^2, \; k > 2, \tag{6.13}$$

with $\tilde{\tilde{\theta}}$ and $\tilde{\tilde{\beta}}$ the estimates of θ and β using the weights $\tilde{w}_i$, $i = 1, \ldots, k$. The distribution of $\tilde{Q}_2(\tau^2)$ can be approximated by a χ^2-distribution with $k-2$ degrees of freedom. Consequently, in analogy to the Mandel-Paule estimator of τ^2 from Chapter 3, an estimator of τ^2 in the random effects meta-regression model with one covariate is given by the solution for τ^2 of the estimating equation

$$\tilde{Q}_2(\tau^2) = k - 2. \tag{6.14}$$

Moreover, an approximate $(1-\alpha)$-confidence region for τ^2 may be defined as

$$\mathrm{CI}(\tau^2) = \left\{ \tau^2 \geq 0 \;\middle|\; \chi^2_{k-2;\alpha/2} \leq \tilde{Q}(\tau^2) \leq \chi^2_{k-2;1-\alpha/2} \right\} \tag{6.15}$$

with $\chi^2_{k-2;\alpha}$ the α-quantile of the χ^2-distribution with $k-2$ degrees of freedom.

Let us now consider the analysis of the fixed effects in the present model. Let $\hat{w}_i = (\hat{\tau}^2 + \hat{\sigma}_i^2)^{-1}$, $i = 1, \ldots, k$, be the consistent estimators of the weights w_i, and by plugging in these estimators in Eqs. (6.4) and (6.5) we obtain the weighted least-squares estimators,

denoted by $\hat{\theta}$ and $\hat{\beta}$. The commonly used (large sample) $(1-\alpha)$-confidence intervals on the parameters θ and β are given by

$$\hat{\theta} \mp \sqrt{\widehat{\mathrm{Var}}(\hat{\theta})}\; z_{1-\alpha/2} \tag{6.16}$$

and

$$\hat{\beta} \mp \sqrt{\widehat{\mathrm{Var}}(\hat{\beta})}\; z_{1-\kappa/2}\,, \tag{6.17}$$

where $\widehat{\mathrm{Var}}(\hat{\theta})$ and $\widehat{\mathrm{Var}}(\hat{\beta})$ are obtained by putting $\hat{w}_i$, $i = 1, \ldots, k$, in Eqs. (6.6) and (6.7), respectively.

Like in the random effects model of meta-analysis in Chapter 3, the use of the standard normal distribution in Eqs. (6.16) and (6.17) is questionable, especially when the number of studies is small. Based on simulation results, Berkey et al. (1995) recommended the use of a t-distribution with $k-4$ degrees of freedom, where they considered the log relative risk as an outcome measure in their simulation study.

Let us consider again the quadratic form Q_2 from Eq. (6.12). Since $(k-2)Q_2$ is χ^2-distributed with $k-2$ degrees of freedom, the expected value of Q_2 is equal to one for known variance components.

Hence, unbiased and non-negative estimators of the variances of $\tilde{\theta}$ and $\tilde{\beta}$ are given by

$$Q_2(\tilde{\theta}) = \frac{1}{k-2} \sum_{i=1}^{k} g_i\,(Y_i - \tilde{\theta} - \tilde{\beta}\,x_i)^2 \tag{6.18}$$

with $\quad g_i = w_i \,/\, [\sum w_j - (\sum w_j\,x_j)^2 / \sum w_j\,x_j^2],\ i = 1, \ldots, k$, and

$$Q_2(\tilde{\beta}) = \frac{1}{k-2} \sum_{i=1}^{k} h_i\,(Y_i - \tilde{\theta} - \tilde{\beta}\,x_i)^2 \tag{6.19}$$

with $\quad h_i = w_i \,/\, [\sum w_j\,x_j^2 - (\sum w_j\,x_j)^2 / \sum w_j],\ i = 1, \ldots, k$, see Knapp and Hartung (2003).

Replacing the unknown variance components in Eqs. (6.18) and (6.19) by appropriate estimates, Knapp and Hartung (2003) proposed the following approximate $(1-\kappa)$-confidence intervals on θ and β:

$$\hat{\theta} \pm \sqrt{\hat{Q}_2(\hat{\theta})}\; t_{k-2,1-\kappa/2} \tag{6.20}$$

and

$$\hat{\beta} \pm \sqrt{\hat{Q}_2(\hat{\beta})}\, t_{k-2,1-\kappa/2}\ , \tag{6.21}$$

where $t_{\nu;\,\kappa}$ denotes the κ-quantile of the t-distribution with ν degrees of freedom.

Using either the MM estimator or the REML estimator of the between-study variance, the confidence intervals (6.20) and (6.21) are smaller than the corresponding intervals (6.16) and (6.17) when the realized value of the quadratic form Q_2 from Eq. (6.12) is less than one given equal test distributions in both cases. Therefore, Knapp and Hartung (2003) considered an ad-hoc modification of the variance estimates $\hat{Q}_2(\hat{\theta})$ and $\hat{Q}_2(\hat{\beta})$ in the limits of the confidence intervals (6.20) and (6.21) to the effect that they force the realized value of Q_2 to be at least one. That is, the modified confidence intervals are given by

$$\hat{\theta} \mp \sqrt{\hat{Q}_2^*(\hat{\theta})}\, t_{k-2,1-\kappa/2} \tag{6.22}$$

with

$$\hat{Q}_2^*(\hat{\theta}) = \frac{\max\left\{1\,;\ \sum_{i=1}^k \hat{w}_i (Y_i - \hat{\theta} - \hat{\beta} x_i)^2/(k-2)\right\}}{\sum_{i=1}^k \hat{w}_i - (\sum_{j=1}^k \hat{w}_j x_j)^2 / \sum_{\ell=1}^k \hat{w}_\ell x_\ell^2},$$

and

$$\hat{\beta} \mp \sqrt{\hat{Q}_2^*(\hat{\beta})}\, t_{k-2,1-\kappa/2} \tag{6.23}$$

with

$$\hat{Q}_2^*(\hat{\beta}) = \frac{\max\left\{1\,;\ \sum_{i=1}^k \hat{w}_i (Y_i - \hat{\theta} - \hat{\beta} x_i)^2/(k-2)\right\}}{\sum_{i=1}^k \hat{w}_i x_i^2 - (\sum_{j=1}^k \hat{w}_j x_j)^2 / \sum_{\ell=1}^k \hat{w}_\ell}.$$

In a simulation study, Knapp and Hartung (2003) considered the log relative risk as outcome measure in a meta-regression setting. The main result of their simulation study is that the intervals (6.22) and (6.23) outperform the other corresponding intervals with the respect to the nominal confidence coefficient.

Recently, Sidik and Jonkman (2005b) considered robust variance estimation in random effects meta-regression. We will describe their approach in the general random effects meta-regression model in the next section.

6.2 Model with More Than One Covariate

The extension of model (6.3) to the case with more than one covariate is given as

$$Y_i \sim N\left(\theta + \boldsymbol{x_i}'\boldsymbol{\beta}, \tau^2 + \sigma_i^2\right) = N\left(\boldsymbol{z_i}'\boldsymbol{\gamma}, \tau^2 + \sigma_i^2\right), i = 1 \ldots, k, \tag{6.24}$$

where $\boldsymbol{x_i}$ is now a vector of covariates, $\boldsymbol{z_i}' = (1, \boldsymbol{x_i}')$, and $\boldsymbol{\beta}$ a vector of corresponding regression parameters, $\boldsymbol{\gamma}' = (\theta, \boldsymbol{\beta}')$.

In matrix notation, the general random effects meta-regression for meta-analysis with $(r-1)$ covariates can be described as

$$\boldsymbol{Y} \sim N(\boldsymbol{Z\gamma}, \tau^2\boldsymbol{I_k} + \boldsymbol{\Delta}) = N(\boldsymbol{Z\gamma}, \boldsymbol{\Lambda}^{-1}), \quad \boldsymbol{\Lambda}^{-1} = \tau^2\boldsymbol{I_k} + \boldsymbol{\Delta} \tag{6.25}$$

with $\boldsymbol{Y} = (Y_1, \ldots, Y_k)'$, $\boldsymbol{Z}$ the $(k \times r)$-dimensional known regressor matrix with $\text{rank}(\boldsymbol{Z}) = r < k-1$, $\boldsymbol{\gamma} = (\theta, \beta_1, \ldots, \beta_{r-1})'$ the unknown parameter vector of the fixed effects, τ^2 stands for the between-study variance, $\boldsymbol{I_k}$ is the $(k \times k)$-dimensional identity matrix, and $\boldsymbol{\Delta}$ is a $(k \times k)$-dimensional diagonal matrix with entries σ_i^2, $i = 1, \ldots, k$, that is, $\boldsymbol{\Delta}$ contains the within-study variances. Note that the case of a factor with more than two levels can be included in model (6.25) by defining appropriate indicator variables equal to the number of factor levels minus one.

In case all the variance components are known in model (6.25), the weighted least squares estimator of $\boldsymbol{\gamma}$ is given as

$$\tilde{\boldsymbol{\gamma}} = (\boldsymbol{Z}'\boldsymbol{\Lambda}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{\Lambda}\boldsymbol{Y} \tag{6.26}$$

with variance-covariance matrix

$$\boldsymbol{\Sigma} = (\boldsymbol{Z}'\boldsymbol{\Lambda}\boldsymbol{Z})^{-1}. \tag{6.27}$$

Usually, each study provides an estimate of the within-study variability σ_i^2, so that an estimate of $\boldsymbol{\Delta}$, say $\hat{\boldsymbol{\Delta}}$, is given. Consequently, we only have to estimate the between-study variance τ^2 to obtain an estimate of $\boldsymbol{\Lambda}^{-1}$. In the general model (6.25) we consider the method of moments estimator of τ^2 following the lines of the DerSimonian-Laird estimator and the restricted maximum likelihood estimator.

The residual sum of squares in Eq. (6.25) with $\tau^2 = 0$ can be expressed as a quadratic form in $\boldsymbol{Y}$ and has the matrix representation

$$Q = \boldsymbol{Y}'\boldsymbol{P}'\boldsymbol{\Delta}^{-1}\boldsymbol{P}\boldsymbol{Y} \quad \text{with} \quad \boldsymbol{P} = \boldsymbol{I_k} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{\Delta}^{-1}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{\Delta}^{-1}. \tag{6.28}$$

Since $\boldsymbol{P}\boldsymbol{Z} = \boldsymbol{0}$, the expected value of Q is given as

$$\begin{aligned} \mathrm{E}(Q) &= \mathrm{tr}[\boldsymbol{P}'\boldsymbol{\Delta}^{-1}\boldsymbol{P}\,\mathrm{Cov}(\boldsymbol{Y})] \\ &= k - r + \tau^2 f(\boldsymbol{Z}, \boldsymbol{\Delta}^{-1}) \end{aligned}$$

with $f(\boldsymbol{Z}, \boldsymbol{\Delta}^{-1}) = \mathrm{tr}(\boldsymbol{\Delta}^{-1}) - \mathrm{tr}[(\boldsymbol{Z}'\boldsymbol{\Delta}^{-1}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{\Delta}^{-2}\boldsymbol{Z}]$, and $\mathrm{tr}(A)$ denotes the trace of a squared matrix A.

Consequently, the method of moments estimator of τ^2 is given in its truncated form as

$$\hat{\tau}^2_{\mathrm{MM}} = \max\left\{ 0\,,\ \frac{Q - (k-r)}{f(\boldsymbol{Z}, \boldsymbol{\Delta}^{-1})} \right\}. \tag{6.29}$$

The (approximate) restricted maximum likelihood estimator (REML) can be determined by solving iteratively the equation (see Thompson and Sharp, 1999)

$$\hat{\tau}^2 = \frac{\sum_{i=1}^{k} \hat{w}_i^2 \left((k/(k-r))(y_i - \hat{\theta} - \boldsymbol{x_i}'\hat{\boldsymbol{\beta}})^2 - \hat{\sigma}_i^2 \right)}{\sum_{i=1}^{k} \hat{w}_i^2}. \tag{6.30}$$

Let $\hat{\boldsymbol{\Lambda}}^{-1} = \hat{\tau}^2 \boldsymbol{I_k} + \hat{\boldsymbol{\Delta}}$ be the estimated variance-covariance matrix in model (6.31), then the estimate of γ is given by

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}'\hat{\boldsymbol{\Lambda}}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\hat{\boldsymbol{\Lambda}}\boldsymbol{Y} \tag{6.31}$$

with estimated variance-covariance matrix

$$\hat{\boldsymbol{\Sigma}}_1 = \left(\boldsymbol{Z}'\hat{\boldsymbol{\Lambda}}\boldsymbol{Z} \right)^{-1}. \tag{6.32}$$

With the estimated variances on the main diagonal of $\hat{\boldsymbol{\Sigma}}_1$, confidence intervals and hypothesis tests on the fixed effects can be constructed in the usual manner. However, as already mention in Section 6.1, Knapp and Hartung (2003) found out that tests of the meta-regression parameters based on the usual variance estimator generally do not hold a test level at its nominal level.

To carry forward the improved variance estimation approach by Knapp and Hartung (2003) to the case of more than one covariate, let us consider the matrix

$$\boldsymbol{P_1} = \boldsymbol{I_k} - \boldsymbol{Z}(\boldsymbol{Z}'\hat{\boldsymbol{\Lambda}}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\hat{\boldsymbol{\Lambda}} \tag{6.33}$$

and calculate the quadratic form

$$\hat{Q}_r = \frac{\boldsymbol{Y}'\boldsymbol{P_1}'\hat{\boldsymbol{\Lambda}}\boldsymbol{P_1}\boldsymbol{Y}}{k-r}\,. \tag{6.34}$$

The improved variance estimate of a fixed effect estimate is then given by multiplying the corresponding diagonal element in $\hat{\boldsymbol{\Sigma}}_1$ with $\hat{Q}_r$, that is, Knapp and Hartung (2003) suggested as estimator of the variance-covariance matrix of $\hat{\boldsymbol{\gamma}}$

$$\hat{\boldsymbol{\Sigma}}_2 = \hat{Q}_r\ \hat{\boldsymbol{\Sigma}}_1. \tag{6.35}$$

For constructing confidence intervals on the fixed effects the t-distribution with $(k-r)$ degrees of freedom should be used.

Sidik and Jonkman (2005b) considered a robust variance-covariance matrix estimator or so-called sandwich variance-covariance matrix estimator used in a wide range of applications under model misspecification for large samples; see Royall (1986). Extending this approach to the general random effects meta-regression model, Sidik and Jonkman (2005b) proposed the following estimator of the variance-covariance matrix of $\hat{\boldsymbol{\gamma}}$

$$\hat{\boldsymbol{\Sigma}}_3 = \left(\boldsymbol{Z}'\hat{\boldsymbol{\Lambda}}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\hat{\boldsymbol{\Lambda}}\ \{\mathrm{diag}(\hat{\epsilon}_1^{*2},\ldots,\hat{\epsilon}_k^{*2})\}\ \hat{\boldsymbol{\Lambda}}\boldsymbol{Z}\left(\boldsymbol{Z}'\hat{\boldsymbol{\Lambda}}\boldsymbol{Z}\right)^{-1}, \tag{6.36}$$

where $\hat{\epsilon}_i^{*2} = (1-\hat{h}_i)^{-1}\ \hat{\epsilon}_i^2$, with $\hat{\epsilon}_i = Y_i - \boldsymbol{z_i}'\hat{\boldsymbol{\gamma}}$ and $\hat{h}_i = \hat{\lambda}_i\ \boldsymbol{z_i}'\left(\boldsymbol{Z}'\hat{\boldsymbol{\Lambda}}\boldsymbol{Z}\right)^{-1}\boldsymbol{z_i}$. Note that $\hat{\lambda}_i$ is the ith diagonal element of $\hat{\boldsymbol{\Lambda}}$.

In a simulation study, Sidik and Jonkman (2005b) compared their approach with the standard approach and the approach by Knapp and Hartung (2003). They concluded that "despite the seeming suitability of the robust estimator for random effects meta-regression, the improved variance estimator of Knapp and Hartung (2003) yields the best performance among the three estimators, and thus may provide the best protection against errors in the estimated weights."

Bibliography

Asiribo, O., Gurland, J. (1990). Coping with variance heterogeneity. *Communications in Statistics – Theory and Methods*, **19**, 4029–4048.

Bashore, T. R., Osman, A., and Heffley, E. F. (1989). Mental slowing in elderly persons: a cognitive psychophysiological analysis. *Psychology and Aging*, **4**, 235–244.

Berkey, C. S., Hoaglin, D. C., Mosteller, F., Colditz, G. A. (1995). A random-effects regression model for meta-analysis. *Statistics in Medicine*, **14**, 395–411.

Berlin, J.A., Antman, E.M. (1994). Advantages and limitations of meta-analytic regressions of clinical trials data. *Online Journal of Current Clinical Trials*, **134**.

Biggerstaff, B. J., Tweedie, R. L. (1997). Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine*, **16**, 753–768.

Birge, R. T. (1932). The calculation of errors by the method of least squares. *Physical Review*, **40**, 207–227.

Böckenhoff, A., Hartung, J. (1998). Some corrections of the significance level in meta-analysis. *Biometrical Journal*, **40**, 937–947.

Bradburn, M. J., Deeks, J. J., Berlin, J. A., Localio, A. R. (2007). Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Statistics in Medicine*, **26**, 53–77.

Breslow, N. E. (1981). Odds ratio estimators when the data are sparse. *Biometrika*, **68**, 73–84.

Brown, M. B., Forsythe, A. B. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics*, **16**, 129–132.

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

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, **10**, 101–129.

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

Cohen, A., Sackrowitz, H. B. (1984). Testing hypotheses about the common mean of normal distributions. *Journal of Statistical Planning and Inference*, **9**, 207–227.

Cooper, H., Hedges, L. V. (1994). *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

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

Draper, D., Gaver, D. P., Jr. , Goel, P. K., Greenhouse, J. B., Hedges, L. V., Morris, C. N., Tucker, J. R., Waternaux, C. M. (1992). *Combining Information: Statistical Issues and Opportunities for Research*. Washington, D.C.: American Statistical Association, National Academy Press.

Fairweather, W. R. (1972). A method of obtaining an exact confidence interval for the common mean of several normal populations. *Applied Statistics*, **21**, 229–233.

Fisher, R. A. (1932). *Statistical Methods for Research Workers*, 4th edition. London: Oliver and Boyd.

Follmann, D. A., Proschan, M. A. (1999). Valid inference in random effects meta-analysis. *Biometrics*, **55**, 732–737.

Gart, J. J., Zweifel, J. R. (1967). On the bias of various estimators of the logit and its variance with application to quantal bioassay. *Biometrika*, **54**, 181–187.

Glass, G.V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, **5**, 3–8.

Glass, G. V., McGaw, B., Smith, M. L. (1981). *Meta-Analysis in Social Research*. Beverly Hills, CA: Sage.

Graybill, F. A., Deal, R. B. (1959). Combining unbiased estimators. *Biometrics*, **15**, 543–550.

Greenland, S. (1982). Interpretation and estimation of summary ratios under heterogeneity. *Statistics in Medicine*, **1**, 217–227.

Greenland, S. (1994). A critical look at some popular meta-analytical methods. *American Journal of Epidemiology*, **140**, 290–296.

Greenland, S., Robins, J. M. (1985). Estimation of common effect parameter from sparse follow-up data. *Biometrics*, **41**, 55–68.

Haldane, J. B. S. (1955). The estimation and significance of the logarithm of a ratio of frequencies. *Annals of Human Genetics*, **20**, 309–311.

Hamza, T. H., van Houwelingen, H. C., Stijnen, T. (2008). The binomial distribution of meta-analysis was preferred to model within-study variability. *Journal of Clinical Epidemiology*, **61**, 41-51.

Hardy, R. J., Thompson, S. G. (1996). A likelihood approach to meta-analysis with random effects. *Statistics in Medicine*, **15**, 619–629.

Hartung, J. (1981). Nonnegative minimum biased invariant estimation in variance component models. *The Annals of Statistics*, **9**, 278–292.

Hartung, J. (1999). An alternative method for meta-analysis. *Biometrical Journal*, **41**, 901–916.

Hartung, J., Argac, D., Makambi, K. H. (2002). Small sample properties of tests on homogeneity in one-way anova and meta-analysis. *Statistical Papers*, **43**, 197–235.

Hartung, J., Böckenhoff, A., Knapp, G. (2003). Generalized Cochran-Wald statistics in combining of experiments. *Journal of Statistical Planning and Inference*, **113**, 215–237.

Hartung, J., Knapp, G. (2001a). On tests of the overall treatment effect in the meta-analysis with normally distributed responses. *Statistics in Medicine*, **20**, 1771–1782.

Hartung, J., Knapp, G. (2001b) A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine*, **20**, 3875–3889.

Hartung, J., Knapp, G. (2003). Confidence regions on variance components in an extended ANOVA model for combining information. *Acta Applicandae Mathematicae*, **78**, 207–221.

Hartung, J., Knapp, G. (2005a). On confidence intervals for the among-group variance in the one-way random effects model with unequal error variances. *Journal of Statistical Planning and Inference*, **127**, 157–177.

Hartung, J., Knapp, G. (2005b). Models for combining results of different experiments: retrospective and prospective. *American Journal of Mathematical and Management Sciences*, **25**, 149–188.

Hartung, J., Knapp, G., Sinha, B. K. (2008). *Statistical Meta-Analysis with Applications*. New York: Wiley.

Hartung, J., Makambi, K. H. (2002). Positive estimation of the between-study variance in meta-analysis. *South African Statistical Journal*, **36**, 55–76.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, **6**, 107–128.

Hedges, L. V. (1982). Estimating effect size from a series of independent experiments. *Psychological Bulletin*, **92**, 490–499.

Hedges, L. V., Gurevitch, J., Curtis, P. S. (1999). The meta-analysis of response ratios in experimental ecology. *Ecology*, **80**, 1150–1156.

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

Heine, B. (1993). Nonnegative estimation of variance components in an unbalanced one-way random effects model. *Communications in Statistics – Theory and Methods*, **22**, 2351–2371.

Higgins, J. P. T., Thompson, S. G. (2004). Controlling the risk of spurious findings from meta-regression. *Statistics in Medicine*, **23**, 1663–1682.

Iyer, H., Wang, J. M., Mathew, T. (2004). Models and confidence intervals for true values in interlaboratory trials. *Journal of the American Statistical Association*, **99**, 1060–1071.

Jones, D. R. (1992). Meta-analysis of observational epidemiological studies: a review. *Journal of the Royal Society of Medicine*, **85**, 165–168.

Jordan, S. M., Krishnamoorthy, K. (1996). Exact confidence intervals for the common mean of several normal populations. *Biometrics*, **52**, 77–86.

Kacker, R. N. (2004). Combining information from interlaboratory evaluations using a random effects model. *Metrologia*, **41**, 132–136.

Knapp, G., Biggerstaff, B. J., Hartung, J. (2006). Assessing the amount of heterogeneity in random-effects meta-analysis. *Biometrical Journal*, **48**, 271–285.

Knapp, G., Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, **22**, 2693–2710.

Krishnamoorthy, L., Lu, Y. (2003). Inferences on the common mean of several normal populations based on the generalized variable method. *Biometrics*, **59**, 237–247.

Li, Y., Shi, L., Roth, H. D. (1994). The bias of the commonly-used estimate of variance in meta-analysis. *Communications in Statistics – Theory and Methods*, **23**, 1063–1085.

Lin, S.-H., Lee, J. C. (2005). Generalized inferences on the common mean of several normal populations. *Journal of Statistical Planning and Inference*, **134**, 568–582.

Mandel, J., Paule, R. C. (1970). Interlaboratory evaluation of a material with unequal number of replicates. *Analytical Chemistry*, **42**, 1194–1197.

Mantel, N., Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, **22**, 719–748.

Mehrotra, D. V. (1997). Improving the Brown-Forsythe solution to the generalized Behrens-Fisher problem. *Communications in Statistics – Simulation and Computation*, **26**, 1139–1145.

Meier, P. (1953). Variance of a weighted mean. *Biometrics*, **9**, 59–73.

Norwood, T. E., Hinkelmann, K. (1977). Estimating the common mean of several normal populations. *Annals of Statistics*, **5**, 1047–1050.

Pal, N., Lin, J.-J., Chang, C.-H., Kumar, S. (2007). A revisit to the common mean problem: Comparing the maximum likelihood estimator with the Graybill-Deal estimator. *Computational Statistics & Data Analysis*, **51**, 5673–5681.

Patnaik, P. B. (1949). The non-central χ^2- and F-distributions and their applications. *Biometrika*, **36**, 202–232.

Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *British Medical Journal*, **2**, 1243–1246.

Pettigrew, H. M., Gart, J. J., Thomas, D. G. (1986). The bias and higher cumulants of the logarithm of a binomial variate. *Biometrika*, **73**, 425–435.

Rao, C. R. (1972). Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association*, **67**, 112–115.

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

Raudenbush, S. W. (1994). Random effect models. In: Cooper, H., Hedges, L.V. (Eds.): *The Handbook of Research Synthesis*. New York: Russell Sage Foundation, pp. 301–321.

Robins, J., Breslow, N., Greenland, S. (1986). Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics*, **42**, 311–323.

Royall, R. M. (1986). Model robust confidence intervals using maximum likelihood estimators. *International Statistical Review*, **54**, 221–226.

Rukhin, A. L., Biggerstaff, B. J., Vangel, M. G. (2000). Restricted maximum likelihood estimation of a common mean and the Mandel-Paule algorithm. *Journal of Statistical Planning and Inference*, **83**, 319–330.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, **2**, 110–114.

Sidik, K., Jonkman, J. N. (2005a). Simple heterogeneity variance estimation for meta-analysis. *Applied Statistics*, **52**, 367–384.

Sidik, K., Jonkman, J. N. (2005b). A note on variance estimation in random effects meta-regression. *Journal of Biopharmaceutical Statistics*, **15**, 823–838.

Silcocks, P. (2005). An easy appraoch to the Robins-Breslow-Greenland variance estimator. *Epidemiologic Perspectives & Innovations*, **2**, 9.

Sinha, B. K. (1985). Unbiased estimation of the variance of the Graybill-Deal estimator of the common mean of several normal populations. *Canadian Journal of Statistics*, **13**, 243–247.

Smith, T. C., Spiegelhalter, D. J., Thomas, A. (1995). Bayesian approaches to random-effects meta-analysis; A comparative study. *Statistics in Medicine*, **14**, 2685–2699.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., Williams, R. M., Jr. (1949). *The American Soldier, Volume I. Adjustment during Army Life*. Princeton, N.J.: Princeton University Press.

Suguira, N., Gupta, A. K. (1987). Maximum likelihood estimates for Behrens-Fisher problem. *Journal of the Japan Statistical Society*, **17**, 55–60.

Sweeting, M. J., Sutton, A. J., Lambert, P. C. (2004). What to add to nothing? Use of avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine*, **23**, 1351–1375.

Thompson, S. G., Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: A comparison of methods. *Statistics in Medicine*, **18**, 2693–2708.

Tippett, L. H. C. (1931). *The Methods of Statistics*. London: Williams and Norgate.

Tsui, K., Weerahandi, S. (1989). Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. *Journal of the American Statistical Association*, **84**, 602–607.

Turner, R. M., Omar, R. Z., Yang, M., Goldstein, H., Thompson, S. G. (2000). A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine*, **19**, 3417–3432.

Viechtbauer, W. (2007). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine*, **26**, 37–52.

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

Weerahandi, S. (1993). Generalized confidence intervals. *Journal of the American Statistical Association*, **88**, 899–905.

Welch, B. L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika*, **38**, 330–336.

Whitehead, A. (2002). *Meta-Analysis of Controlled Clinical Trials.* Chichester: Wiley.

Yates, F., Cochran, W. G. (1938). The analysis of groups of experiments. *Journal of Agricultural Science*, **28**, 556–580.

Yu, Ph. L. H., Sun, Y., Sinha, B. K. (1999). On exact confidence intervals for the common mean of several normal populations. *Journal of Statistical Planning and Inference*, **81**, 263–277.

Yusuf, S., Peto, R., Lewis, J., Collins, R., Sleight, P. (1985). Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Progress in Cardiovascular Disease*, **27**, 335–371.