WORKSHOP ON META ANALYSIS

University of Tampere
University of Turku
October 4-6, 2006

Instructors:

Professor Bimal Sinha, Ph.D.
Presidential Research Professor
University of Maryland Baltimore
USA

Professor Guido Knapp, Ph.D.
University of Dortmund
Germany

Ms. Laura Koskela
University of Tampere
Finland

OVERVIEW OF WORKSHOP TOPICS

- Introduction
- Effect Size
- Combination of Tests
- Combination of Estimates
- Common Mean
- Tests of Homogeneity
- One-Way Random Effects Model
- Publication Bias
- Vote Counting Procedures
- Analysis of Binary Data
- Computational Aspects
- Data Sets

Topics not covered:

** Multivariate Meta Analysis

** Bayesian Meta Analysis

** Regression Meta Analysis

Contents

1 Introduction

Meta-analysis, a term coined by Glass (1976), is intended to provide *the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings.*

Meta-analysis, or *Research Synthesis*, or *Research Integration* is precisely a scientific method to accomplish this goal by applying sound statistical procedures, and indeed it has a long and old history. The very invention of least squares by Legendre (1805) and Gauss (1809) is an attempt to solve just a unique problem of meta analysis: use of astronomical observations collected at several observatories to estimate the orbit of comets and to determine meridian arcs in geodesy (Stigler, 1986). In order to determine the relationship between mortality and inoculation with a vaccine for enteric fever, Pearson (1904) used data from five small independent samples, and computed a pooled estimate of correlation between mortality and inoculation in order to evaluate the efficacy of the vaccine. As an early application of meta analysis 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 across experiments in the agricultural sciences in order to derive estimates of *treatment* effects and test their significance. Likewise, there are plenty of applications of meta analysis in the fields of education, medicine and social sciences, some of which are briefly described below.

In the field of education, meta analysis is useful in combining studies about coaching effectiveness to improve SAT scores in verbal and math (Rubin, 1981; DerSimonian and Laird, 1983), in studying the effect of open education on (i) attitude of students toward school, (ii) student independence and self-reliance, and in combining studies about the relationship between teacher indirectness and student achievement (Hedges and Olkin, 1985). In social science, there is a need to combine several studies of gender differences in separate categories of quantitative ability, verbal ability, and visual-spatial ability (Hedges and Olkin, 1985). For some novel applications of meta analysis in the field of medicine, we refer to Pauler and Wakefield (2000) for three applications involving *Dentrifice* data, *Anti-hypertension* data and *Pre-eclampsia* data, to Berry (2000) for questions about benefits and risks of mammography of women based on six studies, to Brophy and Joseph (2000) for meta analysis involving three studies to compare streptokinase and tissue-plasminogen activator to reduce mortality following an acute myocardial infarction, and lastly to Dominici and Parmigiani (2000) for an application of meta analysis involving studies in which outcomes are reported on continuous variables for some medical outcomes in some studies and on binary variables on similar medical outcomes in some other studies. Of course, there are numerous other diverse applications of meta analysis in many other fields.

As the scope of meta analysis grew over the years, several terminologies also came into existence such as *quantitative research synthesis*, *pooling of evidence* or *creating an overview*. While most of the above early works, including Mosteller and Bush (1954), provided a logical foundation for meta analysis, the appearance of several books, notably Glass et al. (1981), Hunter et al. (1982), Rosenthal (1984), Hedges and Olkin (1985), and the edited volume by Cooper and Hedges (1994) and literally thousands of meta-analytic papers during the last twenty years or so, primarily covering applications in health sciences and education, has made the subject to have a very special role in diverse fields of applications.

The essential character of Meta-Analysis is that it is the statistical analysis of the summary findings of many empirical studies, which are called *primary* analyses, all targeted towards a *common* goal. Meta-Analysis is essentially quantitative in nature, using various statistical methods in a practical way, to extract and analyse relevant information from large masses of data. A common criticism of meta analysis is that it is illogical because it combines results from studies which are not the same (mixing apples with oranges). Nevertheless, it should be clear that the only studies which need to be integrated or synthesized in meta analysis are those which are *different* but share a common goal.

A fundamental assumption behind conducting a meta analysis or pooling of evidence or information or data across studies in order to obtain an average *effect* across all studies is that the size of the effect (basic parameter of interest) reported in each study is an estimate of the common effect size of the whole population of studies. It is therefore essential to test for homogeneity of population effect sizes across studies before conducting meta analysis if obtaining an estimate of average effect or its test is the primary goal of meta analysis.

The notion of *effect size* is central to many meta analysis studies which often deal with comparing two treatments, control and experimental, in an effort to find out if there is a significant difference between the two. In the case of continuous measurements, a standardized mean difference plays an important role to measure such a difference. In the case of qualitative attributes, difference or ratio of two proportions, odds ratio and ϕ coefficient are used to capture such differences. Again, when the objective is to study relationship between two variables, an obvious choice is the usual correlation coefficient.

Recent meta-analytic work however concentrates on discovering and explaining variations in effect sizes rather than assuming that they remain the same across studies, which is perhaps rarely the case owing to uncontrollable differences in study contexts, designs, treatments, and subjects. Scientific literatures are cluttered with repeated studies of the same phenomena because some investigators may be unaware of what others are doing, or may be skeptical about the results of past studies, or because they wish to extend previous findings. In any event, when results of several scientific studies of the same phenomena exist and differ, it is indeed interesting to ponder how science should proceed.

It is of course clear how it should *not* proceed, namely, by pretending that there does *not* exist a problem, or by discarding most of the studies which violently disagree, and keeping only a handful of those which closely agree. If studies which are expected to show similar results do show similarities by conducting an appropriate test of homogeneity and accepting the hypothesis of homogeneity, the case for summarizing results of all studies with a single average effect size can be strengthened and defended. If, however this hypothesis is rejected, no single number can adequately account for the variety of reported results. Thus, if the results from various studies differ either significantly or even marginally, the true scientific instinct should be to investigate methods to account for the variability by further systematic work. This is precisely the spirit of some recent research in meta analysis using random and mixed effects models, allowing inclusion of trial-specific covariates which may explain a part of the observed heterogeneity. In other words, a set of conflicting findings from different studies is looked upon as an opportunity for learning and discovering the sources of variation among the reported outcomes rather than a cause for dismay.

While most common meta analysis applications involve comparison of just one variable (experimental) with another (control), multivariate data can also arise in meta analysis due to several reasons. First, the primary studies themselves can be multivariate in nature because these studies may measure multiple outcomes for each subject, and are typically known as *multiple-endpoint studies*. It should however be noted that not all studies in a review would have the same set of outcomes. For example, studies of Scholastic Aptitude Test (SAT) do not all report math and verbal scores. In fact, only about half of the studies dealt with in Becker (1990) provided coaching results for both math and verbal! Secondly, multivariate data may arise when primary studies involve several comparisons among groups based on a single outcome. As an example, Ryan et al. (1986) studied the effects of practice on motor skill levels on the basis of a five-group design, four different kinds of practice groups and one no-practice group, thus leading to comparisons of multivariate data. This kind of studies are usually known as *multiple-treatment studies*.

As mentioned earlier, although most statistical methods of meta analysis focus on deriving and studying properties of a common estimated effect which is supposed to exist across all studies, when heterogeneity across studies is believed to exist, a meta analyst must estimate the extent and sources of heterogeneity among studies if the hypothesis of homogeneity is not found to be tenable. While fixed effects models discussed in this book under the assumption of homogeneous effects sizes continue to be the most common method of meta analysis, the assumption of homogeneity given variability among studies due to varying research and evaluation protocols may be unrealistic. In such cases, a random effects model which avoids the homogeneity assumption, and models effects as random and coming from a distribution is recommended. The various study effects are believed to arise from a population and random effects models *borrow strength* across

studies in providing estimates of both study-specific effects and underlying population effect.

Whether a fixed effects model or a random effects model, a Bayesian approach considers all parameters (population effect sizes for fixed effects models, in particular) as random and coming from a super population with its own parameters. There are several advantages for a Bayesian approach to meta analysis. The Bayesian paradigm provides in a very natural way a method for data synthesis from all studies by incorporating model and parameter uncertainty. Moreover, a predictive distribution for future observations coming from any study, which may be a quantity of central interest to some decision makers, can be easily developed based on what have been already observed. The use of Bayesian hierarchical models often leads to more appropriate estimates of parameters compared to the asymptotic ones arising from maximum likelihood especially in case of small sample sizes of component studies which is typical in meta analysis.

There are at least two other vital issues with meta analysis procedures. Although it is true that most of the primary studies to be included in a meta analysis provide a complete background of the problem being considered along with relevant entire or summary data, it also happens sometimes that some studies report only the ultimate finding in terms of the sign of the estimated underlying effect size being positive or negative or in terms of the significance or non-significance of the test for the absence of an effect size. It then poses a challenge for the statisticians to develop suitable statistical procedures to take into account this kind of incomplete information to carry out meta analysis. Fortunately, there are techniques under the category of *vote counting procedures* to effectively deal with such situations.

The problem of selection or publication bias is rather crucial in the context of meta analysis since the reported studies on which meta analysis is typically based tend to be mostly significant and there could be many potential nonsignificant studies which are not reported at all simply because of their non-significant findings and hence these studies are not amenable to meta analysis considerations. Such a situation is bound to happen in almost any meta analysis scenario in spite of one's best attempt to get hold of all relevant studies, and statistically valid corrective measures should be developed and followed to deal with such a serious publication bias issue. Again, fortunately, there are some valid procedures to tackle this vital problem.

At this point, let us emphasize that there are four important stages of research synthesis.

(i) **problem formulation** stage

(ii) **data collection** stage

(iii) **data evaluation** stage

(*iv*) **data analysis and interpretation** stage.

We describe below these four stages.

The **formulation** of the research synthesis problem has important implications for the statistical methods to be used, and usually there are two broad considerations: the universe to which generalizations are made (fixed effects model and random effects model), and the nature of the *effect size* parameters to be inferred upon (Hedges, 1994). Research synthesis extends our knowledge through the combination and comparison of primary studies, and an important issue is how the results of the synthesis are to be interpreted. One perspective is the fixed effects model where the universe to which generalizations are made consists of ensembles of studies identical to those in the study sample except for the particular primary units appearing in the studies. The other perspective is the random effects model where the universe to which generalizations are made consists of a population of studies from which the study sample is drawn. Another fundamental issue in problem formulation concerns the *nature* of the effect size parameter to be estimated or tested. The inference about effect size is usually sought to answer the question: *what is the relationship between two variables* X *and* Y *?* The variables X and Y are chosen with only the constraint that their relationship is of interest to us. The answer to this question essentially comes in two parts: (*a*) the estimate of the *magnitude* of the relationship (effect size) along with an indication of the accuracy or the reliability of the estimated effect size (standard error or confidence interval), and (*b*) a test of significance of the difference between the realized effect size and the effect size expected under the null hypothesis of no relation between X and Y. Some common effect size measures are given by *standardized* difference of two *means*, *standardized* difference of two *proportions*, difference of two *correlations*, *ratio* of proportions, *odds ratio*, *risk ratio*, and so on.

Data **collection** or literature **search** stage in research synthesis is very different from primary research, and can be very challenging. There are usually five major modes of searching for sources of primary research, namely, manual and computer search of subject indexes from abstract databases, footnote chasing (references in review/nonreview papers and books), consultation (formal/informal requests, conferences), browsing through library shelves, and manual and computer citation searches (White, 1994). It is hoped that all these search procedures would lead to an exhaustive collection of relevant literature for the problem under study, and encompass books/book chapters, research/technical reports, conference papers, and other possible sources. Sometimes we may have to use special ways and means to retrieve what are known as *fugitive* literature and information appearing in unpublished papers/technical reports, unpublished dissertations/master's theses, and the like. Above all, we may need to deal with the important issue of *publication bias* while doing the research synthesis, bearing in mind the fact that often research leading to *nonsignificant* conclusions are not reported at all or rarely so (the well known

file-drawer problem).

Data **evaluation** stage consists of carefully checking the nature and sources of primary research data, missing observations in primary data, and sources of potential bias in the primary data, all in an attempt to assign suitable weights to the various primary data sources at the time of carrying out meta analysis or data synthesis.

Finally, data **analysis** stage deals with statistically describing and combining various primary studies, and is essentially a wide collection of statistical methods depending on the nature of the underlying problem. Thus, there are ways to *combine* various measures of effect sizes either for estimation or test or confidence interval, and also ways to deal with missing values in primary studies as well as publication bias.

Given the above broad spectrum of topics that can be covered under the umbrella of a workshop on meta analysis, our goal in this workshop is primarily concerned with some statistical aspects of meta analysis. As already mentioned, the heart of the enterprise of carrying out meta analysis or synthesizing research consists of comparing and combining the results of individual primary studies of a particular, focused research question, and the emphasis is essentially on two types of statistical analysis: combining results of tests of significance of *effect size*, and combining estimates of *effect size*. The *effect size*, as explained earlier, is a generic term referring to the magnitude of an *effect* or more generally the *size* of the relation between two variables. Moreover, in case of diverse research findings from comparable studies, an attempt must be made to understand and point out reasons for such differences.

Keeping the above general points in mind, the outline of the workshop is as follows.

Lecture 2 describes various standard measures of **effect size** based on *means*, *proportions*, ϕ *coefficient*, *odds ratio*, and *correlations*. Some illustrative examples to explain the related computations and concepts are included.

Lecture 3 deals with methods of **combining individual tests** based on primary research with plenty of applications. This lecture is exclusively based on combination of P-values mainly because the studies which are meant for meta analysis more often report their P values than other details of the study. The methods described here are exact and it should be mentioned that there are other methods based on suitably combining often independent component test statistics whose sampling distributions may not be readily available.

Lecture 4 describes methods of **combining individual estimates** of effect sizes based on primary research to efficiently estimate the common effect size parameter as well as to construct its confidence interval. The methods suggested in this lecture are mainly asymptotic in nature.

Lecture 5 is devoted to a detailed analysis of a special kind of meta analysis problem, namely, inference about the **common mean** of several univariate normal populations with unknown and unequal variances. This problem has a long and rich history, and very significant in applications.

Lecture 6 describes various tests of the important hypothesis of the **homogeneity of population effect sizes** in some particular models. In the context of statistical meta analysis, one should carry out these tests of homogeneity of effect sizes before applying tools of combining the effect sizes.

One way random effects models, useful when the basic hypothesis of homogeneity of effect sizes does not hold, will be taken up in **Lecture 7**. There is a huge literature on this topic and we will make an attempt to present all the important results in this connection. Typically, there are two scenarios: error variances are all equal (homogeneous case) and error variances are not equal (heterogeneous case). We will deal mostly with the latter more challenging case.

Lecture 8 will be devoted to discussing two important aspects of statistical meta analysis: **publication bias** and **vote counting procedures**. These problems arise when we may not have access to *all* the literature on the subject under study and also when we do so there is not enough evidence in the studies.

A different kind of meta analysis dealing with **combination of polls** will be discussed in **Lecture 9**. This particular topic has applications in market research.

Lecture 10 is designed to address the methods of meta analysis in case of **binary data** involving both binary and ordinal outcomes. Some applications of this useful set up will be presented.

There are many computational aspects of statistical meta analysis. This will be taken up in **Lecture 11**.

There remain many advanced topics such as meta regression, multivariate meta analysis, bayesian meta analysis, recovery of interblock information which will not be discussed in this workshop.

Finally, sample data sets which are analyzed throughout the book are included in the final section. The *Bibliography* at the end contains a long list of papers referred to in this workshop.

We conclude this introductory lecture with the observation that virtually all of the statistical methods described here are based on standard large sample results for the (asymptotic) distributions of sample means, sample proportions, sample correlations, and so on, and hence due caution should be exercised when using these methods. Such results

are listed below for ready reference (see Rao, 1973; Rohatgi, 1976).

1. $X_1, \cdots, X_n$ are *iid* with mean μ and variance σ^2. Then, for large n, $\bar{X} \sim N[\mu, \frac{\sigma^2}{n}]$, i.e., $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim N[0,1]$. This is a standard version of the celebrated Central Limit Theorem (CLT).

2. $X_1, \cdots, X_n$ are *iid* with mean μ and variance σ^2. Then, for large n, $\frac{\sqrt{n}(\bar{X}-\mu)}{s} \sim N[0,1]$ where $s^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$. This is an application of CLT coupled with Cramer's theorem (Slutsky's theorem).

3. $X \sim B[n,P]$. Then, for large n, $\frac{X-nP}{\sqrt{nPQ}} \sim N[0,1]$ where $Q = 1-P$, i.e., $\frac{\sqrt{n}(p-P)}{\sqrt{PQ}} \sim N[0,1]$ where $p = X/n$. This is a standard application of the *Central Limit Theorem* (CLT).

4. $X \sim B[n,P]$. Then, for large n, $sin^{-1}\sqrt{p} \sim N[sin^{-1}\sqrt{P}, \frac{1}{4n}]$. This is a well known version of Fisher's variance-stabilizing transformation applied to the binomial proportion.

5. $(X_1,Y_1), \cdots, (X_n,Y_n)$ are *iid* from a bivariate distribution with means (μ_1,μ_2), variances (σ_1^2,σ_2^2), and correlation ρ. Then, for large n, $r \sim N[\rho, \frac{(1-\rho^2)^2}{n-1}]$ where r is the usual sample correlation defined as

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{[\sum_{i=1}^{n}(X_i-\bar{X})^2.\sum_{i=1}^{n}(Y_i-\bar{Y})^2]^{1/2}}. \tag{1.1}$$

   This is also an application of CLT coupled with Cramer's theorem (Slutsky's theorem; see Rao, 1973).

6. $(X_1,Y_1), \cdots, (X_n,Y_n)$ are *iid* from a bivariate distribution with means (μ_1,μ_2), variances (σ_1^2,σ_2^2), and correlation ρ. Then, for large n, $z \sim N[\zeta, \frac{1}{n-3}]$ where $z = \frac{1}{2}log_e[\frac{1+r}{1-r}]$ and $\zeta = \frac{1}{2}log_e[\frac{1+\rho}{1-\rho}]$. This is a well known version of Fisher's variance-stabilizing transformation applied to the sample correlation coefficient.

2 Measures of Effect Size

Quite often the main objective in a study is to compare two *treatments*: **experimental** and **control**. When these treatments are applied to a set of experimental units, the outcomes can be of two types: **qualitative** and **quantitative**, leading to either *proportions* or *means*. Accordingly, effect sizes are also essentially of these two types: those based on differences of two means, and those based on differences of two proportions. A third type of effect size, namely, *correlation*, arises when the objective in a study is to ascertain the nature and extent of relationship between two variables.

2.1 Effect Size based on Means

An effect size based on means is defined as follows. Denote the population means of the two groups (experimental and control) by μ_1 and μ_2, and their variances by σ_1^2 and σ_2^2, respectively. Then the effect size θ based on means is a standardized difference between μ_1 and μ_2, and can be expressed as

$$\theta = \frac{\mu_1 - \mu_2}{\sigma} \tag{2.1}$$

where σ denotes either the standard deviation σ_2 of the population control group, or an average population standard deviation (namely, an average of σ_1 and σ_2).

The above measure of effect size θ can be easily estimated based on sample values, and this is explained below. Suppose we have a random sample of size n_1 from the first population with the sample mean $\bar{X}_1$ and sample variance S_1^2, and also a random sample of size n_2 from the second population with the sample mean $\bar{X}_2$ and sample variance S_2^2. One measure of the effect size θ, known as Cohen's d (Cohen, 1969, 1977, 1988) is then given by

$$d = \frac{\bar{X}_1 - \bar{X}_2}{S} \tag{2.2}$$

where the standardized quantity S is the pooled sample standard deviation defined as $S = \sqrt{S^2}$ where

$$S^2 = \frac{(n_1 - 1)\, S_1^2 + (n_2 - 1)\, S_2^2}{n_1 + n_2},$$
$$(n_1 - 1)\, S_1^2 = \sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2, \;\; (n_2 - 1)\, S_2^2 = \sum_{i=1}^{n_2} (X_{2i} - \bar{X}_2)^2. \tag{2.3}$$

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

$$g = \frac{\bar{X}_1 - \bar{X}_2}{S^*} \tag{2.4}$$

where the standardized quantity S^* is also the pooled sample standard deviation defined as $S^* = \sqrt{S^{*2}}$ where

$$S^{*2} = \frac{(n_1 - 1)\, S_1^2 + (n_2 - 1)\, S_2^2}{n_1 + n_2 - 2}. \tag{2.5}$$

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

$$\begin{aligned} E(g) &\sim \theta + \frac{3\,\theta}{4N - 9} \\ \sigma^2(g) = var(g) &\sim \frac{1}{\tilde{n}} + \frac{\theta^2}{2(N - 3.94)} \end{aligned} \tag{2.6}$$

where

$$N = n_1 + n_2, \;\; \tilde{n} = \frac{n_1\, n_2}{n_1 + n_2}. \tag{2.7}$$

In case the population variances are identical in both groups, under the assumption of normality of the data, Hedges (1981) shows that $\sqrt{\tilde{n}}\, g$ follows a non–central t–distribution with non–centrality parameter $\sqrt{\tilde{n}}\, \theta$ and $(n_1 + n_2 - 2)$ degrees of freedom. Consequently, the exact mean and variance of Hedges' g are given by

$$\begin{aligned} E(g) &= \sqrt{\frac{N-2}{2}}\, \frac{\Gamma\left(\frac{N-3}{2}\right)}{\Gamma\left(\frac{N-2}{2}\right)}\, \theta \\ \sigma^2(g) = var(g) &= \frac{N-2}{N-4}\,(1 + \theta^2) - \theta^2\, \frac{N-2}{2}\, \frac{\left(\Gamma\left(\frac{N-3}{2}\right)\right)^2}{\left(\Gamma\left(\frac{N-2}{2}\right)\right)^2} \end{aligned} \tag{2.8}$$

and $\Gamma(\cdot)$ denotes the gamma function. As Cohen's d is proportional to Hedges' g, the results in (2.8) can be easily transferred providing mean and variance of Cohen's d.
The exact mean in (2.8) is well–approximated by (2.6) so that an approximately *unbiased* standardized mean difference g^* is given as

$$g^* = \left(1 - \frac{3}{4N - 9}\right)\, g. \tag{2.9}$$

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

$$\Delta = \frac{\bar{X}_1 - \bar{X}_2}{S_2} \tag{2.10}$$

where the standardized quantity is just S_2, 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. Again under the assumption of normality of the data, Hedges (1981) shows that $\sqrt{\tilde{n}}\,\Delta$ follows a non–central t–distribution with non–centrality parameter $\sqrt{\tilde{n}}\,\theta$ and $(n_2 - 1)$ degrees of freedom.

The variances of the above estimates of θ, in large samples, are given by the following.

$$\begin{aligned}
\sigma^2(d) = var(d) &= \left[\frac{n_1 + n_2}{n_1\, n_2} + \frac{\theta^2}{2\,(n_1 + n_2 - 2)}\right] \cdot \left[\frac{n_1 + n_2}{n_1 + n_2 - 2}\right] \\
\sigma^2(g) = var(g) &= \frac{n_1 + n_2}{n_1\, n_2} + \frac{\theta^2}{2\,(n_1 + n_2 - 2)} \\
\sigma^2(\Delta) = var(\Delta) &= \frac{n_1 + n_2}{n_1\, n_2} + \frac{\theta^2}{2\,(n_2 - 1)}
\end{aligned} \tag{2.11}$$

The estimated variances are then obtained by replacing θ in the above expressions by the respective estimates of θ, namely, d, g, and Δ. These are given below.

$$\begin{aligned}
\hat{\sigma}^2(d) = \widehat{var}(d) &= \left[\frac{n_1 + n_2}{n_1\, n_2} + \frac{d^2}{2\,(n_1 + n_2 - 2)}\right] \cdot \left[\frac{n_1 + n_2}{n_1 + n_2 - 2}\right] \\
\hat{\sigma}^2(g) = \widehat{var}(g) &= \frac{n_1 + n_2}{n_1\, n_2} + \frac{g^2}{2\,(n_1 + n_2 - 2)} \\
\hat{\sigma}^2(\Delta) = \widehat{var}(\Delta) &= \frac{n_1 + n_2}{n_1\, n_2} + \frac{\Delta^2}{2\,(n_2 - 1)}
\end{aligned} \tag{2.12}$$

Large sample tests for $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ are typically based on the standardized *normal* statistics

$$Z = \frac{\hat{\theta}}{\hat{\sigma}(\hat{\theta})} \tag{2.13}$$

where $\hat{\theta}$ is an estimate of θ defined above with $\hat{\sigma}(\hat{\theta})$ as its estimated standard error, and H_0 is rejected if $|Z|$ exceeds $z_{\alpha/2}$, the upper $\alpha/2$ cut-off point of the standard normal distribution. Of course, if the alternative is one-sided, namely, $H_2 : \theta > 0$, then H_0 is rejected if Z exceeds z_α, the upper α cut-off point of the standard normal distribution.

Again, if one is interested in constructing confidence intervals for θ, it is evident that, in large samples, the individual confidence intervals are given by

$$1 - \alpha = P\left[\hat{\theta} - z_{\alpha/2}\,\hat{\sigma}(\hat{\theta}) \leq \theta \leq \hat{\theta} + z_{\alpha/2}\,\hat{\sigma}(\hat{\theta})\right]. \tag{2.14}$$

Example 2.1. Consider the data set given below.

Table 2.1 Studies of Gender Difference in Quantitative Ability

Study	Total sample size (N)	Standardized mean difference (g)	Unbiased standardized mean difference (g^*)	95% CI on θ
1	76	0.72	0.71	[0.256 , 1.184]
2	6.167	0.06	0.06	[0.010 , 0.110]
3	355	0.59	0.59	[0.377 , 0.803]
4	1,050	0.43	0.43	[0.308 , 0.552]
5	136	0.27	0.27	[-0.068 , 0.608]
6	2,925	0.89	0.89	[0.814 , 0.966]
7	45,222	0.35	0.35	[0.331 , 0.369]

For each study above, we can carry out the test for $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ as well as construct a confidence interval for θ based on the above discussion. Thus, for study 1, using the standardized mean difference g (Hedges' g) $= 0.72$, and assuming $n_1 = n_2 = 38$, we get

$$Z = \frac{g}{\left[\frac{n_1+n_2}{n_1\,n_2} + \frac{g^2}{2\,(n_1+n_2-2)}\right]^{1/2}} = \frac{0.72}{0.2369} = 3.039 \tag{2.15}$$

and hence reject H_0 with $\alpha = 0.05$. Moreover, based on (2.14), the 95% confidence interval for θ is obtained as $[0.256, 1.184]$. It may be noted that the conclusions based on $g^* = 0.71$ are the same. All the 95% confidence intervals for the seven studies are summarized in the last column of Table 2.1.

When the analysis is to be carry out on the original metric, the difference of μ_1 and μ_2, sometimes called absolute difference between means, is the appropriate measure. The difference between means may be easier to interpret than the dimensionless standardized mean difference. The difference of the sample means, $\bar{X}_1 - \bar{X}_2$, is an unbiased of the parameter of interest in this situation with variance $\sigma_1^2/n_1 + \sigma_2^2/n_2$. By plugging in the sample variances, the estimated variance of $\bar{X}_1 - \bar{X}_2$ is $S_1^2/n_1 + S_2^2/n_2$.

2.2 Effect Sizes based on Proportions

An effect size θ based on proportions is derived as follows. Denote the population proportions of the two groups (experimental and control) by π_1 and π_2. One measure θ_1 of the effect size θ is then given by

$$\theta_1 = \pi_1 - \pi_2 \tag{2.16}$$

which is simply the difference between the two population proportions.
A second measure θ_2 of θ, based on Fisher's variance-stabilizing transformation (of a sample proportion) is defined as

$$\theta_2 = \sin^{-1}\sqrt{\pi_1} - \sin^{-1}\sqrt{\pi_2}. \tag{2.17}$$

A third measure θ_3 of θ, commonly known as the *rate ratio*, also called *relative risk* or *risk ratio*, is given by

$$\theta_3 = \frac{\pi_1}{\pi_2}. \tag{2.18}$$

The measures θ_1 and θ_2 are such that the value 0 indicates no difference, while for the measure θ_3, the value 1 indicates no difference. Often $\theta_3^* = \ln\theta_3$, which is the natural logarithm of θ_3, is used so that the same value 0 indicates no difference in all the three cases. The above measures of θ can be easily estimated. Suppose a random sample of size n_1 from the first population yields a count of X_1 for the attribute under study while a random sample of size n_2 from the second population yields a count of X_2. Then, if $p_1 = X_1/n_1$ and $p_2 = X_2/n_2$ denote the two sample proportions, estimates of θ are obtained as

$$\begin{aligned} \hat{\theta}_1 &= p_1 - p_2 \\ \hat{\theta}_2 &= \sin^{-1}\sqrt{p_1} - \sin^{-1}\sqrt{p_2} \\ \hat{\theta}_3^* &= \ln\frac{p_1}{p_2} \end{aligned} \tag{2.19}$$

with the respective variances as

$$\begin{aligned} \sigma^2(\hat{\theta}_1) = var(\hat{\theta}_1) &= \frac{\pi_1\,(1-\pi_1)}{n_1} + \frac{\pi_2\,(1-\pi_2)}{n_2} \\ \sigma^2(\hat{\theta}_2) = var(\hat{\theta}_2) &= \frac{1}{4\,n_1} + \frac{1}{4\,n_2} \\ \sigma^2(\hat{\theta}_3^*) = var(\hat{\theta}_3^*) &= \frac{1-\pi_1}{n_1\,\pi_1} + \frac{1-\pi_2}{n_2\,\pi_2}. \end{aligned} \tag{2.20}$$

As before, large sample tests for $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ are typically based on the standardized *normal* statistics

$$Z = \frac{\hat{\theta}}{\hat{\sigma}(\hat{\theta})} \tag{2.21}$$

where $\hat{\sigma}(\hat{\theta})$ is the estimated standard error of $\hat{\theta}$, and H_0 is rejected if $|Z|$ exceeds $z_{\alpha/2}$, the upper $\alpha/2$ cut-off point of the standard normal distribution. Of course, if the alternative is one-sided, namely, $H_2 : \theta > 0$, then H_0 is rejected if Z exceeds z_α, the upper α cut-off point of the standard normal distribution. Again, if one is interested in constructing confidence intervals for θ, it is evident that, in large samples, the individual confidence intervals are given by

$$1 - \alpha = P\left[\hat{\theta} - z_{\alpha/2}\,\hat{\sigma}(\hat{\theta}) \leq \theta \leq \hat{\theta} + z_{\alpha/2}\,\hat{\sigma}(\hat{\theta})\right]. \tag{2.22}$$

Example 2.2. Consider a comparative study in which the experimental treatment is applied to a random sample of $n_1 = 80$ subjects and the control treatment is applied to a random sample of $n_2 = 70$ subjects. If the unimproved proportions are $p_1 = 0.60$ and $p_2 = 0.80$, the value of $\hat{\theta}_1$ is -0.20 and its estimated standard error is

$$\hat{\sigma}(\hat{\theta}_1) = \left(\frac{0.60 \times 0.40}{80} + \frac{0.80 \times 0.20}{70}\right)^{1/2} = 0.0727. \tag{2.23}$$

An approximate 95% confidence interval for θ_1 is $\hat{\theta}_1 \pm 1.96 \cdot \hat{\sigma}(\hat{\theta}_1)$, which turns out to be the interval $-0.20 \pm 1.96 \times 0.0727$, or the interval from -0.34 to -0.06. Incidentally, since this interval does *not* contain 0, we reject the null hypothesis $H_0 : \theta_1 = 0$.

For the same data, an estimate of θ_2 is given by $\hat{\theta}_2 = 0.8861 - 1.1071 = -0.2211$ with $var(\hat{\theta}_2) = 0.006696$, resulting in $Z = -2.702$. We therefore reject $H_0 : \theta_2 = 0$. A 95% confidence interval for θ_2 is easily obtained as $[\hat{\theta}_2 - 1.96\,\sigma(\hat{\theta}_2) = -0.501\,,\ \hat{\theta}_2 + 1.96\,\sigma(\hat{\theta}_2) = -0.074]$.

Finally, again for the same data, the estimated rate ratio is $\hat{\theta}_3 = 0.60/0.80 = 0.75$, so group 1 is estimated to be at a risk that is 25 percent less than group 2's risk. To construct a confidence interval for θ_3, one first obtains the value $\hat{\theta}_3^* = \ln\,\hat{\theta}_3 = -0.2877$ and then obtains the value of its estimated standard error (see (2.19))

$$\hat{\sigma}(\hat{\theta}_3^*) = \left(\frac{0.40}{80 \times 0.60} + \frac{0.20}{70 \times 0.80}\right)^{1/2} = (0.0119)^{1/2} = 0.1091. \tag{2.24}$$

An approximate 95% confidence interval for θ_3^* has as its lower limit

$$\ln(\theta_{3L}^*) = -0.2877 - 1.96 \times 0.1091 = -0.5015 \tag{2.25}$$

and as its upper limit

$$\ln(\theta^*_{3U}) = -0.2877 + 1.96 \times 0.1091 = -0.0739. \tag{2.26}$$

The resulting interval for θ_3 then extends from $\exp(-0.5015) = 0.61$ to $\exp(-0.0739) = 0.93$.

2.3 Effect Size based on φ Coefficient and Odds Ratio

This section is patterned after Fleiss (1994). Consider a cross-sectional study in which measurements are made on a pair of binary random variables, X and Y, and their association is of primary interest. Examples include studies of attitudes or opinions (agree/disagree), case-control studies in epidemiology (exposed/not exposed), and intervention studies (improved/not improved).

Table 2.2 presents notation for the underlying parameters and Table 2.3 presents notation for the observed frequencies in the 2×2 table cross-classifying subjects' categories on the two variables X and Y, the levels of both of which are labelled as 0 or 1.

Table 2.2. Probabilities associated with two binary characteristics

	Y		
X	Positive	Negative	Total
Positive	Π_{11}	Π_{12}	$\Pi_{1.}$
Negative	Π_{21}	Π_{22}	$\Pi_{2.}$
Total	$\Pi_{.1}$	$\Pi_{.2}$	1

Table 2.3. Observed frequencies on two binary characteristics

	Y		
X	Positive	Negative	Total
Positive	n_{11}	n_{12}	$n_{1.}$
Negative	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{..}$

Then one measure of association between X and Y can be described as the product moment correlation coefficient between the two numerically coded variables, and is equal to

$$\varphi = \frac{\Pi_{11}\Pi_{22} - \Pi_{12}\Pi_{21}}{\sqrt{\Pi_{1.}\Pi_{2.}\Pi_{.1}\Pi_{.2}}}. \tag{2.27}$$

Based on the data shown in Table 2.3, the maximum likelihood estimator of φ is equal to

$$\hat{\varphi} = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1.}n_{2.}n_{.1}n_{.2}}}, \tag{2.28}$$

which is closely related to the classical chi-square statistic for testing for association in a fourfold table: $\chi^2 = n_{..}\hat{\varphi}^2$. The large sample estimated standard error of $\hat{\varphi}$ is given by (Bishop, Fienberg, and Holland (1975, pp. 381-382))

$$\hat{\sigma}(\hat{\phi}) = \tag{2.29}$$
$$\frac{1}{\sqrt{n_{..}}}\left(1 - \hat{\varphi}^2 + \hat{\varphi}(1 + \frac{\hat{\varphi}^2}{2})\frac{(p_{1.}-p_{2.})(p_{.1}-p_{.2})}{\sqrt{p_{1.}p_{.1}p_{2.}p_{.2}}} - \frac{3}{4}\hat{\varphi}^2\left[\frac{(p_{1.}-p_{2.})^2}{p_{1.}p_{2.}} + \frac{(p_{.1}-p_{.2})^2}{p_{.1}p_{.2}}\right]\right)^{1/2}.$$

A second measure of the association between X and Y is provided by the *odds ratio* (sometimes referred to as the cross-product ratio) defined as

$$\omega = \frac{\Pi_{11}\Pi_{22}}{\Pi_{12}\Pi_{21}}. \tag{2.30}$$

If the observed multinomial frequencies are as displayed in Table 2.3, the maximum likelihood estimator of ω is

$$\hat{\omega} = \frac{n_{11}n_{22}}{n_{12}n_{21}}. \tag{2.31}$$

The motivation for using ω as a measure of association between two binary variables stems from the following observation. Suppose that the study calls for $n_{1.}$ units to be sampled from the population which are positive on X, and for $n_{2.}$ units to be sampled from the population which are negative on X. Then $\Pi_{11}/\Pi_{1.}$ represents the conditional probability that Y is positive given that X is positive, namely, $P(Y+ \mid X+)$, and hence the *odds* for Y being positive, conditional on X being positive, are equal to

$$odds(Y+|X+) = P(Y+|X+)/P(Y-|X+) = (\Pi_{11}/\Pi_{1.})/(\Pi_{12}/\Pi_{1.}) = \Pi_{11}/\Pi_{12}. \tag{2.32}$$

Analogously, the *odds* for Y being positive, conditional on X being negative, are equal to

$$odds(Y+|X-) = P(Y+|X-)/P(Y-|X-) = (\Pi_{21}/\Pi_{2.})/(\Pi_{22}/\Pi_{2.}) = \Pi_{21}/\Pi_{22}. \tag{2.33}$$

The *odds ratio* ω is simply defined as the ratio of these two odds values, leading to

$$\omega = \frac{odds(Y+|X+)}{odds(Y+|X-)} = \frac{\Pi_{11}/\Pi_{12}}{\Pi_{21}/\Pi_{22}} = \frac{\Pi_{11}\Pi_{22}}{\Pi_{12}\Pi_{21}}. \tag{2.34}$$

A value of 1 for ω represents no association between X and Y while values more than 1 (less than 1) mean positive (negative) association. In practice it is customary to use $\omega^* = ln\omega$, the natural logarithm of the odds ratio, and its sample analogue $\hat{\omega}^* = ln\hat{\omega}$, rather than the odds ratio directly. The large sample standard error of $\hat{\omega}^*$ (Woolf, 1955) is given by the equation

$$\hat{\sigma}(\hat{\omega}^*) = \left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right)^{1/2} \tag{2.35}$$

which can be readily used to test hypotheses for ω and also to construct a confidence interval for ω.

Example 2.3. Consider a hypothetical study with the data as shown in the Table 2.4.

Table 2.4. Hypothetical Frequencies in a Fourfold Table

	Y		
X	Positive	Negative	Total
Positive	135	15	150
Negative	40	10	50
Total	175	25	200

The value of $\hat{\varphi}$ for the above frequencies is easily computed as

$$\hat{\varphi} = \frac{135 \times 10 - 15 \times 40}{\sqrt{150 \times 50 \times 175 \times 25}} = 0.130931, \tag{2.36}$$

which represents a modest association. Its estimated standard error, based on the formula (2.29), is obtained as

$$\hat{\sigma}(\hat{\varphi}) = \frac{1}{\sqrt{200}} \left(1.245388\right)^{1/2} = 0.079. \tag{2.37}$$

Similarly, we compute $\hat{\omega} = 2.25$ and hence $\hat{\omega}^* = \ln \hat{\omega} = 0.811$ with $\hat{\sigma}(\omega^*) = 0.4462$.

We can test the null hypothesis of no association, i.e., $H_0 : \varphi = 0$ versus $H_1 : \varphi \neq 0$ based on

$$Z = \frac{\hat{\varphi}}{\hat{\sigma}(\hat{\varphi})} = 1.66 \tag{2.38}$$

which leads to acceptance of H_0 with $\alpha = 0.05$. Also, a 95% confidence interval for φ is obtained as

$$LB = \hat{\varphi} - 1.96\,\hat{\sigma}(\hat{\varphi}) = -0.024, \;\; UB = \hat{\varphi} + 1.96\,\hat{\sigma}(\hat{\varphi}) = 0.286. \tag{2.39}$$

Likewise, we can also test the null hypothesis of no association, i.e., $H_0 : \omega^* = 0$ versus $H_1 : \omega^* \neq 0$ based on

$$Z = \frac{\hat{\omega}^*}{\hat{\sigma}(\hat{\omega}^*)} = 1.82 \tag{2.40}$$

which leads to acceptance of H_0 with $\alpha = 0.05$. Also, a 95% confidence interval for ω^* is obtained as

$$LB = \hat{\omega}^* - 1.96\,\hat{\sigma}(\hat{\omega}^*) = -0.063,\;\; UB = \hat{\omega}^* + 1.96\,\hat{\sigma}(\hat{\omega}^*) = 1.685 \tag{2.41}$$

which yields [0.939 , 5.395] as the confidence interval for ω.

2.4 Effect Size based on Correlations

Finally, an effect size based on correlation is directly taken as the value of the correlation ρ itself, or its well known ζ-value, based on Fisher's variance-stabilizing transformation (of r), given by

$$\zeta = \frac{1}{2}\left[\ln \frac{1+\rho}{1-\rho}\right]. \tag{2.42}$$

These measures are readily estimated by the sample correlation r (for ρ), or its transformed version z (for ζ) given by

$$z = \frac{1}{2}\left[\ln \frac{1+r}{1-r}\right] \tag{2.43}$$

with respective approximate variances as (see Rao, 1973)

$$\begin{aligned} var(r) &\sim (1-\rho^2)^2/(n-1) \\ var(z) &\sim 1/(n-3). \end{aligned} \tag{2.44}$$

Large sample tests for $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$ are typically based on the standardized *normal* statistics

$$\begin{aligned} Z_1 &= \frac{r\sqrt{n-1}}{(1-r^2)} \\ Z_2 &= z\sqrt{n-3} \end{aligned} \tag{2.45}$$

and H_0 is rejected if $|Z_1|$ (or $|Z_2|$) exceeds $z_{\alpha/2}$, the upper $\alpha/2$ cut-off point of the standard normal distribution. Of course, if the alternative is one-sided, namely, $H_2 : \rho > 0$, then

H_0 is rejected if Z_1 or Z_2 exceeds z_α, the upper α cut-off point of the standard normal distribution. Again, if one is interested in constructing confidence intervals for ρ, it is evident that, in large samples, the individual confidence intervals based on r for ρ and z for ζ are given by

$$\begin{aligned} 1-\alpha &= P\left[r - \frac{z_{\alpha/2}\,(1-r^2)}{\sqrt{n-1}} \le \rho \le r + \frac{z_{\alpha/2}\,(1-r^2)}{\sqrt{n-1}}\right] \\ 1-\alpha &= P\left[z - \frac{z_{\alpha/2}}{\sqrt{n-3}} \le \zeta \le z + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right] \end{aligned} \tag{2.46}$$

Clearly, the second equation above can be used to provide a confidence interval for ρ using the relation between ρ and ζ.

Example 2.4. Let us consider the results of the seven studies reported below in Table 2.5.

Table 2.5. Studies of the Relationship between an Observation Measure of Teacher Indirectness and Student Achievement

Study	No. of teachers	Correlation coefficient r	95% CI on ρ	95 % CI on ζ	95 % CI on ρ (re-transformed)
1	15	-0.073	[-0.594 , 0.448]	[-0.639 , 0.493]	[-0.564 , 0.456]
2	16	0.308	[-0.150 , 0.766]	[-0.225 , 0.862]	[-0.222 , 0.697]
3	15	0.481	[0.078 , 0.884]	[-0.042 , 1.090]	[-0.041 , 0.797]
4	16	0.428	[0.015 , 0.841]	[-0.086 , 1.001]	[-0.086 , 0.762]
5	15	0.180	[-0.327 , 0.687]	[-0.384 , 0.748]	[-0.366 , 0.634]
6	17	0.290	[-0.159 , 0.739]	[-0.225 , 0.822]	[-0.222 , 0.676]
7	15	0.400	[-0.040 , 0.840]	[-0.142 , 0.989]	[-0.141 , 0.757]

For the first study, we can test $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$ based on both Z_1 and Z_2. A direct computation gives

$$Z_1 = -0.275, \;\; z = -0.073, \;\; Z_2 = -0.253. \tag{2.47}$$

Taking $\alpha = 0.05$, which means $z_{\alpha/2} = 1.96$, we accept H_0. To construct a confidence interval for ρ with confidence level 0.95, we can use (2.46). The first equation gives $[-0.594\,,\, 0.448]$ as the confidence interval for ρ. On the other hand, the second equation yields $[-0.639\,,\, 0.493]$ as the confidence interval for ζ. Using (2.42), we convert this to the interval for ρ as $[-0.564\,,\, 0.456]$. A similar analysis can be carried out for all the other studies. The results are given in Table 2.5 above.

3 Combination of Tests

3.1 Introduction

Methodology for combining findings from repeated research studies did in fact begin with the idea of combining independent tests back in the 1930's (Tippett, 1931; Fisher, 1932; Pearson, 1933). Here we provide a comprehensive review of the so-called *omnibus* or *nonparametric* statistical methods for testing the significance of combined results.

All the methods of combining tests depend on what is known as a P-value. A key point is that the observed P values derived from continuous test statistics follow a *uniform* distribution under the null hypothesis *regardless* of the form of the test statistic, the underlying testing problem, and the nature of the parent population from which samples are drawn.

Quite generally, suppose $X_1, \cdots, X_n$ is a random sample from a certain population indexed by the parameter θ, and $T(X_1, \cdots, X_n)$ is a test statistic for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$, where θ_0 is a null value, and suppose also that H_0 is rejected for large values of $T(x_1, \cdots, x_n)$. Then if the (continuous) null distribution of $T(X_1, \cdots, X_n)$ is denoted by $g(t)$, the (one-sided) P value based on $T(X_1, \cdots, X_n)$ is defined as

$$P = \int_{T(x_1,\cdots,x_n)}^{\infty} g(t)dt = P[T(X_1, \cdots, X_n) > T(x_1, \cdots, x_n)|H_0] \tag{3.1}$$

which stands for the probability of observing as extreme a value of the statistic $T(X_1, \cdots, X_n)$ as the observed one $T(x_1, \cdots, x_n)$ under the null hypothesis. Here $x_1, \cdots, x_n$ denote the observed realization of the X_i's. Since the null hypothesis H_0 is rejected for large values of $T(x_1, \cdots, x_n)$, this is equivalent to rejecting H_0 for small values of P.

In most meta analysis applications, the P values are computed from the approximate normal distribution of the relevant test statistics. Thus, if $T(X_1, \cdots, X_n)$ is approximately normally distributed with mean $\mu(\theta)$ and variance $\sigma^2(\theta, n)$, the P value is computed as

$$\begin{aligned} P &= P\left[T(X_1, \cdots, X_n) > T(x_1, \cdots, x_n)|H_0\right] \\ &= P\left[N(0,1) > \frac{T(x_1, \cdots, x_n) - \mu(\theta_0)}{\sigma(\theta_0, n)}\right]. \end{aligned} \tag{3.2}$$

General principle:
Consider k different studies in which test problems H_{0i} versus H_{1i} are considered, $i = 1, \ldots, k$.
A combined test procedure tests the *global null hypothesis*

$$H_0 : \text{ All } H_{0i} \text{ true } \ i = 1, \ldots, k$$

versus the alternative

$$H_1: \text{ Some of the } H_{1i} \text{ true}$$

Note: The problem of selecting a test for H_0 is complicated by the fact that there are many different ways in which the omnibus null hypothesis H_0 can be false.

Two general properties of a combined test procedure:

- *admissibility*
  A combined test procedure is said to be admissible if it provides a (not necessarily the only) most powerful test against some alternative hypothesis for combining some collection of tests.

- *monotonicity*
  A combined test procedure is said to be monotone if the combined test procedure rejects the null hypothesis H_0 for one set of P values and it must also reject the hypothesis for any set of componentwise smaller P values.

Birnbaum (1954): every monotone combined test procedure is admissible and therefore optimal for some testing situation.

3.2 Description of Combined Tests

Two broad classes of combined tests based on the P values:

- *uniform* distribution methods, e.g. Tippett's method and Wilkinson's method

- *probability transformation* methods, e.g. Stouffer's method (inverse normal method), a modified (weighted) Stouffer's method, Fisher's method, and logit method

Each of the methods described below satisfies the monotonicity principle and is therefore optimal for some testing situation.

Minimum P Method

Tippett's (1931) minimum P test rejects the null hypothesis H_0 if any of the k P values is less than α^*, where $\alpha^* = 1 - (1-\alpha)^{\frac{1}{k}}$. In other words, we reject H_0 if

$$\min(P_1, \cdots, P_k) = P_{[1]} < \alpha^* = 1 - (1-\alpha)^{\frac{1}{k}}. \tag{3.3}$$

Example 3.1

Consider the P values: 0.015, 0.077, 0.025, 0.045, 0.079. The minimum P value is $P_{[1]}$=0.015. With $\alpha = 0.05$, the cut-off point is $\alpha^* = 1 - (1-0.05)^{\frac{1}{5}}$=1 - 0.9898= 0.0102. Since $P_{[1]} = 0.015 > \alpha^*$ =0.0102, the minimum P test fails to reject H_0 for this set of data.

Wilkinson's method

This method, due to Wilkinson (1951), rejects H_0 if the rth smallest P value, $P_{[r]}$, is *small*, i.e., less than some c for some fixed r. Since under H_0, $P_{[r]}$ follows a beta distribution with the parameters r and $k - r + 1$, it is easy to determine the cut-off point c for this test from the following equation:

$$\alpha = \int_0^c \frac{u^{r-1}(1-u)^{k-r}}{B(r, k-r+1)} du \tag{3.4}$$

where $B(.,.)$ is the usual *beta function*.

Example 3.2

Consider the same set of probabilities 0.015, 0.077, 0.025, 0.045, 0.079. Taking $r = 2$, a direct computation shows $c = 0.077$. Since $P_{[2]} = 0.025$, we reject H_0. Similarly, we can get $c = 0.1892$ for $r = 3$ and $c = 0.3425$ for $r = 4$, and since $P_{[3]} = 0.045$ and $P_{[4]} = 0.077$, we reject H_0 in all these cases.

Stouffer's method

This method is due to Stouffer and his colleagues (Stouffer et al. 1949), also called **inverse normal method**. It is based on the fact that the z value based on the P value, defined as

$$z = \Phi^{-1}(P) \tag{3.5}$$

is a standard normal variable under the null hypothesis H_0, where $\Phi(.)$ is the standard normal *cdf*. Thus, when the P values $P_1, \cdots, P_k$ are converted to the z values $z_1, \cdots, z_k$, we have *iid* standard normal variables under H_0. The combined significance test is essentially based on the sum of these z values, which has a normal distribution under the null hypothesis with mean 0 and variance k. The test statistic

$$Z = \sum_{i=1}^{k} z(P_i)/\sqrt{k} \tag{3.6}$$

is thus a standard normal variable under H_0, and hence can be compared with the critical values in the standard normal table. Since small P values correspond to small (in fact, negative) z values, the combined test rejects H_0 when Z is less than $-z_\alpha$, or, equivalently, $|Z| > z_\alpha$.

Remark Some authors suggest to compute the z scores from the P values by using the formula

$$z = \Phi^{-1}(1 - P). \tag{3.7}$$

If this is done, the resulting Z value, say Z^*, will be large for small values of P, implying thereby that H_0 is rejected when Z^* is large.

Example 3.3 We can compute the value of the sum of z's test for our small data set given above in the above examples. First, we obtain the standard normal deviates for the five P values, namely,

$$z(0.015) = -2.1701, \quad z(0.077) = -1.4257, \quad z(0.025) = -1.9601,$$
$$z(0.045) = -1.6961, \quad z(0.079) = -1.4257$$

The z values are summed, giving $\sum_{i=1}^{5} z(P_i) = -8.6637$. The sum is divided by the square root of $k = 5$, leading to the normal test statistic $Z = -3.8745$. $|Z|$ is compared with the critical value z_α for a one-tailed test at α=0.05, which is 1.645. Thus, the sum of z's test also rejects H_0 for this data set.

Fisher's method

This method, which is a special case of the inverse chi-square transform, was described by Fisher (1932), and is widely used in meta analysis. The method is based on the fact that the variable $-2 \ln P$ is distributed as a chi-square variable with 2 degrees of freedom under the null hypothesis whenever P has a uniform distribution. The sum of k of these values is therefore a chi-square variable with $2k$ degrees of freedom under H_0. The test thus rejects H_0 when $-2\sum_{i=1}^{k} \ln P_i$ exceeds the $100(1-\alpha)\%$ critical value of the chi-square distribution with $2k$ degrees of freedom.

Example 3.4

We compute the sum of logs statistic for the sample data set given above. First, we compute the natural logarithm of each P value:

$$\ln(0.015) = -4.1997, \quad \ln(0.077) = -2.5639, \quad \ln(0.025) = -3.6888,$$
$$\ln(0.045) = -3.1011, \quad \ln(0.079) = -2.5383.$$

These values are summed and multiplied by -2. The value of the test statistic is $-2 \times (-16.0919) = 32.1839$. We compare this value with $\alpha = 0.05$ upper-tail critical value, which is 18.307 for the chi-square distribution with 10 degrees of freedom. Therefore, we reject H_0 using Fisher's procedure.

Logit Method

George (1977) proposed this method using the statistic

$$G = -\sum_{i=1}^{k} \ln\left(P_i/(1-P_i)\right) \left[k\pi^2(5k+2)/3(5k+4)\right]^{-1/2} \tag{3.8}$$

as another combined significance technique. The argument is that the logit (i.e., $\ln(P/(1-P))$ is distributed as a logistic variable under H_0, and further that the distribution of the sum of the logits, suitably normalized, is close to the t distribution. There are usually two approximations of the null distribution of G which can be used. First, we can approximate the null distribution of G with the t distribution based on $(5k+4)$ degrees of freedom. The

test based on this approximation rejects H_0 if G exceeds the $100(1-\alpha)\%$ critical value of the t distribution with $(5k+4)$ degrees of freedom. Another approximation is based on the observation that, under H_0, $\ln(P_i/(1-P_i))$ could be viewed as approximately normal with a zero mean and variance of $\pi^2/3$. The test based on this approximation therefore rejects H_0 when

$$G^* = \Big[-\sum_{i=1}^{k} \ln(P_i/(1-P_i)) \Big] \left[3/k\pi^2\right]^{1/2} \tag{3.9}$$

exceeds z_α.

Example 3.5 We apply the logit test for the same data set as above. We first compute the natural logarithm of $(P/(1-P))$ for each P value. The values are

$$\begin{aligned} &\ln(0.015/0.985) = -4.1846, \quad \ln(0.077/0.923) = -2.4838, \\ &\ln(0.025/0.075) = -3.6636, \quad \ln(0.045/0.955) = -3.0550, \\ &\ln(0.079/0.921) = -2.4560. \end{aligned}$$

These values are summed, which gives $\sum_{i=1}^{5} \ln(P_i/(1-P_i)) = -15.843$. The sum is multiplied by - $\left[5\pi^2(27)/(3\times 29)\right]^{-1/2}$ or -0.2555. The resultant test statistic is 4.048, which is compared with the $100(1-\alpha)$ percentile point of the t distribution with 29 degrees of freedom. The critical value being 1.699 for $\alpha = 0.05$, we reject H_0 on the basis of the logit method.

3.3 Comparisons of Methods, Criticism and Recommendations

There is no general recommendation for the choice of the combination method. All the combination methods are optimal for some testing situations. Hedges and Olkin (1985) summarize some results on the performance of the above combination methods considering criteria as admissibility, monotonicity, and Bahadur-efficiency. They conclude that Fisher's test is perhaps the best one to use if there is no indication of particular alternatives.

Marden (1991) introduces the notions sensitivity and sturdiness to compare the performance of combination test procedures. Based on five combination methods, namely minimum P, maximum P (Wilkinson's test with largest P-value), sum of P's, sum of logs, and sum of z's, again Fisher's test turns out best.

Example 3.6
The following example is taken from Draper et al. (1992).

Table 3.1. Number of patients and mortality rate from all causes, for six trials comparing the use of aspirin and placebo by patients following a heart attack

	Aspirin		Placebo		Comparison			
Study	No. of Pat.	Mort. Rate (%)	No. of Pat.	Mort. Rate (%)	Diff (%)	SE Diff	Z_i	P_i
UK-1	615	7.97	624	10.74	2.77	1.65	1.68	0.047
CDPA	758	5.80	771	8.30	2.50	1.31	1.91	0.028
GAMS	317	8.52	309	10.36	1.84	2.34	0.79	0.216
UK-2	832	12.26	850	14.82	2.56	1.67	1.54	0.062
PARIS	810	10.49	406	12.81	2.31	1.98	1.17	0.129
AMIS	2267	10.85	2257	9.70	-1.15	0.90	-1.27	0.898

The first five trials are in remarkable agreement with each other. However, the last study (AMIS) is by far the largest, and its large P value runs counter to the small values from the other five studies.

Let us first combine only the P values of the first five studies and use Tippett's (Minimum P) and Fisher's method. The smallest P value is 0.028 and, with $\alpha = 0.05$, the cut-off point is $\alpha^* = 1 - (1 - 0.05)^{\frac{1}{5}} = 1 - 0.9898 = 0.0102$. Consequently, we cannot reject the null hypothesis. The corresponding P value of Tippett's method is 0.1324. Using Fisher's method, the value of the test statistic is 25.988 and, with $\alpha = 0.05$, the cut-off point of the χ^2-distribution with 10 degrees of freedom is 18.307. We reject the null hypothesis based on Fisher's method. The corresponding P value is 0.0038. The two methods disagree markedly with respect to statistical significance.

Including the AMIS trial, the smallest P value is again 0.028 and the cut-off point is now $1 - (1 - 0.05)^{\frac{1}{6}} = 0.0085$. The value of Fisher's test statistic is 26.204 and, with $\alpha = 0.05$, the cut-off point of the χ^2-distribution with 12 degrees of freedom is 21.026. Again, the two methods disagree markedly with respect to statistical significance. Still more important, the corresponding P values of Tippett's and Fisher's method, 0.1567 and 0.01, do not differ so much from the results for the first five studies although the combined sample size is increased from 6292 to 10816, that is, by 72% and dramatically different results were observed.

Combining P values can lead to incorrect conclusions because

- acceptance or rejection can depend more on the choice of the statistic than on the data,
- the information in a highly informative experiment can be masked, and thereby largely disregarded.

Recommendations:
A P value itself is not as informative as the estimate and standard error on which it is based. If this more complete summary information about a study is available, it makes good sense to use it and avoid P values altogether. However, methods that combine P values have their place when such precise information is unavailable.

4 Methods of Combining Effect Sizes

Here we describe the standard methods of combining effect sizes from various independent studies for both point estimation as well as confidence interval estimation. We refer to Rosenthal (1994) for further reading.

The general principle is the following. Consider k *independent* studies with the ith study resulting in the estimated effect size T_i, which is an estimate of the population effect size θ_i, and suppose $\hat{\sigma}^2(T_i)$ is the estimated variance of T_i, $i = 1, \cdots, k$. Usually, T_i is based on a random sample of size n_i from the ith population or study, and, in large samples, T_i has an approximate normal distribution with mean θ_i and variance $\sigma^2(T_i) = \sigma^2_{(\theta_i;n_i)}$. In most cases the variance $\sigma^2_{(\theta_i;n_i)}$ indeed depends on θ_i so that it is unknown, and $\hat{\sigma}^2(T_i)$ represents an estimate of $\sigma^2_{(\theta_i;n_i)}$. In some cases, T_i may be stochastically independent of $\hat{\sigma}^2(T_i)$.

We *assume* that

$$\theta_1 = \cdots = \theta_k = \theta \tag{4.1}$$

where θ denotes the common population effect size. Then a combined estimate of θ is given by a weighted combination of the T_i's, namely,

$$\hat{\theta} = \frac{\sum_{i=1}^k w_i T_i}{\sum_{i=1}^k w_i} \tag{4.2}$$

where w_i is a nonnegative weight assigned to the ith study. This very general method of linearly combining T_i's to derive an estimate of a common mean effect dates back to Cochran (1937). Clearly, for any choice of the *nonstochastic* weights w_i's, $\hat{\theta}$ is an unbiased estimate of θ, and the weights which make $var(\hat{\theta})$ the smallest are given by

$$w_i = 1/\sigma^2_{(\theta_i;n_i)}, \;\; i = 1, \cdots, k. \tag{4.3}$$

However, the above *optimum* weights are typically unknown since the variances $\sigma^2_{(\theta_i;n_i)}$ will usually be unknown, and hence cannot be used. When $\sigma^2_{(\theta_i;n_i)}$ is estimated and thus replaced by $\hat{\sigma}^2(T_i)$, this results in the special weighted combination

$$\tilde{\theta} = \frac{\sum_{i=1}^k T_i/\hat{\sigma}^2(T_i)}{\sum_{i=1}^k 1/\hat{\sigma}^2(T_i)} \tag{4.4}$$

with the estimated $var(\tilde{\theta})$ as

$$\hat{\sigma}^2(\tilde{\theta}) = \widehat{var}(\tilde{\theta}) = \frac{1}{\sum_{i=1}^k 1/\hat{\sigma}^2(T_i)}. \tag{4.5}$$

More generally, we can also attach a *quality* index q_i to the ith study along with the nonnegative weights w_i's, thus yielding an unbiased estimate of θ given by

$$\hat{\theta}^* = \frac{\sum_{i=1}^k q_i\, w_i\, T_i}{\sum_{i=1}^k q_i\, w_i} \tag{4.6}$$

with its estimated variance as

$$\hat{\sigma}^2(\hat{\theta}^*) = \widehat{var}(\hat{\theta}^*) = \frac{\sum_{i=1}^k q_i^2\, w_i^2\, \hat{\sigma}^2(T_i)}{\left(\sum_{i=1}^k q_i\, w_i\right)^2}. \tag{4.7}$$

In any event, when a combined estimate of θ, say T, is thus derived along with its estimated standard error given by $\hat{\sigma}(T)$, a confidence interval for θ with confidence level $(1-\alpha)$ is approximated by

$$LB = T - z_{\alpha/2}\, \hat{\sigma}(T),\;\; UB = T + z_{\alpha/2}\, \hat{\sigma}(T) \tag{4.8}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ cut-off point obtained from a standard normal table. Moreover, if the above confidence interval does *not* contain 0, we reject the null hypothesis $H_0 : \theta = 0$ at level α in favor of the alternative $H_1 : \theta \neq 0$. Equivalently, we may test the null hypothesis $H_0 : \theta = 0$ at level α against the alternative $H_1 : \theta \neq 0$ by rejecting H_0 if

$$|Z| = \frac{|T|}{\hat{\sigma}(T)} > z_{\alpha/2}. \tag{4.9}$$

Finally, based on the data from k studies, we can also test the validity of the assumption (4.1) by using a chi-square test. Using $\tilde{\theta}$, this test is based on the large sample chi-square statistic (Cochran, 1937)

$$\chi^2 = \sum_{i=1}^k \frac{(T_i - \tilde{\theta})^2}{\hat{\sigma}^2(T_i)} = \sum_{i=1}^k \frac{T_i^2}{\hat{\sigma}^2(T_i)} - \frac{(\sum_{i=1}^k T_i/\hat{\sigma}^2(T_i))^2}{\sum_{i=1}^k 1/\hat{\sigma}^2(T_i)}, \tag{4.10}$$

and we reject H_0 if $\chi^2 > \chi^2_{k-1,\alpha}$.

We now discuss a few examples to illustrate the applications of the above methods.

Example 4.1. We refer to the data set below dealing with validity correlation studies.

Table 4.1. Validity Studies Correlation Student Ratings of the Instructor with Student Achievement

Study	n	r	Study	n	r
1	10	0.68	11	36	-0.11
2	20	0.56	12	75	0.27
3	13	0.23	13	33	0.26
4	22	0.64	14	121	0.40
5	28	0.49	15	37	0.49
6	12	-0.04	16	14	0.51
7	12	0.49	17	40	0.40
8	36	0.33	18	16	0.34
9	19	0.58	19	14	0.42
10	12	0.18	20	20	0.16

For this data set, using r_i as T_i and recalling that $\widehat{var}(r_i) = \hat{\sigma}^2(T_i) = (1 - r_i^2)^2/(n_i - 1)$, we obtain $\sum_{i=1}^{20} T_i/\hat{\sigma}^2(T_i) = 337.002$, $\sum_{i=1}^{20} 1/\hat{\sigma}^2(T_i) = 847.185$, $\sum_{i=1}^{20} T_i^2/\hat{\sigma}^2(T_i) = 159.687$. This leads to

$$\tilde{\theta} = \left[\sum_{i=1}^{20} T_i/\hat{\sigma}^2(T_i)\right] / \left[\sum_{i=1}^{20} 1/\hat{\sigma}^2(T_i)\right] = 0.3978 \tag{4.11}$$

and

$$\widehat{var}(\tilde{\theta}) = 1/\left[\sum_{i=1}^{20} 1/\hat{\sigma}^2(T_i)\right] = 0.00118. \tag{4.12}$$

Moreover, taking $\alpha = 0.05$, we get

$$LB = \tilde{\theta} - 1.96\sqrt{\widehat{var}(\tilde{\theta})} = 0.3305, \;\; UB = \tilde{\theta} + 1.96\sqrt{\widehat{var}(\tilde{\theta})} = 0.4651. \tag{4.13}$$

For testing $H_0 : \theta = 0$, we compute $|Z| = 11.58$, which implies we reject H_0 at level 0.05. Finally, the test for homogeneity of the θ_i's is carried out by computing $\chi^2 = 25.65$, which when compared with the table value 30.14 of χ^2 with 19 *df* leads to acceptance of the assumption (4.1).

Using Fisher's z–transformation for this data set, that is,
$z_i = 0.5 \ln((1 + r_i)/(1 - r_i))$, and recalling that $\widehat{var}(z_i) = \hat{\sigma}^2(z_i) = 1/(n_i - 3)$, we obtain $\sum_{i=1}^{20} z_i/\hat{\sigma}^2(z_i) = 201.3513$, $\sum_{i=1}^{20} 1/\hat{\sigma}^2(z_i) = 530$, $\sum_{i=1}^{20} z_i^2/\hat{\sigma}^2(z_i) = 97.4695$. This leads to

$$\tilde{\zeta} = \Big[\sum_{i=1}^{20} z_i/\hat{\sigma}^2(z_i)\Big] \Big/ \Big[\sum_{i=1}^{20} 1/\hat{\sigma}^2(z_i)\Big] = 0.3799 \tag{4.14}$$

and

$$\hat{v}ar(\tilde{\zeta}) = 1/\Big[\sum_{i=1}^{20} 1/\hat{\sigma}^2(z_i)\Big] = 0.00189. \tag{4.15}$$

Moreover, taking $\alpha = 0.05$, we get

$$LB = \tilde{\zeta} - 1.96\sqrt{\widehat{var}(\tilde{\zeta})} = 0.2948, \;\; UB = \tilde{\zeta} + 1.96\sqrt{\widehat{var}(\tilde{\zeta})} = 0.4650. \tag{4.16}$$

For testing $H_0 : \zeta = 0$, we compute $|Z| = 8.74$, which implies we reject H_0 at level 0.05. Finally, the test for homogeneity of the ζ_i's is carried out by computing $\chi^2 = 20.97$, which when compared with the table value 30.14 of χ^2 with 19 *df* leads to acceptance of the assumption (4.1).

Converting results from (4.14) and (4.16), we obtain $\tilde{\theta} = 0.3626$ with 95% confidence interval $[0.2865, 0.4342]$.

Example 4.2. Here we examine the data reported in Meier (1953) about the percentage of albumin in plasma protein in human subjects.

Table 4.2. Percentage of albumin in plasma protein

Experiment	n_i	Mean	Variance s_i^2	95% CI on mean
A	12	62.3	12.986	[60.0104 , 64.5896]
B	15	60.3	7.840	[58.7494 , 61.8506]
C	7	59.5	33.433	[54.1524 , 64.8476]
D	16	61.5	18.513	[59.2073 , 63.7927]

For this data set, using the mean as T_i and the variance of T_i as $\hat{\sigma}^2(T_i) = s_i^2/n_i$, we obtain $\sum_{i=1}^4 T_i/\hat{\sigma}^2(T_i) = 238.5492$, $\sum_{i=1}^4 1/\hat{\sigma}^2(T_i) = 3.9110$, and $\sum_{i=1}^4 T_i^2/\hat{\sigma}^2(T_i) = 14553.47$. This leads to

$$\tilde{\theta} = \left[\sum_{i=1}^4 T_i/\hat{\sigma}^2(T_i)\right] \Big/ \left[\sum_{i=1}^4 1/\hat{\sigma}^2(T_i)\right] = 60.9949 \tag{4.17}$$

and

$$\hat{v}ar(\tilde{\theta}) = 1 \Big/ \left[\sum_{i=1}^4 1/\hat{\sigma}^2(T_i)\right] = 0.2557. \tag{4.18}$$

Moreover, taking $\alpha = 0.05$, we get

$$LB = \tilde{\theta} - 1.96\sqrt{\widehat{var}(\tilde{\theta})} = 60.0038, \;\; UB = \tilde{\theta} + 1.96\sqrt{\widehat{var}(\tilde{\theta})} = 61.9860. \tag{4.19}$$

Finally, the test for homogeneity of the θ_i's is carried out by computing $\chi^2 = 3.1862$, which when compared with the table value 7.815 of χ^2 with 3 *df* leads to acceptance of the assumption (4.1).

Example 4.3. This data are quoted from Eberhardt et al. (1989) and deal with the problem of estimation of mean *Selenium* in non-fat milk powder by combining the results of four methods.

Table 4.3. Selenium in non-fat milk powder

Methods	n_i	Mean	Variance s_i^2	95% CI on mean
Atomic absorption spectrometry	8	105.0	85.711	[97.2601 , 112.7399]
Neutron activation:				
1). Instrumental	12	109.75	20.748	[106.8559 , 112.6441]
2). Radiochemical	14	109.5	2.729	[108.5462 , 110.4538]
Isotope dilution mass spectrometry	8	113.25	33.640	[108.4011 , 118.0989]

For this data set, using the mean as T_i and the variance as $\hat{\sigma}^2(T_i) = s_i^2/n_i$, we obtain $\sum_{i=1}^4 T_i/\hat{\sigma}^2(T_i) = 661.9528$, $\sum_{i=1}^4 1/\hat{\sigma}^2(T_i) = 6.0396$, and $\sum_{i=1}^4 T_i^2/\hat{\sigma}^2(T_i) = 72556.6$. This leads to

$$\tilde{\theta} = \left[\sum_{i=1}^4 T_i/\hat{\sigma}^2(T_i)\right] \Big/ \left[\sum_{i=1}^4 1/\hat{\sigma}^2(T_i)\right] = 109.6021 \tag{4.20}$$

and

$$\widehat{var}(\tilde{\theta}) = 1/[\sum_{i=1}^{4} 1/\hat{\sigma}^2(T_i)] = 0.1656. \tag{4.21}$$

Moreover, taking $\alpha = 0.05$, we get

$$LB = \tilde{\theta} - 1.96\sqrt{\widehat{var}(\tilde{\theta})} = 108.8045, \;\; UB = \tilde{\theta} + 1.96\sqrt{\widehat{var}(\tilde{\theta})} = 110.3996. \tag{4.22}$$

Finally, the test for homogeneity of the θ_i's is carried out by computing $\chi^2 = 5.2076$, which when compared with the table value 7.815 of χ^2 with 3 *df* leads to acceptance of the assumption (4.1).

5 Inference about a Common Mean of Normal Populations

In this lecture we consider a very special kind of a meta analysis problem, namely, statistical inference about the common mean of several univariate normal populations with unknown and possibly unequal variances, and provide a review of this rich literature.

One of the oldest and interesting problems in statistical meta analysis is inference about a common mean of several univariate normal populations with unknown and possibly unequal variances. The motivation of this problem comes from a balanced incomplete block design (BIBD) with uncorrelated random block effects and fixed treatment effects. In this set up, one has two estimates—namely, the intra-block estimate $\hat{\tau}$ and the inter-block estimate $\tilde{\tau}$ of the vector τ of treatment contrasts. Under the usual assumption of normality and independence, $\hat{\tau}$ and $\tilde{\tau}$ are independent, following normal distributions with a common mean vector τ but unknown and unequal intra-block and inter-block variances (see Montgomery (1991), p:184-186). The problem thus is to derive an estimate of τ on the basis of $\hat{\tau}$ and $\tilde{\tau}$, and also to provide some tests for hypotheses concerning this common vector of treatment contrasts. This, of course, is a multivariate version of the standard univariate common mean problem which is the subject of discussion of this chapter. The special case of two populations with equal sample sizes is treated with some details.

Another feature of this meta analysis problem which makes it distinct is that it does *not* correspond to the usual set up of combining data from different studies taking place at different sources which are not controlled by the statistician. Rather, here the experiments are *designed* to provide duplicate information about a parameter. Our two examples presented later in this lecture will make this point clear.

This lecture is organized as follows. After some preliminary discussion about the model and the inference problem in this section, we consider in section 5.1 the problem of point estimation of the common mean in details. An asymptotic comparison of some selected estimates of the common mean in the case of two normal populations with equal sample sizes is provided in section 5.2. This section also contains a discussion about the Bayes estimate of μ under Jeffrey's invariant prior. The related problem of test and confidence interval of the common mean is taken up in section 5.3. Two illustrative examples showing computations of our proposed methods are mentioned in section 5.4. We end this lecture with an Appendix containing some technical details.

To be specific, let us assume that in general there are k independent univariate normal populations where the i^{th} population follows $N(\mu, \sigma_i^2)$ distribution, $\mu \in \Re$, $\sigma_i^2 > 0, 1 \leq i \leq k$. Let $X_{ij}, j = 1, 2, \ldots, n_i$ $(n_i \geq 2)$ be *iid* observations from the i^{th} population, $1 \leq i \leq k$. Define $\bar{X}_i$ and S_i^2 as

$$\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij} \qquad \text{and} \qquad S_i^2 = \frac{1}{(n_i-1)}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2, \;\; 1 \le i \le k. \tag{5.1}$$

Note that $(\bar{X}_i, S_i^2, 1 \le i \le k)$ is minimal sufficient for $(\mu, \sigma_1^2, \cdots, \sigma_k^2)$ even though it is not complete. Observe that

$$\bar{X}_i \sim N(\mu, \frac{\sigma_i^2}{n_i}), \quad (n_i-1)S_i^2 \sim \sigma_i^2 \chi^2_{n_i-1}, \quad 1 \le i \le k \tag{5.2}$$

and they are all mutually independent.

Estimation of the common mean μ in the above context has drawn the attention of many researchers over the last four decades from both classical as well as decision theoretic point of view. We now provide a brief historical perspective of the problem of point estimation of μ.

If the population variances $(\sigma_1^2, \cdots, \sigma_k^2)$ are completely known, then the maximum likelihood estimator (MLE) of μ is given as

$$\hat{\mu}(\sigma_1^2, \cdots, \sigma_k^2) = \sum_{i=1}^{k} \frac{n_i}{\sigma_i^2}\bar{X}_i \Big/ \sum_{i=1}^{k} \frac{n_i}{\sigma_i^2}. \tag{5.3}$$

The above estimator is also the unique minimum variance unbiased estimator (UMVUE) under normality as well as the best linear unbiased estimator (BLUE) without normality for estimating μ. Note that in the two populations case and for equal sample sizes (i.e., $k = 2$ and $n_1 = n_2 = n$) we only need to know $\tau = \sigma_2^2/\sigma_1^2$ (apart from $\bar{X}_1$ and $\bar{X}_2$) to obtain $\hat{\mu}(\sigma_1^2, \sigma_2^2)$.

If the population variances are completely unknown, the estimation of the common mean μ becomes nontrivial and more interesting. One can try to find the MLEs of $\mu, \sigma_1^2, \cdots, \sigma_k^2$ by solving the following system of equations:

$$\hat{\mu}\sum_{i=1}^{k} \frac{n_i}{\hat{\sigma}_i^2} = \sum_{i=1}^{k} \frac{n_i}{\hat{\sigma}_i^2}\bar{X}_i \quad \text{and} \quad \hat{\sigma}_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i}(X_{ij}-\hat{\mu})^2, \;\; 1 \le i \le k. \tag{5.4}$$

Clearly the MLE of μ does not have a closed form. However, one can obtain an estimator of μ from the expression (5.3) by replacing σ_i^2 by $\hat{\sigma}_i^2 = S_i^2$. The estimator of μ thus obtained is the well known *Graybill-Deal* estimator given by

$$\hat{\mu}_{GD} = \sum_{i=1}^{k} \frac{n_i}{S_i^2}\bar{X}_i \Big/ \sum_{i=1}^{k} \frac{n_i}{S_i^2}. \tag{5.5}$$

Using the mutual independence of $\bar{X}_i$'s and S_i's, it is readily verified that $\hat{\mu}_{GD}$ is an unbiased estimator of μ.

Even though Graybill and Deal (1959) pioneered the research on common mean estimation, it is probably due to Zacks (1966, 1970) that many researchers paid attention to this age old problem, especially from a decision theoretic point of view. Zacks (1966, 1970) was motivated by the applications and in his own words — " $\cdots$ The best of my papers were motivated by consulting problems. $\cdots$ In 1963, I was approached by a soil engineer. He wanted to estimate the common mean of two populations and he didn't know anything about the variances. But, a priori from his theory he said that the means should be same, and here are the two samples from two different soils. So I thought about this problem a little bit and I started to investigate. I realized that there is room for innovation $\cdots$ " (see 'Research-How to do it : A panel discussion' by Kempthorne et al. (1991)).

A good amount of work has been done dealing with the properties of $\hat{\mu}_{GD}$ or its variations in relation to other estimators. In the next section we review this literature from both classical as well as decision theoretic point of view.

5.1 Results on common mean estimation

Broadly speaking, the research on common mean estimation can be categorized as:

(i) a small sample comparison of $\hat{\mu}_{GD}$ with other estimators;
(ii) properties of $\hat{\mu}_{GD}$.

We address the above two categories separately.

5.1.1 Small sample comparison of $\hat{\mu}_{GD}$ with other estimators

Note that the estimator $\hat{\mu}_{GD}$ is unbiased for μ. But since $\hat{\mu}_{GD}$ uses sufficient statistics, it is expected that this estimator should have smaller variance than the individual sample means. Obviously

$$V(\bar{X}_i) = \frac{\sigma_i^2}{n_i}\ , \ \ 1 \leq i \leq k;$$

and a standard conditional argument yields

$$\begin{aligned} V(\hat{\mu}_{GD}) &= E\left\{V(\hat{\mu}_{GD}|S_1, \cdots, S_k)\right\} + V\left\{E(\hat{\mu}_{GD}|S_1, \cdots, S_k)\right\} \\ &= E\left[\left\{\sum_{i=1}^{k} \frac{n_i \sigma_i^2}{S_i^4}\right\} \Big/ \left\{\sum_{i=1}^{k} \frac{n_i}{S_i^2}\right\}^2\right]. \end{aligned} \tag{5.6}$$

The exact variance expression of $\hat{\mu}_{GD}$ (the expectation in (5.6)) is not easy to get. However Khatri and Shah (1974) derived this exact variance for $k = 2$ in an infinite series form involving hypergeometric functions. Unfortunately, this infinite series form has little use when one wants to compare $V(\hat{\mu}_{GD})$ against individual sample mean variances $(\sigma_i^2/n_i, 1 \leq i \leq k)$. For the two populations case $(k = 2)$, Graybill and Deal (1959) were the first to derive necessary and sufficient conditions such that

$$V(\hat{\mu}_{GD}) \leq \frac{\sigma_i^2}{n_i}\ , \ \ 1 \leq i \leq k \qquad \text{and for all } \ \sigma_1^2, \cdots, \sigma_k^2. \tag{5.7}$$

The following result is due to Graybill and Deal (1959).

Proposition 5.1. For $k = 2$, the inequality (5.7) holds if and only if $n_i \geq 11, i = 1, 2$.

The implication of the above result is far reaching. If either n_1 or n_2 is less than 11, then $\hat{\mu}_{GD}$ does not have a uniformly smaller variance than $\bar{X}_1$ or $\bar{X}_2$ (i.e., $\bar{X}_1$ or $\bar{X}_2$ can *sometimes* be better than $\hat{\mu}_{GD}$ in terms of variance). This was later extended by Norwood and Hinkelmann (1977) for k populations, which is stated below.

Proposition 5.2. The inequality (5.7) holds if and only if

(a) $n_i \geq 11 \ \ \forall\, i$; or

(b) $n_i = 10$ for some i and $n_j \geq 19 \ \ \forall\, j \neq i$.

It is possible to generalize the **Proposition 5.2** further by considering a more general common mean estimator of μ of the form

$$\hat{\mu}_c = \left\{ \sum_{i=1}^{k} \frac{c_i n_i}{S_i^2} \bar{X}_i \right\} \Big/ \left\{ \sum_{i=1}^{k} \frac{c_i n_i}{S_i^2} \right\} \tag{5.8}$$

where $\mathbf{c} = (c_1, \cdots, c_k)$ is a vector of nonnegative real constants. Obviously $\mathbf{c} = (1, \cdots, 1)$ produces the estimator $\hat{\mu}_{GD}$. The following result which is an extension of **Proposition 5.2** is due to Khatri and Shah (1974) (for $k = 2$) and Shinozaki (1978) (for general k).

Proposition 5.3. The estimator $\hat{\mu}_c$ in (5.8) has a uniformly smaller variance than each $\bar{X}_i$ if and only if

$$\begin{aligned}
&\text{(a)} \qquad \frac{c_j}{c_i} \leq 2\frac{(n_i - 1)(n_j - 5)}{(n_i + 1)(n_j - 1)} \quad \forall \ \ i \neq j; \\
&\text{(b)} \qquad n_i \geq 8 \quad \forall \ \ i; \quad \text{and} \\
&\text{(c)} \qquad (n_i - 7)(n_j - 7) \geq 16 \ \ \forall i \neq j.
\end{aligned}$$

Even though the estimators in (5.8) are more general than $\hat{\mu}_{GD}$, for all practical purposes $\hat{\mu}_{GD}$ seems to be the most natural choice in this class. This is more obvious when the sample sizes are all equal, i.e., when $n_1 = \cdots = n_k = n$ (say), because then the **Proposition 5.3** implies that $V(\hat{\mu}_{GD}) \leq \sigma_i^2/n \ \forall\, i$ if and only if $n \geq 11$.

A question which arises naturally is : 'Is it possible to improve over $\bar{X}_i$ $(1 \leq i \leq k)$ for smaller sample sizes by using estimators other than $\hat{\mu}_{GD}$?' Investigation on unbiased estimators other than $\hat{\mu}_{GD}$ was stimulated by the works of Cohen and Sackrowitz (1974), and Brown and Cohen (1974).

Cohen and Sackrowitz (1974) considered the simple case of $k = 2$ and $n_1 = n_2 = n$. Define $T = S_2^2/S_1^2$ and

$$\begin{aligned} G_n(T) &= {}_2F_1(1,(3-n)/2;(n-1)/2;T) \ \ for \ \ 0 \leq T \leq 1; \\ &= ((n-3)/(n-1))T^{-1}{}_2F_1(1,(5-n)/2;(n+1)/2;T^{-1}) \ \text{ for } \ T > 1. \end{aligned} \tag{5.9}$$

where ${}_2F_1$ is a hypergeometric function

Proposition 5.4. For $k = 2$ and $n_1 = n_2 = n$, consider the common mean estimator

$$\hat{\mu}(a_n) = (1 - a_n G_n(T))\bar{X}_1 + a_n G_n(T)\bar{X}_2 \tag{5.10}$$

where $a_n = (n-3)^2/((n+1)(n-1))$ for n odd; $= (n-4)/(n+2)$ for n even. The estimator $\hat{\mu}(a_n)$ is unbiased and minimax for all $n \geq 5$. Also, the estimator $\hat{\mu}(1)$ (i.e., replace a_n by 1) is better than both $\bar{X}_1$ and $\bar{X}_2$ for $n \geq 10$.

As $n \to \infty$, $G_n(T) \to (1+T)^{-1}$ and $a_n \to 1$. Therefore, the weights given to the sample means in (5.10) are converging strongly to the optimal weights in the case where the variances are known. Hence, for large values of n, the estimator $\hat{\mu}(a_n)$ is essentially the same as the estimator $\hat{\mu}_{GD}$. Note that $\hat{\mu}(a_n)$ is better than $\bar{X}_1$ for $n \geq 5$, whereas $\hat{\mu}_{GD}$ is not better than either $\bar{X}_1$ or $\bar{X}_2$ for $n < 11$. For $n = 10$, $\hat{\mu}(1)$ has a smaller variance than $\bar{X}_i (i = 1, 2)$ and this is clearly an advantage over $\hat{\mu}_{GD}$. Cohen and Sackrowitz (1974) also provided some other type of unbiased estimators which are better than $\bar{X}_1$ only for $n \geq 5$.

Brown and Cohen (1974) considered the case of unequal sample sizes for $k = 2$ and obtained the following result.

Proposition 5.5. Assume $k = 2$ and $n_1, n_2 \geq 2$. The estimator

$$\hat{\mu}^a = \bar{X}_1 + a(\bar{X}_2 - \bar{X}_1)\left(\frac{S_1^2}{n_1}\right) \Bigg/ \left\{\frac{S_1^2}{n_1} + \frac{(n_2-1)S_2^2}{n_2(n_2+2)} + \frac{(\bar{X}_1 - \bar{X}_2)^2}{(n_2+2)}\right\} \tag{5.11}$$

is unbiased and has a smaller variance than $\bar{X}_1$ provided $n_2 \geq 3$ and $0 < a \leq a(n_1, n_2)$ where $a(n_1, n_2) = 2(n_2+2)/[nE\{max(V^{-1}, V^{-2})\}]$, where V has F distribution with (n_2+2) and (n_1-1) dfs.

Exact values of $a(n_1, n_2)$ are given in Brown and Cohen (1974) for selected values of (n_1, n_2). It was also shown that when $n_2 = 2$ the estimator $\hat{\mu}^a$ in (5.11) is not better than $\bar{X}_1$ uniformly for any value of a. Brown and Cohen (1974) also considered a slight variation of (5.11) of the form

$$\hat{\mu}_a = \bar{X}_1 + a(\bar{X}_2 - \bar{X}_1)\left(\frac{S_1}{n_1}\right) \Bigg/ \left\{\frac{S_1}{n_1} + \frac{S_2}{n_2}\right\} \tag{5.12}$$

and showed that $\hat{\mu}_a$ has a smaller variance than $\bar{X}_1$ for $n_1 \geq 2$ and $n_2 \geq 6$ whenever $0 < a < a(n_1, n_2 - 3)$.

Unification of all the results presented above appears in an excellent paper by Bhattacharya (1980).

For the two populations equal sample size case, Zacks (1966) considered two quite different classes of estimators. Note that in a decision theoretic set up under the loss function $(\hat{\mu} - \mu)^2 / \max(\sigma_1^2, \sigma_2^2)$, the grand mean $\bar{X} = (\bar{X}_1 + \bar{X}_2)/2$ is admissible as well as minimax (a more general result is due to Kubokawa (1990)). Zacks (1966) combined $\hat{\mu}_{GD}$ and $\bar{X}$ to generate the following two classes of randomized estimators:

$$\hat{\mu}(\tau_o) = I(T, \tau_o)\bar{X} + \{1 - I(T, \tau_o)\}\,\hat{\mu}_{GD} \tag{5.13}$$

and

$$\tilde{\mu}(\tau_o) = I(T, \tau_o)\bar{X} + J_1(T, \tau_o)\bar{X}_1 + J_2(T, \tau_o)\bar{X}_2 \tag{5.14}$$

where

$$I(T, \tau_o) = \begin{cases} 1 & \text{if } \tau_o^{-1} \leq T \leq \tau_o \\ 0 & \text{otherwise;} \end{cases}$$

$$J_1(T, \tau_o) = \begin{cases} 1 & \text{if } T > \tau_o^{-1} \\ 0 & \text{otherwise;} \end{cases}$$

$$J_2(T, \tau_o) = \begin{cases} 1 & \text{if } T < \tau_o^{-1} \\ 0 & \text{otherwise;} \end{cases}$$

and $\tau_o \in [0, \infty)$ is a known constant. The values of τ_o both in $\hat{\mu}(\tau_o)$ and in $\tilde{\mu}(\tau_o)$ are the critical values of the $F-$tests of significance (to compare the variances), according to which one decides whether to apply the estimators $\bar{X}$, $\hat{\mu}_{GD}$, $\bar{X}_1$ or $\bar{X}_2$. Zacks (1966) provided variance and efficiency expressions of $\hat{\mu}(\tau_o)$ and $\tilde{\mu}(\tau_o)$. Somewhat similar classes of estimators have been considered by Mehta and Gurland (1969), but these estimators have very little practical importance.

We now direct our discussion to the second aspect of the problem.

5.1.2 Properties of $\hat{\mu}_{GD}$

Earlier we have seen the variance expression of the unbiased estimator $\hat{\mu}_{GD}$ (see (5.6)). The exact probability distribution of $\hat{\mu}_{GD}$ is somewhat complicated. However, for $k = 2$ and $n_1 = n_2 = n$, Nair (1980) gave an approximate *cdf* of $\hat{\mu}_{GD}$. But for general k if we can find an unbiased estimator $\hat{V}(\hat{\mu}_{GD})$ of $V(\hat{\mu}_{GD})$ then the studentized version $(\hat{\mu}_{GD} - \mu)/\sqrt{\hat{V}(\hat{\mu}_{GD})}$ follows $N(0,1)$ asymptotically (i.e., as $\min_{1\le i\le k} n_i \to \infty$). This can be used for testing as well as interval estimation of μ.

Finding an unbiased estimator $\hat{V}(\hat{\mu}_{GD})$ of $V(\hat{\mu}_{GD})$ is not an easy task. From the expression (5.6), it is enough to have real valued functions $\psi_i = \psi_i(S_1^2, \cdots, S_k^2), 1 \le i \le k$ such that

$$E(\psi_i) = \sigma_i^2 E\left[\left\{S_i^2 \sum_{j=1}^{k} \frac{n_j}{S_j^2}\right\}^{-2}\right]$$

so that an unbiased estimator of $V(\hat{\mu}_{GD})$ is obtained as

$$\hat{V}(\hat{\mu}_{GD}) = \sum_{i=1}^{k} n_i \psi_i. \tag{5.15}$$

Making use of Haff's (1979) Wishart identity for the univariate case, Sinha (1985) derived the expression for ψ_i with the following form

$$\psi_i = \lim_{m\to\infty} \psi_{i,m} \quad , \quad \text{where}$$

$$\psi_{i,m} = \sum_{l=0}^{m-1} \frac{S_i^{2(l+1)} 2^l (l+1)! A_{(-i)}^l}{(n_i+1)^{[l]} (n_i + A_{(-i)} S_i^2)^{l+2}} \quad , \quad m \ge l \tag{5.16}$$

with $A_{(-i)} = \sum_{j\ne i} n_j/S_j^2$ and $(n_i+1)^{[l]} = (n_i+1)\cdots(n_i+2l-1)$ for $l \ge 1$; $= 1$ for $l = 0, i = 1, 2, \cdots, k$. The following result which approximates $\hat{V}(\hat{\mu}_{GD})$ is due to Sinha (1985).

Proposition 5.6. Let $n = \min_{1\le i\le k}(n_i)$. Then using $\psi_{i,m}$ as in (5.16),

$$\left| E\left(\sum_{i=1}^{k} n_i \psi_{i,m}\right) - V(\hat{\mu}_{GD}) \right| = O(n^{-(m+1)}).$$

Using the above result, we get $(\hat{\mu}_{GD} - \mu)/\sqrt{\sum_{i=1}^{k} n_i \psi_{i,m}} \sim N(0,1)$ as $n \to \infty$. A first order approximation to $\hat{V}(\hat{\mu}_{GD})$, say $\hat{V}_{(1)}(\hat{\mu}_{GD})$, is obtained as (by taking $m = 1$)

$$\hat{V}_{(1)}(\hat{\mu}_{GD}) = \left(\sum_{i=1}^{k} \frac{n_i}{S_i^2}\right)^{-1} \left[1 + 4\sum_{i=1}^{k} \frac{n_i}{(n_1+1)S_i^2} \Big/ \left\{\sum_{i=1}^{k} \frac{n_i}{S_i^2} - \frac{n_i^2/S_i^4}{(\sum_{i=1}^{k} n_i/S_i^2)^2}\right\}\right] \tag{5.17}$$

which is comparable to the approximation

$$\hat{V}(\hat{\mu}_{GD}) \approx \left(\sum_{i=1}^{k} \frac{n_i}{S_i^2}\right)^{-1} \left[1 + 4\sum_{i=1}^{k} \frac{n_i}{(n_1-1)S_i^2} \Bigg/ \left\{\sum_{i=1}^{k} \frac{n_i}{S_i^2} - \frac{n_i^2/S_i^4}{(\sum_{i=1}^{k} n_i/S_i^2)^2}\right\}\right] \quad (5.18)$$

due to Meier (1953).

Decision theoretic estimation of the common mean has been addressed by several authors. Zacks (1966) pointed out for $k = 2$ and $n_1 = n_2$ that while $\bar{X}_1$ is minimax under the loss function $(\hat{\mu} - \mu)^2/\sigma_1^2$, a minimax estimator for the loss $(\hat{\mu} - \mu)^2/\max(\sigma_1^2, \sigma_1^2)$ is not $\bar{X}_1$ but $\bar{X} = (\bar{X}_1 + \bar{X}_2)/2$. Kubokawa (1990) extended this result for general k and showed the minimaxity as well as admissibility of the grand mean $\bar{X} = \sum_{i=1}^{k} \bar{X}_i/k$ under the loss function $(\hat{\mu} - \mu)^2/(\max_{1\le i\le k}\ \sigma_i^2)$. Zacks (1970) also derived Bayes and fiducial equivariant estimators for $k = 2$ and gave their variance expressions.

It may be mentioned that, under the standard squared error loss function $(\hat{\mu}-\mu)^2$, the exact admissibility (or otherwise) of $\hat{\mu}_{GD}$ is still an *open* problem. Minimax estimation under the loss $(\hat{\mu} - \mu)^2$ is not meaningful since estimators have unbounded risks under this loss.

Sinha and Mouqadem (1982) considered the special case $k = 2$ and $n_1 = n_2 = n$ and obtained some restricted admissibility results for $\hat{\mu}_{GD}$. Note that $\hat{\mu}_{GD}$ can be written as (with $k = 2$ and $n_1 = n_2 = n$)

$$\hat{\mu}_{GD} = \bar{X}_1 + (\bar{X}_2 - \bar{X}_2)\left(\frac{S_1^2}{S_1^2 + S_2^2}\right) \quad (5.19)$$

which is affine equivariant (i.e., equivariant under the group of transformations $(\bar{X}_1, \bar{X}_2, S_1^2, S_2^2) \to (a\bar{X}_1 + b, a\bar{X}_2 + b, a^2S_1^2, a^2S_2^2), a > 0, b \in \Re$). Let $D = (\bar{X}_2 - \bar{X}_1)$ and define the following four classes of estimators

$$\begin{aligned}
\mathcal{C}_o &= \left\{\hat{\mu} \mid \hat{\mu} = \bar{X}_1 + D\phi_o, \quad 0 \le \phi_o(S_2^2/S_1^2) \le 1\right\}; & (5.20)\\
\mathcal{C}_1 &= \left\{\hat{\mu} \mid \hat{\mu} = \bar{X}_1 + D\phi_1, \quad 0 \le \phi_1(S_1^2, S_2^2) \le 1\right\}; & (5.21)\\
\mathcal{C}_2 &= \left\{\hat{\mu} \mid \hat{\mu} = \bar{X}_1 + D\phi_2, \quad 0 \le \phi_2(S_1^2/D^2, S_2^2/D^2) \le 1\right\}; & (5.22)\\
\mathcal{C} &= \left\{\hat{\mu} \mid \hat{\mu} = \bar{X}_1 + D\phi, \quad 0 \le \phi(S_1^2, S_2^2, D^2) \le 1\right\}. & (5.23)
\end{aligned}$$

Clearly, $\mathcal{C}_o \subset \mathcal{C}_1 \subset \mathcal{C}$ and $\mathcal{C}_o \subset \mathcal{C}_2 \subset \mathcal{C}$. The classes $\mathcal{C}_o$ and $\mathcal{C}_2$ are equivariant under affine transformations whereas the estimators in $\mathcal{C}_1$ and $\mathcal{C}$ are equivariant under location transformations only. The following result is due to Sinha and Mouqadem (1982).

Proposition 5.7. (a) The estimator $\hat{\mu}_{GD}$ is admissible in $\mathcal{C}_o$ and $\mathcal{C}_2$.
(b) The estimator $\hat{\mu}_{GD}$ is extended admissible in $\mathcal{C}$ for $n \geq 5$, i.e., there does not exist any $\hat{\mu}$ such that $E(\hat{\mu} - \mu)^2 \leq V(\hat{\mu}_{GD}) - \epsilon$ for all σ_1^2, σ_2^2 and for any $\epsilon > 0$
(c) An estimator of the form

$$\hat{\mu} = \bar{X}_1 + D\left(\frac{S_1^2 + c_1}{S_1^2 + S_2^2 + c_1 + c_2}\right)$$

is admissible in $\mathcal{C}_1$ for any $c_1, c_2 > 0$.

Extended admissibility of $\hat{\mu}_{GD}$ in $\mathcal{C}$ is a strong indication of the true admissibility of $\hat{\mu}_{GD}$ in $\mathcal{C}$, although this is still open. Incidentally, any estimator $\hat{\mu} \in \mathcal{C}$ has variance given by

$$V(\hat{\mu}) = \frac{\sigma_1^2\sigma_2^2}{n(\sigma_1^2 + \sigma_2^2)} + E\left\{D^2\left(\phi - \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}\right)^2\right\}. \tag{5.24}$$

If we impose the condition that D is independent of ϕ, then $\hat{\mu} \in \mathcal{C}$ becomes an unbiased estimator of μ with variance

$$V(\hat{\mu}) = \frac{\sigma_1^2\sigma_2^2}{n(\sigma_1^2 + \sigma_2^2)} + \frac{(\sigma_1^2 + \sigma_2^2)}{n} E\left\{\left(\phi - \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}\right)^2\right\}. \tag{5.25}$$

Therefore, in this context performance of an unbiased estimator $\hat{\mu} \in \mathcal{C}$ can be judged by the performance of an estimator ϕ of $\sigma_1^2/(\sigma_1^2 + \sigma_2^2)$ which is a rather interesting observation. It is clear from the previous discussion that quite generally we can characterize the unbiased estimators of μ as

$$\hat{\mu}(h_1, h_2) = \bar{X}_1 + Dh_1(D)\phi\left(S_1^2, S_2^2, h_2(D)\right) \tag{5.26}$$

where $h_i(D), i = 1, 2$, are any two even functions. Variance of $\hat{\mu}(h_1, h_2)$ is given as

$$V(\hat{\mu}(h_1, h_2)) = \frac{\sigma_1^2\sigma_2^2}{n(\sigma_1^2 + \sigma_2^2)} + E\left[D^2\left\{h_1\phi - \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}\right\}^2\right]. \tag{5.27}$$

Even though the admissibility of $\hat{\mu}_{GD}$ seems a near certainty, it is inadmissible if we have some prior knowledge about the unknown variances. Consider the simple case of $k = 2$ and $n_1 = n_2 = n$. If it is found out (after data collection) that $\sigma_1^2 \leq \sigma_2^2$ (which can be checked through a suitable hypothesis testing) then one can construct a better estimator of μ as shown by Sinha (1979).

Proposition 5.8. Assume $\sigma_1^2 \leq \sigma_2^2$. For $k = 2$ and $n_1 = n_2 = n$, define

$$\hat{\mu}^* = \bar{X}_1 + (\bar{X}_2 - \bar{X}_1) \min\left\{\frac{1}{2}, \frac{S_1^2}{S_1^2 + S_2^2}\right\}.$$

Then (a) $\hat{\mu}^*$ is an unbiased estimator of μ; and (b) $V(\hat{\mu}^*) \leq V(\hat{\mu}_{GD}), \ \ \forall \sigma_1^2 \leq \sigma_2^2$.

For unequal sample sizes n_1 and n_2 one can have a similar result provided $\sigma_1^2/n_1 \leq \sigma_2^2/n_2$.

5.2 Asymptotic comparison of some estimates of common mean for $k = 2$

In this section we present some recent results due to Mitra and Sinha (2007) on an asymptotic comparison of some selected estimates of the common mean μ for $k = 2$ and $n_1 = n_2 = n$.

Let $\mathcal{C}_u$ be the general class of unbiased estimates of μ, defined as $\mathcal{C}_u = \{\hat{\mu}_\phi : \hat{\mu} = \bar{x} + D\phi(s_1^2, s_2^2, D^2)\}$ where $D = \bar{y} - \bar{x}$. Note that $E[D|D^2] = 0$ [Khuri, A.I., Mathew, T. and Sinha, B.K.(1998), Lemma 7.5.3, page 194-195], which implies that all estimates of μ in $\mathcal{C}_u$ are unbiased. We also consider a subclass of $\mathcal{C}_u$ defined as $\mathcal{C}_0 = \{\hat{\mu}_{\phi_0} : \hat{\mu} = \bar{x} + D\phi_0(s_1^2, s_2^2)\}$. Here we assume that both ϕ and ϕ_0 are smooth in the sense that they admit enough order derivatives with respect to their arguments.

Four popular estimates of μ in this context are given below.

$$\begin{aligned}
\hat{\mu}_1 &= \frac{\frac{\bar{x}}{s_1^2} + \frac{\bar{y}}{s_2^2}}{\frac{1}{s_1^2} + \frac{1}{s_2^2}} \\
\hat{\mu}_2 &= \bar{x} + D\frac{s_1^2 + D^2}{s_1^2 + s_2^2 + D^2} \text{ [Sinha-Mouqadem, 1982]} \\
\hat{\mu}_3 &= \bar{x} + D\min(0.5, \frac{s_1^2}{s_1^2 + s_2^2}) \text{ [Sinha, 1979]} \\
\hat{\mu}_4 &= \bar{x} + D\frac{s_1}{s_1 + s_2} \quad \text{[Sinha-Mouqadem, 1982]}.
\end{aligned}$$

Our comparison of the above estimates is essentially based on an expansion of their large sample variances in n^{-1}. In order for an estimate to be first order efficient (FOE), we expect the leading term of its variance (i.e., coefficient of n^{-1}) to be equal to the Rao-Cramer lower bound (Rao, 1973) which can be obtained by inverting the Fisher information matrix. The coefficient of n^{-2} in the large sample variance of an unbiased

estimate determines the nature of its second order efficiency (SOE). The following result is established in Mitra and Sinha (2007).

Theorem 5.1. In the class $\mathcal{C}_0$, $\hat{\mu}_1$ is unique FOE. In the extended class $\mathcal{C}_u$, $\hat{\mu}_1$ is FOE (though not unique) and the condition of FOE determines second order terms in the expansion of $var(\hat{\mu}_\phi)$.

As a byproduct of the proof of the above theorem, it is observed in Mitra and Sinha (2007) that the estimate $\hat{\mu}_4$ is not FOE. It is also proved there that the estimate $\hat{\mu}_3$ with $\phi = \min(0.5, \frac{s_1^2}{s_1^2+s_2^2})$, though lacks smoothness, is both FOE and SOE. Thus, its small sample dominance over the Graybill-Deal estimate, which holds whenever $\sigma_1^2 \leq \sigma_2^2$, is not really true in large samples.

We now discuss the Bayes estimation of the common mean μ under Jeffrey's noninformative prior [Berger (1980), page 87], $\pi(.)$, on the parameters $\theta = (\mu, \sigma_1^2, \sigma_2^2)$. Under this formulation, $\pi(\theta)$ is given by: $\pi(\theta) = \sqrt{det\, I(\theta)}$ where $I(\theta)$ is the Fisher information matrix.

Note that for a bivariate normal distribution,

$$I(\mu, \sigma_1^2, \sigma_2^2) = \begin{pmatrix} \frac{n(\sigma_1^2+\sigma_2^2)}{\sigma_1^2\sigma_2^2} & 0 & 0 \\ 0 & \frac{n}{2\sigma_1^4} & 0 \\ 0 & 0 & \frac{n}{2\sigma_2^4} \end{pmatrix}.$$

Hence, based on Fisher information matrix, such a prior is given by $p(\mu, \sigma_1^2, \sigma_2^2) \propto (\sqrt{\sigma_1^2 + \sigma_2^2})/(\sigma_1^2\sigma_2^2)^{\frac{3}{2}}$ where $-\infty < \mu < \infty$, σ_1^2, $\sigma_2^2 > 0$.

Combining this prior with the likelihood, and writing $\mu_0 = \left(\frac{\bar{x}}{\sigma_1^2} + \frac{\bar{y}}{\sigma_2^2}\right) / \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)$, the posterior distribution of the parameters $(\mu, \sigma_1^2, \sigma_2^2)$ is given by,

$$\begin{aligned} p(\mu, \sigma_1^2, \sigma_2^2|data) &\propto (\sigma_1^2\sigma_2^2)^{-\frac{n+3}{2}} \sqrt{\sigma_1^2 + \sigma_2^2}\, exp\left[-\frac{n(\bar{x}-\mu)^2}{2\sigma_1^2} - \frac{n(\bar{y}-\mu)^2}{2\sigma_2^2} - \frac{(n-1)s_1^2}{2\sigma_1^2} - \frac{(n-1)s_2^2}{2\sigma_2^2}\right] \\ &= (\sigma_1^2)^{-\frac{n+3}{2}} (\sigma_2^2)^{-\frac{n+3}{2}} \sqrt{\sigma_1^2 + \sigma_2^2}\, exp\left[-\frac{nD^2}{2(\sigma_1^2+\sigma_2^2)}\right] \\ &\cdot\ exp\left[-\frac{n}{2}(\frac{1}{(\sigma_1^2} + \frac{1}{\sigma_2^2})(\mu - \mu_0)^2\right] \\ &\cdot\ exp\left[-\frac{(n-1)s_1^2}{2\sigma_1^2} - \frac{(n-1)s_2^2}{2\sigma_2^2}\right] \end{aligned}$$

The joint posterior of $(\mu, \sigma_1^2, \sigma_2^2)$ can be viewed as:

1. Conditionally given (σ_1^2, σ_2^2), the posterior of μ is $N(\mu_0, \frac{\sigma_1^2\sigma_2^2}{n(\sigma_1^2+\sigma_2^2)})$

2. Joint marginal posterior of σ_1^2, σ_2^2 is given by
   $p(\sigma_1^2, \sigma_2^2|data) \propto (\sigma_1^2)^{-(\frac{n}{2}+1)}(\sigma_2^2)^{-(\frac{n}{2}+1)} exp\left[-\frac{nD^2}{2(\sigma_1^2+\sigma_2^2)} - \frac{(n-1)s_1^2}{\sigma_1^2} - \frac{(n-1)s_2^2}{\sigma_2^2}\right]$.

As a Bayes estimate of μ, we choose the posterior mean which is given by

$$\begin{aligned}\hat{\mu}_B &= E(\mu|data) \\ &= E[E(\mu|\sigma_1^2, \sigma_2^2, data)] \\ &= \bar{x}E[\frac{\theta}{1+\theta}|data] + \bar{y}E[\frac{1}{1+\theta}|data] \quad \text{where } \theta = \frac{\sigma_2^2}{\sigma_1^2}\,.\end{aligned}$$

Hence computation of $\hat{\mu}_B$ boils down to evaluating $E[\frac{1}{1+\theta}|data]$. To compute this term we need to find the posterior density of θ.

Upon making a transformation from $(\sigma_1^2, \sigma_2^2) \mapsto (\sigma_1^2, \theta)$ in (ii), we get the following.

$p(\sigma_1^2, \theta) \propto \theta^{-(\frac{n}{2}+1)}(\sigma_1^2)^{-(n+2)} exp\left[-\frac{nD^2}{1+\theta} - (n-1)s_1^2 - \frac{(n-1)s_2^2}{\theta}\right]$.

Now integrating the above expression with respect to σ_1^2 we get unnormalized posterior density of θ as

$$p(\theta|data) \propto \frac{\theta^{\frac{n}{2}}(\theta+1)^{n+1}}{(a\theta^2+b\theta+c)^{n+1}}$$

where $\theta > 0$ and $a = (n-1)s_1^2$, $b = (n-1)s_1^2 + (n-1)s_2^2 + nD^2$, $c = (n-1)s_2^2$.

This leads to

$$E[\frac{1}{1+\theta}|data] = \frac{\int_0^\infty \frac{\theta^{\frac{n}{2}}(\theta+1)^n}{(a\theta^2+b\theta+c)^{n+1}}\,d\theta}{\int_0^\infty \frac{\theta^{\frac{n}{2}}(\theta+1)^{n+1}}{(a\theta^2+b\theta+c)^{n+1}}\,d\theta} \tag{5.28}$$

The above integral is computed by using importance sampling method by choosing $g(\theta) = exp(-\theta)$ (Gelman et al. (2004)). It is obvious that the Bayes estimate of μ is unbiased. It is also proved in Mitra and Sinha (2007) that $\hat{\mu}_B$ is both FOE and SOE.

We end this section with a reference to Mitra and Sinha (2007) who reported the results of an extensive simulation study to compare bias and variance of five unbiased estimates of μ: $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\mu}_3$, $\hat{\mu}_4$, $\hat{\mu}_B$ for $n = 5, 10, 15$ and $\sigma_1^2 = 1$ and $\sigma_2^2 = 0.2(0.2)2$. Without any loss of generality, $\mu = 0$ is chosen for the simulation purpose. These simulation studies reveal that the Graybill-Deal estimate $\hat{\mu}_1$ and the Sinha-Mouqadem estimate $\hat{\mu}_2$ perform similarly and these two are better than the others. However, quite surprisingly, it turns out that the performance of the Bayes estimate is not satisfactory from the point of view of variance.

5.3 Exact and approximate confidence intervals for μ

In this section we address the problem of constructing exact and approximate confidence intervals for μ. Our discussion is based on combinations of relevant component t or F statistics and also Fisher's P-values, as discussed in Lecture 3. We also provide a comparison of various methods based on their expected lengths of confidence intervals for μ. It should be noted that tests for μ are not separately discussed here because of the well known connection between tests and confidence intervals.

The problem of constructing exact and approximate confidence intervals for the common mean μ of several normal populations with unequal and unknown variances arises in various contexts in statistical applications whenever two or more sources are involved with collecting data on the same basic characteristic of interest. We refer to Meier (1953), Eberhardt et al. (1989), and Skinner (1991) for some applications. However, although a lot of work has been done on point estimation of μ, as mentioned above, much less attention has been given to the problem of providing a meaningful confidence interval for μ. Several papers provide approximate confidence intervals for μ, centered at $\hat{\mu}_{GD}$, which are not quite useful because of the nature of underlying assumptions (see Meier, 1953; Eberhardt et al., 1989). In one particular context of interblock analysis of a balanced incomplete block design, similar approximate confidence intervals centered at some combined estimator are known (see Brown and Cohen, 1974).

Our review of the literature given below includes an old work by Fairweather (1972) and a relatively recent work by Jordan and Krishnamoorthy (1996), which are based on inverting weighted linear combinations of Student's t statistics and F statistics, respectively, which are used to test hypothesis about μ. However, determination of the exact cut-off points of these test statistics can be done only numerically, and it seems to us that the full thrust of *meta analysis* is not quite accomplished in these procedures. We mention below some exact confidence intervals for μ based on inverting exact tests for μ, which are constructed by combining the relevant P-values in a meaningful way (Yu, Sun and Sinha, 2002). We also provide a comparison among them on the basis of their expected lengths. An approximate confidence interval for μ based on an unbiased estimator of $var(\hat{\mu}_{GD})$ (see Sinha, 1985) is also given.

5.3.1 Approximate confidence intervals for μ

Using the Graybill-Deal estimate and its estimated variance given earlier, an approximate $100(1-\alpha)\%$ confidence interval for μ can be constructed on the basis of a suitable normalization of $\hat{\mu}_{GD}$, and can be expressed as $[\hat{\mu}_{GD} - z_{\alpha/2}\sqrt{\{\hat{v}ar(\hat{\mu}_{GD})\}}, \hat{\mu}_{GD} + z_{\alpha/2}\sqrt{\{\hat{v}ar(\hat{\mu}_{GD})\}}]$, where $z_{\alpha/2}$ is the standard normal upper $\alpha/2$ point. In practice, however, one can only use a first few terms from $\hat{v}ar(\hat{\mu}_{GD})$, depending on the sample sizes (see (5.17) and (5.18)). For better accuracy, one can use the fact that (Sinha, 1985) truncation

of $\hat{v}ar(\hat{\mu}_{GD})$ at $(m-1)$th term results in error not exceeding $n_{min}^{-(m+1)}$ (see **Proposition 5.6**).

5.3.2 Exact confidence intervals for μ

We now focus our attention to the construction of exact confidence intervals for μ. Since

$$t_i = \frac{\sqrt{n_i}(\bar{X}_i - \mu)}{s_i} \sim t_{n_i-1} \tag{5.29}$$

or, equivalently,

$$F_i = \frac{n_i(\bar{X}_i - \mu)^2}{s_i^2} \sim F_{1,n_i-1} \tag{5.30}$$

are standard test statistics for testing hypotheses about μ based on the ith sample, suitable linear combinations of $|t_i|$'s or F_i's or other functions thereof can be used as a *pivot* to construct exact confidence intervals for μ. This is precisely what is accomplished in Fairweather (1972), Cohen and Sackrowitz (1984), and Jordan and Krishnamoorthy (1996).

a) Confidence interval for μ based on t_i's

Cohen and Sackrowitz (1984) suggested to use $M_t = max_{1\le i\le k}\{|t_i|\}$ as a test statistic for testing hypotheses about μ. We can use M_t to construct a confidence interval for μ once the cut-off point of the distribution of M_t is known, which is independent of any parameter. Thus, if $c_{\alpha/2}$ satisfies the condition

$$\begin{aligned} 1-\alpha &= P[M_t \le c_{\alpha/2}] \\ &= \prod_{i=1}^{k} P[|t_i| \le c_{\alpha/2}], \end{aligned} \tag{5.31}$$

an exact confidence interval for μ with confidence level $1-\alpha$ is given by

$$\left[\max_{1\le i\le k}\left\{\bar{X}_i - \frac{c_{\alpha/2}s_i}{\sqrt{n_i}}\right\}, \min_{1\le i\le k}\left\{\bar{X}_i + \frac{c_{\alpha/2}s_i}{\sqrt{n_i}}\right\}\right]. \tag{5.32}$$

Determination of the cut-off point $c_{\alpha/2}$ is not easy in applications, and simulation may be necessary. An alternative approach is to use the confidence interval

$$\left[\max_{1\le i\le k}\left\{\bar{X}_i - \frac{c^{(i)}_{\alpha/2}s_i}{\sqrt{n_i}}\right\}, \min_{1\le i\le k}\left\{\bar{X}_i + \frac{c^{(i)}_{\alpha/2}s_i}{\sqrt{n_i}}\right\}\right] \tag{5.33}$$

where $c^{(i)}_{\alpha/2}$ satisfies $P[|t_i| \leq c^{(i)}_{\alpha/2}] = (1-\alpha)^{1/k}$. This latter interval clearly also has an exact coverage probability $1-\alpha$.

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

$$W_t = \sum_{i=1}^{k} u_i t_i, \;\; u_i = \frac{(var(t_i))^{-1}}{\sum_{j=1}^{k}(var(t_j))^{-1}} \tag{5.34}$$

which is also a *pivot*. If $b_{\alpha/2}$ denotes the cut-off point of the distribution of W_t, satisfying the equation

$$1-\alpha = P[|W_t| \leq b_{\alpha/2}], \tag{5.35}$$

then the confidence interval for μ is obtained as

$$\left[\frac{\sum_{i=1}^{k}\sqrt{n_i}u_i\bar{X}_i/s_i}{\sum_{i=1}^{k}\sqrt{n_i}u_i/s_i} - \frac{b}{\sum_{i=1}^{k}\sqrt{n_i}u_i/s_i}, \frac{\sum_{i=1}^{k}\sqrt{n_i}u_i\bar{X}_i/s_i}{\sum_{i=1}^{k}\sqrt{n_i}u_i/s_i} + \frac{b}{\sum_{i=1}^{k}\sqrt{n_i}u_i/s_i}\right] \tag{5.36}$$

It may be noted that

$$var(t_\nu) = \frac{\nu}{\nu-2}, \;\; \nu > 2. \tag{5.37}$$

b) Confidence interval for μ based on F_i's

Jordan and Krishnamoorthy (1996) suggested using a linear combination of the F_i's such as $W_f = \sum_{i=1}^{k} w_i F_i$ for positive weights w_i's, which is again a *pivot*. Hence, if we can compute $a_{\alpha/2}$ such that

$$P[W_f \leq a_{\alpha/2}] = 1-\alpha, \tag{5.38}$$

then, after simplification, an exact confidence interval for μ with confidence level $1-\alpha$ is given by

$$\left[LB = \sum_{i=1}^{k} p_i\bar{X}_i - \Delta, \;\; UB = \sum_{i=1}^{k} p_i\bar{X}_i + \Delta\right] \tag{5.39}$$

where

$$p_i = \frac{w_i n_i/s_i^2}{\sum_{j=1}^{k} w_j n_j/s_j^2} \tag{5.40}$$

and

$$\Delta^2 = \frac{a_{\alpha/2}}{\sum_{i=1}^{k} w_i n_i / s_i^2} - \{\sum_{i=1}^{k} p_i \bar{X}_i^2 - (\sum_{i=1}^{k} p_i \bar{X}_i)^2\}. \tag{5.41}$$

Jordan and Krishnamoorthy (1996) used w_i as inversely proportional to $var(F_i) = 2m_i^2(m_i - 1)/[(m_i - 2)^2(m_i - 4)]$ where $m_i = n_i - 1$, resulting in w_i as

$$w_i = \frac{[(m_i - 2)^2(m_i - 4)]/[m_i^2(m_i - 1)]}{\sum_{j=1}^{k}[(m_j - 2)^2(m_j - 4)]/[m_j^2(m_j - 1)]}. \tag{5.42}$$

Of course, it is assumed that $n_i > 5$ for all the k studies.

c) Confidence interval for μ based on P_i's

Since F_i, defined in (5.31), can be used for testing hypotheses about μ, we define the ith P value, P_i, as

$$P_i = \int_{F_i}^{\infty} h_i(x)dx \tag{5.43}$$

where $h_i(x)$ denotes the *pdf* of the F distribution with 1 and $(n_i - 1)$ *df*. Recalling the fact that $P_1, \cdots P_k$ are *iid* uniformly distributed random variables, we can combine them using any of the methods described earlier in Lecture 3. In particular, we use below Tippett's method, Fisher's method, inverse normal method and logit method.

(1) Tippett's method [Tippett (1931)]

As already explained, if $P_{[1]}$ is the minimum of $P_1, P_2, \cdots, P_k$, then Tippett's method rejects the hypothesis about μ if $P_{[1]} < c_1 = 1 - (1 - \alpha)^{1/k}$. By inverting this rejection region, we have a confidence interval for μ with confidence coefficient $1 - \alpha$, given by

$$\begin{aligned} C.I. &= \{\mu : P_{[1]} \geq c_1\} \\ &= \{\mu : P_i \geq c_1, i = 1, \cdots, k\} \\ &= \{\mu : \int_{n_i(\bar{x}_i - \mu)^2/s_i^2}^{\infty} f_i(x)dx \geq 1 - (1 - \alpha)^{1/k}, i = 1, \cdots, k\}. \end{aligned} \tag{5.44}$$

(2) Fisher's method [Fisher (1932)]

Since Fisher's method rejects hypotheses about μ when $-2\sum_{i=1}^{k} lnP_i > \chi^2_{2k,\alpha}$, the confidence interval for μ obtained by inverting the acceptance region of this test is given by

$$
\begin{aligned}
C.I. &= \{\mu : -2\sum_{i=1}^{k} \log P_i \leq \chi^2_{2k,\alpha}\} \qquad (5.45)\\
&= \{\mu : \prod_{i=1}^{k} P_i \geq e^{-2\chi^2_{2k,\alpha}}\}\\
&= \{\mu : \prod_{i=1}^{k} \int_{n_i(\bar{x}_i-\mu)^2/s_i^2}^{\infty} f_i(x)dx \geq e^{-2\chi^2_{2k,\alpha}}\}
\end{aligned}
$$

(3) Inverse normal method [Stouffer et al., 1949]

Since this method rejects hypotheses about μ when $\frac{\sum_{i=1}^{k}\Phi^{-1}(P_i)}{\sqrt{k}} < -z_\alpha$ at level α, the $(1-\alpha)$ level confidence interval for μ obtained by inverting this acceptance region is given by

$$
C.I. = \{\mu : \frac{\sum_{i=1}^{k}\Phi^{-1}(P_i)}{\sqrt{k}} \geq -z_\alpha\}. \qquad (5.46)
$$

(4) Logit method [George, 1977]

This method rejects H_0 if $\sum_{i=1}^{k}\log(\frac{P_i}{1-P_i}) < c$ where c is a predetermined constant. It was mentioned earlier that the distribution of

$$
G^* = [-\sum_{i=1}^{k}\log(\frac{P_i}{1-P_i})][\frac{3}{k\pi^2}]^{1/2} \qquad (5.47)
$$

can be approximated by a standard normal distribution (see Lecture 3). Therefore a $(1-\alpha)$ level confidence interval for μ can be obtained from

$$
C.I. = \{\mu : G^* < z_\alpha\}. \qquad (5.48)
$$

It is an interesting *research problem* to settle if the confidence regions for μ obtained from the above four methods are actually genuine intervals. The appendix at the end of this section makes an attempt to establish the same for Fisher's method on the basis of an *expansion* technique.

5.4 Two examples

In this section we provide two examples to illustrate the methods described above.

Example 5.1. Here we examine the data reported in Meier (1953) and analyzed in Jordan and Krishnamoorthy (1996) about the percentage of albumin in plasma protein in human subjects. We would like to combine the results of four experiments in order to construct a confidence interval for the common mean μ. The data appear in Table 5.1.

Table 5.1. Percentage of albumin in plasma protein

Experiment	n_i	Mean	Variance
A	12	62.3	12.986
B	15	60.3	7.840
C	7	59.5	33.433
D	16	61.5	18.513

We have applied all the techniques described in this section, and computed the two-sided confidence intervals with $\alpha = 0.05$. These are given below. It is rather interesting to observe that most of the confidence intervals are centered at around the same value, and the one based on F turns out to be the best in the sense of having the smallest observed length.

Table 5.2. Interval estimates for μ

Intervals	Critical values	Weights	Interval
C & S	c=3.043		60.82 ± 1.68
t	$c_i^{'}$ s=2.9702,2.8543, 3.5055,2.8272.		60.78 ± 1.58
F	b ≐ 1.102	$\mu_i^{'}$s=0.2550,0.2671, 0.2708, 0.2701.	61.04± 1.15
J & K	a ≐ 3.191	$p_i^{'}$s =0.2100,0.5245, 0.0181,0.2474.	61.00 ± 1.44
Fisher		$b_i^{'}$s =1.0190,1.0756, 0.8210, 1.0898.	60.9992± 1.4245
		$p_i^{'}$s =0.2289,0.5003, 0.0418,0.2290.	
Normal		$b_i^{'}$s =0.2862,0.2987, 0.2410,0.3019.	60.9986 ± 1.3147
		$p_i^{'}$s =0.2305,0.4982, 0.0440,0.2274.	
Logit		$b_i^{'}$s =0.6678,0.6996, 0.5546,0.7076.	60.9988 ± 1.3478
		$p_i^{'}$s =0.2300,0.4988, 0.0433,0.2279.	

Example 5.2. This is quoted from Eberhardt et al. (1989) and deals with the problem of estimation of mean *Selenium* in non-fat milk powder by combining the results of four methods. Data appear in the table below.

Table 5.3. Selenium in non-fat milk powder

Methods	n_i	Mean	Variance
Atomic absorption spectrometry	8	105.0	85.711
Neutron activation:			
1). Instrumental	12	109.75	20.748
2). Radiochemical	14	109.5	2.729
Isotope dilution mass spectrometry	8	113.25	33.640

Here again we have applied all the techniques described in this section, and computed the two-sided confidence intervals for the common mean μ with $\alpha = 0.05$. These are given below. It is rather interesting to observe that most of the confidence intervals are centered at around the same value, namely, 109.5, and the one based on the normal method turns out to be the best in the sense of having the smallest observed length.

Table 5.4. Interval estimates for μ

Intervals	Critical values	Weights	Interval
C & S	c=3.128		109.5 ± 1.38
t	$c_i^{'}$ s=3.321,2.970,		109.5 ± 1.27
	2.886,3.321.		
F	b $\doteq$ 1.118	$\mu_i^{'}$s=0.2309,0.2645,	109.7± 1.11
		0.2736, 0.2309.	
J & K	a $\doteq$ 3.341	$p_i^{'}$s =0.0068,0.0777,	109.6 ± 1.08
		0.8908,0.0247.	
Fisher		$b_i^{'}$s =0.8795,1.0190,	109.5890± 1.0876
		=1.0594,0.8795.	
		$p_i^{'}$s =0.0130,0.0933,	
		0.8606,0.0331.	
Normal		$b_i^{'}$s =0.2546,0.2862,	109.5915 ± 0.9269
		0.2952,0.2546.	
		$p_i^{'}$s =0.0135,0.0938,	
		0.8584,0.0343.	
Logit		$b_i^{'}$s =0.5884,0.6678,	109.5907 ± 1.2526
		0.6905,0.5884.	
		$p_i^{'}$s =0.0133,0.0937,	
		0.8591,0.0339.	

5.5 Appendix: Theory of Fisher's Method

Let $T_n(\mu) = -2\sum_{i=1}^{k} \ln P_i(\mu) \sim \chi^2(2k)$ under H_0, where $P_i(\mu)$ is the P value defined by

$$P_i(\mu) = P(F_{1,n_i-1} > c_i(\mu)) \tag{5.49}$$

with

$$c_i(\mu) = \frac{n_i(\bar{x}_i - \mu)^2}{s_i^2}. \tag{5.50}$$

Hence, the $100(1-\alpha)\%$ confidence interval for μ is

$$\{\mu : T_n(\mu) \leq \chi^2_\alpha(2k)\}. \tag{5.51}$$

Now we approximate T_n by $\tilde{T}_n$ which is

$$\tilde{T}_n(\mu) = T_n(\hat{\mu}) + \sum_{i=1}^{k} b_i(c_i - \hat{c}_i) \tag{5.52}$$

where

$$\hat{\mu} = \frac{\sum_{i=1}^{k} n_i\bar{x}_i/s_i^2}{\sum_{i=1}^{k} n_i/s_i^2} \quad \text{the Graybill-Deal estimator,} \tag{5.53}$$

$$\hat{c}_i = c_i(\hat{\mu}), \tag{5.54}$$

and b_i is chosen such that $T_n(\mu) \approx \tilde{T}_n(\mu)$.
Suppose there exists an μ_0 such that

$$c_i \equiv c_i^* = F_{\exp(-\frac{\chi^2_\alpha(2k)}{2k})}(1, n_i - 1) \tag{5.55}$$

and define $\varepsilon(\mu) = T_n(\mu) - \tilde{T}_n(\mu)$. Then, we have,

$$\begin{aligned} \varepsilon(\mu) &= T_n(\mu) - \tilde{T}_n(\mu) - \sum_{i=1}^{k} b_i(c_i - \hat{c}_i) \\ &\approx T_n(\mu)|_{c_i=c_i^*} - T_n(\hat{\mu}) - \sum_{i=1}^{k} b_i(c_i^* - \hat{c}_i) + \sum_{i=1}^{k}(\frac{d\,T_n}{d\,c_i} - b_i)|_{c_i=c_i^*}(c_i - c_i^*). \end{aligned} \tag{5.56}$$

If we put

$$b_i = \frac{d\,T_n}{d\,c_i}|_{c_i=c_i^*} = -2\frac{P_i'(c_i)}{P_i(c_i)}|_{c_i=c_i^*}, \tag{5.57}$$

then

$$\varepsilon(\hat{\mu}) = 0 \quad \text{and} \quad \varepsilon(\mu|_{c_i=c_i^*}) \approx 0 \ \ \text{(1st order)}. \tag{5.58}$$

Since

$$\begin{aligned} P_i'(c_i) &= \frac{d}{d\,c_i} P(F_{1,n_i-1} > c_i) \\ &= \frac{d}{d\,c_i} \int_{c_i}^{\infty} \frac{(\frac{1}{n_i-1})^{\frac{1}{2}}}{Beta(\frac{1}{2}, \frac{n_i-1}{2})} \cdot \frac{u^{-\frac{1}{2}}}{(1+\frac{u}{n_i-1})^{\frac{n_i}{2}}}\, d\,u \\ &= -\frac{\Gamma(\frac{n_i}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{n_i-1}{2})} \cdot (\frac{1}{n_i-1})^{\frac{1}{2}} \cdot c_i^{-\frac{1}{2}} \cdot (1+\frac{c_i}{n_i-1})^{-\frac{n_i}{2}} \end{aligned} \tag{5.59}$$

and

$$P_i(c_i^*) = P(F_{1,n_i-1} > c_i^*) = \exp(-\frac{\chi_\alpha^2(2k)}{2k}). \tag{5.60}$$

Therefore,

$$b_i = \frac{2\Gamma(\frac{n_i}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{n_i-1}{2})} (\frac{1}{n_i-1})^{\frac{1}{2}} \cdot c_i^{-\frac{1}{2}} \cdot (1+\frac{c_i}{n_i-1})^{-\frac{n_i}{2}} \cdot \exp(-\frac{\chi_\alpha^2(2k)}{2k}). \tag{5.61}$$

Hence,

$$\begin{aligned} & T_n(\mu) && \le \chi_\alpha^2(2k) \\ \Longrightarrow \quad & \tilde{T}_n(\mu) && \le \chi_\alpha^2(2k) \\ \Longrightarrow \quad & \textstyle\sum_{i=1}^k b_i c_i && \le \chi_\alpha^2(2k) - T_n(\hat{\mu}) + \sum_{i=1}^k b_i \hat{c}_i \equiv a, \text{says.} \end{aligned} \tag{5.62}$$

As a result, the $100(1-\alpha)\%$ confidence interval for μ is

$$\mu \in \sum_{i=1}^k q_i \bar{x}_i \pm \left[\frac{a}{\sum_{i=1}^k b_i n_i / s_i^2} + \left(\sum_{i=1}^k q_i \bar{x}_i \right)^2 - \sum_{i=1}^k q_i \bar{x}_i^2 \right]^{\frac{1}{2}}, \tag{5.63}$$

where

$$q_i = \frac{b_i n_i / s_i^2}{\sum_{j=1}^k b_j n_j / s_j^2}. \tag{5.64}$$

6 Tests of Homogeneity in Meta Analysis

As has been mentioned earlier, meta analysis of results from different experiments or studies is quite common these days. However, as has been emphasized, it is equally important to make sure that the underlying *effect sizes* are indeed homogeneous before performing any meta analysis or pooling of evidence or data so that an inference on a common effect makes sense.

In this lecture we discuss at length the problem of testing homogeneity of means in a one-way fixed effects model. We assume throughout that the observations are drawn from k independent univariate normal populations with means μ_1, ..., μ_k and variances σ_1^2, σ_2^2,..., σ_k^2, and the problem is to test the homogeneity hypothesis with respect to means, given by $H_0 : \mu_1 = \cdots = \mu_k$ against a general alternative. Once H_0 is accepted, we feel quite comfortable in pooling all the data sets in order to make suitable inference about the common unknown mean μ. Later on we will discuss the dual problem of testing homogeneity of means in a one-way random effects model which indeed also has a long and rich history.

The problem of testing the homogeneity of means in a one-way ANOVA is one of the oldest problems in statistics with applications in many diverse fields (Cochran, 1937). Under the classical ANOVA assumption of normality, independence and homogeneous error variances ($\sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$), one uses the standard likelihood ratio F-test which is also known to be the optimum from an invariance point of view. However, when one or more of these basic assumptions are violated, the F-test ceases to be any good, let alone be optimum! This is especially true in the case of non-homogeneous error variances which is often the situation in meta analysis. In the literature (Cochran, 1937; Welch, 1951), several tests of H_0 have been proposed and compared in the presence of heterogeneity of error variances. All these tests are approximate and work quite well in large samples. The main goal of this lecture, mostly based on Hartung, Argac and Makambi (2002), is to present a systematic development of these tests along with results of some simulation studies to compare them. An exact solution based on a relatively new notion of generalized P-values will also be presented. However, a complete understanding of this solution requires a good notion of generalized P-values. It should be noted that, for $k = 2$, the testing problem under consideration boils down to the famous Behrens-Fisher problem!

6.1 Model and Test Statistics

Let X_{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

$$X_{ij} = \mu_i + e_{ij} = \mu + \tau_i + e_{ij}; \quad i = 1, \ldots, k,\ j = 1, \ldots, n_i. \tag{6.1}$$

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

$$E(e_{ij}) = 0, \quad \text{Var}(e_{ij}) = \sigma_i^2, i = 1, \ldots, k, j = 1, \ldots, n_i. \tag{6.2}$$

Under the above set up, we are interested in testing the hypothesis $H_0 : \mu_1 = \cdots = \mu_k$. To test this hypothesis, we propose the following test statistics.

a) **ANOVA F Test**
S_{an}, given by

$$S_{an} = \frac{N-k}{k-1} \cdot \frac{\sum\limits_{i=1}^{k} n_i(\bar{X}_{i.} - \bar{X}_{..})^2}{\sum\limits_{i=1}^{k}(n_i - 1)S_i^2}, \tag{6.3}$$

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

This test was originally meant to test for equality of population means under variance homogeneity, and has an F distribution with $k-1$ and $N-k$ degrees of freedom under the null hypothesis. The test rejects H_0 at level α if $S_{an} > F_{k-1,N-k;1-\alpha}$.

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).

b) **Cochran's Test**
This test suggested by Cochran in 1937 is based on

$$S_{ch} = \sum_{i=1}^{k} w_i \left(\bar{X}_{i.} - \sum_{j=1}^{k} h_j\, \bar{X}_{j.}\right)^2, \tag{6.4}$$

where $w_i = n_i/S_i^2$, $h_i = w_i/\sum_{i=1}^{k} w_i$. Under H_0, the Cochran statistic is distributed approximately as a χ^2-variable with $k-1$ degrees of freedom. The test rejects H_0 at level α if $S_{ch} > \chi^2_{k-1;1-\alpha}$. Cochran's test is often used as the standard test for testing homogeneity in meta analysis. This test has been already introduced in Chapter 4, see (4.10), as the general large sample test of homogeneity.

c) **Welch Test**

The Welch test is given

$$S_{we} = \frac{\sum_{i=1}^{k} w_i \Big(\bar{X}_{i.} - \sum_{j=1}^{k} h_j \, \bar{X}_{j.}\Big)^2}{(k-1) + 2\,\frac{k-2}{k+1} \sum_{i=1}^{k} \frac{1}{n_i - 1} \, (1 - h_i)^2}, \tag{6.5}$$

where $w_i = n_i / S_i^2, h_i = w_i / \sum_{i=1}^{k} w_i$, is an extension of testing the equality of two means to more than two means (see Welch, 1951) in the presence of variance heterogeneity within populations. The Welch test is a modification of Cochran's test. Under H_0, the statistic S_{we} has an approximate F distribution with $k-1$ and ν_g degrees of freedom, where

$$\nu_g = \frac{(k^2 - 1)/3}{\sum_{i=1}^{k} \frac{1}{n_i - 1} \, (1 - h_i)^2}. \tag{6.6}$$

This test rejects H_0 at level α if $S_{we} > F_{k-1,\nu_g;1-\alpha}$.

d) **Brown-Forsythe (B-F) Test**

This test, also known as the modified F test, is based on

$$S_{b-f} = \frac{\sum_{i=1}^{k} n_i \, (\bar{X}_{i.} - \bar{X}_{..})^2}{\sum_{i=1}^{k} (1 - n_i/N) \, S_i^2}. \tag{6.7}$$

When H_0 is true, S_{b-f} is distributed approximately as an F variable with $k-1$ and ν degrees of freedom where

$$\nu = \frac{\Big(\sum_{i=1}^{k} (1 - n_i/N) \, S_i^2\Big)^2}{\sum_{i=1}^{k} (1 - n_i/N)^2 \, S_i^4/(n_i - 1)}. \tag{6.8}$$

The test rejects H_0 at level α if $S_{b-f} > F_{k-1,\nu;1-\alpha}$. Using a simulation study, Brown and Forsythe (1974) demonstrated that their statistic is robust under heterogeneity of variances. If the population variances are close to being homogenous, the B-F test is closer to the ANOVA F-test than Welch' test.

e) **Mehrotra (Modified Brown-Forsythe) Test**

The test statistic

$$S_{b-f(m)} = \frac{\sum_{i=1}^{k} n_i(\bar{X}_{i.} - \bar{X}_{..})^2}{\sum_{i=1}^{k}(1 - n_i/N)S_i^2} \tag{6.9}$$

was proposed by Mehrotra (1997) in an attempt to correct a "flaw" in the B-F-test. Under $H_0, S_{b-f(m)}$ is distributed approximately as an F variable with ν_1 and ν degrees of freedom where

$$\nu_1 = \frac{(\sum_{i=1}^{k}(1 - n_i/N)S_i^2)^2}{\sum_{i=1}^{k} S_i^4 + \Big(\sum_{i=1}^{k} n_i\, S_i^2/N\Big)^2 - 2\,\sum_{i=1}^{k} n_i\, S_i^4/N} \tag{6.10}$$

and ν is defined in B-F test. The test rejects H_0 at level α if $S_{b-f(m)} > F_{\nu_1,\nu;1-\alpha}$.

f) **Approximate ANOVA F Test**

The test statistic

$$S_{aF} = \frac{N-k}{k-1}\,\frac{\sum_{i=1}^{k} n_i\,(\bar{X}_{i.} - \bar{X}_{..})^2}{\sum_{i=1}^{k}(n_i - 1)\,S_i^2}, \tag{6.11}$$

was proposed by Asiribo and Gurland (1990). UnderH_0, the statistic S_{aF} is distributed approximately as an F-variable with ν_1 and ν_2 degrees of freedom where ν_1 is defined under Mehrotra test above and

$$\nu_2 = \frac{\Big(\sum_{i=1}^{k}(n_i - 1)\,S_i^2\Big)^2}{\sum_{i=1}^{k}(n_i - 1)\,S_i^4}. \tag{6.12}$$

The test rejects H_0 at level α if $S_{aF} > \widehat{c}\cdot F_{\nu_1,\nu_2;1-\alpha}$, where

$$\widehat{c} = \frac{N-k}{N(k-1)}\,\frac{\sum_{i=1}^{k}(N - n_i)\,S_i^2}{\sum_{i=1}^{k}(n_i - 1)\,S_i^2}. \tag{6.13}$$

We notice that the numerator degrees of freedom for S_{aF} and $S_{b-f(m)}$ are equal. Further, for $n_i = n, i = 1, \ldots, k$, that is, for balanced data, the test statistic and the degrees of freedom for both the numerator and denominator of these two statistics are also equal.

g) **Adjusted Welch Test**

The Welch test uses weights $w_i = n_i/s_i^2$. We know that

$$E(w_i) = E\left(\frac{n_i}{S_i^2}\right) = c_i \cdot \frac{n_i}{\sigma_i^2}, \tag{6.14}$$

where $c_i = (n_i - 1)/(n_i - 3)$. Therefore, an unbiased estimator of n_i/σ_i^2 is $n_i/(c_i S_i^2)$. Defining $w_i^* = n_i/(c_i S_i^2)$, Hartung, Argac, and Makambi (2002) propose a test they called adjusted Welch test, denoted by $S_a w$, which is given by

$$S_{aw} = \frac{\sum_{i=1}^{k} w_i^* \, (\bar{X}_{i\cdot} - \sum_{j=1}^{k} h_j^* \bar{X}_{j\cdot})^2}{\left((k-1) + 2\,\frac{k-2}{k+1} \sum_{i=1}^{k} \frac{1}{n_i - 1}\,(1 - h_i^*)\right)^2}, \tag{6.15}$$

where $h_i^* = w_i^* / \sum_{j=1}^{k} w_j^*, i = 1, \ldots, k$.

Under H_0, the adjusted Welch statistic, S_{aw}, is distributed approximately as an F-variable with $k - 1$ and ν_g^* degrees of freedom, with

$$\nu_g^* = \frac{(k^2 - 1)/3}{\sum_{i=1}^{k} \frac{1}{n_i - 1}\,(1 - h_i^*)^2}. \tag{6.16}$$

The test rejects H_0 at level α if $S_{aw} > F_{k-1,\nu_g^*;1-\alpha}$. When the sample sizes are large, S_{aw} approaches the Welch test. With small sample sizes, this statistic will help to correct the overshooting of the Welch test with respect to α.

Extensive simulation studies by Hartung, Argac, and Makambi (2002) for both size and power under normal and nonnormal populations, under homogeneous and heterogeneous variances, and under balanced and unbalanced schemes reveal that the modified Brown-Forsythe test and the approximate F test are relatively least affected by changes from normal populations with homogeneous variances.

6.2 An Exact Test of Homogeneity

We conclude this section with a brief discussion of the application of the generalized P-value for solving the underlying testing problem of homogeneity of means in presence of heterogeneity variances. This procedure described below will produce an exact test. For details, we refer to Tsui and Weerahandi (1989), Thursby (1992), and Griffiths and Judge (1992).

We first discuss the case $k = 2$, i.e., the Behrens-Fisher problem. Let $X = (\bar{X}_{1.} - \bar{X}_{2.}, S_1^2, S_2^2)$, $x = (\bar{x}_{1.} - \bar{x}_{2.}, s_1^2, s_2^2)$, $\theta = \mu_1 - \mu_2$ and $\eta = (\sigma_1^2, \sigma_2^2)$. Here S_1^2 and S_2^2 are the two sample variances, which are unbiased estimates of σ_1^2 and σ_2^2, respectively. We then define $T(X; x, \theta, \eta)$ as

$$T(X; x, \theta, \eta) = (\bar{X}_{1.} - \bar{X}_{2.}) \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right)^{-1/2} \left(\frac{s_1^2\, \sigma_1^2}{S_1^2\, n_1} + \frac{s_2^2\, \sigma_2^2}{S_2^2\, n_2} \right)^{1/2}. \tag{6.17}$$

Note that the *observed* value of T is $t = \bar{x}_{1.} - \bar{x}_{2.}$, and that $E(T)$ increases with $\mu_1 - \mu_2$. Hence the generalized P-value can be defined as

$$gen.P = P[T \geq \bar{x}_{1.} - \bar{x}_{2.} \,|\, \mu_1 = \mu_2] = P\left[Z \left(\frac{s_1^2}{U_1\, n_1} + \frac{s_2^2}{U_2\, n_2} \right)^{1/2} \geq \bar{x}_{1.} - \bar{x}_{2.} \right] \tag{6.18}$$

where Z is standard normal, $U_1 \sim \chi^2_{n_1 - 1}$, $U_2 \sim \chi^2_{n_2 - 1}$, and all three are independent. The null hypothesis of equality of two normal means is rejected when the generalized P-value is small.

For $k > 2$, we proceed by defining $a_i = n_i/\sigma_i^2$, $b_i = S_i^2\, n_i / s_i^2\, \sigma_i^2$, and

$$S_0^2 = \sum_{i=1}^{k} a_i \left[\bar{X}_{i.} - \frac{\sum_{i=1}^{k} a_i\, \bar{X}_{i.}}{\sum_{i=1}^{k} a_i} \right]^2 \tag{6.19}$$

$$\tilde{S}_0^2 = \sum_{i=1}^{k} b_i \left[\bar{X}_{i.} - \frac{\sum_{i=1}^{k} b_i\, \bar{X}_{i.}}{\sum_{i=1}^{k} b_i} \right]^2. \tag{6.20}$$

Then, obviously under the null hypothesis of equal means, S_0^2 has a central χ^2 distribution with $k - 1$ df, and will tend to be large under the alternative hypothesis. We now define the test variable as $T = S_0^2/\tilde{S}_0^2$ and, noting that the observed value of T is one, we compute the generalized P-value as $P = P[T > 1|H_0]$. Of course, the computation of the P-value here and in all such problems is carried out, for fixed x, often by simulation.

It is evident from the discussion in this lecture that there are many tests for comparing normal means with unequal within study variances. Many tests have also been compared

in terms of size in Thursby (1992) and Gamage and Weerahandi (1998). We also refer to Weerahandi (1995), Khuri, Mathew and Sinha (1998), and Ananda and Weerahandi (1997) for a discussion on generalized P-values and their applications.

6.3 An Application

We conclude this section with an example from Weerahandi (1995) where the goal is to compare four means of corn yields by four hybrids: A, B, C, D. The data and the standard fixed effects ANOVA table are given below.

Table 6.1. Yield of corn from four hybrids: data, means and standard deviations

Population	Data	$\bar{x}_{i.}$	s_i
Hybrid A	7.4, 6.6, 6.7, 6.1, 6.5, 7.2	6.750	0.435
Hybrid B	7.1, 7.3, 6.8, 6.9, 7.0	7.020	0.172
Hybrid C	6.8, 6.3, 6.4, 6.7, 6.5, 6.8	6.583	0.195
Hybrid D	6.4, 6.9, 7.6, 6.8, 7.3	7.000	0.415

Table 6.2. Anova table for comparing the four hybrids

Source of variation	df	Sum of Squares	Mean Sum of Squares	F-statistic
Between	3	0.728	0.2427	1.841
Error	18	2.372	0.1318	
Total	21	3.1		

The usual P-value based on the assumption of equal population within hybrid variances (F-statistic = 1.841) is 0.176, thus leading to acceptance of the null hypothesis of equal means. It is however clear from the values of the sample standard deviations that the assumption of equal population variances may not be tenable for this data set. The refined approximate test statistics lead to different conclusions in this example. The Brown-Forsythe test and its derivatives yield P-values in the order of magnitude as the usual F-test. Cochran's test produces a highly significant result. The P-value of the Welch test and its adjusted version is 0.045 leading to rejection of the homogeneity hypothesis at level $\alpha = 0.05$. An application of the generalized P-value as explained above yields 0.048, leading to marginal significance as the Welch test. The results of the various test procedures are summarized in Table 6.3.

Table 6.3. Test statistics and P-values

Test	Value of test statistic	P-value
ANOVA F-Test	1.840	0.176
Cochran	13.638	0.003
Welch	3.980	0.045
Brown-Forsythe	1.851	0.191
Mehrotra	1.851	0.179
approximate ANOVA F-Test	1.851	0.178
adjusted Welch test	2.180	0.045
generalized P-value		0.048

We also present below the results of three additional examples.

Example 6.2. Here we examine the data reported in Meier (1953) about the percentage of albumin in plasma protein in human subjects.

Table 6.4. Percentage of albumin in plasma protein

Experiment	n_i	Mean	Variance s_i^2
A	12	62.3	12.986
B	15	60.3	7.840
C	7	59.5	33.433
D	16	61.5	18.513

Test statistics and P-values

Test	Value of test statistic	P-value
ANOVA F-Test	0.991	0.405
Cochran	3.186	0.364
Welch	0.993	0.417
Brown-Forsythe	0.833	0.491
Mehrotra	0.833	0.522
approximate ANOVA F-Test	0.833	0.516
adjusted Welch test	0.804	0.418
generalized P-value		

Example 6.3. Here we examine the data on selenium in non-fat milk powder.

Table 6.5. Selenium in non-fat milk powder

Methods	n_i	Mean	Variance s_i^2
Atomic absorption spectrometry	8	105.0	85.711
Neutron activation:			
1). Instrumental	12	109.75	20.748
2). Radiochemical	14	109.5	2.729
Isotope dilution mass spectrometry	8	113.25	33.640

Test statistics and P-values

Test	Value of test statistic	P-value
ANOVA F-Test	3.169	0.035
Cochran	5.208	0.157
Welch	1.589	0.235
Brown-Forsythe	2.428	0.104
Mehrotra	2.428	0.107
approximate ANOVA F-Test	2.428	0.100
adjusted Welch test	1.137	0.236
generalized P-value		

Example 6.4. Here we examine the data reported in Weerahandi (Generalized Inference in Repeated Measures, page 43).

Table 6.6. Strength of four brands of reinforcing bars

Brand A	21.4, 13.5, 21.1, 13.3, 18.9, 19.2, 18.3
Brand B	27.3, 22.3, 16.9, 11.3, 26.3, 19.8, 16.2, 25.4
Brand C	18.7, 19.1, 16.4, 15.9, 18.7, 20.1, 17.8
Brand D	19.9, 19.3, 18.7, 20.3, 22.8, 20.8, 20.9, 23.6, 21.2

Test statistics and P-values

Test	Value of test statistic	P-value
ANOVA F-Test	1.608	0.211
Cochran	14.439	0.002
Welch	4.385	0.023
Brown-Forsythe	1.616	0.232
Mehrotra	1.616	0.231
approximate ANOVA F-Test	1.616	0.233
adjusted Welch test	3.086	0.023
generalized P-value		

7 One-Way Random Effects Model

7.1 Introduction

As discussed in the previous chapter, tests for homogeneity of means or in general effect sizes are crucial before performing any meta analysis or pooling of data. When tests for homogeneity lead to acceptance of the null hypothesis, thus supporting the evidence that the underlying population means or effect sizes can be believed to be the same, one feels quite comfortable in carrying out the meta analysis in order to draw appropriate inference about the common mean or effect size. When, however, the tests lead to rejection of the null hypothesis of homogeneity of means, it is not proper to do meta analysis of data unless we find out reasons for heterogeneity and make an attempt to explain them. The lack of homogeneity could be due to several covariates which might behave differently for different studies or simply because the means themselves might arise from a so called super population, thus leading to their variability and apparent differences. In this chapter we discuss at length the latter formulation which is often known as the one-way random effects model.

There is a vast literature on the topic of one-way random effects model with its root in meta analysis. Under this model with normality assumption, the treatment means μ_1,...., μ_k corresponding to k different studies or experiments are modelled as arising from a super normal population with an overall mean μ and an overall variability σ_a^2. The parameters of interest are then the overall mean μ and the inter-study variability σ_a^2 in terms of their estimation, tests and confidence intervals.

In the remainder of this chapter we discuss many results pertaining to the above problems. Recalling that in the context of meta analysis ANOVA models, existence of heterogeneous within study variances (also known as error variances) is very much a possibility, we consider the two cases of homogeneous and heterogeneous error variances separately in sections 7.2 and 7.3, respectively. It turns out that, as expected, statistical inference about the parameters of interest under the homogeneous error structure can be carried out much more easily compared to that under a heterogenous error structure. It is also true that the analysis of a balanced model is much easier than the analysis of an unbalanced model. Recall that a balanced model refers to the case when we have an equal number of observations or replications from all the populations.

As will be clear from what follows, this particular topic of research has drawn the attention of many statisticians from all over the world, and has prompted the emergence of new statistical methods. Most notably among them is the method based on generalized P-values which itself has a considerable amount of literature including a few text books. We will mention in the sequel some results based on the notion of generalized P-values. For details on generalized P-values, we refer to Khuri, Mathew and Sinha (1998) and

Weerahandi (1995).

We end this section with a simple description of the model to be analyzed. We consider the case of the one–way random effects model of ANOVA, i. e.

$$y_{ij} = \mu + a_i + e_{ij}\,, \qquad i = 1, \ldots, k,\; j = 1, \ldots, n_i \geq 1\,, \tag{7.1}$$

where y_{ij} denotes the observable variable, μ the fixed, but unknown grand mean, a_i the unobservable random effect with mean 0 and variance σ_a^2, and e_{ij} the error term with mean 0 and variance σ_i^2. We assume that the random variables $a_1, \ldots, a_k, e_{11}, \ldots e_{kn_k}$ are normally distributed and mutually stochastically independent. Furthermore, we denote by $N = \sum_{i=1}^{k} n_i$, the total number of observations.

The basic statistics for the above model are the sample means $\bar{y}_{i.}$ and sample sum of squares S_i^2, defined by

$$\bar{y}_{i.} = \sum_{j=1}^{n_i} y_{ij}/n_i,\;\; S_i^2 = \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2,\;\; i = 1, \cdots, k. \tag{7.2}$$

Then the overall or grand mean and the two well known sums of squares, namely, between sum of squares (BSS) and within sum of squares (WSS) are defined as

$$\bar{y}_{..} = \sum_{i=1}^{k} n_i \bar{y}_{i.}/N,\;\; BSS = \sum_{i=1}^{k} n_i (\bar{y}_{i.} - \bar{y}_{..})^2,\;\; WSS = S_1^2 + \cdots + S_k^2 \tag{7.3}$$

Distributional properties.

Obviously, under the assumption of normality and independence, the distribution of $\bar{y}_{..} \sim N[\mu, \sum_{i=1}^{k} n_i^2(\sigma_a^2 + \sigma_i^2/n_i)/N^2]$. When the homogeneity of error variances hold, this reduces to $\bar{y}_{..} \sim N[\mu, \sum_{i=1}^{k} n_i^2(\sigma_a^2 + \sigma^2/n_i)/N^2]$. Furthermore, in case of balanced models and homogeneity of error variances, we get $\bar{y}_{..} \sim N[\mu, (\sigma_a^2 + \sigma^2/n)/k]$.

For WSS, we readily have

$$WSS \sim \sum_{i=1}^{k} \sigma_i^2 \chi_{n_i - 1}^2 \sim \sigma^2 \chi_{N-k}^2, \tag{7.4}$$

with the latter result holding in case of the homogeneity of error variances.

For BSS, the results are some what complicated except for the balanced case with homogeneous error variances. Quite generally, since BSS can be written as a quadratic form in the sample means (corrected for the mean μ, without any loss of generality), we can conclude that the general distribution of BSS can be written as a linear function of independent chisquare variables with coefficients depending on the variance components and the replications. Under homogeneous error variances and a balanced model, we get

$$BSS \sim (\sigma^2 + n\sigma_a^2)\chi^2_{k-1}. \tag{7.5}$$

Of course, under normality of errors and random effects, independence of BSS and WSS follows immediately.

The between group mean sum of squares $BMS = BSS/(k-1)$, denoted as MS_1, has the expected value given by

$$\mathrm{E}(\mathrm{MS}_1) = \gamma\sigma_a^2 + \sigma^2\,, \qquad \gamma = \frac{1}{k-1} \cdot \frac{N^2 - \sum_{i=1}^k n_i^2}{N}. \tag{7.6}$$

For later reference, we note from (7.4) that, under the assumption of homogeneity of error variances, a $(1-\alpha)$ level confidence interval for σ^2 is given by

$$\mathrm{CI}(\sigma^2) : \left[\frac{(N-k)\mathrm{MS}_2}{\chi^2_{N-k;\,1-\alpha/2}} ; \frac{(N-k)\mathrm{MS}_2}{\chi^2_{N-k;\,\alpha/2}} \right], \tag{7.7}$$

where $MS_2 = WSS/(N-K)$ and $\chi^2_{\nu;\gamma}$ denotes the γ–quantile of a χ^2–distribution with ν degrees of freedom.

It should also be mentioned that the approximation of the distribution of MS_1 by a multiple of a χ^2–distribution in the general case is satisfactory only if the between group variance σ_a^2 is close to 0. This explains why an easy extension of the confidence interval for σ_a^2 in the balanced case independently proposed by Tukey (1951) and Williams (1962), to be discussed later in this chapter, is not possible in the unbalanced case.

7.2 Homogeneous error variances

Under the assumption of homogeneous error variances, i.e., $\sigma_1^2 = \sigma_2^2 = \ldots\cdots = \sigma_k^2 = \sigma^2$, the above model clearly boils down to the familiar intraclass correlation model with just three parameters: overall mean μ, between study variance σ_a^2 and within study or error variance σ^2. For balanced models, i.e., when the treatment replications n_1, ..., n_k are the same, there are exactly three sufficient statistics, namely, the overall sample mean, the between sum of squares and the within sum of squares, and it is easy to derive the UMVUEs of the three parameters. In case of unbalanced models, there are many sufficient statistics and an unbiased estimate of the between study variance σ_a^2 is not unique! Moreover, applying a fundamental result of Lamotte (1973), it turns out that all unbiased estimates of σ_a^2 in both balanced and unbalanced models are bound to assume negative values, thus making them unacceptable in practice. A lot of research has been conducted in order to derive nonnegative estimates of σ_a^2 with good frequentist properties. Because our emphasis here is more on tests and confidence intervals rather

than on estimation, we omit these details and refer to the book by Searle, Casella, Mc Culloch (1992).

7.2.1 Test for $\sigma_a^2 = 0$

Although the one-way random effects model postulates the presence of a random component, it is of interest to test if this random component a_i is indeed present in the model (7.1). Since the null hypothesis here corresponds to the equality of means in a standard ANOVA setup, we can use the regular F-test based on the ratio of between and within sums of squares, i.e., $F = \frac{BSS/(k-1)}{WSS/(N-k)}$. Although this F-test is known to have some optimum properties in the balanced case, in the unbalanced case this F-test though valid ceases to have optimum properties, and a locally best invariant test under a natural group of transformations was derived by Das and Sinha (1987). The test statistic F^*, whose sampling distribution does not follow any known tabulated distribution, is given by

$$F^* = \frac{\sum_{i=1}^{k} n_i^2 (\bar{y}_{i.} - \bar{y}_{..})^2}{WSS}. \tag{7.8}$$

7.2.2 Approximate tests for $H_0 : \sigma_a^2 = \delta > 0$ and confidence intervals for σ_a^2

When the F-test for the nullity of the between study variability is rejected, it is of importance to test for other meaningful positive values of this parameter as well as to construct its appropriate confidence intervals. Quite surprisingly, this particular problem has been tackled by many researchers over the last fifty years. It is clear from (7.5) that even in the case of balanced models, there is no obvious test for a positive value of σ_a^2 and also it is not clear how to construct an *exact* confidence interval for σ_a^2. This is the main reason for a lot of research on this topic. Fortunately, the relatively new notion of a generalized P-value can be used to solve these problems *exactly* even in the case of unbalanced models. We will discuss this solution in section 7.2.3.

In this section however we provide a survey of some main results on the derivation of approximate confidence intervals for σ_a^2 mostly from a classical point of view. Once an appropriate confidence interval is derived, it can be used to test the significance of a suggested positive value of the parameter σ_a^2 in the usual way. It should be noted that most of the procedures discussed below provide approximate solutions to our problem especially in unbalanced models. In the sequel, we discuss the two cases of balanced and unbalanced models separately.

Balanced models.

In the case of a balanced design, one method for constructing a confidence interval for the between group variance σ_a^2 was proposed by Tukey (1951) and also independently

by Williams (1962). The Tukey-Williams method is based upon noting the distributional properties of BSS and WSS, given in (7.4) and (7.5). Since it is easy to construct confidence intervals for σ^2 and $\sigma^2+n\sigma_a^2$, exact $(1-\alpha)$–confidence intervals for these parameters can be easily calculated and by solving the intersection of these two confidence intervals, a confidence interval of the between group variance σ_a^2 can be obtained which has a confidence coefficient at least $(1-2\alpha)$ due to Bonferroni's inequality. The results of simulation studies conducted by Boardman (1974) indicated that the confidence coefficient of the Tukey-Wiliams interval is nearly $1-\alpha$ (cf. also Graybill (1976, p. 620)) and Wang (1990) showed that the confidence coefficient of this interval is even at least $1-\alpha$ for customary values of α.

Unbalanced models

Following the Tukey–Williams approach, Thomas and Hultquist (1978) proposed a confidence interval for the between group variance σ_a^2 in the unbalanced case. This is based on a suitable χ^2 approximation of the distribution of BSS. However, this approximation is not good if the design is extremely unbalanced or if the ratio of the between and within group variances is less than 0.25. To overcome this problem, Burdick, Maqsood and Graybill (1986) considered a conservative confidence interval for the ratio of between and within group variance, which was used in Burdick and Eickman (1986) to construct a confidence interval for the between group variance based on the ideas of the Tukey–Williams method. In Burdick and Eickman (1986), a comparison of the confidence coefficients of the Thomas–Hultquist interval and the Burdick–Eickman interval on the basis of some simulation studies is reported. The results of the simulations studies indicated that the confidence coefficient is near $1-\alpha$ in most cases. If the approximation to a χ^2–distribution in the Thomas–Hultquist approach is not so good, the resulting confidence interval can be very liberal, while in these situations the Burdick–Eickman intervall can be very conservative.

Hartung and Knapp (2000) proposed a confidence interval for the between group variance in the unbalanced design which is constructed from an exact confidence interval for the ratio of between and within group variance derived from Wald (1940), (see also Searle, Casella, and McCulloch (1992, p. 78), Burdick and Graybill (1992, p. 186 f.)), and an exact confidence interval of the error variance.

We describe below all the three procedures mentioned above for constructing an approximate confidence interval for σ_a^2 based on the two familiar sums of squares, namely, between and within sums of squares.

Thomas–Hultquist confidence interval for σ_a^2.

Instead of MS_1 from (7.6), Thomas and Hultquist (1978) considered the sample vari-

ance of the group means given by

$$\mathrm{MS}_3 = \frac{1}{k-1}\sum_{i=1}^{k}\left(\overline{y_{i\cdot}} - \frac{1}{k}\sum_{i=1}^{k}\overline{y_{i\cdot}}\right)^2 . \tag{7.9}$$

They showed that it holds approximately

$$\frac{(k-1)\mathrm{MS}_3}{\sigma_a^2 + \sigma^2/\tilde{n}} \overset{\text{appr.}}{\sim} \chi^2_{k-1}\,, \tag{7.10}$$

where $\tilde{n}$ denotes the harmonic mean of the sample sizes of the k groups.

Combining (7.4) and (7.10), it is then easy to conclude that

$$\frac{\sigma^2}{\sigma_a^2 + \sigma^2/\tilde{n}} \cdot \frac{\mathrm{MS}_3}{\mathrm{MS}_2} \overset{\text{appr.}}{\sim} F_{k-1,\,N-k}\,, \tag{7.11}$$

where $F_{\nu_1,\,\nu_2}$ denotes a F–distributed random variable with ν_1 and ν_2 degrees of freedom.

From (7.10) and (7.11), $(1-\alpha)$ level confidence intervals for $\sigma_a^2 + \sigma^2/\tilde{n}$ and σ_a^2/σ^2 can be constructed and adopting the ideas of constructing a confidence interval by Tukey and Williams to the present situation leads to the following confidence interval for σ_a^2:

$$\mathrm{CI}_{\mathrm{TH}}(\sigma_a^2) : \left[\frac{(k-1)}{\chi^2_{k-1;\,1-\alpha/2}}\left(\mathrm{MS}_3 - \frac{\mathrm{MS}_2}{\tilde{n}}F_{k-1,\,N-k;\,1-\alpha/2}\right)\,;\right.$$
$$\left.\frac{(k-1)}{\chi^2_{k-1;\,\alpha/2}}\left(\mathrm{MS}_3 - \frac{\mathrm{MS}_2}{\tilde{n}}F_{k-1,\,N-k;\,\alpha/2}\right)\right] . \tag{7.12}$$

Due to Bonferroni's inequality the confidence coefficient of (7.12) is at least $(1-2\alpha)$, but one may hope that the actual confidence coefficient is nearly $(1-\alpha)$. However, as mentioned earlier, Thomas and Hultquist (1978) reported that the χ^2–approximation in (7.10) is not good for extremely unbalanced designs where the ratio $\eta = \sigma_a^2/\sigma_e^2$ is less than 0.25. Thus, in such situations the confidence interval (7.12) can be a liberal one, i. e. the confidence coefficient substantially lies below $(1-\alpha)$.

Burdick–Eickman confidence interval for σ_a^2.

Burdick, Maqsood and Graybill (1986) suggested a confidence interval for the ratio $\eta = \sigma_a^2/\sigma^2$ which overcomes the problem with small ratios in the Thomas–Hultquist procedure and has a confidence coefficient of at least $1-\alpha$. This interval is given by

$$\mathrm{CI}(\eta) : \left[\frac{\mathrm{MS}_3}{\mathrm{MS}_2}\cdot\frac{1}{F_{k-1,\,N-k;\,1-\alpha/2}} - \frac{1}{n_{\min}}\,;\,\frac{\mathrm{MS}_3}{\mathrm{MS}_2}\cdot\frac{1}{F_{k-1,\,N-k;\,1-\alpha/2}} - \frac{1}{n_{\max}}\right] \tag{7.13}$$

with $n_{\min} = \min\{n_1, \ldots, n_k\}$ and $n_{\max} = \max\{n_1, \ldots, n_k\}$.

The difference between (7.12) and (7.13) lies in subtracting $1/\tilde{n}$ in both bounds instead of $1/n_{\min}$ and $1/n_{\max}$, respectively, in (7.13).

Using (7.13) and the confidence interval for $\sigma_a^2 + \sigma^2/\tilde{n}$ from (7.10), Burdick and Eickman (1986) investigated the confidence interval for σ_a^2 constructed by the Tukey-Williams method.

This interval is given by

$$\mathrm{CI}_{\mathrm{BE}}(\sigma_a^2) : \left[\left(\frac{\tilde{n}L}{1+\tilde{n}L} \right) \cdot \frac{(k-1)\mathrm{MS}_3}{\chi^2_{k-1;\,1-\alpha/2}} ; \left(\frac{\tilde{n}U}{1+\tilde{n}U} \right) \cdot \frac{(k-1)\mathrm{MS}_3}{\chi^2_{k-1;\,1-\alpha/2}} \right] , \tag{7.14}$$

with

$$L = \max\left\{ 0,\ \frac{\mathrm{MS}_3}{\mathrm{MS}_2} \cdot \frac{1}{F_{k-1,\,N-k;\,1-\alpha/2}} - \frac{1}{n_{\min}} \right\}$$

and

$$U = \max\left\{ 0,\ \frac{\mathrm{MS}_3}{\mathrm{MS}_2} \cdot \frac{1}{F_{k-1,\,N-k;\,\alpha/2}} - \frac{1}{n_{\max}} \right\}.$$

Hartung-Knapp confidence interval for σ_a^2.

Instead of approximative confidence intervals for η as in the Thomas–Hultquist and Burdick–Eickman approach, Hartung and Knapp (2000) considered the exact confidence interval for η given in Wald (1940) to construct a confidence interval for σ_a^2.

Following Wald (1940), we observe that

$$\mathrm{Var}(\overline{y_{i\cdot}}) = \sigma_a^2 + \sigma^2/n_i = \sigma^2/w_i \tag{7.15}$$

with $w_i = n_i/(1+\eta n_i)$, $i = 1, \ldots, k$.

Now, Wald considered the sum of squares

$$(k-1)\mathrm{MS}_4 = \sum_{i=1}^{k} w_i \left(\overline{y_{i\cdot}} - \frac{\sum_{i=1}^{k} w_i \overline{y_{i\cdot}}}{\sum_{i=1}^{k} w_i} \right)^2 \tag{7.16}$$

and proved that

$$(k-1)\mathrm{MS}_4/\sigma^2 \sim \chi^2_{k-1}\,.$$

Furthermore, MS_4 and MS_2 are stochastically independent so that

$$F_w(\eta) = \frac{\mathrm{MS}_4}{\mathrm{MS}_2} \sim F_{k-1,\,n-k}\,. \tag{7.17}$$

Obviously, (7.17) can be used to construct an exact confidence interval for the ratio η.

Wald showed that $(k-1)\mathrm{MS}_4$ is a strictly monotonously decreasing function in η, and so the bounds of the exact confidence interval are given as the solutions of the following two equations:

$$\begin{array}{ll} \text{lower bound:} & F_w(\eta) = F_{k-1,\,N-k,\,1-\alpha/2} \\ \text{upper bound:} & F_w(\eta) = F_{k-1,\,N-k,\,\alpha/2} \end{array} \tag{7.18}$$

Since $F_w(\eta)$ is a strictly monotonously decreasing function in η, the solution of (7.18), if it exists, is unique. But due to the fact that η is nonnegative, $(k-1)\mathrm{MS}_4$ is bounded at $\eta = 0$, namely it holds that

$$(k-1)\mathrm{MS}_4 \le \sum_{i=1}^{k} n_i \left(\overline{y_{i\cdot}} - \frac{\sum_{i=1}^{k} n_i \overline{y_{i\cdot}}}{\sum_{n=1}^{k} n_i} \right)^2 . \tag{7.19}$$

Thus, a nonnegative solution of (7.18) may not exist. If such a solution of one of the equations in (7.18) does not exist, the corresponding bound in the confidence interval is set equal to zero. Note that the existence of a nonnegative solution in (7.18) only depends on the chosen α.

Let us denote by η_L and η_U the solutions of the equations in (7.18). We then propose, using the confidence bounds from (7.7) for σ^2, the following confidence interval for σ_a^2:

$$\mathrm{CI}(\sigma_a^2) : \left[\frac{(N-k)\mathrm{MS}_2}{\chi^2_{N-k;\,1-\alpha}} \cdot \eta_L \,;\, \frac{(N-k)\mathrm{MS}_2}{\chi^2_{N-k;\,\alpha}} \cdot \eta_U \right] , \tag{7.20}$$

which has a confidence coefficient of at least $(1-2\alpha)$ according to Bonferroni's inequality. But due to the fact that the confidence coefficient of $[\sigma^2 \cdot \eta_L, \sigma^2 \cdot \eta_U]$ is exactly $1-\alpha$, the resulting confidence interval (7.20) may be very conservative, i. e. the confidence coefficient is larger than $(1-\alpha)$. So, we also consider a confidence interval for σ_a^2 with the estimator MS_2 for σ^2 instead of the bounds of the confidence interval for σ^2, i. e.

$$\widetilde{\mathrm{CI}}(\sigma_a^2) : [\mathrm{MS}_2 \cdot \eta_L \;;\; \mathrm{MS}_2 \cdot \eta_U] \;. \tag{7.21}$$

Through extensive simulation studies conducted by Hartung and Knapp (2000), the observations of Burdick and Eickman (1986) are confirmed in the sense that the Thomas–Hultquist interval may be very liberal for small σ_a^2, i. e. the confidence coefficient lies considerably below $1-\alpha$. In these situations, the Burdick–Eickman interval has a confidence

coefficient which is always larger than $1-\alpha$, but the interval can be very conservative. If σ_a^2 becomes larger, both intervals are very similar. The confidence interval CI deduced from Wald's confidence interval for the ratio η with the bounds of the confidence interval of the error variance as estimates for the error variance has always a confidence coefficient at least as great as $1-\alpha$, but this interval can be very conservative for large σ_a^2. A good compromise for the whole range of σ_a^2 is the confidence interval $\widetilde{\text{CI}}$ from (7.21), which has a confidence coefficient at least as great as $1-\alpha$ for small σ_a^2, and for growing σ_a^2 the confidence interval only becomes moderately conservative.

7.2.3 Exact test and confidence interval for σ_a^2 based on a generalized P-value approach

.

In this section we describe the relatively new notion of a generalized P-value and its applications to our problem. The original ideas are due to Tsui and Weerahandi (1989) and Weerahandi (1993).

We start with a general description of the notion of a generalized P-value. If X is a random variable whose distribution depends on the scalar parameter θ of interest and a set of nuisance parameters η, and the problem is to test $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$, a generalized P-value approach proceeds by judiciously specifying a test variable $T(X; x, \theta, \eta)$ which depends on the random variable X, its observed value x, and the parameters θ and η, satisfying the following three properties:

(i) The sampling distribution of $T(X; x, \theta, \eta)$ derived from that of X, for fixed x, is free of the nuisance parameter η;
(ii) The observed value of $T(X; x, \theta, \eta)$ when $X = x$, i.e., $T(x; x, \theta, \eta)$ is free of the nuisance parameter η;
(iii) $P[T(X; x, \theta, \eta) \geq t]$ is nondecreasing in θ, for fixed x and η.

Under the above conditions, a generalized P-value is defined by

$$genP = P[T(X; x, \theta_0, \eta) \geq t] \tag{7.22}$$

where $t = T(x; x, \theta_0, \eta)$.

In the same spirit as above, Weerahandi (1993) constructed a one-sided confidence bound for θ based on a test variable $T_1(X; x, \theta, \eta)$ satisfying the above three properties and also the added constraint that the observed value of T_1 is $T_1(x; x, \theta, \eta) = \theta$. Let $t_1(x)$ satisfy the condition:

$$P[T_1 \leq t_1(x)] = 1 - \alpha. \tag{7.23}$$

Then $t_1(x)$ can be regarded as a $(1-\alpha)$ level upper confidence limit for θ.

We now turn our attention to the applications of this concept to our specific problem.

a) **An exact test for $H_0 : \sigma_a^2 = \delta > 0$ in the balanced case**. This testing problem is commonly known in the literature as a non-standard testing problem in the sense that there are no obvious pivots or exact tests for testing this null hypothesis based on the two sums of squares: BSS and WSS. Recall that, in the balanced case, this fact follows from the canonical form of the model based on two independent sums of squares, $BSS \sim (\sigma^2 + n\sigma_a^2).\chi^2_{(k-1)}$ and $WSS \sim \sigma^2.\chi^2_{k(n-1)}$. While it is obvious that an exact test for $\sigma_a^2 = 0$ or even for $\sigma_a^2/\sigma^2 = \delta$ can be easily constructed simply by taking the ratio of BSS and WSS, the same is not true for the null hypothesis $H_0 : \sigma_a^2 = \delta$ for some $\delta > 0$. We now describe a test for this hypothesis in the balanced case based on a generalized P-value.

In our context, taking $X = (BSS, WSS)$, $x = (bss, wss)$, the observed values of X, $\theta = \sigma_a^2$ and $\eta = \sigma^2$, we define

$$T(BSS, WSS; bss, wss, \sigma_a^2, \sigma^2) = \frac{n\sigma_a^2 + wss(\sigma^2/WSS)}{bss[(n\sigma_a^2 + \sigma^2)/BSS]} \tag{7.24}$$

It is easy to verify that T defined above satisfies the conditions $(i), (ii), (iii)$, and hence the generalized P-value for testing $H_0 : \sigma_a^2 = \delta$ or $H_0 : \sigma_a^2 \leq \delta$ versus $H_1 : \sigma_a^2 > \delta$ is given by

$$genP = P[T \geq 1] = P[\frac{n\delta + wss/U_e}{bss/U_a} \geq 1] \tag{7.25}$$

where $U_a = BSS/(\sigma^2 + n\sigma_a^2) \sim \chi^2_{(k-1)}$ and $U_e = WSS/\sigma^2 \sim \chi^2_{k(n-1)}$. The test procedure rejects H_0 if the generalized P-value is small. We should point out that the computation of the generalized P-values in this and similar other problems is facilitated by the software package $XPro$ Software Package (1994) developed by X-Techniques, Inc.

b) **One-sided confidence bound for σ_a^2 in the balanced case**. We define

$$T_1 = T(BSS, WSS; bss, wss, \sigma_a^2, \sigma^2) = \frac{1}{n}[\frac{bss(n\sigma_a^2 + \sigma^2)}{BSS} - \frac{wss\sigma^2}{WSS}] = \frac{1}{n}[\frac{bss}{U_a} - \frac{wss}{U_e}] \tag{7.26}$$

It is then easy to verify that the sampling distribution of T_1, for fixed $x = (bss, wss)$, does not depend on σ^2, and that the observed value of T_1 is indeed σ_a^2. Let $t_1(bss, wss)$ satisfy the condition

$$P[T_1 \leq t_1(bss, wss)] = 1 - \alpha. \tag{7.27}$$

Then $t_1(bss, wss)$ can be regarded as a $(1-\alpha)$ level upper confidence limit for σ_a^2.

We now provide an example from Verbeke and Molenberghs (1997) to illustrate the application of this approach.

Example 7.1. This example demonstrating the application of generalized P-value to find an upper confidence limit of σ_a^2 is taken from Verbeke and Molenberghs (1997). To measure the efficiency of an antibiotic after it has been stored for two years, eight batches of the drug are randomly selected from a population of available batches and a random sample of size two is taken from each selected batch (balanced design). Data representing the concentration of the active component are given below.

Batch:	1	2	3	4	5	6	7	8
Obs:	40	33	46	55	63	35	56	34
	42	34	47	52	59	38	56	29

Employing the very natural one-way balanced random effects model here and doing some routine computations, the following ANOVA table is obtained. It is evident from the ANOVA table that a batch to batch variability is very much in existence in this problem, implying $\sigma_a^2 > 0$.

ANOVA table

	Sum of squares	d.f.	Mean squares	Expected mean squares
Batches	BSS = 1708	7	BMS = 244.1	$2\sigma_a^2 + \sigma^2$
Error	WSS = 32.5	8	WMS = 4.062	σ^2

In order to derive a 95% upper confidence interval for the parameter σ_a^2, it is indeed possible to use the familiar Satterthwaite approximation which is as follows. The estimate of σ_a^2 is $\hat{\sigma}_a^2 = (BMS - WMS)/2$. Consider the approximation $\nu\hat{\sigma}_a^2/\sigma_a^2 \sim \chi_\nu^2$ and equating the second moments yield $\hat{\nu} = (BMS - WMS)^2/(BMS^2/7 + WMS^2/8) = 1.69$. This leads to the interval $\sigma_a^2 \leq \hat{\nu}\hat{\sigma}_a^2/\chi_{0.05;\hat{\nu}}^2 = 3707.50$, which is just useless for this problem.

On the other hand, the application of a generalized confidence interval to this problem, as developed here, based on $T_1 = (1708/U_a - 32.5/U_e)/2$ yields $\sigma_a^2 \leq 392.27$, which is much more informative than the previous bound.

c) **An exact test for $H_0 : \sigma_a^2 = \delta > 0$ and a confidence bound for σ_a^2 in the unbalanced case**.

For testing $H_0 : \sigma_a^2 = \delta > 0$ versus the alternative $H_1 : \sigma_a^2 > \delta$, a potential generalized test variable can be defined as follows.

Define $\rho = \sigma_a^2/\sigma^2$, $w_i(\rho) = n_i/(1+\rho n_i)$,
$\bar{Y}_w(\rho) = \sum_{i=1}^k w_i(\rho)\bar{Y}_{i.}/\sum_{i=1}^k w_i(\rho) \sim N[\mu, \sigma^2/\sum_{i=1}^k w_i(\rho)]$. Let

$$S_{wB}(\rho) = \sum_{i=1}^{k} w_i(\rho)(\bar{Y}_{i.} - \bar{Y}_w(\rho))^2 \sim \sigma^2\chi^2_{k-1}. \tag{7.28}$$

Define $W_1 = WSS/\sigma^2 \sim \chi^2_{N-k}$ and $W_2 = S_{wB}(\rho)/\sigma^2 \sim \chi^2_{k-1}$ which are independent. Finally, let

$$\begin{aligned} T(S_{wB}(\rho), WSS; s_{wB}, wss, \sigma_a^2, \sigma^2) &= \frac{wss S_{wB}(\rho)}{WSS s_{wB}(\frac{\sigma_a^2 WSS}{\sigma^2 wss})} \\ &= \frac{W_2 wss}{W_1 s_{wB}(\frac{\sigma_a^2 W_1}{wss})} \end{aligned} \tag{7.29}$$

It is easily seen from the second equality above that the distribution of the test variable $T(S_{wB}(\rho), WSS; s_{wB}, wss, \sigma_a^2, \sigma^2)$ depends only on the parameter of interest, namely, σ_a^2, and is independent of the nuisance parameter σ^2! It also follows from the first equality above that the observed value of $T(S_{wB}(\rho), WSS; s_{wB}, wss, \sigma_a^2, \sigma^2)$ is one. Hence, the generalized P-value for testing $H_0 : \sigma_a^2 = \delta > 0$ versus the alternative $H_1 : \sigma_a^2 > \delta$ is given by

$$\begin{aligned} P &= Pr[T(S_{wB}(\rho), WSS; s_{wB}, wss, \sigma_a^2, \sigma^2) \geq 1 | \sigma_a^2 = \delta] \\ &= Pr[W_2 \geq \frac{W_1}{wss} s_{wB}(\frac{W_2\delta}{wss})] \end{aligned} \tag{7.30}$$

The generalized confidence bounds for σ_a^2 can be obtained by solving the equations:
$Pr[W_2 \geq \frac{W_1}{wss} s_{wB}(\frac{W_2\delta_1}{wss})] = \frac{\alpha}{2}$
$Pr[W_2 \geq \frac{W_1}{wss} s_{wB}(\frac{W_2\delta_2}{wss})] = 1 - \frac{\alpha}{2}$
in which case $[\delta_2, \delta_1]$ is the $100(1-\alpha)\%$ generalized confidence interval for σ_a^2.

The above P value and the confidence bounds can be conveniently computed using the XPro software package.

7.2.4 Tests and confidence intervals for μ

In this subsection we discuss some tests and confidence intervals for the overall mean parameter μ. Clearly, an unbiased estimate of μ is given by the overall sample mean $\bar{y} = \sum \bar{y}_i/k$ whose distribution is normal with mean μ and variance $\eta^2 = \sum(\sigma_a^2 + \sigma^2/n_i)/k^2$.

In the balanced case when $n_1 = \cdots = n_k = n$, $\bar{y}_i$'s are iid with a common mean μ and a common variance $\sigma_a^2 + \sigma^2/n$ so that a t-test can be carried out to test hypotheses about μ and also the usual t-statistic can be used to derive confidence limits for μ.
In the unbalanced case, however, only some approximate tests and confidence intervals for μ can be developed. We can easily estimate the common within study variance σ^2 by just combining the within sample variances MS_2 with a combined df $N - k$. As for the other variance component, namely, the between study variance σ_a^2, since the usual ANOVA estimate can assume negative values, many modifications of it are available in the literature. A normal approximation is then used for the distribution of the so-called studentized variable $t = (\bar{y} - \mu)/\hat{\sigma}(\bar{y})$ to obtain approximate tests and confidence intervals for μ. Details can be found in Rukhin and Vangel (1998), Rukhin, Biggerstaff and Vangel (2000).

An exact test for μ in the unbalanced case is described in Iyer et al. (2004), using the notion of the generalized P-value. However, the solution is rather complicated and we omit the details.

7.3 Heterogeneous error variances

In this section we discuss the problem of drawing appropriate inferences about the overall mean μ and the between study variability σ_a^2 under the more realistic scenario of heterogeneous error or within study variances. It is obvious that one can estimate an within study variance σ_i^2 from replicated observations from the ith study. However, the associated inference problems here are quite hard and some satisfactory solutions have been offered only recently.

7.3.1 Tests for $H_0 : \sigma_a^2 = 0$

It is of course clear that testing the nullity of the between study variance $H_0 : \sigma_a^2 = 0$ is easy to carry out because, under the null hypothesis, one has the usual fixed effects model with heterogeneous error or within study variances. Thus, all the test procedures described in lecture 6 are applicable here. Argac, Makambi, and Hartung (2001) performed some simulation in the context of this problem in an attempt to compare the proposed tests in terms of power and recommend the use of adjusted Welch test in most cases.

7.3.2 Tests for $H_0 : \sigma_a^2 = \delta > 0$

In many situations it is known a priori that some positive level of between study variability may be present and it is desired to prescribe a test for a designated positive value for this parameter. Due to the heterogeneous error variances, it is clear that the testing problem here is quite difficult. In the following, we provide below two solutions to this problem.

Hartung, Makambi, and Argac (2001) propose a test statistic they call extended ANOVA test statistic, and this is given by

$$F_A^* = \frac{\sum_{i=1}^{k} h_i \left(\bar{y}_{i.} - \sum_{j=1}^{k} h_j \bar{y}_{j.}\right)^2 / (k-1)}{\delta \sum_{i=1}^{k} h_i^2 + \sum_{i=1}^{k} h_i^2 \, S_i^2 / n_i} \tag{7.31}$$

with $h_i = w_i / \sum_{j=1}^{k} w_j$, $w_i = 1/\tilde{\tau}_i^2$, and $\tilde{\tau}_i^2 = \delta + S_i^2/n_i$. Under the null hypothesis $H_0 : \sigma_a^2 = \delta$, the test statistic F_A^* is approximately F-distributed with $(k-1)$ and $\hat{\nu}_A$ degrees of freedom, where

$$\hat{\nu}_A = \frac{\left(\sum_{i=1}^{k} h_i^2 \, \tilde{\tau}_i^2\right)^2}{\sum_{i=1}^{k} h_i^4 \, S_i^4 \, / \, (n_i^2 \, (n_i + 1))} - 2. \tag{7.32}$$

So, we reject $H_0 : \sigma_a^2 = \delta$ at level α, if $F^* > F_{m-1,\hat{\nu}_R;1-\alpha}$.

Hartung and Argac (2002) derive an extension of the Welch test. Their test statistic is given by

$$F_W^* = \frac{\sum_{i=1}^{k} w_i \left(\bar{y}_{i.} - \sum_{j=1}^{k} h_j \bar{y}_{j.}\right)^2}{(k-1) + 2(k-2)(k-1)^{-1} \sum_{i=1}^{k} (1-h_i)^2 / \hat{\nu}_i} \tag{7.33}$$

with $h_i = w_i / \sum_{j=1}^{k} w_j$, $w_i = 1/(\delta + S_i^2/n_i)$, and $\hat{\nu}_i = 2(\delta + S_i^2/n_i)^2 / \{2S_i^2/(n_i^2(n_i+1))$. The test statistics F_W^* is approximately F-distributed under the null hypothesis with $(k-1)$ and $\hat{\nu}_W$ degrees of freedom, where

$$\hat{\nu}_W = \frac{k^2 - 1}{3 \sum_{i=1}^{k} (1-h_i)^2 / \hat{\nu}_i}. \tag{7.34}$$

The null hypothesis is rejected at level α if $F_W^* > F_{k-1,\hat{\nu}_W;1-\alpha}$.

In section 7.4, appropriate confidence intervals of σ_a^2 will be presented, which in turn can also be used to test the significance of a designated positive value of this parameter.

7.3.3 Nonnegative estimation of σ_a^2

In this subsection we discuss various procedures to derive nonnegative estimates of the central parameter of interest, namely, σ_a^2. The estimators are either based on quadratic forms in y or on likelihood methods.

Rao, Kaplan, and Cochran (1981) discuss extensively the parameter estimation in the one-way random-effects models. We present here three estimators of σ_a^2 from this paper that are generally eligible for use in meta analysis.

With the between and within classes sum of squares, the unbiased ANOVA type estimator of σ_a^2 has the form

$$\hat{\sigma}_a^2 = \left(\frac{n}{n^2 - \sum_{i=1}^k n_i^2} \right) \left(\sum_{i=1}^k n_i (\bar{y}_{i.} - \bar{y}_{..})^2 - \sum_{i=1}^k \left(1 - \frac{n_i}{n} \right) S_i^2 \right) \tag{7.35}$$

Based on the unweighted sum of squares, the unbiased ANOVA-type estimator of the between group variance is given by

$$\hat{\sigma}_a^2 = \frac{1}{k-1} \sum_{i=1}^k (\bar{y}_i - \bar{y}^*)^2 - \frac{1}{k} \sum_{i=1}^k \frac{S_i^2}{n_i} \tag{7.36}$$

with $\bar{y}^* = \sum_{i=1}^k \bar{y}_i / k$ the mean of the group means.

Both estimators, (7.35) and (7.36), are unbiased estimators of σ_a^2. However, both estimators can yield negative values. Based on the Rao's (1972) MINQUE principle without the condition of unbiasedness, Rao, Kaplan, and Cochran (1981) provide an always nonnegative estimator of σ_a^2 as

$$\hat{\sigma}_a^2 = \frac{1}{k} \sum_{i=1}^k \ell_i^2 \left(\bar{y}_{i.} - \bar{\bar{y}}_{..} \right)^2, \tag{7.37}$$

where $\ell_i = n_i/(n_i + 1)$ and $\bar{\bar{y}}_{...} = (\sum_{i=1}^k \ell_i \bar{y}_{i.})/(\sum_{i=1}^k \ell_i)$.

In the biomedical literature, an estimator proposed by DerSimonian and Laird (1986) is widely used. Based on Cochran's homogeneity statistic and using the method of moment approach in the one-way random effects model assuming known within-group variances σ_i^2, they derive the estimator

$$\hat{\sigma}_a^2 = \frac{\sum_{i=1}^k w_i (\bar{y}_{i.} - \tilde{y}_{..})^2 - (k-1)}{\sum_{i=1}^k w_i - \sum_{i=1}^k w_i^2 / \sum_{i=1}^k w_i} \tag{7.38}$$

where $\tilde{y}_{..} = \sum_{i=1}^k w_i \bar{y}_{i.} / \sum_{i=1}^k w_i$, and in the present model, $w_i = n_i/\sigma_i^2$. The estimator (7.38) is an unbiased estimator of σ_a^2 given known σ_i^2. In practice, estimates of the within-group variances have to be plugged in and then, the estimator is no longer unbiased. Moreover, like the unbiased ANOVA-type estimators (7.35) and (7.36), the DerSimonian-Laird estimator can yield negative estimates with positive probability.

Using the general approach of nonnegative minimum biased invariant quadratic estimation of variance components proposed by Hartung (1981), Heine (1993) derives the nonnegative minimum biased estimator of σ_a^2 in the present model. If $N - 2n_i \geq 0$, $i = 1, \ldots, k$, this estimator reads

$$\hat{\sigma}_a^2 = \frac{n^2}{\big(\sum\limits_{\ell=1}^{k} n_\ell^2 + 1\big) \sum\limits_{\ell=1}^{k} n_\ell(N - n_\ell) \prod\limits_{\ell' \neq \ell} (N - 2n_{\ell'})} \sum_{i=1}^{k} n_i^2 \prod_{\ell' \neq \ell} (N - 2n_{\ell'})(\bar{y}_{i.} - \bar{y}_{..})^2 \quad (7.39)$$

Maximum likelihood estimation in the present model has been already discussed by Cochran (1954). Rukhin, Biggerstaff, and Vangel (2000) provide the estimation equations of the maximum likelihood (ML) and the restricted maximum likelihood estimator (REML) estimator of σ_a^2.

In the one-way random effects model, all the data are available. Sometimes, only summary statistics are available and then we obtain the following model which can be seen as a special case of the one-way random effects model. Let $(\bar{y}_{1.}, S_1^2), (\bar{y}_{2.}, S_2^2), \ldots, (\bar{y}_{k.}, S_k^2)$ be independent observations representing summary estimates $\bar{y}_{i.}$ of some parameter μ of interest from k independent sources, together with estimates S_i^2/n_i of the variances of $\bar{y}_{i.}$, and n_i denotes the corresponding sample size.

The random effects meta analysis model is given as

$$\bar{y}_{i.} = \mu + a_i + e_i, i = 1, \ldots, k, \quad (7.40)$$

where a_i are normally distributed random variables with mean zero and variance σ_a^2, representing the between-group variance, and e_i are normally distributed random variables with mean zero and variance σ_i^2/n_i. In model (7.40), we assume that the variances σ_i^2 are reasonably well estimated by the S_i^2 within the independent groups. So, we assume the σ_i^2 are known and simply replace them by their estimates s_i^2.

Taking the σ_i^2 as known, the estimating equations of the maximum likelihood estimators of μ and σ_a^2 are given by

$$\mu = \frac{\sum_{i=1}^{k} w_i(\sigma_a^2)\, \bar{y}_{i.}}{\sum_{i=1}^{k} w_i(\sigma_a^2)}\,, \quad (7.41)$$

$$\sum_{i=1}^{k} w_i^2(\sigma_a^2)\, (\bar{y}_{i.} - \mu)^2 = \sum_{i=1}^{k} w_i(\sigma_a^2) \quad (7.42)$$

where $w_i(\sigma_a^2) = (\sigma_a^2 + s_i^2/n_i)^{-1}$. A convenient form of equation (7.42) for iterative solution is given by

$$\sigma_a^2 = \frac{\sum_{i=1}^{k} w_i^2(\sigma_a^2)[(\bar{y}_{i.} - \mu)^2 - s_i^2/n_i]}{\sum_{i=1}^{k} w_i^2(\sigma_a^2)} \quad (7.43)$$

The restricted likelihood estimate of σ_a^2 is found numerically by iterating

$$\sigma_a^2 = \frac{\sum_{i=1}^k w_i^2(\sigma_a^2)[(\bar{y}_{i.} - \hat{\mu}(\sigma_a^2))^2 - s_i^2/n_i]}{\sum_{i=1}^k w_i^2(\sigma_a^2)} + \frac{1}{\sum_{i=1}^k w_i(\sigma_a^2)} \tag{7.44}$$

7.3.4 Confidence intervals for σ_a^2

We now present some very latest work due to Hartung and Knapp (2005) on the confidence interval of σ_a^2. This is based on a suitable quadratic form of the group means and exploiting solutions of some non-linear equations along with a convexity argument. The appproach is on a quadratic form of the group means which can be used for estimating the variance of the weighted least squares estimator (WLSE) of the overall mean without bias. For known variance components this quadratic form is exactly distributed as a multiple of a χ^2–distributed random variable, see Hartung (1999) and section 7.4.

To begin, let us recall again the one–way random effects model with unequal or heterogeneous error variances, which is given by

$$y_{ij} = \mu + a_i + e_{ij}\ ,\ i = 1, \ldots, k > 1\ ; j = 1, \ldots, n_i > 1\ , \tag{7.45}$$

where, as before, y_{ij} denotes the observable random variable, μ the overall mean, and a_i and e_{ij} are unobservable mutually stochastically independent random variables which are normally distributed with mean 0 and variance $\sigma_a^2 \geq 0$ and $\sigma_i^2 > 0$, $i = 1, \ldots, k$, respectively. Furthermore, let $N = \sum_{i=1}^k n_i$ denote the total number of observations.

It is obvious that within each group the arithmetic mean of the observations, $\bar{y}_{i.} = \sum_{j=1}^{n_i} y_{ij}/n_i$, is an unbiased estimator of μ with variance $\sigma_a^2 + \sigma_i^2/n_i$ in model (7.45) and the weighted least squares estimator (WLSE) of the overall mean is given by

$$\hat{\mu} = \sum_{i=1}^k \frac{w_i\, \bar{y}_{i.}}{w_\Sigma} \tag{7.46}$$

with $w_i = (\sigma_a^2 + \sigma_i^2/n_i)^{-1}$, $w_\Sigma = \sum_{i=1}^k w_i$. The WLSE $\hat{\mu}$ is normally distributed with mean μ and variance $1/w_\Sigma$.

As the estimate of the within group variance σ_i^2 we always consider the unbiased estimator

$$\hat{\sigma}_i^2 = \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2/(n_i - 1)\,, \tag{7.47}$$

and in model (7.45) it holds that $(n_i - 1)\,\hat{\sigma}_i^2/\sigma_i^2$ is χ^2–distributed with $(n_i - 1)$ degrees of freedom and, furthermore, the estimator $\hat{\sigma}_i^2$ is stochastically independent of the group mean $\bar{y}_{i.}$.

The approach of Hartung and Knapp (2005) is based on the quadratic form Q of the group means, defined by

$$Q = \sum_{i=1}^{k} w_i \, (\bar{y}_{i\cdot} - \hat{\mu})^2 \tag{7.48}$$

which is χ^2–distributed with $(k-1)$ degrees of freedom. We note that the quadratic form Q given above contains the usually unknown true variance components σ_a^2 and σ_i^2, $i = 1, \ldots, k$. Let us first replace the within–group variances σ_i^2 by their unbiased estimates $\hat{\sigma}_i^2$ from (7.47). So, we obtain the quadratic form

$$\widetilde{Q}(\sigma_a^2) = \sum_{i=1}^{k} \tilde{w}_i \left(\bar{y}_{i\cdot} - \hat{\tilde{\mu}} \right)^2 \tag{7.49}$$

with $\tilde{w}_i = (\sigma_a^2 + \hat{\sigma}_i{}^2/n_i)^{-1}$, $\hat{\tilde{\mu}} = \sum_{i=1}^{k} \tilde{w}_i \, \bar{y}_{i\cdot}/\tilde{w}_\Sigma$, $\tilde{w}_\Sigma = \sum_{i=1}^{k} \tilde{w}_i$. Since $\hat{\sigma}_i^2$ are unbiased estimators of σ_i^2 and stochastically independent of the group means $\bar{y}_{i\cdot}$ in model (7.45), it follows that the weights $\tilde{w}_i$ are consistent estimators of the weights w_i. By considering the first two moments of $\widetilde{Q}(\sigma_a^2)$ we suggest, as a first step, approximating the distribution of $\widetilde{Q}(\sigma_a^2)$ by a χ^2–distribution with $(k-1)$ degrees of freedom. The derivation of the first two moments of $\widetilde{Q}(\sigma_a^2)$ as well as a discussion of the approximation is given Hartung and Knapp (2005).

Hartung and Knapp (2005) show that $\widetilde{Q}(\sigma_a^2)$ is a monotone decreasing function in σ_a^2 and, thus, propose a $(1-\alpha)$–confidence region for the among–group variance defined by

$$C_1(\sigma_a^2) = \left\{ \sigma_a^2 \geq 0 \;\middle|\; \chi^2_{k-1;\alpha/2} \leq \widetilde{Q}(\sigma_a^2) \leq \chi^2_{k-1;1-\alpha/2} \right\} \tag{7.50}$$

where $\chi^2_{\nu;\kappa}$ denotes the κ–quantile of the χ^2–distribution with ν degrees of freedom.

Since $\widetilde{Q}(\sigma_a^2)$ is a monotone decreasing function in $\sigma_a^2 \geq 0$ the function $\widetilde{Q}(\sigma_a^2)$ has its maximal value at $\widetilde{Q}(0)$. For $\widetilde{Q}(0) < \chi^2_{k-1;\alpha/2}$ we define $C_1(\sigma_a^2) = \{0\}$, otherwise the confidence region $C_1(\sigma_a^2)$ is a real interval. Note that the validity of the inequality $\widetilde{Q}(0) < \chi^2_{k-1;\alpha/2}$ only depends on the choice of α.

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

$$\begin{aligned} &\text{lower bound:} && \widetilde{Q}(\sigma_a^2) = \chi^2_{k-1;1-\alpha/2} \\ &\text{upper bound:} && \widetilde{Q}(\sigma_a^2) = \chi^2_{k-1;\alpha/2} \end{aligned} \tag{7.51}$$

This can be easily done, for instance, by using the bisection method.

Likelihood based confidence intervals have been proposed by Hardy and Thompson (1996) and Biggerstaff and Tweedie (1997). We omit the details here.

7.4 Inference about μ

In the last section of this lection, we present some results on estimation, tests and confidence intervals of the overall mean μ.

Let us recall that for the one-way random effects model, $\hat{\mu}_i = \bar{y}_i \sim N(\mu, \sigma_a^2 + \sigma_i^2/n_i)$. Then the standard estimator of μ is given by

$$\hat{\mu} = \frac{\sum_{i=1}^{k} \frac{1}{\hat{w}_i} \cdot \hat{\mu}_i}{\sum_{i=1}^{k} 1/\hat{w}_i}, \tag{7.52}$$

where $\hat{w}_i = \hat{\sigma}_a^2 + \hat{\sigma}_i^2/n_i$, $i = 1, \ldots, k$. Therefore, we have the commonly used test statistic

$$Z = \frac{\hat{\mu}}{(\sum_{i=1}^{K} 1/\hat{w}_i)^{-1/2}} \overset{approx}{\sim} N(0,1) \tag{7.53}$$

In the above, the within study variances σ_i^2 are estimated by their sample counterparts, and the between study variance σ_a^2 is usually estimated by the so-called DerSimonian and Laird (1986) estimate. As is well known, in small samples, which is mostly the case in applications, this test suffers from the same weaknesses as its fixed effects counterpart. Namely, the test is anticonservative, that means it yields too many unjustified significant results.

Several modifications of the above normal test have been suggested in the literature (Hartung, 1999; Hartung and Knapp; 2001a ,b; Sidik and Jonkman, 2002; Hartung, Böckenhoff and Knapp, 2003). We should mention that most of the modifications are very similar. We present below some results from Hartung and Knapp (2001a, b).

The basic results for the improved test are that the quadratic form

$$Q = \sum_{i=1}^{k} w_i \, (\bar{y}_{i\cdot} - \hat{\mu})^2 \tag{7.54}$$

is a χ^2-distributed random variable stochastically independent of $\hat{\mu}$ and that

$$\widehat{\mathrm{var}}(\hat{\mu}) = \frac{1}{k-1} \frac{Q}{\sum_{i=1}^{k} w_i} \tag{7.55}$$

is an unbiased estimator of the variance of $\hat{\mu}$. Consequently, under $\mathrm{H}_0 : \mu = 0$,

$$T = \frac{\hat{\mu}}{\widehat{\mathrm{var}}(\hat{\mu})} \tag{7.56}$$

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

Hartung and Knapp (2001a, b) conducted an extensive simulation study to compare the attained type I error rates for the commonly used test statistic Z from 7.53 and the proposed modified test statistic according (7.56). It turns out that the proposed test greatly improves the attained type I error rate. Moreover, the good performance of the proposed test does not depend on the choice of the the between group variance estimator.

An exact test for μ in the present model is described in Iyer et al. (2004), using the notion of the generalized P-value. However, the solution is rather complicated and we omit the details.

8 Publication Bias and Vote Counting Procedures

In this lecture we discuss two *new* concepts in statistical meta analysis: publication bias and vote counting procedures. Both are relevant when we have incomplete information about either the existing literature on the subject of our study or about details of the studies which are available.

8.1 Publication Bias

As mentioned in the introduction, if a meta-analyst is restricted only to the published studies, then there is a risk that it will lead to biased conclusions because there may be many nonsignificant studies which are often unpublished and hence are ignored, and it is quite possible that their combined effect, significant and nonsignificant studies together, may change the overall conclusion. Publication bias thus results from ignoring unavailable nonsignificant studies and this is the familiar *file-drawer* problem.

A general principle is that one ought to perform a preliminary analysis to assess the chances that publication bias could be playing a role in the selection of studies before the component studies are assembled for the meta analysis purpose. This assessment can be done informally by using what is known as a *funnel graph*, which is merely a plot of the sample size (sometimes the standard error) versus the effect size of the k studies (Light and Pillemer, 1984). If no bias is present, this plot would look like a funnel, with the spout pointing up. This is because there will be a broad spread of points for the highly variable small studies (due to a small sample size) at the bottom and decreasing spread as the sample size increases, with the indication that publication bias is unlikely to be a factor for this meta analysis.

Referring to the **Data Set 1** and the resulting graph, *Figure 8.1*, we observe a funnel graph consistent with the pattern mentioned above. On the other hand, for the **Data Set 2** and the associated graph, *Figure 8.2*, the contrast is clear. The large studies at the top (with small standard errors) are clustered around the null value while the small studies at the bottom (with large standard errors) show a positive effect, suggesting that there could be a number of small studies with positive effects, which might remain unpublished. We refer to Begg (1994) for details.

Figure 8.1. Funnel plot for validity correlation studies (data set 1)

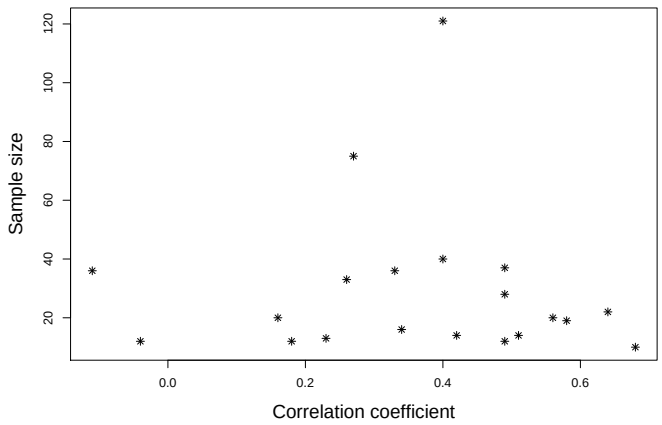
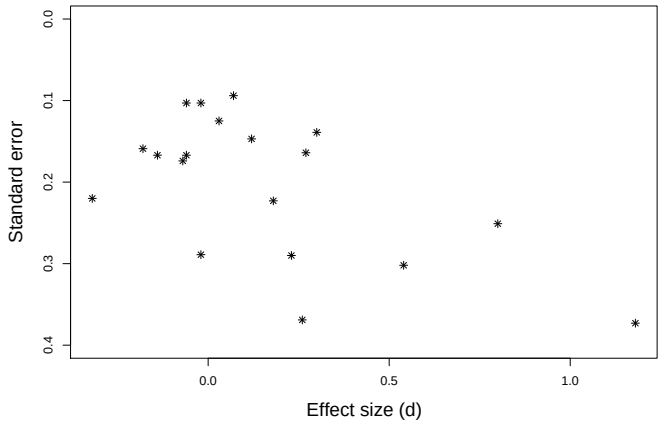


Figure 8.2. Funnel plot for teacher expectancy studies (data set 2)



There are two general strategies to deal with publication bias: *sampling methods* and *analytic methods*. Sampling methods are designed to eliminate publication bias as far as possible by directly addressing the manner in which the studies are selected for inclusion in the meta analysis, and attempting by all reasonable means to get hold of relevant unpublished studies on the topic. This method, which has been advocated by Peto and his colleagues at Oxford (see Collins et al., 1987), strongly suggests following up on published abstracts on the particular topic and contacting leading researchers in the field for leads on relevant studies being conducted worldwide with the hope that such an attempt would reveal many or some hitherto unpublished nonsignificant articles. The

criticism of this method is that accuracy of some of these sought-after studies may be questionable and also the quality of some of these studies may not be acceptable.

The second method, the well known *file-drawer method* (Rosenthal, 1979), is designed to provide a simple qualification on a summary P value from a meta analysis. Assume that the meta analysis of k available studies leads to a significant result, i.e., the combination of k P values by one of the methods described earlier leads to rejection of H_0. Recall the method of computation of the P values described in Lecture 3, and also that small P values lead to a significance of the null hypothesis, i.e., rejection of H_0. We are then wondering if a set of k_0 *nonsignificant* studies, which remain unpublished and hence unknown to us, would have made a difference in the overall conclusion, i.e., would have made rejection of H_0 on the basis of all the $k + k_0$ studies impossible. The *file-drawer* method provides a technique to get some idea about k_0. Once such a value of k_0 is determined, we then use our judgement to see if so many nonsignificant studies on the particular problem under consideration could exist!

Suppose we have used Stouffer's (1949) *normal* method to combine the k P values. This method suggests that we first convert the individual P values, $P_1, \cdots, P_k$ of the k published studies to normal Z scores, $Z_1, \cdots, Z_k$, defined by

$$Z_i = \Phi^{-1}(P_i), \;\; i = 1, \cdots, k \tag{8.1}$$

and then use the overall Z defined by

$$Z = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} Z_i \tag{8.2}$$

to test the significance of H_0. Since Z behaves like $N(0, 1)$ under H_0 and H_0 is rejected for small values of Z, the assumed rejection of H_0 at the significance level α essentially implies that $Z < -z_\alpha$, or, $|Z| > z_\alpha$. To determine a plausible value of k_0, we assume that the average observed effect of the k_0 unpublished (or unavailable) studies is 0, i.e., the sum of the Z scores corresponding to these k_0 studies is 0. Under this assumption, even if these k_0 studies were available, the value of the combined sum of all the Z_i's remains the same as before (i.e., $\sum_{i=1}^{k} Z_i = \sum_{i=1}^{k+k_0} Z_i$). Therefore, this combined sum would have led to the acceptance of H_0, thus reversing our original conclusion, if

$$\frac{1}{\sqrt{k + k_0}} \left| \sum_{i=1}^{k} Z_i \right| < z_\alpha, \tag{8.3}$$

which happens if

$$k_0 > -k + (\sum_{i=1}^{k} Z_i)^2/(z_\alpha)^2. \tag{8.4}$$

The above equation provides us with an idea about the number k_0 of unpublished studies with *nonsignificant* conclusions which, when combined with the results of k published studies, would have made a difference in the overall conclusion. The rationale behind the method is that, considering the relevant research domain, if k_0 is judged to be sufficiently large, it is unlikely that so many unpublished studies exist, and hence we can conclude that the significance of the observed studies is not affected by publication bias.

Example 8.1. We can apply this method to the **Data Set 2** dealing with **Teacher-Expectancy Studies**. The individual z scores are computed by dividing the effect sizes by the corresponding standard errors, details of which appear in Lecture 3. This leads to $\sum_{i=1}^{19} Z_i = 10.615$, which yields $Z = 2.435$, a significant value at the 5% level. Using the formula given above, we find that $k_0 \geq 22$. This means if there are at least 22 unpublished *nonsignificant* studies, then the conclusion obtained by ignoring them would have been wrong. The plausibility of the existence of so many unpublished studies is of course a judgement call, and would depend on the search technique used by the meta-analyst.

Example 8.2. We can also apply this method to the **Data Set 1** dealing with **Validity Correlation Studies**. In this case the individual z scores are computed from P values, which in turn are obtained from the t values. This leads to $\sum_{i=1}^{20} Z_i = 36.632$, which yields $Z = 8.191$, a highly significant value at the 5% level. Again, using the formula given above, we find that $k_0 \geq 476$. This means if there are at least 476 unpublished *nonsignificant* studies, then the conclusion obtained by ignoring them would have been wrong. The plausibility of the existence of so many unpublished *nonsignificant* studies seems very remote, and we can therefore conclude that publication bias is unlikely to make a difference in this problem.

Remark 8.1 If the meta analysis of k available studies leads to a *nonsignificant* conclusion, then of course the issue of publication bias does *not* arise!

The advantage of the file-drawer method is that it is very simple and easily interpretable. A disadvantage is the assumption that the results of the missing studies are centered on the null hypothesis.

More sophisticated methods for adjusting the meta analysis for publication bias have been developed using weighted distribution theory (Patil and Rao, 1977) and a Bayesian data-augmentation approach (Givens et al., 1997). These methods lead to a much more

complicated analysis, and are omitted. We also refer to Iyengar and Greenhouse (1988) for some related results.

8.2 Vote Counting Procedures

We now describe the method of vote counting procedures which is used when we have scanty information from the studies to be combined for statistical meta analysis. The nature of data from primary research sources which are available to a meta-analyst generally falls into three broad categories: (*i*) complete information (e.g., raw data, summary statistics) that can be used to calculate relevant effect size estimates such as means, proportions, correlations, test statistic values, (*ii*) results of hypothesis tests for population effect sizes about statistically significant or nonsignificant relations, and (*iii*) information about the direction of relevant outcomes (i.e., conclusions of significant tests) without their actual values (i.e., without the actual values of the test statistics).

Vote-counting procedures are useful for the second and third types of data, i.e., when complete information about the results of primary studies are *not* available in the sense that effect size estimates cannot be calculated. In such situations often the information from a primary source is in the form of a report of the decision obtained from a significance test (i.e., significant positive relation or nonsignificant positive relation), or in the form of a direction (positive or negative) of the effect without regard to its statistical significance. In other words, all is known is whether a test statistic exceeds a certain critical value at a given significance level (such as $\alpha^* = 0.05$), or if an estimated effect size is positive or negative, which amounts to the observation that the test statistic exceeds the special critical value at significance level $\alpha^* = 0.5$. Actual values of the test statistics are not available.

To fix ideas, recall that often a meta-analyst is interested in determining whether a relation exists between an independent variable and a dependent variable for each study, i.e., whether the effect size is zero for each study. Let $T_1, \cdots, T_k$ be independent estimates from k studies of the corresponding population effect sizes $\theta_1, \cdots, \theta_k$ (i.e., difference of two means, difference/ratio of two proportions, difference of two correlations or z values). Under the assumption that the population effect sizes are equal, i.e., $\theta_1 = \cdots = \theta_k = \theta$, the appropriate null and alternative hypotheses are: $H_0 : \theta = 0$ (no relation) against $H_1 : \theta > 0$ (relation exists). The test rejects H_0 if an estimate T of the common effect size θ, when standardized, exceeds the one-sided critical value ψ_α. Typically, in large samples, one invokes the large sample approximation of the distribution of T, resulting in the normal distribution of T, and we can then use $\psi_\alpha = z_\alpha$, the cut-off point from a standard normal distribution. On the other hand, if a $100(1-\alpha)\%$ level confidence interval for θ is desired, it is usually provided by $T - \psi_{\alpha/2} SE(T) \leq \theta \leq T + \psi_{\alpha/2} SE(T)$

where $SE(T)$ is the (estimated, if necessary) standard error of T. Quite generally, the standard error $S(\theta)$ of T will be a function of θ and can be estimated by $SE(T)$, and a normal approximation can be used in large samples. We refer to Lecture 4 for details.

When the individual estimates $T_1, \cdots, T_k$ as well as their (estimated) standard errors $SE(T_1), \cdots, SE(T_k)$ are available, the solutions to these testing and confidence interval problems are trivial (as discussed in previous lectures). However, the essential feature of vote-counting procedure is that the values of $T_1, \cdots, T_k$ are *not* observed, and hence none of the estimated standard errors of the T_i's is also available. What is known to us is not the exact values of the T_i's, but just the number of them which are positive or how many of them exceed the one-sided critical value ψ_{α^*}. The question then arises if we can test $H_0 : \theta = 0$, or estimate the common effect size θ based on just this very incomplete information.

The sign test, which is the oldest of all nonparametric tests, can be used to test the hypothesis that the effect sizes from a collection of k independent studies are all zero when only the signs of estimated effect sizes from the primary sources are known. If the population effect sizes are all zero, the probability of getting a positive result for the estimated effect size is 0.5. If, on the other hand, the *treatment* has an effect, the probability of getting a positive result for the estimated effect size is greater than 0.5. Hence, the appropriate null and alternative hypotheses can be described as

$$H_0 : \pi = 0.5 \ \ vs. \ \ H_1 : \pi > 0.5 \tag{8.5}$$

where π is the probability of a positive effect size in the population. The test can be carried out in the usual fashion based on a *Binomial* distribution, and rejects H_0 if X/k exceeds the desired level of significance where X is the number of studies out of a total of k studies with positive estimated effect sizes.

Example 8.3. Suppose that a meta-analyst finds exactly 10 positive results in 15 independent studies. The estimate of π is $p = 10/15 = 0.67$, and the corresponding tail area from the binomial table is 0.1509. Thus, we would fail to reject H_0 at the 0.05 overall significance level or even at the 0.10 overall significance level. On the other hand, if exactly 12 of the 15 studies had positive results, the tail area would become 0.0176, and we would reject H_0 at the 0.05 overall level of significance.

The main criticism against the sign test is that it does not take into account the sample sizes of the different studies, which are likely to be unequal, and also it does not provide an estimate of the underlying common effect size θ, nor does it provide a confidence interval for the common effect size. Under the simplifying assumption that each study in a collection of k independent studies has an identical sample size n, we now describe a procedure to establish a point estimate as well as a confidence interval for the common effect size θ based on a knowledge of the number of positive results. If a study involves

an experimental (E) as well as a control (C) group, we assume that the sample sizes for each such group are the same, i.e., $n_i^E = n_i^C = n$ for all k studies. In case k studies have different sample sizes, we may use an *average* value, namely,

$$\bar{n} = [\frac{\sqrt{n}_1 + \cdots + \sqrt{n}_k}{k}]^2. \tag{8.6}$$

Based on a knowledge of the signs of T_i's, an unbiased estimate of π is given by $p = X/k$ where X is the number of positive T_i's. It is also well known that a $100(1-\alpha)\%$ level approximate confidence interval for π (based on the normal approximation) is given by

$$\pi_L = p - z_{\alpha/2}\sqrt{\frac{p(1-p)}{k}} \leq \pi \leq p + z_{\alpha/2}\sqrt{\frac{p(1-p)}{k}} = \pi_U \tag{8.7}$$

where $z_{\alpha/2}$ is the two-sided critical value of the standard normal distribution. A second method uses the fact that

$$z^2 = \frac{k(p-\pi)^2}{\pi(1-\pi)} \tag{8.8}$$

has an approximate chi-square distribution with 1 *df*, which leads to the two-sided interval

$$\pi_L = \frac{(2p+b) - \sqrt{b^2 + 4bp(1-p)}}{2(1+b)} \leq \pi \leq \frac{(2p+b) + \sqrt{b^2 + 4bp(1-p)}}{2(1+b)} = \pi_U \tag{8.9}$$

where $b = \chi^2_\alpha(1)/k$ and $\chi^2_\alpha(1)$ is the upper $100\alpha\%$ point of the chi-square distribution with 1 *df*.

Once a two-sided confidence interval $[\pi_L, \pi_U]$ has been obtained for π, a two-sided confidence interval for θ can be constructed by using the relation

$$\begin{aligned} \pi &= Pr[T > \psi_\alpha] \\ &= Pr[(T-\theta)/S_\theta > (\psi_\alpha - \theta)/S_\theta] \\ &\sim 1 - \Phi[(\psi_\alpha - \theta)/S_\theta] \end{aligned} \tag{8.10}$$

where $\Phi(.)$ is the standard normal *cdf*. Solving the above equation yields

$$\theta = \psi_\alpha - S(\theta)\Phi^{-1}(1-\pi) \tag{8.11}$$

which provides a relation between the effect size θ and the population proportion π of a positive effect size. A point estimate of θ is then obtained by replacing π by $p = X/k$ in

the above equation and solving for θ. To obtain a two-sided confidence interval for θ, we substitute π_L and π_U for π, and solve for the two bounds for θ.

Example 8.4. Let us consider the case when an effect size is measured by the standardized mean difference given by

$$\theta_i = \frac{\mu_i^E - \mu_i^C}{\sigma_i}, \;\; i = 1, \cdots, k \tag{8.12}$$

where μ_i^E is the population mean for the experimental group in the ith study, μ_i^C is the population mean for the control group in the ith study, and σ_i is the population standard deviation in the ith study, which is assumed to be the same for the experimental and the control groups. The corresponding estimates T_i's are given by (Hedges's g)

$$T_i = \frac{\bar{y}_i^E - \bar{y}_i^C}{s_i}, \;\; i = 1, \cdots, k \tag{8.13}$$

where $\bar{y}_i^E$ is the sample mean for the experimental group in the ith study, $\bar{y}_i^C$ is the sample mean for the control group in the ith study, and s_i is the pooled within group sample standard deviation in the ith study. In large samples, the approximate variance $S_i(\theta_i)$ of T_i is given by

$$var(T_i) = S_i(\theta_i) \sim \frac{2}{n} + \frac{\theta_i^2}{4n} \tag{8.14}$$

where n denotes the common sample size for all the studies. The equation (7.11) in this case then reduces to

$$\theta = \psi_\alpha - [\frac{2}{n} + \frac{\theta^2}{4n}]\Phi^{-1}(1-\pi) \tag{8.15}$$

For the **Data Set 2**, n is approximated as 84, and the estimate of π based on the proportion of *positive* results is $p = 11/19 = 0.579$. Solving for θ, using $\psi_\alpha = 0$, we obtain $\hat{\theta} = 0.032$, which is the proposed point estimate of the population effect size. To obtain 95% confidence interval for θ, we note that the same for π based on the equation (8.7) is $[0.357, 0.801]$ and that based on the equation (8.9) is $[0.363, 0.769]$, which is slightly narrower. Using these latter values in (8.15), we find that the 95% confidence interval for θ is given by $[-0.056, 0.121]$. Since this confidence interval contains the value 0, we conclude that we can accept the null hypothesis that the population effect size is 0 for all the studies.

For the same **Data Set 2**, we can also obtain a point estimate and a confidence interval for θ based on the proportion of *significant positive* results. Since 3 of the 19 studies result in statistically significant values at $\alpha = 0.05$, with the corresponding value

of $\psi_\alpha = 1.64$, our estimate of π is $p = 3/19 = 0.158$, and this results in the point estimate of θ as $\hat{\theta} = 0.013$. Again, the confidence interval for π based on the normal theory is obtained as $[-0.006, 0.322]$, and the same based on the chi-square distribution is given by $[0.055, 0.376]$. Using the latter bounds and the equation (8.15), we obtain the 95% confidence bounds for θ as $[0.032, 0.212]$. Since this interval does *not* contain 0, we can conclude that the common effect size θ is significantly greater than 0.

Example 8.5. We next consider the situation when both the variables X and Y are continuous, and a measure of effect size is provided by the correlation coefficient ρ. Typically, the population correlation coefficients $\rho_1, \cdots, \rho_k$ of the k studies are estimated by the sample correlation coefficients $r_1, \cdots, r_k$, which represent the θ_i's and the T_i's, respectively. It is well known that, in large samples, $var(r_i) \sim (1 - \rho_i^2)^2/(n-1)$ where n is the sample size. We thus have all the ingredients to apply the formula (8.11) to any specific problem.

As an example, we consider the **Data Set 1** and suppose we wish to obtain a point estimate and a confidence interval for ρ, the assumed common population correlation, based on the proportion of positive results. Obviously, here $p = 18/20 = 0.9$, and, using (8.6), $\bar{n} = 26$. Taking $\psi_\alpha = 0$, we then get $\hat{\rho} = 0.264$ as the point estimate of ρ. The 95% approximate two-sided confidence interval for ρ based on the normal theory is given by $[0.769, 1.031]$ while that based on the chi-square theory is obtained as $[0.699, 0.972]$. Using the latter, the confidence bounds for ρ turn out as $[0.107, 0.381]$. Because this interval does not contain the value 0, we can conclude that there is a positive correlation between student ratings of the instructor and the student achievement.

For the same **Data Set 1**, we can proceed to obtain point estimate and confidence bounds for ρ based on only *significantly* positive results. Taking $\alpha = 0.05$, so that $\psi_\alpha = 1.64$, and noting that $p = 12/20 = 0.6$, we obtain the point estimate of ρ as $\hat{\rho} = 0.372$. Similarly, using the chi-square-based confidence interval for π, namely, $[0.387, 0.781]$, the bounds for ρ are obtained as $[0.271, 0.464]$, leading to the same conclusion.

9 Combination of Polls

The basic motivation of this lecture, which is taken from Dasgupta and Sinha (2006), essentially arises from an attempt to understand various poll results conducted by several competing agencies and to meaningfully combine such results. As an example, consider the 1996 USA presidential election poll results, which are reproduced below and reported in leading newspapers back then, regarding several presidential candidates.

Table 9.1. Gallup Poll Results
August 18-20, 1996

	President Clinton	Robert Dole	Others
ABC-W.Post	44	40	16
Newsweek	44	42	14
CNN-Gallup	48	41	11
CBS-NY Times	50	39	11

It is clear that depending on which poll one looks at, the conclusion in terms of margin of variation can be different, sometimes widely. Similar phenomena exist in various other contexts such as results of a series of studies comparing different brands of cereals, TV ratings, ratings of athletes by different judges, and so on. In studies of this type, one is bound to observe different levels of margin of variation between two suitably selected leading *candidates*, and one often wonders how much variation between polls would be considered as *normal*, *i.e.*, can be attributed to chance! Clearly, such a question does not arise had there been only one study, and in such a situation one could apply standard statistical techniques to estimate the margin of difference as well as test relevant hypotheses about the margin of difference. In the presence of several *independent* studies all with a common goal, what is needed is a data fusion or data synthesis technique, which can be used to meaningfully combine results of all the studies in order to come up with efficient inference regarding parameters of interest. It is the purpose of this chapter to describe appropriate statistical methods to deal with the above problems.

A general mathematical formulation of the problem involving k *candidates* and m *polls* (judges) is given in section 2. This section also provides a solution to the problem posed earlier, namely, how much variation between two selected candidates can be expected as *normal*.

The major issue of combining polls is addressed in section 3 which has a few subsections. Subsection 3.1 is devoted to estimation of the difference θ between the true proportions P_1 and P_2 of two selected candidates. Subsection 3.2 deals with providing a

confidence interval for θ, and lastly subsection 3.3 is concerned with a test for the significance of θ. All throughout, whenever applicable, we have discussed relevant asymptotics with applications.

We have also discussed the special case of two candidates (i.e., $k = 2$) and noted that often this leads to amazingly simple results.

9.1 Formulation of the problem

Assume that m independent polls are conducted to study the effectiveness of k candidates, and the following results are obtained.

Table 9.2. General Set Up

Study	Subject 1	Subject 2	...	Subject k	Total
1	X_{11}	X_{12}	...	X_{1k}	n_1
2	X_{21}	X_{22}	...	X_{2k}	n_2
.	...	...	...	...	...
.	...	...	...	...	...
.	...	...	...	...	...
m	X_{m1}	X_{m2}	...	X_{mk}	n_m
Total	$X_{.1}$	$X_{.2}$	...	$X_{.k}$	N

Denoting by X_{ij} the number of votes received by the jth candidate (subject) in the ith poll (study), so that $\sum_{j=1}^{k} X_{ij} = n_i,\ i = 1, \cdots, m$, it follows readily that $(X_{i1}, \cdots, X_{ik})$ follows a multinomial distribution with the parameters n_i and $(P_{i1}, \cdots, P_{ik})$ with $\sum_{j=1}^{k} P_{ij} = 1$ for all i. The underlying probability structure is thus essentially m independent multinomials each with k classes and possibly unequal sample sizes $n_1, \cdots, n_m$. Here P_{ij} denotes the chance that a response in the ith study belongs to the jth subject. Clearly, an unbiased estimate of P_{ij} is provided by $p_{ij} = X_{ij}/n_i,\ i = 1, \cdots, m$, $j = 1, \cdots, k$. Moreover, it is well known that

$$\begin{aligned} E(p_{ij}) &= P_{ij} \\ var(p_{ij}) &= \frac{P_{ij}(1 - P_{ij})}{n_i} \\ cov(p_{ij}, p_{ij'}) &= -\frac{P_{ij}P_{ij'}}{n_i},\ \ j \neq j'. \end{aligned} \tag{9.1}$$

Although there are k candidates, quite often we are interested in only two of them, namely, the two leading candidates such as a sitting candidate and a close runner-up.

Assuming without any loss of generality that we are interested in the candidates 1 and 2, poll i reports an unbiased estimate of the difference between P_{i1} and P_{i2} as $p_{i1} - p_{i2} = Y_i$, say, for $i = 1, \cdots, m$. Obviously, $Y_1, \cdots, Y_m$ are independent, but *not* identically distributed random variables. A *measure of variation* among the polls can then be taken as $Z_m = Y_{(m)} - Y_{(1)}$ where $Y_{(m)} = max(Y_1, \cdots, Y_m)$ and $Y_{(1)} = min(Y_1, \cdots, Y_m)$.

We now address the question of how much variation one should expect as *normal*. Obviously, any such measure would require us to compute at least the mean and the variance of Z_m. An exact computation of these quantities seems to be extremely complicated, and we therefore take recourse to asymptotics, which is quite reasonable since the sample sizes of the m polls are typically large in practical applications. We also assume that $P_{1j} = \cdots P_{mj} = P_j$, $j = 1, \cdots, k$, the unknown true values in the entire population. This is justified because the voters usually have a definite opinion about the candidates no matter who is conducting the gallup poll, thus making meta analysis viable and useful. Using (9.1), we then get

$$E(Y_i) = P_1 - P_2, \;\; var(Y_i) = [P_1 + P_2 - (P_1 - P_2)^2]/n_i. \tag{9.2}$$

Hence, under the assumption of a large sample size, we get

$$\sqrt{n_i}[Y_i - (P_1 - P_2)] \sim N[0, P_1 + P_2 - (P_1 - P_2)^2]. \tag{9.3}$$

Let us write $\theta = P_1 - P_2$ and $\sigma_i^2 = \frac{P_1 + P_2 - (P_1 - P_2)^2}{n_i}$. In view of independence and uniform integrability of the Y_i's, for m fixed, we readily get

$$E(Y_{(m)}) \sim E(W_m), \quad E(Y_{(1)}) \sim E(W_1) \tag{9.4}$$

where W_m is a random variable with the *cdf*

$$F_m(w) = \prod_{i=1}^{m} \Phi(\frac{w - \theta}{\sigma_i}) \tag{9.5}$$

and W_1 is a random variable having the *cdf*

$$F_1(w) = 1 - \prod_{i=1}^{m} \bar{\Phi}(\frac{w - \theta}{\sigma_i}) \tag{9.6}$$

where $\Phi(.)$ is the standard normal *cdf* and $\bar{\Phi}(.) = 1 - \Phi(.)$. Moreover,

$$Var(Y_{(m)}) \sim Var(W_m), \quad Var(Y_{(1)}) \sim Var(W_1), \quad Cov(Y_{(m)}, Y_{(1)}) \sim Cov(W_m, W_1).$$

By arguing probabilistically, the above expectations, namely, $E(W_m)$ and $E(W_1)$, can be computed without much difficulty for m up to 4, and are given below.

$$
\begin{aligned}
E(W_m|m=2) &= P_1 - P_2 + [(\frac{1}{2\pi})\{P_1+P_2-(P_1-P_2)^2\}(\frac{1}{n_1}+\frac{1}{n_2})]^{1/2} \\
E(W_1|m=2) &= P_1 - P_2 - [(\frac{1}{2\pi})\{P_1+P_2-(P_1-P_2)^2\}(\frac{1}{n_1}+\frac{1}{n_2})]^{1/2} \\
E(W_m|m=3) &= P_1 - P_2 + [(\frac{1}{8\pi})\{P_1+P_2-(P_1-P_2)^2\}]^{1/2} \\
&\quad \times([\frac{1}{n_1}+\frac{1}{n_2}]^{1/2} + [\frac{1}{n_1}+\frac{1}{n_3}]^{1/2} + [\frac{1}{n_2}+\frac{1}{n_3}]^{1/2}) \\
E(W_1|m=3) &= P_1 - P_2 - [(\frac{1}{8\pi})\{P_1+P_2-(P_1-P_2)^2\}]^{1/2} \\
&\quad \times([\frac{1}{n_1}+\frac{1}{n_2}]^{1/2} + [\frac{1}{n_1}+\frac{1}{n_3}]^{1/2} + [\frac{1}{n_2}+\frac{1}{n_3}]^{1/2}) \\
E(W_m|m=4) &= P_1 - P_2 + [(\frac{1}{2\pi})\{P_1+P_2-(P_1-P_2)^2\}]^{1/2} \\
&\quad \times \sum_i \sum_{j\neq i} \sum_{k<l}{}_{k,l\neq j,i} \sqrt{\frac{n_j}{n_i(n_i+n_j)}}\{\frac{1}{4} + \frac{1}{2\pi} sin^{-1}\sqrt{\frac{n_k n_l}{(n_i+n_k)(n_i+n_l)}}\} \\
E(W_1|m=4) &= P_1 - P_2 - [(\frac{1}{2\pi})\{P_1+P_2-(P_1-P_2)^2\}]^{1/2} \\
&\quad \times \sum_i \sum_{j\neq i} \sum_{k<l}{}_{k,l\neq j,i} \sqrt{\frac{n_j}{n_i(n_i+n_j)}}\{\frac{1}{4} + \frac{1}{2\pi} sin^{-1}\sqrt{\frac{n_k n_l}{(n_i+n_k)(n_i+n_l)}}\}
\end{aligned}
\tag{9.7}
$$

Returning to the original problem, for large sample sizes, we then get the following.

$$
\begin{aligned}
E[Y_{(m)} - Y_{(1)}|m=2] &\sim [(\frac{2}{\pi})\{P_1+P_2-(P_1-P_2)^2\}(\frac{1}{n_1}+\frac{1}{n_2})]^{1/2} \\
E[Y_{(m)} - Y_{(1)}|m=3] &\sim [(\frac{1}{2\pi})\{P_1+P_2-(P_1-P_2)^2\}]^{1/2} \\
&\quad \times([\frac{1}{n_1}+\frac{1}{n_2}]^{1/2}+[\frac{1}{n_1}+\frac{1}{n_3}]^{1/2}+[\frac{1}{n_2}+\frac{1}{n_3}]^{1/2}) \\
E[Y_{(m)} - Y_{(1)}|m=4] &\sim [\frac{2}{\pi}\{P_1+P_2-(P_1-P_2)^2\}]^{1/2} \\
&\quad \times \sum_i \sum_{j\neq i} \sum_{k<l}{}_{k,l\neq j,i} \sqrt{\frac{n_j}{n_i(n_i+n_j)}}\{\frac{1}{4}+\frac{1}{2\pi}sin^{-1}\sqrt{\frac{n_k n_l}{(n_i+n_k)(n_i+n_l)}}\}
\end{aligned}
\tag{9.8}
$$

The following table provides the values of $E[Y_{(m)} - Y_{(1)}]$ for $m = 2, 3, 4$ and for various values of P_1 and P_2 when n_i's are equal.

Table 9.3. Values of E$[Y_{(m)} - Y_{(1)}]$

m	n	$P_1 = .40, P_2 = .35$	$P_1 = .50, P_2 = .40$	$P_1 = .60, P_2 = .30$
2	$n_1 = n_2 =$500	0.04514	0.04754	0.04399
	$n_1 = n_2 =$600	0.04120	0.04340	0.04016
	$n_1 = n_2 =$700	0.03815	0.04018	0.03718
3	$n_1 = n_2 = n_3 =$500	0.06770	0.07131	0.06599
	$n_1 = n_2 = n_3 =$600	0.06180	0.06510	0.06024
	$n_1 = n_2 = n_3 =$700	0.05722	0.06027	0.05577
4	$n_1 = n_2 = n_3 = n_4 =$500	0.09027	0.09508	0.08798
	$n_1 = n_2 = n_3 = n_4 =$600	0.08241	0.08679	0.08032
	$n_1 = n_2 = n_3 = n_4 =$700	0.07629	0.08036	0.07436

Example 9.1. Returning to the data in Table 9.1, we find that $m = 4$, $n_1 = n_2 = n_3 = n_4 = 100$, $Y_1 = 4\%$, $Y_2 = 2\%$, $Y_3 = 7\%$, $Y_4 = 11\%$ so that $Y_{(4)} = 11\%$ and $Y_{(1)} = 2\%$, giving $Y_{(4)} - Y_{(1)} = 9\%$. To check if this amount of variation between polls is *normal*, we note from (9.9) that $E[Y_{(4)} - Y_{(1)}] \sim 0.2133$ when $P_1 + P_2 = .9$, $P_1 - P_2 = .08$. Thus we can conclude that, under this scenario, what we have observed can be treated as below *normal*. Again, if $P_1 + P_2 = .9$ and $P_1 - P_2 = .4$, then $E[Y_{(4)} - Y_{(1)}] \sim 0.1941$ which again suggests that the observed difference can be regarded as below *normal*. On the other hand, if $P_1 = 0.95$ and $P_2 = 0.05$, then $E[Y_{(4)} - Y_{(1)}] \sim 0.098$, implying that the observed difference can be taken as *normal*.

Remark 9.1. Assume $k = 2$ so that $P_1 + P_2 = 1$. In this case, it is interesting to observe that $E(Y_{(m)} - Y_{(1)})$ is a maximum when $P_1 = P_2$, and the above formulae simplify to the following:

$$
\begin{aligned}
E[Y_{(m)} - Y_{(1)}|m = 2] &= [(\frac{2}{\pi})(\frac{1}{n_1} + \frac{1}{n_2})]^{1/2} \\
E[Y_{(m)} - Y_{(1)}|m = 3] &= [(\frac{1}{2\pi})]^{1/2}([\frac{1}{n_1} + \frac{1}{n_2}]^{1/2} + [\frac{1}{n_1} + \frac{1}{n_3}]^{1/2} + [\frac{1}{n_2} + \frac{1}{n_3}]^{1/2}) \\
E[Y_{(m)} - Y_{(1)}|m = 4] &= \\
& [\frac{2}{\pi}]^{1/2} \sum_i \sum_{j \neq i} \sum_{k<l} {}_{k,l \neq j,i} \sqrt{\frac{n_j}{n_i(n_i + n_j)}} \{\frac{1}{4} + \frac{1}{2\pi} sin^{-1} \sqrt{\frac{n_k n_l}{(n_i + n_k)(n_i + n_l)}}\}
\end{aligned}
\tag{9.9}
$$

9.2 Meta analysis of polls

We now describe various meta analysis procedures to estimate θ, provide confidence interval for θ, and to test hypotheses about θ.

9.2.1 Estimation of θ

In this section we discuss the important issue of how to combine the results of independent polls to arrive at some meaningful conclusions. To fix ideas, referring to Table 9.2, we address the problem of combining independent estimates $p_{i1} - p_{i2}$ of $\theta = P_1 - P_2$ based on a sample of size n_i, for $i = 1, \cdots, m$. As already noted, we have assumed that the differences $P_{i1} - P_{i2}$ are the same for all i, and the parameter θ stands for the *common* population difference. Basically, there are two standard ways of combining the $(p_{i1} - p_{i2})$'s to arrive at a pooled estimate of θ. The first, popularly known as *Commentators's* estimate, is given by

$$\hat{\theta}_C = \sum_{i=1}^{m} (p_{i1} - p_{i2})/m \tag{9.10}$$

while the second, which is essentially the uniformly minimum variance unbiased estimate (*UMVUE*) and also the maximum likelihood estimate (*MLE*) based on all the data, is given by

$$\hat{\theta}_{MLE} = \sum_{i=1}^{m} n_i(p_{i1} - p_{i2})/(\sum_{i=1}^{m} n_i). \tag{9.11}$$

It may be noted that the two estimates $\hat{\theta}_C$ and $\hat{\theta}_{MLE}$ coincide when the sample sizes are all equal. Also, the computation of $\hat{\theta}_C$ does not directly require knowledge of the sample sizes and so can be readily used. Using (12.2) and independence of the m studies, we get

$$E(\hat{\theta}_C) = \theta, \;\; var(\hat{\theta}_C) = \frac{(P_1 + P_2 - \theta^2)(\sum_{i=1}^{m} 1/n_i)}{m^2} \tag{9.12}$$

$$E(\hat{\theta}_{MLE}) = \theta, \;\; var(\hat{\theta}_{MLE}) = \frac{P_1 + P_2 - \theta^2}{\sum_{i=1}^{m} n_i}. \tag{9.13}$$

It is therefore easy to verify that the efficiency (E) of $\hat{\theta}_C$ with respect to $\hat{\theta}_{MLE}$, as measured by the ratio of their variances, is given by

$$E = \frac{m^2}{(\sum_{i=1}^{m} n_i)(\sum_{i=1}^{m} 1/n_i)}, \tag{9.14}$$

which is always < 1 by the well known *AM/HM* inequality. Hence the MLE $\hat{\theta}_{MLE}$ is always preferred to $\hat{\theta}_C$.

For large m, by the strong law of large numbers (*SLLN*), one can approximate E by $E \sim 1/[E(n)E(1/n)]$. Thus, assuming that n_i is uniform over $[a, b]$, we readily get $E = \frac{2(b-a)}{(a+b)(lnb-lna)}$. In particular, choosing $[a, b] = [675, 1200]$, we get $E \sim 0.97$, which is very high. The following table provides values of E for $m = 2, 3, 4$ and various values of n_i.

Table 9.5. Values of E

m	n	E
2	$n_1 = 500, n_2 = 600$	0.9917
3	$n_1 = 400, n_2 = 500, n_3 = 600$	0.9730
4	$n_1 = 300, n_2 = 400, n_3 = 500, n_4 = 600$	0.9357

9.2.2 Confidence interval for θ

A more challenging and informative answer to provide in this context is a confidence interval for θ. This can be done on the basis of one of the following two point estimates of θ:

$$\bar{d} = \hat{\theta}_C = \sum_{i=1}^{m} (p_{i1} - p_{i2})/m \tag{9.15}$$

and

$$\hat{d} = \hat{\theta}_{MLE} = \sum_{i=1}^{m} n_i (p_{i1} - p_{i2}) / (\sum_{i=1}^{m} n_i). \tag{9.16}$$

The exact distributions of the above two estimates again are quite difficult, and asymptotics seem to be the only recourse. One can think of two kinds of asymptotics in this context: (i) m fixed and each n_i tends to ∞, and (ii) each n_i is taken as fixed while m tends to ∞. It turns out, however, that under either type of asymptotics, the same result holds, and we get (see (9.13) and (9.14))

$$\frac{m(\bar{d} - \theta)}{\sqrt{(\sum_{i=1}^{m} 1/n_i)}} \sim N[0, P_1 + P_2 - (P_1 - P_2)^2] \tag{9.17}$$

and

$$(\hat{d} - \theta) \sqrt{(\sum_{i=1}^{m} n_i)} \sim N[0, P_1 + P_2 - (P_1 - P_2)^2]. \tag{9.18}$$

It should be noted that the use of $\bar{d}$ for inference purposes for θ requires that we know the sample sizes n_i's (just as for the use of $\hat{d}$) although computation of $\bar{d}$ does not require any direct knowledge of the sample sizes. From (9.18) and (9.19), we find that $P_1 + P_2$ appears as a nuisance parameter for drawing inference about $P_1 - P_2$ unless $k = 2$ in which case $P_1 + P_2 = 1$. For $k = 2$, a two-sided confidence interval for θ based on $\hat{d}$ is easily obtained from the probability statement:

$$1 - \alpha = P[\ |\hat{d} - \theta| < z_{\alpha/2} \sqrt{\frac{1 - \theta^2}{\sum_{i=1}^{m} n_i}}\ ] \tag{9.19}$$

where $1 - \alpha$ is the level of confidence and $z_{\alpha/2}$ is the upper $\alpha/2$ cut-off point from a standard normal distribution. A straightforward computation yields the confidence bounds of θ as

$$
\begin{aligned}
LB &= \frac{N\hat{d} - z_{\alpha/2}[N + z^2_{\alpha/2} - N\hat{d}^2]^{1/2}}{N + z^2_{\alpha/2}} \\
UB &= \frac{N\hat{d} + z_{\alpha/2}[N + z^2_{\alpha/2} - N\hat{d}^2]^{1/2}}{N + z^2_{\alpha/2}}
\end{aligned} \tag{9.20}
$$

where $N = \sum_{i=1}^{m} n_i$. Analogously, for $k = 2$, a two-sided confidence interval for θ based on $\bar{d}$ is obtained from the probability statement:

$$
1 - \alpha = P[\, m|\bar{d} - \theta| < z_{\alpha/2}\sqrt{(1-\theta^2)(\sum_{i=1}^{m} 1/n_i)}\,]. \tag{9.21}
$$

This yields the confidence bounds of θ as

$$
\begin{aligned}
LB &= \frac{\bar{d}m^2N^* - z_{\alpha/2}[z^2_{\alpha/2} + m^2N^* - m^2N^*(\bar{d})^2]^{1/2}}{m^2N^* + z^2_{\alpha/2}} \\
UB &= \frac{\bar{d}m^2N^* + z_{\alpha/2}[z^2_{\alpha/2} + m^2N^* - m^2N^*(\bar{d})^2]^{1/2}}{m^2N^* + z^2_{\alpha/2}}
\end{aligned} \tag{9.22}
$$

where $N^* = 1/[\sum_{i=1}^{m} 1/n_i]$.

Example 9.2. For the polling data in Table 9.1, we compute $\hat{d} = (186 - 162)/400 = 0.06$. Taking $\alpha = 0.05$ so that $z_{\alpha/2} = 1.96$, we find that $LB = -0.038$ and $UB = 0.157$. Hence, a 95% confidence interval for θ based on $\hat{d}$ is given by $-0.038 < \theta < 0.157$. Of course, in this data set, $\hat{d} = \bar{d}$, so that the two methods provide identical confidence intervals. Finally, since this interval contains 0, we accept the null hypothesis $H_0 : \theta = 0$.

For $k > 2$, since $P_1 + P_2 < 1$, we get the same inequality as above (given in (9.20) and (9.22)) with confidence level $\geq 1 - \alpha$. Of course, any known upper bound η of $P_1 + P_2$ can also be used. Alternatively, instead of replacing $P_1 + P_2$ by an upper bound, we can estimate it based on the data by $p_1 + p_2$ where

$$
p_1 = \frac{\sum_{i=1}^{m} n_i p_{i1}}{\sum_{i=1}^{m} n_i}, \;\; p_2 = \frac{\sum_{i=1}^{m} n_i p_{i2}}{\sum_{i=1}^{m} n_i}. \tag{9.23}
$$

Then, in large samples, by Slutsky's theorem (see Rao, 1973)

$$
\frac{m[(\bar{d} - \theta)]}{[(\sum_{i=1}^{m} 1/n_i)(p_1 + p_2 - \theta^2)]^{1/2}} \sim N[0, 1] \tag{9.24}
$$

and

$$\frac{(\hat{d}-\theta)\sqrt{\sum_{i=1}^{m} n_i}}{(p_1+p_2-\theta^2)^{1/2}} \sim N[0,1]. \tag{9.25}$$

The above two results can be readily used to provide an approximate $100(1-\alpha)\%$ confidence interval for θ. These are given below.

$$\begin{aligned} LB &= \frac{N\hat{d} - z_{\alpha/2}[(p_1+p_2)(N+z^2_{\alpha/2}) - N\hat{d}^2]^{1/2}}{N+z^2_{\alpha/2}} \\ UB &= \frac{N\hat{d} + z_{\alpha/2}[(p_1+p_2)(N+z^2_{\alpha/2}) - N\hat{d}^2]^{1/2}}{N+z^2_{\alpha/2}} \end{aligned} \tag{9.26}$$

$$\begin{aligned} LB &= \frac{\bar{d}m^2N^* - z_{\alpha/2}[(p_1+p_2)(z^2_{\alpha/2}+m^2N^*) - m^2N^*(\bar{d})^2]^{1/2}}{m^2N^*+z^2_{\alpha/2}} \\ UB &= \frac{\bar{d}m^2N^* + z_{\alpha/2}[(p_1+p_2)(z^2_{\alpha/2}+m^2N^*) - m^2N^*(\bar{d})^2]^{1/2}}{m^2N^*+z^2_{\alpha/2}} \end{aligned} \tag{9.27}$$

Obviously, one can also use the variance-stabilizing transformation in the above two cases.

Example 9.3. For the same data as in Table 9.1, we compute $p_1 = 186/400 = 0.465$ and $p_2 = 162/400 = 0.405$, and hence, using (9.27), we readily obtain the 95% confidence interval for θ as $[-0.031, 0.150]$.

9.2.3 Hypothesis testing for θ

We now discuss the problem of hypothesis testing about the difference $\theta = P_1 - P_2$ based on all the data given in Table 9.3. Let us consider the problem of testing

$$H_0 : \theta \leq \delta \ \ vs \ \ H_1 : \theta > \delta \tag{9.28}$$

where $\delta \geq 0$ is a given constant. Clearly, for $k = 2$, this is a trivial problem of testing hypothesis about a single binomial proportion, and is well known. In the following, we deal with the case when $k > 2$.

One can consider two classical tests in this context, namely, an intuitive test which rejects H_0 when $p_1 - p_2$, an estimate of θ, is *large*, and the likelihood ratio test (LRT) which rejects H_0 when

$$\lambda = \frac{sup_{\theta \leq \delta} P_1^{X_{.1}} . \cdots . P_k^{X_{.k}}}{sup_{unrestricted} P_1^{X_{.1}} . \cdots . P_k^{X_{.k}}} \tag{9.29}$$

is *small*. As to the choice of $p_1 - p_2$, we can choose either $\bar{d}$ or $\hat{d}$, described in (9.16) and (9.17), respectively. In any event, the intuitive test can be carried out using standard asymptotic theory and by suitably standardizing $p_1 - p_2$ so that the test rejects H_0 when $p_1 - p_2 > c$ where c satisfies:

$$\begin{aligned} \alpha &= sup_{\theta \leq \delta} P[\frac{N^{**}\{(p_1 - p_2) - \theta\}}{\sqrt{P_1 + P_2 - \theta^2}} > \frac{N^{**}(c - \theta)}{\sqrt{P_1 + P_2 - \theta^2}}] \\ &= sup_{\theta \leq \delta} P[N(0,1) > \frac{N^{**}(c - \theta)}{\sqrt{P_1 + P_2 - \theta^2}}]. \end{aligned} \tag{9.30}$$

where N^{**} is a suitable normalizing constant. In the above, α is the level of the test. It can be shown that the *supremum* of the above probability occurs when $P_1 = \frac{1+\delta}{2}$, $P_2 = \frac{1-\delta}{2}$, so that the above equation reduces to

$$\alpha = P[N(0,1) > \frac{N^{**}(c - \delta)}{\sqrt{1 - \delta^2}}]. \tag{9.31}$$

Hence, c is readily obtained as

$$c = \delta + \frac{z_\alpha (1 - \delta^2)^{1/2}}{N^{**}}. \tag{9.32}$$

It may be noted from (9.18) and (9.19) that when $\bar{d}$ is used in place of $p_1 - p_2$, we take $N^{**} = m[N^*]^{1/2}$, while if $\hat{d}$ is used in place of $p_1 - p_2$, we take $N^{**} = N^{1/2}$. Recall that $N = \sum_{i=1}^m n_i$ and $N^* = 1/[\sum_{i=1}^m 1/n_i]$.

Example 9.4. For the data in Table 9.1, to test $H_0 : P_1 - P_2 \leq 0$ *vs.* $H_1 : P_1 - P_2 > 0$ at level 0.05, note from (9.33) that $c = 1.64/N^{**}$. Using $\hat{d} = 0.06$, and $N = 400$, we get $N^{**} = 20$ so that $c = 0.082$. We therefore accept the null hypothesis H_0.

The LRT, on the other hand, is in general highly nontrivial because of the computations involved in the numerator of λ, and we do not pursue it here.

10 Analysis of Binary Data

An important application of meta analysis is the combination of results from comparative trials with binary outcome, especially in biometry and epidemiology. Often in clinical trials or observational studies, the outcome can be generally characterized as success and failure or as positive and negative. The effect measures for binary outcome have been already introduced in Lecture 2. The meta-analytical methods described in Lecture 4 can be generally applied to the case of binary data as well as the methods of the one-way random effects model, see Lecture 7. In the first part of this lecture, we discuss some additional features of meta analysis of binary data. In the second part, we consider the natural extension, namely the meta analysis of outcomes with more than two categories or, in other words, the meta analysis of ordinal data.

10.1 Binary outcome

Recalling from Lecture 2, let π_1 and π_2 denote the population proportions of two groups, say experimental and control group. The observed frequencies on the two binary characteristics can be arranged in a (2×2)-table, see Table 10.1.

Table 10.1. Observed frequencies on two binary characteristics

	Group		
Outcome	Experimental	Control	Total
Positive	n_{11}	n_{21}	$n_{.1}$
Negative	n_{12}	n_{22}	$n_{.2}$
Total	$n_{1.}$	$n_{2.}$	$n_{..}$

10.1.1 Effects estimates

Three prominent parameters of the difference of two groups with binary outcome, namely probability difference, relative risk, and odds ratio, and their estimates have already been introduced in Lecture 2. Standard large-sample meta analysis results are summarized in Lecture 4. In this section, we discuss some properties of the estimates with emphasis on sparse data situations. Given zero cells in Table 10.1, some estimates and their variances cannot be computed.

Probability difference

The probability difference is defined as $\theta_1 = \pi_1 - \pi_2$ and can be unbiasedly estimated by the difference of the observed success probabilities

$$\hat{\theta}_1 = \frac{n_{11}}{n_{1.}} - \frac{n_{21}}{n_{2.}} \tag{10.1}$$

The unbiased estimate of the variance of (10.1) is

$$\widehat{\text{var}}(\hat{\theta}_1) = \frac{n_{11}\, n_{12}}{n_{1.}^2\, (n_{1.} - 1)} + \frac{n_{21}\, n_{22}}{n_{2.}^2\, (n_{2.} - 1)} \tag{10.2}$$

Critical data situations only occur in extreme situations, namely when $n_{11} = n_{21} = 0$ or $n_{11} = n_{1.}$ and $n_{21} = 0$ or $n_{11} = 0$ and $n_{21} = n_{2.}$. In the first case, the estimated difference is 0, in the second case the estimate is $+1$, and in the last case -1. But in all the three cases, the variance estimate is zero. Hence, the inverse of the variance is infinity and a trial with such an extreme data situation cannot be incorporated in the usual way in the meta analysis. The two extreme cases with estimates $+1$ and -1 may be only of theoretical interest. But the case of $n_{11} = n_{21} = 0$ may be of practical interest. Consider a controlled clinical trial and the number of adverse events is of interest. Especially for small sample sizes, the situation might occur that no adverse events were observed in both treatment groups.

Log–relative risk

Setting $\theta_2 = \log(\pi_1/\pi_2)$, the log–relative risk, then an estimate of θ_2 may be

$$\hat{\theta}_2 = \log\left(\frac{n_{11} \,/\, n_{1.}}{n_{21} \,/\, n_{2.}}\right) \tag{10.3}$$

However, the estimate (10.3) cannot be computed when $n_{11} = 0$ or $n_{21} = 0$. 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) discuss the proposed estimators with respect to bias and variance, and there is no optimal solution. The "optimal" solution always depends on the true, but unknown, success probabilities.

One widely used estimate is

$$\hat{\theta}_2 = \log\left(\frac{(n_{11} + 0.5) \,/\, (n_{1.} + 0.5)}{(n_{21} + 0.5) \,/\, (n_{2.} + 0.5)}\right) \tag{10.4}$$

The variance of (10.4) is estimated without bias except for terms of order $O(n^{-3})$ by

$$\widehat{\text{var}}(\hat{\theta}_2) = \frac{1}{n_{11} + 0.5} - \frac{1}{n_{1.} + 0.5} + \frac{1}{n_{21} + 0.5} - \frac{1}{n_{2.} + 0.5} \tag{10.5}$$

This variance estimate is always positive if $n_{11} \neq n_{1.}$ or $n_{21} \neq n_{2.}$. If $n_{11} \neq n_{1.}$ or $n_{21} \neq n_{2.}$, then the value 0.5 will not be added to n_{11} and n_{21} to ensure the positiveness of the variance estimate.

Log–odds ratio

Setting $\theta_3 = \log([\pi_1/(1-\pi_2)]/[\pi_1/(1-\pi_2)])$, the log–odds ratio, then an estimate of θ_3 is

$$\hat{\theta}_3 = \log\left(\frac{n_{11} \,/\, n_{12}}{n_{21} \,/\, n_{22}}\right) = \log\left(\frac{m_T\,(n_C - m_C)}{(n_T - m_T)\,m_C}\right) \tag{10.6}$$

Like in the case of the log–relative risk, the estimate (10.6) cannot be computed when there are no success or only successes in at least one group.

Again, no unbiased estimate of the log–odds ratio exists and Gart and Zweifel (1967) investigate several estimates of this parameter with respect to bias and variance.

One estimate, originally proposed by Haldane (1955), is widely used, namely

$$\hat{\theta}_3 = \log\left(\frac{(n_{11}+0.5) \,/\, (n_{12}+0.5)}{(n_{21}+0.5) \,/\, (n_{22}+0.5)}\right) = \log\left(\frac{(n_{11}+0.5)\,(n_{22}+0.5)}{(n_{12}+0.5)\,(n_{21}+0.5)}\right) \tag{10.7}$$

The variance of (10.7) is unbiasedly estimated except of terms of order $O(n^{-3})$ by

$$\widehat{\text{var}}(\hat{\theta}_3) = \frac{1}{n_{11}+0.5} + \frac{1}{n_{12}+0.5} + \frac{1}{n_{21}+0.5} + \frac{1}{n_{22}+0.5} \tag{10.8}$$

10.1.2 Homogeneity tests

Before combining the results from experiments, a test of homogeneity of treatment effects should be carried out. In experiments with binary outcome, however, the choice of the measure of treatment difference may introduce a variability between the study results. For instance, homogeneity on the risk difference scale does not in generally imply homogeneity on the log odds scale and vice versa.

The test of homogeneity is usually carried out in the framework of the fixed effects model testing the equality of the means, but the hypothesis of homogeneity can be equivalently formulated in the random effects model testing that no between-study variance is present, see Lecture 6 and 7.

The commonly used test of homogeneity in metaanalysis is Cochran's (1954) test, see Lecture 4 and 6. The test is based on a weighted least squares statistic and compares the study-specific estimates of the effect measure with an estimate of the common homogeneous effect measure. For the effect measure log odds ratio, Cochran's test can be very conservative. Consequently, this test does not have sufficient power to detect heterogeneity. However, for the effect measure probability difference, Cochran's test can be very liberal so that the null hypothesis of homogeneity is falsely rejected too often. Based on the random effects meta analysis approach, Hartung and Knapp (2004) suggest another test of homogeneity which is derived from an unbiased estimator of the variance of the common effect measure in the random effects model proposed by Hartung (1999),

see Lecture 7. Hartung and Knapp (2004) discuss both the tests of homogeneity for the two outcome measures probability difference and log odds ratio and work out some improvements with respect to level and power of their new test.

In different areas of application there exists further tests of homogeneity for binary outcome measures. For instance, Lipsitz *et al.* (1998) consider homogeneity tests for the risk difference and Liang and Self (1985) for the (logarithmic) odds ratio. We omit the details here.

10.1.3 Binomial-normal hierarchical models in meta analysis

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 for small sample sizes. When the number of successes in the treatment groups are known, that is, the observed 2×2 is given, one can make direct use of the binomially distributed number of successes. In the random effects approach this can be done in a binomial–normal hierarchical model that can be analysed within the Bayesian framework using Markov Chain Monte Carlo (MCMC) methods. Here we will only present the basic ideas of the model formulations.

Smith *et al.* (1995) first present the formulation for the log–odds ratio that is straight forward. Then Warn *et al.* (2002) also consider the binomial–normal hierarchical model for the risk difference.

All the three models have in common that the number of successes m_{Ti} and m_{Ci} are both binomially distributed with parameters n_{Ti} and p_{Ti}, and n_{Ci} and p_{Ci}, respectively, in each study i, $i = 1, \ldots, k$. Then let $\mu_i = \mathit{logit}(p_{Ci})$ be the logarithmic odds in the control group and 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, θ_i comes from a normal distribution with mean θ, the overall effect of treatment difference, and variance τ^2, the heterogeneity parameter.

Summarized we may write the binomial–normal hierarchical model for the log–odds ratio as

$$\begin{aligned} m_{Ci} &\sim \text{Bin}\,(n_{Ci}, p_{Ci}) \\ m_{Ti} &\sim \text{Bin}\,(n_{Ti}, p_{Ti}) \\ \mu_i &= \text{logit}(p_{Ci}) \\ \text{logit}(p_{Ti}) &= \mu_i + \theta_i \\ \theta_i &\sim \mathcal{N}(\theta, \tau^2) \end{aligned} \tag{10.9}$$

Note that each value of θ_i from the normal distribution yields admissible values of the success probabilities p_T and p_C.

For the log relative risk, we set $\mu_i = \log(p_{Ci})$, that is, the logarithm of success probability in the control group. Then the logarithm of the success probability in the treatment group is parameterized as $\log(p_{Ti}) = \mu_i + \theta_i$ and θ_i is the log–relative risk. Again, 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 *et al.* (2002) this is equivalent to constraining $\log(p_{Ti})$ to the interval $(-\infty, 0]$, achieved by confining θ_i to be less than $-\log(p_{Ci})$. Let θ_i^U be the minimum of θ_i and $-\log(p_{Ci})$, then θ_i^U can take any value in the range $(-\infty, -\log(p_{Ci}))$. The full model can be then summarized as

$$\begin{aligned} m_{Ci} &\sim \text{Bin}\,(n_{Ci}, p_{Ci}) \\ m_{Ti} &\sim \text{Bin}\,(n_{Ti}, p_{Ti}) \\ \mu_i &= \log(p_{Ci}) \\ \log(p_{Ti}) &= \mu_i + \min\,(\theta_i, -\log(p_{Ci})) \\ \theta_i &\sim \mathcal{N}(\theta, \tau^2) \end{aligned} \tag{10.10}$$

Finally, we consider the 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 $p_{Ti} = \mu_i + \theta_i$. Again, θ_i comes from a normal distribution with mean θ, the overall effect of treatment difference, and variance τ^2, the heterogeneity parameter. 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

$$\begin{aligned} m_{Ci} &\sim \text{Bin}\,(n_{Ci}, p_{Ci}) \\ m_{Ti} &\sim \text{Bin}\,(n_{Ti}, p_{Ti}) \\ \mu_i &= p_{Ci} \\ p_{Ti} &= \mu_i + \min\,(\max\,(\theta_i, -p_{Ci}), 1 - p_{Ci}) \\ \theta_i &\sim \mathcal{N}(\theta, \tau^2) \end{aligned} \tag{10.11}$$

For a full Bayesian analysis in the models (10.9), (10.10), and (10.11) appropriate a–priori 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, that may be also called baseline risk. A Bayesian meta analysis can be conducted using the software WinBUGS.

10.1.4 An Example

In this example, we only consider the classical meta analysis approach based on the results from Lecture 4 and 7. Hartung and Knapp (2001 b) put together the results of 13 controlled trials of a drug named cisapride compared to placebo for the treatment of non-ulcer dyspepsia. Table 10.2 contains the data and Table 10.3 the estimates of the three outcome measures probability difference, log relative risk and log odds ratio with corresponding variance estimates.

Table 10.2. Results of 13 cisapride studies
(number of successes/number of patients)

Study	Cisapride	Placebo
1	15 / 16	9 / 16
2	12 / 16	1 / 16
3	29 / 34	18 / 34
4	42 / 56	31 / 56
5	14 / 22	6 / 22
6	44 / 54	17 / 55
7	14 / 17	7 / 15
8	29 / 58	23 / 58
9	10 / 14	3 / 15
10	17 / 26	6 / 27
11	38 / 44	12 / 45
12	19 / 29	22 / 30
13	21 / 38	19 / 38

Table 10.3. Estimates of probability difference, log relative risk, and log odds ratio with estimated variances (in parentheses)

Study	Probability difference		Log relative risk		Log odds ratio	
1	0.3750	(0.0192)	0.4895	(0.0486)	2.0990	(0.9698)
2	0.6875	(0.0156)	2.1203	(0.6255)	3.3570	(1.0334)
3	0.3235	(0.0110)	0.4666	(0.0300)	1.5652	(0.3304)
4	0.1964	(0.0078)	0.2995	(0.0199)	0.8640	(0.1635)
5	0.3636	(0.0196)	0.8023	(0.1339)	1.4656	(0.4011)
6	0.5057	(0.0067)	0.9515	(0.0432)	2.2326	(0.2008)
7	0.3569	(0.0252)	0.5379	(0.0806)	1.5465	(0.6057)
8	0.1034	(0.0084)	0.2274	(0.0423)	0.4125	(0.1385)
9	0.5143	(0.0254)	1.1653	(0.2475)	2.1203	(0.6832)
10	0.4316	(0.0151)	1.0274	(0.1369)	1.8072	(0.3628)
11	0.5970	(0.0070)	1.1472	(0.0615)	2.7647	(0.2897)
12	-0.0782	(0.0143)	-0.1098	(0.0290)	-0.3544	(0.3086)
13	0.0526	(0.0131)	0.0976	(0.0458)	0.2059	(0.2062)

Applying methods of Lecture 4 and 7, we obtain for the probability difference an estimate of 0.3409 (95% CI: [0.2814; 0.4003]) assuming a fixed effects model and an estimate of 0.3380 (95% CI: [0.2026;0.4733]) in the random effects model. For the effect measure log relative risk, the estimates are 0.4422 (95% CI: [0.3197; 0.5646]) in the fixed effects model and 0.5575 (95% CI: [0.2729;0.8421])in the random effects model. Finally, the meta-analytical estimates for the log odds ratio are 1.2305 (95% CI: [0.9325; 1.5286]) in the fixed effects model and 1.4209 (95% CI: [0.7971;2.0446]) in the random effects model. Note that for the calculations of the confidence intervals in the random effects model, the improved method of Hartung and Knapp (2001 b) has been used, see the end of Lecture 7.

10.2 Ordinal outcome

The data from a controlled trial with ordinal outcome can be arranged in a $(2 \times r)$–contingency table like in Table 10.4., where r denotes the number of categories of the response variable. In Table 10.4., n_{1j} denotes the number of patients in the first group with response in the jth category and n_{2j} the corresponding number of patients in the second group. The sample sizes in both groups are $n_{1.} = \sum_{j=1}^{r} n_{1j}$ and $n_{2.} = \sum_{j=1}^{r} n_{2j}$, respectively.

Table 10.4. Data from a controlled trial with ordinal outcome

	Category				
	1	2	$\cdots$	r	Total
Group 1	n_{11}	n_{12}	$\cdots$	n_{1r}	$n_{1.}$
Group 2	n_{21}	n_{22}	$\cdots$	n_{2r}	$n_{2.}$

Let $\pi_{1j} > 0$, $j = 1, \ldots, r$, be the probability observing a response in the jth category in the firs group and $\pi_{2j} > 0$ the corresponding probability in the second group. Note that $\sum_{j=1}^{r} \pi_{1j} = \sum_{j=1}^{r} \pi_{2j} = 1$. We assume that the categories are ordered in terms of desirability: category 1 is the best and category r is the worst.

Let Y_1 denote the response variable in the first sample and Y_2 the one in the second sample, then, in view of the ordering of the categories, the treatment is superior to the control when Y_2 is stochastically larger than Y_1. If Y_2 is stochastically larger than Y_1 then it holds $P(Y_2 > Y_1) \geq P(Y_2 < Y_1)$. However, if the inequality is true then it does not necessarily follow that Y_C is stochastically larger than Y_T. In the following two subsections, we consider two effect measures that may be used to describe the difference of the response variables in a controlled trial with ordinal outcome.

10.2.1 Proportional odds model

The proportional odds model was introduced by McCullagh (1980). Consider the cumulative probabilities $q_{1j} = \sum_{i=1}^{j} \pi_{1i}$ and $q_{2j} = \sum_{i=1}^{j} \pi_{2i}$, respectively, up to category j, $j = 1, \ldots, r-1$, then the odds ratio given cut-off point category j is

$$\theta_j = \frac{q_{1j}\,(1 - q_{2j})}{(1 - q_{1j})\,q_{2j}}, \qquad j = 1, ..., m-1. \tag{10.12}$$

The proportional odds assumption reads

$$\theta_1 = \theta_2 = \cdots = \theta_{r-1} =: \theta. \tag{10.13}$$

If $\theta > 1$, then the treatment is superior to the control, in view of the above ordering of the categories. This implies that Y_2 is stochastically larger than Y_1. But, through the model assumption of proportional odds, the type how Y_2 is stochastically larger than Y_1 is restricted.

The proportional odds model can be analyzed using standard statistical software packages for linear logistic regression. These software packages usually yield the maximum likelihood estimate of the log odds ratio and the corresponding standard error. Additional remarks for the analysis in the proportional odds model can be found in Whitehead and Jones (1994). Note that the proportional odds model can be considered as arising from a latent continuous variable, where this latent variable has a logistic distribution, see Whitehead *et al.* (2001) for further details.

10.2.2 Agresti's α

Agresti (1980) proposed a measure of association, named briefly now as Agresti's α, that, in case of a $(2 \times r)$–contingence table, can be seen as a generalized odds ratio. Agresti's α is the ratio of $P(Y_C > Y_T)$ and $P(Y_C < Y_T)$, that is, in the present context, the probability to observe a worse response in the control group than in the treatment group divided by the probability to observe a better response in the control group than in the treatment group. In formula, Agresti's α can be written as

$$\alpha = \frac{\sum\limits_{j>i} p_{Ti}\, p_{Cj}}{\sum\limits_{j<i} p_{Ti}\, p_{Cj}}. \tag{10.14}$$

Note that, if Y_C is stochastically larger than Y_t than $\alpha > 1$, however, $\alpha > 1$ does not necessarily mean that Y_C is stochastically larger than Y_T. In case, Y_C is stochastically larger than Y_T, Agresti's α is a meaningful measure of the difference of all possible distributions of Y_T and Y_C. If the distributions of Y_T and Y_C are identical then $\alpha = 1$ and $\theta = 1$. However, $\alpha = 1$ does not necessarily mean that the two distribution are identical, it only means, that the two probabilities, $P(Y_C > Y_T)$ and $P(Y_C < Y_T)$, are identical. Of course, for (2×2)–tables, Agresti's α is the odds ratio.

Agresti's α is easily estimated by plugging in the observed proportions $\hat{p}_{Ti} = m_{Ti}/n_T$ and $\hat{p}_{Ci} = m_{Ci}/n_C$, $i = 1, \ldots, r$, in (10.14) and we denote this estimator by $\hat{\alpha}$. Note that $\hat{\alpha}$ does not exist when "zeros occur".

Agresti (1980) provided a large-sample estimator of the variance of the estimator of α. This variance estimator reads

$$\hat{\sigma}(\hat{\alpha}) = \left\{ \frac{1}{n_T} \sum_j \hat{p}_{Tj} \left(\hat{\alpha} \sum_{i<j} \hat{p}_{Ci} - \sum_{i>j} \hat{p}_{Ci} \right)^2 + \frac{1}{n_C} \sum_j \hat{p}_{Cj} \left(\hat{\alpha} \sum_{i>j} \hat{p}_{Ti} - \sum_{i<j} \hat{p}_{Ti} \right)^2 \right\} \Big/ \left(\sum_{i>j} \hat{p}_{Ti} \hat{p}_{Ci} \right)^2 \tag{10.15}$$

For constructing confidence intervals on Agresti's α, it is convenient to make the inference first on $\log(\alpha)$ since the distribution of $\log(\hat{\alpha})$ tends to be more symmetric and to converge faster to normality than the distribution of $\hat{\alpha}$. According to Agresti (1980), the large-sample $(1-\kappa)$ confidence interval on α is then given as $\exp\left(\log(\hat{\alpha}) \pm u_{1-\kappa/2}\, \hat{\sigma}(\hat{\alpha}) \,/\, \hat{\alpha}\right)$ with u_γ the γ-quantile of the standard normal distribution.

10.2.3 An example

For illustration purposes, we take an example from Whitehead and Jones (1994). Thirteen controlled trials were undertaken to investigate whether concurrent treatment with the synthetic prostaglandin, misoprostol, would prevent or at least reduce the degree of

gastrointestinal damage without reducing the anti–inflammatory effect of non–steroidal anti–inflammatory drugs (NSAIDs). Patients suffering from arthritis are often prescribed NSAIDs. In the trials, different scoring systems were used to assess the extent of gastrointestinal damage. The number of categories ranges from two up to five. The data with the different classification schemes are put together in Table 10.5. For a more detailed description of the trials let us refer to Whitehead and Jones (1994).

The definition of best category is different from trial to trial. However, the classification category 1 in Table 10.5 always stands for the best category in each trial. Score tests on proportional odds assumptions do not reveal any violence of this assumptions in all the trials, see Whitehead and Jones (1994).

Table 10.5. Thirteen randomized trials of misoprostol by endoscopic classification

Study	Treatment	Endoscopic classification				
		1	2	3	4	5
1	Misoprostol	21	2	4	2	0
	Placebo	2	2	4	9	13
2	Misoprostol	17	8	3	2	0
	Placebo	0	3	4	10	13
3	Misoprostol	20	4	6	0	0
	Placebo	8	4	9	4	5
4	Misoprostol	20	4	6	0	0
	Placebo	0	2	5	5	17
5	Misoprostol	1	4	5	0	0
	Placebo	0	0	0	4	6
6	Misoprostol	93	5	3	1	1
	Placebo	85	10	10	4	5
7	Misoprostol	61	12	0		
	Placebo	49	28	3		
8	Misoprostol	45	1	0		
	Placebo	65	6	3		
9	Misoprostol	138	1			
	Placebo	121	17			
10	Misoprostol	126	2			
	Placebo	110	21			
11	Misoprostol	30	1	1		
	Placebo	20	11	7		
12	Misoprostol	56	12	8	0	
	Placebo	50	15	12	5	
13	Misoprostol	12	3	1	0	
	Placebo	11	5	2	3	

In Table 10.6 the study-specific estimates are summarized along with standard errors and confidence intervals.

Table 10.6. Study-specific estimates, their standard errors, and 95% confidence intervals

	Proportional odds			Agresti's α		
Study	$\log(\hat{\theta})$	SE($\log(\hat{\theta})$)	95% CI	$\log(\hat{\alpha})$	SE($\log(\hat{\alpha})$)	95% CI
1	3.55	0.66	[2.25 ; 4.84]	3.04	0.56	[1.95 ; 4.14]
2	4.05	0.72	[2.64 ; 5.46]	3.58	0.62	[2.36 ; 4.79]
3	1.91	0.53	[0.87 ; 2.95]	1.69	0.46	[0.78 ; 2.60]
4	3.75	0.69	[2.40 ; 5.10]	3.04	0.59	[1.88 ; 4.19]
5	6.51	2.28	[2.04 ; 10.98]	4.02	1.96	[0.19 ; 7.86]
6	1.18	0.40	[0.40 ; 1.95]	1.12	0.38	[0.38 ; 1.86]
7	1.19	0.39	[0.43 ; 1.96]	1.19	0.39	[0.43 ; 1.94]
8	1.84	1.07	[-0.26 ; 3.94]	1.84	1.07	[-0.25 ; 3.92]
9	2.96	1.04	[0.93 ; 5.00]	2.96	1.04	[0.93 ; 5.00]
10	2.49	0.75	[1.01 ; 3.96]	2.49	0.75	[1.01 ; 3.96]
11	2.57	0.80	[1.00 ; 4.13]	2.37	0.78	[0.83 ; 3.91]
12	0.65	0.34	[-0.02 ; 1.31]	0.60	0.31	[-0.01 ; 1.21]
13	1.11	0.71	[-0.28 ; 2.50]	1.04	0.65	[-0.24 ; 2.31]

For both effect measures, the meta analysis in a random effects model is appropriate. The values of Cochran's homogeneity statistic are 49.49 (log odds ratio) and 42.30 (log α) leading to P values less than 0.0001. The DerSimonian-Laird estimate, see Lecture 7, is 1.1074 on the log odds ratio scale and 0.7672 on the log Agresti's α scale. The combined estimate is 2.2614 (95%CI: [1.4665;3.0561]) for the log odds ratio and 2.0160 (95%CI: [1.3906;2.6413]).

11 Computational Aspects

In this section, we consider various computational aspects of meta-analytical methods. First, we describe some methods of extracting summary statistics from publication. Then, we indicate how to conduct meta analysis using statistical software SAS and R. More details will be presented during the workshop.

11.1 Extracting summary statistics

Usually, the various publications do not deliver the same precise information on the results of trials. The ideal situation would be if each publication reported the estimate of the effect size, say $\hat{\theta}$, and its standard error, say $\hat{\sigma}(\hat{\theta})$. Whereas one can expect that $\hat{\theta}$ will in general be reported, the information on the precision of this estimate is often given indirectly.

Consider the situation in which $\hat{\theta}$ and a $100(1-\alpha)\%$ confidence interval, say $[\hat{\theta}_L; \hat{\theta}_U]$, are reported. Assuming that the confidence interval is based on (approximate) normality, that is,

$$\hat{\theta} \pm \hat{\sigma}(\hat{\theta}) z_{\alpha/2} \tag{11.1}$$

we can extract the information on the standard error by

$$\hat{\sigma}(\hat{\theta}) = \frac{\hat{\theta}_U - \hat{\theta}_L}{2\, z_{\alpha/2}} \tag{11.2}$$

In case only the estimate of the effect size in combination with a *one-sided* P value is reported we can proceed as follows. Assuming that the calculation of the P value is based on the (approximate) normal test statistic $\hat{\theta}/\hat{\sigma}(\hat{\theta})$ and large values of the test statistic are in favor of the alternative, that is,

$$P = P\left(N(0,1) > \frac{\hat{\theta}}{\hat{\sigma}(\hat{\theta})} \,\middle|\, \mathrm{H}_0 \right) \tag{11.3}$$

Then we extract the standard error as

$$\hat{\sigma}(\hat{\theta}) = \frac{\hat{\theta}}{z_{1-P}} \tag{11.4}$$

Given a *two-sided* P value and the effect size estimate $\hat{\theta}$, the standard error can be computed as

$$\hat{\sigma}(\hat{\theta}) = \frac{|\hat{\theta}|}{z_{1-P/2}} \tag{11.5}$$

since

$$P = 2\,P\left(N(0,1) > \frac{|\hat{\theta}|}{\hat{\sigma}(\hat{\theta})} \mid \mathrm{H}_0\right) \tag{11.6}$$

11.2 Combining tests

The combined test procedures can be easily calculated using standard statistical software like R or SAS. Both software packages have implemented the required function for *normal*, t, χ^2 and *beta* distributed random variables. The following table contains a summary of the syntax of the necessary functions in both software packages.

Probability and quantile functions in R and SAS

Distribution	R function	SAS function
beta	pbeta(x, a, b)	cdf('beta', x, a, b)
	qbeta(prob, a, b)	quantile('beta', x, a, b)
χ^2	pchisq(x, df)	cdf('chisquare', x, df)
	qchisq(prob, df)	quantile('chisquare', prob, df)
normal	pnorm(x, mean, sd)	cdf('normal', x, mean, sd)
	qnorm(prob, mean, sd)	quantile('normal', prob, mean, sd)
t	pt(x,df)	cdf('t', x)
	qt(prob,df)	quantile('t', prob)

Consider the six P values from example 3.6: 0.047, 0.028, 0.216, 0.062, 0.129, 0.898.

Tippett's, Stouffer's (inverse normal) and Fisher's method in R:

```
pvalues <- c(0.047, 0.028, 0.216, 0.062, 0.129, 0.898)
k <- length(pvalues)       # number of trials
# test statistics
tippett <- min(pvalues)
stouffer <- sum(qnorm(pvalues)) / sqrt(k)
fisher <- sum(-2 * log(pvalues))
# alternative calculation
fisher.2 <- sum(qchisq(1 - pvalues, 2))
# P values of the three tests
pv.tippett <- pbeta(tippett, 1, k)
pv.stouffer <- pnorm(stouffer)
pv.fisher <- 1 - pchisq(fisher, 2*k)
```

11.3 Combining effect sizes

Obtaining the statistics via weighted least-squares regression

The results for combining effect size described in Lecture 4 can be obtained by statistical software using weighted least-squares regression. Often, the general method in Lecture 4 is denote as *generic inverse variance* method.

Whitehead (2002) shows the use of SAS PROC GLM for fitting a weighted least-squares regression. For k trials, the observed responses are the study estimates, say $\hat{\theta}_i$, $i = 1, \ldots, k$, and there are no explanatory variables, only a constant term. The weights are the inverse of the estimated variances, say $w_i = 1/\widehat{var}(\hat{\theta}_i)$. The estimated intercept of this weighted least-squares regression is the estimate of the common effect size. However, as Whitehead (2002) noted, the standard error and the test statistics displayed for the intercept parameter are incorrect for the required model, because the model assumption is $var(\hat{\theta}) = \sigma^2/w_i$, where σ^2 is to be estimated from the data, instead of equal to 1. This will also be the case for other statistical packages.

Van Houwelingen et al. (2002) show the use of SAS PROC MIXED for fitting a weighted least-squares regression. Moreover, they show how to use SAS PROC MIXED for the meta-analysis in the random effects model, that is, how to implement the standard method from Lecture 7, Section 7.4. However, in the random effects model, SAS PROC MIXED computes (restricted) maximum likelihood estimates of the between-trial variance. Other estimates of the between-trial variance, like the DerSimonian-Laird estimate, are not available. Van Houwelingen also discuss advanced method in meta analysis like multivariate meta analysis and meta-regression and the implementation of these methods in SAS PROC MIXED.

R packages

There are two **R** packages in which several meta analysis methods are implemented. The packages *rmeta* by Thomas Lumley and *meta* by Guido Schwarzer provides methods for simple fixed and random effects meta analysis for two-sample comparisons and cumulative meta-analyses and computes summaries and tests for association and heterogeneity. In both packages, functions are implemented for conducting a random effects model meta analysis with the DerSimonian-Laird estimate as between-study estimate. More or less, the functions of both packages are identical. Additionally, in *rmeta*, combining binary data via Mantel-Haenszel method is possible, whereas *meta* provides a test for funnel plot asymmetry. Both packages provide standard graphics for meta analysis described below.

11.4 Graphics

A graphical representation of the results of a meta analysis is the *confidence interval plot*, sometimes also referred as *forest plot*. The confidence interval plot displays study-specific estimate and corresponding 100(1-α)% confidence interval for each study as well as the meta-analytical estimate of the common effect size and and corresponding 100(1-α)%

confidence interval. The two above mentioned R packages provide functions for drawing confidence interval plots.

A funnel plot, described in Lecture 8, for assessing publication bias is implemented in both R packages.

Data Sets

Data Set 1: Validity Studies Correlating Student Ratings of the Instructor with Student Achievement

Study	Sample	n	r
Bolton et al. 1979	General psychology	10	0.68
Bryson 1974	College algebra	20	0.56
Centra 1977[1]	General biology	13	0.23
Centra 1977[2]	General psychology	22	0.64
Crooks and Smock 1974	General physics	28	0.49
Doyle and Crichton 1978	Introductory communications	12	-0.04
Doyle and Whitely 1974	Beginning French	12	0.49
Elliot 1950	General chemistry	36	0.33
Ellis and Richard 1977	General psychology	19	0.58
Frey, Leonard and Beatty 1975	Introductory calculus	12	0.18
Greenwood et al. 1976	Analytic geometry and calculus	36	-0.11
Hoffman 1978	Introductory math	75	0.27
McKeachie et al. 1971	General psychology	33	0.26
Marsh et al. 1956	Aircraft mechanics	121	0.40
Remmer et al. 1949	General chemistry	37	0.49
Sullivan and Skanes 1974[1]	First-year science	14	0.51
Sullivan and Skanes 1974[2]	Introductory psychology	40	0.40
Sullivan and Skanes 1974[3]	First-year math	16	0.34
Sullivan and Skanes 1974[4]	First-year biology	14	0.42
Wherry 1952	Introductory psychology	20	0.16

Source: Cohen, P.A. (1983). Comment on selective review of the validity of student ratings of teaching. *Journal of Higher Education*, 54, 449-458.

Data Set 2: Studies of the Effects of Teacher Expectancy on Pupil IQ

Study	Estimated Weeks of Teacher-Student Contact Prior to Expectancy Induction	d	Standard Error
Rosenthal et al. 1974	2	0.03	0.125
Conn et al. 1968	21	0.12	0.147
Jose and Cody 1971	19	-0.14	0.167
Pellegrini and Hicks 1972[1]	0	1.18	0.373
Pellegrini and Hicks 1972[2]	0	0.26	0.369
Evans and Rosenthal 1968	3	-0.06	0.103
Fiedler et al. 1971	17	-0.02	0.103
Claiborn 1969	24	-0.32	0.220
Kester 1969	0	0.27	0.164
Maxwell 1970	1	0.80	0.251
Carter 1970	0	0.54	0.302
Flowers 1966	0	0.18	0.223
Keshock 1970	1	-0.02	0.289
Henrikson 1970	2	0.23	0.290
Fine 1972	17	-0.18	0.159
Greiger 1970	5	-0.06	0.167
Rosenthal and Jacobsen 1968	1	0.30	0.139
Fleming and Anttonen 1971	2	0.07	0.094
Ginsburg 1970	7	-0.07	0.174

Source: Raudenbush, S.W. and Bryk, A.S. (1985). Empirical Bayes meta–analysis. *Journal of Educational Statistics*, 10, 75-98.

Data Set 3: Results of eight randomized controlled trials comparing the effectiveness of amlodipine and a placebo on work capacity

	Amlodipine 10 mg (E)			Placebo (C)		
Protocol	n_{Ei}	$\bar{x}_{Ei}$	d^2_{Ei}	n_{Ci}	$\bar{x}_{Ci}$	d^2_{Ci}
154	46	0.2316	0.2254	48	-0.0027	0.0007
156	30	0.2811	0.1441	26	0.0270	0.1139
157	75	0.1894	0.1981	72	0.0443	0.4972
162A	12	0.0930	0.1389	12	0.2277	0.0488
163	32	0.1622	0.0961	34	0.0056	0.0955
166	31	0.1837	0.1246	31	0.0943	0.1734
303A	27	0.6612	0.7060	27	-0.0057	0.9891
306	46	0.1366	0.1211	47	-0.0057	0.1291

Source: Li, Y., Shi, L. and 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.

Data Set 4: Placebo-controlled trials on the effect of cisapride in the treatment of non–ulcer dyspepsia

Study	Cisapride	Placebo
1	15 / 16	9 / 16
2	12 / 16	1 / 16
3	29 / 34	18 / 34
4	42 / 56	31 / 56
5	14 / 22	6 / 22
6	44 / 54	17 / 55
7	14 / 17	7 / 15
8	29 / 58	23 / 58
9	10 / 14	3 / 15
10	17 / 26	6 / 27
11	38 / 44	12 / 45
12	19 / 29	22 / 30
13	21 / 38	19 / 38

Source: Hartung, J.and Knapp, G. (2001). A refined method for the meta–analysis of controlled clinical trials with binary outcome. *Statistics in Medicine*, 20, 3875-3889.

Data Set 5: Secondhand Smoking

For 19 case–control studies, number of cases of lung cancer in women who did not actively smoke cigarettes and estimated relative risk of lung cancer in relation exposure to environmental tobacco smoke

	Number of Cases	Estimated Relative Risk (95% Confidence Interval)
Akiba, Kato, Blot (1986)	94	1.52 (0.88 - 2.63)
Brownson et al. (1987)	19	1.52 (0.39 - 5.99)
Buffler et al. (1984)	41	0.81 (0.34 - 1.90)
Chan et al. (1979)	84	0.75 (0.43 - 1.30)
Correa et al. (1983)	22	2.07 (0.82 - 5.25)
Gao et al. (1978)	246	1.19 (0.82 - 1.73)
Garfinkel, Auerbach, Joubert (1985)	134	1.31 (0.87 - 1.98)
Geng, Liang, Zhang (1988)	54	2.16 (1.08 - 4.29)
Humble, Samet, Pathak (1987)	20	2.34 (0.81 - 6.75)
Inoue, Hirayama (1988)	22	2.55 (0.74 - 8.78)
Kabat, Wynder (1984)	24	0.79 (0.25 - 2.45)
Koo et al. (1987)	86	1.55 (0.90 - 2.67)
Lam et al. (1987)	199	1.65 (1.16 - 2.35)
Lam (1985)	60	2.01 (1.09 - 3.71)
Lee, Chamberlain, Alderson (1986)	32	1.03 (0.41 - 2.55)
Pershagen, Hrubec, Svensson (1987)	67	1.28 (0.76 - 2.15)
Svensson, Pershagen, Klominek (1988)	34	1.26 (0.57 - 2.82)
Trichopoulos, Kalandidi, Sparros (1983)	62	2.13 (1.19 - 3.83)
Wu et al. (1985)	28	1.41 (0.54 - 3.67)
Summary relative risk		1.42 (1.24 - 1.63)

Source: Environmental Protection Agency (1990); tables references cited there.

Based on this information, an Advisory Committee of EPA designated Environmental Tobacco Smoke as a Carcinogen.

Data Set 6: Randomized trials of misoprostol by endoscopic classification

Study	Treatment	Endoscopic classification				
		1	2	3	4	5
1	Misoprostol	21	2	4	2	0
	Placebo	2	2	4	9	13
2	Misoprostol	17	8	3	2	0
	Placebo	0	3	4	10	13
3	Misoprostol	20	4	6	0	0
	Placebo	8	4	9	4	5
4	Misoprostol	20	4	6	0	0
	Placebo	0	2	5	5	17
5	Misoprostol	1	4	5	0	0
	Placebo	0	0	0	4	6
6	Misoprostol	93	5	3	1	1
	Placebo	85	10	10	4	5
7	Misoprostol	61	12	0		
	Placebo	49	28	3		
8	Misoprostol	45	1	0		
	Placebo	65	6	3		
9	Misoprostol	138	1			
	Placebo	121	17			
10	Misoprostol	126	2			
	Placebo	110	21			
11	Misoprostol	30	1	1		
	Placebo	20	11	7		
12	Misoprostol	56	12	8	0	
	Placebo	50	15	12	5	
13	Misoprostol	12	3	1	0	
	Placebo	11	5	2	3	

Source: Whitehead, A. and Jones, N.M. (1994). A meta–analysis of clinical trials involving different classifications of response into ordered categories. *Statistics in Medicine*, 13, 2503-2515.

Bibliography

Agresti A. (1980). Generalized odds ratios for ordinal data. *Biometrics*, 36, 59-67.

Ananda, M.M.A and Weerahandi, S. (1997). Two–way anova with unequal cell frequencies and unequal variances. *Statistica Sinica*, 7, 631-646.

Argac, D., Makambi, K.H., and Hartung; J. (2001). A note on testing the nullity of the between group variance in the one–way random effects model under variance heterogeneity. *Journal of Applied Statistics*, 28, 215-222.

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

Becker, B.J. (1990). Coaching for the scholastic aptitude test: further synthesis appraisal. *Review of Educational Research*, 60, 373-417.

Begg, C.B. (1994). Publication bias. In *The Handbook of Research Synthesis*, H. Cooper and L.V. Hedges (Eds.). New York: Russell Sage Foundation.

Berger, J.O. (1980). *Statistical Decision Theory and Bayesian Analysis (2nd ed.)* Berlin: Springer.

Berry, S.C. (2000). Meta-analyisis versus large trials: Resolving the controversy. In *Meta-analysis in Medicine and Health Policy*, D. Stangl and D.A. Berry (Eds.). New York: Marcel Dekker.

Bhattacharya, C.G. (1980). Estimation of a common mean and recovery of interblock information. *Annals of Statistics*, 8, 205-211.

Biggerstaff, B.J. and 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.

Birnbaum, A. (1954). Combining independent tests of significance. *Journal of the American Statistical Association*, 49, 559-575.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Boardman, T.J. (1974). Confidence intervals for variance components – a comparative Monte Carlo study. *Biometrics*, 30, 251-262.

Brophy, J. and Joseph, L. (2000). A Bayesian meta-analysis of randomized mega-trials for the choice of thrombolytic agent in acute myocardial infarction. In *Meta-analysis in Medicine and Health Policy*, D. Stangl and D.A. Berry (Eds.). New York: Marcel Dekker.

Brown, L.D. and Cohen, A. (1974). Point and confidence estimation of a common mean and recovery of interblock information. *Annals of Statistics*, 2, 963-976.

Brown, M.B. and Forsythe, A.B. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics*, 16, 129-132.

Burdick, R.K. and Eickman, J. (1986). Confidence intervals on the among group variance component in the unbalanced one-fold nested design. *Journal of Statistical Computation and Simulation*, 26, 205-219.

Burdick, R.K. and Graybill, F.A. (1992). *Confidence Intervals on Variance Components*. New York: Dekker.

Burdick, R.K., Maqsood, F. and Graybill, F.A. (1986). Confidence intervals on the intra-class correlation in the unbalanced one-way classification. *Communication in Statistics – Theory and Methods*, 15, 3353-3378.

Cochran, W.G. (1937). Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society (Supplement)*, 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, J. (1977). *Statistical Power Analysis for the Behavioral Sciences* (rev. ed.). New York: Academic Press.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, A. and Sackrowitz, H.B. (1974). On estimating the common mean of two normal distributions. *Annals of Statistics*, 2, 1274-1282.

Cohen, A. and Sackrowitz, H.B. (1984). Testing hypotheses about the common mean of normal distributions. *Journal of Statistical Planning and Inference*, 9, 207-227.

Collins, R., Gray, R., Godwin, J., and Peto, R. (1987). Avoidances of large biases and large random errors in the assessment of moderate treatment effects: The need for systematic overviews. *Statistics in Medicine*, 6, 245-250.

Cooper, H. and Hedges, L.V. (1994). *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Das, R. and Sinha, B. K. (1987). Robust optimum invariant unbiased tests for variance components. *Proceedings of Second International Tampere Conference in Statistics*, 317-342.

Dasgupta, A. and Sinha, B.K. (2006). On some statistical aspects of combining gallup poll results. Technical Report, Department of Mathematics and Statistics, University of Maryland, Baltimore County.

DerSimonian, R. and Laird, N.M. (1983). Evaluating the effect of coaching on SAT scores: a meta-analysis. *Harvard Educational Review*, 53, 1-15.

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

Dominici, F. and Parmigiani, G. (2000). Combining studies with continuous and dichotomous responses: A latent variables approach. In *Meta-analysis in Medicine and Health Policy*, D. Stangl and D.A. Berry (Eds.). New York: Marcel Dekker.

Draper, D., Gaver, D.P., Jr., Goel, P.K., Greenhouse, J.B., Hedges, L.V., Morris, C.N., Tucker, J.R., and Waternaux, C.M. (1992). *Combining Information: Statistical Issues and Opportunities for Research*. Washington, D.C.: American Statistical Association, National Academy Press.

Eberhardt, K.R., Reeve, C.P., and Spiegelman, C.H. (1989). A minimax approach to combining means, with practical examples. *Chemometrics and Intelligent Laboratory Systems*, 5, 129-148.

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 ed.)*. London: Oliver & Boyd.

Fleiss, J.L. (1994). Measures of effect size for categorical data. In *The Handbook of Research Synthesis*, H. Cooper and L.V. Hedges (Eds.). New York: Russell Sage Foundation.

Gamage, J. and Weerahandi, S. (1998). Size performance of some tests in one-way anova. *Communications in Statistics - Simulation and Computation*, 27, 625-640

Gart, J.J. and 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.

Gauss, C.F. (1809). *Theoria motus corporum coelestium in sectionis conicis solem ambientum.* Hamburg: Perthes et Besser.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2004). *Bayesian Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC Press.

George, E.O. (1977). Combining independent one-sided and two-sided statistical tests — Some theory and applications. *Doctoral dissertation*, University of Rochester.

Givens, G.H., Smith, D.D., and Tweedie, R.L. (1997). Publication bias in meta-analysis: A Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Statistical Science*, 12, 221-247.

Glass, G.V. (1976). Primary, secondary, and meta-analysis. *Educational Researcher*, 5, 3-8.

Glass, G.V., McGaw, B., and Smith, M.L. (1981). *Meta-Analysis in Social Research.* Beverly Hills, CA: Sage.

Graybill, F.A. (1976). *Theory and Application of the Linear Model.* North Scituate: Duxbury Press.

Graybill, F.A. and Deal, R.B. (1959). Combining unbiased estimators. *Biometrics*, 15, 543-550.

Griffiths, W. and Judge, G. (1992). Testing and estimating location vectors when the error covariance matrix is unknown. *Journal of Econometrics*, 54, 121-138.

Haff, L.R. (1979). An identity for the Wishart distribution with applications. *Journal of Multivariate Analysis*, 9, 531-544.

Hardy, R.J. and 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. *Annals of Statistics*, 9, 278-292.

Hartung, J. (1999). An alternative method for meta–analysis. *Biometrical Journal*, 41, 901-916.

Hartung, J. and Argac, D. (2002). Generalizing the Welch test to non-zero hypotheses on the variance component in the one-way random effects model under variance heterogeneity. *Statistics*, 36, 89-99.

Hartung, J., Argac, D., and 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. and Knapp, G. (2003). Generalized Cochran–Wald statistics in combining of experiments. *Journal of Statistical Planning and Inference*, 113, 215-237.

Hartung, J. and Knapp, G. (2000). Confidence intervals for the between group variance in the unbalanced one-way random effects model of analysis of variance. *Journal of Statistical Computation and Simulation*, 65, 311-323.

Hartung, J. and 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. and 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.and Knapp, G. (2004). Improved tests of homogeneity in randomized controlled multi–center trials with binary outcome. *Far East Journal of Theoretical Statistics*, 13, 101-126.

Hartung, J. and Knapp, G. (2005). 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., Makambi, K.H., and Argac, D. (2001). An extended ANOVA F-test with applications to the heterogeneity problem in meta-analysis. *Biometrical Journal*, 43, 135-146.

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. (1994). Fixed Effects Models. In *The Handbook of Research Synthesis*, H. Cooper and L.V. Hedges (Eds.). New York: Russell Sage Foundation.

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

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.

Hunter, J.E., Schmidt, F.L., and Jackson, G.B. (1982). *Meta-analysis: Cumulating Research Finding across Studies*. Beverly Hills, CA: Sage.

Iyengar, S. and Greenhouse, J.B. (1988). Selection models and the file drawer problem. *Statistical Science*, 3, 109-135.

Iyer, H., Wand, J., Mathew, T. (2004). Models and confidence intervals for true values in interlaboratory trials. *Journal of the American Statistical Association*, 99, 1060-1071.

Jordan, S.M. and Krishnamoorthy, K. (1996). Exact confidence intervals for the common mean of several normal populations. *Biometrics*, 52, 77-86.

Kempthorne, O., Mukhopadhyay, N., Sen, P.K., and Zacks, S. (1991). Research–How to do it: A panel discussion. *Statistical Science*, 6, 149-162.

Khatri, C.G. and Shah, K.R. (1974). Estimation of location parameters from two linear models under normality. *Communications in Statistics*, 3, 647-663.

Khuri, A.I., Mathew, T., and Sinha, B.K. (1998). *Statistical Tests for Mixed Linear Models.* New York: Wiley.

Kubokawa, T. (1990). Minimax estimation of common coefficients of several regression models under quadratic loss. *Journal of Statistical Planning and Inference*, 24, 337-345.

Lamotte, L.R. (1973). On non-negative quadratic unbiased estimation of variance components. *Journal of the American Statistical Association*, 68, 728-730.

Legendre, A.M. (1805). *Nouvelles méthodes pour la détermination der orbites des comètes.* Paris: Courcier.

Liang, K.Y. and Self, S.G. (1985). Test for homogeneity of odds ratio when the data are sparse. *Biometrika*, 72, 353-358.

Light, R.J., and Pillemer, D.B. (1984). *Summing up: The Science of Reviewing Research.* Cambridge, MA: Harvard University Press.

Lipsitz, S.R., Dear, K.B.G., Laird, N.M., and Molenberghs, G. (1998). Tests for homogeneity of the risk difference when data are sparse. *Biometrics*, 54, 148-160.

Marden, J.I. (1991). Sensitive and sturdy p-values. *The Annals of Statistics*, 19, 918-934.

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

Mehrotra, D.V. (1997). Improving the Brown-Forsythe solution to the generalized Behrens-Fisher problem. *Communications in Statistics - Simulation and Computation*, 26, 1139–1145.

Mehta, J.S. and Gurland, J. (1969). Combination of unbiased estimates of the mean which consider inequality of unknown variances. *Journal of the American Statistical Association*, 64, 1042-1055.

Meier, P. (1953). Variance of a weighted mean. *Biometrics*, 9, 59-73.

Mitra, P.K. and Sinha, B.K. (2007). On some aspects of estimation of a common mean of two independent normal populations. *Journal of Statistical Planning and Inference* 137, 184-193.

Montgomery, D. (1991). *Design and Analysis of Experiments (3rd ed.)* New York: Wiley

Mosteller, F. and Bush, R. (1954). Selected quantitative techniques. In *Handbook of Social Psychology: Theory and Method, Vol.1*, G. Lindsey (Ed.). Cambridge, MA: Addison-Wesley.

Nair, K.A. (1980). Distribution of an estimator of the common mean of two normal populations. *Annals of Statistics*, 8, 212-216.

Norwood, T. E. and Hinkelmann, K. (1977). Estimating the common mean of several normal populations. *Annals of Statistics*, 5, 1047-1050.

Patil, G.P. and Rao, C.R. (1977). The weighted distributions: A survey of their applications. In *Applications of Statistics*, Krishnaiah (Ed.). Amsterdam: North-Holland.

Pauler, D.K. and Wakefield, J. (2000). Modeling and implementation issues in Bayesian meta-analysis. In *Meta-analysis in Medicine and Health Policy*, D. Stangl and D.A. Berry (Eds.). New York: Marcel Dekker.

Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *British Medical Journal*, 2, 1243-1246.

Pearson, K. (1933). On a method of determining whether a sample of given size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika*, 25, 379-410.

Pettigrew, H.M., Gart, J.J., and 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, 122-115.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*. New York: Wiley.

Rao, P.S.R.S, Kaplan, J., and 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.

Rohatgi, V.K. (1976). *An Introduction to Probability Theory and Mathematical Statistics*. New York: Wiley.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, 86, 638-641.

Rosenthal, R. (1984). *Meta-Analytic Procedures for Social Research*. Beverly Hills, CA: Sage.

Rosenthal, R. (1994). Parametric measures of effect size. In *The Handbook of Research Synthesis*, H. Cooper and L.V. Hedges (Eds.). New York: Russell Sage Foundation.

Rubin, D.B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics*, 6, 337-401.

Rukhin, A.L., Biggerstaff, B.J., and 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.

Searle, S.R., Casella, G. and McCulloch, C.E. (1992). *Variance components.* New York: Wiley.

Shinozaki, N. (1978). A note on estimating the common mean of k normal populations and the Stein Problem. *Communications in Statistics*, A 7(15), 1421-1432.

Sidik, K. and Jonkman, J.N. (2002). A simple confidence interval for meta–analysis. *Statistics in Medicine*, 21, 3153-3159.

Sinha, B.K. (1979). Is the maximum likelihood estimate of the common mean of several normal populations admissible? *Sankhyā*, B, 40, 192-196.

Sinha, B.K. (1985). Unbiased estimation of the variance of the Graybill-Deal estimator of the common mean of several normal distributions. *Canadian Journal of Statistics*, 13, 243-247.

Sinha, B.K. and Mouqadem, O. (1982). Estimation of the common mean of two univariate normal populations. *Communications in Statistics, Theory & Methods*, 11, 1603-1614.

Skinner, J.B. (1991). On combining studies. *Drug Information Journal*, 25, 395-403.

Smith, T.C., Spiegelhalter, D.J., and Thomas, A. (1995). Bayesian approaches to random–effects meta–analysis: a comparative study. *Statistics in Medicine*, 14, 2685-2699.

Stigler, S.M. (1986). *The history of statistics — The measurement of uncertainty before 1900.* Cambridge, Ma: Harvard University Press.

Stouffer, S.A., Suchman, E.A., DeVinney, L.C., Star, S.A., and Williams, R.M., Jr. (1949). *The American Soldier, Volume I. Adjustment during Army Life.* Princeton, N.J.: Princeton University Press.

Thomas, J.D. and Hultquist, R.A. (1978). Interval estimation for the unbalanced case of the one-way random effects model. *Annals of Statistics*, 6, 582-587.

Thursby, J.G. (1992). A comparison of several exact and approximate tests for structural shift under heteroscedasticity. *Journal of Econometrics*, 53, 363-386.

Tippett, L.H.C. (1931). *The Methods of Statistics.* London: Williams & Norgate.

Tsui, K. and 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.

Tukey, J.W. (1951). Components in regression. *Biometrics*, 7, 33-69.

Van Houwelingen, H.C, Arends, L.R., Stijnen, T. (2002). Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine*, 21, 589-624.

Verbeke, G. and Molenberghs, G. (1997). *Linear Mixed Models in Practice.* New York: Springer.

Wald, A. (1940). A note on the analysis of variance with unequal class frequencies. *Annals of Mathematical Statistics*, 11, 96-100.

Wang, C.M. (1990). On the lower bound of confidence coefficients for a confidence interval on variance components. *Biometrics*, 46, 187-192.

Warn, D.E., Thompson, S.G., and 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.

Weerahandi, S. (1995). *Exact Statistical Methods for Data Analysis.* New York: Springer.

Welch, B.L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika*, 38, 330-336.

White, H.D. (1994). Scientific communication and literature retrieval. In *The Handbook of Research Synthesis*, H. Cooper and L.V. Hedges (Eds.). New York: Russell Sage Foundation.

Whitehead, A. (2002). *Meta-Analysis of Controlled Clinical Trials.* Chicester: Wiley.

Whitehead, A., Jones, N.M. (1994). A meta–analysis of clinical trials involving different classifications of response into ordered categories. *Statistics in Medicine*, 13, 2503-2515.

Whitehead, A., Omar, R.Z., Higgins, J.P.T., Savaluny, E., Turner, R.M., and Thompson, S.G. (2001). Meta–analysis of ordinal outcomes using individual patient data. *Statistics in Medicine*, 20, 2243-2260.

Wilkinson, B. (1951). A statistical consideration in psychological research. *Psychological Bulletin*, 48, 156-158.

Williams, J.S. (1962). A confidence interval for variance components. *Biometrika*, 49, 278-281.

Woolf, B. (1955). On estimating the relation between blood group and disease. *Annals of Human Genetics*, 19, 251-253.

Yates, F. and Cochran, W.G. (1938). The analysis of groups of experiments. *Journal of Agricultural Science*, 28, 556-580.

Yu, P.L.H., Sun, Y., and Sinha, B.K. (2002). Estimation of the common mean of a bivariate normal population. *Annals of the Institute of Statistical Mathematics*, 54, 861-878.

Zacks, S. (1966). Unbiased estimation of the common mean of the two normal distributions based on small samples of equal size. *Journal of American Statistical Association*, 61, 467-476.

Zacks, S. (1970). Bayes and fiducial equivariant estimators of the common mean of two populations. *Annals of Mathematical Statistics*, 41, 59-69.