AN INTRODUCTION TO META ANALYSIS

by

ALLA BOYKOVA

B.A., Penza State Technical University, Ministry of Education of Russian Federation,1991
M.S., Kansas State University,2005

A REPORT

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts And Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2008

Approved by:

Major Professor
Dr. D.Johnson

ABSTRACT

Meta analysis is a statistical technique for synthesizing of results obtained from multiple studies. It is the process of combining, summarizing, and reanalyzing previous quantitative research. It yields a quantitative summary of the pooled results.

Decisions of the validity of a hypothesis cannot be based on the results of a single study, because results typically vary from one study to the next. Traditional methods do not allow involving more than a few studies. Meta analysis provides certain procedures to synthesize data across studies. When the treatment effect (or effect size) is consistent from one study to the next, meta-analysis can be used to identify this common effect. When the effect varies from one study to the next, meta-analysis may be used to identify the reason for the variation.

The amount of accumulated information in fast developing fields of science such as biology, medicine, education, pharmacology, physics, etc. increased very quickly after the Second World War. This lead to large amounts of literature which was not systematized. One problem in education might include ten independent studies. All of the studies might be performed by different researchers, using different techniques, and different measurements. The idea of integrating the research literature was proposed by Glass (1976, 1977). He referred it as the meta analysis of research.

There are three major meta analysis approaches: combining significance levels, combining estimates of effect size for fixed effect size models and random effect size models, and vote-counting method.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I am grateful to my supervisor Professor Dallas Johnson for his guidance, advice, corrections, forbearance, and deep insight throughout the report.

I wish to thank my Committee members Professor John Boyer and Professor Shie-Shien Yang for their suggestions and comments.

INTRODUCTION

Meta analysis is concerned with pooling or combining results from several different studies.

The term “meta analysis” was first proposed by Glass who called it "analysis of analysis" (Glass, 1976). Glass suggested that there are three levels of data analysis. The first level or *primary analysis* corresponds to an original data analysis in a research study. The *secondary analysis* (second level) is a re-analysis of data with regards to original research questions using the most appropriate statistical techniques or answering a new question using old data. And finally, an advanced secondary analysis (the third level) is the *meta analysis of research* or analysis of analysis. It is the statistical analysis of a collection of analysis results that come from individual different studies. The purpose of meta analysis is to choose appropriate techniques to integrate or combine different studies to better answer an original question.

The need for the meta analysis of research studies seemed to be clear 30-40 years ago because of rapidly growing collections of research literature in social science fields. Fast developing fields such as medicine and pharmacology need advanced statistical methodologies as well. Each field of science contains hundreds of unsolved problems with dozens of papers devoted to each of them. Usually each study involves more than one topic. The importance of choosing the right topic and the corresponding collection of studies arises immediately after determining a question of interest. Even if the topic is the same, techniques and measurements may vary from one study to another.

Assume that a question of importance is determined. What is the next step? To determine the study or topic, or to collect literature? One study may contain several topics. How does one recognize whether a study topic contains important information? Or if one has several studies involved, how does one decide which studies to include? There is no a single method that can answer all these questions in general. Meta analysis techniques allow one to describe quantitative data and combine evidence across studies.

One problem concerns the standardization of different studies. Published studies may

come from different research laboratories, different centers, etc. The studies are almost always performed independently of one another. Unfortunately there are no standardized methods or commonly used report forms under which such studies are published.

Difficulties in determining the methodology of meta analysis starts with the assumptions that define what studies should be included. Usually studies involve many different subjects that produce different numbers or kinds of findings. Most studies produce more than one finding. Moreover, different studies usually use different scales, measures, etc. So, a big issue is how one can combine many different findings that may have used different measurement scales? If one study produces ten findings, and another study produces a hundred findings, should one average findings within each study? If the answer to this question is yes, should one average the number of subjects first and then find the average of the findings? Or should one assign a weight to each study? If one is going to weight each study, then how should the weights be obtained? Should it be some number or should it be some weight function? All these questions arise at the first level of meta analysis.

A simple example considers the analysis of the effectiveness of open classrooms in the education of students (Hedges, 1985). Students from traditional schools were compared with students from experimental open classroom schools. About 200 studies were involved. They classified 16 different dependent variables using a variety of different outcomes. Here are few of the variables considered: anxiety, attitude toward teacher, cooperativeness, creativity, curiosity, general mental ability, mathematical achievement, reading achievement, etc.

An important question is: how does one recognize poorly designed studies among hundreds of studies if one only has the exact findings from previous studies? There is a paradox that was popular at early stages of meta analysis. Many added weak studies (with poor design, say) may lead to a strong conclusion. But even if it works, one should recognize the "weaknesses" in each study and avoid consistently repeating weaknesses from one study to another one. Assume one has 10 studies. Suppose that the first two are weak with respect to data analysis but strong in other components (representative samples, measurements utilized). Suppose another two studies are weak in the way that samples were collected. The point is to avoid repeating weakness in sampling in all 10 studies. Sampling weaknesses would lead one to question the trustworthiness of the design, its description, and conclusions made from the study.

How many studies should be involved to answer a particular question using known statistical methods to determine “aggregate findings”? Collecting results from a thousand studies could lead to the same answer as collecting results from ten representative studies. Typical meta analysis of research studies is to formulate a conception of the topic at the stage of literature collection (Glass 1977). The researcher may then narrow the topic concept at the meta analysis stage.

Designed experiments produce some outcomes or “findings”. Researchers carrying out their own experiments follow their own interests. A researcher’s interest is to get a desired result and he may not think of additional experiments that would make his report clear for ensuing investigations, i.e. include detailed information about their experiment. Many published reports are full of limitations on such aspects as study and design descriptions, measurements, data analysis (primary and/or secondary analysis of research in this context). In such cases it is very difficult to decide whether a study and/or findings are appropriate for research integration and further investigations. So, it can be confusing when one investigates a certain topic and uses published studies and findings even if previous designs were not perfect and published reports contain limitations. Another possible situation occurs when a study "fails" desired criteria or some conditions and the study is eliminated from consideration. "The researcher does not want to conduct a poor study ... but it hardly follows that after a less-than-perfect study has been done, its findings should not be considered " (Glass 1977).

Are there some commonly used criteria to justify a "grade" of a design? Probably not. Nevertheless, there are some ways to improve the design. One way is to study "the covariation between design characteristics and findings" (Glass, 1977). Hence, research integration can help one perform a better design. It may help to avoid some of the problems indicated above. A detailed description of the study design and analysis may clarify some limitations. Further study of covariation between findings and analysis may lead to a determination of the number of findings and better descriptions of the findings.

The next issue is combining or “integrating” studies. A point of interest is to integrate different studies and find methods for combining them. For example, a suppose a researcher investigates several cattle diets. He picks eight farms in Kansas. After performing a completely randomized design, he gets some results or findings. Then he picks six farms in Iowa and

produces randomized complete block design to investigate the same diets. He is testing the same hypothesis but the two designs are different and therefore these particular studies can be classified as different studies trying to answer the same question.

To be able to combine results from different studies, the results from the different studies should be comparable. If they differ too much, it will not be possible to combine the studies. Another issue is: How does one integrate different studies that are not easily compared, i.e. those having different structures, different measurements, or different scales? It is necessary that the different studies attempt to answer the same question or serve as parts of the same problem. So, the question is: How does one make inadequate studies adequate? For example (Glass, 1977), a researcher wants to find evidence of the relative effectiveness of unequal studies on computer-assisted instruction (CAI) and cross-age tutoring (CAT). Assume that 100 studies in CAI were divided into two groups such as 25 were in science and 75 were in math. Meanwhile 100 studies in CAT consisted of two groups such as 25 studies were in math and 75 were in science.

The problem of comparison is obvious. Each field has the same number of studies, but they have different sizes! Suppose that one is interested in the effectiveness of installing CAI in a traditional school (Glass, 1977). Then it is obvious that the researcher should have evidence of using CAI instruction for math more often (say, three times) than for science. But, if the researcher is interested in "effective medium" CAI versus CAT, the necessity of having some technique to make adequate size measurements for both fields would be eliminated.

The first attempts to integrate several individual studies used classifications of studies by type and then interpreted statistical significance. Historically, Tippett first proposed a test of statistical significance of combined results (the minimum *p* method) in 1931. Then Fisher (1932) and Pearson (1933) independently derived a test of statistical significance of combined results (now called Fisher or Pearson methods or *p*-value across the study). Next Cochran (1937) proposed a method based on numerical estimates of treatment effect. Many researchers used the methods mentioned above but all of them have disadvantages. We will consider some of the disadvantages in Chapter 1.

The next step was taken in the 1970s. This approach could be briefly described as that which consisted of finding some deficiencies when analyzing the collection of studies and then

developing one or two of the most acceptable studies. Most criticism of this approach was that it seemed to be hardly possible to compare significance of results coming from poorly-designed and well-designed experiments.

Glass suggested that one should group studies by "quantification and measurement of study characteristics, by experimental outcomes, by correlation outcomes, and by problems of statistical inference".

Quantification of study characteristics requires the presentation of descriptions of findings in quantitative terms. It is not always easy because some findings are categorical. One has to have a bridge. Even if quantification is possible, problems with using reports of studies that omit important information still remain and missing data methods are necessary.

To resolve the issue of outcomes of experimental and correlation studies, two free-scale values to measure effect magnitude were proposed by Glass (1976, 1977).

The first is called effect-size and was derived by Cohen (1969). He wrote "we need a 'pure' number, one free of our original measurement unit. This is accomplished by standardizing the raw effect size as expressed in the measurement unit of the dependent variable by dividing it by the (common) standard deviation...

$$\delta = \frac{\mu_a - \mu_b}{\sigma} \qquad (0.1)$$

where Φ is the standard deviation of either population (they are assumed equal)".

The effect size is used to combine the results of studies and to measure the effectiveness of the experimental treatments.

Another commonly used free-scale index of effect magnitude is the product-moment correlation coefficient. Glass (1977) suggested that a correlation analysis may be carried out in the metric of r_{xy} or r_{xy}^2. The usual approach is to obtain a Pearson correlation coefficient or its approximation from reported statistics. Glass (1977) also gave "guidelines" for converting various summary statistics into product-moment correlations.

This report will concentrate on methods that involve effect size estimations.

One of the techniques used to estimate effect size across studies involves computing the

effect size for each of individual studies and then averaging them. Also regression analysis and analysis of variance have been used (Hedges and Olkin, 1985).

The inferential statistical problems are complex. In fact, data are usually independent statistically. Two suggestions were proposed (Glass 1977). One is based on considering independent findings. It is wrong, but practical, because it reduces standard errors. Otherwise, one can not use some studies that yield enormous standard errors. Another method known as the jackknife method was proposed by Mosteller and Tukey (1968). This method is not discussed in this report. The interested reader should refer to their paper.

There are two meta-analysis approaches that investigate an effect size. One of them is a so-called traditional approach proposed by Glass (1976, 1977), Cohen (1969), and Hedges and Olkin (1985) is based on investigating the standardized mean difference, its estimation, distribution, distribution of estimates, different types of effect models, hypotheses testing, etc .

Another one is based on measuring the absolute mean difference in two groups of study. This is common in the field of medicine. The absolute difference in the means is defined as

$$\theta = | \mu^E - \mu^C | \qquad (0.2)$$

where μ^E is a mean of experimental population and μ^C is a mean of control population.

Meta-analysis methodology is widely used in medicine. Most clinical research studies are based on randomized controlled trials. The forms and amount of data may vary but what makes such research special is the presence of individual patient data. Meta analysis methods are conducted by using individual patient data as well as summary statistics obtained from individual clinical trials. Statistical packages are very useful, especially in cases where obtaining an exact analytical solution is difficult. In this report some SAS® procedures for the analysis of clinical trials are presented. Methodology for conducting meta-analysis for clinical trials with detailed explanations and examples including SAS® codes are given in Whitehead (2002).

Data for conducting meta analyses in clinical research may be provided in the form of summary information obtained from clinical trial reports or from studies when individual patient data are available. Three forms of data are commonly used: i) an estimate of the treatment difference and its variance or standard error; ii) summary statistics for each treatment group; and

iii) individual patient data. In general, there are five different types of outcome data: normally distributed data, binary data, survival data, interval-censored survival data, and ordinal data. In this report normally distributed and binary data and methods of their analyses are considered.

A particular interest for researchers performing clinical experiments is to investigate absolute mean differences between two groups in studies. To conduct analyses for individual patient data researchers usually use Student's two sample *t*-test, F- tests, and maximum likelihood approaches. Examples of models for different types of outcome and statistical analyses are given in detail in Whitehead (2002). In her book she also describes the traditional statistical approach based on summary statistics information proposed by Hedges and Oklin (1985) and refers to applications in clinical trials.

CHAPTER 1 - TESTS OF SIGNIFICANCE OF COMBINED RESULTS

This chapter is devoted to statistical methods for testing the statistical significance of combined results. These methods are based on combining significance levels or *p*-values obtained from different independent studies testing the same directional hypotheses. Such procedures are called *omnibus* or *nonparametric* procedures (Hedges and Olkin, 1985) because they do not depend on the distribution of the data but only on observed significance levels called *p*-values. Moreover, the distributions of the test statistics might be unknown. In fact, continuous test statistics yield *p*-values that are distributed uniformly under the null hypothesis regardless of the distribution from which they arise, (Casella, Berger, 2002), (Hedges and Olkin, 1985).

The first publications that combined significance tests belonged to Tippett (1931), Fisher (1932), and Pearson (1933). Wallis (1942) continued working on Fisher's method and described important discrete cases. Further investigations were continued by Wilkinson (1951), Birnbaum (1954), Littell and Folks (1971), Rosental (1978).

The problem of producing a specific statistical procedure for quantitative synthesis is as

follows. There are sets of null hypotheses, test statistics, and *p*-values for some parameters obtained from independent experiments (studies). In order to combine results one has to develop a common null hypothesis as well as a common test statistic for the whole set of experiments. There are two possibilities i) either values of test statistics or their distribution are unknown or ii) even if such information is available it is impossible to make up an appropriate single test. For example, a slight simplification of the example stated by Birnbaum (1954) is as follows: Two independent experiments to measure a certain drug effect are performed. At least one of the possible effects may be asserted: a) an increase in the mean of a certain measurable physiological quantity; b) an increase in the variance (within a subject) of the same or a second measurable physiological quantity. Suppose that the tests for each of these two independent experiments are based on two statistics T_1 and T_2 . The goal is to produce a single test based on some combination of the two test statistics. Unfortunately there is no single optimal method of combining independent test statistics.

1.1 Preliminaries and Notations

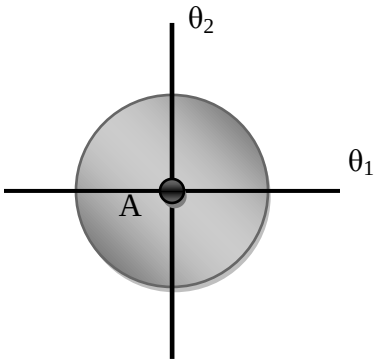
Consider *k* independent studies. Each study is characterized by one parameter $\theta_i, i = 1,\ldots,k$ such as a mean, a difference between two means, or a correlation coefficient. Therefore, for *k* studies, there are *k* parameters $\theta_1,\ldots,\theta_k$ to be investigated (Hedges and Olkin, 1985). There are *k* null hypotheses to be tested such as $H_{0i} : \theta_i = 0, i = 1,\ldots,k$. Assume that the *i*th study produces a test statistic T_i . It is not necessary that all *k* null hypotheses have the same meaning and/or the corresponding test statistics have the same distributions. The composite hypothesis $H_0 : \theta_1 = \theta_2 = \cdots = \theta_k = 0$ is valid if each of the H_{0i} being true implies that none of the θ_i is significantly different from zero.

The *p*-value for the *i*th study is defined as follows $p_i = \Pr\{T_i \geq t_{i0}\}$ where t_{i0} is the value of the statistic that was obtained in the *i*th study. If H_{0i} is true, then the p_i's are uniformly distributed in the interval (0,1) (Hedges and Olkin, 1985).

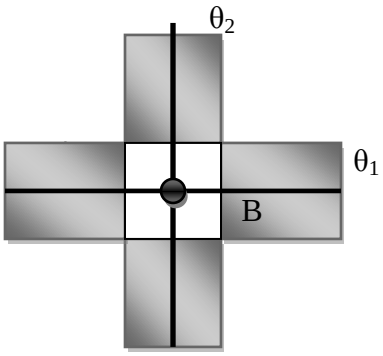
The question "which test produces false H_0" does not have a direct answer. All parameters $\theta_i, i = 1,...,k$ greater than zero yield false H_0 and one parameter greater than zero, i.e.

$\theta_1 = \theta_2 = \cdots = 0$ with some $\theta_k > 0$ also gives a false null hypothesis. One test does not appear to be sensible to all possible alternatives. An illustration of null and alternative hypothesis variation is described for two-dimensional space below.

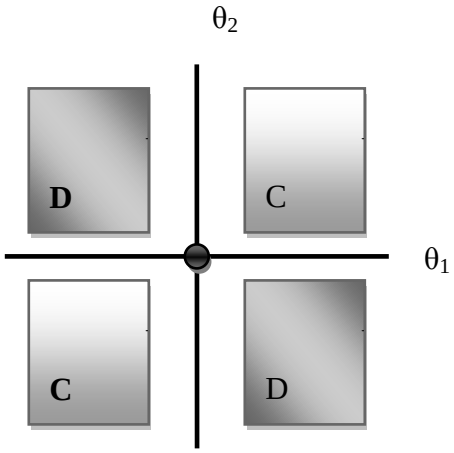
Let $\theta_i, i = 1,2$ be parameters to be tested.



$\{\theta_1, \theta_2 : \theta_1^2 + \theta_2^2 \le C\}$



$\{\theta_1, \theta_2 : |\theta_1| > C, |\theta_2| < C\}$ or
$\{\theta_1, \theta_2 : |\theta_1| < C, |\theta_2| > C\}$



$\{\{\theta_1, \theta_2 : \theta_1 > C,\ \theta_2 < -C\}$ or $\{\theta_1, \theta_2 : \theta_1 < -C,\ \theta_2 > C\}\}$ and
$\{\{\theta_1, \theta_2 : \theta_1 > D,\ \theta_2 > D\}$ or $\{\theta_1, \theta_2 : \theta_1 < -D,\ \theta_2 > -D\}\}$

Fig 1. Examples of alternative hypotheses in two-dimensional parameter spaces (Hedges and Olkin, 1985).

The null hypothesis H_0 corresponds to the origin (0,0) (region A) implies both θ_1 and θ_2 are close to zero in region *A.* In region *B* just one of the θ's is close to zero. In regions *C* and *D* both θ_1 and θ_2 are far from zero.

There are three general alternative hypotheses. The first one implies that there is one known direction of all deviations from H_0. The alternative hypothesis would be

$H_1 : \theta_i \geq 0, i = 1,...,k$ and at least one $\theta_i > 0$. Such an alternative hypothesis is appropriate in the case of *F*-statistics in an analysis of variance or for a chi-square statistic where one rejects for large values of the test statistics.

A second alternative hypothesis is that $H_2 : \theta_i \leq 0$ or $\theta_i \geq 0$ and at least one $\theta_i \neq 0$. Such an alternative hypothesis would result in the case of *t*-statistics or a correlation coefficient.

A third alternative hypothesis is given by $H_3 :$ at least one $\theta_i \neq 0$. The hypothesis is relevant in the case when effects that arise from different studies need not have the same sign.

Choosing an appropriate alternative hypothesis depends on the problem.

1.2 Combined Test Procedures

This section is devoted to using tests of significance to combine results. Consider continuous test statistics.

1.2.1 Methods Based on the Uniform Distribution

Tippett (1931) first proposed a test of the significance of combined results. The procedure involves ordered independent *p*-values $p_1,...,p_k$ that are distributed uniformly on the (0,1) interval under H_0. Let $p_{[1]}$ be the minimum of $p_1,...,p_k$, then a size α test procedure is

reject H_0 if $p_{[1]} < 1-(1-\alpha)^{1/k}$.

Wilkinson (1951) generalized Tippett's procedure. Let ordered *p*-values, $p_1,...,p_k$, be obtained from *k* independent studies satisfy the condition that

$p_{[1]} \leq p_{[2]} \leq \cdots \leq p_{[k]}$.

He used the *r*th smallest *p*-value as a test statistic and compared it to a critical value $p_{r,\alpha}$. Because $p_{[r]} \sim beta(r, k-r+1)$, the critical values can be obtained from the tables of the incomplete beta distribution function for a desirable size α (Hedges and Olkin, 1985).

1.2.2 The Inverse Chi-square Method

The inverse chi-squared method is the most widely used test of significance for combining results based on *p*-values. It was proposed by Fisher (1932). Fisher used the product of the *p*-values obtained from *k* independent studies. Recall that if *U* is distributed uniformly on (0,1), then $-2\log U$ has a chi-square distribution with 2 degrees of freedom. Therefore, since the *p*-values are distributed uniformly under true H_{0i}, $-2\log p_i, i = 1,...,k$ has a chi-square distribution with 2 degrees of freedom. Then if H_0 is true,

$$-2\log(p_1 p_2 \cdots p_k) = -2\log p_1 - 2\log p_2 - \cdots - 2\log p_k$$

has a chi-square distribution with 2*k* degrees of freedom. Fisher's test is to

reject H_0 if $P = -2\sum_{i=1}^{k} \log p_i \geq \chi^2_{\alpha,2k}$.

A modification of Fisher's method was proposed by Good (1955). The modification combines the *p*-values as $P_w = p_1^{v_1} p_2^{v_2} \cdots p_k^{v_k}$, where $v_1, v_2, \ldots, v_k$ are nonnegative weights chosen such that the test becomes more sensitive. When $v_1 = \cdots = v_k = 1$, one gets the Fisher method.

The distribution of P_w was obtained by Robbins (1948) and Good (1955) for the case when all weights are distinct.

They obtained the cumulative distribution function as

$$\Pr\{P_w \leq q\} = \frac{q^{1/v_1}}{a_1} + \cdots + \frac{q^{1/v_k}}{a_k}$$

where $a_i = \dfrac{(v_i - v_1)(v_i - v_2)\cdots(v_i - v_{i-1})(v_i - v_{i+1})\cdots(v_i - v_k)}{v_1^{k-1}}$.

No general expression for the distribution of P_w has been obtained for the case when weights are not distinct. Note, that under H_0, $-2\log P_W$ is distributed as a weighted sum of chi-squared variables, and this has complicated representations (Hedges and Olkin, 1985).

1.2.3 Pearson's Method

A method known as Pearson's method was proposed by Pearson (1933). He combined *p*-values as the product $(1-p_1)(1-p_2)\cdots(1-p_k)$. His test is to

reject H_0 if $(1-p_1)(1-p_2)\cdots(1-p_k)\geq C$, where C is a critical value corresponding to a desired significance level and obtained by following Fisher's method.

1.2.4 The Inverse Normal Method

Stouffer, Suchman, DeVinney, Star and Williams (1949) and Liptak (1958) independently proposed the inverse normal method. Define Z_i by $p_i = \Phi(Z_i)$, where $\Phi(x)$ is the standard normal cumulative distribution function. The test statistic is a transformation of the *p*-values to a standard normal score as

$$Z = \frac{Z_1 + Z_2 + \cdots + Z_k}{\sqrt{k}} = \frac{\Phi^{-1}(p_1) + \cdots + \Phi^{-1}(p_k)}{\sqrt{k}},$$

where Z has the standard normal distribution. The test is to

reject H_0 if $Z \geq C$ where C is a critical value obtained from standard normal distribution.

1.2.5 The Weighted Inverse Normal Method

The weighed inverse normal method was proposed by Mosteller and Bush (1954). The test statistic was derived as follows

$$Z_w = \frac{v_1 Z_1 + v_2 Z_2 + \cdots + v_k Z_k}{\sqrt{v_1^2 + v_2^2 + \cdots + v_k^2}} = \frac{v_1\Phi^{-1}(p_1) + \cdots + v_k\Phi^{-1}(p_k)}{\sqrt{v_1^2 + v_2^2 + \cdots + v_k^2}},$$

where $v_1,..,v_k$ are nonnegative weights. Note that Z_w has the standard normal distribution. When it exceeds the corresponding critical value of the standard normal distribution, a null hypothesis

is rejected. No general procedure for computing weights has been obtained.

1.2.6 The Logit Method

The method based on logarithm transformation for the *p*-values was proposed by Mudholkar and George (1979). The test statistic was derived as follows

$$L = \log\frac{p_1}{1-p_1} + \cdots + \log\frac{p_k}{1-p_k}.$$

It was difficult to obtain the distribution of *L* and Mudholkar and George (Hedges and Olkin, 1985) showed that the Student's *t*-distribution with *5k+4* degrees of freedom could approximate the distribution of *L* closely up to a constant. They suggested the following test procedure

$$\text{reject } H_0 \text{ if } L^* = |L|\sqrt{\frac{(3/\pi^2)(5k+4)}{k(5k+2)}} > t_{\alpha,5k+4}.$$

For large *k*, $\sqrt{\frac{(3/\pi^2)(5k+4)}{5k+2}} \approx 0.55$ and $L^* = (0.55/\sqrt{k})|L|$.

The weighted modification is

$$L_w = v_1\log\frac{p_1}{1-p_1} + \cdots + v_k\log\frac{p_k}{1-p_k}$$

where $v_i, i = 1,...k$ are nonnegative weights. L_w also has an approximate *t*-distribution. More precisely $L_w = L/\sqrt{c_w}$ has approximate *t*-distribution with *m* degrees of freedom where

$c_w = 3m/(m-2)\pi^2(v_1^2 + \cdots + v_k^2)$ and $m = 4 + 5(v_1^2 + \cdots + v_k^2)^2/(v_1^4 + \cdots + v_k^4)$. The test becomes

$$\text{reject } H_0 \text{ if } L_w \geq t_{\alpha,m}.$$

Both the inverse normal and the logit methods are symmetric in the sense of a *p*-values property. The *p*-values are accumulated about zero in the same way as they are near unity. Both of these tests are appropriate when the direction of deviation from H_0 is not known, i.e. an H_3 type alternative hypothesis.

Comparisons among the above methods involve some "goodness of test" criteria. Two criteria are generally used. The *admissibility* criterion proposed by Birnbaum (1954) consists of two principles: *monotonicity* and *convexity.* A complete discussion is given in Birnbaum (1954) and by Hedges and Olkin (1985).

Another criterion is asymptotic Bahadur optimality (ABO) proposed by Bahadur (1967). The description of ABO using a conception of Bahadur relative efficiency was given by Littell and Folks (1971) and Berk and Cohen (1979).

Bahadur efficiency is formulated (Littell and Folks, 1971) as follows: Let $(x_1, x_2, ...)$ denote an infinite sequence of independent observations of a random variable *X*, whose probability P_θ distribution depends on a parameter $\theta \in \Theta$.

Let *H* be a null hypothesis $H: \theta \in \Theta_0$ and *A* be an alternative $A: \theta \in \theta - \Theta_0$. Let $T_n, n = 1,2,...$ be a real valued test statistic depending on the first *n* observations $x_1, ..., x_n$. Large values of T_n will be considered critical for testing *H.* Assume T_n is continuous, and its probability distribution is the same for all $\theta \in \Theta_0$, and that $F_n(t) = P_\theta\{T_n < t\} = P_0\{T_n < t\}$. The significance level attained by T_n is defined by $L_n = 1 - F_n(T_n)$ and for $\theta \in \Theta_0$, L_n is distributed uniformly on the (0,1). There is a positive valued function $c(\Theta)$, called the *exact slope* of $\{T_n\}$, such that for $\theta \in \Theta - \Theta_0$, $-(2/n)\log L_n \to c(\theta)$ with probability one. Let $\{T_n^{(1)}\}$ and $\{T_n^{(2)}\}$ be two sequences of test statistics with exact slopes $c_1(\theta)$ and $c_2(\theta)$, respectively. The exact Bahadur efficiency of $T_n^{(1)}$ relative to $T_n^{(2)}$ is as the ratio $\varphi_{12}(\theta) = c_1(\theta)/c_2(\theta)$. If $\varphi_{12}(\theta) > 1$, the sequence $\{T_n^{(1)}\}$ is judged superior to $\{T_n^{(2)}\}$ at θ. The calculation of exact slopes is given in Littell and Folks (1971) and Berk and Cohen (1979). Littell and Folks carried out a comparison of four methods: Fisher's method, the normal inverse method, the maximum significance method, and the minimum significance method. (The latter two methods are not discussed in this report). They claimed that according to Bahadur efficiency, the Fisher method is the most efficient.

1.2.7 Lancaster's Method

Berk and Cohen (1979) described some specific methods of combining *p*-values. Lancaster (1961) proposed giving weights to the individual statistics and is ABO. Let

$W = \Gamma_{\alpha_i}^{-1}\left[\frac{1 - L_n(T_n, n_i)}{n_i}\right]$, where Γ_{α_i} is a gamma cumulative distribution function with parameters $(\alpha_i, 1/2)$, where the α_i play a role as weights and the choice of weights is flexible. The statistic is $W = \sum W_i$ such that $W \sim \Gamma(\sum \alpha_i, 1/2)$. Critical values are obtained from chi-square distribution tables if $\sum \alpha_i$ is an integer. Berk and Cohen (1979) claimed that the Lancaster's method is ABO.

They also established that the method proposed by Good (1955) (a weighted Fisher method) is not ABO.

Rosental (1985) compared nine methods of combining independent tests by computing *p*-values obtained from five independent studies. He compared seven basic methods such as Fisher's method, Edgington's method (1972), a method of adding *ts* proposed by Winer (1971), the inverse normal method, the weighted inverse normal method, testing the mean *p* proposed by Edgington (1972), method of testing the mean *Z* proposed by Mosteller and Bush (1954). He also compared two additional methods such as counting and blocking methods.

Results of five methods are presented in Table 1.1 (Rosental, 1985). The first column gives the calculated *t*-statistic. The sign (+) means that the difference was consistent with a majority of the results, the sign (-) means that the difference was not consistent. The second column presents the degrees of freedom for each *t*-test. The third column gives the one-tailed *p* associated with each *t*. The column labeled *Z* is associated with a standard normal deviate for each *p*. The final column presents the natural logarithms of the one-tailed *p*'s in column 3 multiplied by 2 that is, $-2\log p \sim \chi_2^2$.

Table 1-1. Statistics from five independent experiments

Study	***t***	***df***	***p* (one tailed)**	***Z***	$-2\log p$
1	+1.19	40	.12	+1.17	4.24
2	+2.39	60	.01	+2.33	9.21
3	-0.60	10	.72	-0.58	0.66
4	+1.52	30	.07	+1.48	5.32
5	+0.98	20	.17	+0.95	3.54
Σ	+5.48	160	1.09	+5.35	22.97
Mean	+1.10	32	.22	+1.07	4.59
Median	+1.19	30	.12	+1.17	4.24

1.2.8 Fisher's Method

Fisher's test statistic and overall *p*-value is $\chi_2^2(df = 2k) = \sum(-2\log p) = 22.97, p = .006$, and it is a one tailed test.

One disadvantage for a simple sign test (*t* or *Z* columns) is inconsistency. Thus the null hypothesis may be rejected by the sign test if consistent *p*-values are not below .05 by very much. Another property of the Fisher's test is the possibility of supporting significant results in any direction. If two studies show strong significant results in opposite directions, Fisher's method may support the significance of either outcome. Despite all of its limitations (Rosental, 1985), Fisher's method remains the best known and the most discussed of all the methods of combining independent tests.

1.2.9 The Edgington Method

The Edgington method is useful but is limited to small sets of studies, since it requires that the sum of *p*-values do not exceed unity by very much. It gives an overall *p*-value as

$P = \frac{(\sum p)^k}{k!} = \frac{(1.09)^5}{5!} = .006$ and it is also a one tailed test.

1.2.10 The Method of Adding t's

The method of adding *t*'s was proposed by Winer (1971). Winer's test statistic and overall one-tailed *p*-value is

$$Z = \frac{\sum t}{\left\{\sum \left[df/(df-2)\right]\right\}^{1/2}} = \frac{5.48}{\left(40/38 + 60/58 + 10/8 + 30/28 + 20/18\right)^{1/2}} = 2.33,\ p = .01.$$

The method is free of the disadvantages of two methods described above. A limitation is that the method can not be used if the sample size is less than three (division by zero in the denominator).

1.2.11 The Inverse Normal Method

The test statistics for the inverse normal method and its corresponding one-tailed overall *p*-value is $Z = \frac{\sum Z}{k^{1/2}} = \frac{5.35}{5^{1/2}} = 2.39,\ p = 0.009.$

1.2.12 The weighted inverse method

The test statistic for the weighted inverse method and its corresponding one-tailed overall *p*-value is

$$Z = \frac{df_1 Z_1 + df_2 Z_2 + \cdots + df_n Z_n}{(df_1^2 + df_2^2 + \cdots + df_n^2)^{1/2}} = \frac{(40)(1.17) + \cdots + (20)(0.95)}{[40^2 + \cdots + 20^2]^{1/2}} = 3.01,\ p = .0013.$$

Lancaster noted (Rosental, 1985) that when weighting is employed this method is preferable to the weighted Fisher method for reasons of computational convenience and because the final sum obtained is again a normal variable. It also shows the smallest *p*-value.

1.2.13 Method of Testing Mean p

The method of testing proposed by Edgington (1972) uses the mean of the added probabilities values. The test statistic and its corresponding one-tailed overall *p*-value is

$Z = (.50 - \bar{p})(\sqrt{12k}) = (.50 - .22)(\sqrt{12(5)}) = 2.17, p = .015$, where $\bar{p}$ is the mean of *k*

p-values. The presence of 1/12 in the denominator is derived from the fact that the variance of the population of the *p*-values is 1/12 (Rosental, 1978). The test is appropriate for four or more combined studies.

1.2.14 Method of Testing the Mean Z

In the method of testing the mean of *Z*, the test statistic and overall one-tailed *p*-value are

$$t = \frac{\sum Z / k}{\left(MS_Z / k\right)^{1/2}} = \frac{1.07}{\left(.22513\right)^{1/2}} = 2.26, df = 4, p < .05; \text{ or}$$

$t = \frac{\left(\sum Z\right)^2}{k\left(MS_Z\right)} = 5.09, df = 1{,}4, p < .05$. It yields the largest combined *p*-value of all methods.

1.2.15 Counting Method

The binomial model can be used for evaluating the probability of obtaining the results completely by chance (Brozek and Tiede (1952); Jones and Fiske (1953), Wilkinson (1951)). In a series of 15 experiments, the probability of obtaining 3 or more results which exceed the significance level p=0.05 completely by chance can be evaluated as $P = \sum_{j=3}^{15} \binom{15}{j} (0.05)^j (0.95)^{15-j} = 0.036$ and equal to 3.6%, that is less than 5% level of significance.

Thus, if 12 of 15 studies are consistent in either direction, i.e. p-values are less or greater than 0.05, the probability of obtaining 12 consistent results by chance is 3.6%.

The sign test is simple to apply. It can be used as an additional method for probability counting and for checking the consistency of the results.

1.2.16 Blocking Method

The blocking method was suggested by Snedecor and Cochran (1967) (Rosental, 1985) and it requires one to construct the means, sample sizes, and mean squares within each condition for each of the studies and then combine the data into an overall analysis of variance (ANOVA) in which studies are regarded as a blocking variable. Because of differences among studies on their means and variance, it requires one to put the dependent variables on a common scale (e.g.

zero mean and unit variance). The only real disadvantage in this approach is that it may involve more work than some of the other methods especially when there are a large number of studies.

A procedure of choosing an appropriate method depends on special circumstances. Most of the methods described above give satisfactory results. A counting method gives a quick result but it is not powerful. The blocking method often requires too much work without any special benefits. Edgington's method is bounded with small sets of studies but it is preferable for a few studies to the method of testing the mean Z and the counting method. There is no the best method under all conditions (Birnbaum, 1954, Rosental, 1985), but the one that seems the most serviceable under the largest range of conditions is the inverse normal method with or without weighting. The chi-square test might be chosen as the best one since this test is both admissible and ABO (Hedges and Olkin, 1985). When the number of studies is small, the inverse normal method might be suggested and compared with at least two other procedures. When the number of studies is large, it can be combined with one or more of the counting methods to check. It should be mentioned that if *p*-value is very small, it is hard to say anything about the typical size of the examining effect.

CHAPTER 2 - ESTIMATION OF EFFECT SIZE FROM A SINGLE EXPERIMENT

In this chapter estimators of effect size for a single two-group experiment are discussed. Both normally distributed data and binary data are considered. Several different standardizations of the difference in the group means are described in the first section of this chapter. The first section also consider estimators for the absolute difference between means. The second section of the chapter is devoted to estimates of effect size for binary data.

2.1 Normally Distributed Data

2.1.1 Standardized Mean Difference

This section is devoted to several point estimators of the effect size δ from a single two-group experiment. Estimators considered in this section are based on the sample standardized

mean difference for normally distributed data and have identical large sample properties. They differ by constants that depend on the sample size, they also differ in terms of small sample properties (Hedges and Olkin, 1985).

Let $Y_1^E,...,Y_{n^E}^E$ represent the data collected from an experimental group and let $Y_1^C,...,Y_{n^C}^C$ represent the data collected from a control group. Both sets of data are assumed to be distributed normally, so

$$\begin{aligned} Y_j^E &\sim \text{i.i.d. } N(\mu^E, \sigma^2),\ j = 1,...,n^E, \\ Y_j^C &\sim \text{i.i.d. } N(\mu^C, \sigma^2),\ j = 1,...,n^C \end{aligned} \quad . \tag{2.1}$$

The standardized mean difference effect size δ is defined as

$$\delta = (\mu^E - \mu^C)/\sigma. \tag{2.2}$$

The effect size δ is the standardized *z* score of the experimental group mean in the control group distribution, $\Phi(\delta)$ represents the proportion of *control group scores* that are less than the *average score in the experimental group.* For example, if the effect size is $\delta = 0.5$, then $\Phi(\delta) = 0.69$, so that 69% of the individuals in the control group have values that are smaller than the mean of the experimental group. Positive effect size implies the average score in the experimental group is greater than the average score in the control group. Thus the score of the average individual in the experimental group exceeds that of 69% of the individuals in the control group. A negative effect size of $\delta = -0.5$ implies that the only 31% of the individuals in the experimental group have values that exceed the mean of the control group (Hedges and Olkin, 1985).

Another interpretation of effect size is to convert δ to an estimate of a correlation coefficient as to

$$\rho^2 = \frac{\delta^2}{\delta^2 (n_a - n_b - 2)/\tilde{n}}$$

where $\tilde{n} = n_a n_b / (n_a + n_b)$. This is primarily used to summarize the relationship between two continuous variables.

2.1.2 Estimators of Effect Size Based on the Standardized Mean Difference

The idea of estimating an effect size δ with standardized mean difference as

$$(\bar{Y}^E - \bar{Y}^C)/s,$$

where $\bar{Y}^E$ and $\bar{Y}^C$ are the observed experimental and control group sample means, respectively, and s is a standard deviation estimate was proposed by Glass (1976). Different choices of a standard deviation estimate yield different estimators of the effect size.

Glass (1976) proposed to using s^C, the standard deviation of the control group, and then the estimate of effect size is

$$g' = (\bar{Y}^E - \bar{Y}^C)/s^C.$$

The idea of using s^C is obvious when the assumption of different sample standard deviation for each treatment group holds. Indeed different sample standard deviations lead to different estimator values. Assume two treatment groups with the different quantity of the standard deviation s^{E1}, s^{E2}. Using one or the other would yield different values of the estimator g'. For an equal variances case, the estimator might be changed for

$$g = (\bar{Y}^E - \bar{Y}^C)/s, \tag{2.3}$$

where s is the pooled sample standard deviation defined by

$$s = \sqrt{\frac{(n^E - 1)(s^E)^2 + (n^C - 1)(s^C)^2}{n^E + n^C - 2}}$$

where n^E and n^C are the experimental and control group sample sizes, respectively.

For the two sample statistics g and g', Hedges and Olkin (1985) derived their sampling distributions, and showed that they are close to non-central *t*-distributions. They showed that $\sqrt{\tilde{n}}g \sim t'(n^E + n^C - 2, \sqrt{\tilde{n}}\delta)$ and $\sqrt{\tilde{n}}g' \sim t'(n^C - 1, \sqrt{\tilde{n}}\delta)$, where $\tilde{n} = \frac{n^E n^C}{n^E + n^C}$.

It is an important fact (for a detailed discussion see Hedges, 1981) that the bias and variance of g is smaller than that of g'. Therefore g is a (uniformly) better estimator of δ than g' and the latter estimator is omitted from further discussion.

Hedges and Olkin (1985) showed that

$$E(g) \approx \delta + \frac{3\delta}{4\mathrm{N}+9}, \tag{2.4}$$

where $N = n^E + n^C$.

The exact mean is

$$E(g) = \frac{\delta}{J(N-2)}, \tag{2.5}$$

where $J(m)$ is a constant closely approximated by

$$J(m) = 1 - \frac{3}{4\mathrm{m}-1}. \tag{2.6}$$

The variance of g is approximately

$$Var(g) \approx \frac{1}{\tilde{n}} + \delta^2 \frac{1}{2(N-3.94)}. \tag{2.7}$$

It follows from (2.4) that the bias in estimating δ by g turns out to be

$$Bias(g) \approx \frac{3\delta}{(4\mathrm{N}-9)}.$$

For small sample sizes (N<12) the bias is 0.08δ with the bias getting larger as the value of the effect size increases.

2.1.3 An unbiased estimator of effect size

An unbiased estimator of δ is defined by

$$d = J(N-2)g = J(N-2)\frac{\bar{Y}^E - \bar{Y}^C}{s}, \tag{2.8}$$

and $$d \approx \left(1 - \frac{3}{4\mathrm{N}-9}\right)g.$$

Both the bias and the variance of d are smaller than that of g. (Indeed, for $N > 3$, the value $J(N-2) = 1 - 3/(4\mathrm{N}-9)$ is smaller than one.) Therefore d has a smaller mean squared

error than *g*. For $n^E = n^C$, *d* is also a unique minimum variance unbiased estimator (Hedges, 1981). Consequently, for small *N*, *d* turns out to be preferable to *g* as an estimator of δ. For large *N*, *d* and *g* are approximately equal.

2.1.4 The maximum likelihood estimator (MLE) of effect size

The MLE of $\mu^E - \mu^C$ is $\bar{Y}^E - \bar{Y}^C$. The MLE of the pooled within group variance is $\hat{\sigma} = s\sqrt{(N-2)/N}$. Therefore the maximum likelihood estimator $\hat{\delta}$ of the effect size δ is given by

$$\hat{\delta} = \sqrt{\frac{N}{N-2}}\,\frac{\bar{Y}^E - \bar{Y}^C}{s} = \sqrt{\frac{N}{N-2}}\,g \tag{2.9}$$

For large samples, the asymptotic distributions of the estimators *g*, *d*, and $\hat{\delta}$ are approximately normal.

The MLE may be obtained numerically using SAS® PROC GLM as follows

PROC GLM;

MODEL y= treat;

In the output the estimate of effect size turns out to be in the 'treat' statement and the value of s^2 appears as the error mean square.

A shrunken estimator of effect size is defined by (Hedges and Olkin (1985) as

$$\tilde{g} = \frac{N-4}{N-2}\,\frac{g}{J(N-2)} = \frac{N-4}{N-2}\,\frac{d}{\left[J(N-2)\right]^2}. \tag{2.10}$$

It has smaller mean-squared error than *d*.

2.1.5 Comparing parametric estimators of effect size

Four estimators of the effect size have been discussed above. The result of their ordering is as follows

$$\hat{\delta}^2 \geq g^2 \geq d^2 \geq \tilde{g}^2 .$$

The order of their variance is

$Var(\hat{\delta}^2) \geq Var(g^2) \geq Var(d^2) \geq Var(\tilde{g}^2)$.

The best estimator by mean squared error criterion is $\tilde{g}$ (Hedges and Olkin, 1985). The differences among these estimators are largest when the total sample size is small.

2.1.6 Distribution Theory and Confidence Intervals for Effect Sizes.

The asymptotic distribution of estimators of effect size.

Hedges and Olkin (1985) showed that if $\frac{n^E}{N}$ and $\frac{n^C}{N}$ are fixed (i.e. n^E and n^C increase at the same rate), the asymptotic distribution of *d* is

$$d \sim N(\delta, \sigma_\infty^2(d)) \tag{2.11}$$

where $\sigma_\infty^2(d) = \frac{n^E + n^C}{n^E n^C} + \frac{\delta^2}{2(n^E + n^C)}$. (2.12)

This asymptotic distribution can be used to obtain a large sample approximation to the variance of d which is obtained by substituting d for δ in (2.12). The *estimated variance* is

$$\hat{\sigma}^2(d) = \frac{n^E + n^C}{n^E n^C} + \frac{d^2}{2(n^E + n^C)}. \tag{2.13}$$

A $100(1-\alpha)$ percent confidence interval (δ_L, δ_U) for δ is given by

$\delta_L = d - C_{\alpha/2}\hat{\sigma}(d)$ and $\delta_U = d + C_{\alpha/2}\hat{\sigma}(d)$ where $C_{\alpha/2}$ is the two-tailed critical value of the standard normal distribution. These exact and asymptotic distributions were examined and described by Johnson and Welch (1939).

Confidence Intervals for Effect Sizes Based on Transformations.

Since the variance of d depends on the unknown parameter δ (equation (2.12)), one can use the variance-stabilizing transformation

$$h(d) = \sqrt{2}\sinh^{-1}\left(\frac{d}{a}\right) = \sqrt{2}\log\left(\frac{d}{a} + \sqrt{\frac{d^2}{a^2} + 1}\right) \tag{2.14}$$

where $a = \sqrt{4 + 2\left(n^E / n^C\right) + 2\left(n^C / n^E\right)}.$

Denote the transformed value of the estimate by $h = h(d)$ and of the parameter by $\eta = h(\delta)$. Then $\sqrt{N}(h - \eta) \sim N(0,1)$, where $N = n^E + n^C$. Therefore, a $100(1-\alpha)$ percent confidence interval is (η_L, η_U) where

$$\eta_L = h - C_{\alpha/2}\sqrt{N} \text{ and } \eta_U = h + C_{\alpha/2}\sqrt{N} \text{ and}$$

where $C_{\alpha/2}$ is a two-tailed critical value of the standard normal distribution. Thus a confidence interval (δ_L, δ_U) for δ is

$$\delta_L = h^{-1}(\eta_L), \delta_U = h^{-1}(\eta_U),$$

where $h^{-1}(x) = a\sinh\left(x/\sqrt{2}\right).$

Exact confidence intervals for effect sizes

Asymptotic confidence intervals for effect sizes can be used for large sample sizes ($N \geq 20$). For small sample sizes exact confidence intervals are obtained from the exact distribution of the effect size estimator *g*.

$$\sqrt{\frac{n^E n^C}{N}}\, g \sim t\left(N - 2, \delta\sqrt{\frac{n^E n^C}{N}}\right), \quad \text{where } N = n^E + n^C.$$

The cumulative distribution function of *g* has a complicated analytical form. Denote it by $F(g; N-2, \delta)$. Unfortunately it difficult to compute the distribution function by hand. The confidence interval for δ are solutions of the equations

$$F(g; N-2, \delta_L) = \alpha/2, \text{ and } F(g; N-2, \delta_U) = 1 - \alpha/2. \qquad (2.15)$$

2.1.7 Absolute Difference Between Means Estimation

The meta analyses methods applied for investigating the absolute difference of the two mean parameters is a particular point of interest in medicine. The theory of estimating the absolute difference between two mean parameters, the distribution of the estimate, and analyses of obtained results are given in Whitehead (2002).

The absolute difference between means

$$\theta = | \mu^E - \mu^C |$$

is estimated using the likelihood approach and the MLE is to

$$\hat{\theta} = | \bar{Y}^E - \bar{Y}^C |.$$

The variance is given by

$$Var(\hat{\theta}) = \sigma^2 \left(\frac{1}{n^E} + \frac{1}{n^C} \right).$$

The maximum likelihood estimate of $Var(\hat{\theta})$ is $\hat{\sigma}_M^2 = \frac{(n^E - 1)(s^E)^2 + (n^C - 1)(s^C)^2}{n^E + n^C}$.

2.2 Binary Data

A binary variable is scored as either 1 or 0 and is often referred to as a "success" or a "failure". Such an outcome may be recorded for each patient. A typical clinical experiment/study involves two groups; one is a treated group and one is a control group. Outcome data are individual records of the patients in each group. A binary outcome is recorded for each patient. The probability of a success may be denoted by p_E and p_C for the experimental and control groups, respectively. Assume that n_E and n_C subjects are involved in the experimental and the control groups, respectively. The number of successes and failures in each group are denoted by s_E and f_E and s_C and f_C respectively.

Table 2-1. Data for two groups study with a binary outcome

Outcome	Experimental group	Control group	Total
Success	s_E	s_C	s
Failure	f_E	f_C	f
Total	n_E	n_C	n

There are three widely used measures for binary data. One is the probability difference, $p_E - p_C$. A second is the log-odds ratio, $\ln\left(\frac{p_E(1-p_C)}{p_C(1-p_E)}\right)$. And a third is the log-relative risk, $\ln\left(\frac{p_E}{p_C}\right)$. The log-odds ratio is preferred because the corresponding test statistic has the closest asymptotic approximation to a normal and/or a chi-square distribution (Whitehead, 2002).

2.2.1 Log-odds ratio

Let the log-odds ratio be denoted by

$$\theta = \ln\left(\frac{p_E(1-p_C)}{p_C(1-p_E)}\right)$$

which is the log-odds of success on the treatment relative to the control. Methods of analyzing binary data are based on the binomial distribution (Whitehead, 2002). The MLE of the log-odds ratio is commonly obtained by using a linear logistic regression model. The *linear logistic regression model* for binomially distributed data $Y_{ij} = B(p_{ij}, n_{ij}), i = 1,\ldots,m, j = 1,\ldots,k$, with known numbers of Bernoulli trials n_{ij} and unknown probability of success p_{ij} is given by

$$logit(\hat{p}_{ij}) = \ln\left(\frac{\hat{p}_{ij}}{(1-\hat{p}_{ij})}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}$$

where $\hat{p}_{ij} = Y_{ij} / n_{ij}$ and β_0 represents an intercept, and $\beta_j, j = 1,\ldots,k$ are unknown parameters usually estimated by the maximum likelihood method . The x_{ji} denote *explanatory variables* one of which is an indicator variable that represents the treatment received . Suppose the indicator variable is equal to "1" for the treatment group and "0" for the control group.

One possibility of obtaining the MLE for θ is to use SAS®-GENMOD. For this procedure the data for each patient should be entered separately. The response coded as "1" is indicates a "success" and "0 " indicates "failure". The MODEL statement is

MODEL resp = treat / dist = bin link = logit;

The ”dist” option indicates that the distribution of the data is binomial. The “link” option specifies the link function to use in the model. The estimate of θ appears as the “treat” parameter estimate in the output.

Another option is to enter the data as the number of success and the number of trials format. In this case,

MODEL succ/tot = treat / dist = bin link=logit;

is used as the model statement.

The MLE of the sample log-odds ratio is given by

$$\hat{\theta} = \ln\left(\frac{s_E f_C}{s_C f_E}\right). \tag{2.16}$$

The asymptotic estimate of the variance of $\hat{\theta}$, obtained by the delta method, and is

$$Var(\hat{\theta}) = \frac{1}{s_E} + \frac{1}{s_C} + \frac{1}{f_E} + \frac{1}{f_C}. \tag{2.17}$$

An asymptotic two-sided $100(1-\alpha)$percent confidence interval for the parameter θ based on a Wald test is

$$\hat{\theta} \pm z_{\alpha/2}\sqrt{Var(\hat{\theta})}.$$

2.2.2 Probability difference

Now let θ denote a probability difference as

$$\theta = p_E - p_C .$$

The MLE turns out to be

$$\hat{\theta} = \frac{s_E}{n_E} - \frac{s_C}{n_C}. \tag{2.18}$$

The asymptotic estimate of variance derived by delta method is to

$$\hat{Var}(\hat{\theta}) = \frac{s_E f_E}{n_E} + \frac{s_C f_C}{n_C}. \tag{2.19}$$

2.2.3 Log-relative risk

Let θ denote the log-relative risk as to

$$\theta = \ln\left(\frac{p_E}{p_C}\right).$$

The MLE of θ is given by

$$\hat{\theta} = \ln\left(\frac{s_E / n_E}{s_C / n_C}\right). \tag{2.20}$$

The asymptotic estimate of variance derived by delta method is

$$\hat{Var}(\hat{\theta}) = \frac{f_E}{s_E n_E} + \frac{f_C}{s_C n_C}. \tag{2.21}$$

For additional methods that analyze binary data, see Whitehead (2002).

CHAPTER 3 - PARAMETRIC ESTIMATION OF EFFECT SIZE FROM A SERIES OF EXPERIMENTS

In this chapter some methods of obtaining estimates of the standardized mean difference effect size from a series of experiments are discussed. It is assumed that the data are distributed normally.

Suppose a series of *k* studies share a common effect (a standardized difference of two means) δ, it is necessary to have a combined estimate of δ. The sample sizes in these studies may vary from moderate to large.

One method is based on computing the average of the estimated effect size obtained from each study. It is easy to compute a common estimate when all studies have a common sample size. For unequal sample sizes some weighting procedures proposed by Hedges and Olkin (1985) are described. “Optimal” combinations of estimates appear to be (i) a direct weighted linear combination of estimators from different studies; (ii) a maximum likelihood estimator.

Both estimators have the same asymptotic distributions, and therefore they are asymptotically equivalent. Other methods are based on transformations of the effect size estimators.

3.1 Model and Notation

Suppose the data are obtained from a series of *k* independent studies and that each study involves a comparison of an experimental group (E) with a control group (C). The effect size δ proposed by Cohen (1969) was described in Chapter 2. Typical statistical analyses for mean differences involve Student's two-sample *t*-test or an *F*- test. If the assumptions of these tests are met, i.e. the data arise from normal distributions and variances for two groups are equal, the estimator of δ can be computed directly as $\hat{\delta} = \dfrac{\overline{Y}^{E} - \overline{Y}^{C}}{s}$.

Assume that for the *i*th study in the experimental group (E) the observations $Y_{i1}^{E},\ldots,Y_{ik}^{E}$ are distributed normally with a common mean μ_i^E and a common variance σ_i^E, $i = 1,\ldots,k$. Assume also that for the *i*th study the control group (C) observations $Y_{i1}^{C},\ldots,Y_{ik}^{C}$ are distributed normally with a common mean μ_i^C and a common variance σ_i^C, $i = 1,\ldots,k$ as indicated in Table 3.1 and Table 3.2. Table 3.1 lists the experimental observations $Y_{i,j}^{E}, i = 1,\ldots,k,\ j = 1,\ldots,n_i^E$ and the control observations $Y_{i,j}^{C}, i = 1,\ldots,k,\ j = 1,\ldots,n_i^C$ for the *i*th study, $i = 1,\ldots,k$, where n_i^E and n_i^C are the samples sizes in the experimental group and the control group studies, respectively.

Table 3-1 Data arise from a series of *k* experiments, in which each study is a comparison of an experimental group (E) and a control group (C) :

	Observations	
Study	Experimental	Control
1	$Y_{11}^{E},\ldots,Y_{kn_1}^{E}$	$Y_{11}^{C},\ldots,Y_{kn_1}^{C}$
⋮	⋮	⋮
k	$Y_{k1}^{E},\ldots,Y_{kn_k}^{E}$	$Y_{k1}^{C},\ldots,Y_{kn_1}^{C}$

The corresponding parameters for each study such as the mean μ_i^E and the variance σ_i^E, $i = 1,\ldots,k$ for experimental group and mean μ_i^C and variance σ_i^C for control group are presented in Table 3.2. The last column of Table 3.2 lists the effect sizes $\delta_i, i = 1,\ldots,k$ for the *i*th study.

Table 3-2 Parameters such as the mean and the variance for the experimental group and the control group for each study indicated in Table 3:

	Experimental		Control		
Study	Mean	Variance	Mean	Variance	Effect size
1	μ_1^E	σ_1^2	μ_1^C	σ_1^2	$\delta_1 = \left(\mu_1^E - \mu_1^C\right)/\sigma_1$
⋮	⋮	⋮	⋮	⋮	⋮
k	μ_k^E	σ_k^2	μ_k^C	σ_k^2	$\delta_k = \left(\mu_k^E - \mu_k^C\right)/\sigma_k$

In other words,

$$\begin{aligned} Y_{ij}^E &\sim N\left(\mu_i^E, \sigma_i^2\right),\ j = 1,\ldots,n_i^E,\ i = 1,\ldots,k, \\ Y_{ij}^C &\sim N\left(\mu_i^C, \sigma_i^2\right),\ j = 1,\ldots,n_i^C,\ i = 1,\ldots,k \end{aligned}. \tag{3.1}$$

The effect size for the *i*th experiment is given by

$$\delta_i = \left(\mu_i^E - \mu_i^C\right)/\sigma_i\ . \tag{3.2}$$

The assumption that each study measures the same effect implies that $\delta_1 = \delta_2 = \cdots = \delta_k = \delta$.

3.2 Weighted Linear Combinations of Estimates

If the sample sizes of the studies are different, the studies with large sample sizes give more precise estimators of the effect size than the studies with small sample sizes. To obtain a better estimator of the common effect size using data from studies with different sample sizes,

one may use a weighted estimator as

$$d_w = w_1 d_1 + \cdots + w_k d_k \tag{3.3}$$

where $w_1, \ldots, w_k$ are nonnegative weights that sum to unity. Recall that an unbiased estimate of δ from a single study is given by $d = J(N-2)\dfrac{\overline{Y}^E - \overline{Y}^C}{s}$ (see equation (2.8), Chapter 2).

3.2.1 Estimating Weights

It is recommended that the weights be given by $w_i = \dfrac{1}{\sigma^2(d_i)} / \sum_{j=1}^{k} \dfrac{1}{\sigma^2(d_j)}$ where $\sigma^2(d_i)$ is the variance of d_i (see equation (2.13)). Using large sample theory, the weights are

$$w_i = \frac{1}{\sigma_\infty^2(d_i)} / \sum_{j=1}^{k} \frac{1}{\sigma_\infty^2(d_j)} \tag{3.4}$$

where $\sigma_\infty^2(d_i)$ is the large sample variance given in (2.13).

The weights can be approximated by

$$w_i \approx \frac{\tilde{n}_i}{\sum_{j=1}^{k} \tilde{n}_j} \tag{3.5}$$

where $\tilde{n}_i = n_i^E n_i^C / (n_i^E + n_i^C)$. The approximate weights are close to optimal when δ is near zero and the $\tilde{n}_i$ are large.

The weighted estimator of δ is given by

$$d_+ = \sum_{i=1}^{k} \frac{d_i}{\hat{\sigma}^2(d_i)} / \sum_{j=1}^{k} \frac{1}{\hat{\sigma}^2(d_j)} \tag{3.6}$$

where $\hat{\sigma}^2(d_j)$ is defined in (2.13). As stated in Chapter 2, d is an unbiased estimator. The bias of d_+ tends towards zero as the sample sizes get large.

Hedges and Olkin (1985) showed that if the sample sizes of the experimental and control groups in each of the k studies, $n_1^E, \ldots, n_k^E$ and $n_1^C, \ldots, n_k^C$ become large at the same rate

so that n_i^E / N and n_i^C / N remain fixed where $N = n_1^E + \cdots + n_k^E + n_1^C + \cdots + n_k^C$. Then the distribution of d_+ tends to normality with a mean

$$\delta_+ = \sum_{i=1}^{k} \frac{\delta_i}{\sigma_\infty^2(d_i)} / \sum_{i=1}^{k} \frac{1}{\sigma_\infty^2(d_i)} \tag{3.7}$$

and a variance

$$\sigma_\infty^2(d_+) = \left(\sum_{i=1}^{k} \frac{1}{\sigma_\infty^2(d_i)} \right)^{-1} . \tag{3.8}$$

Under the assumption that $\delta_1 = \delta_2 = \cdots = \delta_k = \delta$, $\delta_+ = \delta$ and a $100(1-\alpha)$percent confidence interval for δ turns out to be

$$\delta_L = d_+ - C_{\alpha/2}\hat{\sigma}(d_+), \delta_U = d_+ + C_{\alpha/2}\hat{\sigma}(d_+), \tag{3.9}$$

where $C_{\alpha/2}$ is the two-tailed critical value of the standard normal distribution and $\hat{\sigma}(d_+)$ is the sample estimate of the variance of d_+ given by

$$\hat{\sigma}^2(d_+) = \left(\sum_{i=1}^{k} \frac{1}{\hat{\sigma}_\infty^2(d_i)} \right)^{-1} . \tag{3.10}$$

3.3 The Maximum Likelihood Estimator of Effect Size from a Series of Experiments

Let $\delta_1 = \delta_2 = \cdots = \delta_k = \delta$. The maximum likelihood estimator $\hat{\delta}$ based on observed effect sizes $g_{1,}\ldots, g_k$ defined in (2.3) is the solution of the equation

$$A\hat{\delta} + B_1\sqrt{\hat{\delta}^2 + c_1} + \cdots + B_k\sqrt{\hat{\delta}^2 + c_k} = 0 \tag{3.11}$$

where $A = \tilde{n}_1(2 - L_1) + \cdots + \tilde{n}_k(2 - L_k)$, $B_i = (signg_i)\tilde{n}_i L_i$, $\tilde{n}_i = n_i^E n_i^C / N_i$, $N_i = n_i^E + n_i^C$, $L_i = \tilde{n}_i g_i^2 / (\tilde{n}_i g_i^2 + N_i - 2)$, and $c_i = 4N_i / \tilde{n}_i L_i$, $i = 1,\ldots,k$.

In general, it is not possible to obtain the exact formula for *k>2*. However it is possible to obtain approximate numerical solutions of equation (3.11) using statistical software. Since

the properties of d_+ and $\hat{\delta}$ for large sample sizes are equivalent (Hedges and Olkin, 1985), $\hat{\delta}$ tends to normality (for large samples) with a mean δ and a variance of $\hat{\sigma}^2(\hat{\delta}) = \left(\sum_{i=1}^{k} \frac{1}{\hat{\sigma}_\infty^2(d_i)}\right)^{-1}$.

3.4 Estimators of Effect Size Based on Transformed Estimates

When the sample sizes of both experimental and control groups are equal within each study, i.e. $n_j^E = n_j^C = n_j$, $j = 1,\ldots,k$, then a variance-stabilizing transformation for d is given by

$$h(d) = \sqrt{2}\sinh^{-1}\left(d / 2\sqrt{2}\right). \tag{3.12}$$

Let $h_1 = h(d_1),\ldots, h_k = h(d_k)$ be transformed estimates and $\eta = h(\delta)$ be the transformed effect size parameter. The parameter η is assumed to be the same for all studies. Each of the transformed estimates h_i has an approximate normal distribution with mean η and a variance of $1/(2n_i)$. The linearly weighted estimate of η with the smallest variance (Hedges and Olkin, 1985) is given by

$$h_+ = 2\sum_{i=1}^{k} \frac{n_i h_i}{N} \tag{3.13}$$

where $N = 2\sum n_i$ is the total sample size. A $100(1-\alpha)$ percent confidence interval for η is given by

$$\eta_L = h_+ - C_{\alpha/2}/\sqrt{N}, \eta_U = h_+ + C_{\alpha/2}/\sqrt{N}, \tag{3.14}$$

and a confidence interval for δ is to

$$\delta_L = 2\sqrt{2}\sinh\left(\eta_L/\sqrt{2}\right), \delta_U = 2\sqrt{2}\sinh\left(\eta_U/\sqrt{2}\right). \tag{3.15}$$

3.5 Testing for Homogeneity of Effect Sizes

A statistical test for the homogeneity of effect size is a test of the hypothesis $H_0 : \delta_1 = \delta_2 = \cdots = \delta_k$ versus $H_a : \delta_i \neq \delta_j$ for some $i \neq j$. For large sample sizes the test statistic is

$$Q = \sum_{i=1}^{k} \frac{(d_i - d_+)^2}{\hat{\sigma}^2(d_i)} \tag{3.16}$$

where $\hat{\sigma}^2(d_i)$ is defined in (2.13). If all k studies have the same effect size, i.e. H_0 is true, then $Q \sim \chi^2(k-1)$ (Hedges and Olkin, 1985). Therefore to produce a statistical test or construct a confidence interval, one can use a critical value from the χ^2 distribution with k-1 degrees of freedom.

The statistic Q may be obtained by using the weighted least-squares regression method which is available in SAS® package as follows:

```
PROC GLM;
MODEL y= / inverse;
WEIGHT w;
```

where $w_i = 1/\hat{\sigma}^2(d_i)$ and w is a $k \times k$ matrix whose diagonal elements consist of the w_i's. There is no variable in the right hand side of the `MODEL` statement which implies that the "intercept" value in the output is equal to d_+. The `inverse` option displays the matrix of $(X^T W X)^{-1}$ where for this case X is a $k \times 1$ vector with components equal to 1. The `WEIGHT` option requests minimization of a weighted residual sum of squares.

3.5.1 Small Sample Significance Levels for the Homogeneity Test Statistics

For small sample sizes an exact test statistic is unknown. The Q-test is accurate when the sample sizes are at least 10 per group. See Hedges and Olkin (1985).

3.5.2 Other Procedures for Testing Homogeneity of Effect Sizes

Since the likelihood ratio test involves rather difficult calculations, Hedges and Olkin (1985) recommend that one should use the Q-test.

If the groups in each study have the same size, i.e. an experiment is balanced, one can use a transformation method. Let $a = 1$ in (2.14). Then transform $d_1, \ldots, d_k$ to $h_1, \ldots, h_k$ and $\delta_1, \ldots, \delta_k$ to $\eta_1, \ldots, \eta_k$ via

$$h_i = \sqrt{2}\sinh^{-1}\left(d_i / 2\sqrt{2}\right) \text{ and } \quad \eta_i = \sqrt{2}\sinh^{-1}\left(\delta_i / 2\sqrt{2}\right) \tag{3.17}$$

The equality of $\delta_1,\ldots,\delta_k$ is equivalent to the equality of $\eta_1,\ldots,\eta_k$. To test $H_0 : \delta_1 = \cdots = \delta_k$ vs. $H_a : \delta_i \neq \delta_j$, for some $i \neq j$, calculate

$$Q_1 = 2\sum_{i=1}^{k} n_i\left(h_i - h_+\right)^2, \text{ then} \tag{3.18}$$

reject H_0 if $Q_1 > C$, where C is a critical value obtained from chi-square distribution with k-1 degrees of freedom.

3.6 Estimation of Effect Size for Small Sample Sizes

The large sample theory is not accurate for sample sizes less than 10. Another option for obtaining asymptotic results is to use a large number of studies. This requires a different version of normal theory. While the results are not the same as the results obtained for large sample sizes, they are very close.

There are several methods to estimate the effect size from a large series of studies when each study has small sample size.

3.6.1 Estimation Effect Size from a Linear Combination of Estimates

One of the simplest methods of estimating a common effect size is based on a weighted mean. The weighted mean with the smallest variance (Hedges and Olkin, 1985) is given by

$$\bar{d}_w = w_1 d_1 + \cdots + w_k d_k \tag{3.19}$$

where the optimal weights are given by

$$w_i = \frac{1}{v_i\left(\bar{d}\right)} \Big/ \sum_{j=1}^{k} \frac{1}{v_j\left(\bar{d}\right)} \tag{3.20}$$

where $\bar{d}$ is the mean of $d_1,\ldots,d_k$,

$$v_i\left(\bar{d}\right) = a_i + b_i \bar{d}^{\,2}, \tag{3.21}$$

$a_i = (N-2)[J(N_i-2)]^2 / [\tilde{n}_i(N_i-4)]$ and $b_i = \{(N_i-2)[J(N_i-2)]^2 - (N_i-4)\}/[(N_i-4)]$ (3.22)

and *J(m)* is given in equation (2.6).

Hedges and Olkin (1985) noted that $\bar{d} \sim N(\delta, v)$ where *v* is the estimated variance given by $v = \left(\sum_{k=1}^{k} \frac{1}{v_i(\bar{d})} \right)^{-1}$. (3.23)

A $100(1-\alpha)$ percent confidence interval for the effect size δ is given by

$\delta_L = \bar{d}_w - C_{\alpha/2}\sqrt{v}$ and $\delta_U = \bar{d}_w + C_{\alpha/2}\sqrt{v}$. (3.24)

CHAPTER 4 - PARAMETRIC FIXED EFFECT MODELS

4.1 Categorical Models

4.1.1 Normally Distributed Data

Model and Notation

Assume that the studies are sorted into *p* disjoint classes and that there are m_i studies in the *i*th class, $i = 1, \ldots, p$. Let Y_{ijl}^E and Y_{ijl}^C be the *l*th experimental and control group observations in the *j*th experiment in the *i*th class. Sample sizes of the experimental and the control groups for the *j*th study in the *i*th class are denoted by n_{ij}^E and n_{ij}^C, respectively. The set of observations, parameters, and their estimators are described above and summarized in Table 4.1.

Table 4-1. Parameters and Estimates for the Control and Experimental Groups

Class	Study	Experimental				Control			
		Parameters		Estimates		Parameters		Estimates	
		Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
1	1	μ_{11}^E	σ_{11}^2	$\bar{Y}_{11}^E$	$\left(s_{11}^E\right)^2$	μ_{11}^C	σ_{11}^2	$\bar{Y}_{11}^C$	$\left(s_{11}^C\right)^2$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	m_1	$\mu_{1m_1}^E$	$\sigma_{1m_1}^2$	$\bar{Y}_{1m_1}^E$	$\left(s_{1m_1}^E\right)^2$	$\mu_{1m_1}^C$	$\sigma_{1m_1}^2$	$\bar{Y}_{1m_1}^C$	$\left(s_{1m_1}^C\right)^2$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
p	1	μ_{p1}^E	σ_{p1}^2	$\bar{Y}_{p1}^E$	$\left(s_{p1}^E\right)^2$	μ_{p1}^C	σ_{p1}^2	$\bar{Y}_{p1}^C$	$\left(s_{p1}^C\right)^2$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
p	m_p	$\mu_{pm_p}^E$	$\sigma_{pm_p}^E$	$\bar{Y}_{pm_p}^E$	$\left(s_{pm_p}^E\right)^2$	$\mu_{pm_p}^C$	$\sigma_{pm_p}^2$	$\bar{Y}_{pm_p}^C$	$\left(s_{pm_p}^C\right)^2$

Suppose that

$$\begin{aligned} Y_{ijl}^E &\sim N(\mu_{ij}^E, \sigma_{ij}^2), l = 1,\ldots,n_{ij}^E,\ j = 1,\ldots,m_i, i = 1,\ldots,p \\ Y_{ijl}^C &\sim N(\mu_{ij}^C, \sigma_{ij}^2), l = 1,\ldots,n_{ij}^C,\ j = 1,\ldots,m_i, i = 1,\ldots,p \end{aligned} \quad \text{and} \qquad (4.1)$$

The effect size for the *j*th experiment in the *i*th class is given by (Hedges and Olkin (1985)) $\delta_{ij} = \left(\mu_{ij}^E - \mu_{ij}^C\right)/\sigma_{ij}$. (4.2)

Three methods of testing hypotheses are considered:

i) The studies from *different* classes share a common but unknown effect size δ . An hypothesis of interest is

$$H_0: \begin{array}{l} class1: \delta_{11} = \delta_{12} = \cdots = \delta_{1m_1} = \delta, \cdots, \\ classp: \delta_{p1} = \delta_{p2} = \cdots = \delta_{pm_p} = \delta \end{array} . \qquad (4.3)$$

ii) The effect sizes *within* classes are equal, but are not the same for all classes. A hypothesis of interest might be

$$H_1: \begin{array}{l} class1: \delta_{11} = \delta_{12} = \cdots = \delta_{1m_1} = \delta_1, \cdots, \\ classp: \delta_{p1} = \delta_{p2} = \cdots = \delta_{pm_p} = \delta_p \end{array}. \quad (4.4)$$

iii) All effect sizes may be different. In this case the hypothesis is given by

$$H_2: \quad \delta_{ij} \text{ unrestricted.} \quad (4.5)$$

The test of H_1 vs. H_2 is a test of homogeneity of effect size *within* classes. The test of H_0 vs. H_1 is a test of homogeneity *between* classes, given that there is a homogeneity within classes.

An unbiased estimator of the effect size δ_{ij} is given by

$$d_{ij} = J\left(N_{ij} - 2\right)\frac{\bar{Y}_{ij}^{E} - \bar{Y}_{ij}^{C}}{s_{ij}},\ j = 1,\ldots,m_i,\ i = 1,\ldots,p,\ N_{ij} = n_{ij}^{E} + n_{ij}^{C} \quad (4.6)$$

where s_{ij} is estimated pooled sample standard deviation.

For large sample sizes $d_{ij} \sim N\left(\delta_{ij}, \sigma_{\infty}^2\left(d_{ij}\right)\right)$, where the asymptotic variance is given by

$$\sigma_{\infty}^2(d_{ij}) = \frac{n_{ij}^{E} + n_{ij}^{C}}{n_{ij}^{E} n_{ij}^{C}} + \frac{\delta_{ij}^2}{2(n_{ij}^{E} + n_{ij}^{C})}, \quad (4.7)$$

and the asymptotic variance is estimated by

$$\hat{\sigma}^2(d_{ij}) = \frac{n_{ij}^{E} + n_{ij}^{C}}{n_{ij}^{E} n_{ij}^{C}} + \frac{d_{ij}^2}{2(n_{ij}^{E} + n_{ij}^{C})}. \quad (4.8)$$

4.1.2 Some Tests of Homogeneity

4.1.2.1 Testing homogeneity of effect sizes across classes when all studies have a common effect

For a test of H_0 versus H_2 the test statistic is given by

$$Q_T = \sum_{i=1}^{p}\sum_{j=1}^{m_i}\frac{(d_{ij} - d_{++})^2}{\hat{\sigma}^2\left(d_{ij}\right)} \quad (4.9)$$

where $\hat{\sigma}^2(d_{ij})$ is defined in (4.8) and

$$d_{++} = \sum_{i=1}^{k}\sum_{j=1}^{m_i} \frac{d_{ij}}{\hat{\sigma}^2(d_{ij})} / \sum_{i=1}^{k}\sum_{j=1}^{m_i} \frac{1}{\hat{\sigma}^2(d_{ij})}. \tag{4.10}$$

The approximate distribution of Q_T is $Q_T \sim \chi^2(m_1 + m_2 + \cdots + m_p - 1)$.

4.1.2.2 Testing homogeneity of effect sizes across classes

To test H_0 versus H_1, the between class goodness of fit test statistic is

$$Q_B = \sum_{i=1}^{p} \frac{(d_{i+} - d_{++})^2}{\hat{\sigma}^2(d_{i+})} = \sum_{i=1}^{p}\sum_{j=1}^{m_i} \frac{(d_{i+} - d_{++})^2}{\hat{\sigma}^2(d_{ij})} \tag{4.11}$$

where

$$d_{i+} = \sum_{i=1}^{m_i} \frac{d_{ij}}{\hat{\sigma}^2(d_{ij})} / \sum_{j=1}^{m_i} \frac{1}{\hat{\sigma}^2(d_{ij})}. \tag{4.12}$$

The approximate distribution of Q_B is $Q_B \sim \chi^2(p-1)$.

4.1.2.3 Testing homogeneity of effect sizes within classes

To test H_1 versus H_2, the within class goodness of fit test statistic is

$$Q_W = \sum_{i=1}^{p}\sum_{j=1}^{m_i} \frac{(d_{i+} - d_{i+})^2}{\hat{\sigma}^2(d_{ij})}. \tag{4.13}$$

The approximate distribution of Q_W is $Q_W \sim \chi^2((m_1 - 1) + \cdots + (m_p - 1))$.

Since $Q_T = Q_B + Q_W$ and since each of these statistics has a chi-square distribution, one can obtain a summary table that is analogous to an Analysis of Variance table. See Table 4.2 where $k = m_1 + m_2 + \cdots + m_p$

Table 4-2. An Analogy to an Analysis of Variance table

Source	Statistics	Degrees of freedom
Between classes	Q_B	p-1
Within classes	Q_W	k-p
Total	Q_T	k-1

4.2 Meta Analysis for Fixed Effect Models Based on Individual Patient Data

Traditional meta-analysis methods described in Chapters 2 and 3 are also available when one is lucky enough to have individual patient data. Typical statistical approaches on modeling when one has individual patient data are based on likelihood theory (Whitehead, 2002). For individual patient data, meta-analyses models are extensions of linear models for a single study. Numerical analyses may be conducted by using SAS® as a statistical package. In this section, both normally distributed data and binary data are considered. The theory of obtaining analytical expressions for likelihood statistics is omitted in this report in favor of application examples related to clinical trials and using the SAS® package.

4.2.1 Normally Distributed Data

4.2.1.1 Model and Notation

Let the random variables Y_{ij} be normally distributed with means μ_{ij} and common variance σ^2. That is, $Y_{ij} \sim N\left(\mu_{ij}, \sigma^2\right), i = 1,\ldots,n_r, j = 1,\ldots,n_i$. Let y_{ij} denote the response (observation) from patient *j* in study *i*, moreover, let $n = \sum_{i=1}^{r} n_i$ be the total number of patients in all studies. The general linear model is

$$y_{ij} = \mu_{ij} + \varepsilon_{ij}$$

where ε_{ij} are the error terms that are distributed normally $\varepsilon_{ij} \sim N\left(0, \sigma^2\right)$.

Without loss of generality, assume that

$$\mu_{ij} = \alpha + \eta_{ij}$$

where α represents an intercept. Also suppose that $\beta_k, k = 1,\ldots,q$ are unknown parameters and that $\eta_{ij} = \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_q x_{qij}$. Explanatory variables $x_{kij}, k = 1,\ldots,q$ can be quantitative variables such as age. They also can be qualitative factors and have fixed factor levels. For example, if a qualitative variables x_{kij} represents a particular study, only two levels such as "1" and "0" are needed.

The model provides the fixed effect of the absolute mean difference between the two treatment

$$\theta = \left|\mu^E - \mu^C\right|$$

defined in (0.3) is given by

$$\mu_{ij} = \alpha + \beta_{0i} + \beta_1 x_{1ij}, \tag{4.14}$$

where α represents the effect in the control group in *r*th study, $\alpha + \beta_{0i}$ represents the effect in the control group since x_{1ij}=0 for the control group in the *i*th study, and β_1 represents the absolute treatment mean differences between experimental and control groups since x_{1ij}=1 for the treatment group .

4.2.1.2 Estimation and Hypothesis Testing

The null hypothesis is that the treatment difference in all studies is 0, i.e. $\theta = 0$ and in terms of model (4.14) , $\beta_1 = 0$. Therefore, model (4.14) is compared to the model

$$\mu_{ij} = \alpha + \beta_{0i}. \tag{4.15}$$

Model (4.14) is called the full model with *r*+1 degrees of freedom, and (4.15) is the reduced model with *r* degrees of freedom (Whitehead, 2002). The estimator of σ^2 has *n*-*r*-1 degrees of freedom. Therefore the F test for comparing the full model to the reduced model is $\frac{SSE(R) - SSE(F)}{1} / \frac{SSE(F)}{n-r-1} \sim F(1, n-r-1)$. To obtain test results numerically, one may use the SAS®-GLM procedure as

PROC GLM;

CLASS study;

MODEL y= study treat / ss1 solution;

where "treat" represents x_{1ij} which is the explanatory variable defined in models (4.14), (4.15). The solution option allows one to obtain the parameter estimates, standard errors, and the estimate of β_1 appears as the "treat" parameter estimate. It is also possible to include "treat" into the CLASS statement.

4.2.1.3 Testing for Heterogeneity in the Absolute Mean Difference Across Studies

For testing the treatment difference parameter θ across all studies the model is

$$\mu_{ij} = \alpha + \beta_{0i} + \beta_{1i} x_{1ij}, \tag{4.16}$$

where β_{1i} varies from study to study. The F statistic has *r-1* d.f. in the numerator and *n-2r* d.f. in denominator. Using SAS®-GLM, the commands are

PROC GLM;

MODEL y=study treat study*treat / ss1 solution;

where the desired *F* statistic is associated with the "study*treat" term.

4.2.2 Binary Data

Model and Notation

Let the random variables Y_{ij} be distributed binomial such as $Y_{ij} \sim B(n_{ij}, p_{ij}), i = 1\ldots, r, j = 1,\ldots, k,$ and let Y_{ij} be the number of successes for the *j*th treatment in the *i*th study , and let $n = \sum_{i=1}^{r} n_i$. The parameters p_{ij} represent the probability of success for a patient in the *j*th treatment group in the *i*th study. The model that yields an overall fixed effect estimate of treatment difference (Whitehead 2002) is to

$$\ln\left(\frac{\hat{p}_{ij}}{(1-\hat{p}_{ij})}\right) = \alpha + \beta_{0i} + \beta_1 x_{1ij} \tag{4.17}$$

where $\hat{p}_{ij} = Y_{ij} / n_{ij}$ and β_1 represents the common log-ratio of success on treatment relative to control.

4.2.2.1 Estimation and Hypotheses Testing

Parameters are estimated by using the maximum likelihood method. PROC GENMOD in SAS® that fits a linear logistic regression model is appropriate.

To test the absolute difference between treatment means, one has to state the null hypothesis which is $H_0 : \theta = 0$ which implies there is no difference versus the alternative that is $H_a : \theta \neq 0$. The reduced model is defined by

$$\ln\left(\frac{\hat{p}_{ij}}{(1-\hat{p}_{ij})}\right) = \alpha + \beta_{0i}, i = 1,2,\ldots,r \qquad (4.18)$$

and the likelihood ratio test may be obtained by using

PROC GENMOD;

CLASS study;

MODEL y= study treat / type1 dist = bin link = logit waldci;

The parameter β_1 is associated with "treat" in the output, "waldci" option gives a Wald CI, the "lrci" option might also be used to obtain CIs based on the maximum likelihood method.

Another possibility to enter data is a binomial form. For each treatment group in each study the total number of patients *n* is available as well as the total number of successes, $y(s)_{ij}$. The MODEL statement in this case appears to be

MODEL s/n = study treat / type1 dist = bin link = logit waldci; .

4.2.2.2 Testing for Heterogeneity in the Log-odds Ratio Across Studies

An appropriate model for testing heterogeneity of the treatment difference parameter across studies includes the study by treatment interaction term that is

$$\ln\left(\frac{\hat{p}_{ij}}{(1-\hat{p}_{ij})}\right) = \alpha + \beta_{0i} + \beta_{1i}x_{1i} . \qquad (4.19)$$

This test makes a comparison between models (4.17) and (4.19) using likelihood method and SAS®-GENMOD, the MODEL statements are

MODEL y=study treat study*treat / type1 dist = bin link = logit;

the $\chi^2(r-1)$ statistics are associated with the "study*treatment" term in the output, β_{1r} represents the "treat" effect and $\beta_{1i} - \beta_{1r}$ relates to "study i * treat".

CHAPTER 5 - RANDOM EFFECT MODELS FOR EFFECT SIZES

In this chapter a brief description of the process of estimating the standardized effect size for random models is given. The theory of obtaining estimators and hypotheses tests as well as confidence intervals for desired parameters is very close to the theory of obtaining estimates of effect sizes for the fixed effect models described in Chapters 2-4. It is assumed that the data are distributed normally. An example using SAS® to obtain numerical results is given.

As previously mentioned in both the Introduction and Chapter 2 of this report, Cohen (1969) proposed a population measure δ of effect size in connection with the *t- test* for the difference between means. Glass (1976) proposed *g* as the quantitative estimator of the results of a collection of experimental/control group studies by estimating δ for each study. Assume that requirements for the validity of the two-sample *t*- test are met by each study.

In the previous chapters the effect sizes $\delta_1, \ldots, \delta_k$ were assumed to be fixed but unknown parameters. In this Chapter the effect sizes $\delta_1, \ldots, \delta_k$ are treated the same way and δ_i is considered as a population parameter for the *i*th study. At the same time $\delta_1, \ldots, \delta_k$ are "sample realizations" of the random variable Δ because the studies are considered as a sample from a population of studies with a distribution of δ_i values.

5.1 Model and Notation

Suppose (Hedges, 1983) that the data arise from a series of *k* independent studies, where each study compares an experimental group (E) with a control group (C). Let Y_{ij}^E and Y_{ij}^C be the *j*th observations from the *i*th experiment for the experimental and control groups, respectively. Assume that for fixed *i*, Y_{ij}^E and Y_{ij}^C are independently normally distributed such as

$$Y_{ij}^{E} \sim N\left(\mu_i^{E}, \sigma_i^{E}\right),\ j = 1,\ldots,n_i^{E},\ i = 1,\ldots,k, \tag{5.1}$$

$Y_{ij}^{C} \sim N\left(\mu_i^{C}, \sigma_i^{C}\right),\ j = 1,\ldots,n_i^{C},\ i = 1,\ldots,k$ as presented in Table 3.1.

The effect size for the *i*th study was defined in Chapter 2 by

$$\delta_i = \frac{\left(\mu_i^{E} - \mu_i^{C}\right)}{\sigma_i}. \tag{5.2}$$

An unbiased estimator d_i of the effect size (4.2) is

$$d_i = J\left(N_i - 2\right)\left(\bar{Y}_i^{E} - \bar{Y}_i^{C}\right)/ s_i \tag{5.3}$$

where $N_i = n_I^{E} + n_I^{C}, \bar{Y}_i^{E}$, $\bar{Y}_i^{C}$, and s_i are the experimental and control group sample sizes, means and the pooled within group standard deviation from the *i*th study, respectively, and *J(m)* is the correction factor defined in (2.6).

5.2 Estimating the Mean Effect Size

Let the mean effect size, that is the mean of the populations of δ, be denoted by $\bar{\Delta}$. The most precise weighted estimator $w_1 d_1 + \cdots + w_k d_k$ of $\bar{\Delta}$ has weights as

$$w_i\left(\Delta,\delta\right) = \frac{1}{v_i^2} / \sum_{j=1}^{k} \frac{1}{v_j^2} \tag{5.4}$$

where $v_i^2 = \sigma^2\left(\Delta\right) + \sigma^2\left(d_i \mid \delta_i\right), i = 1,\ldots,k$ and (5.5)

$$\sigma^2\left(d_i \mid \delta_i\right) = a_i / \tilde{n}_i + \left(a_i - 1\right)\delta_i^2, \tag{5.6}$$

where $\tilde{n}_i = n_i^{E} n_i^{C} / N_i, N_i = n_i^{E} + n_i^{C}$, and $a_i = \left(N_i - 2\right)\left[J\left(N_i - 2\right)^2 \right] / \left(N_i - 4\right)$. (5.7)

Since the parameters $\sigma\left(\Delta\right)$ and $\delta_1,\ldots,\delta_k$ are unknown, it is necessary to estimate the weights in (4.4). The estimated weights are given by

$$\hat{w}_i\left(\Delta,\delta\right) = \frac{1}{\hat{v}_i^2} / \sum_{j=1}^{k} \frac{1}{\hat{v}_j^2} \tag{5.8}$$

where $\hat{v}_i^2 = \sigma^2(\Delta) + \hat{\sigma}^2(d_i \mid \delta_i), i = 1,\ldots,k$, (5.9)

$$\hat{\sigma}^2(d_i \mid \delta_i) = a_i / \tilde{n}_i + (a_i - 1)d_i^2, \quad (5.10)$$

a_i is defined in (4.7), and

$$\hat{\sigma}^2(\Delta) = \sum_{i=1}^{k} \frac{(d_i - \bar{d})^2}{k-1} - \frac{1}{k}\sum_{i=1}^{k} a_i / \tilde{n}_i + (a_i - 1)/a_i d_i^2 = s^2(d) - \frac{1}{k}\left(a_i / \tilde{n}_i + (a_i - 1)/a_i d_i^2\right). \quad (5.11)$$

The test procedure is similar to the one described in Chapter 3 where the formula of the test statistic involves $\hat{\sigma}^2(d_i \mid \delta_i)$instead of $\hat{\sigma}^2(d_i)$. Discussion and details are given in Hedges and Olkin (1985).

There are a lot of applications using random effect models for medical problems given in Whitehead (2002). She not only discusses different types of test procedures applicable for different models but also writes SAS® code with detailed explanations for the models.

One example of a simple meta analysis based on individual patient data is given in Higgins *et.al* (2001). They discussed a two-level model so that patients correspond to level one units and trials corresponds to level two units. Observations y_{ij} denote the outcome of patient *j* in trial *i*. The variable x_{1ij} represents a treatment group with a value of 1 for the treated group and 0 for the control group.

The random effects meta analysis model for normally distributed responses y_{ij} is given by

$$y_{ij} = \alpha + \beta_{0i} + \beta_1 x_{1ij} + v_{1i} x_{1ij} + \varepsilon_{ij} \quad (5.12)$$

where $v_{1i} \sim N(0, \sigma_\tau^2)$and $\varepsilon_{ij} \sim N(0, \sigma_i^2)$are random terms corresponding to level two and level one, respectively (Higgins *et.al* 2001). The covariances between different levels are assumed to be zero. Model (4.12) is a general linear mixed model.

For example, to analyze the model in (4.12) with SAS®–MIXED, one can use

PROC MIXED;

CLASS study;

MODEL y = study treat / htype = 1 ddfm = kenwardroger solution;

RANDOM treat / subject = study;

The fixed term appear to be in the MODEL statement and the random effect term is in the RANDOM statement. The "subject= study" option indicates that the random effect "treat" varies from study to study.

CHAPTER 6 - VOTE-COUNTING METHODS

The conventional vote-counting or box-score methods synthesize results across studies by sorting studies into categories and counting outcomes (consistent or not) of tests of hypotheses found in literature. Like combining independent tests described in Chapter 1, vote-counting methods require little information about the individual studies. The idea is based on knowing the signs of mean differences or correlations or an assumption that a hypothesis test yields a statistically significant result (Hedges and Olkin, 1985). All studies are divided into three categories: the first one contains those studies yielding significant results with a positive mean difference, the second category contains those studies yielding significant results with a negative mean difference, and the third category contains those studies that did not yield a significant result.

In this chapter methods of obtaining confidence intervals for parameters based on asymptotic theory (Hedges and Olkin, 1985) and methods yielding exact confidence intervals for parameters (Molenaar, 1970, Blom, 1954) are described. Estimators of effect size defined for vote-counting methods are given.

6.1 Preliminaries

Suppose that one wants to integrate k independent "identical" studies. Suppose a statistic T (for instance Student's t-test statistic) can be obtained for each study. Assume that the standardized mean difference is the same for all k studies. A positive significant result occurs if a trial is a success, a negative significant result or no significant result implies a failure of a trial. The probability that a study yields a positive significant result is

$$p \equiv \Pr\{signific.result \mid \delta, n\} = \int_{C_\alpha}^{\infty} f(t;\delta,n)dt$$

where $f(t;\delta,n)$ is the probability density function of the statistic *T* in samples of size *n* with effect magnitudes δ, and the critical value C_α of the statistic *T*. It is known the number of successes has a binomial distribution.

An effect δ turns out to be positive if the proportion of the studies with positive significant results is greater than 1/3 (the cutoff value C_0). Let *X* be the number of success,

$$\Pr\{proportionofsuccess > C_0\} = \Pr\left\{\frac{X}{k} > C_0\right\} = \sum_{i=\lfloor C_0 k \rfloor + 1}^{k} \binom{k}{i} p^i (1-p)^{k-i},$$

where [*a*] is the greatest integer less than or equal to *a*, $0 \le C_0 \le 1$.

Assume one wants to estimate a common parameter θ for all *k* studies. One obtains *k* test statistics $T_1,\ldots,T_k$ which represent *k* parameters $\theta_1,\ldots,\theta_k$. The null hypothesis for the *i*th study is $H_{0i}: \theta_i = 0$. One rejects H_{0i} if $T_i > C$ where *C* is a critical value obtained from the distribution of T_i. Usually the test statistics $T_1,\ldots,T_k$ are not known. The only known information that is known is the number *U* of successful results (positive result, null hypothesis is rejected) and the number of failures (negative result, null hypothesis is not rejected) in the *k* independent trials. Therefore the sample proportion of successes U/k is available, that is (Hedges and Olkin, 1985) the maximum likelihood estimate of the probability $p_C(\theta)$ of success is U/k. The maximum likelihood estimator $\hat{\theta}$ of θ is obtained from the maximum likelihood estimator of $p_C(\hat{\theta})$ by solving the equation $p_C(\hat{\theta}) = U/k$ for $\hat{\theta}$. Since the power function $p_C(\theta)$ is a monotone function of θ, confidence intervals for $p_C(\theta)$ can be translated to confidence intervals for θ.

6.2 Confidence Intervals for Parameters

There are several methods for obtaining confidence intervals for the parameter $p_C(\theta)$ (Hedges and Olkin, 1985). One approach uses simpler asymptotic theory for the distribution of *U/k* based on the large sample normal approximation to the binomial distribution. Another

approach based on approximations to the distribution of *U/k* allows obtaining exact confidence intervals for $p_C(\theta)$. Different approximation methods to the distributions of *U/k* obtained by different researchers independently are described in Molenaars's monograph (1970). One of the methods for obtaining exact confidence intervals for a desired parameter is given in Blom (1954).

6.2.1 Confidence Intervals Based on Asymptotic Theory

6.2.1.1 Use of normal theory

Any consistent estimator of p, $\hat{p}$ may be used to estimate the variance of $\hat{p}$, $p(1-p)/k$ by $\hat{p}(1-\hat{p})/k$ and a $100(1-\alpha)$ percent confidence interval (p_L, p_U) for p is $p_{L,U} = \hat{p} \mp C_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/k}$ where $C_{\alpha/2}$ is the two-tailed critical value of the standard normal distribution. The confidence interval (p_L, p_U) for *p* can be translated to the confidence interval (θ_L, θ_U) for θ by solving $p_C(\theta_L) = p_L$ and $p_C(\theta_U) = p_U$ (Hedges and Olkin 1985).

6.2.1.2 Use of chi-square theory

It is a well known fact that $z^2 = k(\hat{p}-p)^2 / p(1-p) \sim \chi^2(1)$ for large *k* (Hedges and Olkin, 1985). To obtain a confidence interval for *p* one needs to solve the equation

$$k(\hat{p}-p)^2 / p(1-p) = C_\alpha \quad \text{for } p$$

where C_α is the upper critical value of the chi-square distribution with one degree of freedom. Set $g(\tilde{p}) = C_\alpha / k$. An analytical solution allows obtaining two points $\tilde{p}_L, \tilde{p}_U$ is as follows

$$\tilde{p}_L = \frac{2\hat{p} + b - \sqrt{b^2 + 4b\hat{p}(1-\hat{p})}}{2(1+b)} \text{ and } \tilde{p}_U = \frac{2\hat{p} + b + \sqrt{b^2 + 4b\hat{p}(1-\hat{p})}}{2(1+b)} \text{ where } b = C_\alpha / k.$$

6.3 Estimating an Effect Size

Let each study consist of two groups: an experimental (E) group and a control (C) group that have the same sample sizes such as $n_i^E = n_i^C = n$ for the whole collection of *k* studies. Let $T_i > C$ for all *k* studies. Let Y_{ij} denote the score of the individual *j* in the *i*th study. Assume that

$Y_{ij}^E \sim N(\mu_i^E, \sigma_i^2)$ and $Y_{ij}^C \sim N(\mu_i^C, \sigma_i^2), i = 1,...,k, j = 1,2,...,n.$ The effect size δ_i for the *i*th experiment is $\delta_i = \dfrac{\mu_i^E - \mu_i^C}{\sigma_i}, i = 1,...,k.$

Assume that the effect size is the same for all studies $\delta_1 = \cdots = \delta_k = \delta.$ The estimate of δ_i is the Glass effect size defined by

$$g_i = \frac{\bar{Y}_i^E - \bar{Y}_i^C}{s_i}, i = 1,...k,$$

where $\bar{Y}_i^E$ and $\bar{Y}_i^C$ are the experimental and control sample means, and s_i is the pooled within group sample standard deviation in the *i*th experiment. Then

$t_i = g_i\sqrt{n/2} \sim t(2n-2, \delta\sqrt{n/2})$. To estimate an effect size, one counts the number of times that $t_i > C_\alpha$. Thus (Hedges and Olkin, 1985) the probability $p_C(\delta)$ of success is the probability that a noncentral *t*-variate exceeds C_α. For example, if $\alpha = 0.50$, the critical value $C_{0.5} = 0.0$, then g_i are positive.

6.4 Limitations of the vote-counting estimators.

The estimators have several limitations that restrict their application (Hedges and Olkin, 1985). One limitation relates to the asymptotic theory that holds as *k* gets large. Therefore vote-counting estimators depend on having a large number of studies.

Another one relates to the issue of averaging identical sample sizes. If sample sizes of studies are not very different, Hedges and Olkin (1985) recommend an average value such as

$$n = \left(\frac{\sqrt{n_1} + \cdots + \sqrt{n_k}}{k} \right)^2.$$

The next limitation relates to the case when $U = 0$ or $U = k$. This means that the estimate of $p_C(\theta)$ turns out to be zero or unity. If $p_C(\theta_0) = 1$ for some θ_0, then $p_C(\theta) = 1$ for all $\theta \geq \theta_0$ and therefore it is impossible to define a unique θ.

6.5 Vote-counting Method for Unequal Sample Sizes.

Using the same notation as in the case of equal sample sizes, let $T_1,...,T_k$ be independent estimators of parameters $\theta_1,...,\theta_k$ obtained from experiments with sample sizes $n_1,...,n_k$. The critical values C_i may differ from study to study (Hedges and Olkin, 1985). The probability that $T_i > C_i$ is

$$p(\theta_i, n_i) \equiv \Pr\{T_i > C_i \mid \theta_i, n_i\}.$$

The idea of estimating a parameter θ_i in each study is based on the fact that the probability function is a function of θ_i, n_i for $\theta_1 = \cdots = \theta_k = \theta$.

Suppose $X_i = 0$ or 1. Then $\Pr\{X_i = 1 \mid \theta, n_i\} = \Pr\{T_i > C_i \mid \theta, n_i\} = p(\theta(n_i)$.

Maximum likelihood methodology can be used to estimate θ and the log likelihood function is

$$L(\theta \mid X_1,..., X_k) = X_1 \log p(\theta, n_1) + (1 - X_1)\log[1 - p(\theta, n_1)] + \cdots +$$
$$X_k \log p(\theta, n_k) + (1 - X_k)\log[1 - p(\theta, n_k)].$$

Since n_i and $X_i, i = 1,...,k$ are known, the likelihood function is a function of θ and can be maximized over θ to obtain an estimator $\hat{\theta}$. It is difficult to get the estimator in a closed form, but one method to get the estimator numerically is to obtain a grid of possible values for θ and then select $\hat{\theta}$ in the grid so that it yields the greatest value for the likelihood function.

To estimate an effect size for unequal sample sizes one may observe whether $\bar{Y}^E - \bar{Y}^C > 0$ for each study. Under condition of a homogeneous effect, i.e. $\delta_1 = \cdots = \delta_k = \delta$, the model turns out to be

$$\bar{Y}_i^E - \bar{Y}_i^C \sim N(\delta\sigma_i, \sigma_i^2 / \tilde{n}_i) \text{ where } \tilde{n}_i = n_i^E n_i^C / (n_i^E + n_i^C).$$

The probability function of positive result is to

$$p(\delta(\tilde{n}_i) = \Pr\{\bar{Y}_i^E - \bar{Y}_i^C > 0\} = 1 - \Phi(-\sqrt{\tilde{n}_i}\delta).$$

The likelihood is

$$L(\delta \mid X_1,...,X_k) = \sum_{i=1}^{k} \left\{ X_i \log[1 - \Phi(-\sqrt{\tilde{n}_i}\,\delta)] + (1 - X_i) \log \Phi(-\sqrt{\tilde{n}_i}\,\delta) \right\}$$

And it must be maximized numerically to obtain the maximum likelihood estimator $\hat{\delta}$.

The report introduces many of the basic techniques used in meta analysis. One should consider the references given for a more in depth study of meta analysis.

REFERENCES.

R. R. Bahadur. Rates of convergence of estimates and test statistics. *Annals of Mathematical Statistics,* 38, 303-324, 1965.

A. Birnbaum. Combining independent tests of significance. *Journal of American Statistical Association,* 49, (1954), 559-74.

R. Berk, A. Cohen. Asymptotically optimal methods of combining tests. *Journal of the American Statistical Association,* Vol. 74, N.368. (Dec.1979), 812-814.

G. Blom. Transformations of the binomial, negative binomial, poisson and χ^2 distributions, *Biometrica,* Vol. 41, (1954), 302-316.

J. Brozek and K. Tiede. Reliable and questionable significance in a series of statistical tests. *Psychological Bulletin,* Vol. 49, (1952), 339-341.

G. Casella, R. L. Berger. *Statistical inference.* 2nd.Edition. Duxbury. 2001.

J. Cohen. *Statistical power analysis for the behavioral sciences.* New-York : Academic press. 1969.

W. G. Cochran. Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society (Suppl.),* 4, (1937), 102-118.

T. D. Cook. The potential and limitations of secondary evaluations. In M.W.Apple, M.J.Subkoviak, U.S.Lufler (Eds.), *Educational Evaluation : Analysis and Responsibility.*

Berkeley, Cal.:McCutchan, 1974.

E. S. Edington. A normal curve method for combining probability values from independent experiments. *Journal of Psychology*, 82, (1972), 85-89.

R. A. Fisher. *Statistical methods for research workers* (4th ed.). London : Oliver and Boyd, 1932.

G. V. Glass. Primary, secondary and meta analysis of research, *Educational Researcher.* Vol.5. N.10. (1976), 3-8.

G. V. Glass. Integrating findings. *Review of Research of Education.* Vol.5. (1977), 351-379.

J. Good. On the weighted combination of significance tests. *Journal of the Royal Statistical Society, Series B,* 17, (1955), 264-265.

L. V. Hedges, I. Olkin. Vote-counting methods in research synthesis. *Psychological Bulletin*, Vol. 88, (1980), 359-369.

L. V. Hedges, I. Olkin. *Statistical Methods for Meta Analysis.* Academic Press, INC. 1985.

L. V. Hedges. Distribution theory of Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, Vol.6, (1981), 107-128.

L. V. Hedges. Fitting categorical models to effect size from a series of experiments. *Journal of Educational Statistics,* Vol.7, (1982), 119-137.

L. V. Hedges. A random effect model for effect sizes. *Psychological Bulletin,* Vol. 93, (1983), 388-395.

J. P. T. Higgins, A. Whitehead, R. M. Turner, R. Z. Omar, S.G. Thompson. Meta-analysis of continuous outcome data from individual patients. *Statistics in Medicine*, 20, (2001),

2219-2241.

L. W. Jones, D.W. Fiske. Models for testing the significance of combined results. *Psychological Bulletin,* Vol. 50, (1953), 375-382.

H. O. Lancaster. The combination of probabilities: An application of orthonormal functions. *Australian Journal of Statistics,* 3, (1961), 20-33.

T. Liptak. On the combination of independent tests. *Magyar Tudomanyos Akademia Matematikai Kutato Intezetenek Kozlemenyei,* 3, (1958), 1971-1977.

R. Littell, J. Folks. Asymptotic optimality of Fisher's methods of combining independent tests. *Journal of the American Statistician Association,* 66, (1971), 802-806.

W. Molenaar. *Approximations to the Poisson, binomial, and hypergeometric distribution functions.* Amsterdam: Mathematical Centre Tracts 31, 1970.

F. M. Mosteller and R. R. Bush. Selected quatitative techniques. In *Handbook of social psychology* (Ed. by G. Lindsey). Cambridge, MA : Addison-Wesley, 1954, p. 289-334.

F. M. Mosteller, J. W. Tukey. Data analysis including statistics. In G.Lindzey and E. Aronson (Eds.) *Handbook of social psychology*. (2nd ed) Reading Mass:Addinson-Wesley, 1968.

G. S. Mudholkar, E. O. George. The logit method for combining probabilities. In J. Rustagi (Ed.) *Symposium on optimizing methods in statistics* (pp. 345-366). New York : Academy Press, 1979.

K. Pearson. 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 probability drawn at random. *Biometrika,* 25, (1933), 379-410.

H. E. Robbins. The distribution of a definite quadratic form. *Annals of Mathematical Statistics,* 19, (1948), 266-270.

G. W. Snedecor, W. G. Cochran. *Statistical methods* (6th ed.) Ames: Iowa State University Press, 1967.

R. Rosental. Combining results of independent studies. *Psychological Bulletin,* Vol.85.

N 1.(1985), 185-193.

R. Rosenthal. Interpersonal expectations. In R. Rosenthal and R. L. Rosnow (Eds.), *Artifact in Behavioral research.* New York: Academic Press, 1969.

S. A. Stouffer, E. A. Suchman, L. C. DeVinney, S. A. Star and R. M. Williams. *The American soldier, Volume I, Adjustment during Army life.* Princeton, NJ: Princeton University Press.,1949.

L. H. Tippett. *The methods of statistics.* London: Williams & Norgate, 1931

W. Wallis. Compounding Probabilities from Independent Significance Tests. *Econometrica*, 10. (1942), 229-48.

A. Whitehead. *Meta-Analysis of Controlled Clinical Trials (Statistics in Practice).* Wiley, Chichester, 2002.

B. Wilkinson. A Statistical Consideration in Psychological Research, *Psychological Bulletin*, 48, (1951), 156-7.

B. J. Winer. *Statistical principles of experimental design* (2nd ed.). New York:MCGraw-Hill, 1971.