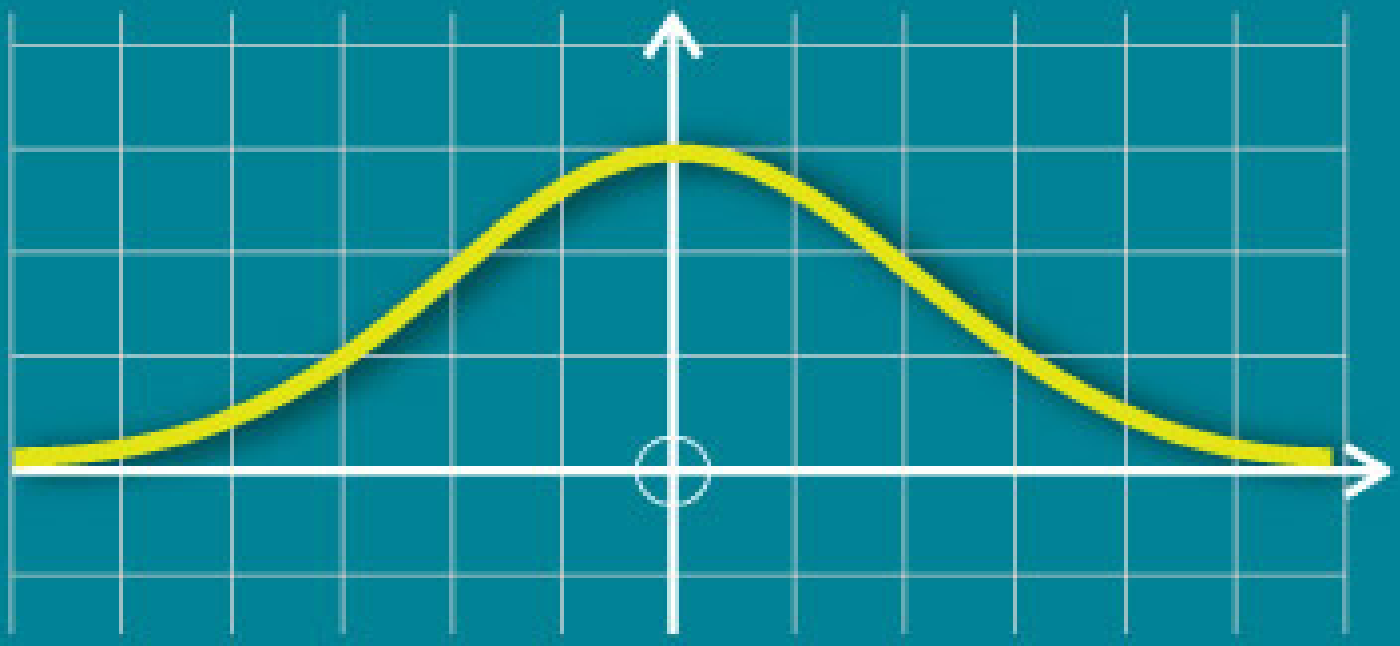
RALF SCHULZE · HEINZ HOLLING
DANKMAR BÖHNING (EDITORS)

META-ANALYSIS

NEW DEVELOPMENTS AND APPLICATIONS IN MEDICAL AND SOCIAL SCIENCES



Hogrefe & Huber

META-ANALYSIS

New Developments and Applications in Medical and Social Sciences

META-ANALYSIS

New Developments and Applications in Medical and Social Sciences

Edited by

Ralf Schulze, Heinz Holling, & Dankmar Böhning

Hogrefe & Huber

Library of Congress Cataloging-in-Publication Data

is now available via the Library of Congress Marc Database under the

LC Control Number: 2003104139

National Library of Canada Cataloguing-in-Publication

Meta-analysis: new developments and applications in medical and social sciences / Ralf Schulze, Heinz Holling, Dankmar Böhning (editors).

Includes bibliographical references and index.
ISBN 0-88937-266-7

1. Meta-analysis. 2. Medicine--Research--Evaluation. 3. Clinical trials--Statistical methods. 4. Social sciences--Research--Evaluation. 5. Social sciences--Statistical methods. I. Schulze, Ralf II. Holling, Heinz III. Böhning, Dankmar

R853.S7M48 2003 610'.7'27 C2003-902748-1

PUBLISHING OFFICES
USA: Hogrefe & Huber Publishers, 44 Brattle Street, 4th Floor, Cambridge, MA 02138
Phone (866) 823-4726, Fax (617) 354-6875, E-mail info@hhpub.com
Europe: Hogrefe & Huber Publishers, Rohnsweg 25, D-37085 Göttingen, Germany,
Phone +49 551 49609-0, Fax +49 551 49609-88, E-mail hh@hhpub.com

SALES & DISTRIBUTION
USA: Hogrefe & Huber Publishers, Customer Services Department, 30 Amberwood Parkway, Ashland, OH 44805,
Phone (800) 228-3749, Fax (419) 281-6883, E-mail custserv@hhpub.com
Europe: Hogrefe & Huber Publishers, Rohnsweg 25, D-37085 Göttingen, Germany,
Phone +49 551 49609-0, Fax +49 551 49609-88, E-mail hh@hhpub.com

OTHER OFFICES
Canada: Hogrefe & Huber Publishers, 12 Bruce Park Avenue, Toronto, Ontario M4P 2S3
Switzerland: Hogrefe & Huber Publishers, Länggass-Strasse 76, CH-3000 Bern 9

Hogrefe & Huber Publishers
Incorporated and registered in the State of Washington, USA, and in Göttingen, Lower Saxony, Germany

Printed and bound in Germany
ISBN 0-88937-266-7

Preface

Meta-analysis as a systematic method to integrate empirical findings has become a widely adopted technique in various scientific fields. Among the major areas of application of the method are medicine and the social sciences. New statistical developments and methodological advances often happen unrecognized in different substantive fields, or are assimilated with considerable delay. The present volume is intended to bring scholars from medical and social sciences together to present their theoretical advances as well as new applications of the method.

The book is divided in two parts. The first part consists of a collection of chapters that address various important theoretical issues. These chapters focus on the evaluation and systematization of existing procedures that are used in practice, present new developments regarding statistical procedures, describe techniques for the detection of bias in meta-analysis, and provide detailed expositions of the methodological viewpoints on meta-analysis in pharmaceutical, medical as well social science research.

In Chapter 1, Hartung, Argaç, and Makambi present a series of homogeneity tests that are known within the framework of ANOVA but have not been widely adopted in applications of meta-analysis. They expound the underlying logic of the tests and evaluate their performance in a simulation study. Hartung et al. address the problem of testing the homogeneity assumption that is often made in practical applications of meta-analysis, and they show which tests perform best under several conditions.

Schulze, Holling, Großmann, Jütting, and Brocke present a comparison of two meta-analytical approaches for the analysis of correlation coefficients in Chapter 2. It is shown that parallel statistical developments in different subdisciplines of psychology have lead to diverse procedural details in approaches often used in practice. These details can in turn lead to differences in results on the basis of the same database. This is demonstrated in a Monte-Carlo study of different homogeneous situations for which the procedures of the approaches – and fixed effects models in general – are supposed to be appropriate.

Random and fixed effects models in meta-analysis play an important role for the data analytic strategy and the interpretation of results. In recent years, the random effects model has been favored over the fixed effects model for theoretical reasons but only few procedures have been proposed for the estimation of the heterogeneity variance. This variance is an important component in the random effects model. Malzahn presents a general principle for its estimation in several meta-analytical models in Chapter 3.

The choice between the random and fixed effects model of meta-analysis has been subject of several debates. Although the random effects model was focused in theoretical discussions of the topic, in practical applications of meta-analysis, especially in the social sciences, the fixed effects model still prevails (see Chapter 2). Several authors have argued that the choice between these models has to be based on theoretical reasons and the inference that is intended with a meta-analysis. Hartung and Knapp present the basics of both the random and fixed effects model as well as commonly used methods in these models in Chapter 4. They also show that there are *theoretical* deficiencies in these models and propose an alternative test procedure which is presented in detail from an analytical point of view. Furthermore, the results of a simulation study that evaluates the performance of this new test procedure is reported.

The issue of bias in meta-analysis poses considerable problems to the interpretation of meta-analytical results. Often, the so-called publication bias is of particular interest. In Chapter 5, Schwarzer, Antes, and Schumacher review several procedures – graphical methods as well as test procedures – for the detection of bias in meta-analysis. They also present the results of a simulation study to evaluate the performance of two statistical tests for the identification of bias.

Apart from statistical issues in a narrower sense like those addressed in the first five chapters, more general methodological discussions have reoccurred in the literature since the advent of meta-analysis. Such methodological issues are addressed in the following four chapters. The different perspectives of medical research and the social sciences are reflected in these chapters and it is shown how analogous problems are dealt with in these areas of research.

In Chapter 6, Sauerbrei and Blettner review and compare different methods for summarizing empirical results from observational studies, including narrative reviews, meta-analysis of literature, meta-analysis of patient data, and prospective meta-analysis. Focusing on applications to medical research problems, the utility of meta-analysis for the evaluation of medical treatments is critically assessed. In addition to a theoretical analysis of the different review methods, several examples from the medical literature are presented. These examples support their arguments for a sceptical view on the utility of meta-analyses that are based on summary reports from the literature.

Koch and Röhmel concentrate in Chapter 7 on the use of meta-analysis in the process of new drug applications, where the method has not played a major role to date. They point out obstacles for the acceptance of meta-analytical results in this area. An analysis of the evaluation process for outcomes from randomized clinical trials on the comparison of different drugs for the same indication is presented, and references to relevant guidelines are given. Also, problems as well as benefits in using meta-analysis are illustrated by giving concrete examples. The characteristics that influence the credibility of meta-analyses in this field of application are highlighted as well. Thereby, Koch and Röhmel provide a constructive account for the enhancement of meta-analytical design.

In the subsequent chapter, Matt presents a comprehensive treatment on the possibilities to draw generalized causal inferences based on the results of meta-analysis. Here, like in other chapters in this volume, it is acknowledged that methods of meta-analysis are comparable to quasi-experiments or observational studies in methods of primary research. Drawing on principles developed in the context of generalization in quasi-experimentation, he demonstrates how these principles can be fruitfully applied to methods of meta-analysis. In his detailed exposition Matt also refers to general principles of generalization and provides examples of their successful application in practice. The presentation in Chapter 8 by Matt shows how questions of generalization are treated in the social sciences, and this view stands – at least partly – in contrast to treatments from the perspective of medical research (see e.g., Chapter 6 by Sauerbrei and Blettner).

The last chapter of the first part addresses the utility of tests of moderator hypotheses in meta-analysis. In Chapter 9 by Czienskowski, an example from social cognition research on the so-called self-reference effect is given to illustrate the application of moderator-analysis. Potential conclusions on the basis of the results are discussed, and it is shown how and why moderator analyses can and should be supplemented by follow-up experiments.

In the second part of the book applications of meta-analysis to different problems in medical, pharmaceutical and social science research are presented. A series of six chapters illustrates the breath of potential fields of application for meta-analytic methods.

An innovative field of application for meta-analysis is quality control in pharmaceutical production. In Chapter 10, Böhning and Dammann provide an overview and an example on how methods of meta-analysis can be applied in this new area of application. They extend an approach of mixture modeling of heterogeneity in meta-analysis and show its potential for an improvement of production processes in pharmaceutical industry.

In the following Chapter 11 by Greiner, Wegscheider, Böhning, and Dahms, an application of meta-analysis to explore and identify factors that influence the sensitivity and specificity of a medical test for the detection of trichinella antibodies is presented. They illustrate how adequate statistical methods of meta-analysis (e.g., mixed logistic regression) can contribute new knowledge that is of practical concern.

In Chapter 12, Dietz and Weist introduce a method based on finite mixed generalized linear models as a means for modeling heterogeneity in meta-analytic data. They present a detailed account of the model, methods for the estimation of parameters, and also give two examples of its application. The authors thereby demonstrate how advanced flexible methods of meta-analysis can provide useful results for the explanation of heterogeneity that go well beyond information gained from ordinary applications of meta-analysis.

Franklin also uses the generalized linear model in Chapter 13 to assess the impact of explanatory variables on the variability in a meta-analytical database. He examines, among other influential factors, the differences between treatment results in paediatric and adult clinical trials on Hodgkin's disease.

In a meta-analysis on the results of controlled clinical trials on antidepressants, Schöchlin, Klein, Abrahm-Rudolf, and Engel examine the potential moderating influence of design variables. They report results in Chapter 14 that stress the important role of design variables – especially the inclusion of placebo conditions – in this area of clinical applications.

One of the major research fields in social psychology, attitude research, is the subject of Chapter 15 by Schulze and Wittmann. The authors first provide an exposition of the two most often applied theories in this area. Additionally, moderator hypotheses concerning the relationships between the theory's components are substantiated that reflect standard assumptions of the theories as well as new hypotheses not previously tested in a meta-analytical framework. The results of a meta-analysis are also presented to assess overall effects as well as tests of pertinent moderator hypotheses in a random effects model.

Finally, Schlattmann, Malzahn, and Böhning present a new software package called META for the application of meta-analysis in Chapter 16. META enables the user to perform not only standard analysis to integrate research results but also includes procedures to apply the latest developments in mixture modeling of heterogeneity in meta-analysis as presented in this volume (see also Chapter 10).

The new developments and applications described in these chapters are contributions from different fields of research. Our hopes are that bringing together the contributions from these scholars in a single volume adds new knowledge to the different fields, counteracts fragmentation of statistical and substantial developments, and encourages potential users of the procedures to apply the latest methods of meta-analysis in their field of interest.

RALF SCHULZE

HEINZ HOLLING

DANKMAR BÖHNING

Contents

Part I

Theory

1

Homogeneity Tests in Meta-Analysis

Joachim Hartung
Doğan Argaç
Department of Statistics[†]
University of Dortmund

Kepher Makambi
Department of Mathematics and Statistics
Jomo Kenyatta University of Agriculture and Technology

Summary

For the homogeneity problem in meta-analysis, the performance of seven test statistics is compared under homogeneity and heterogeneity of the underlying population (study, group) variances. These are: the classical ANOVA F test, the Cochran test, the Welch test, the Brown-Forsythe test, the modified Brown-Forsythe test, the approximate ANOVA F test and as a proposal, an adjusted Welch test. At the whole, the Welch test proves to be the best one, but for small sample sizes and many groups, it becomes too liberal. In this case the adjusted Welch test is recommended to correct this anomaly. The other tests prove to have changing advantages dependent on the sizes of the parameters involved.

[†]Project "Meta-Analysis in Biometry and Epidemiology" (SFB 475) of the Deutsche Forschungsgemeinschaft (DFG).

1.1 INTRODUCTION

Meta-analysis of results from different experiments (groups, studies) is a common practice nowadays. In the framework of a one-way ANOVA model, serving generally as supporting edifice for meta-analysis, one may be interested in testing the homogeneity hypothesis. However, when the underlying population variances in different populations (studies, groups) are different, the ANOVA F-statistic attains significance levels which are very different from the nominal level (see for example, De Beuckelaer, 1996). In the rubric of the (generalized) Behrens-Fisher problem, a number of alternatives have been suggested.

Using simulation studies for various constellations of number of populations, sample sizes and within population error variances, we compare the actual attained sizes of the classical ANOVA F test, the Cochran test, the Welch test, the Brown-Forsythe test, the modified Brown-Forsythe test, the approximate ANOVA F test and, by adopting an idea of Böckenhoff and Hartung (1998), an adjusted Welch test, simultaneously.

1.2 MODEL AND TEST STATISTICS

Let y_{ij} be the observation on the jth subject of the ith population/study, $i = 1, \ldots, K$ and $j = 1, \ldots, n_i$

$$\begin{aligned} y_{ij} &= \mu_i + e_{ij} \\ &= \mu + a_i + e_{ij}\,; \quad i = 1, \ldots, K,\ j = 1, \ldots, n_i, \end{aligned}$$

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

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

That is, $e_{ij} \sim \mathcal{N}(0, \sigma_i^2);\ i = 1, \ldots, K,\ j = 1, \ldots, n_i$.

Interest is in testing the hypothesis $H_0 : \mu_1 = \cdots = \mu_K = \mu$. To test this hypothesis we will make use of the following test statistics:

a) The ANOVA *F* Test

S_{an}, given by

$$S_{an} = \frac{N-K}{K-1} \cdot \frac{\sum_{i=1}^{K} n_i(\bar{y}_{i.} - \bar{y}_{..})^2}{\sum_{i=1}^{K}(n_i - 1)s_i^2}, \tag{1.1}$$

with $N = \sum_{i=1}^{K} n_i$, $\bar{y}_{i.} = \sum_{j=1}^{K} y_{ij}/n_i$, $\bar{y}_{..} = \sum_{i=1}^{K} n_i \bar{y}_{i.}/N$.

This test was originally meant to test for equality of population means under variance homogeneity and has an F distribution with $K-1$ and

$N - K$ degrees of freedom.

Test: Reject $H_0 : \mu_1 = \cdots = \mu_K$ at level α if $S_{an} > F_{K-1,N-K;1-\alpha}$.

The ANOVA test has the weakness of not being robust with respect to heterogeneity in the intra-population error variances (Brown & Forsythe, 1974).

b) The Welch Test

$$S_{we} = \frac{\sum_{i=1}^{K} w_i(\bar{y}_{i.} - \sum_{j=1}^{K} h_j \bar{y}_{j.})^2}{\left((K-1) + 2 \cdot \frac{K-2}{K+1} \cdot \sum_{i=1}^{K} \frac{1}{n_i - 1}(1-h_i)^2\right)}, \tag{1.2}$$

where $w_i = n_i / s_i^2, \quad h_i = w_i / \sum_{k=1}^{K} w_k$, was an extension of testing the equality of two means to more than two means (see Welch, 1951) in the presence of variance heterogeneity within populations.

Under H_0, the statistic S_{we} has an approximate F distribution with $K-1$ and ν_g degrees of freedom, where

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

Test: Reject H_0 at level α if $S_{we} > F_{K-1,\nu_g;1-\alpha}$.

c) Cochran's Test

$$S_{ch} = \sum_{i=1}^{K} w_i(\bar{y}_{i.} - \sum_{j=1}^{K} h_j \bar{y}_{j.})^2, \tag{1.3}$$

was proposed by Cochran (1937) and then modified by Welch. We take it into our comparisons in order to get better comprehension and insight of the behavior of both statistics.

Under H_0, the Cochran statistic is distributed approximately as a χ^2-variable with $K-1$ degrees of freedom.

Test: Reject H_0 at level α if $S_{ch} > \chi^2_{K-1;1-\alpha}$.

d) Brown-Forsythe (B-F) Test

This one is also known as the modified F test and is given by

$$S_{b-f} = \frac{\sum_{i=1}^{K} n_i(\bar{y}_{i.} - \bar{y}_{..})^2}{\sum_{i=1}^{K}(1 - n_i/N)s_i^2}. \tag{1.4}$$

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

$$\nu = \frac{\left(\sum_{i=1}^{K}(1 - n_i/N)s_i^2\right)^2}{\sum_{i=1}^{K}(1 - n_i/N)^2 s_i^4/(n_i - 1)}. \tag{1.5}$$

Test: Reject H_0 at level α if $S_{b-f} > F_{K-1,\nu;1-\alpha}$.

Using a simulation study Brown and Forsythe (1974) demonstrated that their statistic is robust under inequality of variances. If the population variances are homogeneous, the B-F test is closer to ANOVA than Welch.

e) Mehrotra (Modified Brown-Forsythe) Test

$$S_{b-f(m)} = \frac{\sum_{i=1}^{K} n_i(\bar{y}_{i.} - \bar{y}_{..})^2}{\sum_{i=1}^{K}(1 - n_i/N)s_i^2}, \tag{1.6}$$

was proposed by (Mehrotra, 1997) in an attempt to correct a "flaw" in the B-F test.

Under H_0, $S_{b-f(m)}$ is distributed approximately as an F variable with ν_1 and ν degrees of freedom where

$$\nu_1 = \frac{\left(\sum_{i=1}^{K}(1 - n_i/N)s_i^2\right)^2}{\sum_{i=1}^{K} s_i^4 + \left(\sum_{i=1}^{K} n_i s_i^2/N\right)^2 - 2 \cdot \sum_{i=1}^{K} n_i s_i^4/N} \tag{1.7}$$

and ν is given in Equation 1.5 above.

Test: Reject H_0 at level α if $S_{b-f(m)} > F_{\nu_1,\nu;1-\alpha}$.

The flaw mentioned above is in the estimation of the numerator degrees of freedom by $K-1$ instead of ν_1.

f) The Approximate ANOVA *F* Test

$$S_{aF} = \frac{N-K}{K-1} \cdot \frac{\sum_{i=1}^{K} n_i(\bar{y}_{i.} - \bar{y}_{..})^2}{\sum_{i=1}^{K}(n_i-1)s_i^2}, \tag{1.8}$$

by Asiribo and Gurland (1990). This test gives an approximate solution to the problem of testing equality of means of normal populations in case of heteroscedasticity by making use of the classical ANOVA test.

Under H_0, the statistic S_{aF} is distributed approximately as an F-variable with ν_1 and ν_2 degrees of freedom where ν_1 is as given in Equation 1.7 above and

$$\nu_2 = \frac{\left(\sum_{i=1}^{K}(n_i-1)s_i^2\right)^2}{\sum_{i=1}^{K}(n_i-1)s_i^4}. \tag{1.9}$$

Test: Reject H_0 at level α if $S_{aF} > \hat{c} \cdot F_{\nu_1,\nu_2;1-\alpha}$, where

$$\hat{c} = \frac{N-K}{N(K-1)} \frac{\sum_{i=1}^{K}(N-n_i)s_i^2}{\sum_{i=1}^{K}(n_i-1)s_i^2}. \tag{1.10}$$

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

$$S_{aF} = S_{b-f(m)} = \frac{nK}{K-1} \cdot \frac{\sum_{i=1}^{K}(\bar{y}_{i.} - \bar{y}_{..})^2}{\sum_{i=1}^{K} s_i^2},$$

and

$$\nu = \nu_2 = (n-1) \cdot \frac{\left(\sum_{i=1}^{K} s_i^2\right)^2}{\sum_{i=1}^{K} s_i^4}.$$

g) The Adjusted Welch Test

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

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

where $c_i = (n_i - 1)/(n_i - 3)$, see Patel, Kapadia, and Owen (1976, pages 39-40). Therefore, an unbiased estimator of n_i/σ_i^2 is $n_i/c_i s_i^2$.

Now, let $\varphi_i = (n_i + \delta_1)/(n_i + \delta_2)$, where δ_1 and δ_2 are arbitrary real numbers; and then define the general weights by $w_i^* = n_i/\varphi_i s_i^2$. That is, for the Welch test, $w_i = w_i^*$ with $\varphi_i = 1$ ($\delta_1 = 0$, and $\delta_2 = 0$) and if we take the unbiased weights, $w_i = n_i/c_i s_i^2$, then $\varphi_i = c_i$, ($\delta_1 = -1$ and $\delta_2 = -3$).

For small samples in the groups, the Welch test becomes too liberal especially with increasing number of groups. Also, in our experience, using the unbiased weights in the Welch test makes the test too conservative. A reasonable compromise in this situation is to choose φ_i such that $1 \leq \varphi_i \leq c_i$.

This defines a new class of Welch type test statistics whose properties can be adjusted accordingly by choosing the control parameter, φ_i, appropriately. Our proposed test, which we shall henceforth call the *adjusted Welch test*, uses the weights $w_i^* = n_i/\varphi_i s_i^2$ in the Welch test, where $1 \leq \varphi_i \leq c_i$. That is the adjusted Welch test, S_{aw}, is given by:

$$S_{aw} = \frac{\sum_{i=1}^{K} w_i^* (\bar{y}_{i.} - \sum_{j=1}^{K} h_j^* \bar{y}_{j.})^2}{\left((K-1) + 2 \cdot \frac{K-2}{K+1} \cdot \sum_{i=1}^{K} \frac{1}{n_i - 1}(1 - h_i^*)^2\right)}, \tag{1.11}$$

where $h_i^* = w_i^* / \sum_{i=1}^{K} w_i^*$, $i = 1, \ldots, K$.
Under H_0, the adjusted Welch statistic, S_{aw}, is distributed approximately as an F-variable with $K - 1$ and ν_g^* degrees of freedom, with

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

Test: Reject H_0 at α level if $S_{aw} > F_{K-1, \nu_g^*; 1-\alpha}$.

When the sample sizes are large, S_{aw} approaches the Welch test, that is, $(n_i + \delta_1)/(n_i + \delta_2) \overset{n_i \to \infty}{\longrightarrow} 1$. With small sample sizes, our statistic will help correct the liberality witnessed in the Welch test.

To assess the relative performance of these test statistics in terms of the actual levels of significance attained, we will consider levels between 4% and 6% to be satisfactory, that is, following Cochran's rule of thumb (cf. Cochran, 1954).

1.3 SIMULATION STUDY AND DISCUSSION

In order to see the effect of balancedness and unbalancedness, as well as variance homogeneity and heterogeneity, a simulation study was conducted with sampling experiments determined by the number of studies, sample sizes and the variances in each study. In the first sampling experiment the following patterns and combinations of the number of studies, sample sizes and variances were considered (cf. Tables 1.1, 1.2, 1.3, and 1.4): Balanced samples and homogeneous variances, unbalanced samples combined with homogeneous variances. The next experiment investigated the effect of variance heterogeneity on the empirical Type I error rates. We matched balanced and unbalanced sample sizes with heterogeneous variances. In the unbalanced sample size cases, large sample sizes were separately paired with small and large variances. To investigate the effect of a large number of studies, we started with $K = 3$ studies and made independent replications to give $K = 6, 2 \times (.)$, $K = 9$, $3 \times (.)$, and $K = 18, 6 \times (.)$. We will use the term small sample to refer to $n_i = 5$, and moderate for $n_i = 10, 15, i = 1, \ldots, K$. However, if any of the sample sizes, n_i, is greater or equal to 20, then the constellation will be taken to be of large samples.

Table 1.1 reports the actual significance levels for $K = 3$, Table 1.2 for $K = 6$, Table 1.3 for $K = 9$ and Table 1.4 for $K = 18$. For the adjusted Welch test, S_{aw}, we have taken $\varphi_i = (n_i + 2)/(n_i + 1), i = 1, \ldots, K$. From these Tables, we make the following observations in order of the various tests presented in Section 1.2 above:

a) The ANOVA *F* Test

In the case when the number of populations, $K = 3$:

i. for balanced samples sizes and homoscedastic cases, the test, as expected, keeps the nominal level;

ii. for balanced and heterogeneous variance cases, the test keeps control of the significance level. This trend is maintained with increasing sample sizes;

iii. for unbalanced and homoscedastic cases, the test keeps the nominal level;

iv. for the unbalanced and heterogeneous cases, if small samples are matched with small variances, the test tends to be conservative. However, when small sample sizes are paired with large variances, the test becomes liberal. This pattern remains largely unchanged even if the sample sizes are increased.

For $K = 6$, $K = 9$ and $K = 18$, the observations made in i. to iv. above still hold; except for balanced designs and heterogeneous variances where the test becomes more liberal with increasing number of populations.

Table 1.1 Actual Simulated Significance Levels (Nominal Level 5%) for $K = 3$

Sample Sizes	Variances	$\hat{\alpha}\%$						
(n_1, n_2, n_3)	$(\sigma_1^2, \sigma_2^2, \sigma_3^2)$	S_{an}	S_{we}	S_{ch}	S_{b-f}	$S_{b-f(m)}$	S_{aF}	S_{aw}
(5,5,5)	(4,4,4)	5.0	4.8	12.2	4.1	3.8	3.8	3.3
	(1,3,5)	6.0	5.0	13.5	4.6	4.2	4.2	3.6
(10,10,10)	(4,4,4)	5.1	4.9	8.4	4.9	4.6	4.6	3.9
	(1,3,5)	5.7	4.7	8.2	5.1	4.5	4.5	3.9
(20,20,20)	(4,4,4)	5.1	4.9	6.5	5.0	4.9	4.9	4.2
	(1,3,5)	5.6	4.8	6.4	5.4	4.7	4.7	4.2
(40,40,40)	(4,4,4)	4.9	4.9	5.6	4.8	4.8	4.8	4.5
	(1,3,5)	5.9	5.2	5.8	5.8	5.0	5.0	4.8
(5,10,15)	(4,4,4)	5.0	5.3	10.2	5.1	4.8	5.4	4.2
	(1,3,5)	2.4	4.9	8.9	5.6	4.7	4.5	3.8
	(5,3,1)	12.3	5.4	11.5	5.3	5.0	6.2	4.4
(10,20,30)	(4,4,4)	5.2	5.3	7.7	5.1	4.9	5.3	4.5
	(1,3,5)	2.2	4.9	6.5	5.5	4.6	4.5	4.2
	(5,3,1)	12.9	5.5	8.1	5.6	5.2	5.9	4.5
(20,40,60)	(4,4,4)	4.8	4.9	5.9	4.9	4.7	4.9	4.4
	(1,3,5)	2.1	5.1	5.8	5.7	4.7	4.6	4.5
	(5,3,1)	12.5	4.9	6.4	5.5	5.0	5.4	4.4

Note. For a definition of S_{an}, S_{we}, S_{ch}, S_{b-f}, $S_{b-f(m)}$, S_{aF}, and S_{aw} see Equations 1.1, 1.2, 1.3, 1.4, 1.6, 1.8, and 1.11.

Table 1.2 Actual Simulated Significance Levels (Nominal Level 5%) for $K = 6$

Sample Sizes	Variances	$\hat{\alpha}\%$						
$2\times$ (n_1, n_2, n_3)	$2\times$ $(\sigma_1^2, \sigma_2^2, \sigma_3^2)$	S_{an}	S_{we}	S_{ch}	S_{b-f}	$S_{b-f(m)}$	S_{aF}	S_{aw}
(5,5,5)	(4,4,4)	5.2	6.2	22.1	4.1	3.3	3.3	4.1
	(1,3,5)	6.6	6.1	22.4	4.8	3.7	3.7	4.3
(10,10,10)	(4,4,4)	5.1	5.1	11.4	4.8	4.2	4.2	3.7
	(1,3,5)	6.3	5.2	12.0	5.6	4.3	4.3	3.7
(20,20,20)	(4,4,4)	4.8	4.7	7.7	4.7	4.3	4.3	3.8
	(1,3,5)	6.0	4.8	7.7	5.7	4.4	4.4	4.0
(40,40,40)	(4,4,4)	4.7	4.6	6.0	4.6	4.4	4.4	4.2
	(1,3,5)	6.8	5.4	6.9	6.6	5.0	5.0	4.9
(5,10,15)	(4,4,4)	5.0	6.3	15.5	4.7	4.0	4.5	4.7
	(1,3,5)	2.4	5.5	13.1	5.9	4.3	4.2	3.8
	(5,3,1)	16.3	6.7	16.7	5.7	4.6	5.5	5.0
(10,20,30)	(4,4,4)	5.5	5.7	9.7	5.2	4.7	4.9	4.8
	(1,3,5)	2.3	5.2	8.3	6.5	4.8	4.7	4.2
	(5,3,1)	16.3	5.7	10.2	6.3	4.8	5.5	4.7
(20,40,60)	(4,4,4)	5.2	5.3	7.2	5.2	4.8	5.0	4.6
	(1,3,5)	2.6	5.5	7.1	6.7	5.1	5.0	4.7
	(5,3,1)	15.3	4.8	6.7	6.3	4.9	5.2	4.1

Note. For a definition of S_{an}, S_{we}, S_{ch}, S_{b-f}, $S_{b-f(m)}$, S_{aF}, and S_{aw} see Equations 1.1, 1.2, 1.3, 1.4, 1.6, 1.8, and 1.11.