

Statistical Methods for Meta-Analysis

LARRY V. HEDGES
INGRAM OLKIN

Statistical Methods for Meta-Analysis

Larry V. Hedges

Department of Education
University of Chicago
Chicago, Illinois

Ingram Olkin

Department of Statistics
and School of Education
Stanford University
Stanford, California

1985



ACADEMIC PRESS, INC.

(Harcourt Brace Jovanovich, Publishers)

Orlando San Diego New York London
Toronto Montreal Sydney Tokyo

COPYRIGHT © 1985, BY ACADEMIC PRESS, INC.
ALL RIGHTS RESERVED.
NO PART OF THIS PUBLICATION MAY BE REPRODUCED OR
TRANSMITTED IN ANY FORM OR BY ANY MEANS, ELECTRONIC
OR MECHANICAL, INCLUDING PHOTOCOPY, RECORDING, OR
ANY INFORMATION STORAGE AND RETRIEVAL SYSTEM, WITHOUT
PERMISSION IN WRITING FROM THE PUBLISHER.

ACADEMIC PRESS, INC.
Orlando, Florida 32887

United Kingdom Edition published by
ACADEMIC PRESS INC. (LONDON) LTD.
24-28 Oval Road, London NW1 7DX

LIBRARY OF CONGRESS CATALOGING IN PUBLICATION DATA

Hedges, Larry V.
Statistical methods for meta-analysis.

Includes index.
I. Social sciences—Statistical methods. I. Olkin,
Ingram. II. Title. III. Title: Meta-analysis.
HA29.H425 1985 300'.72 84-12469
ISBN 0-12-336380-2 (alk. paper)

PRINTED IN THE UNITED STATES OF AMERICA

Preface

Methodology for combining findings from repeated research studies has a long history. Early examples of combining evidence are found in replicated astronomical and physical measurements. Agricultural experiments particularly lend themselves to replication and led to the development of statistical techniques for merging results.

In recent years a plethora of meta-analyses have emerged in social science research. The need to arrive at policy decisions affecting social institutions fostered the momentum toward summarizing research. But, as with most methodologies, abuse frequently accompanies use. Two central aspects of meta-analysis were quickly recognized. One involved methods for collecting the body of information to be summarized. This pinpointed a variety of problems, pitfalls, and questions. For example, what steps should be taken to guarantee objectivity? Should some studies be omitted because of inadequacies in design or execution?

The second aspect of meta-analysis assumes as a starting point that we have available a set of reasonably well-designed studies that address the same question using similar outcome measures and focuses on the methodology needed for summarizing the data. Because classical statistics primarily addresses the analysis of single experiments, new formulations, models, and methods are required.

The main purpose of this book is to address the statistical issues for integrating independent studies. There exist a number of papers and books that discuss the

mechanics of collecting, coding, and preparing data for a meta-analysis, and we do not deal with these.

It is not unusual in the early development of a field for terms to be used that later may be less than adequate or more restrictive than need be. In particular, the term *effect size* has been used to refer to standardized mean differences. In the beginning this usage was very natural in that the particular studies of interest did indeed involve differences between means. However, with more elaborate experimentation and more diverse applications, differences between treatments may depend not only on means but also on variances, medians, correlations, order statistics, distances, etc. Thus it would behoove us to now use the term *effect size* to refer to any such indices. But because this would be contrary to much of the existing literature, we do not, somewhat regrettably, do so. Instead we introduce the term *effect magnitude* to refer to measures in general.

The problem is further compounded. For large samples, quantities such as variances, medians, correlations, etc., will frequently have a normal distribution in which some of the population parameters will indeed be means. Consequently, although we may begin with statistics that are not means, we often end up with statistics that are effect sizes in the original sense.

Because this book concerns methodology, the content necessarily is statistical, and at times mathematical. In order to make the material accessible to a wider audience, we have not provided proofs in the text. Where proofs are given, they are placed as commentary at the end of a chapter. These can be omitted at the discretion of the reader.

We make a number of technical statements such as “The statistic Q has a chi-square distribution with degrees of freedom,” or “The statistics Q_1 and Q_2 are independent.” Each of these statements warrants a proof or a reference. However, for the sake of simplicity, readability, and accessibility to the materials, we have taken the liberty of omitting many proofs and references. However, as a compromise, we include a few proofs and references for statements that might be considered typical.

At times, mathematical expressions are needed. When possible we provide tables and graphs to make these expressions simple to use.

In our writing we have in mind a prototypical reader who is familiar with basic statistics at an applied level. This normally means the completion of a one-year sequence in statistics at a noncalculus level, and includes the ideas of statistical inference, regression and correlation, and analysis of variance. Concepts such as distribution (cumulative distribution function), expected value, bias, variance, and mean-squared error are generally defined in standard introductory statistics textbooks and are not defined in this book.

Occasionally, we use more advanced statistical concepts, such as consistency, efficiency, invariance, or asymptotic distributions. Because these concepts are used infrequently, we do not give formal definitions and on occasion give only a

brief explanation or no explanation at all. Elementary expositions of these concepts can be found in more advanced statistics books. Our main reason for not discussing these concepts is that they are not essential for an understanding of the principles. A lengthy explanation would destroy readability. On the other hand, these comments can be useful for those readers familiar with the concepts.

Throughout the book we describe computational procedures whenever required. Many computations can be completed on a hand calculator, whereas some require the use of a standard statistical package such as SAS, SPSS, or BMD. Readers with experience using a statistical package or who conduct analyses such as *multiple regression* or *analysis of variance* should be able to carry out the analyses described with the aid of a statistical package.

Because of the inclusion of so many tables, a commentary on interpolation may be in order. For any two-way table (see figure), the simplest method of interpolation is linear in each direction; and in general, linear interpolation will suffice for most practical purposes. For more accurate interpolation, the values can be plotted horizontally or vertically. Thus, for example, vertical plots give interpolated values between c_1 and d_1 and between c_2 and d_2 , denoted by crosses. Similarly, horizontal plots given interpolated values between c_1 and c_2 and between d_1 and d_2 , denoted by dots. Subsequent linear interpolation in the other direction will provide a more accurate result than two-way linear interpolation. Of course, plotting a complete row or column of interpolated values gives still more accurate results, but this may be more effort than is warranted in practice.

A comment is in order concerning the calculations presented in the examples. Individual terms are presented after rounding, whereas totals are computed without roundoff. Consequently, discrepancies in the last decimal may exist in some of the computations.

There is an inherent difficulty in trying to use a single letter to denote a particular characteristic. For example, if we denote a population mean by μ , then how shall we denote the mean of the means of an experiment and control group? If we label the two means as μ^E and μ^C and $\mu = \frac{1}{2}(\mu^E + \mu^C)$, then we have denoted by μ two different types of means. Such inconsistencies are inherent in the subject and occur throughout the book, so it is important to make quite explicit the underlying context and thereby remove potential confusion. Because

	c_1	•	c_2
	X		X
	d_1	•	d_2

We write $a \approx b$ to mean that a and b are “approximately” equal. We use the symbol $a \equiv b$ to signify a definition.

Independent multiple opinions and replications of experiments are but two examples of corroborative evidence which are currently in vogue, and which we believe will increase in frequency during the next decade. We hope that the present development of methodology will provide some guidelines for the rigorous interpretation of data from independent sources.

Acknowledgments

It is a pleasure to express our appreciation to Dr. Betsy Becker (Michigan State University), Mr. Brad Hanson (Stanford University), Dr. Nan Laird (Harvard University), Ms. Therese Pigott (University of Chicago), Dr. Marie Tisak (University of California, Berkeley), and Dr. Margaret Uguroglu (DeVry, Inc.) for reading the manuscript and providing us with numerous comments and suggestions. The authors are particularly grateful to Dr. Betsy Becker for carrying out the simulation studies and for the computations in many of the tables. Therese Pigott (University of Chicago) checked the computations in all of our examples. Although it would be comforting to be able to blame these readers for errors, honesty compels us to say that any shortcomings are solely ours.

Unfortunately, the first draft of this book was not typed onto a word processor, thereby necessitating many retypings. For this task we thank Julie Less, Irene Miura, Jerri Rudnick and Tonda West who managed to maintain a calm exterior at all times.

We are grateful to the Spencer Foundation and to the National Science Foundation for their continued support of our research. Their generous support has been essential in making this work possible.

The authors wish to acknowledge, with thanks, the following associations and societies for kindly granting us permission to use published materials:

The American Psychological Association for Tables 1, 2, 3, and Fig. 1 of Chapter 4; Tables 4, 5, 6 and Figure 3 of Chapter 5; Tables 3, 4, and 7 of Chapter

6; Tables 1, 2, 3, 4, 5, and 6 of Chapter 13; and Table 4 of Chapter 14.

The American Educational Research Association for Table 2 of Chapter 5, Tables 1, 2 and 3 of Chapter 8, and Table 1, 2 and 3 of Chapter 14.

The Institute of Mathematical Statistics for Table 1 of Chapter 11.

The Biometrika Trust for Appendix E and Appendix G.

Contents

Preface

xv

1. Introduction

A. The Use of Statistical Procedures for Combining the Results of Research Studies in the Social Sciences	2
A.1 The Misuse of Statistical Significance in Reviewing Research	3
A.2 Statistical Procedures for Combining Estimates of Effect Magnitude	6
B. Failings in Conventional Statistical Methodology in Research Synthesis	7
B.1 Goals of Statistical Procedures in Research Synthesis	7
B.2 Conventional Analyses for Effect Size Data	9
B.3 Conceptual Problems with Conventional Analyses	10
B.4 Statistical Problems with Conventional Analyses	11
C. Statistics for Research Synthesis	12

2. Data Sets

A. Cognitive Gender Differences	16
B. Sex Differences in Conformity	21
C. The Effects of Open Education	23

D. The Relationship between Teacher Indirectness and Student Achievement	26
3. Tests of Statistical Significance of Combined Results	
A. Preliminaries and Notation	28
B. General Results on Tests of Significance of Combined Results	31
C. Combined Test Procedures	33
C.1 Methods Based on the Uniform Distribution	34
C.2 The Inverse Chi-Square Method	37
C.3 The Inverse Normal Method	39
C.4 The Logit Method	40
C.5 Other Procedures for Combining Tests	42
C.6 Summary of Combined Test Procedures	43
D. The Uses and Interpretation of Combined Test Procedures in Research Synthesis	45
E. Technical Commentary	46
4. Vote-Counting Methods	
A. The Inadequacy of Conventional Vote-Counting Methodology	48
B. Counting Estimators of Continuous Parameters	52
C. Confidence Intervals for Parameters Based on Vote Counts	53
C.1 Use of Normal Theory Approach	54
C.2 Use of Chi-Square Theory	54
D. Choosing a Critical Value	56
E. Estimating an Effect Size	56
F. Estimating a Correlation	63
G. Limitations of Vote-Counting Estimators	67
H. Vote-Counting Methods for Unequal Sample Sizes	69
H.1 The Large Sample Variance of the Maximum Likelihood Estimator	70
H.2 Estimating Effect Size	71
5. Estimation of a Single Effect Size: Parametric and Nonparametric Methods	
A. Estimation of Effect Size from a Single Experiment	76
A.1 Interpreting Effect Sizes	76
A.2 An Estimator of Effect Size Based on the Standard Mean Difference	78

C.2 Estimators of Effect Size Based on Transformed Estimates	119
D. Testing for Homogeneity of Effect Sizes	122
D.1 Small Sample Significance Levels for the Homogeneity Test Statistic	124
D.2 Other Procedures for Testing Homogeneity of Effect Sizes	124
E. Computation of Homogeneity Test Statistics	127
F. Estimation of Effect Size for Small Sample Sizes	128
F.1 Estimation of Effect Size from a Linear Combination of Estimates	129
F.2 Modified Maximum Likelihood Estimation of Effect Size	131
G. The Effects of Measurement Error and Invalidity	131
G.1 The Effects of Measurement Error	132
G.2 The Effect of Validity of Response Measures	138
 7. Fitting Parametric Fixed Effect Models to Effect Sizes: Categorical Models	
A. An Analogue to the Analysis of Variance for Effect Sizes	149
B. Model and Notation	149
C. Some Tests of Homogeneity	153
C.1 Testing Whether All Studies Share a Common Effect Size	153
C.2 Testing Homogeneity of Effect Sizes Across Classes	154
C.3 Testing Homogeneity within Classes	155
C.4 An Analogy to the Analysis of Variance	156
C.5 Small Sample Accuracy of the Asymptotic Distributions of the Test Statistics	156
D. Fitting Effect Size Models to a Series of Studies	157
E. Comparisons among Classes	159
E.1 Simultaneous Tests for Many Comparisons	160
F. Computational Formulas for Weighted Means and Homogeneity Statistics	163
 8. Fitting Parametric Fixed Effect Models to Effect Sizes: General Linear Models	
A. Model and Notation	168
B. A Weighted Least Squares Estimator of Regression Coefficients	169

A.3 An Unbiased Estimator of Effect Size	81
A.4 The Maximum Likelihood Estimator of Effect Size	81
A.5 Shrunk Estimators of Effect Size	82
A.6 Comparing Parametric Estimators of Effect Size	82
B. Distribution Theory and Confidence Intervals for Effect Sizes	85
B.1 The Asymptotic Distribution of Estimators of Effect Size	85
B.2 Confidence Intervals for Effect Sizes Based on Transformations	88
B.3 Exact Confidence Intervals for Effect Sizes	91
C. Robust and Nonparametric Estimation of Effect Size	92
C.1 Estimates of Effect Size that are Robust Against Outliers	93
C.2 Nonparametric Estimators of Effect Size	93
C.3 Estimators Based on Differences of Control versus Treatment Proportions	95
C.4 Estimators Based on Gain Scores in the Experimental Group Relative to the Control Group	97
C.5 Nonparametric Estimators Involving Only Posttest Scores	98
C.6 Relationships between Estimators	99
D. Other Measures of Effect Magnitude	100
D.1 The Correlation Coefficient and Correlation Ratio	101
D.2 The Intraclass Correlation Coefficient	102
D.3 The Omega-Squared Index	103
D.4 Problems with Variance-Accounted-For Measures	103
E. Technical Commentary	104
 6. Parametric Estimation of Effect Size From a Series of Experiments	
A. Model and Notation	108
B. Weighted Linear Combinations of Estimates	109
B.1 Estimating Weights	110
B.2 Efficiency of Weighted Estimators	113
B.3 The Accuracy of the Large Sample Approximation to the Distribution of Weighted Estimators of Effect Size	114
C. Other Methods of Estimation of Effect Size from a Series of Experiments	117
C.1 The Maximum Likelihood Estimator of Effect Size from a Series of Experiments	118

C. Testing Model Specification	172
D. Computation of Estimates and Test Statistics	173
E. The Accuracy of Large Sample Approximations	174
F. Other Methods of Estimating Regression Coefficients	183
F.1 Maximum Likelihood Estimators of Regression Coefficients	183
F.2 Estimators Based on Transformations of Sample Effect Sizes	184
G. Technical Commentary	187
9. Random Effects Models for Effect Sizes	
A. Model and Notation	191
B. The Variance of Estimates of Effect Size	193
C. Estimating the Effect Size Variance Component	193
D. The Variance of the Effect Size Variance Component	195
E. Testing That the Effect Size Variance Component Is Zero	197
F. Estimating the Mean Effect Size	198
G. Empirical Bayes Estimation for Random Effects Models	200
10. Multivariate Models for Effect Sizes	
A. Model and Notation	206
B. The Multivariate Distribution of Effect Sizes	209
C. Estimating a Common Effect Size from a Vector of Correlated Estimates	210
C.1 Testing Homogeneity of Correlated Effect Sizes	210
C.2 Estimation of Effect Size From Correlated Estimates	212
D. Estimating a Common Effect Size and Testing for Homogeneity of Effect Sizes	213
D.1 An Estimator of Effect Size	213
D.2 Testing Homogeneity of Effect Sizes	215
E. Estimating a Vector of Effect Sizes	216
F. Testing Homogeneity of Vectors of Effect Sizes	219
G. Is Pooling of Correlated Estimators Necessary?	221
11. Combining Estimates of Correlation Coefficients	
A. Estimating a Correlation from a Single Study	224
A.1 Point Estimation of a Correlation from a Single Study	224
A.2 Approximations to the Distribution of the Sample Correlation Coefficient	226

A.3 Exact Confidence Intervals for Correlations	228
B. Effects of Measurement Error	228
C. Estimating a Common Correlation from Several Studies	229
C.1 Weighted Estimators of a Common Correlation	230
C.2 The Maximum Likelihood Estimator of a Common Correlation	232
D. Testing Homogeneity of Correlations across Studies	234
D.1 A Test of Homogeneity Based on Fisher's z-Transform	235
D.2 The Likelihood Ratio Test of Homogeneity of Correlations	236
E. Fitting General Linear Models to Correlations	237
E.1 Model and Notation	237
E.2 Estimating Regression Coefficients	238
E.3 Testing Model Specification	240
E.4 Computation of Estimates and Test Statistics	241
F. Random Effects Models for Correlations	242
F.1 Model and Notation	242
F.2 Testing That the Variance of Population Correlations is Zero	243
F.3 An Unbiased Estimate of the Correlation Variance Component	244
 12. Diagnostic Procedures for Research Synthesis Models	
A. How Many Observations Should Be Set Aside	249
B. Diagnostic Procedures for Homogeneous Effect Size Models	251
B.1 Graphic Method for Identifying Outliers	251
B.2 The Use of Residuals to Locate Outliers	253
B.3 The Use of Homogeneity Statistics to Locate Outliers	256
C. Diagnostic Procedures for Categorical Models	257
D. Diagnostic Procedures for General Linear Models	257
D.1 The Use of Residuals to Locate Outliers in General Linear Models	258
D.2 Changes in Regression Coefficients	260
D.3 The Use of Model Specification Statistics to Locate Outliers	261
E. Diagnostic Procedures for Combining Estimates of Correlation Coefficients	262
F. Technical Commentary	262

13. Clustering Estimates of Effect Magnitude

A. Theory for Clustering Unit Normal Random Variables	266
A.1 Disjoint Clustering	267
A.2 Overlapping Clustering	269
B. Clustering Correlation Coefficients	271
C. Clustering Effect Sizes	273
D. The Effect of Unequal Sample Sizes	276
D.1 The Effect on the Disjoint Clustering Procedure	277
D.2 The Effect on the Overlapping Clustering Procedure	278
D.3 An Alternative Method for Handling Unequal Sample Sizes	279
E. Relative Merits of the Clustering Procedures	281
F. Computation of Tables	282
F.1 The Significance of Gaps	282
F.2 Significance Values for the Overlapping Clustering Procedure	283

14. Estimation of Effect Size When Not All Study Outcomes Are Observed

A. The Existence of Sampling Bias in Observed Effect Size Estimates	286
B. Consequences of Observing Only Significant Effect Sizes	287
B.1 Model and Notation	288
B.2 The Distribution of the Observed Effect Size	288
C. Estimation of Effect Size from a Single Study When Only Significant Results Are Observed	290
C.1 Estimation of Effect Size	291
C.2 The Distribution of the Maximum Likelihood Estimator	292
D. Estimation of Effect Size from a Series of Independent Experiments When Only Significant Results are Observed	297
D.1 Estimation of Effect Size Using Counting Procedures	298
D.2 The Maximum Likelihood Estimator of Effect Size	301
D.3 Weighted Estimators of Effect Size	302
D.4 Applications of the Methods That Assume Censoring	303
E. Other Methods for Estimating Effect Sizes When Not All Study Outcomes Are Observed	304
E.1 Assessing the Number of Studies (with Null Results) Needed to Overturn a Conclusion	305
E.2 Random Effects Models When Not All Study Outcomes Are Observed	306

F. Technical Commentary	307
15. Meta-Analysis in the Physical and Biological Sciences	
A. A Multiplicative Model	314
B. Estimating Displacement and Potency Factor	318
C. A Multiplicative Model with Scaling	321
D. An Additive Model for Interlaboratory Differences	322
E. An Additive Model With a Control	323
Appendix	
A. Table of the Standard Normal Cumulative Distribution Function	328
B. Percentiles of Chi-Square Distributions	330
C. Values of $z = \frac{1}{2} \log[(1+r)/(1-r)]$ and $r = (e^{2z} - 1)/(e^{2z} + 1)$ and Graph of Fisher's z -Transformation	333
D. Values of $\sqrt{2} \sinh^{-1} x$ and $\sinh(x/\sqrt{2})$	335
E. Confidence Intervals of the Parameter p of the Binomial Distribution	337
F. Nomographs for Exact Confidence Intervals for Effect Size δ When $2 \leq n \leq 10$	341
G. Confidence Intervals for the Correlation Coefficient for Different Sample Sizes	342
References	347
Index	361