ED 319 775 TM 014 988

AUTHOR	Harwell, Michael R.
TITLE	Summarizing Monte Carlo Results in Methodological Research.
PUB DATE	Apr 90
NOTE	30p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Data Collection; *Data Interpretation; Educational Research; Mathematical Models; *Meta Analysis; *Methods Research; *Monte Carlo Methods; Psychological Studies
IDENTIFIERS	Power (Statistics); Type I Errors

ABSTRACT

Monte Carlo studies of statistical tests are prominently featured in the methodological research literature. Unfortunately, the information from these studies does not appear to have significantly influenced methodological practice in educational and psychological research. One reason is that Monte Carlo studies lack an overarching theory to guide their interpretation. Also, the impressionistic nature of Monte Carlo studies can lead different readers to different conclusions. These shortcomings can be addressed using meta-analytic methods to summarize the results of Monte Carlo studies. Summarizing Monte Carlo studies in this fashion will generate guidelines for the appropriate use of particular statistical tests under specific assumption violations, and will allow an evaluation of the validity of the statistical results of previously published studies. This paper provides a methodological framework for, and an example of, quantitatively summarizing empirical Type I error and power values from Monte Carlo studies. The four-step summarization strategy covers problem formulation, data collection, data evaluation, and data analysis and interpretation. The example of a quantitative summary involves seven Monte Carlo studies of the parametric Bartlett test of the quality of independent variances. (Author/TJH)

Summarizing Monte Carlo Results in Methodological Research

Michael R. Harwell

University of Pittsburgh

April, 1990

Paper Presented at the Annual Meeting of the American Educational Research Association, Boston, April, 1990

Abstract

Monte Carlo studies of statistical tests are prominently featured in the methodological research literature. Unfortunately, the information from these studies does not appear to have significantly influenced methodological practice in educational and psychological research. One reason is that Monte Carlo studies lack an overarching theory to guide their interpretation. Also, the impressionistic nature of Monte Carlo studies can lead different readers to different conclusions. These shortcomings can be addressed using meta-analytic methods to summarize the results of Monte Carlo studies. Summarizing Monte Carlo studies in this fashion will generate guidelines for the appropriate use of particular statistical tests under specific assumption violations, and will allow an evaluation of the validity of the statistical results of previously published studies. This paper provides a methodological framework for, and an example of, quantitatively summarizing empirical type I error and power values from Monte Carlo studies.

Summarizing Monte Carlo Results in Methodological Research

Introduction

Quantitative methods of inquiry play a key role in educational and psychological research by providing evidence about the plausibility of substantive models (e.g., models specifying the form of the relationship between time on task and learning mathematics). Parametric statistical tests are among the most popular of these methods, and require that certain assumptions (e.g., normality of a population of scores) be tenable for the tests to yield valid conclusions. If these assumptions are not tenable the results may lack validity and therefore lead to incorrect inferences (Cook & Campbell, 1979). The results of critical examinations of educational and psychological data suggest that these data often fail to satisfy underlying assumptions (Stigler, 1977; Micceri, 1989). This suggests a need to identify the effects of assumption violations on statistical tests for data conditions (e.g., nonnormality) that are present in empirical studies in educational and psychological research.

Whenever possible exact statistical theory is used to determine the mathematical properties a test will have when its underlying assumptions are not tenable. This is seldom possible since most exact statistical theory requires normality of the population distribution of scores, an assumption which educational data rarely satisfies. Instead, researchers have examined the effects of assumption violations on parametric statistical tests using computer simulation or Monte Carlo (MC) studies.

In the typical MC study of a given statistical test the following process is repeated for a large number of samples: data are simulated which reflect a specified relationship among variables (but which do not usually conform to the assumptions required for correct application of the test), the statistical test is computed for the data, and the value of the statistical test is recorded. The collection of values of the statistical test provide information on its properties (e.g., the proportion of "significant" values of the test). If the underlying assumptions of the test were satisfied, exact statistical theory would guarantee that the test would have a specified type I error rate and would permit the probability of rejecting a false statistical hypothesis to be computed; MC studies permit these characteristics to be examined when underlying assumptions are violated.

An unfortunate characteristic that MC studies share with many empirical studies in educational and psychological research is their lack of an overarching theory to guide their interpretation. The absence of such a theory implies that each MC result is limited to the particular conditions of that study (e.g., the way data was simulated, sample sizes used). This problem is exacerbated by the impressionistic nature of MC studies, which can lead different readers to different conclusions regarding a particular MC study or a series of MC studies investigating the same statistical test. On the whole, MC results seem to have had little effect on methodological practice in educational and psychological research.

These shortcomings can be addressed using quantitative methods of research synthesis (e.g., meta-analysis), conceptualized by Glass (1976). Meta-analytic methods can be used to summarize the results of MC studies of any statistical test (e.g., ANOVA and ANCOVA F-tests) for which a sufficient body of literature exists. The goal of these methods in a MC setting is the correct modeling of the relationships between the MC results (i.e., empirical type I error and power values) and the characteristics (i.e., explanatory variables) of the simulation studies themselves.

An important outcome of such summaries would be the generation of a context within which to place these studies, i.e., a network of empirical (MC) results that would provide a context for interpreting past and future MC studies. The generation of a network of empirical MC results would lead to more comprehensive, definitive, and valid guidelines for the appropriate use of statistical tests (such as those associated with the ANOVA and ANCOVA models) than are presently available, and would permit frequently asked questions about these tests to be addressed. It would also enable educational and psychological researchers to better evaluate the validity of published statistical results involving these tests. For example, published research articles that used ANCOVA could be evaluated considering these guidelines.

This paper provides a rationale and a methodological framework for applying meta-analytic methods to summarizing MC results. The intent is to encourage methodological researchers to use these methods to quantitatively summarize MC results of particular statistical tests. This should contribute to improved methodological practice in educational and psychological research.

The organization of the paper is as follows: First, previous attempts to summarize the results of MC studies are reviewed. Second, issues related to study selection and variable definitions are discussed. Next, statistical procedures that can be used to quantitatively summarize MC results are presented. A small sample of MC studies of the parametric

Bartlett test of the equality of independent variances is used to illustrate the preceding issues and methods. The rationale for choosing the Bartlett test is that the availability of both analytic and empirical evidence of the effects of particular assumption violations (e.g., nonnormality of the population score distribution) will permit an evaluation of the usefulness of the proposed methods. Put another way, the effect of certain assumption violations on the Bartlett test of equality of independent variances are known a priori and thus the performance of the meta-analytic methods can be evaluated considering the relationships between the MC results and characteristics of the MC studies that should or should not be detected. Finally, the need for summaries of MC results for particular statistical tests is emphasized.

Previous Attempts to Synthesize Monte Carlo Results

The early 1970's witnessed the emergence of a substantial body of MC literature of the performance of particular statistical tests. Among the best known and most influential summaries of MC results from this period is that due to Glass, Peckham, and Sanders (1972). These authors narratively summarized the available MC studies of the F-test associated with the oneway fixed-effects ANOVA and ANCOVA models. Glass et al. concluded that the associated F-tests were quite robust to departures from the assumptions of normality and homogeneity of variance (the exception for the latter occurs when large and unequal samples are paired with small variances or when small and unequal samples are paired with large variances). A strength of the review, and one that should be emulated in constructing a network of empirical MC studies, is the effort that Glass et al. made to relate MC results to exact statistical theory.

Despite its comprehensive nature (for that time) and its influence on statistical practice, the Glass et al. review suffers from two shortcomings. Perhaps the most serious shortcoming is its impressionistic nature. Blair (1981) pointed out that a careful reading of the studies summarized by Glass et al. could allow different conclusions to be reached. This shortcoming is shared by recent narrative summaries of MC results in the educational and psychological research literature (e.g., Harwell & Serlin, 1988 (F-test in ANCOVA); Conover, Johnson, & Johnson, 1979 (F-test of equality of independent variances); Blair & Higgins, 1985 (matched-pair t-test)).

A second shortcoming of the Glass et al. review is that the results of recent MC studies of ANOVA and ANCOVA models suggest that certain of their recommendations need to be re-examined. For example, Glass et al. concluded that if sample sizes are equal, variance heterogeneity will have a negligible effect on the nominal type I error rate (i.e., a few hundredths of a percent). Yet Tomarkin and Serlin (1987) found that realistic combinations of heterogeneous variances noticeably affect the nominal type I error rate of the F-test even if sample sizes are equal.

Despite their shortcomings, qualitative reviews like Glass et al. should not be discounted; on the contrary, narrative summaries can provide valuable information about the performance of a statistical test across a sample of MC studies. However, there is a need to complement qualitative reviews like Glass et al. with quantitative methods of research synthesis (e.g., meta-analysis). A five-stage strategy for quantitatively summarizing the results of MC studies and associated issues are presented next.

A Four Step Strategy For Summarizing Monte Carlo Results

Quantitatively summarizing MC results for a particular statistical test can be represented in four stages following the framework of Cooper (1982): 1) problem formulation, 2) data collection, 3) data evaluation, and 4) data analyses and interpretation.

Stage One: Problem Formulation

The goal of quantitatively summarizing MC results is to generate guidelines for the appropriate use of a statistical test (e.g., F-test in the oneway fixed-effects ANOVA model) in educational and psychological research when underlying assumptions are violated. Operationally, the guidelines result from investigating the relationship between the simulation factors defining MC studies (e.g., type of population score distribution) and the empirical type I error and power values of a test. The latter are used to evaluate the performance of the test being investigated.

Stage Two: Data Collection

In the next stage an accessible population of relevant studies (e.g., available MC studies of the F-test in the oneway fixed-effects ANOVA model) is identified. This requires searching a variety of literature sources, e.g., Current Index to Statistics, Dissertation Abstracts International, Psychological Abstracts, ERIC. Several issues must be addressed in this process.

1. Issues of Study Selection

Identification of an accessible population of MC studies should be followed by formulation of a sampling plan. If the population of accessible MC studies is large, simple random sampling seems reasonable. Little information is available regarding the minimum number of studies that should be sampled to ensure a specified power for the statistical tests used in quantitatively combining MC results.

If the population of accessible studies is small, the sampling pian may reduce to simply using all available studies in the research synthesis. This was the stratcgy adopted by Harwell, Hayes, Olds, and Rubenstein (1990), who attempted to summarize MC studies of the F-test in the oneway fixed-effects ANOVA model. An extensive literature search by these authors, including the Current Index to Statistics, ERIC, and Dissertation Abstracts International, yielded approximately thirt" accessible studies. Given the small population of studies, Harwell et al. opted to use every available study in their meta-analysis.

A consequence of this decision could be the introduction of one or more study selection biases th may predispose the results of the meta-analysis. One manifestation of study selection bias occurs when the potentially nonrandom sample of studies to be summarized differ systematically from the population of studies. For example, differences among published and unpublished MC studies of the same statistical test might be due to a study selection bias. This kind of bias is often critical in substantive meta-analysis in which statistically significant results are more likely to be published (Rosenthal, 1979). However, this seems unlikely to pose a major problem in a meta-analysis of MC studies because publication of these studies is unrelated to notions of statistical significance. In addition, the nature of MC studies (e.g., researcher controlled data generation) suggests that, as a group, MC studies of the same statistical test are more homogeneous than empirical studies of the same phenomenon. This reduces the likelihood that an accessible sample of MC studies will systematically differ from a target population of studies, but does not release methodological researchers from the responsibility of checking for this kind of bias.

A study selection bias may also arise when a study is excluded from a meta-analysis because of perceived methdological flaws. This raises the question of whether all sampled studies should be included in the meta-analysis or whether a screening process should be employed to detect and remove those studies that are judged to be methodologicaly flawed. Unless there is convincing evidence that the results of a MC study are invalid there is little reason to exclude any of the sampled MC studies (random or not) from the meta-analysis.

As an example, suppose that a MC study investigated the performance of the F-test in the oneway fixed-effects ANOVA model across several conditions (e.g., sample sizes of 15, 25, or 40 per group). Suppose further that a nominal type I error rate of .05 was used and that the reported empirical type I error rates for a normal distribution, homogeneous variances, and sample sizes of 15, 25, and 40 per group were .06, .08, and .10, respectively. If the statistical assumptions underlying the F-test are satisfied (which is the case here), the empirical type I error rate of the F-test should converge to .05 as sample size increases. Barring typographical errors in reporting the MC results, the empirical error rates (i.e., .06, .08, .10) suggest a computer programing error. If this was the case the results would likely be invalid and might provide grounds for excluding this study from the sample of studies to be summarized. In any event, the decision to exclude a study because of percieved methodological flaws is judgemental in nature. Any decision to exclude a MC study because of suspected methodological flaws should be accompanied by an explanation of the basis of the decision and an indication of how the results might change if the study were included (Light & Pillemer, 1984, p.32).

2. Equatability of Outcome Variables

A frequent concern in meta-analysis is the requirement that the outcome variables in the studies to be summarized be linearly equatable (Hedges & Olkin, 1985, p. 108). In MC studies the outcomes (i.e., empirical proportions of rejections) share the same underlying metric and thus satisfy the requirement of being linearly equatable.

3. Estimating Effect Magnitude

The choice of an appropriate effect magnitude (EM) depends on the nature and purpose of a research synthesis. Glass (1976) defined EM through a standardized mean difference (i.e., effct size) between two samples (Glass, McGaw, & Smith, 1981). Other definitions of EM include measures of explained variance when the number of samples is greater than two (cf. Hedges & Olkin, 1985, pp. 100-103) and nonparametric effect size estimates (cf. Hedges & Olkin, 1985, pp. 92-100), and the dropout rate of students over time measured in proportions.

The results of a MC study are usually reported in terms of empirical proportions of rejections. These proportions can serve as EM estimates.

Consider the results of a MC study of the F-test in the oneway fixed-effects ANOVA model in which the effect of three types of population score distributions and three sample sizes are investigated. Suppose that the nominal type I error rate was .05 and that the empirical type I error rate for a given combination of conditions (e.g., small sample size and a nonnormal population score distribution) was .065. Compared with the nominal type I error rate, this value estimates the effect of the combination of MC conditions upon the type I error rate of the statistical test; the greater the difference between the empirical and nominal type I error rates the greater the effect.

Before continuing two comments are merited. First, the difference between an empirical type I error rate and the nominal error rate represents an index of the fit of an empirical sampling distribution of a statistical test under the null hypothesis and a theoretical reference distribution; the larger the difference the poorer the fit (and the less valid the probability statements) and vice versa. Empirical type I error rates that are (equally) above or below the nominal value suggest an equally poor fit. Second, the issues and methods described in this paper apply to both empirical type I error and power values.

4. Defining the Explanatory Variables

An issue that has plagued meta-analysis has been varying definitions of explanatory variables across studies of the same phenomenon (Strube & Hartmann, 1982). In a meta-analysis of MC studies the simulation factors serve as explanatory variables (e.g., type of population score distribution, sample size). For example, MC studies of the F-test in the oneway fixed-effects ANOVA model would usually include one or more of the following simulati i factors: type of population score distribution, number of groups, patterns of group sample sizes, patterns of group sample sizes, patterns of variance heterogeneity, and patterns of mean differences. Variation i the outcome variables is modeled as a function of these kinds of explanatory variables.

Adequately catagorizing the exp'anatory variables is a critical part of any meta-analysis. Consider the explanatory variable type of population score distribution. To examine the effects c this variable on the outcomes, proxy variables must be used that capture the effect of type of population score distribution. For example, suppose that each of twenty MC studies of the F-test in the oneway fixed-effects ANOVA model used varying skewness and kurtosis values to define population score distributions. Under this scheme, a population of scores having a skewness and a kurtosis of zero would be catagorized as a normal distribution, a population of scores with a skewness of zero and a kurtosis of three as a double exponential distribution, etc. (Kendall & Stuart, 1977, Vol. I). If sufficient information on the data generation process is available (e.g., skewness and kurtosis values), the effects of various population score distributions could be represented by the explanatory variables of skewness and kurtosis.

Occasionally somewhat arbitrary decisions must be made concerning the catagorization of particular explanatory variables. Consider a MC study that uses a variety of unequal sample sizes. Coding the pattern of unequal sample calls for some scheme that adequately captures this information. It might be sensible to limit the number of unequal sample size patterns or employ several coding schemes and look for a convergence in the results.

The next step is to enter the empirical type I error, power, and coded explanatory variable values into a computer data file in preparation for statistical analysis. Each line in a data file could correspond to a combination (i.e., cell) of simulation factors (i.e., explanatory variables) in a MC study. For example, the first line of data associated with a cell of a MC study of the F-test in the oneway fixed-effects ANOVA model might look like

```
01 00.00 00.00 3 10 .065 .777 2000                         (1)
```

From left to right, 01 is a study identification code, 00.00 and 00.00 are skewness and kurtosis indices (here the data was generated from a normal population of scores), 3 is the number of groups, 10 is the sample size per group, .065 is an empirical type I error value, .777 an empirical power value for a particular noncentrality structure, and 2000 is the number of samples used in computing the empirical type I error and power values.

The fact that the number of simulation factors is usually small (e.g., three or four), and the nature and specification of these factors (e.g., sample sizes used), usually clarifies what conditions were examined. This suggests, but does not ensure, that explanatory variables in MC studies are somewhat easier to define than their substantive counterparts.

5. Internal Validity

A particularily serious threat to the validity of a meta-analysis is the (lack of) internal validity of the results of each of the studies to be summarized, i.e., the extent to which the results of a study can be attributed to the explanatory variable(s) and not to other (confounding) factors (Campbell & Stanley, 1963, p.5). If there is evidence that the data generated in a MC study have the desired properties (e.g., are pseudo-random), this problem should not arise because the outcome variable values (i.e., empirical type I error and power values) can then be unambiguously attributed to the explanatory variables (i.e., simulation factors). This minimizes the role of potentially confounding variables and permits strong causal statments to be made (e.g., increasingly nonnormal population score distributions result in increasingly inflated type I error values).

6. Independence of Effect Magnitudes

The use of inferential statistical methods in meta-analysis requires that the EMs to be summarized are independent. Yet the same MC study will typically yield a large number of EMs which are to be summarized. For example, a MC study of the parametric ANCOVA model by Harwell and Serlin (1988) yielded sixty four EMs (i.e., empirical proportions of rejections) reflecting the performance of the F-test at a nominal error rate of .05 when the associated null hypothesis was true. The use of pseudo-random number generators to generate MC data implies that EMs like those in Harwell and Serlin (1988) are independent.

In short, MC studies do not appear to suffer from the range or magnitude of difficulties (e.g., study selection bias, lack of internal validity) that often plague substantive meta-analysis. This is due to the control that methodological researchers exercises over the data generation process in a MC setting and bolsters the credibility of a carefully conducted summary of MC results.

Stage Three: Data Evaluation

The complexity of many MC studies (e.g., Tomarken & Serlin, 1987), the difficulty of adequately catagorizing explanatory variables, and the possibly large number of EMs to be summarized suggests the need to employ competent reviewers to examine each MC study to ensure a consensus in what was studied, how assumption violations were modeled, etc. Each set of MC results should be examined considering 1) the data generation procedure, 2) evidence of the success of the data generation, and 3) reported type I error and power results when the assumptions underlying a statistical test are satisfied prior to coding.

The nature and amount of MC data to be coded and entered into a computer data file in preparation for statistical analysis virtually guarantees errors. This requires the development of a system in which these errors are detected and corrected. For example, the agreement among teams of reviewers coding the same MC data could be assessed with indices of inter-rater consistency (Cooper, 1982). Given the nature of MC studies, only complete agreement in what is being coded is acceptable.

Stage Four: Data Analyses and Interpretation

In the fourth stage, appropriate statistical methods for summarizing MC results are applied to the EMs. Good statistical practice calls for a thorough examination of the MC data to help to identify patterns in the data and to suggest particular analyses; however, this is unlikely to provide a complete and accurate description of the relationships between the outcome and explanatory variables. More formal procedures are needed to model variation in the outcomes (i.e., empirical type I error and statistical power values) as a function of key explanatory variables (e.g., type of population score distribution).

A primary reference for statistical procedures in meta-analysis is Hedges and Olkin (1985). These authors discuss a procedure in which a fixed-effects regression model is fitted to EMs. This procedure can be used to quantitatively summarize MC results.

Consider the population model

$$d_k = X_{k1}b_1 + X_{k2}b_2 + \ldots + X_{KT}b_T \quad k=1,..,K \qquad (2)$$

$$D = X\beta$$

In (2) d_k is the k(th) EM (proportion of rejections) which depends on a set of T fixed explanatory (i.e., predictor) variables X_{KT}, B_T is a regression coefficient which captures the relationship between the t(th) explanatory variable and d_k, D is a K x 1 vector of EMs, X is a K x T matrix of explanatory variable values, and β is a T x 1 vector of regression coefficients (see Hedges & Olkin, 1985, p. 169). The model in (2) could contain interaction and nonlinear predictor terms. Errors of prediction using (2) can be represented by $e_k = d_k - p_k$, where p_k is an empirical proportion of rejections from a cell (i.e., particular combination of conditions) in a MC study.

As an example of the use of (2), suppose that the empirical type I error values of twenty MC studies of the F-test in the oneway fixed-effects ANOVA model were to be summarized and that each study examined the same simulation factors (e.g., three types of population score distribution crossed with three sample sizes). Associated with each cell in each MC study are coded explanatory variables (see (1)) and an empirical type I error value for a given nominal error rate. Thus the total number of EMs to be summarized is 9 (cells) x 20 (MC studies) = 180 EMs. Concerning model (2), T = 3 and K = 180 in this example.

Tests for the presence of a relationship between the T explanatory variables and the outcome variable(s), and whether the explanatory model is correctly specified (i.e., whether all the variables needed to explain variation in the p_k are in the model), can be performed using procedures presented in Hedges and Olkin (1985, pp. 168-174).

The first step is to estimate the β_T in (2). Ordinary least squares estimation of the β_T is likely to be inappropriate because the EMs (i.e., empirical proportion of rejections) will have different variances (a violation of ordinary least squares estimation) when the number of samples (i.e., replications) differs across MC studies. Instead, weighted least squares estimation of the β_T can be performed using a diagonal weighting matrix with elements $1/\sigma^2_p = 1/[p_k(1-p_k)/S_k] = S_k/p_k(1-p_k)$, S_k = number of samples associated with each p_k (Hedges & Olkin, 1985, pp. 169-174). For hypothesis testing purposes the distribution of the e_k is assumed to be normal with a mean of zero and a diagonal covariance matrix given by Σ_p. This means that the errors are uncorrelated across the p_k but do not necessarily share the same variance.

Weighted least squares estimation of the β_T permits the regression model in (2) to be fitted to the p_k. A test for a relationship between the set of T explanatory variables and the outcome variable can be determined by testing the hypothesis

$$Ho: \beta_1 = \beta_2 = \ldots = \beta_T = 0 \qquad (3)$$

using the Q_R statistic presented in Hedges and Olkin (1985, p. 171). This statistic is equal to the weighted sum of squares due to regression associated with fitting a model like (2) to an outcome variable. Under the truth of (3), Q_R is distributed as a chi-square variable with T degrees of freedom (assuming the weighted least squares program includes an intercept in the regression model; see Hedges & Olkin, 1985, p. 174). Rejection of the hypothesis in (3) implies a relationship between the set of T predictors and the p_k; retention of this hypothesis implies that there is no relationship. Hedges and Olkin also provide expressions for constructing a confidence interval about individual β_T parameters.

An important use of the Q_R statistic is to test competing explanatory models. For example, an explanatory model might yield a Q_R statistic based on the variable total sample size. A competing explanatory model might yield a Q_R statistic based on the variables in model one plus skewness and kurtosis predictors. A chi-square test of the difference in the two Q_R statistics based on degrees of freedom ($T_{model\ 2} - T_{model\ 1}$) provides evidence about the role of the skewness and kurtosis predictors in accounting for variation in the p_k. The multiple R^2 provides evidence about the explanatory power of the model.

A key characteristic of the regression procedure discussed in Hedges and Olkin is the ability to test model specification (i.e., whether all the explanatory variables contributing to variation in the p_k are in the model; if they are not the model is misspecified and the results subject to model misspecification bias). A test of model specification can be performed when

K > T and the assumptions of uncorrelated errors which have been sampled from a normal distribution are tenable. An error sum of squares associated with (2) is computed using the statistic $Q_E = D'\Sigma_p\ D - Q_R$, where D, defined earlier, is a K x 1 vector of empirical proportions of rejections. If the model is correctly specified Q_E is approximately distributed as a chi-square variable with K-T-1 degrees of freedom. Rejection of the hypothesis that the model is correctly specified implies that the (weighted) error variance is larger than expected (i.e., the error term likely contains variation due to explanatory variables which should appear in (2)) (Hedges & Olkin, 1985, p.173). The difference between Q_E statistics for competing models can also be assessed. The Q_R and Q_E statistics are illustrated in the next section.

The information provided by the R^2, Q_R, and Q_E statistics has important implications for constructing explanatory models to summarize MC results. These statistics permit methodological researchers to tease out important relationships among explanatory variables and an outcome variable. They also provide evidence about model misspecification.

An Example of Quantitatively Summarizing Monte Carlo Results

A small data set consisting of seven MC studies of the parametric Bartlett (1937) test of the equality of independent variances is used to illustrate the process of quantitatively summarizing MC results. The fact that the behavior of this test is well known through analytic (Box, 1953) and empirical work (e.g., Conover, Johnson, & Johnson, 1981) will provide a standard against which to compare the results of the meta-analysis. For example, analytic and empirical work suggests that the shape of the population score distribution should play a key role in the ability of any explanatory model to predict variation in type I errors. Note that this example is not intended to be comprehensive; the purpose is to discuss issues raised earlier in the paper and to illustrate how MC results can be quantitatively summarized.

The population of accessible studies was defined to be all published MC studies of the behavior of the parametric Bartlett test of the equality of independent variances. The Bartlett test tests the hypothesis

$$Ho : \sigma^2_1 = \sigma^2_2 = \ldots = \sigma^2_J, \; j = 1,2,\ldots,J \qquad (4)$$

under the assumptions of independent and normally distributed errors. A nonrandom sample of seven MC studies investigating the performance of this test were used in the meta-analysis. Hence the results should be regarded as preliminary. The empirical proportions of rejections (i.e., p_k) served as the outcome variable. Finally, only the type I error case was investigated for a nominal error rate of .05.

Data from each of the seven MC studies were checked to ensure that the Bartlett test behaved as expected under certain conditions (e.g., empirical type I error rates converged to .05 as sample size increased when the assumptions underlying the Bartlett test were satisfied). No irregularities were noted. Four of the seven studies provided limited information about the random number generator.

Only a few simulation factors were examined in the seven MC studies and thus it was relatively easy to code the explanatory variables. The following variables were coded:

Skewness [SKEW] (actual value coded)

Kurtosis [KURT] (actual value coded)

Number of Groups [NUMGRPS] (2, J)

Total Sample Size [TOTALN] (actual value coded)

Number of Monte Carlo samples or replications [REPS1]

Empirical type I error values [TYPEI]

Two procedures were used to detect coding errors. First, the data file was scanned for obvious errors and corrections were immediately made. Second, the relatively small number of MC studies and EMs (K = 71) permitted each study to be reviewed a second time, with the coding checked on a line by line basis. Errors detected in this fashion were immediately corrected.

Research Question and Rationale

A single research question was addressed using the weighted least squares model outlined earlier: Does the explanatory power of a model predicting p_k using the predictors NUMGRPS, TOTALN, and REPS1 (model 1, T=3) differ from that of a model containing these same predictors plus SKEW and KURT (model 2, T=5)? The rationale for this question is that the analytic work of Box (1953) and subsequent MC results provide uncompromising evidence that the type I error behavior of this test is quite sensitive to nonnormal skewness and kurtosis. This suggests that the power of explanatory models with and without the predictors SKEW and KURT will differ dramatically.

Analyses

A nominal type I error rate of .05 was used for all hypothesis tests. The p_k were assumed to be uncorrelated and the normality assumption was judged to be tenable after examining the residuals from the analyses.

The hypothesis associated with model 1 was that the $T = 3$ explanatory variables (i.e., NUMGRPS, TOTALN, SAMPLE) were not associated with the outcome variable p_k, i.e., Ho: $\beta_1 = \beta_2 = \beta_3 = 0$. The regression statistic $Q_{R1} = 1066.5$ was computed for model 1, compared to a chi-square value based on $T = 3$ degrees of freedom, and found to be statistically significant. However, the squared multiple correlated coefficient, adjusted for the number of predictors, was $R^2_{adj\ model1} = .02$ (see Marascuilo & Serlin, 1988). this suggests a model with litle explanatory power, i.e., virtually no relationship between type I errors and the set of predictors.

The associated fit statistic for model 1, Q_{E1}, was computed as $D' \Sigma_p D - Q_{R1} = 50126 - 1066.5 = 49059.5$, where p_k is a $(K = 71) \times 1$ vector of empirical type I error values. This value was compared to a chi-square value based on $K-T-1 = 71-3-1 = 67$ degrees of freedom and was significant (i.e., the model is misspecified). This suggests that the explanatory power of model 1 could be improved by including additional predictors (e.g., SKEW, KURT).

The hypothesis associated with model 2 is that there is no relationship between the outcome variable p_k and the set of $T = 5$ predictors NUMGRPS, TOTALN, REPS1, SKEW, and KURT, i.e., Ho: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$. The regression statistic Q_{R2} was computed and found to be significant when compared to a chi-square value based on 5 degrees of freedom. The statistic $Q_{R2} - Q_{R1} = 43856.7$ provides evidence about the power of the explanatory model with and without the SKEW and KURT variables. This difference is compared to a chi-square with 5 - 3 = 2 degrees of freedom and is significant. Even more compelling is $R^2_{adj\ model2}$ = 97. This values suggests indicates there is a quite strong relationship between type I error and the model including the skewness and kurtosis predictors.

The fit statistic for model 2 was Q_{E2} = 48544 - 44359.7 = 4184.3, which was significant when compared to a chi-square value based on 71-6 = 65 degrees of freedom. Despite the size of the $Q_{R2} - Q_{R1}$ difference and the multiple R^2, the $Q_{E2} - Q_{E1}$ difference (significant compared to a chi-square based on 67 - 65 = 2 degrees of freedom) suggests that the explanatory model containing T = 5 predictors is still misspecified. Additional analyses in which interaction terms and nonlinear predictors were included in the explanatory model did little to reduce model misspecification.

Additional information about the contribution of the individual predictors is contained in the estimated regression coefficients, obtained from model 2 above. These coefficients and their standard errors (corrected following Hedges & Olkin, 1985, p.174), were

Predictor	Estimated Regression Coefficient	Estimated Standard Error
NUMGRPS	-.002	.0007
TOTALN	.000	.0000
SAMPLE	-.003	.0007
SKEW	.203	.0029
KURT	.058	.0007

These values provide additional evidence about the importance of SKEW and KURT in the explanatory model above.

In short, the meta-analysis clearly detected the relationship predicted by analytic work and supported by available MC results: that the shape of a population score distribution has a strong influence upon the type I rate of the Bartlett test for independent variances. Factors such as total sample size and number of MC samples appear to have little effect on the type I error rate of the Bartlett test. The misspecification of the explanatory model containing all five predictors is statistically significant but seems negligable in light of the associated R^2 for model 2.

Conclusion

The application of quantitative methods of research synthesis to summarize Monte Carlo results shows great promise for improving methodological practice. The goal is to produce an empirical network of Monte Carlo results of a particular statistical test that will generate guidelines for the appropriate use of a test under specific assumption violations. This will also permit previous statistical analyses to be evaluated considering these guidelines.

Conceptually, the process of quantitatively summarizing Monte Carlo results follows is quite similar to that of substantive meta-analysis, i.e., problem formulation, data collection, data evaluation, and data analysis and interpretation. The use of the weighted least square regression model provides a powerful tool for investigating the relationship between an outcome variable (e.g., type I error) and a set of explanatory variables.

The small meta-analysis for Monte Carlo results for the Bartlett test of independent variances suggest that these techniques can play a key role in constructing an empirical framework of Monte Carlo studies of a statistical test. These procedures should be especially useful for addressing longstanding questions about the behavior of particular statistical tests when underlying assumptions are violated. For example, the proposed methodology might prove to be especially valuable in examining the behavior of the F-test in the ANCOVA model and various multivariate tests (e.g., Rao F, Pillai-Bartlett) in the MANOVA model when assumptions are violated.

References

Bartlett, M.S. (1937). Properties of sufficiency and statistical tests. Proceedings of the Royal Society Series A, 160, 268-282.

Blair, C. (1981). A reaction to "Consequences of failure to meet assumptions underlying fixed effects analysis of variance and covariance." Review of Educational Research, 51, 499-507.

Campbell, D.T., & Stanley, J.C. (1963). Experimental and quasi-experimental designs for research. In Handbook of Research on Teaching. Chicago, Rand-McNally.

Cook, T.D., & Leviton, L.C. (1980). Reviewing the literature: A comparison of traditional methods with meta-analysis. Journal of Personality, 48, 449-472.

Cooper, H.M. (1982). Scientific guidelines for conducting integrative research reviews. Review of Educational Research, 52, 291-302.

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

Glass, G.V., McGaw, B., & Smith, M.L. (1981). Meta-analysis in social research. Beverly Hills, CA: Sage.

Glass, G.V., Peckham, P.D., & Sanders, J.R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. Review of Educational Research, 42, 237-288.

Harwell, M.R., Hayes, W., Olds, C., & Rubinstein, E. (1990). Summarizing Monte Carlo results in methodological research: The oneway fixed-effects ANOVA case. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, April.

Harwell, M.R., & Serlin, R.C. (1988). An empirical study of a proposed test of nonparametric analysis of covariance. Psychological Bulletin, 104, 268-281.

Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis. New York: Harcourt, Brace, Jovanovich:

Jackson, G. B. (1980). Methods for integrative reviews. Review of Educational Research, 50, 438-460.

Kendall, M., & Stuart, A. (1977). The advanced theory of statistics (Vol. I). New York: Macmillan.

Kocher, A. T. (1974). An investigation of the effects of nonhomogenecus within-group regression coefficients upon the F-test of analysis of covariance. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, January.

Marascuilo, L.A., & Serlin, R.C. (1988). Statistical methods for the social and behavioral sciences. New York: Freeman.

Micceri, T. (1989). The unicorn, the normal distribution, and other improbable creatures. Psychological Bulletin, 105, 156-166.

Rosenthal, R. (1978). Combining results of independent studies. Psychological Bulletin, 85, 185-193.

Stigler, S. (1977). Do robust estimators work with real data? The Annals of Statistics, 5, 1055-1098.

Strube, M.J., & Hartman, D.P. (1983). A critical appraisal of meta-analysis. British Journal of Clinical Psychology, 51, 14-27.

Tomarkin, A., & Serlin, R.C. (1986). A comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. Psychological Bulletin, 99, 90-99.

Monte Carlo Studies Used in the Meta-analysis

Brown, M.B., & Forsythe, A.B. (1974). Robust tests for the equality of variances. Journal of the American Statistical Association, 69, 364-367.

Conover, W.J., Johnson, M.E., & Johnson, M.M. (1981). A comparative study of tests for homogeneity of variances with applications to the outer continental shelf bidding data. Technometrics, 23, 351-361.

Gartside, P.S. (1972). A study of methods for comparing several variances. Journal of the American Statistical Association, 67, 342-346.

Keselman, H.J., Games, P.A., & Clinch, J.J. (1979). Tests for homogeneity of variance. Communications in Statistics - Simulation and Computation, B8, 113-129.

Layard, M.W.J. (1973). Robust large-sample tests for homogeneity of variances. Journal of the American Statistical Association, 68, 195-198.

Levy, K.J. (1975). An empirical comparison of several multiple range tests for variances. Journal of the American Statistical Association, 70, 180-183.

Penfield, D.A., & Koffler, S.L. (1985). A power study of selected nonparametric K-sample tests. Paper presented at the Annual Meeting of the American Educational Research Association, March, Chicago.