UCLA
Department of Statistics Papers

Title
Statistical Inference and Meta-Analysis

Permalink
https://escholarship.org/uc/item/1186950j

Author
Berk, Richard

Publication Date
2006-06-16

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

DRAFT
Statistical Inference and Meta-Analysis

Richard Berk
Department of Statistics, UCLA

May 16, 2006

1 Introduction

Statistical inference is an important feature of meta-analysis. Estimation is often a central goal, with hypothesis tests and confidence intervals used to address uncertainty. Expositions of meta-analysis make statistical inference a major theme (Hedges and Olkin, 1985; Hedges, 1993; Raudenbush, 1983; Lipsey, 1992; 1997; Fleiss, 2003: chapter 10). Indeed, a significant motivation for meta-analysis can be improving the precision of the estimates produced and increasing the power of any hypothesis tests.

In the pages ahead, the use of statistical inference in meta-analysis will be examined. The intent is to consider the statistical models employed and the data with which they are used. Building on some previous work (Wachter, 1988, Berk and Freedman, 2003, Briggs, 2005), a key issue will be whether the data were produced by the mechanisms that the models require. The paper begins by describing some popular meta-analysis models. An assessment of their use follows.

2 The Basic Meta-Analysis Model

Statistical inference is a process by which information contained in a data set is used to drawn conclusions about unobservables. For meta-analysis, there is a model representing how the studies to be summarized came to be. This model has unobservable parameters. Investigators use information from the studies on hand to estimate the values of these parameters.

Consider the basic meta-analysis model. There are $m = 1, 2, \ldots, M$ studies to be summarized. Within each study, there is a treatment group and a control group. The treatment group is exposed to some intervention. The control group gets an alternative, often just the status quo. Interest centers on the difference between the mean response of the experimentals and the mean response of the controls. Under the basic model, such differences are the result of a single, treatment effect shared by all studies and a within-study random error.

More formally, if α is the common treatment effect, and ϵ_m is the random error associated with each study, the basic "fixed effects model" is

$$\delta_m = \alpha + \epsilon_m, \tag{1}$$

where δ_m is the realized treatment effect for study m. It can be measured the units in which the outcome is measured (e.g., the homicide rate) or in standard deviation units. The units in which δ_m is measured determine the units of α and ϵ_m. More will be said about this later.

The model requires that $E(\epsilon_m) = 0$ and that the disturbances represented by ϵ_m are independent of one another. One imagines a very large number of hypothetical studies in which the value of α is shifted up or down by an additive, chance perturbation to produce δ_m. These chance perturbations are independent across studies, which makes the δ_m independent across studies as well.[1] The variance of ϵ_m, often represented by σ_m^2, is usually assumed to be the same for the experimental group and control group. All three assumptions about ϵ_m are weighty, and their implications will be addressed shortly.

Researchers never get to see any of equation 1. Equation 1 is a *theory* about how the results of a given study are produced. All that follows only makes sense if this theory is a good approximation of what really happened.

Equation 1 can be elaborated in many ways. Very commonly, a set of covariates is included, so that equation 1 applies conditional on these covariates. In the absence of random assignment, for example, the covariates

[1] The assumption that $E(\epsilon_m) = 0$ implies that α alone is responsible for any systematic differences between the experimentals and controls. The assumption of ϵ_m independence implies that for estimates of α, hypothesis tests can be undertaken and confidence intervals can be constructed with conventional expressions.

might provide information on how membership in the experimental or control group was determined. For the points to be made in this paper, nothing fundamentally changes. Other elaborations of the basic model will be considered later.

2.1 Using the Basic Model

Data to inform the basic model must include for all study subjects a response measure and an indicator for membership in the experimental or control group. The data will often include covariates as well. Interest usually centers on the difference between the mean response for the experimentals and mean response for the controls.

In many settings, the difference between means shown in equation 1 is taken to be in standard deviation units. Then, the observed difference in means is divided by the standard deviation of the outcome variable. Specifically, for each of $m = 1, 2, \ldots, M$ studies

$$d_m = \frac{\bar{y}_m^T - \bar{y}_m^C}{s_m}, \tag{2}$$

where d_m is the standardized effect size for study m, $\bar{y}_m^T$ is the mean response for the treatment group, $\bar{y}_m^C$ the mean response for the control group, and s_m is the standard deviation of the response in study m. For ease of exposition and consistent with most studies in criminal justice settings, standardized treatment effects will be employed from here forward.

The standard deviation of any given study is computed as.

$$s_m = \sqrt{\frac{V_m^T(n_m^T - 1) + V_m^C(n_m^C - 1)}{n_m^T + n_m^C - 2}}. \tag{3}$$

V_m^T and V_m^C stand for the variance of the response variable for the treatment group and the control group respectively, and n_m^T and n_m^C are the corresponding sample sizes. Equation 3 can be justified by assuming that $V_m^T = V_m^C$. Whether this equivalence is reasonable is addressed below.

Because the goal is to summarize the findings across studies, an average effect (over studies) is computed as an estimate of α. It is a weighed average, with weights the inverse of the variance of each study's standardized treatment effect. Studies with larger chance variation are given less weight when the weighted average is computed.

More formally, the standard error for each study is

$$se_m = \frac{s_m}{\sqrt{n_m^T + n_m^C}}. \tag{4}$$

Then, the weight for each study is the inverse of each study's squared standard error:

$$w_m = \frac{1}{se_m^2}. \tag{5}$$

To compute for M studies the weighted average standardized effect, one uses,

$$\overline{d} = \frac{\sum_{m=1}^{M}(w_m \times d_m)}{\sum_{m=1}^{M} w_m}. \tag{6}$$

Finally, the standard error of the weighted average is computed as

$$se_{\overline{d}} = \sqrt{\frac{1}{\sum_{m=1}^{M} w_m}}. \tag{7}$$

With the weighted average and its standard error in hand, confidence intervals and hypothesis tests can be produced as usual.

2.2 A Simple Illustration

To help make the preceding discussion more concrete, consider a very simple example. Table 1 shows some summary statistics from three hypothetical studies. There is an experimental and control group in each. For each study, the table reports in the first three columns from left to right the sample size, the difference between the mean of the experimental group and the mean of the control group (in their original units), and the standard deviation of the response variable.

N	$\bar{y}_m^T - \bar{y}_m^C$	s_m	d_m	se_m	w_m
100	30	17.0	1.76	1.70	0.35
80	20	11.0	1.81	1.23	0.66
120	40	21.0	1.90	1.92	0.27

Table 1: Summary Statistics for Three Studies

The observed standardized effects, computed using equations 2 and 3, are found in the fourth column. Using equation 4, the standard errors for each study are shown in the fifth column. In the sixth column are the weights, computed using equation 5. From equation 6, the weighted average is then 1.82. This is the estimate of α.

The standard error of the weighted average, computed using equation 7, is .88. The 95% confidence interval is then 0.10 to 3.54. A test of the null hypothesis that the treatment effect is 0, leads to a t-value of 2.06. Because the p-value is smaller than .05, the null hypothesis is rejected. The same null hypothesis would not have been rejected for each study by itself. This is the kind of result that can be used to justify meta-analysis. Three studies with null findings, when properly combined, lead to a rejection the null hypothesis of no treatment effect.

3 Interpreting the Basic Model

The basic meta-analysis model is simple, and the computations that follow are simple as well. Simplicity is good. However, because the model connects statistical inference to how the studies were produced, simplicity by itself is insufficient. We need to ask whether that connection performs as advertised. A useful place to begin is with a discussion of the common treatment effect α. What is the nature of the key parameter meta-analysts are trying to estimate?

3.1 Causality

Meta-analysis is commonly used to estimate the impact of an intervention. Interventions are manipulable within the real world in which the summarized studies have been done. Thus, treatment effects are causal effects and in this context, α is the "true" effect of some cause.

Some researchers apply meta-analysis to draw causal inferences when the "intervention" is a fixed attribute of an individual, such as sex (Archer, 2000) or race (Mitchell, 2005). Regardless of the statistical procedure used, it makes little sense to ask, for instance, whether a prison sentence would change if the offender's race were different, when it could not have been different (Holland, 1986; Rubin, 1986; Berk, 2003; Freedman, 2004).

But, there is nothing in equation 1 that requires cause and effect. One may treat α as a description of a difference between the mean of one group and the mean of the other. It is possible to conduct a meta-analysis of the *association* between race and the length of prison sentence as long as no causal interpretations are imposed.

In short, although meta-analysis is usually motivated by a desire to estimate a causal effect, description will suffice. Problems arise when variables that cannot be manipulated are analyzed as causes. Further confusion can result when the language of experiments is used when it does not apply. We will proceed from here onward within a causal framework because cause and effect are usually key concerns in a meta-analysis. However, description is not precluded.

3.2 Standardization

If all of the studies to be summarized have outcomes in common and interpretable units, estimates of α are in those units as well. There is no need to standardize the units across studies; α is the single, common treatment effect in natural units. Suppose, for example, that α is three fewer burglaries committed by individuals in a certain kind of drug treatment program compared to individuals under conventional parole supervision. Then, there will be exactly three fewer burglaries for *each* such drug treatment program studied, save for random error.

In a wide variety of social science and criminal justice investigations, the units in which the response is measured vary fundamentally. For example, the outcome may be the homicide rate in one study and the number of homicides in another study. Then, standardization is undertaken so that the treatment effects can be more appropriately compared.

Under this scenario, the basic meta-analysis model requires the intervention has the same effect in each study in standard deviation units (e.g., -1.2), except for random error. The only systematic difference across studies is a difference in scale. This means that if the response variable in each study had the same standard deviation, the differences between the mean of the experimentals and the mean of the controls would be exactly the same. Then, Equations 2 and 3 follow directly.

However, standardization depends on two assumptions about how the real world functions. Recall that the basic model requires that in each study the variance of the experimentals is the same as the variance of the controls: $V^T_m = V^C_m$. This is the first assumption, and it depends on a theory of how the treatment affects the response.

For example, when the response to the treatment is *proportional* to the response to the control condition, *both* the mean and the variance will differ between the experimentals and controls. If, for example, the variance of the controls is V_m^C, and if the treatment increases the response by multiplicative factor of k, the variance of the treatment group, V_m^T is multiplied by k^2: $V_m^T = k^2 V_m^C$. However, if the intervention functions by adding a constant k, which is the usual assumption, the equal variance assumption may be credible. Credibility will depend on the precise nature of the study, what is being measured, and the composition of the experimental and control groups.[2]

The equal variance assumption is not a mere technicality. If the effect is really additive, computing a standardized difference between means is may not be misrepresenting impact of the treatment. But if the effect is multiplicative, the treatment impact is being misrepresented from the start. Therefore, a strong rationale must be provided in subject-matter terms for why an additive effect (or some other effect) is being assumed. Why does the phenomenon being studied work that way?

The second assumption is that the intervention has exactly the same effect, in standard deviation units, in the different studies. This is a very strong statement about how treatment effects are produced and requires very strong justifications. Why does the phenomenon being studied work that way?

To help fix these ideas, suppose the three studies in Table 1 report the impact of "hot spot" policing. Hot spot policing is introduced in three cities. In each of the three cities, some precincts get hot spot policing and some do not. The short-term response is the number of arrests. But, why might that intervention increase the standardized number of arrests in the treatment precincts by, in effect, adding a constant to the standardized number of arrests in the control precincts? Perhaps just as plausibly, there is a proportional increase. But even assuming that the additive model is correct, what is it about the social processes responsible for arrests and the social processes inherent in hot spot policing that require the exact same increase

[2] An alternative assumption is that the variance of the experimental group's response is the same as the variance of the control group's response under the null hypothesis that $\alpha = 0$. Then the variance of the control group's response can be used as the "common" variance. But if the null hypothesis is rejected, then what? One needs a new model that can be justified in which the variances are not the same.

in the standardized number of crimes in the three cities? The burden is on the meta-analyst to make the case and for this illustration at least, it would be a difficult case to make.

In many social science applications of meta-analysis, what is being measured is arguably quite different in different studies. For example, whether a parolee commits a parole violation is rather different from whether a parolee commits a felony. At a higher level of abstraction, both are parole failures that for purposes of a meta-analysis some might treat as the same. However, then the assumption of a common effect across studies, except for scale, seems especially tenuous.[3]

In summary, α in equation 1 is the parameter of main interest. It is by assumption a common, treatment effect, usually taken to be a causal. Whether α corresponds to anything real depends on how the intervention affects the response in real life. The value of α is what meta-analysts want to estimate. Statistical inference typically addresses various properties of α.

3.3 Statistical Inference

Researchers never see α. They see d_m, which by equation 1 is α perturbed by how the data in each study are generated. Statistical inference is a process by which researchers use the d_m to learn about α. Consider now some specifics.

From Table 1, it is clear that despite the assumption of an common effect, the three effects are a bit different: 1.76, 1.81 and 1.90 standard deviations. Perhaps the simplest explanation is that the differences result from random error introduced by each study's research design.

If in each study, the units are assigned to the experimental and control conditions by random assignment, random assignment is to blame. Were it possible to go back in time and randomly assign the units a second time, the results would be a little different, even if nothing else had changed. By the luck of the draw, the composition of the experimental and control groups would likely be altered, which would tend to alter the observed difference between the mean of the experimental group and the mean of the control group. By similar reasoning, the three studies have observed effect sizes that vary from one to another because of random assignment alone. The fact that the effect sizes are really equal in standardized units is obscured by an assignment process that shuffles the composition of the experimental and

[3] For a further discussion see Berk and Freedman (2003) and Briggs (2005).

control groups from study to study. In one study, for example, higher crime neighborhoods may be a bit over-represented in the experimental group, and in another study, higher crime neighborhoods may be a bit over-represented in the control group.

If the studies are observational, the same basic logic can apply. It is common to assume that after conditioning on the set of covariates thought to be related to the response and the intervention, nature undertakes the equivalent of random assignment. Once again, chance is introduced into the observed response and is the only reason why the observed effect sizes differ.[4]

Given the existence of α, random assignment to experimental and control conditions, whether by a researcher or by nature, makes the assumptions about ϵ_m reasonable. The perturbations that turn α into δ_m tend to cancel out over a large number of studies and are independent of one another. It follows that the δ_m they are independence *across studies*. Thus, the mean treatment effect of a sufficiently large number of studies will have about the same value as α, and if for one study the observed treatment effect is by chance too large, the likelihood that the next study's treatment effect will be too large or too small is unaffected. These are very attractive consequences of the basic meta-analysis model and permit easy construction of proper statistical tests and confidence intervals.

There an alternative interpretation of equation 1 based on random sampling instead of random assignment (Hedges and Olkin, 1985). It begins with a population of experimentals exposed to the treatment condition and a population of controls exposed to the control condition. In the population, α is the standardized difference between the mean of the experimental population and the mean of the control population. The variance of response is the same in both populations. Then, the data for each study is a simple random sample for each of the two populations. The standardized difference between the two observed means is an estimate of α. Variation in these estimates across studies results from the random sampling. Like random assignment, random sampling will alter the composition of the experimental group and control group from study to study.

Just as for the random assignment interpretation, the random sampling can be an act of researchers or and act of nature. In the first case, there

[4] Whether it is plausible to proceed as if nature conducts the equivalent of a randomized experiment, conditional on a set of covariates, must be examined with great care (Rosenbaum, 2002). How that conditioning is done matters as well (Berk, 2003). Covariance adjustment via regression are one popular option. Matching is another.

is a population of experimentals and a population of controls, and for each study researchers sample at random from the two populations. For example, researchers might take a random sample from a population of school drop outs and a random sample from a population of individuals who completed high school. Comparisons might be made using the number of arrests between the ages of 18 and 25.

In the second case, there is a population of experimentals and a population of controls, and for each study nature provides a random sample from each of the two populations. Such study might also compare individuals who dropped out to those who did not within a given school district. A meta-analyst would have to argue that because of the way social processes allocate children to school districts, the experimentals and controls from the given district are effectively simple random samples from some well defined population of students (e.g., all students from school districts with the same mix of income and racial groups).[5]

The random sampling interpretation of the basic meta-analysis model has essentially the same consequences for statistical inference as the random assignment interpretation. However, the random assignment rationale seems better suited for thinking about interventions and is somewhat more straightforward. We will continue to emphasize the random assignment formulation.

In summary, the random assignment assignment and random sampling formulations both depend on α. There exists a single treatment effect for a well-defined collection of studies. Thinking back to the policing hot spots illustration, it is possible to define a set of cities, interventions, and circumstances in which the standardized number of arrests is increased by exactly the value of α, except for random variation due to within-study random assignment or random sampling (by researchers or nature). But, what if no such case can be made? What are the options? We turn to some now.

4 Extensions and Elaborations of the Basic Model

The basic model can be made a more elaborate by allowing for many treatment effects. On its face, this seems reasonable. It is difficult in practice

[5] In both instances, researchers would likely condition on a variety of covariates such as race and gender.

argue convincingly for a single treatment effect, even if standardized.

4.1 Random Effects Model

In one formulation, the unobserved treatment effects differ because of random variation in the "true" treatment effect. One has a "random effects model." Thus,

$$\delta_m = \alpha_m + \epsilon_m. \tag{8}$$

The random effects model requires that the α_m are independent of one another, and that α_m is independent of ϵ_m. All of the original assumptions for ϵ_m continue to apply. The usual goal of the meta-analysis is to estimate the overall mean of the many standardized treatment effects, taking their random variability into account when confidence intervals and statistical tests are performed.

Translating equation 8 into something real takes a bit of thought, and it is possible to construct more than one account. Under the basic model, there was an unobserved common treatment effect. What we got to see was random variation over studies because of how the data for each study were generated (e.g., by random assignment). By one random effects account, each study now has its own unobserved treatment effect that varies randomly across studies. This random variation is not a result of how the *data* are generated, but inherent in how in each study social processes link the intervention with the response. This link contains some noise.

This is necessarily a theory about how the scientific community functions. The usual null hypothesis that the average treatment effect is zero, for example, requires a remarkable balancing act. The random variation in standardized treatment effects exactly balances so that positive perturbations cancel out negative perturbations. How does the scientific community manage to do that?[6] Independence of the α_m means, for instance, that if one study has a larger effect than average, the next study is no more or less likely to be above average. Random variation in the true effect sizes is unrelated to the chance process by which the data for any particular study are produced. How plausible this is depends on the fine print of each study's research design and how it might be related to size of each study's true treatment effect. Finally, there remains the requirement of independence between studies en-

[6] It is hard enough to get competent peer reviews done in a timely manner.

forced now by the α_m as well as the ϵ_m. Working through these details in a meta-analysis requires a very rich theory of the collective scientific enterprise.

A second popular account for equation 8 begins with a population of studies that vary in the true treatment effects. In the population, each study has its own *fixed* α_m. The set of studies on hand is taken to be a random sample from this population. That is, in addition to the within-study chance process represented by ϵ_m, there is a second chance process that results from the sampling studies at random. The weighted mean over the sampled studies is usually treated as an estimate of the average standardized treatment effect in the population, but there is not a single α to be estimated. This account is also consistent with equation 8, and no easier to justify. What population are we talking about? Where does the random sampling come from? And again, there is the requirement of independence across studies.

Statistical tools for the analysis of random effects models can come in several forms. In the most simple case, the primary difference between how the basic model is analyzed and how the random effects model is analyzed is that for the latter, the standard error for the estimated overall mean has to take two chance processes into account, one represented by ϵ_m and one represented by α_m. Empirical Bayes methods go several steps farther and also allow one to capitalize on information in the overall mean to obtain better estimates of the effects in each of the individual studies. Under either method, it is usually possible to partition the overall variance into a component due to between studies variability and a component due to within studies variability, a "components of variance" approach.

4.2 Systematic Fixed Effects Models

Sometimes meta-analysts are more comfortable treating the study to study variation not as random, but as systematic. Results may differ because of the setting in which the study was done, the mix of study subjects, the research design used, or a host of other reasons. The α_m in equation 8 are now interpreted as fixed. Each study has an unobservable, stable, treatment effect of its own that is not a product of a chance process. One also can think of the α_m as representing bias inherent in each study.

If the variation in the unobserved treatment effect is a function of a set of measurable variables, the pull to some form of regression analysis is almost irresistible. The formulation that results is another kind of "fixed effects model" and will often have the following structure.

$$\delta_m = \alpha_m + \epsilon_m, \tag{9}$$

where

$$\alpha_m = X\beta. \tag{10}$$

There are now a set of p predictors contained in the $M \times p$ matrix X and a set of $p+1$ regression coefficients including a constant term.[7]

At best, a regression approach only postpones the day of reckoning. The many pitfalls of regression analysis (Berk, 2003, Freedman, 2005) are introduced, and one still must make sense of ϵ_m, but more so. One must still explain why the ϵ_m are independent of one another and have the same variance. Then, one must have of a theory of the scientific enterprise that explains why ϵ_m is independent of (or at least uncorrelated with) any of the predictors. Unless all of the studies are randomized experiments, this is a daunting task. For example, perhaps weaker designs are associated with larger chance variation that is likely to produce desirable results.

4.3 Models for Dependence between Studies

There is sometimes a concern that there are natural groups of studies within which they may be dependence. For example, what may first seem to be ten distinct studies, may actually be two outcomes each within five distinct studies. There is then reason to worry about within-study dependence. This would be a clear threat to the usual requirement of independence across studies.

An illustration of more than one outcome per study might be parole failure measured by the elapsed time to an arrest for a person crime and parole failure measured by the elapsed time to an arrest for a property crime. In effect, there is a new level of nesting: outcomes within studies.

The nesting principle generalizes. For example, the nesting may result from studies done by researchers in the same research institution, or it may result from studies done in common settings such as particular police departments or political jurisdictions. If it is possible to know in advance that nesting exists, and if there is information on which studies should be grouped,

[7] It is possible to get fancier. One can add an error term to $X\beta$. Such formulations are sometimes called multilevel. However, the basic concerns are unchanged. There is just a new layer of complexity to explain.

there are models that in principle can adjust for the dependence (Gleser and Olkin, 1994). However, these models are extensions of one or more of the models already described and require even more elaborate theoretical understandings about the sociology of science. For example, there is now a mandatory independence between the groupings of studies.

4.4 The Special Case of Randomized Experiments

There is one situation it which statistical inference can be easily justified in a meta-analysis. If all of the studies to be summarized are randomized experiments, and if there is a null hypothesis of no effect, all of the models reduce to $\delta_m = \epsilon_m$ under the null hypothesis. Then, the hypothesis test follows directly.

However, it is not clear what should be done if the null hypothesis is rejected. One may conclude that the null hypothesis is incorrect, but then what? In order to decide what to estimate from the data, a model of the treatment effects is necessary. For example, is the overall standardized mean of the studies an estimate of a common standardized treatment effect? And if so, what is it about the intervention and the response that make such a claim plausible? In short, all of the earlier concerns reappear.

4.5 Other Variations

Meta-analyses can take a larger number of other twists and turns. Perhaps most important, summaries of the outcomes for the experimentals and controls are not limited to means. One can use proportions, risk ratios, odds ratios and other calculations. All of the models discussed above can be altered to take these variations into account. But for purposes of this discussion, nothing significant changes. The problems raised are only altered around the edges.[8]

In summary, a fundamental concern with any meta-analysis model is how well it represents the manner in which the studies were produced. Understandings of how the relevant scientific community functions must be consistent with the meta-analysis model. More complicated meta-analysis models are not necessarily more realistic. And if the model is not realistic, there can

[8] But, a number of quite daunting technical problems sometimes surface (Berk and Freedman, 2003).

be serious consequences. We turn now to a more explicit discussion of those consequences.

5 A Primer on Sampling Distributions for Meta-Analysis

All statistical inference depends on the concept of a sampling distribution.[9] For meta-analysis, the nature of the sampling distribution depends on how the studies to be summarized came to be. If the correspondence between the study generation process and the sampling distribution is poor, any confidence intervals and statistical tests will likely be misleading.

To appreciate why this is true, we consider a simple illustration based on the random effects model discussed as equation 8. We build on the account in which there is a population of studies that differ in their underlying treatment effects α_m. In the population, there is also random variation across studies resulting from the chance process by which subjects are assigned to the experimental or control groups. The studies on hand are a simple random sample from that population. Other models and other stories would likely lead to more a complicated discussion, and would not materially alter the points to made.

Suppose the population consists of five studies. A simple random sample of three studies is chosen. For the three studies, a weighted average of the standard effects is computed, by the methods discussed earlier. It is desirable that the weighted mean computed from the sample of three studies be an unbiased estimate of the mean of the population. What does unbiasedness require?

In the population of five studies, each study has a probability of 1/5 of being selected in the first draw. With one study chosen, each of the remaining studies has a probability of 1/4 of being selected. Finally, with two studies chosen, the remaining studies each have a probability of 1/3 of being selected.

[9] In meta-analysis, occasional reference is made to Bayesian statistical inference in which sampling distributions play no role. But real applications are difficult to find, and the inferential goals are very different (Lewis and Zelterman, 1994). Moreover, one still has to specify a credible likelihood function, which can raise the same kinds of issues discussed in this paper. Suffice it say, the alternative of Bayesian inference has some real strengths, but in the end trades one set of problems for another (Barnett, 1999).

The sampling is, therefore, without replacement and at each draw, all of the remaining studies have the same probability of selection.

Whether simple random sampling leads to an unbiased estimate depends not on the result from a single sample, but what happens over all possible samples from the population. Table 2 contains all possible samples of size three from a population of five. The first column shows the studies chosen; the studies are indexed as $1, 2, 3, 4, 5$. The second column shows the standardized effect size associated with each study; the effect sizes in order are $3, 1, 0, 1, 5$. The third column shows the weighted average. For ease of exposition and with no impact on the points to be made, all of the studies are given equal weight.

Study Index	Sample Values	Sample Mean
1,2,3	3,1,0	1.33
1,2,4	3,1,1	1.67
1,2,5	3,1,5	3.00
1,3,4	3,0,1	1.33
1,3,5	3,0,5	2.67
1,4,5	3,1,5	3.00
2,3,4	1,0,1	0.67
2,3,5	1,0,5	2.00
2,4,5	1,1,5	2.33
3,4,5	0,1,5	2.00

Table 2: Illustrative Sampling Distribution: Mean Standardized Effect Size = 2.0

There are ten possible samples that result from choosing three studies by simple random sampling from a population of five studies. Because of simple random sampling, each sample has the same probability of being selected. That probability is 1/10. It follows that each weighed mean also has a probability of 1/10. The ten weighted means and their associated probabilities constitute a sampling distribution. This is the key to all that follows.

If one multiplies each sample's weighted average by 1/10 and sums them, the result is 2.0. For the population of studies, mean is also 2.0. By definition, therefore, the weighted average from any *one* of the ten possible samples is an unbiased estimate of the population mean. This is good because under the scenario by which the population of studies were produced, one might be interested in what an average treatment effect might be.

There are several points to be taken from this illustration.

1. The five studies constituted a real population; all of the studies in the population were identified and all could have been reviewed, at least in principle.

2. The studies in the population differed in their standardized treatment effects and also differed because of within-study random variation.

3. A single sample of three studies was chosen by probability sampling, here, simple random sampling.

4. The thought experiment all possible samples, derived from simple random sampling, then followed directly and led to the theoretical construct of a sampling distribution.

5. From this theoretical sampling distribution, it was possible to illustrate how the weighted average from any sample of the three studies was an unbiased estimate of the population mean.

More generally, a properly weighted mean computed from a set of studies is an unbiased estimate of the population mean when coupled with simple random sampling. Building on Thompson's exposition (2002, section 2.6), let each sample of n observations be indexed by g so that the probability that sample g is selected is $P(g)$. Then,

$$E(\bar{d}) = \sum \bar{d}_g P(g) = \sum_{i=1}^{N} \bar{d}_i \frac{\left(\begin{array}{c} N-1 \\ n-1 \end{array} \right)}{\left(\begin{array}{c} N \\ n \end{array} \right)} = \frac{1}{N} \sum_{i=1}^{N} \bar{d}_i, \tag{11}$$

where the number of combinations of n studies from a population of N studies is

$$\left(\begin{array}{c} N \\ n \end{array} \right) = \frac{N!}{n!(N-n)!}. \tag{12}$$

In equation 11, $\left(\begin{array}{c} N-1 \\ n-1 \end{array} \right)$ is number of samples in which study i is included and $\left(\begin{array}{c} N \\ n \end{array} \right)$ is the total number of distinct samples. So, the ratio of the two is the probability of study i appearing in any sample. This ratio is essential for an unbiased estimate of the mean and clearly depends on simple random sampling from a population.

The overall message is this. The way the studies were sampled determined the particular thought experiment undertaken. Had a different form of probability sampling been used, the thought experiment would have been different as well. But without some form of probability sampling, it is impossible to consider full the set of samples that could be drawn, let alone the probabilities associated with each. And without that information, it would be impossible to determine if the summary statistic computed for a set of studies is unbiased. Indeed, it is not even clear if the concept of bias is defined.

These conclusions may be generalized in three ways. First, the population size can in principle be limitless. However, there is no such thing as a limitless set of studies. No more real is a population of "all possible studies" or a population of studies "like" the studies on hand. Such formulations, while providing an apparent population from which a set of studies could be a probability sample, on closer inspection are not within the purview of science. Generalizations are being made to entities that are entirely fictional or are not clearly defined. It is hard to imagine how science is advanced in this manner.[10]

Second, one does not have to be limited to simple random sampling. All probability sampling is in play, although representing the probabilities associated with each sample can sometimes be very difficult. For purposes of this paper, there is no need to consider these complications.

[10] For readers who find the idea of "all possible studies" seductive, consider the following. All possible when: this year, over the next 3 years, over the next decade, over the next next century? All possible by whom: current meta-analysis practitioners, all such practitioners and their students, applied statisticians familiar with multilevel models, social scientists winning grants and contracts to do meta-analysis, or all individuals who read the Lispey and Wilson textbook? All possible with respect to what: collections of randomized experiments, randomized experiments and strong quasi-experiments, any study with a comparison group, any study at all? The phrase "all possible studies" is vacuous. The same sort of exercise for all studies "like" the studies on hand would be no more compelling.

Third, the notion that a set of real studies to be summarized is literally a random sample from a well-defined population is typically contradicted by the facts. As an alternative rationale for the sampling distribution, some favor model-based sampling (Thompson, 2002: section 2.7).[11] For model-based sampling, there is no population from which a random sample is drawn. There is, therefore, no sampling distribution generated from all possible random samples of the population. There are natural processes, captured in a model, capable of generating data. In meta-analysis, these are the processes by which the relevant scientific or policy communities function, and the data are studies. In this discussion, the studies would need to have the same properties as a simple random sample from a population of studies, consistent with equation 8. Then, the sampling distribution is a derived from standardized treatment effects across all possible studies that could be produced by the forces represented in the model. Inferences are made back to features of the model, not to a population from which the studies were drawn.

Crafting such a model would be a challenge. The model would have to include features that explained any dependence between studies. Are studies funded by the same agencies and published in the same journal more alike than studies funded by different agencies and published in different journals? Are later studies affected by earlier studies? If so, how? It cannot be overemphasized that to rely on model-based sampling is to make strong assertions about how science actually works. Ideally, there already exists strong evidence for that theory. Alternatively, the theory at least must be testable.

5.1 Standard Errors

Conventional confidence intervals and statistical tests depend on appropriate estimates of standard errors. A standard error is the standard deviation of the sampling distribution for a sample statistic. If there is no sampling distribution, there can be no standard deviation of that sampling distribution.

For Table 2, the standard error for the weighted mean is .77. It is the standard deviation of the 10 sample means. The standard error of .77 indicates that the mean from a simple random sample of three studies from this

[11] A simple random sample is sometimes said to be a special case of "design-based sampling."

population will fall on the average about .77 standard deviation units from the population mean of 2.0.

In practice, the standard error must be estimated from a single sample. Estimates of the standard error for each sample in Table 2 vary considerably around .77 because each sample size is so small, but whatever the value computed, multiplying it by 1.96 and alternatively adding it to and subtracting it from the estimated mean produces a 95% confidence interval. For hypothesis tests, the estimated standard error is used in the denominator when a t-statistic is computed.

With a sample of only three observations, one might not take a confidence interval or an hypothesis test very seriously. But that is not the point here. The point is that unless there is a sampling distribution derived appropriately from how the studies were produced, there can be no standard error; there is no good answer to the question, standard deviation of what? And without standard errors, there can be no confidence intervals or hypothesis tests.

6 Conclusions

Tests and confidence intervals depend on a credible sampling distribution. In meta-analysis, a credible sampling distribution depends an accurate representation of the way the studies being summarized were produced. Figure 1 shows these relationships. The last link in the chain is relatively straightforward; it is mostly a matter of following the right recipes. The first link is where serious problems commonly arise. This is usually where meta-analysis stumbles.

It is usually difficult to square how treatment effects are defined with any credible account of how an intervention affects a response. It may be telling that researchers rarely bother to even try. No more convincing are the typical assumptions about how the studies to be summarized were generated. In particular, any reliance on independence between studies is a very long stretch. As David Freedman and I have observed elsewhere, (Berk and Freedman, 2003),

> "Investigators are trained in similar ways, read the same papers, talk to one another, write proposals for funding to the same agencies, and publish the findings after peer review. Earlier studies beget later studies, just as each generation of Ph.D. students trains the next. After the first few million dollars are committed,

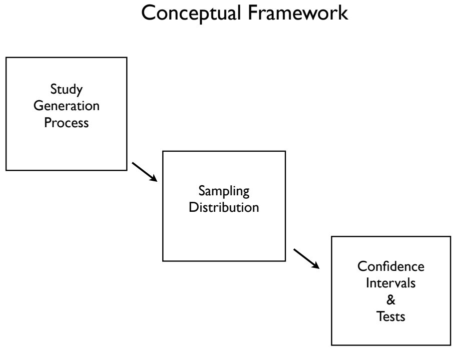

Figure 1: The Framework for Statistical Inference

> granting agencies develop agendas of their own, which investigators learn to accommodate. Meta-analytic summaries of past work further channel the effort. There is, in short, a web of social dependence inherent in all scientific research."

If the meta analysis model does not accurately represent how the studies were actually generated, a credible sampling distribution cannot be constructed. Indeed, the very idea of a sampling distribution may not apply. The problem is not with the meta-analysis model itself. The problem is that the model has little or no correspondence to how the set of studies were actually done. And If the model does not apply, the statistical inference that follows is not likely to help.

There appear to be three common responses to the mismatch between a meta-analysis model and anything real. In some cases, the problems are not mentioned. Even if recognized as the meta-analysis was done, they go missing when the results are written. In other cases, the requisite assumptions are listed, but not defended. A list of the assumptions by itself apparently inoculates the meta-analysis against modeling errors. In yet other cases, the modeling assumptions are specified and discussed, but the account is not convincing. In perhaps the most obvious illustration, the population to which generalizations are being made is "all possible studies."

It is also possible to dismiss the issues raised in this paper as statistical purity that should not be taken literally in real research. But that position is to fundamentally misunderstand the message. This paper is not a call for perfection. It is a call for rigor. At some point, moreover, the correspondence between what the formal mathematics require and the way the studies were generated is so out of kilter that the mathematical results do not usefully apply. The conclusions reported are just plain wrong.

How should researchers proceed? First, there is always the option of a conventional literature review. These are certainly not flawless, but they have served science well for a very long time. Also, readers do not have to cut through pages of statistical razzle-dazzle to understand what is really being said. Second, a meta-analysis can stop short of statistical inference. Good description alone can make a contribution. Third, one can reconsider the uncertainty that statistical inference is supposed to address, and seek alternative approaches. For example, if a concern is the stability of the results had the set of studies summarized been a bit different, a "drop-one" analysis can be helpful. If there are, for instance, ten studies, ten meta-analyses can be done using nine of the studies each time. Each study is dropped in turn. If the effect sizes vary dramatically over the ten meta-analyses, there are ample grounds for caution.[12] This and some other possibilities are considered by Greenhouse and Iyenger (1994), but there is a lot more that could be done. That discussion will be saved for another time.

References

Archer J. (2000) "Sex Differences in Aggression Between Heterosexual Partners: A Meta-analytic Review," *Psychological Bulletin* 126 (5): 651–680.

Barnett, V. (1999) *Comparative Statistical Inference.* Third Edition. New York: John Wiley and Sons.

Berk, R.A. (2003) *Regression Anaysis: A Constructive Critique.* Newbury Park, Sage Publications

Berk, R.A., and D.A. Freedman (2003) "Statistical Assumptions as Empirical Commitments" In T.G. Blomberg and S. Cohen (eds.), *Punishment and Social Control: Essays in Honor of Sheldon Messinger.*, second edition: 235-254. New York: Aldine de Gruyter.

Briggs, D.C., (2005) "Meta-Analysis: A Case Study." *Evaluation Review* 29: 87-127.

[12] This is the start of a resampling procedure called the "jackknife."

Cohen J. (1998) *Statistical Power Analysis for the Behavioral Sciences.* Second Edition. Hillsdale, NJ: Lawrence Erlbaum.

Fleiss, J. L., Levin, B., and M.C. Paik (2003) *Statistical Methods for Rates and Proportions.* New York: John Wiley and Sons.

Freedman, D.A. (2004) "Graphical Models for Causation and the Identification Problem." *Evaluation Review* 28: 267-293.

Freedman, D.A. (2005) *Statistical Models: Theory and Practice.* Cambridge: Cambridge University Press.

Garrett, C.J. (1985) " The Effects of Residential Treatment on Adjudicated Delinquents: A Meta-Analysis," *Journal of Research on Crime and Delinquency*, 45: 287-308.

Gleser, L.J. and I. Olkin (1994) "Statistically Dependent Effects," in H. Cooper and L.V. Hedges (eds.) *The Handbook of Research Synthesis.* New York: Russell Sage Foundation.

Greenhouse, J.B and S. Iyengar (1994) "Sensitivity Analysis and Diagnostics," in H. Cooper and L.V. Hedges (eds.) *The Handbook of Research Synthesis.* New York: Russell Sage Foundation.

Hedges L. V. and I. Olkin (1985) *Statistical Methods for Meta-Analysis.* New York: Academic Press.

Hedges, L.V. (1994) "Fixed Effects Models," in H. Cooper and L.V. Hedges (eds.) *The Handbook of Research Synthesis.* New York: Russell Sage Foundation.

Holland, P.W. (1986) "Statistics and Causal Inference." *Journal of the American Statistical Association* 81: 945-960.

Lewis, T.A. and D. Zelterman (1994) "Bayesian Approaches to Research Synthesis," in H. Cooper and L.V. Hedges (eds.) *The Handbook of Research Synthesis.* New York: Russell Sage Foundation.

Lipsey M. W. and D. Wilson (2001) *Practical Meta-Analysis.* Newbury Park, CA: Sage Publications.

Lipsey M. W. (1997) “What can You Build with Thousands of Bricks? Musings on the Cumulation of Knowledge in Program Evaluation,” *New Directions for Evaluation*, 76 (Winter): 7–24.

Lipsey M. W. (1992) “Juvenile Delinquency Treatment: A Meta-Analysis Inquiry into the Variability of Effects,” in T. C. Cook, D. S. Cooper, H. Hartmann, L. V. Hedges, R. I. Light, T. A. Loomis and F. M. Mosteller (eds.), *Meta-Analysis for Explanation*. New York: Russell Sage: 83–127.

Mitchell, O. (2005) “A Meta-Analysis of Race and Sentencing Research: Explaining the Inconsistencies.” *Journal of Quantitative Criminology* 21(4): 439-466.

Petitti D. B. (1999) *Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis*, 2nd ed. New York: Oxford University Press.

Raudenbush, S.W. (1994) “Random Effects Models,” in H. Cooper and L.V. Hedges (eds.) *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Rosenbaum, P.R. (2002) *Observational Studies*,second edition. New York: Springer-Verlag.

Rubin, D. B. (1986) “Which Ifs Have Causal Answers.” *Journal of the American Statistical Association* 81: 961-962.

Shadish, W.R. and C.K. Haddock (1994) ”Combining Estimates of Effect Size,” in H. Cooper and L.V. Hedges (eds.) *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Thompson, S. (2002) *Sampling*, second edition. New York: John Wiley and Sons.

Wachter, K.W. (1988) “Disturbed about Meta-Analysis?” *Science* 241: 1407-1408.