Donald Green // June 20, 2018

Ten things guide: Meta-Analysis

1. What is meta-analysis?

Meta-analysis is a method for summarizing the statistical findings from a research literature. For example, if five experiments have been conducted using the same intervention and outcome measure, one might imagine pooling these five studies together into a single dataset and analyzing them jointly, as though the studies came from five blocks within a single experiment. In essence, a meta-analysis renders a weighted average of the five studies’ results. As explained below, this method is also used to summarize research literatures that comprise a diverse array of interventions and outcomes, under the assumption that the interventions are theoretically similar and the outcome measures tap into a shared underlying trait.

1. How does a meta-analysis differ from a literature review?

Meta-analysis is often characterized as a form of systematic review insofar as it involves a specific set of data collection and analysis procedures. These procedures are spelled out in great detail here [Campbell Collaboration] and here [Lind Library]. Particular attention is paid to gathering both published and unpublished studies (see below). By comparison, most conventional literature reviews cite the most noteworthy theoretical or empirical contributions but rarely attempt to be comprehensive or to summarize the findings quantitatively. Critics of conventional reviews point out the possibility that the most noteworthy or memorable studies present findings that are unrepresentative of the broader research literature and therefore may be a poor guide for policy. On the other hand, critics of meta-analysis point out that the flaws of individual studies are often lost sight of when their estimates are blended together to generate an overarching conclusion. As Uri Simonsohn once quipped, “Meta-analysis is a sausage factory that uses sausages by other factories as inputs.”

1. What types of data are used as input for a meta-analysis?

In principle, meta-analysis could be applied to the original data from each study, but in practice such data are seldom available for all relevant studies. Instead, researchers typically cull estimated treatment effects from research papers. This process presents scholars conducting a meta-analysis with an array of decisions when the research papers present results involving multiple outcomes, treatments, and estimation approaches. Often meta-analysis focuses on the “main” results, but identifying the main results can be a judgment call. The reproducibility of meta-analysis hinges on careful documentation of such decisions. In addition to locating the key estimates, the meta-analyst must also track down measures of statistical uncertainty (e.g., standard errors, confidence intervals), as these statistics will be used to assign weights to each of the study. As noted below, meta-analysis tends to assign more weight to studies with less sampling variability.

1. What is the estimand in meta-analysis?

Some meta-analyses are narrowly tailored to specific treatments and outcomes. For example, a vast literature dating back to the 1920s focuses on the extent to which mailings that encourage voting in fact cause people to vote. In this case, one could imagine a population parameter that represents the average causal effect of mailed voting encouragements in a specified region over some specified time period.

Other meta-analyses are more abstract, focusing on a broad class of treatments and outcomes. For example, the vast literature on prejudice reduction comprises hundreds of studies on the effects of interpersonal contact between people with different racial, ethnic, religious, age, or gender backgrounds (Pettigrew and Tropp 2006). Contact ranges from a brief conversation to a year of co-habitation in a college dormitory. Outcomes also range widely from overt behaviors to self-reported feelings about in-groups and out-groups. Since these interventions and outcomes are on different scales, the underlying population parameter is ambiguous. Researchers try to sidestep this issue by standardizing the outcomes (e.g., by dividing the outcome by the standard deviation in the control group), but there remains the problem of what to make of treatments that vary in intensity and duration. In effect, the population parameter becomes the average extent to which an ad hoc collection of interventions changes putative measures of prejudice within some location and time period. Notice that this parameter reflects the particular characteristics of the research literature – for example, a literature dominated by laboratory experiments conducted in the United States will implicitly assess an underlying average treatment effect in which the “average” gives disproportionate weight to American undergraduates.

1. Meta-analysis and Bayes’ Rule

Consider the simple case in which two experiments are conducted using the same treatments and outcomes. Imagine that both studies draw their subject pool from the same population. In this case, if we had the individual data for the experiments, we could pool them together into a single dataset and analyze them as though they were part of the same block-randomized experiment. But suppose that we did not have the individual data; instead, we only have the estimated ATE and estimated standard error from each study. How might we combine the two studies to form our best guess of the average treatment effect in the population from which the subjects were drawn? With some simplifying assumptions, we could apply Bayes’ Rule. Let’s assume that the sampling distribution of each experimental estimate is normal. (This is a reasonable assumption under the Central Limit Theorem, assuming that each experiment has at least a few dozen subjects and if the outcome distribution is not too skewed.) Since these experiments are independent of one another, Bayes’ Rule takes a simple form: take a weighted average of the two estimates, where the weights are the inverse of each study’s squared standard error.

This formula turns out to be the same as a so-called “fixed effects” meta-analysis. This formula is sometimes called a “precision-weighted average,” where the term “precision” refers to the inverse of the squared standard error.

1. Fixed effects versus random effects estimation

Most meta-analysis software presents users with a choice between fixed effects estimation and random effects estimation. Fixed effects estimation is simply a precision-weighted average. The studies with the smallest standard errors are accorded the most weight. Random effects estimation applies a different set of weights depending on the extent to which the estimates vary more than would be expected by chance under a fixed effects model. Therefore, the weights for the random effects estimation not only consider the variance within each study but also an estimate of the between-study variance ().

The more heterogeneous the estimated effects – perhaps due to variations in experimental techniques, outcome measurement, or context – the more the resulting weighted average represents a simple average rather than a precision-weighted average. Typically, researchers start with a fixed effects meta-analysis, test whether the estimates are significantly overdispersed given the fixed effects model, and, if so, estimate and report a random effects meta-analysis.

1. Is it okay to summarize both experimental and observational research findings?

Beware of meta-analyses that combine experimental and observational estimates. When properly executed, experiments provide unbiased estimates of the average treatment effect. An observational study, on the other hand, is prone to bias insofar as the treatments are not randomly assigned. The nominal standard errors associated with observational studies ignore the potential for bias; the standard errors are biased downward because they assume the best-case scenario, namely, that nature assigned treatments in a manner that was as good as random. Gerber, Green, and Kaplan (2004) show that merely being uncertain about the bias of an observational study is equivalent to according it a larger standard error. Although many prominent meta-analyses include both experiments and observational studies (e.g., Lau and Sigelman CITE; Pettigrew and Tropp 2006), this practice is frowned upon by leading scholars conducting biomedical meta-analyses [CITE TO COCHRANE COLLABORATION ADMONITION ABOUT THIS]

1. Publication bias as a threat to meta-analysis

Because meta-analyses draw their data from reported results, publication bias presents a serious threat. If the only results that see the light of day are splashy or statistically significant, meta-analysis may simply amplify publication bias. Methodological guidance to meta-analytic researchers therefore places special emphasis on conducting and carefully documenting a broad-ranging search for relevant studies, whether published or not, including languages other than English. This task is, in principle, aided by pre-registration of studies in public archives; unfortunately, pre-registration in the social sciences is not sufficiently comprehensive to make this a dependable approach on its own.

When assembling a meta-analysis, it is often impossible to know whether one has missed relevant studies. Some statistical methods have been developed in order to detect publication bias, but these tests tend to have low power and therefore may give more reassurance than is warranted. For example, one common approach is to construct a scatterplot to assess the relationship between study size (whether measured by the N of subjects or the standard error of the estimated treatment effect) and effect size. A telltale symptom of publication bias is a tendency for smaller studies to produce larger effects (as would be the case if studies were published only if they showed statistically significant results; to reach the significance bar, small studies (with large standard errors) would need to generate larger effect estimates. Unfortunately, this test often produces ambiguous results (CITE), and methods to correct publication bias in the wake of such diagnostic tests (e.g., the trim-and-fill method) may do little to reduce bias. Given growing criticism of statistical tests for publication bias and accompanying statistical correctives, there is an increasing sense that meta-analysis hinges on whether research reports in a given domain can be assembled in a comprehensive manner.

1. Modeling inter-study heterogeneity using meta-regression

Researchers often seek to investigate *systematic* sources of treatment effect heterogeneity. These systematic sources may reflect differences among subjects (Do certain drugs work especially well for men or women?), contexts (Do lab studies of exposure to mass media produce stronger effects than field studies?), outcomes (Are treatment effects especially large when outcomes are measured via opinion surveys as opposed to direct observation of behavior?), or treatments (Are partisan messages more effective at mobilizing voters than nonpartisan messages?). Quite often, these investigations are best studied directly, via an experimental design. For example, variation in treatment may be studies by randomly assigned different treatment arms. Variation in effects associated with different outcome measures may also be studies in the context of a given experiment by gathering data on more than one outcome or by randomly assigning how outcomes are measured.

A second-best approach is to compare studies that differ on one or more of these dimensions (subjects, treatments, context, or outcomes). The drawback of this approach is that it is essentially descriptive rather than causal – the researcher is basically characterizing the features of studies that contribute to especially large or small effect sizes. That said, this exercise is conducted via meta-regression: the estimated effect size is the dependent variable, while study attributes (e.g., whether outcomes were measured through direct observation or via survey self-reports) constitute the independent variables. Note that meta-regression is a generalization of random effects meta-analysis, with measured predictors of effect sizes as well as unmeasured sources of heterogeneity.

1. Methods for assessing the accuracy of meta-analytic results

A skeptic might ask whether meta-analysis improves our understanding of cause-and-effect. Do we learn anything from pooling existing studies via a weighted average? One way to assess the value-added of meta-analysis is to conduct a leave-one-out analysis: how much more accurately can one forecast the left-out study by knowing the results of the other studies? Recently, this type of analysis was conducted in real time in the context of Metaketa #1. [MATT, CAN YOU INSERT A DESCRIPTION OF THE D.C. ROLLOUT OF THE RESULTS?]