Simpson's Paradox; Meta-Analysis

Psychology (Statistics) 484

Statistics, Ethics, and the Social and Behavioral Sciences

June 18, 2013

Beginning Quotations

Old Statisticians never die—they just get broken down by age and sex.
– Anon.

Alice laughed: 'There's no use trying,' she said; 'one can't believe impossible things.' 'I daresay you haven't had much practice,' said the Queen. 'When I was younger, I always did it for half an hour a day. Why, sometimes I've believed as many as six impossible things before breakfast.'
– Lewis Carroll, *Alice in Wonderland* (1865)

When the Okies left Oklahoma for California, the average intelligence was improved in two states.
– Will Rogers

Week 11: Simpson's Paradox; Meta-Analysis

— the ubiquity of Simpson's Paradox in the (mis)interpretation of data; when a relationship that appears to be present at an aggregated level disappears or reverses when disaggregated and viewed within levels

— meta-analysis and the controversies it engenders in childhood sexual abuse and other medically relevant research summarizations

Required Reading:
SGEP (333-357) —

Popular Article —
Meta-Analysis at 25; Gene V. Glass, January 2000

Suggested Reading:
Suggested Reading on Simpson's Paradox
Suggested Reading on Meta-analysis

Film: *Sacco and Vanzetti* (82 minutes)

Introduction: Simpson's Paradox

An unusual phenomenon occurs so frequently in the analysis of multiway contingency tables that it has been given the label of "Simpson's Paradox".

Basically, various relations that appear to be present when data are conditioned on the levels of one variable, either disappear or change "direction" when aggregation occurs over the levels of the conditioning variable.

A well-known real-life example is the Berkeley sex bias case applicable to graduate school (Bickel, Hammel, & O'Connell, 1975).

The table below shows the aggregate admission figures for the fall of 1973:

	Number of applicants	Percent admitted
Men	8442	44
Women	4321	35

Given these data, there appears to be a *primae facie* case for bias because a lower percentage of women than men is admitted.

Although a bias seems to be present against women at the aggregate level, the situation becomes less clear when the data are broken down by major.

No department is significantly biased against women;

in fact, most have a small bias against men.

Apparently, women tend to apply to competitive departments with lower rates of admission among qualified applicants (for example, English);

men tend to apply to departments with generally higher rates of admission (for example, Engineering).

A question exists as to whether an argument for bias "fall apart" because of Simpson's paradox?

Interesting, in many cases like this, there is a variable that if interpreted in a slightly different way would make a case for bias even at the disaggregated level.

Here, why do the differential admission quotas interact with sex?

In other words, is it inherently discriminatory to women if the majors to which they apply most heavily are also those with the most limiting admission quotas?

Death Penalty Example

A different example showing a similar point can be given using data on the differential imposition of a death sentence depending on the race of the defendant and the victim.

These data are from twenty Florida counties during 1976-1977 (Radelet, 1981):

	Death Penalty	
Defendant	Yes	No
White	19 (12%)	141
Black	17 (10%)	149

Because 12% of white defendants receive the Death penalty and only 10% of blacks, at this aggregate level there appears to be no bias against blacks.

But when the data are disaggregated, the situation appears to change:

Victim	Defendant	Death Penalty Yes	Death Penalty No
White	White	19 (13%)	132
White	Black	11 (17%)	52
Black	White	0 (0%)	9
Black	Black	6 (6%)	97

When aggregated over victim race, there is a higher percentage of white defendants (12%) receiving the death penalty than black defendants (10%), so apparently, there is a slight race bias against whites.

But when looking within the race of the victim, black defendants have the higher percentages of receiving the death sentence compared to white defendants (17% to 13% for white victims; 6% to 0% for black victims).

The conclusion is disconcerting: the value of a victim is worth more if white than if black, and because more whites kill whites, there appears to be a slight bias against whites at the aggregate level.

But for both types of victims, blacks are more likely to receive the death penalty.

Simpson's Paradox is a very common occurrence, and even through it can be "explained away" by the influence of differential marginal frequencies, the question remains as to why the differential marginal frequencies are present in the first place.

Generally, a case can be made that gives an argument for bias or discrimination in an alternative framework, for example, differential admission quotas or differing values on a life.

Although not explicitly a Simpson's Paradox context, there are similar situations that appear in various forms of multifactor analysis of variance that raise cautions about aggregation phenomena.

The simplest dictum is that "you cannot interpret main effects in the presence of interaction."

This admonition is usually softened when the interaction is not disordinal, and the graphs of means don't actually cross.

In these instances it may be possible to eliminate the interaction by some relatively simple transformation of the data, and produce an "additive" model.

Because of this, noncrossing interactions might be considered "unimportant."

Similarly, the absence of parallel profiles (that is, when interaction is present) may hinder the other tests for the main effects of coincident and horizontal profiles.

Reversal Paradoxes

Simpson's Paradox is part of a larger class of reversal paradoxes (Messick & van de Geer, 1981).

(1) Based on the algebraic constraints for correlations given earlier, suppose that performance on each of two developmental tasks has a positive correlation with age;

an observed positive correlation between the two tasks could conceivably reverse when a partial correlation is computed between the two tasks that "holds age constant."

(2) Illusory (positive) correlations that result from a "lurking" or confounding variable might be reversed when that variable is controlled.

One particularly memorable example given below is from Messick and van de Geer (1981):

A well-known example is the apparent paradox that the larger the number of firemen involved in extinguishing a fire, the larger the damage. Here the crucial third variable, of course, is the "severity of the fire"; for fires of equal severity, one would hope that the correlation would have a reversed sign.

Weighted Average Explanation

A common way to explain what occurs in Simpson's Paradox is to use contingency tables.

For convenience, we restrict discussion to the simple $2 \times 2 \times 2$ case, and use the "death penalty" data as an illustration.

There are two general approaches based on conditional probabilities. One involves weighted averages; the second relies on the language of events being conditionally positively correlated, but unconditionally negatively correlated (or the reverse).

We only do the weighted average explanation here. See the required readings for the other approach.

To set up the numerical example, define three events: A, B, and C:

A: the death penalty is imposed;
B: the defendant is black;
C: the victim is white.

For reference later, we give a collection of conditional probabilities based on frequencies in the $2 \times 2 \times 2$ contingency table:

$P(A|B) = .10$; $P(A|\bar{B}) = .12$; $P(A|B \cap C) = .17$;

$P(A|\bar{B} \cap C) = .13$; $P(A|\bar{B} \cap \bar{C}) = .00$;
$P(C|B) = .38$; $P(\bar{C}|B) = .62$; $P(C|\bar{B}) = .94$;
$P(\bar{C}|\bar{B}) = .38$; $P(C) = .66$; $P(\bar{C}) = .34$.

The explanation for Simpson's Paradox based on a weighted average begins by formally stating the paradox through conditional probabilities: It is possible to have

$$P(A|B) < P(A|\bar{B}) \ ,$$

but

$$P(A|B \cap C) \geq P(A|\bar{B} \cap C) \ ;$$

$$P(A|B \cap \bar{C}) \geq P(A|\bar{B} \cap \bar{C}) \ .$$

So, conditioning on the C and $\bar{C}$ events, the relation reverses. In labeling this reversal as anomalous, people reason that the conditional probability, $P(A|B)$, should be an average of

$$P(A|B \cap C) \text{ and } P(A|B \cap \bar{C}) \ ,$$

and similarly, that $P(A|\bar{B})$ should be an average of

$$P(A|\bar{B} \cap C) \text{ and } P(A|\bar{B} \cap \bar{C}) \ .$$

Although this is true, it is not a simple average but one that is weighted:

$$P(A|B) = P(C|B)P(A|B \cap C) + P(\bar{C}|B)P(A|B \cap \bar{C}) \ ;$$

$$P(A|\bar{B}) = P(C|\bar{B})P(A|\bar{B} \cap C) + P(\bar{C}|\bar{B})P(A|\bar{B} \cap \bar{C}) \ .$$

If B and C are independent, $P(C|B) = P(C|\bar{B}) = P(C)$ and $P(\bar{C}|B) = P(\bar{C}|\bar{B}) = P(\bar{C})$. Also, under such independence, $P(C)$ and $P(\bar{C})$ $(= 1 - P(C))$ would be the weights for constructing the average, and no reversal would occur.

If B and C are not independent, however, a reversal can happen, as it does for our "death penalty" example:

$.10 = P(A|B) = (.38)(.17) + (.62)(.06);$
$.12 = P(A|\bar{B}) = (.94)(.13) + (.62)(.00).$

So, instead of the weights of .66 $(= P(C))$ and .34 $(= P(\bar{C}))$, we use .38 $(= P(C|B))$ and .62 $(= P(\bar{C}|B))$; and .94 $(= P(C|\bar{B}))$ and .06 $(= P(\bar{C}|\bar{B}))$.

Figure 12.1 Interpretation of Simpson's Paradox

Figure 12.1 (in the required reading) provides a convenient graphical representation for the reversal paradox in our "death penalty" illustration.

This representation generalizes to any $2 \times 2 \times 2$ contingency table.

The x-axis is labeled as percentage of victims who are white; the y-axis has a label indicating the probability of death penalty imposition.

This probability generally increases along with the percentage of victims that are white.

Two separate lines are given in the graph reflecting this increase, one for Black defendants and one for white defendants.

Note that the line for the black defendant lies wholly above that for the white defendant, implying that irrespective of the percentage of victims that may be white, the imposition of the death penalty has a greater probability for a black defendant compared to a white defendant.

Although Simpson's Paradox has been known by this name only rather recently (as coined by Colin Blyth in 1972), the phenomenon has been recognized and discussed for well over a hundred years;

in fact, it has a complete textbook development in Yule's *An Introduction to the Theory of Statistics*, first published in 1911.

We give the section of Yule's text in the endnotes of the required reading that discusses Simpson's Paradox, but obviously without the name.

Introduction: Meta Analysis

In his Presidential Address to the American Educational Research Association, Glass (1976) introduced the idea of "meta-analysis," referring to a statistical integration of a set of studies that ask a common research question.

The title of his address, "Primary, Secondary, and Meta-Analysis of Research," distinguishes three types of data analysis.

A *primary analysis* is the initial data analysis for an original research study.

A *secondary analysis* is a reexamination of an existing dataset, possibly with different statistical and/or interpretative tools than originally used or available.

Mosteller and Moynihan's (1972) reanalysis of the Coleman (1966) data on equality of educational opportunity is a famous example of a secondary data analysis.

Finally, a *meta-analysis* combines the analyses (both primary and secondary) for a number of studies into a coherent statistical review.

This is in contrast to the more usual discursive literature review that was common up to the time of Glass's address.

The three decades that followed Glass's introduction of meta-analysis has seen an explosion of such studies published in journals in the behavioral and social sciences, education, health, and medicine.

In the behavioral sciences, meta-analyses appear regularly in the field's premier journals (e.g., *Psychological Bulletin*); for medical- and health-related topics, we now have the extensive internationally organized Cochrane Collaboration, founded in 1993.

The handbook produced by this latter consortium, the *Cochrane Handbook for Systematic Reviews of Interventions* (Higgins & Green, 2008), although designed for researchers in medicine, is also useful in other areas.

The Campbell Collaboration, an organization similar to the Cochrane Collaboration, was founded in 1999 and dedicated to Donald Campbell. It is devoted to systematic reviews of interventions in the social, behavioral, and educational areas.

Many of these reviews are now published electronically in the free online journal, *Campbell Systematic Reviews*

A new journal, *Research Synthesis Methods*, sponsored by the Society for Research Synthesis Methodology, was introduced by Wiley in 2010. This outlet has the interdisciplinary goal of following work on all facets of research synthesis of the type represented by both the Cochrane and Campbell Collaborations.

A meta-analysis involves the use of a common measure of "effect size" that is then aggregated over studies to give an "average" effect.

The heterogeneity in the individual effects can be used to generate confidence intervals for an assumed population effect parameter, or related to various study characteristics to assess why effect sizes might systematically vary apart from the inherent random variation present within any single study.

The common effect measures for dichotomous outcomes are odds ratios, risk ratios, or risk differences.

For continuous data, there are Pearson correlation coefficients or Cohen effect sizes defined by between-group mean differences divided by within-group standard deviations.

Psychological Bulletin Article Controversy

The use of meta-analysis has been involved in several publicly controversial cases in the recent past.

Probably the most sensationalized was a meta-analysis published in *Psychological Bulletin* (1998, *124*, 22–53) by Rind, Tromovitch, and Bauserman, entitled "A Meta-Analytic Examination of Assumed Properties of Child Sexual Abuse Using College Samples."

As might have been expected, this meta-analysis caused quite an uproar, including a unanimous condemnation resolution from Congress.

In the required readings we give an OpEd item from the *Los Angeles Times* by Carol Tavris ("The Politics of Sex Abuse," July 19, 1999). The first paragraph follows:

I guess I should be reassured to know that Congress disapproves of pedophilia and the sexual abuse of children. On July 12, the House voted unanimously to denounce a study that the resolution's sponsor, Matt Salmon (R-Ariz.), called "the emancipation proclamation of pedophiles." In a stunning display of scientific illiteracy and moral posturing, Congress misunderstood the message, so they condemned the messenger.

Study Inclusion/Exclusion Criteria

One of the first demands in carrying out any meta-analysis is a preliminary selection of the studies to be included, and in turn, who exactly are the individuals to be studied.

Usually, this is implemented by an explicit set of exclusionary and inclusionary criteria.

Given this selection, it is then obviously important to temper one's overall conclusion to what studies were actually integrated.

Some difficulty with carrying out this admonition occurs whenever the cable news networks demand fodder to fill their airtime.

Witness the recent controversy about the effectiveness of antidepressants as judged by a particular meta-analysis published in the *Journal of the American Medical Association* (Fournier, et al., 2010).

To ignore the acute or chronic nature of the depression for those individuals included in the studies evaluated, and more generally, to ignore who the subjects actually are in a statistical integration, can turn meta-analysis into an unethically motivated strategy of data analysis and persuasion.

Publication Bias

In conducting a meta-analysis, the goal is to include all relevant studies that pass the set of inclusionary criteria, whatever they may be.

This can involve finding or considering studies that may, for whatever reason, remain unpublished.

Folk wisdom and/or experience tells us that negative results and those that are just "nonsignificant" might not be publishable because of suppression, say, by Big Pharma (or Big Tobacco), or the inherent reluctance of editors to use valuable pages to publish noninteresting (that is, statistically nonsignificant) results.

The tendency to see a larger than what might be expected proportion of significant results in the published literature is *prima facie* evidence for a publication bias toward statistically significant results.

Funnel Plots

Both the file-drawer problem (characterized by negative or nonsignificant results being tucked away in a cabinet), and bias resulting from explicit decisions not to place studies in the public domain are problematic.

If present, these tendencies would falsely inflate the effect estimates; thus, various ways of detection (and hopefully, correction) have been proposed.

One relatively simple diagnostic is labeled a "funnel plot," where effect magnitudes (on the horizontal axis) are plotted against some measure of sample size (or an estimate of precision that depends on sample size) on the vertical axis.

If sample size is unrelated to the true magnitude of effect, as it should be, and there is no publication bias, the plot should resemble a funnel where estimates associated with the smaller sample sizes spread out over a wider range at the bottom of the funnel.

The degree to which symmetry is not present around some vertical line through the funnel may indicate that publication bias has occurred (that is, for the smaller sample sizes, there is an asymmetry in that there are more published results than might be expected).

Again, one can be luckier for small samples in getting a spuriously significant and larger effect estimate than for larger sample sizes (this is sometimes discussed under the rubric of "small-study effects").

Gray Literature

The type of unpublished literature alluded to here has led to another issue in meta-analysis and elsewhere, in how to deal generally with "gray literature," or material that has not been subjected to the usual standards of peer review.

The major issue with gray literature is a lack of vetting, which can be a major problem when biased, wrong, fraudulent, and so on.

Witness the recent case of the IPCC Climate Change Report (2007) merely lifting the Himalayan glacier melt claim from an unsubstantiated news source, and then one editor's decision not to remove it because of its potential for dramatic persuasion.

The Decline Effect

The problems generated from various forms of publication bias are legion.

One recent and readable exposé appeared in the *New Yorker* by Jonah Lehrer (December 13, 2010), "The Truth Wears Off: Is There Something Wrong With the Scientific Method?"

The basic issue is that some supposed "big result," upon replication, generally declines.

In fact, this phenomenon appears to be so universal it has now been labeled the "decline effect" (by Lehrer and others).

Study Inclusion

Any meta-analysis requires a decision as to what studies to include.

Gene Glass's original idea was to be very inclusive and to "bring on all comers."

If effects varied widely, these could then be related to study characteristics (for example, randomized or not, study settings, general age and sex distribution of the subjects).

Other entities (such as the Cochrane Collaboration) emphasize the need to be very selective, and to include only randomized clinical trials.

The inclusion of nonrandomized observational studies, they might argue, just moves the bias up to a different level when subjects partially self-select into the various treatment groups.

But even when restricted to randomized clinical trials, bias can creep in by some of the mechanisms we mention later in Experimental Design and the Collection of Data.

In general, we must be ever vigilant to the effects of confirmation bias, where we decide, possibly without really knowing that we are doing so, on what the "truth" should be before we begin to amass our studies.

If the inclusionary criteria are set to show what we know should be there (or, to set the exclusionary criteria to eliminate any inconvenient nonconforming studies), the "truth" is being constructed and not discovered.

The need to be explicit about the inclusionary criteria for studies in a meta-analysis may be obvious, but some of the resulting interpretative differences can run very deep indeed.

There are problems inherent in the variation of who the subjects are in the various studies, how the treatments might vary, what type and form the measures take that are used to evaluate the treatments, and so on.

As an example of what difficulties can happen, a study reported in the *New England Journal of Medicine* some years ago raised quite a sizable kerfuffle about these issues: "Discrepancies Between Meta-Analyses and Subsequent Large, Randomized Controlled Trials" (LeLorier, Grégoire, Benhaddad, Lapierre, & Derderian; *New England Journal of Medicine*, 1997, *337*, 536–542).

The "Applies and Oranges" Criticism

One commonly heard criticism of meta-analysis is that you can't meaningfully mix "apples and oranges" together, but that's exactly what a meta-analysis tries to do.

On the face of it such a complaint may sound reasonable, but more often it is a red herring.

Think about how the practice of modern statistics proceeds.

In a multiple regression that attempts to predict some output measure from a collection of predictor variables, the latter can be on any scale whatsoever (that is, differing means and variances, for example).

If that bothers the sensibilities of the analyst, everything can be reduced to z-scores, and the results (re)interpreted in terms of all variables now being forced to be commensurable (with means of zero and variances of 1).

Typically, it is not necessary in a meta-analysis to move to this level of enforced z-score commensurability.

If we are measuring behavior change, for example, with a number of manifestly different observed measures that may vary from study to study, then as long as a strong common factor underlies the various measures (in the traditional Spearman sense), it is reasonable to normalize the measures and include them in a meta-analysis.

If the observed measures are more factorially complex, a "multivariate meta-analysis" might make sense using several normalized measures that tap the distinct domains (in the tradition of Thurstone [1935] group factors).

Individual Differences

When our cognitively oriented colleagues recruit subjects for their studies, they are usually not very picky.

Subjects are more or less interchangeable for the processes they are interested in, and all are equally representative of what is being studied.

Some of our other colleagues with more of an individual differences emphasis (for example, those in the industrial/organization, social/personality, and developmental fields), are typically more concerned with variety and diversity because what is being studied is probably related to who the subjects are, such as age, education, political leanings, attitudes more generally.

And commonly, that is part of what is being studied.

Meta-analysis in its usual vanilla form is not concerned with individual differences to any great extent.

What is analyzed are averages at a group level.

But averages cannot do justice to what occurs internally within an intervention or study.

Who benefits or doesn't? A zero overall effect could result from and mask the situation where some do great, some do badly, and many just don't change at all.

Vote-Counting Meta-Analysis

One strategy suggested for reconciling possibly conflicting studies is to count the number of times that led to a rejection/acceptance of the null hypothesis (at say, the .05 level) over all available studies.

Then, according to some variation on majority rule, a conclusion of "effect" or "no effect" is made.

There are several problematic aspects to this strategy:

the sample sizes for the individual studies are a primary determinant of significance;
the actual size of an effect plays a more secondary role when it should be utmost;
even if we want to rely on p-values to make a conclusion, there are much better ways of aggregating the p-values over studies to get one such overall p-value.

For a cautionary tale about the inadvisability of vote-counting methods, we suggest the article by Hedges, Laine, and Greenwald in the *Educational Researcher* (1994, *23*(3), 5–14): "An Exchange: Part I: Does Money Matter? A Meta-Analysis of Studies of the Effects of Differential School Inputs on Student Outcomes."