Statistical Assumptions as Empirical Commitments

Richard A. Berk
David A. Freedman

Introduction

Researchers who study punishment and social control, like those who study other social phenomena, typically seek to generalize their findings from the data they have to some larger context: in statistical jargon, they generalize from a sample to a population. Generalizations are one important product of empirical inquiry. Of course, the process by which the data are selected introduces uncertainty. Indeed, any given dataset is but one of many that could have been studied. If the dataset had been different, the statistical summaries would have been different, and so would the conclusions, at least by a little.

How do we calibrate the uncertainty introduced by data collection? Nowadays, this question has become quite salient, and it is routinely answered using well-known methods of statistical inference, with standard errors, t-tests, and P-values, culminating in the "tabular asterisks" of Meehl (1978). These conventional answers, however, turn out to depend critically on certain rather restrictive assumptions, for instance, random sampling.[1] When the data are generated by random sampling from a clearly defined population, and when the goal is to estimate population parameters from sample statistics, statistical inference can be relatively straightforward. The usual textbook formulas apply; tests of statistical significance and confidence intervals follow.

If the random-sampling assumptions do not apply, or the parameters are not clearly defined, or the inferences are to a population that is only vaguely defined, the calibration of uncertainty offered by contemporary statistical technique is in turn rather questionable.[2] Thus, investigators who use conventional statistical technique

[1] "Random sampling" has a precise, technical meaning: sample units are drawn independently, and each unit in the population has an equal chance to be drawn at each stage. Drawing a random sample of the U. S. population, in this technical sense, would cost several billion dollars (since it requires a census as a preliminary matter) and would probably require the suspension of major constitutional guarantees. Random sampling is not an idea to be lightly invoked.

[2] As we shall explain below, researchers may find themselves assuming that their sample is a random sample from an imaginary population. Such a population has no empirical existence, but is defined in an essentially circular way—as that population from which the sample may be assumed to be randomly

turn out to be making, explicitly or implicitly, quite restrictive behavioral assumptions about their data collection process. By using apparently familiar arithmetic, they have made substantial empirical commitments; the research enterprise may be distorted by statistical technique, not helped. At least, that is our thesis, which we will develop in the pages that follow.

Random sampling is hardly universal in contemporary studies of punishment and social control. More typically, perhaps, the data in hand are simply the data most readily available (e.g., Gross and Mauro, 1989; MacKenzie, 1991; Nagin and Paternoster, 1993; Berk and Campbell, 1993; Phillips and Grattet, 2000; White 2000). For instance, information on the use of prison "good time" may come from one prison in a certain state. Records on police use of force may be available only for encounters in which a suspect requires medical attention. Prosecutors' charging decisions may be documented only after the resolution of a law suit.

"Convenience samples" of this sort are not random samples. Still, researchers may quite properly be worried about replicability. The generic concern is the same as for random sampling: if the study were repeated, the results would be different. What, then, can be said about the results obtained? For example, if the study of police use of force were repeated, it is almost certain that the sample statistics would change. What can be concluded, therefore, from the statistics?

These questions are natural, but may be answerable only in certain contexts. The moment that conventional statistical inferences are made from convenience samples, substantive assumptions are made about how the social world operates. Conventional statistical inferences (e.g., formulas for the standard error of the mean, t-tests, etc.) depend on the assumption of random sampling. This is not a matter of debate or opinion; it is a matter of mathematical necessity.[3] When applied to convenience samples, the random sampling assumption is not a mere technicality or a minor revision on the periphery; the assumption becomes an integral part of the theory.

In the pages ahead, we will try to show how statistical and empirical concerns interact. The basic question will be this: what kinds of social processes are assumed by the application of conventional statistical techniques to convenience samples? Our answer will be that the assumptions are quite unrealistic. If so, probability calculations that depend on the assumptions must be viewed as unrealistic too.[4]

---

drawn. At the risk of the obvious, inferences to imaginary populations are also imaginary.

[3] Of course, somewhat weaker assumptions may be sufficient for some purposes. However, as we discuss below, the outlines of the problem stay the same.

[4] We use the term "parameter" for a characteristic of the population. A "sample statistic" or "estimate" is computed from the sample to estimate the value of a parameter. As indicated above, we use "random sampling" to mean sampling with replacement from a finite population: each unit in the population is selected independently (with replacement) and with the same probability of selection. Sampling without replacement (i.e., simple random sampling) may be more familiar. In many practical situations, sampling without replacement is very close to sampling with replacement. Stratified cluster samples are often more cost-effective than purely random samples, but estimates and standard errors then need to be computed taking the sample design into account. Convenience samples are often treated as if they were random

Treating the Data as a Population

Suppose that one has data from spouse abuse victims currently residing in a particular shelter. A summary statistic of interest is the proportion of women who want to obtain restraining orders. How should potential uncertainty be considered?

One strategy is to treat the women currently residing in the shelter as a population; the issue of what would happen if the study were repeated does not arise. All the investigator cares about are the data now in hand. The summary statistics describe the women in the dataset. No statistical inference is needed since there is no sampling uncertainty to worry about.

Treating the data as a population and discarding statistical inference might well make sense if the summary statistics are used to plan for current shelter residents. A conclusion that "most" want to obtain restraining orders is one thing; a conclusion that a "few" want to obtain such orders has different implications. But there are no inferences about women who might use the shelter in the future, or women residing in other shelters. In short, the ability to generalize has been severely restricted.

Assuming a Real Population and Imaginary Sampling Mechanism

Another way to treat uncertainty is to define a real population and assume that the data can be treated as a random sample from that population. Thus, current shelter residents could perhaps be treated as a random sample drawn from the population of residents in all shelters in the area during the previous 12 months. This "as-if" strategy would seem to set the stage for statistical business as usual.

An explicit goal of the "as-if " strategy is generalizing to a specific population. And one issue is this: are the data representative? For example, did each member of the specified population have the same probability of coming into the sample? If not, and the investigator fails to weight the data, inferences from the sample to the population will likely be wrong.[5]

More subtle are the implications for estimates of standard errors.[6] The usual formulas require the investigator to believe that the women are sampled independently of one another. Even small departures from independence may have serious consequences, as we demonstrate later. Furthermore, the investigator is required to

---

samples, and sometimes as if they were stratified random samples—that is, random samples drawn within subgroups of some poorly-defined super-population. Our analysis is framed in terms of the first model, but applies equally well to the second.

[5]Weighting requires that the investigator know the probability of selection for each member of the population. It is hard to imagine that such precise knowledge will be available for convenience samples. Without reweighting, estimates will be biased, perhaps severely.

[6]The standard error measures sampling variability; it does not take bias into account. Our basic model is random sampling. In the time-honored way, suppose we draw women into the sample one after another (with replacement). The conventional formula for the standard error assumes that the selection probabilities stay the same from draw to draw; on any given draw, the selection probabilities do not have to be identical across women.

assume constant probabilities across occasions. This assumption of constant probabilities is almost certainly false. Family violence has seasonal patterns. (Christmas is a particularly bad time.) The probabilities of admission therefore vary over the course of the year. In addition, shelters vary in catchment areas, referral patterns, interpersonal networks, and admissions policies. Thus, women with children may have low probability of admission to one shelter, but a high probability of admission to other shelters. Selection probabilities depend on a host of personal characteristics; such probabilities must vary across geography and over time.

The independence assumption seems even more unrealistic. Admissions policies evolve in response to daily life in the shelter. For example, some shelter residents may insist on keeping contact with their abusers. Experience may make the staff reluctant to admit similar women in the future. Likewise, shelter staff may eventually decide to exclude victims with drug or alcohol problems.

To summarize, the random sampling assumption is required for statistical inference. But this assumption has substantive implications that are unrealistic. The consequences of failures in the assumptions will be discussed below.

An Imaginary Population and Imaginary Sampling Mechanism

Another way to treat uncertainty is to create an imaginary population from which the data are assumed to be a random sample. Consider the shelter story. The population might be taken as the set of all shelter residents that could have been produced by the social processes creating victims who seek shelter. These processes might include family violence, as well as more particular factors affecting possible victims, and external forces shaping the availability of shelter space.

With this approach, the investigator does not explicitly define a population that could in principle be studied, with unlimited resources of time and money. The investigator merely *assumes* that such a population exists in some ill-defined sense. And there is a further assumption, that the dataset being analyzed can be treated *as if* it were based on a random sample from the assumed population. These are convenient fictions. Convenience will not be denied; the source of the fiction is two-fold: (i) the population does not have any empirical existence of its own, and (ii) the sample was not in fact drawn at random.

In order to use the imaginary-population approach, it would seem necessary for investigators to demonstrate that the data can be treated as a random sample. It would be necessary to specify the social processes that are involved, how they work, and why they would produce the statistical equivalent of a random sample. Handwaving is inadequate. We doubt the case could be made for the shelter example or any similar illustration. Nevertheless, reliance on imaginary populations is widespread. Indeed, regression models are commonly used to analyze convenience samples: as we show later, such analyses are often predicated on random sampling from imaginary populations. The rhetoric of imaginary populations is seductive precisely

because it seems to free the investigator from the necessity of understanding how data were generated.

When the Statistical Issues are Substantive

Statistical calculations are often a technical side-show; the primary interest is in some substantive question. Even so, the methodological issues need careful attention, as we have argued. However, in many cases the substantive issues are very close to the statistical ones. For example, in litigation involving claims of racial discrimination, the substantive research question is usually operationalized as a statistical hypothesis: certain data are like a random sample from a specified population.

Suppose, for example, that in a certain jurisdiction there are 1084 probationers under federal supervision: 369 are black. Over a six month period, 119 probationers are cited for technical violations: 54 are black. This is disparate impact, as one sees by computing the percents: in the total pool of probationers, 34% are black; however, among those cited, 45% are black.

A t-test for "statistical significance" would probably follow. The standard error on the 45% is $\sqrt{.45 \times .55/119} = .046$, or 4.6%. So, $t = (.45 - .34)/.046 = 2.41$, and the one-sided P is .01. (A more sophisticated analyst might use the hypergeometric distribution, but that would not change the outlines of the problem.) The null hypothesis is rejected, and there are at least two competing explanations: either blacks are more prone to violate probation, or supervisors are racist. It is up to the probation office to demonstrate the former; the t-test shifts the burden of argument.

However, there is a crucial (and widely ignored) step in applying the t-test: translating the idea of a race-neutral citation process into a statistical null hypothesis. In a race-neutral world, the argument must go, the citation process would be like citing 119 people drawn at random from a pool consisting of 34% blacks. This random-sampling assumption is the critical one for computing the standard error.

In more detail, the t-statistic may be large for two reasons: (i) too many blacks are cited, so the numerator in the t-statistic is too big, or (ii) the standard error in the denominator is too small. The first explanation may be the salient one, but we think the second explanation needs to be considered as well. In a race-neutral world, it is plausible that blacks and whites should have the same overall citation probabilities. However, in any world, these probabilities seem likely to vary from person to person and time to time. Furthermore, dependence from occasion to occasion would seem to be the rule rather than the exception. As will be seen below, even fairly modest amounts of dependence can create substantial bias in estimated standard errors.

In the real world of the 1990's, the proportion of federal probationers convicted for drug offenses increased dramatically. Such probationers were often subjected to drug testing and required to participate in drug treatment programs. The mix of

offenders and supervision policies changed dramatically. The assumption of probabilities constant over time is, therefore, highly suspect. Likewise, an assumption that all probationers faced the same risks of citation is almost certainly false. Even in a race-neutral world, the intensity of supervision must be in part determined by the nature of the offender's crime and background; the intensity of supervision obviously affects the likelihood of detecting probation violations.

The assumption of independence is even more problematic. Probation officers are likely to change their supervision policies, depending on past performance of the probationers. For example, violations of probation seem likely to lead to closer and more demanding supervision, with higher probabilities of detecting future violations. Similarly, behavior of the probationers is likely to depend on the supervision policies.

In short, the translation of race neutrality into a statistical hypothesis of random sampling is not innocuous. The statistical formulation seems inconsistent with the social processes on which it has been imposed. If so, the results of the statistical manipulations—the P-values—are of questionable utility.

This example is not special. For most convenience samples, the social processes responsible for the data likely will be inconsistent with what needs to be assumed to justify conventional formulas for standard errors. If so, translating research questions into statistical hypotheses may be quite problematic: much can be lost in translation.

Does the Random Sampling Assumption Make Any Difference?

For criminal justice research, we have tried to indicate the problems with making statistical inferences based on convenience samples. The assumption of independence is critical, and we believe this assumption will always be difficult to justify (Kruskal, 1988). The next question is whether failures of the independence assumption matter. There is no definitive answer to this question; much depends on context. However, we will show that relatively modest violations of independence can lead to substantial bias in estimated standard errors. In turn, the confidence levels and significance probabilities will be biased too.

Violations of Independence

Suppose the citation process violates the independence assumption in the following manner. Probation officers make contact with probationers on a regular basis. If contact leads to a citation, the probability of a subsequent citation goes up, because the law enforcement perspective is reinforced. If contact does not lead to a citation, the probability of a subsequent citation goes down (the law enforcement perspective is not reinforced). This does not seem to be an unreasonable model; indeed, it may be far more reasonable than independence.

More specifically, suppose the citation process is a "stationary Markov chain." If contact leads to a citation, the chance that the next case will be cited is .50. On the other hand, if contact does not lead to a citation, the chance of a citation on the next contact is only 0.10. To get started, we assume the chance of a citation on the first contact is .30; the starting probability makes little difference for this demonstration.

Suppose an investigator has a sample of 100 cases, and observes 17 citations. The probability of citation would be estimated as $17/100 = .17$, with a standard error of $\sqrt{.17 \times .83/100} = .038$. Implicitly, this calculation assumes independence. However, Markov chains do not obey the independence assumption. The right standard error, computed by simulation, turns out to be .058. This is about 50% larger than the standard error computed by the usual formula. As a result, the conventional t-statistic is about 50% too large. For example, a researcher who might ordinarily use a critical value of 2.0 for statistical significance at the .05 level should really be using a critical value of about 3.0.

Alert investigators might notice the breakdown of the independence assumption: the first-order serial correlation for our Markov process is about .40. This is not large, but it is detectable with the right test. However, the dependencies could easily be more complicated and harder to find, as the next example shows.

Consider a "four-step Markov chain." The probation officer judges an offender in the light of recent experience with similar offenders. The officer thinks back over the past four cases and finds the case most like the current case. If this "reference" case was cited, the probability that the current case will be cited is .50. If the reference case was not cited, the probability that the current case will be cited is .10. In our example, the reference case is chosen at random from the four prior cases. Again, suppose an investigator has a sample of 100 cases, and observes 17 citations. The probability of citation would still be estimated as $17/100 = .17$, with a standard error of $\sqrt{.17 \times .83/100} = .038$. Now, the right standard error, computed by simulation, turns out to be .062. This is about 60% larger than the standard error computed by the usual formula.

Conclusions are much the same as for the first simulation. However, the four-step Markov chain spreads out the dependence so that it is hard to detect: the first-order serial correlation is only about .12.[7] Without a priori knowledge that the data were generated by a four-step Markov chain, a researcher is unlikely to identify the dependence.

Similar problems come about if the Markov chain produces negative serial correlations rather than positive ones. Negative dependence can be just as hard to detect, and the estimated standard errors will still be biased. Now the bias is upward so the null hypothesis is not rejected when it should be: significant findings are missed.

Of course, small correlations are easier to detect with large samples. Yet probation officers may use more than four previous cases to find a reference case; they

[7]The standard error is affected not only by first-order correlations, but also by higher-order correlations.

may draw on their whole current case load, and on salient cases from past case loads. Furthermore, transition probabilities (here, .50 and .10) are likely to vary over time in response to changing penal codes, administrative procedures, and mix of offenders. As a result of such complications, even very large samples may not save the day.

The independence assumption is fragile. It is fragile as an empirical matter because real world criminal justice processes are unlikely to produce data for which independence can be reasonably assumed. (Indeed, if independence were the rule, criminal justice researchers would have little to study.) The assumption is fragile as a statistical matter, because modest violations of independence may have major consequences while being nearly impossible to detect. The Markov chain examples are not worst case scenarios, and they show what can happen when independence breaks down. The main point: even modest violations of independence can introduce substantial biases into conventional procedures.

Dependence in Other Settings

Spatial Dependence

In the probation example, dependence was generated by social processes that unfolded over time. Dependence can also result from spatial relationships rather than temporal ones. Spatial dependence may be even more difficult to handle than temporal dependence.

For example, if a researcher is studying crime rates across census tracts in a particular city, it may seem natural to assume that the correlation between tracts depends on the distance between them. However, the right measure of distance is by no means obvious. Barriers such as freeways, parks, and industrial concentrations may break up dependence irrespective of physical distance. "Closeness" might be better defined by travel time. Perhaps tracts connected by major thoroughfares are more likely to violate the assumption of independence than tracts between which travel is inconvenient. Ethnic mix and demographic profiles matter too, since crimes tend to be committed within ethnic and income groups. Social distance rather than geographical distance may be the key. Our point is that spatial dependence matters. Its measurement will be difficult, and may depend on how distance itself is measured. Whatever measures are used, spatial dependence produces the same kinds of problems for statistical inference as temporal dependence.

Regression Models

In research on punishment and social control, investigators often use complex models. In particular, regression and its elaborations (e.g., structural equation modeling) are now standard tools of the trade. Although rarely discussed, statistical assumptions have major impacts on analytic results obtained by such methods.

Consider the usual textbook exposition of least squares regression. We have n observational units, indexed by $i = 1, \ldots, n$. There is a response variable y_i, conceptualized as $\mu_i + \epsilon_i$, where μ_i is the theoretical mean of y_i while the disturbances or errors ϵ_i represent the impact of random variation (sometimes of omitted variables). The errors are assumed to be drawn independently from a common (gaussian) distribution with mean 0 and finite variance. Generally, the error distribution is not empirically identifiable outside the model; so it cannot be studied directly—even in principle—without the model. The error distribution is an imaginary population and the errors ϵ_i are treated as if they were a random sample from this imaginary population—a research strategy whose frailty was discussed earlier.

Usually, explanatory variables are introduced and μ_i is hypothesized to be a linear combination of such variables. The assumptions about the μ_i and ϵ_i are seldom justified or even made explicit—although minor correlations in the ϵ_i can create major bias in estimated standard errors for coefficients. For one representative textbook exposition, see Weisberg (1985). Conventional econometric expositions are for all practical purposes identical (e.g., Johnston, 1984).

Structural equation models introduce further complications (Freedman, 1987, 1991, 1995, 1997, 1999; Berk, 1988, 1991). Although the models seem sophisticated, the same old problems have been swept under the carpet, because random variation is represented in the same old way. Why do μ_i and ϵ_i behave as assumed? To answer this question, investigators would have to consider, much more closely than is commonly done, the connection between social processes and statistical assumptions.

Time Series Models

Similar issues arise in time series work. Typically, the data are highly aggregated; each observation characterizes a time period rather than a case; rates and averages are frequently used. There may be T time periods indexed by $t = 1, 2, \ldots, T$. The response variable y_t is taken to be $\mu_t + \epsilon_t$ where the ϵ_t are assumed to have been drawn independently from a common distribution with mean 0 and finite variance. Then, μ_t will be assumed to depend linearly on values of the response variable for preceding time periods and on values of the explanatory variables. Why such assumptions should hold is a question that is seldom asked let alone answered.

Serial correlation in residuals may be too obvious to ignore. The common fix is to assume a specific form of dependence between the ϵ_t. For example, a researcher might assert that $\epsilon_t = \alpha\epsilon_{t-1} + \delta_t$, where now δ_t satisfy the familiar assumptions: the δ_t are drawn independently from a common distribution with mean 0 and finite variance. Clearly, the game has not changed except for additional layers of technical complexity.

Meta-Analysis

Literature reviews are a staple of scientific work. Over the past 25 years, a new kind of review has emerged, claiming to be more systematic, more quantitative,

more scientific: this is "meta-analysis." The initial step is to extract "the statistical results of numerous studies, perhaps hundreds, and assemble them in a database along with coded information about the important features of the studies producing these results. Analysis of this database can then yield generalizations about the body of research represented and relationships within it" (Lipsey, 1997: 15). Attention is commonly focused on the key outcomes of each study, with the hope that by combining the results, one can learn what works. For example, Lipsey (1992) assesses the efficacy of a large number of juvenile delinquency treatment programs, while Sherman and his colleagues (1997) consider in a similar fashion a wide variety of other criminal justice interventions. Meta-analysis is discussed in any number of accessible texts (e.g., Lipsey and Wilson, 2001). Statistical inference is usually a central feature of the exposition.

A meta-analysis identifies a set of studies, each of which provides one or more estimates of the effect of some intervention. For example, one might be interested in the impact of job training programs on prisoner behavior after release. For some studies, the outcome of interest might be earnings; do inmates who participate in job training programs have higher earnings after release than those who do not? For other studies, the outcome might be the number of weeks employed during the first year after release. For a third set of studies, the outcome might be the time between release and getting a job. For each outcome, there would likely be several research reports with varying estimates of the treatment effect. The meta-analysis seeks to to provide a summary estimate over all of the studies.

We turn to a brief description of how summary estimates are computed. We follow Hedges and Olkin (1985, §4E, §A), but relax some of their assumptions slightly. Outcomes for treated subjects ("experimentals") are denoted Y_{ij}^E, while the outcomes for the controls are denoted Y_{ij}^C. Here, i indexes the study and j indexes subject within study. Thus, Y_{ij}^E is the response of the jth experimental subject in the ith study. There are k studies in all, with n_i^E experimentals and n_i^C controls in the ith study. Although we use the "treatment-control" language, it should be clear that meta-analysis is commonly applied to observational studies in which the "treatments" can be virtually any variable that differs across subjects. In Archer's (2000) meta-analysis of sex differences in domestic violence, for example, the "treatment" is the sex of the perpetrator.

One key assumption is that for each $i = 1, \ldots, k$,

(A) Y_{ij}^E are independent and identically distributed for $j = 1, \ldots, n_i^E$; these variables have common expectation μ_i^E and variance. σ_i^2.

Similarly,

(B) Y_{ij}^C are independent and identically distributed for $j = 1, \ldots, n_i^C$; these variables have common expectation μ_i^C and variance σ_i^2. .

Notice that μ_i^E, μ_i^C, and σ_i^2 are parameters—population-level quantities that are unobservable. Notice too that the variances in (A) and (B) are assumed to be equal.

Next, it is assumed that

(C) The responses of the experimentals and controls are independent.

Assumptions (A) and (B) specified within-group independence; (C) adds the assumption of between-group independence. Finally, it is assumed that

(D) studies are independent of one another.

Let $\bar{Y}_i^E$ be the average response for the experimentals in study i, and let $\bar{Y}_i^C$ be the average response for the controls. These averages are statistics, computable from study data. It follows from (A) and (B) that, to a reasonable approximation,

$$\bar{Y}_i^E \sim N(\mu_i^E, \sigma_i^2/n_i^E), \qquad i = 1, \ldots, k \tag{1}$$

and

$$\bar{Y}_i^C \sim N(\mu_i^C, \sigma_i^2/n_i^C), \qquad i = 1, \ldots, k. \tag{2}$$

For the ith study, the "effect size" is

$$\eta_i = \frac{\mu_i^E - \mu_i^C}{\sigma_i}. \tag{3}$$

It is assumed that

$$\eta_1 = \eta_2 = \ldots = \eta_k = \eta. \tag{4}$$

The goal is to estimate the value of η. For instance, if $\eta = .20$, the interpretation would be this: treatment shifts the distribution of responses to the right by 20% of a standard deviation.[8]

There are a number of moves here. Assumptions (A), (B), and (C) mean that treatment and control subjects for each study are drawn as independent random samples from two different populations with a common standard deviation. The standardization in (3) eliminates differences in scale across studies.[9] After that, (4) requires that there is but a single parameter value for the effect size over all of the studies: there is only one true treatment effect, which all of the studies are attempting to measure.

Now the common effect can be estimated by taking a weighted average:

$$\hat{\eta} = w_1\hat{\eta}_1 + \ldots + w_k\hat{\eta}_k, \tag{5}$$

where

$$\hat{\eta}_i = (\bar{Y}_i^E - \bar{Y}_i^C)/\hat{\sigma}_i. \tag{6}$$

[8]We are not quite following the notation in Hedges and Olkin (1985): our standardized effect size is η rather than δ, corresponding to d in Cohen (1988).

[9]Temperature can measured in degrees Celsius or degrees Fahrenheit. The two temperature scales are different, but they are linearly related: $F^\circ = \frac{9}{5}C^\circ + 32^\circ$. The Hedges-Olkin model for meta-analysis described above does not account for transformations more complicated than the linear one. In short, units do not matter; but anything more substantive than a difference in units between studies is beyond the scope of the model.

In (6), the statistic $\hat{\sigma}_i$ estimates the common standard deviation from the sample; the weights w_i adjust for differences in sample size across studies. (To minimize variance, w_i should be inversely proportional to $1/n_i^E + 1/n_i^C$; other weights are sometimes used.) Moreover, we can compute standard errors for $\hat{\eta}$, because this estimator is the product of a convenient and well-defined chance process. For details, see Hedges and Olkin (1985, Chapter 6).

The outcome is both pleasing and illusory. The subjects in treatment and control (even in a randomized controlled experiment, as discussed below) are not drawn at random from populations with a common variance; with an observational study, there is no randomization at all. It is gratuitous to assume that *standardized* effects are constant across studies: it could be, for instance, that the average effects themselves are approximately constant but standard deviations vary widely. If we seek to combine studies with different kinds of outcome measures (earnings, weeks worked, time to first job), standardization seems helpful. And yet, *why* are standardized effects constant across these different measures? Is there really one underlying construct being measured, constant across studies, except for scale? We find no satisfactory answers to these critical questions.

The assumed independence of studies is worth a little more attention. 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, 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. Does social dependence compromise statistical independence? Only if you think that investigators' expectations, attitudes, preferences, and motivations affect the written word—and never forget those peer reviewers.[10]

The basic model represented in equations (1–4) can be—and often is—extended in one way or another, although not in any way that makes the model substantially more believable. Perhaps the most common change is to allow for the possibility of different effect sizes. That is, equation (4) no longer applies; there is no longer an η characterizing all of the studies. Under a "random-effects model," the η_i's are assumed to be drawn as a random sample from some population of η's. Now the goal is to estimate the grand mean μ of this population of η's. However, insofar as meta-analysis rests on a convenience sample of studies, if not a whole population,

[10] Meta-analysts deal with publication bias by making the "file-drawer" calculation: How many studies would have to be withheld from publication to change the outcome of the meta-analysis from significant to insignificant? Typically, the number is astronomical. This is because of a crucial assumption in the procedure—that the missing estimates are centered on zero. The calculation ignores the possibility that studies with contrarian findings—significant or insignificant—are the ones that have been withheld. There is still another possibility, which is ignored by the calculation: study designs may get changed in midstream, if results are going the wrong way. See Rosenthal (1979), Oakes (1990, p.158), or Petitti (1999, p.134).

the random-effects model is at a considerable distance from the facts.[11]

But wait. Perhaps the random-effects model can be reformulated: the ith study measures η_i, with an intrinsic error whose size is governed by equations (1), (2) and (3). Then, in turn, η_i differs from the sought-for grand mean μ by some random error; this error (i) has a mean value of 0 across all potential studies, and (ii) a variance that is constant across studies. This second formulation (a "components of variance" model) is equally phantasmagorical. Why would these new assumptions be true? Which potential studies are we talking about,[12] and what parameter are we estimating? Even if we could agree on answers to those questions, it seems likely—particularly with nonexperimental data—that each study deviates from truth by some intrinsic bias, whose size varies from one study to another. If so, the meta-analytic machine grinds to a halt.

There are further variations on the meta-analytic model, with biases related to study characteristics through some form of regression analysis. The unit of analysis is the study, and the response variable is the estimated effect size. Statistical inference is driven by the sort of random sampling assumptions discussed earlier, when regression analysis was initially considered. However, with research studies as the unit of analysis, the random sampling assumption becomes especially puzzling. The interesting question is why the technique is so widely used. One possible answer is this. Meta-analysis would be a wonderful method *if* the assumptions held. However, the assumptions are so esoteric as to be unfathomable and hence immune from rational consideration: the rest is history. For other commentaries, see Oakes (1990) or Petitti (1999).

Observational studies and experiments

We return to the basic assumptions (A–C) above. How are these to be understood? Meta-analysis is on its most secure footing with experiments, so we begin there. By way of example, consider an experiment with 1000 subjects. Each subject has two possible response. One response will be manifest if the subject is put into the treatment condition; the other, in the control condition. For any particular subject, of course, one and only one of the two responses can be measured: the subject can be put into treatment or control, but not both.

Suppose 500 out of our 1000 subjects are chosen at random, and put into treatment; the other 500 are put in the control condition; the treatment and control averages will be compared. This is the cleanest of study designs. Do assumptions (A-B-C) hold? No, they do not—as a moment's reflection will show. There are two samples of size 500 each, but these are dependent, precisely because a subject assigned to treatment cannot be assigned to control, and vice versa. Thus, (C) fails.

[11] The model now requires two kinds of random sampling: a random sample of studies and then a random sample of study subjects.

[12] If the answer is "all possible studies," then the next question might be, with what assumptions about government spending in fiscal 2025? or for that matter, in 1975? What about the respective penal codes and inmate populations? The point is that hypothetical super-populations don't generate real statistics.

Similarly, the treatment group is drawn at random without replacement, so there is dependence between observations within each group: the first subject drawn cannot appear also as the second subject, and so forth. So the independence assumption in (A) fails, as does the corresponding assumption in (B).

To secure assumptions (A-B-C) in an experimental setting, we need an extremely large pool of subjects, most of whom will not be used. Suppose, for instance, we have 10,000 subjects: 500 will be chosen at random and put into treatment; another 500 will be chosen at random for the controls; the remaining 9,000 will be ignored. In this unusual design, we have the independence required by (A-B-C), at least to a first approximation. But we're not there yet. Assumptions (A) and (B) require that the variance be the same in treatment and control. In effect, treatment is only allowed to add one number—the same for all subjects—to the control response. If different subjects show different responses to treatment, then the constant-variance assumption is likely to be wrong.

To sum up, (A-B-C) hold—to a good approximation—for an experiment with a large pool of subjects, where a relatively small number are chosen at random for treatment, another small number are chosen at random for controls, and the only effect of treatment is to add a constant to all responses. Few experiments satisfy these conditions.[13]

Typically, of course, a meta-analysis starts not from a set of experiments, but from a set of observational studies. Then what? The basic conceit is that each observational study can be treated *as if* it were an experiment; not only that, but a very special kind of experiment, with the sampling structure described above. This is exactly the sort of unwarranted assumption whose consequences we have explored earlier in this essay. In brief, standard errors and P-values are liable to be quite misleading.

[13] With a binary response variable—"success" or "failure"—there does seem to be a logical contradiction in the model: changing the probability p of success automatically changes the variance $p(1-p)$. Naturally, other models can then be used, with different definitions for η. But then, combining binary and continuous responses in the same meta-analysis almost seems to be a logical contradiction, because the two kinds of studies are measuring incommensurable parameters.

For example, in Lipsey (1992), half the studies use a binary response variable (item 87, p. 111). Following Cohen (1988), Lipsey (p. 91) handles these binary responses by making the "arcsine transformation" $f(x) = 2\arcsin\sqrt{x}$. In more detail, suppose we have n independent trials, each leading to success with probability p and failure with the remaining probability $1-p$. We would estimate p by $\hat{p}$, the proportion of successes in the sample. The sampling variance of $\hat{p}$ is $p(1-p)/n$, which depends on the parameter p and the sample size n. The charm of the arcsine transformation—which is considerable—is that the asymptotic variance of $f(\hat{p})$ is $1/n$, and does not depend on the unknown p.

If now $\hat{p}^T$ is the proportion of successes in the treatment group, while $\hat{p}^C$ is the proportion of successes in the control group, $f(\hat{p}^T) - f(\hat{p}^C) = f(p^T) - f(p^C)$, up to an additive random error that is asymptotically normal with mean 0 and variance $1/n^T + 1/n^C$. Lipsey—like many others who follow Cohen—would define the effect size as $f(\hat{p}^T) - f(\hat{p}^C)$. But why is a reduction of 0.20 SDs in time to rearrest—for instance—comparable to a reduction of 0.20 in twice the arcsine of the square root of the recidivism rate, i.e., a reduction of 0.10 in the arcsine itself. We see no rationale for combining studies this way, and Lipsey does not address such questions, although he does provide a numerical example on pp. 97–98 to illustrate the claimed equivalence.

There is one situation where the assumptions underlying meta-analysis can be shown to give reasonable results, namely, combining a series of properly designed randomized controlled experiments, run with a common protocol, to test the global null hypothesis (treatment has no effect in any of the experiments).[14] Of course, even if the global null hypothesis is rejected, so the treatment has some effects on some subjects in some studies, the model underlying meta-analysis is far from demonstrated: the treatment may have different effects on different people, depending on context and circumstance. Indeed, that seems more plausible a priori than the hypothesis of a constant additive effect.[15]

Recommendations for Practice

Convenience samples are a fact of scientific life in criminal justice research; so is uncertainty. However, the conventional techniques designed to measure uncertainty assume that the data are generated by the equivalent of random sampling, or probability sampling more generally.[16]

Real probability samples have two great benefits: (i) they allow unbiased extrapolation from the sample; (ii) with data internal to the sample, it is possible to estimate how much results are likely to change if another sample is taken. These benefits, of course, have a price: drawing probability samples is hard work. An investigator who assumes that a convenience sample is like a random sample seeks to obtain the benefits without the costs—just on the basis of assumptions.

If scrutinized, few convenience samples would pass muster as the equivalent of probability samples. Indeed, probability sampling is a technique whose use is justified because it is so unlikely that social processes will generate representative samples. Decades of survey research have demonstrated that when a probability sample is desired, probability sampling must be done. Assumptions do not suffice. Hence, our first recommendation for research practice: whenever possible, use probability sampling.

[14] Although (A-B-C) are false, as shown above, the statistic $\hat{\eta}_i$ in (6) should be essentially normal. Under the global null hypothesis that all the η_i are zero, the expected value of $\hat{\eta}_i$ is approximately zero, and the variance of $\hat{\eta}_i \Big/ \sqrt{1/n_i^E + 1/n_i^C}$ is approximately 1, by a combinatorial argument. Other tests are available too. For example, the χ^2-test is a more standard, and more powerful, test of the global null. Similar calculations can be made if the treatment effect is any additive constant—the same for all subjects in the study. If the treatment effect varies from subject to subject, the situation is more complicated; still, conventional procedures often provide useful approximations to the (correct) permutation distributions—just as the χ^2 is a good approximation to Fisher's exact test.

[15] Some readers will, no doubt, reach for Occam's razor. But this is a two-edged sword. (i) Isn't it simpler to have one number than 100? (ii) Isn't it simpler to drop the assumption that all the numbers are the same? Finally, if the numbers are different, Occam's razor can even cut away the next assumption—that the studies are a random sample from a hypothetical super-population of studies. Occam's razor is to be unsheathed only with great caution.

[16] A probability sample starts from a well-defined population; units are drawn into the sample by some objective chance mechanism, so the probability that any particular set of units falls into the sample is computable. Each sample unit can be weighted by the inverse of the selection probability to get unbiased estimates.

If the data-generation mechanism is unexamined, statistical inference with convenience samples risks substantial error. Bias is to be expected and independence is problematic. When independence is lacking, the P-values produced by conventional formulas can be grossly misleading. In general, we think that reported P-values will be too small; in the social world, proximity seems to breed similarity. Thus, many research results are held to be statistically significant when they are the mere product of chance variation.

We are skeptical about conventional statistical adjustments for dependent data. These adjustments will be successful only under restrictive assumptions whose relevance to the social world is dubious. Moreover, adjustments require new layers of technical complexity, which tend to distance the researcher from the data. Very soon, the model rather than the data will be driving the research. Hence another recommendation: do not rely on post hoc statistical adjustments to remove dependence.

No doubt, many researchers working with convenience samples will continue to attach standard errors to sample statistics. In such cases, sensitivity analyses may be helpful. Partial knowledge of how the data were generated might be used to construct simulations. It may be possible to determine which findings are robust against violations of independence. However, sensitivity analysis will be instructive only if it captures important features of the data-generation mechanism. Fictional sensitivity analysis will produce fictional results.

We recommend better focus on the questions that statistical inference is supposed to answer. If the object is to evaluate what would happen were the study repeated, real replication is an excellent strategy (Freedman, 1991; Berk, 1991; Ehrenberg and Bound, 1993). Empirical results from one study can be used to forecast what should be found in another study. Forecasts about particular summary statistics, such as means or regression coefficients, can be instructive. For example, an average rate of offending estimated for teenagers in one neighborhood could be used as a forecast for teenagers in another, similar neighborhood. Using data from one prison, a researcher might predict which inmates in another prison will be cited for rule infractions. Correct forecasts would be strong evidence for the model.

Cross validation is an easier alternative. Investigators can divide a large sample into two parts. One part of the data can be used to construct forecasting models which are then evaluated against the rest of the data. This offers some degree of protection against bias due to over-fitting or chance capitalization. But cross validation does not really address the issue of replicability. It cannot, because the data come from only one study.

Finally, with respect to meta-analysis, our recommendation is simple: just say no. The suggested alternative is equally simple: read the papers, think about them, and summarize them.[17] Try our alternative. Trust us: you will like it. And if you

[17] Descriptive statistics can be very helpful in the last-mentioned activity. For one lovely example out of many, see Grace, Muench and Chalmers (1966).

can't sort the papers into meaningful categories, neither can the meta-analysts. In the present state of our science, invoking a formal relationship between random samples and populations is more likely to obscure than to clarify.

Conclusions

We have tried to demonstrate that statistical inference with convenience samples is a risky business. While there are better and worse ways to proceed with the data at hand, real progress depends on deeper understanding of the data-generation mechanism. In practice, statistical issues and substantive issues overlap. No amount of statistical maneuvering will get very far without some understanding of how the data were produced.

More generally, we are highly suspicious of efforts to develop empirical generalizations from any single dataset. Rather than ask what would happen in principle if the study were repeated, it makes sense to actually repeat the study. Indeed, it is probably impossible to predict the changes attendant on replication without doing replications. Similarly, it may be impossible to predict changes resulting from interventions without actually intervening.

References

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

Berk R. A. (1988) "Causal Inference for Statistical Data," in N. J. Smelser (ed.), *Handbook of Sociology*, Beverly Hills: Sage Publications.

Berk R. A. (1991) "Toward a Methodology for Mere Mortals," in P. V. Marsden (ed.), *Sociological Methodology*, Volume 21, Washington, D. C.: The American Sociological Association.

Berk R. A. and A. Campbell (1993) "Preliminary Data on Race and Crack Charging Practices in Los Angeles,"*Federal Sentencing Reporter* 6 (1): 36–38.

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

Ehrenberg A. S. C. and Bound J. A. (1993) "Predictability and Prediction," *Journal of the Royal Statistical Society*, Series A, 156 Part 2: 167–206.

Freedman D. A. (1999) "From Association to Causation: Some Remarks on the History of Statistics."*Statistical Science* 14: 243–58.

Freedman D. A. (1997) "From Association to Causation via Regression," In V. McKim and S. Turner (eds.), *Causality in Crisis?* University of Notre Dame Press, 113–82 (with discussion).

Freedman D. A. (1995) "Some Issues in the Foundation of Statistics," *Foundations of Science* 1: 19–83 (with discussion). Reprinted in B. van Fraasen (ed.), *Some Issues in the Foundation of Statistics*, Kluwer, Dordrecht.

Freedman D. (1991) "Statistical Models and Shoe Leather," in P. V. Marsden (ed.) *Sociological Methodology*, Volume 21, Washington, D. C.: The American Sociological Association.

Freedman D. (1987) "As Others See Us: A Case Study in Path Analysis," (with discussion) *Journal of Educational Statistics*, 12: 101–223.

Fromby T. B., R. C. Hill and S. R. Johnson (1984) *Advanced Econometric Methods*, New York: Springer Verlag.

Grace N. D., H. Muench and T. C. Chalmers (1966) "The present status of shunts for portal hypertension in cirrhosis," *Gastroenterology* 50: 684–91.

Gross S. R. and R. Mauro (1989) *Death and Discrimination*, Boston, Northeastern University Press.

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

Johnston J. (1984) *Econometric Methods*, New York: McGraw Hill.

Kreft I. and J. de Leeuw (1998) *Introducing Multilevel Modeling*, Newbury Park, Ca: Sage Publications.

Kruskal W. (1988) "Miracles and Statistics, The Casual Assumption of Independence," *Journal of the American Statistical Association*, 83 (404): 929–940.

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.

MacKenzie D. L. (1991) "The Parole Performance of Offenders Released from Shock Incarceration (Boot Camp Prisons): A Survival Time Analysis," *Journal of Quantitative Criminology*, 7(3): 213–236.

Meehl P. E. (1978) "Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology,"*Journal of Consulting and Clinical Psychology*, 46: 806–834.

Nagin D. S. and R. Paternoster (1993) "Enduring Individual Differences and Rational Choice Theories of Crime," *Law & Society Review*, 27(3): 467–496.

Oakes M. W. (1990) *Statistical Inference*, Chestnut Hill, MA: Epidemiology Resources Inc.

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

Phillips S. and R. Grattet (2000) "Judicial Rhetoric, Meaning-Making, and the Institutionalization of Hate Crime Law," *Law & Society Review*, 34(3): 567–606.

Rosenthal R. (1979) "The 'file drawer' and tolerance for null results," *Psychological Bulletin* 86: 638–41.

Sherman, L. W., D. Gottfredson, D. MacKenzie, J. Eck, P. Reuter and S. Bushway (1997) *Preventing Crime: What Works, What Doesn't, What's Promising?* Washington, DC: U.S. Department of Justice.

Weisberg, S. (1985) *Applied Linear Regression*, New York, John Wiley.

White, M.D. (2000) "Assessing the Impact of Administrative Policy on the Use of Deadly Force by On- and Off-Duty Police," *Evaluation Review*, 24(3): 295-318.

Published as Chapter 10 in T. G. Blomberg and S. Cohen (eds.), *Law, Punishment, and Social Control: Essays in Honor of Sheldon Messinger*, 2nd ed. (2003), Aldine de Gruyter, pp. 235–254.