

DEVIATING FROM THE MEAN: THE DECLINING SIGNIFICANCE OF SIGNIFICANCE

MICHAEL D. MALTZ

Most of the methods we use in criminology to infer relationships are based on mean values of distributions. This essay explores the historical origins of this issue and some counterproductive consequences: relying too heavily on sampling as a means of ensuring "statistical significance"; ignoring the implicit assumptions of regression modeling; and assuming that all data sets reflect a single mode of behavior for the entire population under study. The essay concludes by suggesting that we no longer "make do" with the standard methodologies used to study criminology and criminal justice, and recommends developing categories that more accurately reflect behavior and groupings than the ones we currently use; looking at alternative sources of data, including qualitative data such as narrative accounts; and developing alternative methods to extract and analyze the data from such sources.

Most of the statistical methods in use in criminology today are based in whole or in part on comparing the value of a sample-derived statistic with a standard. In an experimental situation, a t test is used to compare the experimental sample's mean with the mean for the control group; in an analysis of variance, an F test is used to compare a number of different means; in a regression model, the value of a beta weight is compared with zero. In fact, the very term *deviance* implies that there is a norm--of behavior--to which we compare others' behavior. Although we recognize that different people have different standards of behavior that are just as valid as our own, with the implicit understanding that a great deal of variation in behavior is permissible, when it comes to our statistical

I have discussed the ideas embodied in this article with many of my colleagues, although they do not necessarily agree with them. I am grateful to Alfred Blumstein, Carolyn Block, Jan Chaiken, Marcia Chaiken, Jacqueline Cohen, Edward Day, Marcia Farr, David Fogel, Lisa Frohmann, Paul Goldstein, Handle Lazarus-Black, Matthew Lippman, Patrick McAnany, Lonnie Morris, Albert Reiss, Dennis Rosenbaum, Robert Sampson, Stephen Stigler, and David Weisburd for their comments. Much of the material in this article is drawn from ASC presentations in 1990 and 1992.

JOURNAL OF RESEARCH IN CRIME AND DELINQUENCY, Vol. 31 No. 4, November 1994 434-463 © 1994 Sage Publications, Inc.

methods, we are not quite so tolerant: Most statistical analyses are based on means.

The reason for this is partly historical and goes back to the origins of statistical analysis. The tension between what I call "reification of the mean" and a (statistical) accommodation of variation is one of long standing and is still being played out today in the criminological (and other social science) literature. I attempt herein to explain the origins of this tension and the impact it has had on theory-building in criminology.

This essay continues the discussion about research in criminology in the November 1993 issue of the *Journal of Research in Crime and Delinquency*. In the editor's introduction, Fagan (1993, p. 38 1) wrote of the paradox "where our methods grow more powerful and precise as we move ever further away from our data and the complex realities they represent." Responses in that issue to this paradox were quite thoughtful and dealt with the development of theories in criminology, the methods used to test those theories, and the data analyzed by those methods. As an unabashedly quantitative researcher, I have also tried to deal with these issues but from a different standpoint than the authors in that issue. I find, however, that my own concerns strongly resonate with those of Braithwaite (1993) in his concern for the need to deal with specific contexts rather than attempt to generalize to all situations; McCord (1993) in her addressing the same topic as well as the limitations of regression and the need to go beyond testing theories; and Sampson (1993) in his call for methods that incorporate dynamic contextualism and narrative positivism.

I first describe the historical origins of the research paradigms and methods used in the social sciences. The validity of some of their original assumptions is then scrutinized, pointing out some of the problems of these methods and research paradigms. I then show how the assumptions still permeate criminological research to this day, calling into question some of the findings. I conclude by suggesting that we reconsider using methods that are based primarily on estimating the mean of a distribution and use and develop methods that accommodate diversity.

DISCOVERING AND REIFYING THE MEAN

One of the more interesting chapters in the history of statistics concerns how the normal distribution was first derived and developed, based on the needs of ocean navigation. In the 17th and 18th centuries, astronomical observations were being used to develop navigation tables. Astronomers were confronted with the problem of reconciling many different (and somewhat inconsistent) observations of the same object

(e.g., the moon, the North Star) to determine its exact current location and to predict its exact future location. The observations were inconsistent because of inherent errors in the observations caused by, inter alia, variation in the optical characteristics of the atmosphere, mechanical hysteresis (i.e., play) in the measuring instruments, and human error. Consequently, there was a need to develop methods to reconcile the measurements and improve the ability to predict the location of the moon, stars, and planets for navigational purposes.

Stigler (1986, pp. 55-61) recounts how Legendre, building on the work of Euler, Mayer, Boscovich, and Laplace, developed the method of least squares (i.e., minimization of the sums of squares of the error terms in a system of linear equations) as a means of combining error-laden measurements to provide a solution to the equations. This line of research also established that if errors from the true value were distributed normally, then the arithmetic mean was the estimate that minimized the mean square error (Porter 1986, p. 96). Because astronomical observations were (generally) distributed normally, there was a good fit between the theoretical basis for the method and the empirical reality, *p*. Selecting the arithmetic mean as the "true" value of a set of normally distributed observations minimized the mean square error. The proof by Laplace in 1810 of the central limit theorem, showing that the mean of a series of measurements with independent errors would be normally distributed (Stigler 1986, pp. 137, 143), solidified the concept that this curve was the distribution to use when one had error-laden data.

As a consequence, the normal distribution became known as the "error law" and *the bell-shaped curve became indelibly associated with errors*. It is no wonder that when similar curves were found in other situations, the processes that generated them were considered to consist of a mean value plus error terms, that is, a "true" value plus unwanted deviations.

This is exactly what happened in the middle of the 19th century. In 1844, the Belgian astronomer Adolphe Quetelet showed that this same error law distribution applied to the distribution of human physical features, such as height and girth: "This surprising testimony to the power of the method of celestial mechanics bolstered Quetelet's longstanding claim that the social sciences could do no better than to imitate the physical" (Porter 1986, p. 100). Even further, he took this similarity of distributions to signify that these variables had a true value exemplified by the mean, and that deviations from that value were errors: He wrote, "One may ask if there exists, in a people, *un homme type*, a man who represents this people by height, and in relation to which all other men of the same nation must be considered as offering deviations" (quoted in Hacking 1990, p. 105).

Quetelet also used these techniques in the analysis of judicial statistics. He saw in the relative constancy of "moral statistics," for example, conviction rates, the equivalent of a natural law. In other words, he concluded that there was an innate "*penchant au crime*" shared by all; that there was such an entity as the "average man," or even more to the point, the "average moral man" (*l'homme moyen morale*). That is, he felt that "[i]f an individual at any given epoch of society possessed all the qualities of the average man, he would represent all that is great, good, or beautiful" (quoted in Stigler 1986, p. 171). This was apparently a *direct consequence of his interpreting the variation in individual characteristics as exemplifying error*, in much the same way that an astronomical observation would consist of the true value corrupted by errors in measurement.

Émile Durkheim also saw in such distributions error rather than natural variation, concluding that there were social laws that acted on all individuals "with the same inexorable power as the law of gravity" (Hacking 1991, p. 182; see also Duncan 1984, p. 99). For example, he wrote of "suicidogenetic currents" in society, as if there was an innate propensity of the same magnitude for suicide within every individual (Hacking 1990, p. 158).¹

Others, especially Francis Galton, criticized this approach. He saw variation as natural and desirable and criticized those who focused on averages as being "as dull to the charm of variety as that of the native of one of our flat English counties, whose retrospect of Switzerland was that, if its mountains could be thrown into its lakes, two nuisances would be got rid of at once" (Galton 1889, quoted in Porter 1986, p. 129). Yet Galton's view was not shared by most of the social scientists of the time. In fact, there was widespread belief that those far from the mean were deviant in the pejorative sense of the word, that they were "errors." Statistics was not seen so much as the *science of variation* (Duncan 1984, p. 224) as it was the *science of averages*.

Thus, with some exceptions, the *zeitgeist* at the turn of the century was to use statistics to estimate mean values and to consider the residuals of those estimates to be errors. That is, it was believed that statistical variation reflected not potentially beneficial diversity but, rather, unwanted deviations from the "true" value, exemplified by the mean.²

Accounting for Variance: The Experimental Paradigm

It was recognized that the mean value of a group could be changed by changing the conditions imposed on it. Different interventions could be tested to see what effect they had on outcomes. The drawback to

conducting these tests was the lack of a rational basis to determine how large a group had to be tested to determine whether it was the intervention or just random fluctuations that changed the mean value. The experimental paradigm proposed and promoted by Sir Ronald A. Fisher in the 1920s addressed this issue. It was the first practical application of the analysis of variance, which Fisher used in the design of agricultural experiments, to be given a wide audience, as it was in his 1935 classic, *The Design of Experiments* (Fisher 1935; see Gigerenzer 1987, p. 19).

Groups were to be compared, some given treatments (the experimental groups), another not (the control group). Groups were needed because individual units (persons, plots of land) had inherent variation, caused by different factors that were not held constant (e.g., rainfall, soil fertility, and sunlight in the case of agricultural experiments), as well as by the experimental treatment (e.g., type of seed or fertilizer). The experimental design permitted the researcher to test the hypothesis that the difference in outcomes between the experimental and control groups was attributable to the treatment and not to extraneous factors.

Fisher's book and methodology revolutionized the way experiments were conducted. The concepts of random sampling and of selecting a sample size on the basis of the expected extent of variation in the data were great steps forward and permitted the rationalization of experimental design. Inferences could be made about the entire population using a sample, the size of the sample being determined by the expected variance and the degree of confidence desired in the results. Thus methodology in the social sciences borrowed from the agricultural sciences, as well as astronomy.

The overwhelming acceptance of statistical methods in the social sciences may also have been encouraged by the finding by Meehl (1954) that clinical judgments are inferior to actuarial or statistical judgments in a wide variety of domains (Faust 1984).³ This led, in part, to an increase in the authority accorded a quantitative analysis over a qualitative one. It reinforced the belief that statistical analyses of surveys were superior to case history methods in attitude research, an issue studied by Stouffer in 1930 (Bennett 1981, p. 135).

WEAKNESSES IN THE STATISTICAL METHODS

These advances in and testimonials to statistical methodology led social scientists to collect and analyze data using the new statistical tools. Yet the tools had a number of weaknesses that were unrecognized, at least initially, by the social scientists who used them. They include overreliance on hypothesis testing, the reliance on the asymptotic normality of

sampling distributions, reliance on the robustness of regression, and the manner in which we validate the models that result from our analyses. Each is discussed below.

Hypothesis Testing and Statistical Significance

Although hypothesis testing is a mainstay of social science research, it has a number of deficiencies. One is the lack of inclusion of an alternative hypothesis, or the consideration of Type II errors;⁴ these were concepts that came from Neyman and Pearson and were subsequently incorporated into a hybrid system of inferential statistics that continues to plague students to this day (Gigerenzer 1987, p. 21). Others include:

- Too much attention is paid to the "ritualistic reliance on statistical significance" (Mason 1991, p. 344). In 1962, the editor of the *Journal of Experimental Psychology* stated that he was strongly reluctant to publish papers whose "results were significant [only] at the .05 level, whether by one or two-tailed test!" (quoted in Gigerenzer 1987, p. 21). This had the unfortunate consequence of rewarding researchers "for the exorcism of the null hypothesis, rather than for a careful experimental design or a phenomenological preanalysis" (Gigerenzer 1987, p. 23).
- The assumptions on which statistical significance tests are based—that the group(s) under study are random samples of a larger population—are often not met. Very often the samples are samples of convenience. And as Freedman, Pisani, Purves, and Adhikari (1991, p. 506) underscore: "If a test of significance is based on a sample of convenience, watch out." Yet social scientists who deal with samples of convenience often use significance tests.
- The meaning of significance is not well understood. In fact, there have been cases of researchers, with data on a full population, sampling it just so that they could apply the asymptotic normality property to the mean of the sampling distribution—as if, by the magic of statistical analysis, using *fewer* data points would provide more useful findings than those obtainable using all the data.
- It changes the focus of research from measurement to the analysis of numbers, resulting in a "marriage between sophisticated statistics and naive procedures for measurement" (Gigerenzer 1987, p. 23). As Duncan (1984, p. 207) questioned, "Is it really credible that social scientists in 1965 already knew how to *measure* 2,080 distinct sociological quantities?" (emphasis in the original).
- As McCord (1993, p. 412) points out, researchers mistakenly try to find more support for their theories by using a larger N. Lieberman (1985, p. 105) notes, "One does not really improve a situation with poor data by increasing the 'N' and then hoping to deal with more variation." Despite the *Journal of Experimental Psychology's* dictum, not all small-sample experiments should be thrown out because of the lack of statistical significance. Mosteller (1981, p. 881) describes an experiment conducted in 1747, with sailors afflicted with scurvy. A physician administered one of six treatments—vinegar, sea water, cider, vitriol elixir, citrus fruit, and nutmeg—to two sailors each, for a total N of 12. Only the two who ate the citrus fruit were cured, in just a few days, leading the physician to believe that he had found the cure, which is a logical conclusion.⁵

But this result is doubtless not significant at the .05 level. [What would be the fate of a crime analyst who told the police chief, "We only have two cases in which the victim was dismembered; this is too small an N to infer a pattern"?]⁶

- Statistical significance does not imply substantive significance, and most researchers know this-but this does not stop them from implying that it does. In other words, there are (conceptually) Type I and Type II errors that distinguish statistical significance from substantive significance: not all statistically significant findings are substantively significant, and not all substantively significant findings are statistically significant. It has been suggested that it would be better to use the word *discernible* or *distinguishable*⁷ instead of significant in describing statistical tests, because these tests are designed to estimate the probability that one parameter (e.g., a sample mean) can be distinguished from another.

Why all this focus by researchers on statistical significance, despite these well-known faults with it? The problem confronted by researchers is how to draw inferences from a large data set, and the only inferences with which most are familiar are those relating a sample to a parent population. The problem thus is one of inference and interpretation. Statistical significance refers only to inferences beyond the sample at hand, to some (often hypothetical) population from which the sample was drawn. If the group under study is not a random sample, it does not mean that inferences cannot be drawn, but only that they will not be related statistically to a parent population: one can still calculate standard errors and within and between-group variance, and make confidence statements, but using descriptive statistics, not inferential statistics. The validity of making inferences beyond the sample at hand then must be based on the arguments of the researcher and not on the sampling distribution.

In summary, the experimental design proposed by Fisher permitted a rational determination of sample size, and it made use of variance rather than just discarding it as error. It provided researchers with an unequivocal procedure for making inferences from equivocal data and has been instrumental in moving the social sciences onto a more logical and quantitative footing. Yet the use of statistical significance is so often used inappropriately that, in all too many cases, it is a meaningless exercise (or, at the very least, its meaning is not the one claimed by the researchers). But one implicit assumption about the nature of the variance did carry through from the past: that there was one "true" value for the outcome, and if the different mediating factors (i.e., the independent variables) were precisely known and measured, this true value could be calculated.

Sampling Distributions

One of the advantages of Fisher's experimental paradigm was the ability to take a random sample of a population and, regardless of the shape of the population distribution, calculate the sample mean and be confident that it was arbitrarily close to the true mean (the larger the sample, the closer the sample mean to the true mean). For example, Figure 1(a) depicts the income distribution of a hypothetical sample that is bimodal, with about half of the population having incomes distributed around \$4,000 and the other half with incomes distributed around \$16,000, with the overall mean value around \$10,000.

According to sampling theory, if this were a random sample drawn from a parent population, the sample mean would be close to the actual population mean, its closeness depending on the sample size. Not only that, but were we to take numerous random samples (of the same size) from this population, the means of the different samples would be all nestled about the true mean and distributed approximately normally, as depicted in Figure 1(b). The fact that the sample means are normally distributed (given a sufficiently large sample size) means that, if comparisons are to be made of two different distributions, one can determine the likelihood that the two population means are within a specified range of each other, with a level of confidence determined by the sample sizes.

It is often considered too difficult to consider the actual population distribution, especially when (as in this example) the distribution is irregular. As a consequence, analysts often focus on the comparison of the sample means without really investigating the nature of the distributions from which they were generated. But relying on the sample mean alone can be very misleading: note that, for the given example, the mean value is held by very few members of the sample. One wonders, in a case like this, about the advisability of using the mean as a "measure of central tendency"—since there is no central tendency!

To make this point more concrete, consider the case of an experimental intervention that is to be evaluated using two samples, experiment and control, randomly selected from a parent population. Suppose that the outcomes of the experimental group were distributed as in Figure 1(a) and the outcomes of the control group were distributed as in Figure 1(c), in which the distribution of incomes is roughly unimodal. Suppose also that neither the means nor the variances of the two samples are (statistically) significantly different from each other, so that the sampling distribution in Figure 1(b) applies to both.⁸ According to accepted doctrine, the null hypothesis cannot be rejected, leading to the

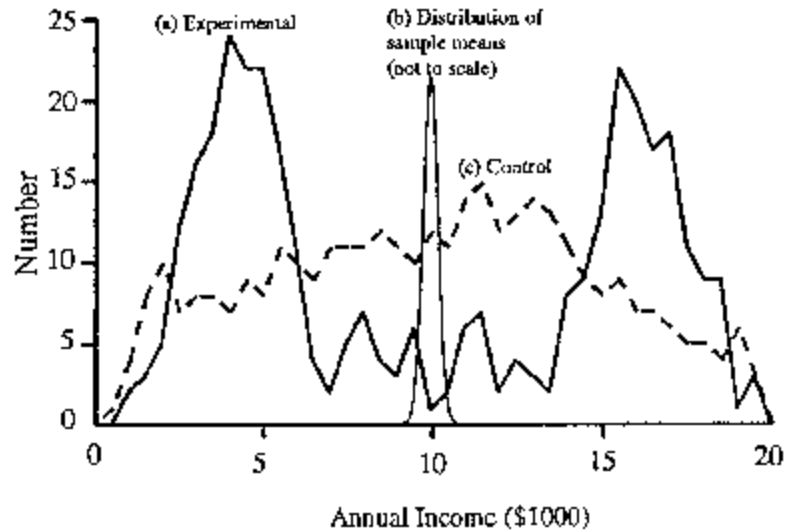


Figure 1: Data from a Hypothetical Experiment

conclusion that the experiment had no effect. Yet if one goes beyond merely testing the null hypothesis and considers the sample distributions, not just the distribution of sample means, one would see that the experiment had a (substantively) significant effect, decreasing the outcomes for one segment of the sample and increasing outcomes for another segment. That is, one can look at experimental data not only in terms of *testing* hypotheses, but in terms of *generating* hypotheses as well.

This type of problem was apparently at the root of the disagreement between the late Hans Zeisel (1982a, 1982b) and Peter Rossi and Richard Berk (1982): Rossi, Berk, and Lenihan (1980; see also Berk, Lenihan, and Rossi 1980) evaluated an experiment in which ex-offenders in Georgia and Texas were randomly assigned to a treatment group--that was given financial aid for a period of time to help the ex-offenders in their transition to the outside world--or to a control group that was not given such aid. After a finding of no difference in income earned after release from prison between the two groups, Berk et al. (1980) analyzed the data further. They found that there was evidence to suggest that one subgroup of offenders was helped by the transitional aid (i.e., earned more than expected), but another group of offenders did not take advantage of the financial "breathing room" and used the transitional aid as an excuse for

not working. Zeisel (1982a) contended that this runs counter to the logic behind the experimental paradigm, which relies on the difference of sample means, whereas Rossi and Berk (1982) maintained that the experimental paradigm does not prevent an analyst from examining the data at greater length to extract additional findings.

Focusing only on the hypothesis that the means are different is another form of reification of the mean: it ignores all characteristics of the population distribution other than the mean. In some sense, sampling is equivalent to squinting at an object instead of opening one's eyes fully; it permits the viewer to see some broad features, for example, the mean. However, this is done at the expense of missing the details, the very features that make research an exciting art rather than a formulaic "science."

Implicit Assumptions of Regression

Experimentalists collect their own data and have control over experimental conditions and data, but other social scientists do not have control over the variables they use. They more often use data from field settings, questionnaires, and official records, and employ quasi-experimental or correlational designs. As a consequence, much of the work in criminology consists of attempting to control for variation in the population under study by doing multivariate regression analyses, that is, by attempting to account for part of the variation in outcomes by it to the variation in independent variables. There are a number of problems with the way regression has been applied in the analysis of social phenomena.

- When the number of independent variables is large, analysts often use the shotgun approach to find which of the collected variables are key ones. That is, the data are regressed against as many independent variables as are available, and those that are the most statistically significant are selected. Freedman (1983) has shown the flaws in this approach: he simulated some random uncorrelated data, ran a regression and found some relationships, eliminated the weaker covariates from the regression equation and produced an even better model than before-and all with data that had no inherent relationship.⁹
- Lieberman (1985) describes an even more fundamental flaw in regression analysis, assuming that the functional relationship between variables is symmetric. That is, a positive correlation between variables A and B implies that increasing variable A causes variable B to increase *and* that decreasing variable A will cause variable B to decrease. This would only be true if variables A and B did not interact with other variables that, in turn, affect A and B; in a

social system, this is virtually never the case.

- Regression analyses, for the most part, imply linear (or log-linear) models. But no one would call offending, for example, a linear process. The rationale that makes it justifiable to employ linear models for nonlinear phenomena is that, if a functional relationship among variables is fairly well behaved (i.e., its slope is continuous) in the neighborhood of a point, it can be approximated by a linear function in that neighborhood. That is, continuous processes are linear "in the small." This leads to the use of regression in one form (linear) or another (log-linear, logistic, etc.). Abbott (1988) argues that this use implies a (mistaken) belief in a "general linear reality."
- In studies that have used these forms of regression, rarely are the method's inherent assumptions (e.g., Malinvaud 1970, p. 82) tested against the data to see if they are met (Freedman 1991 a, p. 305). Moreover, the implicit assumption of linearity is sometimes sorely stretched, as when it is assumed that the same regression equation can be used to relate homicide rates to a number of independent variables (Ehrlich 1975) over the four decades from 1930 to 1970. Such a formulation disregards the possible nonlinearities introduced by a world war, civil rights legislation, the increasing availability of handguns, urbanization, changes in internal migration patterns, the growth in automobile ownership, improvements in medical technology and emergency room procedures, and other such "nonlinear" factors.
- The reason given for disregarding the assumptions of regression analysis is that regression is considered a "robust" method of analysis; that is, it can withstand substantial violations of its assumptions and still produce useful results. However robust it is, no method can or should be used without determining the extent to which it can be used; yet this does not occur very often.
- How the variables are entered into the regression equation can affect the outcome, especially when using nonlinear, log-linear, or logistic regression. For example, Ehrlich's (1975) log-linear model of homicide incorporated the probability of arrest, rather than the probability of avoiding arrest. This can make a substantial difference, especially when modeling deterrence: the probability of arrest increases by 100% when it doubles from 1% to 2%; but to the offender the probability of avoiding arrest changes imperceptibly, from 99% to 98%.
- Although regression methods downplay the importance of the mean by showing how its variation can be attributed to the independent variables, these methods *still* reify mean values. Consider the output of these methods: coefficients and their corresponding t statistics. Thus there is an implicit assumption that the same regression equation (and its beta weights) applies to every member of the group under study. This assumption should be tested; see the Unimodality section below.

Thus there are many valid criticisms of the use of regression in the social sciences, especially Lieberman (1985) and Freedman (1991a, 1991b). Yet many researchers seem to pay no attention to them because the number of papers in which regression is misused has not diminished

since the criticisms were first discussed. The fault may lie with the way we teach statistics in graduate schools (Mason 1991) and the often counterproductive suggestions of referees ("Instead of using OLS here, try logit") who were trained in these graduate schools.

Model Validation

Regression modeling is often misused in another way. Freedman (1985, p. 348) notes:

In social-science regression analysis, usually the idea is to fit a curve to the data, rather than figuring out the process that generated the data. As a matter of fact, investigators often talk about "modeling the data." This is almost perverse: surely the object is to model the phenomenon, and the data are interesting only because they contain information about that phenomenon. Whatever it is that most social scientists are doing when they construct regression models, discovering natural laws does not seem to be uppermost in their minds.

Models are often validated by randomly splitting the population under study and using one of the subpopulations, the estimation or construction sample, to develop the model-by determining which independent variables are significantly associated with the outcome variable(s), and the relative weight assigned to each independent variable. This model is then applied to the other subpopulation, the validation sample, to see how well the model predicts the outcome for this sample.

This technique does not actually validate a model. First, it does not test the model's inherent construct validity: if the process under study is actually nonlinear, but is modeled as a linear process, then the resultant model will not be the best model of the process but the best *linear* model.

Second, a model that fits the validation sample as well as the construction sample is not necessarily a validated model. For example, suppose that the two subpopulations are exactly alike (e.g., pairs of identical twins). In this case, the two samples will obviously give identical results. And just as obviously, this would not validate the model because the identical results would occur regardless of the validity of the model.

The split-sample validation technique validates the fact that the statistics of the two subpopulations are sufficiently similar that they cannot be distinguished. That is, *validation* using a split sample may only show that the random assignment procedure used to split the population did, in fact, split it into two indistinguishable subpopulations. Although this

procedure "can provide a useful antidote to overfilling" (Berk 1991, p. 319), it does not validate the model.

A more useful approach was taken by Berk (1991, p. 321) in attempting to validate a model built from data in one jurisdiction by testing it with data from another jurisdiction. He found this procedure to be useful in documenting the strengths and weaknesses of the original model, but "there was no simple or straightforward way to determine how well the model fared in the new dataset.... [M]uch rested on difficult judgment calls." In other words, model validation cannot always be based on an off-the-shelf technique and a set of rules.

PROBLEMS WITH THE RESEARCH PARADIGM

Aside from problems with specific methods, there are also problems with the research paradigms within which these assumptions are imbedded. In particular, there is a strong bias (perhaps a legacy of the editorial stance of the *Journal of Experimental Psychology*) to sample all populations regardless of the need to do so. In addition, implicit in the use of regression models is the belief that the distribution under study is unimodal: that is, it has only one point around which cluster all members of the population. These issues are discussed below.

Sampling and "Connectivity"

The statistical methodologies of the social sciences, as discussed earlier, are based in large measure on sampling, permitting the use of significance tests. By relying so much on sample-based data, however, we implicitly reject the possibility of examining phenomena that cannot be studied by random sampling. In criminology, co-offending behavior is one phenomenon that cannot be studied using random samples.

That juveniles commit offenses more often with others than by themselves was detailed by Shaw and McKay (1969); see also Hindelang (1976) and Zimring (1981). But if a population of juvenile offenders is sampled, most of the relationships among the co-offenders will not be captured in the sample. Even a cursory review of juvenile offense records would show numerous instances of three or four individuals charged concurrently with the same offense; Reiss (1986, 1988) shows how this

penchant among juveniles for co-offending affects crime rates and criminal careers. In particular, to investigate co-offending among juveniles, Reiss and Farrington (1991, p. 368) studied "*all the boys* then [in 1961-62] aged eight or nine who were on the registers of six state primary schools within a one mile radius" of the study's research office in London [emphasis added]. They showed how co-offending varied, inter alia, with age of offender and with offense types; tracking their co-offending behavior would not be possible in a random sample.

This problem with sampling relates to the implicit assumption of sampling theory, of the independence of all individuals in the sample; that is, knowledge of A's behavior has no effect (statistically) on B's behavior. But we know that they are not always independent in criminology-and other social sciences as well (Abbott 1988, p. 178). In fact, Coleman (1990, pp. 318-20) considers one of the most important factors in studying a social system to be the closure among its actors, that is, their interrelationships.

Moreover, not only do *offenders* act jointly and nonindependently, but *offenses* can be related to each other as well. Consider what occurs in the conduct of field experiments in crime prevention or control, in which the count of events serves as the outcome measure. For example, suppose that some of the more active offenders in a community are responsible for about 300 burglaries a year, a rate not out of the ordinary (Chaiken and Chaiken 1982). This means that burglaries are not committed independently, which is a key assumption for making inferences from a random sample. A reduction in burglaries by 300, which might be statistically significant under the assumption of independence, would not necessarily be significant given the connected nature of the events. The problem here is in the unit of analysis chosen: Perhaps we should be counting offenders, not offenses. Although we normally do not know who commits which offense(s), it may be possible to make reasoned estimates.¹⁰

Organized crime and white-collar crime are two other areas where sampling unduly restricts the view of the phenomenon under study. Organized crimes differ from common crimes specifically because they are connected (Maltz 1990); similarly, the scams committed by white-collar offenders are amenable to study as network phenomena (Waring 1992).

Unimodality

A focus on the mean of a population contains an implicit assumption that the population all acts about the same. This assumption of unimodality

is reflected not only in assumptions about the populations under study but is also reflected in the way categories are constructed and in our expectations about the efficacy of treatments. These assumptions are explored in this section.

Unimodality and models. Regression models implicitly assume that there is a mean value for each of the variables and that the individual cases cluster around this mean in a multivariate normal distribution.¹¹ This formulation of the model implies that only one mechanism is working, that the cases under study represent variations on a single theme, ignoring the possibility that "important subsets are inappropriately being merged as though they represented homogeneous populations" (McCord 1993, p. 413). Therefore, many studies ignore the possibility that individuals may cluster into different types, each with different motivational patterns and different behavior patterns: all too often all individuals are thrown into the same hopper and analyzed (implicitly, if not explicitly) as if they are of the same general type—the computerized crank is turned, coefficients are obtained, and a new "model" is generated or another theory tested.¹² This puts the modeling cart before the data horse.

Regression analyses produce coefficients for a single equation that "explains" the contribution of each variable. These tests purport to show which single value of coefficient is best according to some criterion (e.g., minimum mean square error, maximum likelihood), irrespective of the fact that there may be two distinct values corresponding to two different types of individual. Studies using these analytic methods, whether they are of birth cohorts (e.g., Wolfgang, Figlio, and Sellin 1972) or releasees (e.g., Rossi, Berk, and Lenihan 1980), analyze different types of individuals (white-collar offenders, petty thieves, sex offenders, violent offenders) as if the same regression model can accommodate all of them. As Sampson and Laub (1992, p. 71) point out, there are dangers in "relying on measures of central tendency that mask divergent subgroups."

Consider a cohort of prisoners that has within it a homicidal pedophile like John Wayne Gacy, a junk bond swindler like Michael Milken, a streetgang leader like Jeff Fort, and an organized-crime capo like John Gotti, as well as the usual assortment of armed robbers, drug dealers, and petty thieves. Assuming, a priori, that they are all governed by a single behavioral pattern is absurd on its face. This has no more merit than studying a cohort of cancer patients solely on the basis of their having entered the hospital in the same year: the causes and course of bone, breast, colon, lung, and skin cancer are quite different from each other and should be investigated separately. Similarly with criminality, some may be driven by family deficiencies that can be treated by treating the family; others may be driven by peer-related motivations or environmental factors that are amenable to treatment by changing their environment; still

others may have deep-rooted psychological or physiological problems that manifest themselves in violent behavior, that may not even be treatable. Yet instead of considering these types separately, they are usually grouped together in one regression "model."

Grouping them together implies that there is only one mechanism operating, and that individuals or crimes can be seen as varying around the mean value that defines that mechanism. In other words, *the assumption is implicitly made that the data are unimodal*, that there is one "grand unified theory" of crime or criminal behavior. This is another aspect of the problem of reification of the mean discussed earlier, albeit in a multivariate context. If it is actually the case that more than one mechanism is operating (and that the study variables can actually be used to distinguish these modes), then we are modeling a multimodal situation with a unimodal model of the process(es). To use the same metaphor as Galton, but bringing it closer to home, this is equivalent to fitting the (multimodal) Sierra Nevada with a (unimodal) model of Mount Whitney.

Although there are certainly benefits to grouping them together—for example, to test a treatment that may be effective for a variety of problems—doing so also masks potential effects. The fact that a graph of age of onset of criminality (for most types of crime) is similar for populations in different contexts and from different countries may be interesting (Gottfredson and Hirschi 1990), but what it shows more than anything else is that crime is a young person's game—just as a similar graph of age of onset of cancer (for most types of cancer) would show that cancer is an old person's disease. The fact that most (or even all) offenders lack self-control (Gottfredson and Hirschi 1990) may also be interesting, but it is no more helpful than the observation that all cancer cells multiply uncontrollably. In neither case is this finding as helpful as would be a more intensive analysis of similar cases.

Unimodality and categorizing data. Unimodality is also manifested in the way data are categorized. Social scientists tend to employ categories that are almost cartesian in nature, with sharply defined edges that do not overlap. But there are often culturally biased visions of reality implicit deep within in these categories. For example, we use terms like *dropout*, *minority*, *broken homes*, *lower class*, *working mothers*, and even *age*, *school*, and *education level* as if they are unitary phenomena, notwithstanding the fact that some intact homes are so stressful that a broken home is a welcome alternative;¹³ that a 20-year-old single mother with three children is not comparable to a 20-year-old coed;¹⁴ that schools affect individuals differently, depending on their temperaments and on family interactions (Hinde 1988, p. 375); and that some schools are so bad that the only rational act is to drop out.

These categories can be considered idealized cognitive models (or ICMs; see Lakoff 1987, p. 68ff) that contain within them stereotypes, often based on the culturally imbedded assumptions about their meaning. With regard to school, for example, is the individual's school a place where students exchange ideas—or exchange drugs; where students meet and enjoy friends—or where they cannot go to the bathroom for fear of being assaulted? One ICM may be more common in schools attended by the children of those who collect the data, the other more common in schools attended by those who furnish the data.

Crime categories also draw upon ICMs, despite the fact that we know that they are legal, not behavioral, categories. Thirty years ago Sellin and Wolfgang (1964, p. 42) underscored the fact that the variation within crime categories is almost as great as the variation among categories and developed the crime seriousness index to account for this variation. Yet we still use these categories to assess programs, predict offender dangerousness, and set sentences. Frohmann (1991) has shown how prosecutors try to fit rape cases into ICMs (or typifications), to decide whether to press charges.

We cannot call ourselves social scientists if we ignore the considerable differences within categories that putatively reflect one activity or phenomenon. This does not mean that because every individual or case is *sui generis* that generalizations cannot be made—after all, finding patterns in data is the stuff of science. But it suggests that the generalizations should be made on the basis of the known variation within the categories, rather than on the basis of the ICMs and other prior (mis)conceptions we carry with us.

Unimodality and outcomes. Another potential problem with the assumption of unimodality is associated with Martinson's (1974) conclusion that "nothing works" in corrections. This finding was based on an analysis of experiments that were evaluated, for the most part, by standard methods—that is, comparing means. Palmer (1978) pointed out, however, that almost half of the studies had positive or partly positive findings—for some types of offenders. That no treatment was shown to work across the board should have been expected—different folks may require different strokes.

Summary. The methods and interpretations currently in vogue in criminology and criminal justice research, which make implicit assumptions of unimodality, have a number of problems. They engender a predisposition to ignore multiple modes of behavior, to treat all situations as having common origins, to embrace a single cognitive model of reality, and to overlook any treatments that do not apply to the whole population. Although studies based on mean values are often quite useful, especially as a first step, they should not be considered the only focus of statistical analyses.

TOWARD BETTER DATA AND METHODS

Thus far we have seen that methods considered to be "tried and true" are often "tried and found wanting." Not always: many researchers, including myself, have employed tests of significance, sampling, regression, and model validation techniques to great benefit. But they have often been employed without regard for their limitations and have contained the implicit assumption that the data they study represent a single phenomenon. It sometimes appears that criminologists have grown accustomed to focusing only on certain types of data; and there seems to be a similar mindset on using only those methods that are readily available in statistical packages.¹⁵ It is possible to broaden our consideration of the nature of the data we collect and the way we choose to analyze them. In particular, I discuss the nature of the data we choose to collect and highlight one or two analytic methods that can be used to deal with such data.

Selecting Variables

Stark (1987, p. 894) decries the fact that much current social science research seems to have "lost touch with significant aspects of crime and delinquency. Poor neighborhoods disappeared [from the social science literature] to be replaced by individual kids with various levels of family income, but no detectable environment at all.... Yet through it all, social scientists somehow still knew better than to stroll the streets at night in certain parts of town or even to park there." There are two interesting aspects to this quote. The first relates to how and why we select variables for inclusion in our analyses. When we "round up the usual variables" to characterize an individual¹⁶ or a community,¹⁷ we may realize that we are missing some important characteristics, that we may have a one- or two-dimensional picture of the entity we are studying, but go on anyway: better to do something with the information we have available than to do nothing at all.¹⁸

Variable selection as a sampling process. We should consider why we are collecting such data in the first place: normally to characterize the important attributes of, for example, a delinquent. I suggest that we consider these data to be the output of a sampling process, not the sampling discussed earlier, but sampling in the information theoretic sense, as is involved in the digital recording of sound. In this process, the acoustic waveform generated by a musician is sampled and digitized to permit compression of the waveform for storing and transmitting it more efficiently. It is reconstituted at the receiving end, and the triumph of this technology is that the output sounds the same as the original input.

In social science, we also attempt to sample in this way, by selecting variables that we hope will fully characterize an individual. But few would say that there is sufficient information in these variables to reconstitute the individuals from which they were sampled. This is not to say that this reconstitution process is impossible: consider the skilled clinical psychologist, social worker, psychiatric diagnostician (or even novelist) who, in perhaps two pages of narrative, is able to describe an individual in sufficient detail that he or she would not only be recognizable to another observer, but one might even be able to forecast how the individual would react to different stimuli.

It is not beyond the realm of possibility, therefore, to suggest that one might be able to capture the essence of an individual (and the significant aspects of his or her life course) in a two-page narrative "sample." If narratives can do a better job of it than our standard variables, this may suggest that we need to concentrate our methodological efforts not on developing new statistical algorithms but, rather, on how to extract from such narratives their important features in ways that permit subsequent—and perhaps, eventually, even quantitative—analysis (Sampson 1993, p. 430).¹⁹

Using observational data. Another means of getting a more complete picture of a situation is to collect more meaningful types of data. For example, studies of decision making in juvenile court often consider age, race, age at first arrest, prior record, and type of offense as affecting sentence type and length. All of these data are available from official documents. Observing a juvenile hearing, however, might suggest additional variables: did the parent(s) attend, how was the juvenile dressed, what was his or her demeanor, was she or he respectful to the judge, is the judge inclined to be lenient because she or he handed down fairly harsh dispositions in earlier cases, and so on. I suspect that inclusion of variables such as these would considerably reduce the unexplained variation in outcomes. The problem with this is, of course, that there is no easy way to collect such data without an observer being there, so too often the researcher makes do with the available data instead of collecting more relevant data. But looking under the lamppost will not help us to understand the phenomena under study.

Using contextual data. The second aspect of the Stark quote points out that the context is all-important in giving an individual clues as to (in his example) a neighborhood's dangerousness. Braithwaite (1993, p. 384) stresses the overwhelming importance of context "in deciding whether a crime will occur"; and context is not necessarily handled by adding "community" variables. It is possible, however, to use data from other (i.e., noncriminal justice) agencies in other ways to understand the context of the criminal justice data. A rationale for and a means of combining such data is given in Maltz (1994).

Text can also provide context. A great deal of contextual information is carried in the choice of words, including complicity, social status, and power relationships, whether in oral histories (Bennett 1981), interviews (Briggs 1986), wiretaps (Shuy 1993), or trials (Frohmann 1991; Matoesian 1993). This information needs to be made explicit, and linguistic analysis of such narratives (e.g., presentence investigations) may be valuable for this purpose.²⁰ This would, in some sense, be a combination of clinical and statistical judgment and should be much more useful in making predictions than statistical analyses of the usual data.

Life Course Analysis and Longitudinal Data

Events and experiences occurring to an individual through his or her life course (e.g., Bennett 1981; Farrington 1988; Robins and Rutter 1990; McCord 1992; Sampson and Laub 1993) also can provide a rich source of information for analysis. It is not just the nature of the events occurring to an individual but when they occur that matters. It makes an enormous difference in an individual's propensity for criminality, for example, if she or he suffered physical abuse from a parent at age 2 or at age 14, or fell behind his or her school cohort in first grade or tenth grade. Gottfredson and Hirschi (1990) forcefully make the point that the early influences on children have strong implications for their self-control and, therefore, their potential for criminality. The saying, "As the twig is bent, so grows the tree," is relevant here: we need to develop and employ methods that describe when and how the twig is bent.

Longitudinal research is not without its own pitfalls. Collinearity is one particularly troublesome problem (McCord 1993, p. 417). Moreover, because the information characterizing an individual is arrayed over time, different expedients have been used to manipulate the data into a more manageable form. At least two different methods have been used by researchers; the first method is to "clump" the data over time, the second is to fit the data into smooth and well-behaved chronological patterns of events.

Clumping. When a number of events occur over a period of time, and the number and time period vary from person to person, a researcher would like to be able to (a) take cognizance of the variation in number and time period, and (b) relate the extent of variation to the characteristics of the individuals. One means of doing this is to count the number of events within a fixed period of time and use this as the variable describing the frequency and timing of events. This is how Farrington (1993) dealt with the variation in offenses in the group of youths he studied: variables like "number of offenses between ages 10-14" were used. Whether most of the offenses occurred at age 10 or age 14 was not noted, nor was

whether the interoffense time was increasing or decreasing, nor was the seriousness of the offenses.²¹ Similarly, in the study by Wood et al. (1993) of the effect of moving (i.e., changing residences) on youths' development and behavior, they did not look at when the moves took place in the children's lives, or the *length of time between* moves; rather, they looked at the *number* of moves.

Thus a great deal of potentially relevant information concerning the events under study is compressed into just a few categories. This is understandable from the standpoint of the researcher; she or he wants to find out, at least initially, if there is an effect and may not be overly concerned with the finer structure of the events, how the effect varies from individual to individual.

Smoothing. A second method of summarizing event data over time is to make assumptions about the chronological patterns they generate and fit them into specific patterns that are more easily manipulable. The choice of pattern most often used is the one generated by a Poisson process (Cox and Isham 1980), one in which the rate of event occurrence is on average constant. But even if the occurrence rate is constant, a given individual's pattern is not expected to be regularly spaced for the same reason that tossing a fair coin is not expected to produce a sequence of alternating heads and tails.

When the events in question are crimes or arrests, the rates are known as lambda and mu, respectively (Blumstein, Cohen, Roth, and Visher 1986), and can be estimated by dividing the number of events in a time period by the length of the time period. However, a Poisson model can fit virtually any pattern of events, so any offender's sequence of arrests can be modeled by a (constant) mu. Thus this method of summarizing chronological data does not do a good job of distinguishing between increasing and decreasing interarrest times, either, and also masks potentially relevant information.²² In fact, even though longitudinal data show much promise in terms of finding causal mechanisms in criminal behavior, the analytic methods that are brought to bear on analyzing the data are, for the most part, quite limited. One alternative to using analytic methods is to depict the events and processes occurring in a life course in a time line (e.g., Post, Roy-Byrne, and Uhde 1988; Maltz forthcoming), and let the viewer infer patterns, as Tufte (1990, p. 50) suggests (see the Displaying Data section below).

Multimodal Analysis

In studying life courses, one quickly becomes aware that single explanations often do not work. In some cases, the same stimulus (e.g., military service) will produce different responses (e.g., turning some lives

around, disrupting others' lives) and the same response (e.g., terminating a criminal career) can be produced by different stimuli (e.g., marriage, becoming a parent, employment, an inspiring teacher) (Sampson and Laub 1993, p. 222; Farrington 1993, p. 366). Thus different types of individuals react differently to external influences. Genetic and environmental factors also have a substantial influence on development. One means of distinguishing among the different types is to determine whether all members of the group being studied can be grouped together, or whether they can be clustered into different subgroups.

One method of doing this is to perform a cluster analysis of the data, which is useful in making distinctions among individuals based on the traits they share in common.²³ Cluster analysis has not been used to any great extent in criminology in the past. (Some exceptions are Hindelang and Weis [1972], Schwendinger and Schwendinger [1985], and Fagan [1989].) This seems to be changing, especially in the literature on developmental problems and delinquency, where some useful comparisons with other methods have been made. Hinde and Dennis (1986) show how categorizing cases "provides evidence for relations between variables not revealed by correlational methods"; Magnusson and Bergman (1988) show how a "systematic relationship of aggressiveness to adult criminality and to adult alcohol abuse disappeared completely" when a small group of subjects with severe adjustment problems was analyzed separately from the rest of the subjects; and Farrington, Loeber, and van Kammer (1990) show that different types of delinquency problems are part of "different causal chains or different developmental sequences leading to offending."

Whether or not the variables in a particular study reflect one or many modes is not necessarily the issue: to my mind, it is always worth testing data to see if the data appear to be multimodal or unimodal. Although it may turn out to be true that one "grand unified theory" of crime or criminal behavior will explain everything we need to know, one should not make this assumption a priori and analyze the data as if this were the case; more effort should go into testing this hypothesis.

Other Methods of Analysis

In addition, there are a number of worthwhile methods that are ignored by social scientists, although this is changing. Abbott (1988) notes that there are alternatives to the straitjacket of what he terms "general linear reality"; he cites, in particular, the methodologies associated with demography, sequential analysis, and network analysis.

Recently there have been more and more instances of new methods being brought to bear on problems of crime and justice. They include those used in ecology, epidemiology, ethnography, event history and

survival analysis, sociolinguistics, life course analysis, geography and computer mapping, markov models and processes, and simulation methods. The number of studies using these methods is growing; many have their own limitations that need not be detailed here. Textbooks that have attempted to go beyond the standard statistical techniques, including Coleman (1964) and Greenberg (1979), have not met with much success in the past; perhaps this will change in the future.

Displaying Data

As Wild (1994) notes, "We seem to be heading for an era in which the primary language for promoting the human understanding of data will be sophisticated computer graphics rather than mathematics." No one who has seen the superb examples of data displays in Tufte's (1983, 1990) books can ignore their relevance to this prediction and to the arguments presented herein. Although for the most part researchers attempt to simplify the presentation of their data, Tufte (1990, p. 37) takes the opposite approach: "Simplicity of reading derives from the context of detailed and complex information, properly arranged. A most unconventional design strategy is revealed: *to clarify, add detail*." He feels that standard statistical analyses hide too much from the analyst: to quote Tufte (1990, p. 50) again, "High-density designs also allow viewers to select, to narrate, to recast and personalize data for their own uses. The control of information is given over to viewers, not editors, designers, or decorators [or, I would add, programmers and statisticians]. Data-thin, forgetful displays move viewers toward ignorance and passivity and, at the same time, diminish the credibility of the source. Thin data rightly prompts suspicions: "What are they leaving out? Is that really everything they know? What are they hiding? Is that all they did?"

Although Tufte focuses for the most part on how information is displayed in tables and charts, his examples are instructive for statistical analysis as well. The next frontier in statistical analysis may well be in the development of better ways of displaying data, ways that would permit analysts and not algorithms to infer patterns from the data. This has been shown to be useful in analyzing geographical crime patterns (e.g., Harries 1980; Rengert and Wasilchick 1985; Roncek and Pravatiner 1991) and has also been taken up by police departments where computer-based crime mapping is being used for tactical analysis (Block 1991; Maltz, Gordon, and Friedman 1990). These display techniques can be adapted to other criminal justice areas as well, to *let the data speak for themselves* instead of being filtered through a statistical algorithm.

CONCLUSION

In this essay, I suggest that we can and should do a great deal to improve the state of research in criminology and criminal justice. Collecting a large data set, pouring it into a computer, and turning an analytic crank may provide results that are statistically significant and, perhaps, even worthy of publication, but more can be done with the data we collect to improve the state of our knowledge about crime and its correction. Part of the reason for this, I have argued, is historical—we have been conditioned to assume that a single data set will produce a single pattern that can be characterized by mean values. This implicit assumption of unimodality—one mode of behavior for the entire population under study—is also found in our assumptions about the unitary influence of various processes (e.g., schooling) on all individuals and in our search for treatments that benefit all individuals. Measures should be taken to test these assumptions. Data sets should be analyzed to see if more than one type of individual is contained within the data, if more than one mode of behavior appears to be manifested, if more than one outcome might result from a treatment.

We have made do with the assumptions and limitations of standard statistical techniques because, up until only very recently, we have had no realistic alternatives for analyzing large data sets. To handle them, we have had to employ inexact models whose primary virtue is that they are tractable. But we no longer have to "model the data." The increased availability of high-speed, high-capacity computers and excellent data analysis and graphics programs means that we can let the data speak for themselves.

Further, we should consider the possibility that the data we currently collect and the categories we currently use may be lacking, and attempt to collect different kinds of data, using more relevant categories, recognizing full well that this may require the application of judgment on the part of the data collectors. We should also do more research into how we can extract data from the richer sources that are at our disposal, the narrative accounts that provide a much clearer picture of offenders, offenses, and communities than do the standard types of data that are used in criminological research. We will go much further toward understanding the problems of crime if we relinquish our death-grip on our old beliefs and open ourselves to new methods and paradigms.

The goal of this essay is to provoke thought and debate, preferably in that order. It will be of little value if it does not promote discussion. But it is my hope that we begin to make more penetrating observations of the entire research process, what we choose to collect and why, and what we do with what we collect.

NOTES

1. Note that this interpretation argues for determinism and against free will; as Hacking (1990, p. 116) puts it, "if there were statistical laws of crime and suicide, then criminals could not help themselves."
2. Hacking (1990, p. 120) suggests that the movement toward eugenics was a direct consequence of this statistical interpretation of variation. It is interesting to speculate on the impact this had on the development of the concept of the master race in Nazi Germany, in which genetic purity was to be striven for.
3. The assertion that statistically based judgments are better than clinically based assessments may not hold to the same extent now that it did when it was first reported. In the past, clinicians were rarely given useful information on outcomes that would be necessary for them to gauge the consequences of their decisions; one of the major deficiencies in the criminal justice system is this lack of feedback (e.g., to judges). Furthermore, even when statistical feedback is provided, Gottfredson and Gottfredson (1988) showed that it is not a foregone conclusion that the information will be used. More recent studies (Fong, Lurigio, and Stalans 1990; Bunn and Wright 1991), however, have shown that the judicious combination of judgmental and statistical information can be used to improve the ability of clinicians to forecast outcomes.
4. In a Type I error, the null hypothesis is falsely rejected (a false alarm); in a Type II error, the null hypothesis is falsely accepted (a missed detection). A related factor, statistical power (Cohen 1988, 1990; Weisburd 1993), is useful in determining the extent to which a statistically significant finding is practically significant.
5. Mosteller notes that evidence as to the efficacy of citrus fruit in curing scurvy was first demonstrated in 1601, 146 years earlier, and the British Navy did not implement the findings until 1795, 48 years after the experiment!
6. Bayesian statistics would provide more useful inferences in these cases.
7. Alfred Blumstein, personal communication. It is with some degree of ambivalence that I continue to use the word "significance" herein, but the word is too deeply imbedded in the literature to be extricated at this point.
8. The mean for the experimental sample is 9,927 and standard error of the sample means, for $N = 350$, is 307; for the control sample the mean is 9,887 and the standard error of the sample means, again for $N = 350$, is 256.
9. According to Freedman (1983, p. 152), this often occurs "in a context where substantive theory is weak. To focus on an extreme case, suppose that in fact there is no relationship between the dependent variable and the explanatory variables. Even so, if there are many explanatory variables, the R^2 will be high. If explanatory variables with small t statistics are dropped and the equation refitted, the R^2 will stay high and the overall F will become highly significant."
10. One way of obtaining an estimate of this is to speak with police detectives, to find out the degree of connectivity they see among offenses. Although they are not statisticians, they do have a feel for this; they often see a burglary, purse-snatching, or robbery problem diminish drastically with the arrest of one or two individuals.
11. Of course, regression methods can accommodate heteroskedasticity because they are "robust," but rarely is the extent of heteroskedasticity reviewed to see if it is possible that more than one phenomenon is included in the data.
12. I do not intend this to be a blanket indictment of all who use statistical packages; much of the research employing them is insightful and reliable. No one can dispute, however, that many studies apply statistical methods without due concern for their applicability.
13. As McCord (1990, p. 120) notes, "living in discordant homes with two parents could be more damaging than living in homes with solo mothers," and Farrington (1993, p. 388) states, "early separation from a parent (usually a father) was harmful to boys from average or high-income families but beneficial to boys from low-income families."
14. Rutter (1989) points out that age reflects at least four developmental components: cognitive level, biological maturity, duration of experiences, and types of experiences—in addition to biological and chronological components (Sampson and Laub 1992, p. 81; 1993, p. 253).
15. As regression, then factor analysis, then LISREL, then structured equation modeling found their way into the more popular statistical packages, it was interesting to see the proliferation of studies that used these techniques. The pioneers may have found problems for which the methods were uniquely apt, but

those who followed often used them without really considering whether they were appropriate—see Blalock (1991, p. 329).

16. For example, race, age, SES, family income, education level, prior record, age at first arrest, number of adults in the household, number of siblings.

17. For example, percentage owner-occupied housing, percentage single-parent households, unemployment rate, racial/ethnic composition, income distribution, age distribution.

18. The list of social indicator variables suggested by Felson (1993, p. 407) also speaks to the issue of broadening the type of information we consider relevant.

19. Narratives are helpful in other criminal justice contexts as well. Detectives maintain that they need to read case narratives in order to understand the nature of the offense; see Maltz et al. (1990, p. 102). Perhaps their cognitive models of crimes would be better as a means of categorizing crimes than the legal categories we now use.

20. In fact, analysis of trial transcripts may give a misleading impression. Conversational analysis focuses on not just the words and their meanings but on the pauses and intonations in the speech events (Gumperz 1982). Such verbal and nonverbal cues are sensed (and perhaps acted upon) by judges and jurors but do not find their way into trial transcripts.

21. This is not to say that researchers do not first inspect the data to determine which variables might best characterize the data. In other words, the category "Offenses 10- 14" may have been chosen because it characterized the cohort's behavior better than, say, "Offenses 11-15" or some other means of aggregating the data.

22. Of course, the value of a method depends on the uses to which it is to be put. If the purpose is to analyze the flow of offenders through a system (i.e., for criminal justice research purposes), then the Poisson approximation may be more than adequate. If, however, the purpose is to understand why some offenders terminate their offending careers earlier than others (i.e., for criminology research purposes), then this approximation may be less than adequate.

23. Cluster analysis should be distinguished from a related method, factor analysis. In factor analysis, data are analyzed to determine how variables can be classified into groups; whereas, in cluster analysis, data are analyzed to determine how individuals can be classified into groups.

REFERENCES

- Abbott, Andrew. 1988. "Transcending Linear Reality." *Sociological Theory* 6:169-86.
- Bennett, James. 1981. *Oral History and Delinquency: The Rhetoric of Criminology*. Chicago: University of Chicago Press.
- Berk, Richard A. 1991. "Toward a Methodology for Mere Mortals." Pp. 315-24 in *Sociological Methodology*, edited by Peter V. Marsden. Oxford, UK: Basil Blackwell.
- Berk, Richard A., Kenneth H. Lenihan, and Peter H. Rossi. 1980. "Crime and Poverty: Some Experimental Evidence From Ex-Offenders." *American Sociological Review* 45:766-86.
- Blalock, Hubert M., Jr. 1991. "Are There Really Any Constructive Alternatives to Causal Modeling?" Pp. 325-35 in *Sociological Methodology*, edited by Peter V. Marsden. Oxford, UK: Basil Blackwell.
- Block, Carolyn R. 1991. "Early Warning System for Street Gang Violence Crisis Areas." Proposal to the Bureau of Justice Statistics. Chicago: Illinois Criminal Justice Information Authority.
- Blumstein, Alfred, Jacqueline Cohen, Jeffrey A. Roth, and Christy A. Visher, eds. 1986. *Criminal Careers and "Career Criminals"*. Vols. 1 and 2. Washington, DC: National Academy of Sciences.
- Braithwaite, John. 1993. "Beyond Positivism: Learning From Contextual Integrated Strategies." *Journal of Research in Crime and Delinquency* 30:383-99.
- Briggs, Charles L. 1986. *Learning How to Ask: A Sociolinguistic Appraisal of the Role of the Interview in Social Science Research*. Cambridge, UK: Cambridge University Press.
- Bunn, Derek and George Wright. 1991. "Interaction of Judgemental and Statistical Forecasting: Methods, Issues & Analysis." *Management Science* 37:501-18.

- Chaiken, Jan M. and Marcia R. Chaiken. 1982. *Varieties of Criminal Behavior*. Report R-2814NIJ. Santa Monica, CA: RAND.
- Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum.
- , 1990. "Things I Have Learned (So Far)." *American Psychologist* 45:1304-12.
- Coleman, James S. 1964. *Introduction to Mathematical Sociology*. New York: Free Press.
- , 1990. *Foundations of Social Theory*. Cambridge, MA: Belknap/Harvard University Press.
- Cox, David R. and Valerie Isham. 1980. *Point Processes*. London: Chapman and Hall.
- Duncan, Otis Dudley. 1984. *Notes on Social Measurement: Historical and Critical*. New York: Russell Sage.
- Ehrlich, Isaac. 1975. "The Deterrent Effect of Capital Punishment: A Question of Life and Death." *American Economic Review* 65:397-417.
- Fagan, Jeffrey. 1989. "The Social Organization of Drug Use and Drug Dealing Among Urban Gangs." *Criminology* 27(4):633-70.
- , 1993. "Editor's Introduction." *Journal of Research in Crime and Delinquency* 30:381-2.
- Farrington, David P. 1988. "Studying Changes Within Individuals: The Causes of Offending." Pp. 158-83 in *Studies in Psychosocial Risk: The Power of Longitudinal Data*, edited by Michael D. Rutter. Cambridge, UK: Cambridge University Press.
- , 1993. "Interactions Between Individual and Contextual Factors in the Development of Offending." Pp. 366-89 in *Adolescence in Context*, edited by R. Silbereisen and E. Todt. New York: Springer-Verlag.
- Farrington, David P, Rolf Locher, and Welmoet B. van Kammer. 1990. "Long-Term Criminal Outcomes of Hyperactivity-Impulsivity-Attention Deficit and Conduct Problems in Childhood." Pp. 62-81 in *Straight and Devious Pathways From Childhood to Adulthood*, edited by Lee N. Robins and Michael Rutter. Cambridge, UK: Cambridge University Press.
- Faust, David. 1984. *The Limits of Scientific Reasoning*. Minneapolis: University of Minnesota Press.
- Felson, Marcus. 1993. "Social Indicators for Criminology." *Journal of Research in Crime and Delinquency* 30:400-11.
- Fisher, Ronald A. 1935. *The Design of Experiments*. Edinburgh, Scotland: Oliver and Boyd.
- Fong, Geoffrey T., Arthur J. Lurigio, and Loretta Stalans. 1990. "Improving Probation Decisions Through Statistical Training." *Criminal Justice and Behavior* 17:370-88.
- Freedman, David A. 1983. "A Note on Screening Regression Equations." *American Statistician* 37:152-55.
- , 1985. "Statistics and the Scientific Method." Pp. 343-66 in *Cohort Analysis in the Social Sciences: Beyond the Identification Problem*, edited by William M. Mason and Stephen E. Fienberg. New York: Springer-Verlag.
- , 1991 a. "Statistical Models and Shoe Leather." Pp. 291-313 in *Sociological Methodology*, edited by Peter V. Marsden. Oxford, UK: Basil Blackwell.
- , 1991b. "A Rejoinder to Berk, Blalock, and Mason." Pp. 353-8 in *Sociological Methodology*, edited by Peter V. Marsden. Oxford, UK: Basil Blackwell.
- Freedman, David A., Robert Pisani, Roger Purves, and Ani Adhikari. 1991. *Statistics*. 2d ed. New York: Norton.
- Frohmann, Lisa. 1991. "Discrediting Victims' Allegations of Sexual Assault: Prosecutorial Accounts of Case Rejections." *Social Problems* 38:213-26.
- Gigerenzer, Gerd. 1987. "Probabilistic Thinking and the Fight Against Subjectivity." Chapter I in *The Probabilistic Revolution*. Vol. 1, *Ideas in History*, edited by Lorenz Krüger, Lorraine J. Daston, and Michael Heidelberger. Cambridge, MA: MIT Press.
- Gottfredson, Michael R. and Don M. Gottfredson. 1988. *Decision Making in Criminal Justice: Toward the Rational Exercise of Discretion*. New York: Plenum.
- Gottfredson, Michael R. and Travis Hirschi. 1990. *A General Theory of Crime*. Stanford, CA: Stanford University Press.

- Greenberg, David F. 1979. *Mathematical Criminology*. New Brunswick, NJ: Rutgers University Press.
- Gumperz, John. 1982. *Discourse Strategies*. Cambridge, UK: Cambridge University Press.
- Hacking, Ian. 1990. *The Taming of Chance*. Cambridge, UK: Cambridge University Press.
- , 1991. "How Shall We Do the History of Statistics?" Pp. 181-95 in *The Foucault Effect: Studies in Governmentality*, edited by Graham Burchell, Colin Gordon, and Peter Miller. Chicago: University of Chicago Press.
- Harries, Keith D. 1980. *Crime and the Environment*. Springfield, IL: Charles C Thomas.
- Hinde, Robert A. 1988. "Continuities and Discontinuities: Conceptual Issues and Methodological Considerations." Pp. 367-83 in *Studies in Psychosocial Risk: The Power of Longitudinal Data*, edited by Michael Rutter. Cambridge, UK: Cambridge University Press.
- Hinde, Robert A. and Amanda Dennis. 1986. "Categorizing Individuals: An Alternative to Linear Analysis." *International Journal of Behavioral Development* 9:105-19.
- Hindelang, Michael J. 1976. "With a Little Help From Their Friends: Group Participation in Reported Delinquent Behavior." *British Journal of Criminology* 16:109-25.
- Hindelang, Michael J. and Joseph G. Weis. 1972. "Personality and Self-Reported Delinquency: An Application of Cluster Analysis." *Criminology* 10:268-94.
- Lakoff, George. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. Chicago: University of Chicago Press.
- Lieberson, Stanley. 1985. *Making It Count: The Improvement of Social Research and Theory*. Berkeley: University of California Press.
- Magnusson, David and L. R. Bergman. 1988. "Individual and Variable-Based Approaches to Longitudinal Research on Early Risk Factors." Pp. 45-61 in *Studies in Psychosocial Risk: The Power of Longitudinal Data*, edited by Michael Rutter. Cambridge, UK: Cambridge University Press.
- Malinvaud, E. 1970. *Statistical Methods of Econometrics*. New York: North-Holland/American Elsevier.
- Maltz, Michael D. 1990. *Measuring the Effectiveness of Organized Crime Control Efforts*. Chicago: Office of International Criminal Justice, University of Illinois at Chicago.
- , Forthcoming. "Space, Time, and Crime: Operationalizing Dynamic Contextualism." In *Crime and Place*, edited by John Eck and David Weisburd. Monsey, NY: Criminal Justice Press.
- Maltz, Michael D., Andrew C. Gordon, and Warren Friedman. 1990. *Mapping Crime in Its Community Setting: Event Geography Analysis*. New York: Springer-Verlag.
- Martinson, Robert. 1974. "What Works? Questions and Answers About Prison Reform." *The Public Interest*, Spring, 22-54.
- Mason, William M. 1991. "Freedman Is Right as Far as He Goes, but There Is More, and It's Worse. Statisticians Could Help." Pp. 337-51 in *Sociological Methodology*, edited by Peter V. Marsden. Oxford, UK: Basil Blackwell.
- Matoesian, Gregory M. 1993. *Reproducing Rape: Domination Through Talk in the Courtroom*. London, UK: Polity Press, and Chicago: University of Chicago Press.
- McCord, Joan. 1990. "Long-Term Perspectives on Parental Absence." Pp. 116-34 in *Straight and Devious Pathways From Childhood to Adulthood*, edited by Lee N. Robins and Michael Rutter. Cambridge, UK: Cambridge University Press.
- , 1992. "The Cambridge-Somerville Study: A Pioneering Longitudinal Experimental Study of Delinquency Prevention." Pp. 196-206 in *Preventing Antisocial Behavior: Interventions From Birth to Adolescence*, edited by Joan McCord and Richard E. Tremblay. New York: Guilford.
- , 1993. "Descriptions and Predictions: Three Problems for the Future of Criminological Research." *Journal of Research in Crime and Delinquency* 30:412-25.
- Meehl, Paul. 1954. *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. Minneapolis: University of Minnesota Press.
- Mosteller, Frederick. 1981. "Innovation and Evaluation." *Science* 211:881-6.
- Palmer, Ted. 1978. *Correctional Intervention and Research: Current Issues and Future Prospects*. Lexington, MA: Lexington Books.

- Porter, Theodore M. 1986. *The Rise of Statistical Thinking, 1820-1900*. Princeton, NJ: Princeton University Press.
- Post, Robert M., Peter P. Roy-Byrne, and Thomas W. Uhde. 1988. "Graphic Representation of the Life Course of Illness in Patients With Affective Disorder." *American Journal of Psychiatry* 145:844-8.
- Reiss, Albert J., Jr. 1986. "Co-offending Influences on Criminal Careers." Pp. 121-60 in *Criminal Careers and "Career Criminals"*, edited by Alfred Blumstein, Jacqueline Cohen, Jeffrey A. Roth, and Christy A. Visher. Vol. 2. Washington, DC: National Academy of Sciences.
- , 1988. "Co-offending and Criminal Careers." In *Crime and Justice: A Review of Research*. Vol. 10, edited by Michael Tonry and Norval Morris. Chicago: University of Chicago Press.
- Reiss, Albert J., Jr. and David P. Farrington. 1991. "Advancing Knowledge About Co-Offending: Results From a Prospective Survey of London Males." *Journal of Criminal Law and Criminology* 82:360-95.
- Rengert, George and John Wasilchick. 1985. *Suburban Burglary: A Time and a Place for Everything*. Springfield, IL: Charles C Thomas.
- Robins, Lee N. and Michael Rutter, eds. 1990. *Straight and Devious Pathways From Childhood to Adulthood*. Cambridge, UK: Cambridge University Press.
- Roncek, Dennis W. and Mitchell A. Pravatiner. 1989. "Additional Evidence That Taverns Enhance early Crime." *Sociology and Social Research* 73:185-8.
- Rossi, Peter H. and Richard A. Berk. 1982. "Saying It Wrong With Figures: A Comment on Zeisel." *American Journal of Sociology* 88:390-6.
- Rossi, Peter H., Richard A. Berk, and Kenneth J. Lenihan. 1980. *Money, Work and Crime: Some Experimental Results*. New York: Academic Press.
- , 1989. "Age as an Ambiguous Variable in Developmental Research: Some Epidemiological Considerations From Developmental Psychopathology." *International Journal of Behavioral Development* 12:1-34.
- Sampson, Robert J. 1993. "Linking Time and Place: Dynamic Contextualism and the Future of Criminological Inquiry." *Journal of Research in Crime and Delinquency* 30:426-44.
- Sampson, Robert J. and John H. Laub. 1992. "Crime and Deviance in the Life Course." *Annual Review of Sociology* 18:63-84.
- , 1993. *Crime in the Making: Pathways and Turning Points Through Life*. Cambridge, MA: Harvard University Press.
- Schwendinger, Herman and Julia R. Siegel Schwendinger. 1985. *Adolescent Subcultures and Delinquency*. Research edition. New York: Praeger.
- Sellin, Thorsten and Marvin E. Wolfgang. 1964. *The Measurement of Delinquency*. New York: Wiley.
- Shaw, Clifford and Henry McKay. 1969. *Juvenile Delinquency and Urban Areas*. Chicago: University of Chicago Press.
- Shuy, Roger W. 1993. *Language Crimes: The Use and Abuse of Language Evidence in the Courtroom*. Cambridge, UK: Blackwell.
- Stark, Rodney. 1987. "Deviant Places: A Theory of the Ecology of Crime." *Criminology* 25:893-909.
- Stigler, Stephen M. 1986. *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge, MA: Belknap/Harvard University Press.
- Tufte, Edward. 1983. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.
- , 1990. *Envisioning Information*. Cheshire, CT: Graphics Press.
- Waring, Elin. 1992. "Co-Offending in White Collar Crime: A Network Approach." Unpublished doctoral dissertation, Department of Sociology, Yale University.
- Weisburd, David, with Anthony Petrosino and Gail Mason. 1993. "Design Sensitivity in Criminal Justice Experiments: Reassessing the Relationship Between Sample Size and Statistical Power." Pp. 337-79 in *Crime and Justice*. Vol. 17, edited by Norval Morris and Michael Tonry. Chicago: University of Chicago Press.
- Wild, C. J. 1994. "Embracing the 'Wider View' of Statistics." *The American Statistician* 48:163-71.
- Wolfgang, Marvin E., Robert M. Figlio, and Thorsten Sellin. 1972. *Delinquency in a Birth Cohort*. Chicago:

University of Chicago Press.

- Wood, David, Heal Halfon, Debra Scarlata, Paul Newacheck, and Sharon Nessim. 1993. "Impact of Family Relocation on Children's Growth, Development, School Function, and Behavior." *Journal of the American Medical Association* 270(11):1334-8.
- Zeisel, Hans. 1982a. "Disagreement Over the Evaluation of a Controlled Experiment." *American Journal of Sociology* 88:378-89.
- , 1982b. "Hans Zeisel Concludes the Debate." *American Journal of Sociology* 88:394-6.
- Zimring, Franklin E. 1981. "Kids, Groups and Crime: Some Implications of a Well-Known Secret." *Journal of Criminal Law and Criminology* 72:867-85.