3

Quantitative Methods

Russell K. Schutt

It is hard to overemphasize the importance of quantitative methods in sociology and impossible to describe all of the specific methods that can reasonably be termed "quantitative." Given its importance and scope, it is tempting to enter the universe of quantitative methods and never look back, outside, or into the future. There is more than a lifetime of quantitative methods to review. But to ignore the history and philosophy of quantitative methods, to overlook current debates and emerging trends, is to fail to understand the nature of quantitative methods. By understanding how quantitative methods arose, what problems they have sought to solve and how well they have done so, and what new directions have emerged, we can better understand quantitative methods themselves. So before reviewing different quantitative methods and the problems they have both solved and overlooked, this essay will begin with a review of the history and philosophy of quantitative methods and an overview of the major goals and strategies by which quantitative methods can be distinguished. In addition, after reviewing specific techniques of each type, selected new developments will be considered.

History and Philosophy

Quantitative methods are a collection of techniques that rely on numbers to represent empirical reality and that presume a positivist philosophy in which it is presumed that the social world is knowable by observers who quantify its characteristics. Quantitative methods were part of sociology at its inception, infusing the discipline with the prevailing spirit of discovery and situating it within the historical advance of science. That advance can fairly be linked to the Renaissance, with its willingness to challenge prevailing orthodoxies and its spirit of discovery.

The Wiley-Blackwell Companion to Sociology, First Edition. Edited by George Ritzer.

For the purpose of understanding the subsequent privileged role of quantitative methods in the advance of science, one event stands out above all others. With the development and (nearly) posthumous publication of his heliocentric theory of the solar system, Nicolaus Copernicus shattered the prevailing assumptions that either common sense or traditional beliefs were reliable guides to understanding the world in which we live (Tarnas 1991). Copernicus's *De Revolutionibus* ([1543] 1978), by contrast, demonstrated that years of careful quantitative measurement of the physical world and commitment to a theory that best explained those measurements could reveal fundamental principles about how the world operates – socially as well as physically – irrespective of the preferences of its human occupants (Adamczewski 1974).

It was not that Copernicus's theory immediately transformed understanding, for it was almost two centuries before religious opposition finally abated. What Copernicus succeeded in doing was to change the focus of debate about the physical world from what was most consistent with Church teachings or prevailing philosophies to what best predicted observable phenomena. The positivist spirit had triumphed; the Scientific Revolution had begun and quantitative methods were its foundation (Tarnas 1991: 248–71).

The equation of science with quantitative methods did not diminish for 300 years. Speaking to the Institution of Civil Engineers in London in 1883, William Thomson Kelvin explained:

> I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of Science, whatever the matter may be. (Scripture 1892: 127)

Sir Francis Galton translated Lord Kelvin's enthusiasm into a simple maxim for scientists: "Whenever you can, count" (Newman 1956: 1169). Social scientists were advised to resist the tendency "to rest contented with merely qualitative results where quantitative measurements could be made with the exercise of brains and patience" (Scripture 1892: 127).

Neither the triumph of the scientific worldview in the West nor its expression in quantitative methods was universal or permanent. Philosophers were troubled by the realization that human observations of the world could not be free of imposed conceptual judgments, that causal effects could not actually be observed (Tarnas 1991: 368). Certainty that empirical investigation in general and quantitative methods in particular could yield a verifiable understanding of the natural world also began to erode in the twentieth century due to the advance of science itself. Einstein's (1921) recognition of the interchangeability of matter and energy and of the relativity of space and time (Calder 1979), as well as Heisenberg's identification in the subatomic world of the impossibility of measuring simultaneously the position and momentum of a particle – the famous "uncertainty principle" – forever called into question the belief that quantitative measurement provides the necessary foundation for explaining natural phenomena (Reece 1977). Thomas Kuhn's (1970) reconceptualization of scientific progress as successive "revolutions" that overthrow

prevailing paradigms further undermined the positivist belief that scientific methods were gradually increasing understanding of the world as it "really is."

These professional developments and philosophical debates both laid the foundation for quantitative methods in sociology and infused its subsequent construction. Two trends emerged. First, positivists continued to extend the reach of quantitative methods so as to describe an ever larger fraction of the social world in ways that reflected growing awareness of its complexity. Their efforts often took account of less quantifiable social phenomena and accepted a "post-positivist" assumption that the observations – including numbers – recorded by human beings inevitably will reflect to some extent the subjective orientations of the observers, but they remained committed to the core positivist belief in a knowable, objective external reality.

Relativist challenges to positivism encouraged some qualitative researchers to adopt an alternative "interpretivist" approach that focused on understanding the meanings people attached to their experiences and abjured the concept of an objective "real world." Quantitative methodologists rejected this trend, but some began to incorporate qualitative techniques, developing "mixed methods" that promised to yield greater insight into subjective as well as objective social phenomena.

Goals and Strategies

Specific quantitative research methods can be classified by the primary goal they are designed to achieve and by the general strategy they employ.

Research goals

A positivist perspective on the social world privileges one paramount goal for quantitative research: to understand the external world as it "really" is. This is the goal of validity and it presumes that each methodological technique should be evaluated by its ability to reveal empirical phenomena without distortion. It has three aspects – measurement validity, generalizability, and causal validity – which themselves each subsume multiple quantitative techniques (Schutt 2009).

Measurement validity is achieved when data collected with the specified operational procedures reflect the empirical status of or variation in the phenomenon that the measure was designed to capture. Measurement validity is the cornerstone for quantitative methods, since unless the researcher has measured what he or she thinks has been measured, all analyses and statements based on those measures will mischaracterize empirical reality. Developing valid measurement begins with conceptualization – defining clearly what is meant by the term of interest – and then continues with operationalization – the process of specifying the operations that will yield empirical evidence of variation in a particular aspect of that concept.

Generalizations about findings based on a sample of observed units are valid if they apply to the larger population or collection of units about which they are made. Of course this aspect of validity is not a concern if we have collected data about every unit of interest and make no attempt to generalize findings about these units to some larger population. But this limited purpose is rarely the case in quantitative research. Whether they have collected data from a small group of college students,

a large sample of employed persons, or a census of an entire nation, researchers usually want to draw conclusions about larger populations, a more complete collection of units, and sometimes other times and nations than those from which data were actually obtained. The likelihood of such generalizations being valid is related to the specific methods used.

Causal validity is achieved when conclusions about causal effects – what causes variation in the empirical phenomenon of interest – reflect the operation in empirical reality of the influence understood as causal. So, for example, the conclusion that receiving more education results in higher income would be causally valid if, among the persons studied, those who received more education earned higher income as a result. By contrast, the same conclusion would be invalid if higher parents' education led to higher respondents' education and to higher respondents' income, but higher respondents' education itself had no bearing on respondents' income once parents' education was taken into account.

Research strategies

Specific quantitative methods also differ in their research strategy. A *deductive* research project begins with a formal theory: a set of logically interrelated propositions about empirical reality. A specific hypothesis – a tentative statement about empirical reality involving the relations among two or more variables – is deduced from that theory and then tested. If the results of the test support the hypothesis, then the theory from which the hypothesis was deduced is considered to be on stronger ground. If the test does not support the hypothesis, then the theory from which the hypothesis was deduced is considered to have been weakened. This *hypothetico-deductive* mode of inquiry is considered by many to be the quintessential scientific method.

An *inductive* research project differs from a deductive project first and foremost in terms of its starting point. Inductive research begins with observations about empirical reality – observations that may include both empirical phenomena and relations between them. An explanatory framework is then induced from what has been observed. Many inductive research projects stop at this point, but some generate new propositions that are subsequently tested using a deductive strategy.

Purely *descriptive* research involves a more limited research strategy. If the goal of a research project is only to describe the empirical phenomena of interest, the project may be concluded when the phenomena of interest have been measured and the findings reported. No attempt is made to relate the observations to a general explanatory framework or even to evaluate the support for any specific predictions.

Although these three strategies can be viewed as alternatives, they may be combined in research projects. In fact, most quantitative research projects include an element of descriptive research, since what has been measured is often reported as a description of the setting and/or people studied. In addition, many quantitative projects designed with a deductive research strategy add an inductive component as the researchers try to make sense of unanticipated patterns they observe in their data.

Specific Quantitative Methods

Specific quantitative methods are presented in the following sections in relation to the aspect of validity on which they focus. After the aspect of validity is introduced, the problem that achieving this aspect of validity poses for researchers is discussed and the traditional solution or solutions to this problem offered by specific quantitative methods are explained. Finally, new directions being taken to achieve the goal are highlighted. The new approaches often reflect more of an inductive strategy as compared to the deductive strategy underlying the traditional solutions.

Measurement

Quantitative measurement was a key element in the Scientific Revolution and many of that era's scientists as well as subsequent historians emphasized its singular importance: "it is possible that the deepest meaning and aim of . . . the whole scientific revolution of the seventeenth century . . . is just to abolish the world of the 'more or less,' the world of qualities and sense perception . . . and to replace it by the . . . universe of precision, of exact measures" (Koyré 1965: 4–5). Three hundred years after the Scientific Revolution, still associating a positivist philosophy with enthusiasm about quantitative measurement, Nobel laureate Max Planck declared, "An experiment is a question which science poses to nature, and a measurement is the recording of Nature's answer" (1949: 110).

Hubert M. Blalock, Jr. (1982: 7), one of sociology's foremost quantitative methodologists in the late twentieth century, tied sociological research methods directly to this tradition: "there is a sufficient number of commonalities so that we can hardly afford to throw away whatever success models may be available to us." He defined measurement as "the general process through which numbers are assigned to objects in such a fashion that it is also understood just what kinds of mathematical operations can legitimately be used, given the nature of the physical operations that have been used to justify or rationalize the assignment of numbers to objects" (Blalock 1982: 11).

The problem

It is at the point of measurement that fundamental differences between the physical and social sciences are most apparent. Whereas in the physical sciences, the properties of measured phenomena – whether atoms of carbon or cords of wood – are often homogeneous, key phenomena in the social sciences are notoriously variable – whether people in different cultures, or people in the same culture at different times. As a result, it cannot be assumed that a measure will perform in the same way across different studies. The complications that this measurement heterogeneity creates are legion (Blalock 1982: 17–20).

In practical terms, recognizing the challenge posed by the goal of measurement validity has meant focusing attention on the inability of the researcher to observe directly what he or she seeks to measure. The problem that this creates is represented in classical test theory as the difference between the variation captured by a specific measure and the underlying phenomenon – the true score or latent variable – that

it is designed to measure. In algebraic form, this problem is captured as the size of the error term in the following classical equation:

$$\text{Observed Score} = \text{True Score} + \text{Error}$$

The larger the error term relative to the true score, the less the observed score tells us about the phenomenon of interest. Since there is no *a priori* way to distinguish the true score from the error component of an observed score, and since there are many sources of error, from systematic bias in response to misleading questions to random variation in response to unclear terms, improving and confirming measurement validity – that an observed score reflects what it is intended to measure (the true score) – becomes a central methodological concern (Viswanathan 2005).

In survey research, for example, to generate a quantitative measurement by asking a single question is to do what comes naturally in the course of conversation and thus to engage in what might seem an effortless task. Yet the inherent appeal of this straightforward measurement approach is also its Achilles' heel, for too many survey instruments are constructed without considering the many ways in which questions can be poorly constructed and the multiple adverse consequences of poor question design.

The problem with approaching the design of survey questions as if it was no more demanding than formulating conversational questions is that every survey respondent must interpret each survey question in the same way. If questions or their response choices include ambiguous words or phrases, or convey biased sentiments, the odds increase that different respondents will interpret them differently. The more heterogeneous the sample of survey respondents, the greater the problem is likely to be.

The solution

How can the size of the error term be reduced relative to that of the true score component in obtained measures? In survey research, question wording can be improved through pretesting strategies and systematic experimentation. Questions designed to measure the same concept can also be combined so that the resulting composite index scores represent better the corresponding latent trait, or true score.

Single questions

Pretesting methods can reveal problems with question wording and response choices. Simply asking experienced interviewers to administer a planned questionnaire to a small number of respondents similar to those to be included in a study can identify difficulties that cause hesitancy or confusion. A more systematic technique such as cognitive interviewing can reveal the extent to which potential respondents may adduce different meanings from key questions than the meanings intended by the researcher. In cognitive interviewing, the interviewer asks a question of one respondent and then probes with follow-up questions to elucidate the respondent's interpretation of the question and response choices (Schaeffer & Presser 2003: 82). Respondents may also be asked to "think aloud" as they answer questions in order to determine their mental operations as they formulate an answer.

Pretesting methods should result in greater question clarity through a process of successive refinement of question wording, but they are not a panacea for survey design. Research on cognitive interviewing and other pretesting methods indicates inconsistent results. There is no one method that seems most able to identify problems with questions and so there is no assurance that any particular method will eliminate most of the problems in question design or suggest changes that actually will improve the subsequent responses (Presser, Couper, Lessler, et al. 2004). Although much has been accomplished in improving methods of pretesting, progress continues to be inadequate.

Systematic experiments by Stanford's Jon Krosnick (1999) and others have advanced considerably awareness of the consequences of different ways of wording and organizing key questions and their response choices. The resulting improved guidelines range from choices for wording responses to achieve equal intensity intervals between them to cautions against using so-called Likert-style response choices ("strongly agree," "agree," "disagree," "strongly disagree") which elicit agreement bias and so lead to a 10–20 percent overestimate of the level of agreement.

Indexes

Writing clear and unambiguous questions is a sufficient challenge to provide limitless headaches for survey researchers. Nonetheless, it is only the starting point for concern when designing question-based measures. Single questions are simply not sufficient for measuring many of sociology's most important concepts. How satisfied are respondents with their marriage? To what extent are citizens alienated from conventional politics? What is the level of delinquency in a sample of juveniles? To what extent do respondents support abortion rights? For these and many other abstract concepts, no single question will suffice as an adequate measure.

An *index* combines responses to questions or other measures that are each indicators of a common concept. The logic of index construction from the standpoint of classical test theory is that a person's test (or index) score represents some combination of their true score, or value on the underlying latent variable, and an error component. Given assumptions about the error term (that errors will be normally distributed and independent of each other and the true scores), the more items included in an index, the more the errors will cancel each other out and the closer the average score will be to the subject's true score (Viswanathan 2005: 16–18).

The traditional solution to the problem of measurement can end at this point, following the construction of a multi-question index with a test for the measure's reliability – a test-retest evaluation or an inter-item correlation test – and for its validity, using a criterion validation or a construct validation approach (Schutt 2009: 132–41). However, index construction often includes an inductive stage in which the researcher analyzes inter-item correlations for evidence of multiple dimensions within the index. For example, an index measuring job satisfaction may include some questions that focus on intrinsic sources of satisfaction and so are more highly inter-correlated and other questions that focus on extrinsic sources of satisfaction. Different approaches to identifying subdimensions include exploratory factor analysis, multidimensional scaling, and cluster analysis (Jacoby 1991).

New directions

Classical test theory provides a conceptual framework within which measures can be developed, tested, and refined. Such work continues to dominate measurement textbooks, research protocols, and review standards. However, belief in the value of Classical Test Theory (CTT) rests on acceptance of the reasonableness of its assumptions. With respect to the error term, these assumptions require that the expected value of the error terms is zero, that they are uncorrelated with the true scores, that they are uncorrelated over successive observations, and that they are uncorrelated with all other variables in whatever system of equations is being investigated (such as a multiple regression model) (Blalock 1982: 31). Are these assumptions reasonable? Some investigators have concluded that they are not and so have turned to alternative approaches. The most prominent alternative approach is item response theory (IRT), with Rasch models as the most common such technique (Embretson & Hershberger 1999).

IRT focuses attention on differences between items in a test or index. Rather than treating all items as equivalent and creating a total score by adding them together, IRT distinguishes items in terms of their difficulty with respect to the latent trait that is being measured. The expectation is that fewer respondents will be able to pass more difficult items (or, similarly, fewer respondents will answer positively to items at a more extreme position on the underlying trait). Estimating a respondent's trait level thus depends on the characteristics of the items included in a test or index.

If items can be graded by their difficulty, testing can then be made more efficient by varying the items presented depending on responses to preceding items. There is no point in asking more "easy" questions of a respondent who has already answered a more difficult question; more accurate measurement will come from presenting such a respondent with more questions that are difficult to answer. Conversely, presenting difficult questions to a respondent who has difficulty answering easier questions is pointless (Embretson & Hershberger 1999: 9–10).

Computerized testing facilitates measurement based on IRT principles by making it easier to use adaptive testing, in which the presentation of items varies with responses to preceding items. As a result, fewer questions can be asked yet still provide a reliable assessment of the respondent's level on the latent trait. In contrast to CTT, which leads to the conclusion that longer tests are more reliable, shorter IRT-based tests can be more reliable because items are selected at the appropriate difficulty level for respondents. A related contrast is that while CTT presumes that test parallelism is required for comparing test scores across multiple forms, IRT presumes that different forms of a test developed through adaptive testing and administered to persons with different levels of a trait will maximize testing accuracy (Embretson & Hershberger 1999: 11–14).

The logic of IRT also suggests that the difficulty level of items within a test (or index) can be determined by the fraction of respondents who answer correctly (or in one direction). This in turn makes it possible to specify the numerical difference in difficulty between any two items, thus measuring the variation in true scores in terms of an interval scale. For example, Woodcock (1974) constructed a "word recognition ruler" using IRT principles. Words on a reading mastery test can be placed on this ruler and compared in standard intervals in terms of difficulty:

"away" is one unit less difficult than "drink," while "equestrian" is two units more difficult (Wright 1999).

IRT focuses attention on the possibility of differential functioning of items between different populations and provides a means for identifying such differences. As an example, Rania Tfaily (2010) estimated the extent to which different questions about the acceptability of family dissolution were understood differently by respondents of different gender, ethnicity, and nationality in 56 communities in South and Southeast Asia. Wives and then husbands were asked whether a husband (and then wife) is justified in leaving his wife given conditions ranging from infertility to infidelity. Statistical analysis identified several items as lacking measurement equivalence for respondents across regions and nations. The seriousness of behaviors like alcoholism or infidelity was viewed differently by respondents in some countries and with respect to wives compared to husbands, even among those with the same trait level (acceptance of family dissolution). Items about such behaviors should not be used in cross-country comparative studies with the assumption that their meaning is constant.

The IRT approach to measurement design thus reflects an inherently inductive strategy, in which the methodologist learns from research how to construct an efficient instrument and how to alter that instrument for different contexts.

Generalization

Quantitative methods begin with measurement and it might be said that they end with generalization. The best measures are of little use if the descriptions they provide cannot be generalized to the people, groups or other entities about which the researcher had hoped to learn. Of course if it were possible to measure every entity of interest, there would be no need to consider methods for improving researchers' ability to generalize their findings, but that is rarely – and, practically speaking, never – the case.

The problem

We need go no further than the US Constitution to begin to understand the problem. The decennial census is mandated by article 1, section 1 of the Constitution of the United States of America; it is planned and administered by a massive government agency, citizens are legally required to complete their census form, and there are tangible benefits to census completion in the resulting allocation of government services and political representatives. Yet in the 2010 US Census, only 67 percent of US households returned their mailed census form, leaving the rest of the population's participation to be elicited by 635,000 temporary workers (US Census Bureau 2010). In 2000, a comparable army of employees missed an estimated 3.3 million persons, compared to 281.4 million who were counted (ESCAP, US Census Bureau 2001). It is not certain what the future holds: the mailback response rate declined from 78 percent in 1970 to 65 percent in 1990, but regained some of that lost ground due to an extensive mobilization campaign in Census 2000 (Hillygus, Nie, Prewitt & Pals 2006: 20).

If the $4.5 billion spent to count almost every person in the 2000 US Census – about $15 per person – still leaves some uncertainty about generalizability to groups that fail to respond at high rates, what can researchers do who have only a tiny fraction of the resources of the US government at their disposal (Gauthier 2002: A1)? One response is to rely on Census data itself when planning analyses of sociological research questions. This is a very reasonable option for those whose research questions can be addressed with Census data, and there are also many datasets available for reanalysis from other government and private data collection efforts (Schutt 2009: ch. 13). However, in the usual circumstance when new data must be collected to answer previously unaddressed questions or to study different populations, researchers must rely on samples – subsets of the population – to which they can devote many more resources in order to achieve a higher rate of response.

The solution

The methodology of random sampling and associated statistics provides a means both for selecting a sample that is likely to be representative of the population from which it was selected and for estimating how probable it is that findings obtained with the sample are true for the larger population. A truly random sampling process results in the selection of cases without bias from the population of interest. In other words, each element in the population has an equal probability of selection in a simple random sample: no particular types of elements are more or less likely to be selected. More complex sampling strategies may involve stratifying the population in terms of key characteristics, such as income or race, and then sampling randomly within strata in order to ensure appropriate representation of each stratum, or sampling first clusters of elements, such as states, cities or blocks, and then sampling elements within those clusters. These strategies may be combined and they may be adjusted so as to over-represent within a sample elements in relatively small strata or clusters, but even in these samples the probability of selection of each element is known and can be adjusted *ex post facto* to yield a representative sample of the larger population.

Mastering the methods of probability-based sampling is the most important step to achieve generalizable results, but three problems remain. First, random selection of elements from a population will not itself produce a representative sample if many of those elements decline to participate in the survey or otherwise are unavailable to the researcher. Second, the method of statistical inference allows estimation of the odds that a statistical finding is due to chance, but it can be and is frequently misunderstood and misused. Third, researchers often seek to generalize beyond the population from which their samples were drawn and so the methods of statistical inference do not apply.

The problem of survey nonresponse is a major impediment to confident generalization of results obtained with an otherwise randomly selected sample. A typical mailed survey will elicit at best a response from 30 percent of the selected respondents, unless extensive follow-up efforts are made with nonrespondents and the survey is designed to exacting specifications so that it is clear, attractive, and credible. Phone surveys used to elicit a much higher rate of response, but the growth of

cell phone-only households, the use of answering machines and caller ID as screening devices, and the negative penumbra emanating from excessive telemarketing have combined to drastically reduce the telephone survey response rate: from 80 percent in 1979 to 60 percent in 2003 for the general population, and from 35 percent in 1990 to 20 percent in 2006 among those aged 18 to 34 years (Keeter 2008; Rainie 2010). Response rates remain high for in-person interview studies, but costs are prohibitive for all but well-funded projects.

Once data have been collected with a probability sampling strategy and a reasonable response rate achieved, the process of statistical inference begins with the recognition that the value of a statistic calculated with the obtained sample – such as a mean (arithmetic average) – is only one of an infinite number of values for the statistic that would have been obtained if random samples were selected from the population, *ad infinitum* (presuming replacement of all elements back to the population after each random sample is drawn). Just by chance, some of these sample statistics will have a value close to the true value of the statistic in the population – the population parameter – and others will be far from that value. The hypothetical distribution of all possible values of a statistic for samples of a particular size is termed a sampling distribution.

Recognition of the inherent uncertainty in knowing that a sample statistic comes from an unknown point on a sampling distribution would paralyze efforts to generalize from random samples were it not for the discovery that the sampling distribution for each statistic with a sample of a particular size has a characteristic, and knowable, shape. For the mean and many other statistics, that shape is a normal distribution. Since the normal distribution has a known shape and properties – particularly, the area under the normal curve is invariant with respect to multiples of the sampling distribution's standard deviation (termed the "standard error") and the standard error decreases as sample size increases – knowing that the sample statistic comes from a distribution of known shape means knowing a lot. Armed with this knowledge, a statistician can then estimate the degree of confidence that can be placed in an estimate that the population parameter falls within a particular range of values. Thus does a statistical description of a sample become a generalization to a population: "We can be 95 percent confident that the mean level of anxiety in the population is between 3.7 and 5.2." Bertrand Russell (1962: 63–4) expressed the paradoxical result: "Although this may seem a paradox, all exact science is dominated by the idea of approximation. When a man tells you that he knows the exact truth about anything, you are safe in inferring that he is an inexact man. Every careful measurement in science is always given with the probable error . . . every observer admits that he is likely to be wrong, and knows about how much wrong he is likely to be."

A similar inferential process occurs when statistics are compared between two or more groups, or in fact whenever a statistic is used to characterize the relationship between two or more variables. Could the difference in means between two (or more) groups have been due to chance? Could the increasing average income that was observed as the average level of education increased have been a chance association?

Tests of significance are used to answer such questions. They proceed in three stages: first, specify what a truly random process would produce, such as no

divergence between the means of two groups. Second, calculate a statistic capturing the difference between that result and what was actually obtained (such as a difference of 17.5 between the two groups rather than a difference of 0). Third, locate the value of the obtained statistic on the sampling distribution for that statistic (given the number of cases and the variability of the sample) and determine how likely it is that the obtained value of the statistic could have diverged as much as it did from the value expected on the basis of chance. Seeking to avoid a rush to the conclusion that a difference – that is, "something" – has been found, quantitative methodologists have settled on a convention that a difference is not considered "statistically significant" unless it is likely to have occurred on the basis of chance no more than five times in 100. However, even when used appropriately, inferential statistics cannot support generalizations to populations that have not been sampled: a study of students at one school cannot be generalized with any knowable degree of confidence to all students, nor a sample of residents of one city to all urban residents, or a sample from one nation to all nations. Perhaps the results in such situations are generalizable in the way desired, but that possibility cannot be estimated from the sample results themselves.

New directions

One might easily conclude, after reading many (although not all) quantitative research articles, that that is all there is to it. In other words, once the data analyst knows whether the possibility that a given result is due to chance can safely be rejected, the story is finished. We conclude that there is (or is not) a difference between two means, a relationship between two variables, perhaps a divergence between two trend lines. But such a result, stopping with the establishment (or not) of statistical significance, is a mistake. To maintain a singular focus on whether an effect "exists" in the probabilistic sense while ignoring the magnitude of that effect is to neglect substantive results and turn sociology and other social sciences into a "sizeless science." "To cease measuring *oomph* [or effect size; emphasis added] and its relevant sampling and nonsampling error is to wander off into probability spaces, forgetting – commonly forever – that your interest began in a space of economic or medical or psychological or pharmacological [or sociological] significance" (Ziliak & McCloskey 2008: 9). Growing recognition of the problem of overlooking the strength of relationships in quantitative research has led to more attention to "effect size" statistics – standardized estimates of the amount of change or difference (Lipsey & Wilson 2001).

New directions are also being charted in the effort to improve survey participation rates. As the percentage of households without internet access continues to drop (it was about one-quarter of households in the US in 2009), web-based surveys have become an increasingly attractive alternative to phone and mailed designs (Rainie 2010). The most rigorous web-based survey method involves samples of respondents who are recruited at the household level without regard to internet access and then provided with a free computer and internet access if they are not already connected (Couper 2000; Heeren, Edwards, Dennis, Rodkin, Hingson & Rosenbloom 2008). Internet-based surveys of populations in which internet use is almost universal, such as college students, can also be very successful (Dillman 2007). The increasing numbers of international surveys and the growing number of

available datasets collected in diverse countries are improving the possibilities for research whose results can be generalized beyond traditional boundaries.

Causation

The central task of science is to explain how the natural world works and many quantitative methodologists in sociology accept this same causal mandate to guide their research and theorizing. More specifically, quantitative researchers develop and test "nomothetic causal explanations," in which a common influence is identified on variation in some phenomenon across a number of cases. From this standpoint, a causal effect occurs when variation in an independent variable is followed by variation in a dependent variable, *ceteris paribus* (all else being equal). Learning how to formulate hypotheses positing causal influences of an independent variable on a dependent variable is an essential part of any sociology research methods course.

The problem

What makes establishing causal effects a challenge – and even an impossibility, from the perspective of some philosophers of science – is the stipulation that a causal effect has only been identified when all else that might have caused variation in the phenomenon of interest is equal. This *ceteris paribus* assumption presumes a "counterfactual" situation that can only exist hypothetically. Whatever the posited cause of an outcome, the other circumstances associated with its occurrence cannot be replicated in exactly the same place, at exactly the same time, with exactly the same people, but now without that cause being present.

Lacking a perfect counterfactual comparison, most quantitative methodologists accept a causal assertion as being justified if the evidence meets three criteria: *association* – there must be an association between the presumed cause and effect; *time order* – variation in the presumed cause must occur prior to the variation in its presumed effect; *non-spuriousness* – the variation in the effect must not be due to some influence other than the presumed cause. The extent to which each of these criteria can be met is determined by the research design and the statistical procedures used to analyze data collected with that design.

The solution

Experimental design is widely accepted as the "gold standard" for meeting the three criteria for confirming a causal effect. In a true experimental design, two (or more) groups are assigned to receive different levels of the independent variable, or hypothesized cause (often treatment compared to no treatment), and after the treatment their scores are measured on the dependent variable of interest. Thus, the criterion of an association between the presumed cause and effect can be achieved. Scores on the dependent variable are measured in the different groups both before and after the treatment, thus allowing determination of whether the criterion of time order has been met. Most importantly, cases are assigned randomly to the two (or more) groups, thus allowing confidence that nothing but chance has influenced the value of the independent variable experienced by each case – thus ensuring (within

a margin of statistical error) that there is no "selection bias" in assignment to the groups, so any association found between the independent variable and the dependent variable is not spurious due to the effect of some preexisting difference between the groups.

Of course many sociological hypotheses that involve causal relationships cannot be tested with a true experiment. People cannot be randomly assigned to a race or gender, or to have a specific income or to have been abused as children. In lieu of control for potential sources of spuriousness through experimental design, quantitative methodologists often turn to survey designs and multivariate statistical analyses to reduce the threat of spuriousness. The basic approach is to measure the other variables that may affect the causal relation of interest and then to hold variation in these variables constant while testing the relationship between the independent variable of interest and the dependent variable.

There are many multivariate statistical techniques that can be used for this purpose, with their appropriateness depending on the particular characteristics of the variables and the specific analytic problem. Multiple regression analysis is the most widely used multivariate statistical approach in sociology and it is the foundation for many more advanced techniques.

New directions

As quantitative methods for identifying causal effects, both experimental design and multivariate statistics have been extended with approaches that focus on two additional factors in addition to the three traditional criteria for establishing causality: causal mechanism and causal context. Causal mechanism can be defined as the process by which a treatment has its effect on a dependent variable, and with respect to experimental research it is often termed "opening the black box" of the treatment–outcome relationship. For example, reanalysis of qualitative observational data collected in the experimental study by Sherman and Berk (1984) of the police response to domestic violence allowed them to identify how police officers' interactions with the suspect influenced the extent to which the "treatment" imposed by the police officer (arresting or warning the suspect) changed the suspect's likelihood of reoffending (Paternoster, Brame, Bachman & Sherman 1997).

The effect of context can be evaluated in experimental research by replicating an experiment in diverse contexts. Again, the Sherman and Berk research provides a useful example. In order to determine whether the beneficial effect on recidivism of mandatory arrest in cases of domestic violence occurred in other contexts, the original study in Minneapolis was replicated in five other cities (Sherman 1992). The disparate results obtained made it clear that requiring arrest rather than less severe punishment in domestic violence cases did not have a consistent effect on recidivism. Thus, this variation in police practices could not be understood as by itself a cause of variation in recidivism; other conditions had to be met.

"Intervening variable" is the term used in multivariate statistics for what is called "causal mechanism" in experimental research. In multivariate non-experimental studies, the effort to identify variables that transmit a causal effect from an independent to a dependent variable can lead to complex causal models in which one or more paths of influence are proposed and then statistically tested. Path analysis

apportions correlations between hypothesized causal variables between direct paths that lead to one or more dependent variables and indirect paths through hypothesized intervening variables to those same dependent variables. Structural equation modeling proposes latent variables to capture the shared variation between multiple indicators and then tests the relations between these latent variables (Goldberger & Duncan 1973).

Multilevel modeling, or hierarchical linear models, is an increasingly popular multivariate statistical method that takes account of clustering of individuals within groups and improves estimates of effects of context. The primary motivation for multilevel modeling is the realization that individuals who are clustered together in such units as schools, classes, or blocks will tend to be more similar to each other than to individuals in other such clusters. When a random sample of individuals is selected through a multistage process in which first the clusters and then individuals within the clusters are randomly sampled, this clustering must be taken into account when calculating significance tests (inferential statistics). If ordinary multiple regression analysis is used to estimate effects with such multilevel samples, the results can be totally misleading (Hox 1998).

Propensity score methods are also gaining in popularity as a quantitative approach to lessening the risk of selection bias in non-experimental research. An individual's propensity score is "the conditional probability of being treated given the individual's covariates" (D'Agostino 1998: 2265). The propensity score is calculated for each individual using discriminant analysis or logistic regression to estimate the likelihood that individuals would be in the treatment group or control group based on their characteristics ("covariates"). Once the propensity scores are calculated, they can be used to equate individuals in the treatment and control groups using techniques such as matching pairs of cases in the two groups or controlling for the propensity scores in a regression analysis. The result can be a closer equivalence of the treatment and control groups than otherwise is obtained, with a concomitant reduction of the risk of spurious conclusions about the treatment effect. However, the propensity score can only take into account potential influences on treatment selection that have been measured, so it is not a substitute for careful design of the original research.

Conclusions

Quantitative methods continue to play a central role in sociological research, but without the unfettered adulation of quantification that characterized the Scientific Revolution and much of American sociology in the early to mid-twentieth century. Increasingly sophisticated approaches have been developed in response to recognition of the limitations of what previously were considered to be adequate solutions to research problems. Recognition of non-homogeneity of measured units is increasing attention to item response theory as a guide in quantitative measurement; the limitations of tests of statistical significance fuel growing attention to effect size statistics; awareness of the inherent ambiguity of causal assertions has resulted in greater attention to causal mechanisms and causal context; evidence of misleading results due to common violations of multivariate statistical assumptions has led to greater use of multilevel modeling and other more sophisticated analytic techniques.

The growth and formalization of qualitative research methods and greater acceptance of mixed methods have also infused quantitative methods with greater sensitivity to the importance of inductive research strategies and the potential contribution of in-depth qualitative data for improving measures and specifying causal influences (Clark & Creswell 2008).

Increasingly powerful computational facilities, more sophisticated statistical procedures, and the challenges of investigating an increasingly diverse and interconnected social world will continue to fuel these trends. There is no more hope in the twenty-first century than there was in the twentieth that quantitative methods in general or statistics in particular will provide the key envisioned by Florence Nightingale to "the plan of God" (McDonald 2003: 74), but it is certain that they will continue to enrich sociology's contributions to understanding the social world.

References

Adamczewski, Jan (1974) *Nicolaus Copernicus and his Epoch*. In cooperation with Edward J. Piszek. New York: Scribner.

Blalock, Hubert M., Jr. (1982) *Conceptualization and Measurement in the Social Sciences*. Beverly Hills: Sage.

Calder, Nigel (1979) *Einstein's Universe*. New York: Viking Press.

Clark, Vicki L. Plano and John W. Creswell (2008) *The Mixed Methods Reader*. Thousand Oaks, CA: Sage.

Copernicus, Nicholas ([1543] 1978) *Nicholas Copernicus Complete Works*, Vol. II, *Nicholas Copernicus On the Revolutions*. Ed. Jerzy Dobrzycki. Trans. with commentary Edward Rosen. Warsaw-Cracow: Polish Scientific Publishers.

Couper, Mick P. (2000) "Web Surveys: A Review of Issues and Approaches." *Public Opinion Quarterly* 64: 464–94.

D'Agostino, Ralph B., Jr. (1998) "Tutorial in Biostatistics: Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Non-Randomized Control Group." *Statistics in Medicine* 17: 2265–81.

Dillman, Don A. (2007) *Mail and Internet Surveys: The Tailored Design Method*, 2nd edn. Update with New Internet, Visual, and Mixed-Mode Guide. Hoboken, NJ: Wiley.

Einstein, Albert (1921) *Relativity: The Special and General Theory*. Trans. Robert W. Lawson. New York: Henry Holt.

Embretson, Susan E. and Scott L. Hershberger (eds.) (1999) *The New Rules of Measurement: What Every Psychologist and Educator Should Know*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gauthier, Jason G. (2002) *Measuring America: The Decennial Censuses from 1790–2000*. Washington, DC: Department of Commerce, US Bureau of the Census.

Goldberger, Arthur S. and Otis Dudley Duncan (eds.) (1973) *Structural Models in the Social Sciences*. New York: Seminar Press.

Heeren, Timothy, Erika M. Edwards, J. Michael Dennis, Sergei Rodkin, Ralph W. Hingson, and David L. Rosenbloom (2008) "A Comparison of Results from an Alcohol Survey of a Prerecruited Internet Panel and the National Epidemiologic Survey on Alcohol and Related Conditions." *Alcoholism: Clinical and Experimental Research* 32: 222–9.

Hillygus, D. Sunshine, Norman H. Nie, Kenneth Prewitt, and Heili Pals (2006) *The Hard Count: The Political and Social Challenges of Census Mobilization*. New York: Russell Sage Foundation.

Hox, Joop (1998) "Multilevel Modeling: When and Why." In Ingo Balderjahn, Rudolf Mathar, and Martin Schader (eds.), *Classification, Data Analysis, and Data Highways*. New York: Springer Verlag, pp. 147–54.

Jacoby, William G. (1991) *Data Theory and Dimensional Analysis*. Thousand Oaks, CA: Sage.

Keeter, Scott (2008) "Survey Research and Cell Phones: Is There a Problem?" Presentation to the Harvard Program on Survey Research Spring Conference, New Technologies and Survey Research. Cambridge, MA: Institute of Quantitative Social Science, Harvard University, May 9.

Koyré, Alexandre (1965) *Newtonian Studies*. Cambridge, MA: Harvard University Press.

Krosnick, Jon A. (1999) "Survey Research." *Annual Review of Psychology* 50: 537–67.

Kuhn, Thomas (1970) *The Structure of Scientific Revolutions*, 2nd edn. Chicago: University of Chicago Press.

Lipsey, Mark W. and David B. Wilson (2001) *Practical Meta-Analysis*. Thousand Oaks, CA: Sage.

McDonald, Lynn (ed.) (2003) *Florence Nightingale on Society and Politics, Philosophy, Science, Education and Literature*. Waterloo, Ontario: Wilfrid Laurier University Press.

Newman, James R. (1956) "Commentary on Sir Francis Galton." In James R. Newman (ed.), *The World of Mathematics*, Vol. 2. New York: Simon and Schuster, pp. 1167–72.

Paternoster, Raymond, Robert Brame, Ronet Bachman, and Lawrence W. Sherman (1997) "Do Fair Procedures Matter? The Effect of Procedural Justice on Spouse Assault." *Law and Society Review* 31: 163–204.

Planck, Max (1949) *Scientific Autobiography and Other Papers*. Trans. Frank Gaynor. New York: Philosophical Library.

Presser, Stanley, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, Jennifer M. Rothgeb, and Eleanor Singer (2004) "Methods for Testing and Evaluating Survey Questions." *Public Opinion Quarterly* 68: 109–30.

Rainie, Lee (2010) *Internet, Broadband, and Cell Phone Statistics*. Pew Internet and American Life Project. www.pewinternet.org/~/media//Files/Reports/2010/PIP_December09_update.pdf (accessed June 24, 2010).

Reece, Gordon (1977) "In Praise of Uncertainty." In William C. Price and Seymour S. Chissick (eds.), *The Uncertainty Principle and Foundations of Quantum Mechanics: A Fifty Years' Survey*. New York: John Wiley and Sons, pp. 7–12.

Russell, Bertrand (1962) *The Scientific Outlook*. New York: Norton.

Schaeffer, Nora Cate and Stanley Presser (2003) "The Science of Asking Questions." *Annual Review of Sociology* 29: 65–88.

Schutt, Russell K. (2009) *Investigating the Social World: The Process and Practice of Research*, 6th edn. Thousand Oaks, CA: Pine Forge Press/Sage.

Scripture, E. W. (1892) "The Need of Psychological Training." *Science* 19: 127–8.

Sherman, Lawrence W. (1992) *Policing Domestic Violence: Experiments and Dilemmas*. New York: Free Press.

Sherman, Lawrence W. and Richard A. Berk (1984) "The Specific Deterrent Effects of Arrest for Domestic Assault." *American Sociological Review* 49: 261–72.

Tarnas, Richard (1991) *The Passion of the Western Mind: Understanding the Ideas That Have Shaped Our World View*. New York: Ballantine Books.

Tfaily, Rania (2010) "Cross-Community Comparability of Attitude Questions: An Application of Item Response Theory." *International Journal of Social Research Methodology* 13: 95–110.

US Census Bureau, ESCAP (2001) http://govinfo.library.unt.edu/cmb/cmbp/reports/final_report/fin_sec3_evaluation.pdf

US Census Bureau (2002) Census 2000 Basics. www.census.gov/mso/www/c2000basics/00Basics.pdf

US Census Bureau (2010) "The Numbers Are In." www.census.gov/newsroom/releases/archives/2010_census/cb10-cn61.html

Viswanathan, Madhu (2005) *Measurement Error and Research Design*. Thousand Oaks, CA: Sage.

Woodcock, Richard W. (1974) *Woodcock Reading Mastery Tests*. Circle Pines, MN: American Guidance Service.

Wright, Benjamin D. (1999) "Fundamental Measurement for Psychology." In Susan E. Embretson and Scott L. Hershberger (eds.), *The New Rules of Measurement: What Every Psychologist and Educator Should Know*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 65–104.

Ziliak, Stephen T. and Deirdre N. McCloskey (2008) *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor, MI: University of Michigan Press.