of the analytical insights of formalism, while at the same time holding out the proviso that meaningful explanations of important Balinese social phenomena will require detailed description of the culturally specific norms, values, and worldviews that formalism overlooks.

NOTES

1. The infelicitous terms apparently come from Max Weber. "Formalism" denotes the idea that explanation can proceed on the basis of an abstract (formal) description of the environment of choice of participants, whereas "substantivism" denotes the idea that social explanation requires concrete (substantive) knowledge of participants' norms and values over and above the circumstances of choice in which they find themselves. See Weber's discussion in *Economy and Society* (1978:85). Formalists include Samuel Popkin, Manning Nash, Theodore Schultz, Ramon Myers, and Kang Chao; substantivists include James Scott, Karl Polanyi, George Dalton, Marshall Sahlins, and Clifford Geertz.

2. See Stephen Stich (1983) for a discussion of the role of folk psychology within scientific psychological theory.

3. Recent economists and philosophers have offered a number of proposals on this subject, including A. K. Sen, John Harsanyi, Howard Margolis, and Donald Regan.

SUGGESTIONS FOR FURTHER READING

Dalton, George. 1971. *Economic Anthropology and Development.*
Hardin, Russell. 1982. *Collective Action.*
Little, Daniel. 1989. *Understanding Peasant China: Case Studies in the Philosophy of Social Science.*
Nash, Manning. 1966. *Primitive and Peasant Economic Systems.*
Polanyi, Karl. 1957. *The Great Transformation.*
Popkin, Samuel L. 1979. *The Rational Peasant.*
Russell, Clifford S., and Norman K. Nicholson, eds. 1981. *Public Choice and Rural Development.*
Sahlins, Marshall. 1972. *Stone Age Economics.*
Scott, James C. 1976. *The Moral Economy of the Peasant.*
Shanin, Teodor. 1985. *Russia as a 'Developing Society.'*
Taylor, Michael. 1982. *Community, Anarchy and Liberty.*

Little, Daniel
Varieties of Social Explanation. An Introduction to the Philosophy of Social Science. Boulder: Westview, 1991.

13

8
STATISTICAL ANALYSIS

A common mode of explanation in social science is *statistical*, wherein the scientist explains a phenomenon in terms of its correlation with other variables. Why has Korean economic growth been so remarkable since 1965? Because that nation has had a stable political environment, a relatively equal distribution of assets, and an educated labor force—variables that are positively correlated with economic growth in the countries of the less-developed world. But what sort of explanation is this? This chapter will attempt to answer this question more fully, but a brief reply may be presented in advance. Statistical correlation is *explanatory* to the extent that it provides evidence of a credible causal process underlying the variables being analyzed. Statistical explanations, that is, must be accompanied by a *causal story* indicating the mechanisms through which observed correlations evolve, if the analysis is to be explanatory at all. (The causal story may be provided in greater or lesser detail.)

The explanation of Korean economic growth in the previous paragraph is couched in terms of correlations between a dependent variable (growth) and several independent variables (political stability, equality, and education levels), but this claim is intended to show that there is a *causal* relation between the latter factors and economic growth. In this case the causal story is not difficult to construct. Political stability is causally relevant to growth because growth requires investment and investors are more likely to invest if they are confident that existing institutions will continue. Equality is causally relevant to growth because it stimulates smooth structural change (by creating strong consumer demand for commodities). And education levels are causally relevant to growth because they are a measure of the human capital available to a society.[1] The correlations identified here are explanatory, then, because they identify causal factors that influence the rate of economic growth through credible causal mechanisms.

Thus statistical explanation is a form of causal explanation, and the conclusion of Chapter 2—that causal explanations require hypotheses about underlying causal mechanisms—is equally pertinent here as well. Mere evidence of statistical correlation between a pair of variables does not constitute an explanation of the behavior of either. Instead we can properly rest our explanatory inquiries only when we have established in a credible way the causal relation between the variables.

Consider some of the following research topics in the social sciences:

- What factors stimulate rapid economic growth in the less-developed world?
- What explains the distribution of earnings in the U.S. economy?
- Is the religious identity of a population relevant to the presence or absence of democratic institutions?

Each topic involves the search for an association between two or more variables. In the first instance the researcher aims to discover one or more factors—social, political, or economic—that are positively associated with rapid economic growth. In the second the researcher hopes to discover features of individuals that correlate positively with variations of earnings—for example, gender, skill level, race, or education level. And in the third the researcher aims at evaluating a simple causal hypothesis about regime types—that religious identity (Catholic, Protestant, Islamic) is a causal variable in the development and stability of democratic political institutions.

Statistical explanations raise two sorts of problems that are relevant to the concerns of this book. First, we must have at least a rudimentary understanding of the statistical concepts on which such explanations depend—for example, correlation, regression, association, and conditional probabilities. But second, we must ask a more philosophical set of questions. What have we learned when we have uncovered a statistical relationship between several variables? Is this discovery itself explanatory or is it an empirical circumstance that demands theoretical explanation (hypothesis formation about causal relations)? How do statistical findings support causal inferences? And what are the limitations of statistical analysis in social science?

An attractive example of a causal argument that combines statistical analysis with a hypothesis about the underlying causal mechanisms is a recent study of English demographic history (Example 8.1). Wrigley and Schofield's argument supports a causal hypothesis about economic and demographic change on the basis of a statistical correlation over time between demographic variables and the standard of living. The hypothesis is that growing real wages will result in a rise in the rate of population increase, both through higher fertility and lower mortality, and falling trends in real wages will bring a slowing of the rate of population increase. The greatest difficulty in evaluating this hypothesis empirically is the need for detailed time-series data on real income and fertility and mortality rates; the large data set compiled through parish registers permits such an evaluation. If we graph real wages and rate of population increase over time, the results resemble Figure 8.1, and the correlation between the two variables is evident. Real wages fall between 1577 and 1617; with a short lag the population growth rate begins to fall in the late 1670s. Real wages begin to rise again in the 1630s, and, with a 30- to 40-year lag, the rate of population increase begins to rise as well.

Example 8.1 Historical demography

What factors influence demographic change—fertility rates, age of marriage, rise and fall of population, or age structure of the population? A major effort to answer this family of questions is the recent study of English population (1541–1871) by Wrigley and Schofield and the Cambridge Group for the History of Population and Social Structure. The study provides a detailed description of the year-to-year state of the English population over a three-hundred-year period, based on data drawn from hundreds of Anglican parish registers that record baptisms (births), burials (deaths), and marriages. Using the statistical summary of this data, the authors attempt to empirically evaluate the Malthusian hypothesis that European demography is highly sensitive to economic variations (rising and falling real incomes). "We, therefore, preferred to follow Malthus in taking a wider view of economic opportunity in relation to the preventive check and to consider marriage to have been responsive to the level of real incomes rather than determined solely by access to a niche" (Schofield 1986:15). To evaluate this hypothesis Wrigley and Schofield provide a time series for real-wage levels for the period 1500–1912 and compare the results with the patterns of fluctuation in demographic variables over the same period. Their analysis shows that fertility varied substantially over long periods of time and that fertility fluctuations were approximately twice as important in population change as fluctuations in mortality rates (Schofield 1986:27). And their analysis also corroborates Malthus's general hypothesis that preventive checks (behavioral limitations on fertility), dependent on movements of the real wage, were of primary importance in controlling the English population increase.

Data: population data (births, deaths, and marriages) from English parish registers, 1541–1871

Explanatory model: evaluation of causal hypotheses about demographic change based on analysis of large-scale quantitative analysis of demographic data

Sources: E. Anthony Wrigley and Roger S. Schofield, *The Population History of England, 1541–1871: A Reconstruction* (1981); Roger S. Schofield, "Through a Glass Darkly: *The Population History of England* as an Experiment in History" (1986); Robert I. Rotberg and Theodore K. Rabb, eds., *Population and Economy Population and History from the Traditional to the Modern World* (1986)

QUANTITATIVE REASONING IN SOCIAL ANALYSIS

In this section I will provide an abstract account of statistical reasoning in social science. This discussion is not intended to replace detailed mathematical study of the subject, but it is possible to present enough of the elements of statistical reasoning in this context to identify some of the problems of social explanation in this area.

A *data set* involves the following structure:

- a domain of items, events, or individuals (persons, countries, riots, crimes, suicides);

Fig. 8.1 Real wage and population growth rate

Source: Adapted from Lee 1986:89

- a set of properties of these items that are of interest to the researcher (variables);
- a specification of the state of each item with respect to each property.

We may think of a data set as a table listing individuals in the rows and the properties in the columns. Figure 8.2 is an example of a data set providing economic and social data for twenty-two less-developed countries. This table gives information on ten variables for the countries under study.

There is an implicit temporal structure in empirical studies; data may be collected either to capture change over time or to characterize the state of a set of individuals at a given time. The data provided in Figure 8.2 is a *cross-sectional* data set—one providing a "snapshot" of the state of the items at a moment in time with respect to the properties. A *time-series* data set is a study organized over a time sequence, involving data points for one item at successive moments in time. It is possible to design a study that

Country	GNP per capita* ($)	Energy use per capita*	Labor force in agric.* (%)	Life expec-tancy at birth* (years)	Adult literacy* (%)	Growth rate in GNP per capita* (%)	Gini coef.*	Ratio of top 20% to bottom 40% of income*	PQLI**	Infant mortality /1000 live births**
Ethiopia	120	19	80	43	15	0.50			20	181
Bangladesh	130	36	74	50	26	0.50	0.389	2.74	35	132
Mali	160	22	73	45	10	1.20			15	188
Tanzania	240	38	83	51	79	0.90	0.42	3.15	31	162
India	260	182	71	55	36	1.50	0.407	3.05	43	122
China	300	455	74	67	66	4.40			69	78
Ghana	310	111	53	59		-2.10			35	156
Sri Lanka	330	143	54	69	85	2.90	0.345	2.26	82	45
Kenya	340	109	78	57	47	2.30	0.55	6.79	39	119
Pakistan	390	197	57	50	24	2.50			38	121
Senegal	440	151	77	46	10	-0.50			25	159
Bolivia	510	292	50	51	63	0.60			43	108
Indonesia	560	204	58	54	62	5.00	0.43	3.43	48	137
Egypt	700	532	50	58	44	4.20	0.403	2.91	43	116
Philippines	760	252	46	64	75	2.90	0.459	3.80	71	74
Nigeria	770	150	54	49	34	3.20			25	180
Guatemala	1120	178	55	60		2.10			54	80
Colombia	1430	786	26	64	81	3.20	0.53	6.32	71	97
Malaysia	1860	702	50	67	60	4.50	0.5	5.01	66	75
Brazil	1880	745	30	64	76	5.00	0.605	9.51	68	82
South Korea	2010	1168	34	67	93	6.70	0.378	2.68	82	47
Mexico	2240	1332	36	66	83	3.20	0.52	5.83	73	66

Fig. 8.2 Third World country data set

incorporates both cross-sectional and temporal dimensions, consisting of coordinated time-series studies for a group of items.

How do social scientists collect data? A great volume of statistical data is gathered by national governments and governmental institutions, available for analysis by social scientists. For example, much of the data in Figure 8.2 was compiled by the World Bank. But many social science research topics require that the investigator collect data that has not been previously gathered by formal reporting agencies. A *survey* is a study designed to elicit information concerning the properties of items within the universe of items (or a sample of them), keeping certain hypotheses in mind. The designer must have some idea about what factors are potentially relevant to the occurrence of the phenomena of interest—that is, a range of causal factors. The goal of the study is to collect data indicating the strength (or absence) of correlation among the factors.

A central problem in designing an empirical study is that of *sampling*. Many social features can only be determined through studies that single out a subset (sample) of the whole population and then draw inferences about the population properties. And two problems arise in sampling: size and randomness. If the sample is too small, then there is no reason to expect that the features of the sample correspond to those of the population. For example, if we were interested in the party affiliation of steelworkers in Pennsylvania but interviewed only twenty workers, we would not be able to draw a statistically significant inference from our study.

Sample bias is a different sort of problem. For a sample to be a good indicator of the population as a whole, the individuals selected must be randomly drawn from the population. A survey rests upon a *biased* sample if the individuals studied were chosen according to criteria that make it likely they will share features that are not representative of the population as a whole. Age, gender, place of residence, social class, type of employment, or ethnic identity can all lead to a biased sample. The problem of bias often arises not because the researcher favors one hypothesis over another but because collecting information is difficult and costly, and it will sometimes be possible to collect data from a subpopulation that is conveniently at hand. This subpopulation, however, is not randomly drawn from the whole population. An example of this possibility is found in the important land surveys of China conducted by John Lossing Buck in the 1930s. Buck sent Chinese investigators into a number of different regions of China with a detailed survey to complete concerning land ownership. However the investigators had no easy access to absentee landlords—landlords who owned property in a village but lived elsewhere. So the survey excluded these individuals. Absentee landlords, however, owned larger-than-average pieces of land, and their exclusion led to a downward bias in estimates of the average size of landholdings (Esherick 1981).

The variables defining the information collected in the data set may be either discrete or continuous. Discrete variables include religion, marital status, and class membership. Continuous variables include income, population size, and unemployment rates. (The latter examples make it plain that continuous variables may include quantities for which only integer values are possible.) This distinction is important because the techniques of data analysis are different in the two cases. Continuous variables may be analyzed using regression techniques, whereas associations among discrete variables require different tools.

Once we have a data set we need some way of aggregating the data. The distinction between discrete and continuous variables is particularly important here for the techniques available for aggregating and analyzing discrete data are different from those available for continuous data. Discrete data primarily involves the use of probability tools (discussed in Chapter 2). We can count individuals having a certain property or set of properties, and we can use that information to derive incidence rates (for example, the suicide rate among Protestant widowers). The concept of conditional probabilities is the central tool in this type of case. Continuous variables, on the other hand, permit more extensive forms of mathematical analysis, including particularly the attempt to identify functional relations between variables (discussed below). Several types of *descriptive statistics* are particularly useful in characterizing such a data set. We can calculate mean (average) values for a set of individuals with respect to a given property—for example, the average life expectancy in Figure 8.2 is 57 years. And it is useful to provide measures of the variance of each variable—that is, a measure of the amount that the variable changes over the population. (The *variance* of a set of values is the average squared deviation of each value around the mean of all the values; the *standard deviation* is the square root of the variance.)

Once we have a data set[2] we must try to extract some order from it. At this point our interest becomes explicitly *explanatory*. We want to know whether the data is patterned and whether there are unexpected probability distributions, correlations, or functional relations between variables. A tool of general utility in science is the *null hypothesis*—the hypothesis that there is no relationship between two or more variables. The null hypothesis concerning smoking and cancer, for example, is that smoking is not causally involved in the production of cancer. The null hypothesis concerning economic growth and political stability is that the processes of economic growth do not affect political stability, either positively or negatively. The null hypothesis converts rather directly into various mathematical expectations deriving from the expectation of randomness in the behavior of the variables with respect to each other. In the case of probabilities, it implies that the conditional probability of an event in the presence of a specified condition will be equal to the absolute probability of the event. And in the case of continuous variables, it implies that the *correlation coefficient* of two variables will be 0—that is, there will be no observable pattern in the values of the two variables. (Correlation is the subject of the next section.)

If we find that the null hypothesis is not borne out—that is, that there is a nonzero correlation between two or more variables or that the conditional

probability of an event given a condition is different from the absolute probability of the event—then we may consider whether there is some causal process underlying the behavior of the variables. It may be a case of direct causation—smoking does cause cancer; rising per capita GNP does cause a fall in infant mortality. Or it may be indirect—a common set of factors may be influencing the behavior of the variables under observation, without the variables causally interacting. An example of this possibility is seen in the fact that energy use per capita is negatively correlated with the infant mortality rate. The best explanation of this fact is *not* that more energy use is good for infant health; it is rather that rising per capita income in a society gives rise to *both* rising energy use and falling infant mortality, thus inducing a correlation between the latter variables.

CORRELATION AND REGRESSION

The central idea underlying statistical explanation is the notion of a *correlation* between two or more variables. This concept describes covariance among variables: The variables in question take different values in different circumstances, and there is a tendency for the variables to vary together. A positive correlation means that an increase in one variable is associated with an increase in the other; a negative correlation means that an increase in one variable is associated with a decrease in the other. Two central questions must be posed in considering whether specific variables are correlated. What is the functional relationship between the variables? And how strong is the correlation between them—that is, how much dispersion is there in the data set? The first question asks how the dependent variable behaves in relation to changes in the independent variable(s); the second asks how much the dependent variable varies from what we would expect assuming a strict correlation between the variables.

Turn once more to the country data set in Figure 8.2. The variable describing energy use per capita is positively correlated with per capita incomes across countries; that is, the higher a country's per capita income, the greater the amount of energy used per capita. (Conversely the greater the energy use per capita, the greater the country's per capita income is likely to be.) An example of a negative correlation is found in the relation between infant mortality rates and per capita income; countries with higher per capita incomes generally have lower infant mortality rates.

There is a simple test for correlation between continuous variables. We can construct a *scatterplot* of the data—a graph that represents each variable along an axis and plots each data point in terms of the values of the variables for that point. (This can be done for two or more variables, but the most common application is the two-variable case.) Then we can inspect the resulting chart for an orderly progression among the data points. If the points are randomly scattered over the field, we may conclude that there is no correlation among the variables; if the points fall along a *trend-line*, we may conclude that a correlation does exist.

The statistical technique of *regression analysis* provides a quantitative method for analyzing covariance among two or more variables. Regression analysis, which underlies much quantitative reasoning in social science, is a mathematical technique designed to assess the claim that two continuous variables **x** and **y** are correlated. This means that each variation in the independent variable leads to a regular alteration in the dependent variable. There is, then, a functional relationship between these variables:

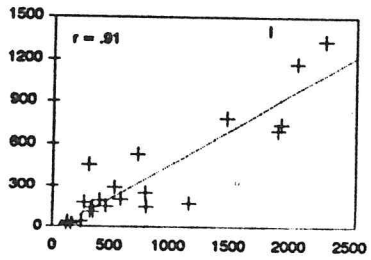
$$y_i = f(x_i) + e_i$$

Here we have broken down the behavior of the dependent variable **y** into a functional component $f(x_i)$ and an error component e_i. In principle the function may take any form whatsoever, but there are several simple functions that suggest themselves. The function may be *linear*, representing a straight-line relationship between the variables. Or it may be *curvilinear*—for example, logarithmic, exponential, or quadratic. Linear and logarithmic functions are the most common forms for representing relations among social variables. A linear function represents the dependent variable as rising or falling at a constant rate with respect to the independent variable, and a logarithmic function represents a declining rate of change as the independent variable increases. Linear functions have the form "$y = a + bx$," while logarithmic functions have the form "$y = a + b*\log(x)$."

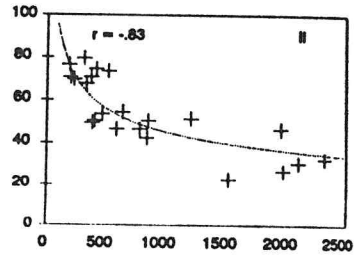
Linear regression on two values is the simplest form of regression analysis; this technique finds a straight-line function that passes through the data set and minimizes the variance around the line.[3] That is, the regression analysis provides a "best-fit" curve (a straight-line, in this case) that passes through the data points. (The regression curve is constructed according to the "least-squares" rule: It is the function that minimizes the sum of the squares of the distance off the trend-line of all points.) Figure 8.3 provides linear and logarithmic regressions on six pairs of variables drawn from the country study. Each panel represents a scatterplot of the data, along with the regression line for each data set. And each panel provides the computed correlation coefficient **r** (explained below), measuring the strength of the correlation of the data for this regression line.

Once we have performed a regression on our data, we have answered one of the questions posed above—what is the functional relationship between the variables? But we have not addressed the second question: How well does this functional relation characterize the data set? That is, how much scatter is there in the data around the regression line? Here we need an indicator of the closeness of fit between the data set and the regression equation—a measure of the degree of variance of the data around the function. There are several statistical measures of this variance, but the simplest is the coefficient of variation R^2. R^2 is defined as:

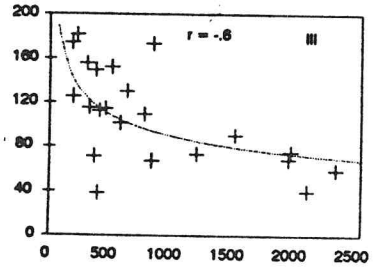
$$R^2 = 1 - \Sigma e_i^2 / \Sigma v_i^2,$$



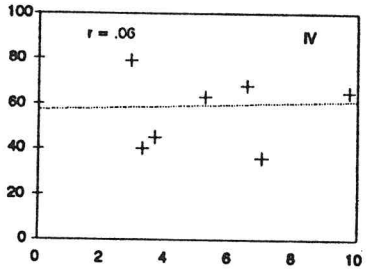
Energy consumption per capita vs. GNP per capita



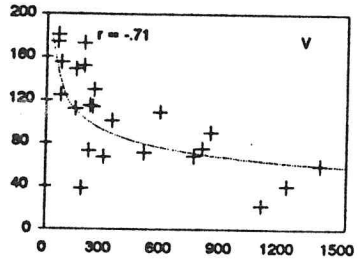
Percent of labor force in agriculture vs. GNP per capita



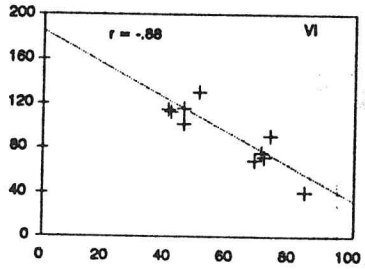
Infant mortality vs. GNP per capita



PQLI vs. income ratio



Infant mortality vs. energy use per capita



Infant mortality vs. PQLI

Fig. 8.3 Regression lines for selected variables in country study

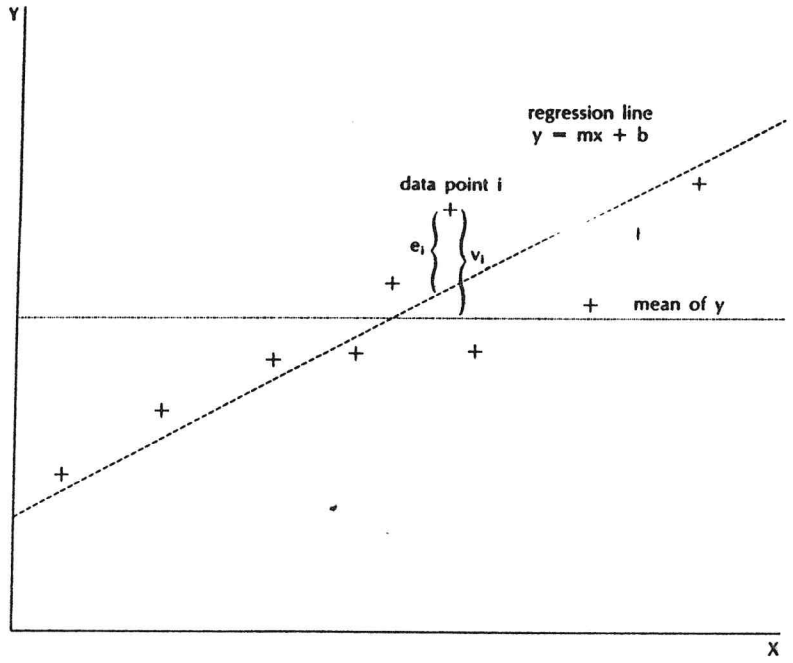


Fig. 8.4 Computation of R^2 for linear regression

where e_i are the error terms for each data point (deviations from the value predicted by the regression equation) and v_i are the variations of each data point around the mean value of y (Figure 8.4). R^2 is a *proportional reduction of error* statistic; it measures the percentage of variation in the dependent variable that can be explained by the regression function. Related to R^2 is the correlation coefficient r, the square root of the coefficient of determination. The correlation coefficient indicates the direction and strength of the relationship between the two variables, and it can vary between 1 and −1. A value of 1 corresponds to the case where the dependent variable always falls on the regression line; a value of 0 corresponds to the case where the dependent variable is randomly scattered around the regression line; and a negative value indicates that the regression line has negative slope (in other words there is a negative correlation between the two variables).

Figure 8.3 shows that there are significant correlations between various pairs of variables in the country study. Which, if any, of these correlations provide the basis for an explanatory relation between two variables? Panel I shows that there is a positive linear relation between gross national product (GNP) per capita and energy consumption; that is, economies with higher GNP per capita strongly tend to consume more energy per capita. The

regression shows that the relation between these variables is linear, and the correlation is high (r=.91). (In other words, 83 percent of the variation in per capita energy consumption can be accounted for on the basis of per capita GNP.) But how should we account for this correlation? Does growth in GNP cause growth in energy consumption (in the way that it plainly causes growth in luxury good consumption)? The causal story in this case is more complex. Growth in GNP unavoidably involves industrialization, and industrialization is energy intensive. Therefore rising energy consumption per capita is a necessary condition for rising per capita income. What panel I identifies, then, is the behavior of two variables within a complex causal process of economic growth, neither of which is the cause of the other. Much the same may be said about panel II. In this case we find a negative logarithmic correlation between GNP and the percent of the labor force in agriculture. Here again the correlation between the variables is high (r= −.83). The curve fits the data well. And the relationship between these variables is explanatory, reflecting the generalization that economic growth involves the structural transformation of an economy away from agriculture and toward industrial expansion. But once again it would be most plausible to construe this correlation as one between collateral effects within a complex process of economic growth: Industrialization leads both to rising per capita GNP and a structural transformation from a rural economy to an increasingly industrial one. In each case, then, the correlations we have discovered are best understood as correlations between collateral effects within a complex causal process.

Panels III and IV, by contrast, attempt to evaluate direct causal relations between pairs of variables. In panel III we find that there is a negative logarithmic relationship between GNP per capita and infant mortality rates, with a moderate correlation coefficient (r=−.6). Countries with higher income generally have lower infant mortality rates. And a simple causal hypothesis can account for this correlation: Both household and public health care expenditures rise quickly as incomes rise, and rising health care expenditures have major effects on infant mortality at low income levels. (This effect falls off as health care expenditures rise; there is a falling marginal effect of health care dollars on health status.) This analysis permits us to assert that rising income levels cause falling infant mortality levels. (However only 36 percent of the variance in infant mortality can be explained on this basis, so there must be other causes of variance that exercise influence as well.) Panel IV, however, may be construed as a refutation of a causal hypothesis concerning the relation between inequalities and physical quality of life. The measure of inequalities used in this panel is the ratio of the share of income going to the top 20 percent of income earners to that going to the bottom 40 percent. Physical quality of life is measured here by PQLI—an index that incorporates data about infant mortality, life expectancy at age one, and literacy (Morris 1979). One might reason that greater inequalities imply greater poverty, which in turn implies lower average physical quality of life. Thus one might entertain the following hypothesis: Greater inequalities will cause lower PQLI. However panel IV shows that there is little significant

relationship between these variables; the correlation coefficient is low (.17) and trends in the opposite direction to that predicted by the hypothesis. Greater inequality is therefore weakly associated with higher physical quality of life (presumably because both are correlated with per capita GNP).

Finally, consider panels V and VI. These panels identify two strong correlations: the first between infant mortality and energy use per capita and the second between infant mortality and PQLI. In each case we find a high degree of correlation between the variables. However, these correlations are spurious. Panel V shows that infant mortality falls with rising energy use per capita, but this correlation is induced by the common correlation between these factors and rising GNP per capita. In fact there is no direct causal connection between these variables at all. And the correlation between PQLI and infant mortality is purely artifactual: The PQLI index is defined in terms of the infant mortality level (along with two other factors). So it is a matter of definition rather than contingent association that these two factors are correlated.

Let us return now to an example discussed in several contexts above—James Tong's analysis of banditry in the Ming dynasty (Example 2.2). Tong believes his data set supports the hypothesis that the incidence of rebellion is correlated with the circumstances of risk found in various places and times. His own analysis rests on the observation that the incidence of banditry reported in Figure 2.2 varies in the direction predicted by the rational choice model. In this chapter, however, we have seen that it is desirable to ask two sorts of questions that Tong does not raise. What is the functional relation between risk and rebellion? And how high is the correlation between incidence of banditry and the two variables? The first question parallels the problem of determining how the incidence of rebellion should be expected to vary in response to an alteration in the level of risk; the latter involves determining how much of the variation of the dependent variable (incidence of banditry) is explained by the two independent variables.

We can use the technique of multiple regression to assess the degree of correlation in this analysis. Consider Tong's original table (Figure 2.2). Here we have two independent variables (hardship survival and outlaw survival) and one continuous dependent variable (incidence of banditry per 100 county-years). The dependent variable rests on observations of 630 cases and 303,869 county-years covered by the survey. It is reasonable to construe the survival levels as continuous variables measured discretely. Let us then assign a value of 3 for maximum survival, 2 for moderate survival, and 1 for minimum survival and perform a multiple regression analysis on the resulting data set. This produces the following functional relation between the risk variables (PEASANT and BANDIT) and the incidence of banditry (INCID):

$$\text{INCID} = -.48\ \text{PEASANT} + .39\ \text{BANDIT} + .86$$

This equation indicates that the incidence of rebellion rises as the prospects of survival as a peasant fall and the prospects of survival as a bandit rise.

Example 8.2 Sources of economic growth

What factors influence the rate of economic growth in various countries in the less-developed world? Adelman and Morris consider a list of 41 features of the social and economic organization that are potentially relevant to economic growth. They then construct a large study of 74 countries with respect to these features. Once the data is collected they make use of a statistical technique similar to regression analysis (factor analysis) designed to determine (1) the correlations among these variables across countries and (2) the correlations among groups of these variables (factors) and the rate of economic growth in the countries studied. Adelman and Morris conclude that two clusters of economic and social variables are significantly correlated with economic growth: variables that reflect processes of change in attitudes and institutions and variables corresponding to political regime type (Adelman and Morris 1967:153, 155). This analysis provides a basis for explaining why Korea experienced a very rapid rate of growth, India an intermediate rate, and Kenya a slow one.

Data: 74-country data set including 48 socioeconomic, political, and economic variables

Explanatory model: Statistical analysis of a large number of variables causally relevant to economic growth, based on data from 74 countries

Source: Irma Adelman and Cynthia Taft Morris, *Society, Politics, and Economic Development: A Quantitative Approach* (1967)

Moreover this function succeeds in explaining a high percentage of the variation in the dependent variable; there is a high multiple correlation between incidence of banditry and the two risk variables (multiple $R = .93$). (Multiple **R** is the multivariate equivalent to the correlation coefficient **r** discussed above.) This analysis gives us a more adequate way to estimate the degree to which Tong's data supports the rational choice hypothesis about banditry and rebellion than he himself provides. It shows that his data implies a correlation coefficient greater than .90 between the independent variables and the incidence of banditry, explaining over 80 percent of the variance in the data.

Consider an example that illustrates many features of the statistical analysis of a complex social phenomenon—an analysis by Adelman and Morris of the factors that influence economic growth in the less-developed world (Example 8.2). This study by Adelman and Morris is a major statistical undertaking, involving as it does the identification of a large number of potentially relevant variables, the collection of a vast quantity of data, and the application of a powerful computational method to sort the sources of variation within the data set. They arrive at a set of statistically significant correlations between a number of the variables and economic growth. What precisely is the significance of this effort, however? It should be noted that there is enormous variation across the countries surveyed. To make sense of this argument, we must postulate that economic growth is a structured process that is responsive to a variety of social, political, and economic

factors. And here we encounter a serious shortcoming in the Adelman-Morris study: Their analysis attempts to identify causally relevant variables without using a theory of the mechanisms through which various factors influence the growth rate.

Turn now to a more pervasive problem—the quality and comparability of the data on which a statistical argument depends. The Adelman-Morris study raises several problems in this regard. First, in some cases there are conceptual problems in the measurements that they attempt to collect. (They try to classify political regimes on the basis of the types and varieties of political parties that are active, although it is not clear that this is a useful comparative framework.) Second, some measures are conceptually clear but difficult to collect, and as a result there is little correspondence between the measure and the real value of the variable in the population. (Infant mortality figures, for example, are poorly collected in impoverished countries, implying that these figures may underestimate the extent of infant mortality.) Finally, some important economic variables that are monitored by national governments—e.g., the savings rate—may be defined differently by different national statistical agencies, resulting in a meaningless cross-country comparison.

This study thus illustrates both the strength and some of the shortcomings of a purely inductive approach to social causation. It succeeds in identifying some causal factors that influence the phenomenon to be explained, but at the same time further causal analysis is required to distinguish between genuine and spurious correlational data. We must identify the causal mechanisms that are at work among the various factors.

PHILOSOPHICAL GROUNDS OF STATISTICAL EXPLANATION

The preceding discussion provides a basic knowledge of some of the statistical concepts used in the social sciences to analyze and explain social phenomena. We may begin a more philosophical discussion of statistical reasoning by asking a fundamental question: What is the role of statistical arguments in social science? There are several general answers to this question. First, statistical tools may be used to empirically evaluate causal hypotheses—that is, statistical analysis can be a method of *hypothesis testing*. Suppose that we hold, on theoretical grounds, that rapid social change is a cause of third-world civil unrest. If this hypothesis is true, there should be a correlation between rapid social change and civil unrest. A statistical study of a sample of countries is a particularly direct way to test if this theoretical expectation is borne out.

Rapid social change and civil unrest are not directly observable, however, so to conduct such a study we need observable variables that correspond to these concepts. We need, that is, to *operationalize* the theoretical hypothesis in terms of observable variables. We can empirically evaluate the hypothesis that **A** is a cause of **B** through a study that operationalizes **A** and **B** in terms of observable variables A^* and B^* and then determines whether B^*

is correlated with A*. Suppose, then, that we take "rate of population movement from rural to urban residence" as a measure of the rapidity of social change and "violent incidents involving five or more participants" as a measure of civil unrest. If we find that there is a correlation between the variables and if we judge that these are plausible measures of the causal factors in question, then we have some confirmation for the causal hypothesis. (Note, however, that such a finding does not establish the truth of the causal hypothesis for it is possible that both variables are the collateral effects of some third causal factor.)

Suppose, however, that we find that A* and B* are not correlated; it is not true that societies with a higher value for A* also have a higher value for B*. Does this refute the claim that A is a cause of B? It does not for there are several other possible explanations of this null discovery. It may be that we have not chosen appropriate measures of the causal variables; A* and B* may not be good surrogates for A and B. Or there may be a causal relation between A and B, but it is one that is highly context-dependent: There is a third factor C that, if present, facilitates the causal process from A to B and, if absent, prevents this causal process. Now, suppose that C is present in about half the cases under study. If this is true we will not find that countries with high values for A* will tend to have high values for B*; there will be a low correlation between A* and B*. A new study incorporating C would, on these assumptions, show that A in the presence of C is highly correlated with B, and A absent C is not. But if we happen not to identify C as a causally relevant variable, our statistical analysis will produce a false inference of no causal relation between A and B.

The lesson of this example is an important one: A statistical study can provide empirical grounds for accepting or rejecting a causal hypothesis, but the statistical findings themselves are not final or conclusive. Study of covariance among factors is a useful tool for investigating causal hypotheses, but it is always possible that the causal hypothesis is true although the corresponding statistical test is negative.

Consider a second common use of statistical analysis—as a preliminary way of probing a complex range of social phenomena for underlying regularities. Here the goal is to discover regularities as a first step in establishing causal relations. The Adelman and Morris study (Example 8.2) represents a good illustration of this use of statistical analysis; they cast a wide net over a range of potentially relevant social, economic, and political variables and then attempt to see which of these are significantly correlated with the dependent variable under scrutiny (the rate of economic growth per capita). Their discovery that a small handful of features is highly correlated with economic growth then indicates where further theoretical analysis should be conducted. The next task is to formulate a theory of the causal processes through which these social variables influence the rate of economic growth. In this approach a statistical study may be seen as an exploratory work aimed at uncovering the patterns present in the empirical phenomena. The resulting patterns can then serve as the basis for hypothesis formation about underlying causal processes.

STATISTICAL ANALYSIS AND COMPARATIVE STUDIES

We have now examined causal investigations in a range of levels of concreteness. In Chapter 2 we examined the case-study method and the comparative method. In the case-study method the investigator considers a particular event or process—for example, the occurrence of the Chinese revolution—and examines the history of this event in detail, attempting to identify the causally significant factors that led to the occurrence and particular character of the event. The comparative method, we found, is a common and powerful technique through which social scientists attempt to identify social causes. This method isolates a small number of cases—for example, the Chinese, Russian, and German revolutionary movements—and then tries to identify the causal processes that lead to different outcomes in these cases. The statistical method, by contrast, involves a large number of cases—perhaps a hundred countries—and a set of abstract variables designed to apply across the highly diverse settings of the various cases. What is the relation between case studies, comparative studies, and statistical studies?

The general perspective I adopt here—in line with the methodological pluralism urged throughout—is that each approach has its own strengths and weaknesses, each is strengthened by the resources of the other, and each has a range of research topics to which it is best suited. At the same time I will hold that the comparative method is in one sense more fundamental than either the case-study or the statistical method. It permits the investigator to identify causal processes that are basic to explanation without unavoidable reference to either case or statistical studies. Case studies, by contrast, require background theoretical beliefs that could only come through knowledge of a number of different social settings, implying either comparative or statistical studies. And statistical studies require knowledge of particular cases in order to sort out the causal mechanisms that are in play, underlying the regularities discovered by statistical analysis. Either explicitly or implicitly, then, statistical studies must be supplemented by hypotheses about causal mechanisms that can only come from case or comparative studies.

We can best study the relation between statistical and comparative studies by considering an example of each. Recall Atul Kohli's comparative study of poverty reform in India (Example 2.6). Kohli tries to identify one or more variables that account for the varying successes of poverty reform in the context of a theory of the politics of reform in India. His work, based on a detailed comparative examination of the politics of reform in three Indian states, reflects a high level of detailed knowledge about the economic arrangements in each state, the political organizations active in each state, the ways in which state bureaucracies function in each state, and the official policies of each state government. In other words this is a very detailed study of the particulars of the politics of poverty in these three settings.

Using this analysis, Kohli comes to the conclusion that the governing variable in determining whether state poverty policies are successful is tied to the ideology of the regime in power, its internal coherence and discipline, and its organizational reach and competence.

Now contrast this study with that of Adelman and Morris (1973). (This is the sequel to the study presented in Example 8.2.) Adelman and Morris are concerned with a problem that is closely related to Kohli's: What are the economic and social features that determine the distributive characteristics of economic development processes across countries? They employ a large country study, involving 74 nations and 48 indicators. And, using a statistical technique (discriminant analysis), they determine that a small cluster of these indicators are closely correlated with distributive consequences—rate of improvement in human resources, direct government economic activity, socioeconomic dualism, potential for economic development, per capita GNP, and strength of the labor movement (Adelman and Morris 1973:184). This conclusion is clearly related to that reached by Kohli, who finds that the regimes that exert disciplined policy attention to the problem of the poor have the greatest effect on poverty alleviation. The first two factors listed by Adelman and Morris are associated with this feature of regime policy choice. And the final factor listed by them (the strength of the labor movement) is consistent with the idea that regimes that take power in the environment of a significant labor movement will be more prone to poverty reduction than others.

So there is a substantial degree of coherence between Kohli's findings and those of Adelman and Morris. However, there are several important shortcomings to the second study that can best be addressed through the comparative method. First, there is the problem of the coarseness of the analytical net that Adelman and Morris throw over the data: Their study ranges over several decades and dozens of countries. It may be, however, that there are regionally or temporally specific processes that must be identified if we are to understand the mechanisms of economic growth and distribution but that operate more finely than the factors measured by their study. Second, their study unavoidably employs highly abstract descriptive categories, although there is little reason to expect that these will capture the causally relevant factors. Finally, at its best their study identifies causally relevant variables without determining the causal mechanisms that connect them. Take their first factor (rate of improvement in human resources). Their study shows that countries that effectively direct social efforts toward improving human resources have more equal distributive outcomes than those that do not. But this finding sheds no light on the underlying question: Why do some countries in fact exert this sort of effort toward improving human resources? By contrast, Kohli's conclusions, if true, do provide an answer to this question: The regimes that make poverty reduction a high-priority goal select improvement in human resources as an instrument to reach this goal.

The two approaches also differ in their treatment of the causal field as related to distributive justice. The relevant factors may be impossible to

discern without in-depth study of a small number of cases. Kohli's investigation does not presuppose that we can identify the causally relevant variables in advance; instead, they emerge as findings. He brings with him a set of theoretical hypotheses about political processes and economic change, but the set of causal factors that he arrives at—leftist ideology, political competence, leadership, intraparty coherence—emerges from his study of the data. Adelman and Morris, however, are forced to identify the causal field in advance because their study requires that they collect data in predefined categories. Consider the following thought experiment: If Kohli's hypothesis is correct, how would Adelman and Morris's analysis come out? They cast their net widely, but if they have not identified left-right regimes as a variable, they will have missed a strong correlation that is present among the phenomena. Indeed, if Kohli is right, it is the decisive correlation. The variables that they consider do not include what turn out to be the causally salient factors: regime type, ideology, and organizational competence. None of the variables used by Adelman and Morris correspond closely enough to these features of state activity to allow them to arrive at the most fundamental causal relations. In general, then, it is unjustified to suppose that we can identify the causal field a priori for a particular circumstance, which suggests that large statistical studies must be supplemented by smaller-grain comparative studies that shed light on the social and political mechanisms underlying the large-scale processes identified through statistical analysis.

ARE THERE AUTONOMOUS STATISTICAL EXPLANATIONS?

It is sometimes held that the discovery of statistical regularities is the beginning and end of explanation. In other words to explain a phenomenon is to show that its occurrence in observed circumstances conforms to an underlying statistical regularity. This view differs sharply from the one argued in this book because it denies the importance of identifying underlying causal mechanisms. So let us confront this position head on. Are there autonomous statistical explanations—that is, explanations that do not depend on a causal story? Is the discovery of a strong statistical correlation ever an explanation of an event or pattern? Is it ever reasonable to say that we have explained an event or regularity by showing that it is a member of a class that has different conditional probabilities from absolute probabilities? Sickle cell anemia is relatively rare among the general U.S. population but substantially more common among Afro-Americans. We find that Tony has this illness, but have we explained this circumstance by discovering that he is Afro-American? Or suppose we find that GNP per capita is strongly correlated with the proportion of college-educated adults in the society. Does this finding constitute an explanation of *either* the fact that France has a high proportion of college-educated adults *or* the fact that it has (by international standards) high GNP per capita? In each case it is most reasonable to suggest that the discovery of the correlation or abnormal

probability distribution is the beginning of an explanation but almost never the final explanation. In general I propose the following principle on probabilistic and statistical explanation: A discovery that refutes the null hypothesis is explanatory only if it leads us to a plausible causal mechanism producing the significant relationship between the variables. When we have discovered a pattern in a data set, we have laid the basis for an explanation of the phenomena in question. But to have a satisfactory explanation we must be able to identify, at least approximately, the causal mechanisms that underlie the statistical regularities.

To begin with, there is the problem of spurious correlation. Some statistical techniques can help to exclude this possibility (by collecting data for other conditional probabilities—for example, the incidence of cancer among non-smokers with nicotine stains).[4] But the most direct way of excluding the possibility of a spurious correlation is to discover the causal mechanism connecting the variables. Correlation between x and y is prima facie evidence of causal connection, but to establish a causal connection it is necessary to exclude the possibility that both are the effects of some third condition. Analysis of the mechanism of causation is an effective way of supporting judgments of causation based on correlational data for once we have a theory of the process by which condition x produces condition y in typical circumstances, we also have a theoretical basis for judging that the correlation is genuinely causal rather than spurious.

There is a deeper consideration as well that militates against rock-bottom statistical explanations. Recall the discussion in Chapter 1 about "why-necessary" questions. The demand for an explanation of an event or regularity typically involves this question: Why did this event come about, given the circumstances at the time of its occurrence? This is a demand for a causal story, which in turn requires an account of the laws and mechanisms through which antecedent conditions brought about the explanandum. We are satisfied with an explanation when we are confident that we have identified the antecedent conditions C_i and laws L_i that brought it about. An adequate causal story permits us to make counterfactual judgments—if conditions C_i had not been present, E would not have occurred. And it provides the basis for making judgments of causal irrelevance about other factors: If condition D was present or not, E would still have occurred.

Whether there are autonomous statistical explanations, then, reduces to this question: Does a well-established statistical association between two variables yield a causal story? It does not. The statistical association does not establish the presence or character of a set of causal mechanisms connecting the variables. There may be no causal mechanism leading from one variable to the other; each may be the effect of some third variable (as was the case in panel I, Figure 8.3) or the association between them may be artifactual (panel VI, Figure 8.3). The statistical association also does not establish the basis for counterfactual judgments because collateral effects are not necessary or sufficient conditions for the occurrence of one another. Finally the statistical result does not establish the causal irrelevance

of other factors. On these grounds, then, it is reasonable to conclude, as I do, that statements of statistical regularities are not in themselves explanatory.

The discovery of a statistical regularity among variables rather constitutes an empirical description of social phenomena that itself demands explanation. As we have seen above it is possible that the correlations reflect experimental artifact or collateral causation, so the statistical findings themselves do not permit us to conclude what explanatory relations obtain among the variables. (Note the similarity between this conclusion and the parallel findings concerning functional and structural explanation in Chapter 5.)

NOTES

1. For discussion of Korean development along these lines, see Mason et al. (1980).
2. This is not to imply that the data comes before the theorizing for it is plain that the researcher has had to employ some set of provisional hypotheses about the causal structure of the domain he is working with in order to identify appropriate features to study.
3. It is also possible to perform regressions on three or more variables; the task is more computationally demanding but in principle the same. *Multivariate regression analysis* for n variables involves constructing an n-dimensional plane (corresponding to the two-dimensional line in the two-variable case) that best fits the n-dimensional scatterplot of data. And it is possible to perform nonlinear regressions on two or more variables. Nonlinear regression finds a curvilinear function of specified form (e.g., exponential, quadratic, or logarithmic) that best fits the data set (once again by minimizing the variance around the function).
4. Herbert Simon addresses this problem in "Spurious Correlation: A Causal Interpretation" (1971).

SUGGESTIONS FOR FURTHER READING

Blalock, H. M., Jr., ed. 1971. *Causal Models in the Social Sciences.*
Bohrnstedt, George W., and David Knoke. 1988. 2d ed. *Statistics for Social Data Analysis.*
Suppes, Patrick. 1984. *Probabilistic Metaphysics.*
Tufte, Edward. 1974. *Data Analysis for Politics and Policy.*