EDITED BY

TUKUFU ZUBERI AND EDUARDO BONILLA-SILVA

WHITE LOGIC, WHITE METHODS

RACISM AND METHODOLOGY

7

Deracializing Social Statistics

Problems in the Quantification of Race

Tukufu Zuberi

This study should seek to ascertain by the most approved methods of social measurement. . . .

—W. E. B. Du Bois, "The Study of The Negro Problems" (1899)

In scholarly circles, demographic and statistical interpretation of racial differences has taken on an almost sacred quality. As a result, demographers and other scholars have forgotten—or perhaps have never realized—that the social concept of race affects *how* we interpret quantitative representations of racial reality. Moreover, many quantitative studies of racial differences fail to place race within a social context, thus allowing the faulty assumption that the existence of race relations could be benign.

In the beginning of this century, empirical social scientists took a eugenic perspective toward race. Du Bois was an exception to the accepted view about race among empirical social scientists. Du Bois was of the opinion that the best minds should study the problem of race according to the best methods. He thought that statistical analysis could help us gain a concrete understanding of the social status of the African American population. He formulated the first empirical refutation of eugenic and social Darwinist thought. After conducting an empirical study of African American life in a modern city in *The Philadelphia Negro*, Du Bois illustrated how *biological notions of African inferiority were grounded only in ideology.* However, Du Bois's contribution has been ignored by most sociologists, and its theoretical significance to understanding modern society continues to be underplayed. This chapter demonstrates the theoretical significance of Du

Bois's tradition of scholarship on our understanding of racial statistics, particularly his contributions to the understanding of quantitative data in societies where race is an "essential" variable.[1] Unfortunately, among social statisticians, including demographers there has developed an implicit tendency to accept the underlying logic of racial reasoning. In part the statistical logic of justifying racial stratification has resulted from a lack of critical theory among social statisticians, and a tendency to avoid reflexive discourse with statisticians in general and with other areas of the social sciences, such as African and African Diaspora studies in particular. This chapter describes the scientific birth of racial reasoning in statistical analysis, what I have learned from reading statistics and practicing demography, and how this shapes a new logic for the quantitative study of racial stratification that has developed between Du Bois's *Philadelphia Negro* and the reflexive discourse among statisticians.

BIRTH OF SOCIAL STATISTICS

Francis Galton was a key intellectual power behind the modern statistical revolution in the social sciences (Stigler 1986, chapter 8; Kevles 1985, chapter 1). His imaginative ideas are the conceptual foundation of eugenic thought and inspired much of the early work in social statistics. Galton's research in *Hereditary Genius* (1892), *English Men of Science* (1874), and *Natural Inheritance* (1889) all suggested that genius and success are inherited and that this process could be measured statistically.

Galton used statistical analysis to make general statements regarding the superiority of different classes within England and of the European-origin race, statements that were consistent with his eugenic agenda. In the 1892 edition of *Hereditary Genius* he outlined that "the natural ability of which this book mainly treats, is such as a modern European possesses in a much greater average share than men of the lower races. There is nothing either in the history of domestic animals or in that of evolution to make us doubt that a race of sane men may be formed, who shall be as much superior mentally and morally to the modern European, as the modern European is to the lowest of the Negro races" (27). While this statement may be considered insignificant in the context of Galton's overall statistical contribution, it is fundamental in understanding the direction and purpose of his causal explanations, and placing Du Bois's empirical response in a historical context (Zuberi 2000).

In 1875, Galton wrote an article, "Statistics by Intercomparison, with Remarks on the Law of Frequency of Error," that suggested that measurement of two values—the median and the quartile—was sufficient to characterize

or compare populations. For Galton this meant different populations could be represented in a bell curve of all populations.

In 1885, Francis Ysidro Edgeworth (1845–1926) developed a test to ascertain whether different populations existed within the bell curve (Edgeworth 1885). Edgeworth's test adapted the bell curve to assess the "significance" of differences between the subpopulations. He used Galton's 1875 formulation as a vehicle to employ classical statistical theory in understanding social statistics.

In 1889, Pearson met W. F. R. Weldon, the chair of Zoology at University College, London. Weldon was attempting to adapt Galton's methods to the study of evolution in wild populations. Weldon turned to Pearson with a series of questions to which Pearson responded in a series of papers known as *Contributions to the Mathematical Theory of Evolution*. Published between 1894 and 1916, the series was retitled *Mathematical Contributions to the Theory of Evolution*, after the second in what became a series of about nineteen papers.[2] Pearson's elaboration of Edgeworth's theorems advanced correlation theory into the main of social statistics.[3] In Pearson's third and fourth papers of the *Mathematical Contribution to the Theory of Evolution*—"Regression, Heredity and Panmixa" and "On the Probable Errors of Frequency Constants and on the Influence of Random Selection on Variation and Correlation" (which he coauthored with L. N. G. Filon)—he provided a basic formula for estimating the correlation coefficient and a test of its accuracy.

In 1897, George Udny Yule provided the conceptual and statistical expression that completed Galton's project to apply statistics to the study of society.[4] Yule was one of the first social statisticians to demonstrate the relationship between regression and least squares estimates. Yule was exceptional in that his application of statistics focused on causation in social sciences. Yule extended the application of regression in one of the first regression analyses of poverty. Interestingly, Yule's analysis provided support for the conservative position advocated by Malthus at the beginning of the century (Yule 1899; Stigler 1986, 355–57). He argued that providing income relief outside the poorhouse increased the number of people on relief. By 1920, his approach to multiple correlation and regression predominated in social science research.

Social statistics took another intellectual leap when Ronald A. Fisher published *Statistical Methods for Research Workers* in 1925, and *The Design of Experiments* in 1935. Both books had a tremendous impact on the teaching and practice of statistics. His work clarified the distinction between a sample statistic and population value, and he emphasized the derivation of exact distributions for hypotheses testing. He is also credited with introducing the modern experimental design and statistical methods to social sciences. These innovations remain as the basis of social statistics.

CAUSAL REASONING AS RACIAL REASONING

The experimental notion of causal inference is the implicit guide in the selection of observations in quasi-experimental research and in the model selection, design, and statistical analysis of nonexperimental data from sample surveys and samples from census data. Most researchers, however, do not appear to appreciate the consequences of adopting the experimental model as a guide in the design, collection, analysis, and interpretation of social science data. Most social scientists use experimental language when interpreting empirical results, thereby entailing a commitment to the experimental mode of analysis introduced by Fisher (Holland 1986; Cox 1992; Sobel 1995; Smith 1990). I am intentionally explicit in this section so that I can present the fundamental elements of the causal process in statistical modeling.

Because most social science researchers study causal effects for the purpose of making inferences about the effects of manipulations to which groups of individuals in a population have been or might be exposed, causes are only things that can, in theory, be manipulated or altered. This recognition forces us to consider the individuals or units we study and our ability to alter or treat these individuals. This type of clarity is essential yet absent in most policy-oriented social research. In most policy research, decisions to manipulate the real world often depend on social researchers' causal inferences. A lack of clarity in the statistical analysis of racial processes has contributed in great measure to the confusion about how to resolve issues of racial stratification.

Cultural studies in anthropology, history, literary criticism, sociology, philosophy, and African studies have questioned and criticized the concept of race. Statisticians are in the process of an important discussion on the issue of using attributes like race in social statistics (See Holland 1986; Cox 1992; Sobel 1995; Rosenbaum 1984). This discussion has not focused on the issue of the conceptualization of race. It nevertheless places a considerable theoretical burden on social statisticians who use race as a variable in their statistical analysis to predict social outcomes. Most social statisticians have not yet integrated the latest statistical research in their statistical analysis of race.

Statistical populations consist of observed measures of some characteristic; yet no observational record can capture completely what it is to be a human being. Researchers employ observational records to define abstract concepts like race. The researchers or the subject (as in self-administered surveys or censuses) can make the observation; however, the researchers and the purpose of their study determine the meaning of the record. This is how empirical research reifies race. If we have records of racial classification, the population of races rather than the population of persons is open

to statistical investigation; yet in social statistics it is always a mistake to think of a population of races as a genetic population. A population of races in this sense is a statistical concept based on a politically constructed measure. Deriving a statistical model of social relationships requires an elaborate theory that states explicitly and in detail the variables in the system, how these variables are causally interrelated, the functional form of their relationships, and the statistical quality and traits of the error terms. Once we have this theoretical model, we can estimate a regression model.[5] Rarely, however, does social science research provide the level of theoretical detail necessary to derive a statistical model in this manner.

The alternative is a data-driven process. To derive a statistical model from data we assume the model is a black box and "test" it against our empirical results. Both the theory- and data-driven statistical models attempt to provide a parsimonious and generalizable account for the phenomenon under investigation. Statistical models attempt to provide a rigorous basis—rooted in abstract statistical theory—for determining when a causal relationship exists between two or more variables in a model. However, unless we start with prior knowledge about the causal relationship, the calculation of the regression equation refers to a regression model and its system of equations, not to the "real" world the model purports to empirically define!

The language of causation originates in the experimental framework of modeling causal inference (Holland 1986; Cox 1992; Sobel 1995; Rosenbaum 1984). Statistically, causation has a particular meaning. Measuring the effects of causes is done in the context of another cause, hence **X** causes **Z** relative to some other cause that includes everything but **X**. In the context of causal inference each individual in the population must be potentially exposable to any of the causes. And, as the statistician Paul Holland notes, "the schooling a student receives can be a cause, in our sense, of the student's performance on a test, whereas the student's race or gender cannot" (Holland 1986, 946; also see Holland, chapter 5 in this volume). For example, being an African American should not be understood as the cause of a student's performance on a test, despite the fact that being African American can be a very reliable basis for predicting test performance. The logic of causal inference itself should give every nonpartisan scholar reason to avoid flamboyant rhetoric about the genetic-based cognitive causes for racial and gender stratification.

Race and gender as unalterable characteristics of individuals are inappropriate variables for inferential statistical analysis (see Holland, chapter 5 in this volume). Statisticians are beginning to question and criticize the use of such attributes—unalterable properties of individuals—in inferential statistical models (see Holland, chapter 5 in this volume; Cox 1992; Rosenbaum 1984; Sobel 1994, 1995). Most social statisticians, however, continue to treat race and sex as an individual attribute in their inferential

models. *Statistical models that present race as a cause are really statements of association between the racial classification and a predictor or explanatory variable across individuals in a population. To treat these models as causal or inferential is a form of racial reasoning.*

FROM CAUSATION TO ASSOCIATION AS AN ASPECT OF DERACIALIZED STATISTICS

I suggest that we reconsider the notion of causation in the study of race. It may be better to interpret these impacts of race as the association of racial stratification on an individual's mortality outcome. In this context the attenuation of the race variable that occurred via the introduction of wealth into Menchik's (1993) model would imply the way that race interacts with wealth in its association with mortality.

Association may not prove causation; however, it may provide the basis for support of a causal theory. Association is evidence of causation when it is buttressed with other knowledge and supporting evidence. When we discuss the "effect of race" we should be more mindful of the larger world in which the path to success or failure is routinely influenced by other contingencies or circumstances.

CONCLUSION

Some will argue that causal models are the best way for social scientists to make public policy statements. They will say that statistic models used in this way allow us to elaborate on how and what produces particular effects. This may all be true, but human knowledge is uncertain and imperfect and it is not clear how statistical models contribute to this uncertainty (Freedman 1987). Interpreting the results of a causal inference is validated by an underlying causal theory. If the theory is rejected, the interpretations have no foundation. Decision makers may like causal models that appear to support their position on important questions, but it is the continued misuse of the statistical models by scholars that gives the process scientific credibility.

Some will argue that the causal language used by many social scientists is not reflective of an unarticulated causal model but is simply the careless use of language. However, this tendency has significant implications for how results are interpreted in policy circles and within the professional discursive mode.

Race, or more specifically the process of racialization, may be the stimulus for how other individuals respond or interact with persons so characterized. The examination of discrimination and prejudice provide a solu-

tion to the trap of racial reasoning. One example of such research is the study of racial attitudes (see Bobo and Smith 1998). Du Bois was aware of this process and gave it direct attention in his work. Another example of research of this process is found in the economic study of statistical racism and the demographic study of environmental racism.[6] The causal factors in statistical racism are discriminatory practices by employers not the races of the people involved. The causal factors in environmental racism are discriminatory practices by institutions in determining the location of hazardous wastes cites.

The solutions to the problem in the analysis of unalterable characteristics like race and gender in inferential statistics requires a shift in perspective. We should describe race and gender as events that represent the acquisition of the attributes for each individual. Thus, being classified one way or another might have a particular impact. As our understanding of what race or gender means changes, so too does our statistical analysis.

The National Liberation movements to decolonize Africa, Asia, and Latin America, the civil rights movement to deracialize civic society in the United States, and the antiapartheid movement in South Africa all questioned the Eurocentric division of humanity. We would benefit by continuing to question the racialization of identity. Using racialized census, survey, or other social data is not in and of itself problematic. But the racialization of data is an artifact of both the struggles to preserve and to destroy racial stratification. Before the data can be deracialized we must deracialize the social circumstances that created race. Statistical research can go beyond racial reasoning if we dare to apply the methods to the data appropriately.

NOTES

1. For a more extensive examination of the social statistical analysis in *The Philadelphia Negro: A Social Study* see my paper on this classic community study (McDaniel 1998; Zuberi 2004).

2. The long series of memoirs entitled *Mathematical Contributions to the Theory of Evolution* was reissued by the Trustees of Biometrika in 1948 in a single volume. The selected papers consist of articles published in the *Philosophical Transactions of the Royal Society*, in the Drapers' Company Research Memoirs, and in *Philosophical Magazine*. Most of these papers cover Pearson's *Mathematical Contribution to the Theory of Evolution* series. The article published in *Philosophical Magazine*, "On the Criterion That a Given System of Deviations from the Probable in the Case of a Correlated System of Variables Is Such That It Can Be Reasonably Supposed to Have Arisen from Random Sampling," presents the first derivation of the distribution referred to as the chi-square. The chi-square test, a goodness-of-fit test, is considered Pearson's most significant contribution to statistical theory. In this work Pearson greatly expanded social statistics. He also expanded on Edgeworth's

significance test by measuring difference in terms of standard deviations. See: Pearson (1948), Stigler (1986).

3. Pearson's correlation coefficient, **r**, continues to be the most commonly used measure of correlation. When people use the term "correlation" without any other specification this is what they mean.

4. Yule's paper "On the Theory of Correlation" reconciled the theory of correlation with the method of least squares from the traditional theory of errors. See Yule (1897) and Stigler (1986, 348–58) for an excellent discussion of the importance of this paper for social statistics. Linking least squares and regression made the developments in simplifying the solutions of normal equations and the calculation of the probable errors of coefficients by astronomers and geodesists available to regression analysis among social statisticians.

5. I use the term "regression" in a broad sense to include logistic regression, regression analysis, regression analysis of survival data, ordinary least-squares regression, and so on.

6. It is important to distinguish between the "statistical" reasoning by an employer, and the statistical reasoning discussed in this chapter that focuses on social scientists engaged in a very different sort of statistical reasoning from that of an employer. The econometric "theory" of statistical racism maintains that racial preference of an employer for a "White" job candidate over a "Black" job candidate who is not known to differ in other respects might stem from the employer's previous statistical experience with the two groups (Phelps 1972). This analysis has been extended to the examination of the impact of affirmative action on employer beliefs and worker productivity (see Lundberg 1991; Coate and Loury 1993).

According to my argument, statistical racism is a difficult process to examine. An employer's assessment of the expected productivity of employees in less-favored groups may be wrong in a way that a longitudinal study could effectively demonstrate; however, such a study is still prohibitively expensive. We might argue that an employer's perspective is invalid because it does not incorporate certain systemic mechanisms of other types of prejudice, such as the effect of higher rewards themselves on the productivity of employees in a less-favored "race."

The analysis of environmental racism examines whether facilities for treatment, storage, and disposal of hazardous wastes are located disproportionately in communities of the less-favored race (See Bullard 1990; Anderton et al. 1994). While the study of statistical racism focuses on intentional prejudice and the study of environmental racism has focused on dumping prejudice, both depend on inequitable distributions as evidence of intentional prejudice.