Preface

Much of social science research is concerned with group differences and comparisons. When the attribute of interest is continuous, for example the differences in life expectancy between racial groups, or comparisons of earnings between men and women, we often summarize the comparisons in terms of means or medians. The usual parametric analysis of location and variation, however, provides a weak and unnecessarily restrictive framework for comparison. Consider the earnings distribution in the United States. Over the past 30 years, median real earnings have declined by about 10% and the variance in earnings has risen dramatically. Hidden behind these summary statistics are a range of important questions. Have the upper and lower tails of the earnings distribution grown at the same rate? Can we determine the role played by the decade-long freeze in the minimum wage? Is there anything more to the narrowing of the gender wage gap than the convergence in median earnings between the two groups? The information we need to answer these questions is there in the data, but inaccessible using standard statistical methods such as regression and Gini index summaries.

Inequality is a good example in this context, because it is a property of a distribution, rather than an individual. So it would be natural to expect that the statistical methods we use to analyze inequality should be focused on distributional analysis. In general, they are not. The traditional statistical methods used in the social sciences – based on the linear model and its extensions – are not designed to represent the rich detail of distributional patterns in data. They instead focus on modeling the conditional mean, with the residual variation often assumed to be homogeneous, and treated as a nuisance parameter. As a result, these methods leave most of the distributional information in the data untapped. The Lorenz curve and the Gini index, which do represent distributional patterns associated with inequality, are a special case of the methods outlined in this monograph.

With the emergence of Exploratory Data Analysis (EDA, Chambers, *et al* 1983; Tukey 1977) and the development of high speed computing and graphical user interfaces, there has been a movement towards more nonparametric and distribution-oriented analytic methods. A prominent feature of these methods is the use of graphical displays. This is not surprising, as the visual display is the analogue to the numerical summary once one leaves the world of parametric assumptions behind. For those social scientists who have made the transition from reams of output containing various summary statistics to the simple visual summary of the boxplot and the world of Chernoff faces, data will never look the same. Graphics exploit the power of our visual senses to convey information in a direct and unambiguous way. The running boxplot, empirical P-P plot and Q-Q plot provide substantial help for comparing distributions, but do not in themselves provide a comprehensive framework for analysis.

The methods developed in this monograph seek to bridge the gap between exploratory tools and parametric restrictions to put comparative distributional analysis on a firm statistical footing and make it accessible to social scientists. We start with a general nonparametric framework that draws on the principles of EDA. The framework is based on the concept of a "relative distribution," a transformation of the data from two distributions into a single distribution that contains all of the information necessary for scale-invariant comparison. The relative distribution is the set of percentile ranks that the observations from one distribution would have if they were placed in another distribution. An example would be the set of ranks that women earners would have if they were placed in the men's earnings distribution. The relative distribution turns out to have a number of properties that make it a good basis for the development of a general analytic framework. It lends itself naturally to simple and informative graphical displays that reveal precisely where and by how much two distributions differ. An example would be graphs that show the proportion of women in the bottom decile of the men's earnings distribution (47% in 1967 versus 20% in 1997 for full-time, full-year workers). The relative distribution can be decomposed into location and shape differences, and can also be adjusted in a fully distributional way for changes in covariate composition. One can thus examine whether the difference in men's and women's earnings is simply a location shift, or something more, and what impact the age composition has on the difference in the two distributions at every point of the earnings scale. The relative distribution provides principles for the development of summary statistics that are often more sensitive to detailed theoretical hypotheses about distributional difference. It does this all in a framework that can be exploited for statistical inference. The relative distribution can provide this general framework for analysis because it represents a theoretically rich and substantively meaningful class of data in a fundamental statistical form: the probability distribution.

The goal of this monograph is to present the concepts, theory and practical aspects of the relative distribution in a coherent fashion. We thus alternate the chapters on theory and methodological development with chapters that provide an in-depth practical application. Many of the application chapters are based on papers that have appeared in recent academic journals, including the *American Journal of Sociology*, the *American Sociological Review*, the *Journal of Labor Economics*, and *Sociological Methodology*.

These chapters perform the dual role of clarifying the intuition behind the techniques and highlighting how they can be used in contemporary theoretical and empirical debates in the social sciences.

There are several audiences that we hope will find this monograph useful. As written, the monograph is mainly intended for quantitative researchers in the social sciences – demographers, economists, sociologists, and those involved in prevention research – and statisticians who focus on methodology. Social scientists will find connections to many standard methods made here, including Lorenz curves, quantile regression and regression decomposition. For the statistical methodologist, this monograph pulls together a wide range of earlier developments that are related to the relative distribution, for example, probability plots (Wilk and Gnanadesikan 1968), comparison change analysis (Parzen 1977; Parzen 1992), the "grade transformation" (Cwik and Mielniczuk 1989; Cwik and Mielniczuk 1993), and the two-sample vertical quantile comparison function (Li, *et al* 1996). Because the comparison of distributions is fundamental in any quantitatively oriented discipline, however, the methods here will also be of interest to a broad group of non-social scientists. Biomedical scientists, for example, will find that the relative CDF is related to the receiver operating characteristics (ROC) curves used in the evaluation of the performance of medical tests for separating two populations (Begg 1991; Campbell 1994, and the references therein). The prerequisite background in mathematical statistics is relatively low, though the notation representing distributional concepts may be unfamiliar and somewhat daunting on first sight. The monograph is designed for use in a one semester course, and contains exercises at the end of each chapter. It can also be used for independent study by practitioners with a solid quantitative background.

We would like to acknowledge first and foremost the contributions that Annette D. Bernhardt has made to the development of these methods. The first seeds of this book were planted by a question she emailed to us nearly a decade ago. She was working on her dissertation then, a study of the impact of economic restructuring on the growth in earnings inequality in the United States. Finding the standard summary measures like the Gini index too blunt to discriminate between inequality caused by job growth at the top or the bottom of the wage distribution, she asked us if we knew of any better methods. The result was the development of the median relative polarization index (and its siblings, the upper and lower indices) now discussed in Chapter 5. Eventually, we came to recognize that the summand in the index was actually the more interesting quantity: the relative distribution itself. Almost all of the subsequent developments of the relative distribution framework were made in collaboration with Annette over the years, as attested by the journal articles on which the application chapters are based.

Our research during the writing of this book has been supported in part by the Russell Sage and Rockefeller Foundations. The effect can be seen throughout the book, but particularly in Chapter 8.

Many of the new results in Chapters 9, 10 and the appendices are due to the work of Paul Janssen. We have also benefitted greatly from interactions about distributional approaches with William Alexander, Mark Hayward, James Heckman, Eric Holmgren, Paul Janssen, Diane McLaughlin, Manny Parzen, Jeffrey Simonoff, and Marc Scott. Jeffrey Simonoff and Paul Janssen gave comments on (close to) final drafts of the manuscript. Charles Kooperberg provided the log-spline density estimation program. We would also like to acknowledge the support and encouragement provided by Ron Brieger, the late Clifford Clogg, Douglas Massey, Adrian Raftery, and Eric Wanner over the years. Their interest in this work helped to convince us that it was worth making the effort to develop new methods and place them in a broader context. Stefan Jonsson has provided truly heroic research assistance, with Icelandic assiduity. Finally, we would like to thank our editor at Springer, John Kimmel, for his patience and encouragement throughout the publication process.

The software for implementing a relative distribution analysis is available in two sets of macros: one for the S-PLUS statistical program, and the other for SAS. Both can be downloaded from the Relative Distribution website maintained at

`http://www.stat.washington.edu/handcock/RelDist`

This site also contains many of the data sets used in application chapters of the book, so that the reader can reconstruct the graphics and results presented here.

The authors can be reached via electronic mail at the Internet address `handcock@stat.washington.edu`.

Croton-on-Hudson, N.Y.

Mark S. Handcock
Martina Morris