Statistics for Social Science and Public Policy

Mark S. Handcock
Martina Morris

Relative Distribution Methods in the Social Sciences

Springer

Statistics for Social Science and Public Policy

Advisors:
S.E. Fienberg D. Lievesley J. Rolph

Springer
New York
Berlin
Heidelberg
Barcelona
Hong Kong
London
Milan
Paris
Singapore
Tokyo

Statistics for Social Science and Public Policy

Devlin/Fienberg/Resnick/Roeder (Eds.): Intelligence, Genes, and Success: Scientists Respond to *The Bell Curve.*
Handcock/Morris: Relative Distribution Methods in the Social Sciences.
Johnson/Albert: Ordinal Data Modeling.
Zeisel/Kaye: Prove It with Figures: Empirical Methods in Law and Litigation.

Mark S. Handcock
Martina Morris

Relative Distribution Methods in the Social Sciences

With 41 Figures

Springer

Mark S. Handcock
Department of Statistics
The Pennsylvania State University
311 Thomas Building
University Park, PA 16802-2111
USA
handcock@stat.psu.edu

Martina Morris
Department of Sociology
The Pennsylvania State University
614 Oswald Tower
University Park, PA 16802-6211
USA
morris@pop.psu.edu

Advisors

Stephen E. Fienberg
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
USA

John Rolph
Department of Information and
Operations Management
Graduate School of Business
University of Southern California
Los Angeles, CA 90089
USA

Denise Lievesley
Institute for Statistics
Room H. 113
UNESCO
7 Place de Fontenoy
75352 Paris 07 SP
France

Library of Congress Cataloging-in-Publication Data
Handcock, Mark Stephen, 1961–
Relative distribution methods in the social sciences / Mark S. Handcock, Martina Morris.
p. cm. — (Statistics for social science and public policy)
Includes bibliographical references and index.
ISBN 0-387-98778-9 (alk. paper)
1. Social sciences—Statistical methods. 2. Distribution (Probability theory) I. Morris, Martina, 1955– II. Title. III. Series.
HA29.H2488 1999
519.5—dc21 99-14664

ISBN 0-387-98778-9 Springer-Verlag New York Berlin Heidelberg SPIN 10557211

To sir, with love

——— W.M.M.

To Mary Cicerello and Gilbert McIntosh Handcock:
for showing the way.

——— M.S.H.

This page intentionally left blank

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. A link to the website is maintained by the publisher at

`http://www.springer-ny.com/stats`

under the heading "Author/Editor Home Pages." 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.psu.edu`.

Croton-on-Hudson, N.Y. Mark S. Handcock
Martina Morris

Contents

This page intentionally left blank

Chapter 1

Introduction and Motivation

1.1 Motivation

In an increasing number of social science applications, the comparison of an attribute across groups requires consideration of more than the usual summary measures of location and variation. Survey and census data on attributes, such as earnings, test scores, birth weights, and survival times, all contain a wealth of distributional information. Traditional methods for the analysis of such data rely heavily on measures that capture only differences in averages between groups or rough measures of dispersion over time. Such summary measures leave much of the information inherent in a distribution untapped. More recent exploratory data analysis techniques have provided important complementary tools for traditional methods, and have helped to change the way we look at data, check the assumptions of our models, and evaluate their performance. But methods that combine both the exploratory power of EDA and a framework for complex statistical inference and estimation remain rare. Our motivation for developing the relative distribution approach is based on this gap in existing statistical methodology.

The relative distribution is a statistical tool for fully representing differences between distributions. It provides a general integrated framework for analysis: a graphical component that simplifies exploratory data analysis and display; a statistically valid basis for the development of hypothesis-driven summary measures; and the potential for decomposition that enables one to examine complex hypotheses regarding the origins of distributional changes within and between groups. We demonstrate the use of the relative distribution for each of these analytic tasks in this book. The integration of the different analytic components in the context of full distributional information helps to clarify complex patterns and relationships in data, making the relative distribution approach well suited to emerging research questions in many fields.

The gender wage gap provides a good example of the limitations of traditional summary measures. Analyses of the earnings gap typically focus on statistics which summarize the location differential between women's and

men's earnings, e.g., the median earnings ratio graphed in Figure 1.1. The women's median is in the numerator, so the ratio represents the fraction of a dollar the median woman earned relative to the median man – about 55 to 60 cents by this measure for much of this period. While the earnings gap was stable from the late 1960s through the 1970s (and had actually been stable for close to 50 years), it began to narrow in the 1980s. This new trend generated predictions that gender equality might finally be moving within reach (Nasar 1992). Numerous articles in the popular and academic press chronicled this historic *upgrading* in women's earnings, speculating on its origins, and highlighting the breakthroughs women were making in high-profile professional occupations. But is the upgrading of women's earnings the real story here?

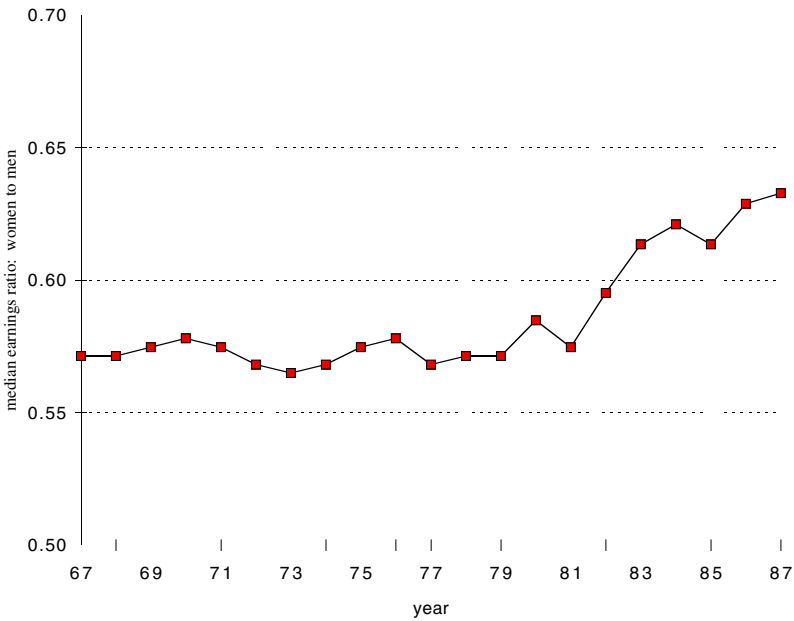

Fig. 1.1. The ratio of the median of women's wages to the median of men's wages for 1967–1987, full-time, full-year workers only.

A different picture emerges if the full distribution of women's earnings relative to men's is examined. This is presented as a relative decile series in Figure 1.2. The relative distribution graphed here is essentially a rescaled density ratio: the ratio of women's to men's probability of falling at each level of the earnings scale. In effect, each woman's earnings is assigned the rank it would have had in the men's distribution for that year, and

these ranks are plotted as a histogram. The histogram bin cutpoints are defined by the deciles of the men's distribution, so the frequency in each bin represents the fraction of women falling into each decile of the men's earnings scale over time. (The formal definition of the relative distribution is presented in Chapter 2.) If the women's and men's earnings distributions were the same, the relative deciles would take a uniform value of 10% over the earnings scale, because 10% of women earners would fall into each men's decile.

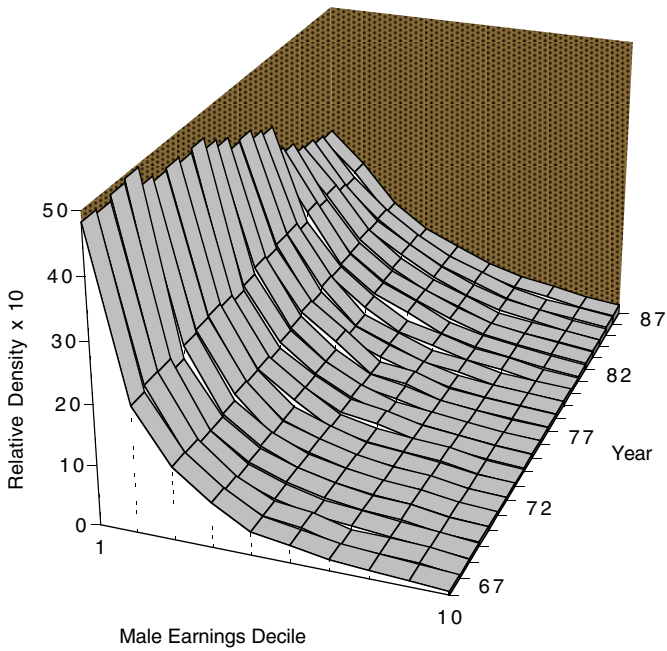

Fig. 1.2. The relative distribution of women's to men's wages 1967–1987. The relative deciles are plotted, see text for details.

In this case the relative distribution is far from uniform: nearly all of the mass in the women's distribution is concentrated in the lower tail of the men's distribution, and this does not change much over the 20–year period. In 1967, nearly half of all women earners were in the bottom *decile* of the men's distribution, and over 90% (the cumulative sum of all those in deciles 1–5) earned less than the median male worker. By 1987, this had changed somewhat, but over a quarter of the women still remained in the bottom decile of the men's distribution, and over 80% still earned less than the median male worker. The persistent absence of women in the upper tail of the men's earnings distribution is equally striking: less than 1% of

women fell in the top decile in 1967, less than 2% twenty years later.

While the median ratio graphed in Figure 1.1 suggests that women made progress during this period, the relative distribution makes it clear that progress was largely limited to women at the bottom end of the earnings distribution: three-quarters of the total change in relative density occurred below the male earnings median, half of it in the lowest decile alone. If upgrading is the story here, it is not the high-profile top earners that this story is about, but rather the lower profile earners at the bottom end of the distribution. The simple median wage trends in Figure 1.1 thus provide a very incomplete picture of the changes in earnings for men and women; obscuring the key features of the trend, inviting misinterpretation, and focusing research agendas on the wrong end of the earnings scale.

The patterns revealed by the relative distribution in Figure 1.2 provide substantially more information about key aspects of these changes. At the same time, this figure is more complicated to interpret because it represents the combined outcome of several factors: a baseline median earnings differential between the two groups, changes in this differential over time (the information conveyed by the median ratio in Figure 1.1), and changes in the shape of the men's and women's earnings distribution. The relative distribution can be decomposed into pieces representing each of these effects (Chapter 3). Decomposition makes it clear that the gains made by women at the bottom of the distribution are due more to the downgrading of men's earnings than to the upgrading of women's.

The substantive trends of interest, and the ones that need to be explained, are often neither visible nor statistically accessible when using techniques that are restricted to summary measures. Distributional methods enhance our understanding of the data and the phenomenon they represent, and our ability to pose the questions that should guide further research.

1.2 Principles of comparison

Suppose we wish to compare two distributions. What principles should be considered as the basis for comparison? Under what circumstances should one distributional comparison be defined as equivalent to another distributional comparison?

One important principle concerns the issue of scale invariance. For example, in the comparison of earnings in Figures 1.1 and 1.2, no adjustments were made for inflation. We can remove the effects of inflation, however, by transforming all of the earnings into 1967 (or 1987) "real dollars." So the earnings comparison can be based on one of several scales: the raw earnings scale, the 1967 real dollar scale, or the 1987 real dollar scale. Will the comparison be the same on these three scales? That depends on the measure chosen. The median ratio, for example, will be the same for all three scales. The median difference, on the other hand, will not.

The choice of measurement scale is a substantive choice, rather than a statistical one. Much of the work on economic inequality is based on measures which obey the "principle of (proportionate) scale invariance" (Schwartz and Winship 1980), which states that the comparison should not be affected by multiplying each individual's earnings by a positive constant. This principle preserves percentage changes in earnings, an approach that is consistent with an underlying assumption that a 5% change in earnings for someone at the bottom of the earnings scale is equivalent to a 5% change for someone at the middle or top. The Lorenz curve (Lorenz 1905) and associated Gini index are standard measures of inequality that satisfy the principle of (proportionate) scale invariance. In some cases, however, one could argue that an inequality measure should preserve the absolute dollar difference. For example, a 200% increase in earnings may not raise a person above the poverty line if their starting level is very low. In such cases, the true value of the dollar is measured by what it can (or cannot) purchase, and the proportionate change does not capture what needs to be measured (Rae 1981).

Dalton (1920) has argued that comparison of inequality should be approached by considering social welfare, as expressed via the form of a social welfare function. Given a social welfare function $U(g)$ that is an additively separable and symmetric function of individual earnings, we would prefer individual distributions according to their expected mean welfare ($E_Y[U(Y)]$). The measurement scale defined by the social welfare function would be the scale for absolute comparison.

If we knew the actual value (or *utility*) an attribute like earnings had for an individual, then the appropriate analysis would be based on the attribute data transformed to the scale in which units represented equal measures of utility. But the true utility scale of an attribute is hard to establish. The approaches described above assume that the utility scale is either a linear or logarithmic version of the original scale, but the true underlying scale may not have this level of regularity. We may, therefore, wish to consider methods that impose weaker assumptions on the underlying utility curves.

Suppose, for example, that all individuals share a common but unknown utility function for earnings, and we wish to compare distributions of utilities across groups rather than the distributions of raw earnings. If the utility function is unknown, it is useful for comparison measures to be invariant to different transformations of the data. Under what conditions will this invariance be met? Using the Lorenz curves, the conditions are quite restrictive. Unless the underlying common utility function is linear, the Lorenz curves for the utilities will be different than the Lorenz curves for the raw earnings. Thus, comparative analyses of inequality based on the Lorenz curves, or on indices of inequality derived from them, may lead to different conclusions for the utilities than for the earnings themselves. It can be shown that the ordering of distributions in terms of inequality by the Lorenz curves is not invariant to *any* (nondegenerate) transformation

to utiles; the Lorenz curve, and summary measures such as the Gini index, Pietra index, coefficient of variation, and Kakwani index are intrinsically tied to the original scale of measurement, up to a proportionate scale.

The relative distribution, by contrast, is invariant to all monotonic transformations of the original measurement scale. The utility scale will thus be accurately and equivalently represented by comparisons of the raw earnings, the log-earnings, or any other monotonic transformation of the earnings, as long as there is a common monotonic underlying utility function in the population. We shall call this the principle of *strong* scale invariance. When this principle holds, the relative distribution plays the primary role in comparisons, in the sense that it contains all the information necessary for comparing distributions, making the minimal assumptions necessary for valid comparison. Holmgren (1995) shows that under appropriate technical conditions the relative distribution is the maximal invariant – loosely speaking, any other quantity that contains more information does not satisfy the principle of strong invariance (cf, Lehmann 1983). This does not mean the relative distribution is inappropriate when the assumptions are not known to hold, only that comparisons may exist that cannot be exclusively expressed in terms of the relative distribution and may require additional characteristics of the original distributions.

An important issue for between-group comparisons is how the *comparisons* are to be ordered. Suppose we compare women's to men's earnings in 1967 and again in 1987. Have the two groups become more equal in 1987 than they were in 1967? One approach would be to compute a measure of within-group inequality, such as the Gini index, and compare the four resulting measures (one for each sex-time distribution). A more succinct approach, however, is to start with a measure that captures the between-sex comparison directly, and then compare the change in this measure over time. This is the approach taken by relative distribution methods. Holmgren (1995) shows that any preference ordering between pairs of distributions can be expressed in terms of preference between their relative distributions, under appropriate technical conditions. In this sense the relative distribution plays the same role for between-group comparisons as the Lorenz curve plays for within-group comparisons.

1.3 Description and summarization

The description, summarization, or analysis of a population (or data sampled from one) cannot proceed without making *some* assumptions about the underlying process. Imposing assumptions on the data carries risks as well, so the challenge is to find the right balance. Parametric approaches to modeling data impose a particular mathematical form on the underlying distribution. This parametric form allows concise descriptions and

summarization of the population and provides access to a statistical framework for estimation and inference. For the parametric representations of the data to be (at least approximately) valid, relatively strong – and often implicit – assumptions are required. If these assumptions are not met, substantively interesting features of the data may be obscured, and statistical inference invalidated. If weaker explicit assumptions can be made instead (e.g., smoothness of the underlying distribution) then one is free to estimate – rather than assume – the more detailed characteristics of the population, such as the distributional quantiles.

The key is to avoid making unnecessary or unjustified assumptions; to represent the data using approaches that are both flexible and robust to violations of the assumptions made. Relative distribution methods were developed with this philosophy in mind.

1.4 Graphical displays

A good graphical image conveys a remarkable amount of information, and the development of accessible graphical methods has dramatically changed the way we analyze data. The techniques for visualization have come a long way since the first simple hand-drawable tools proposed by Playfair (1786) and Tukey (1977). But the principles remain much the same as those articulated in these early works. Looking at the data permits the analyst to discover features that have both substantive and statistical importance, and to model these features in an informed way. Visual perception is a powerful tool to enlist in the service of data analysis. In some cases the perceptual task can be translated into an algorithm and automated (e.g., outliers and other case statistics). In many cases, however, direct visual inspection remains the most efficient and effective way to assimilate information and identify potential statistical problems.

Visualization techniques are at the heart of distributional comparisons, so it is not surprising that they will play a large role here. The amount of information contained in a distribution cannot be conveyed in any other way unless restrictive parametric assumptions are met. Even then, intuition benefits enormously from a simple graphical display. There has been a great deal of work on the development of graphical techniques for distributional comparison. Displays have evolved from simple density overlays to running boxplots, back-to-back stem and leaf plots, and percentile and quantile plots (both theoretical and empirical). Principles for effective display have also been systematically examined and defined. Some of these, such as parsimony (or absence of "chartjunk") and emphasizing the key features of the data, are similar to the principles that apply in traditional statistical analyses. Others, such as the preference for displays that code information into deviations from a straight horizontal line rather than a

sloped line, are a function of (hypothesized) perceptual competencies and are exclusive to visual displays.

Relative distribution methods include a set of graphical displays for comparing distributions that draw on much previous literature. Many of the principles of good visual display have been adopted and married to the techniques for interdistributional comparison. The techniques for relative distribution visualization range from simple back-of-the-envelope calculations for decile-based displays, to computer-intensive resampling methods for discrete data and imputation schemes for heaped continuous data. In our experience, even the simplest versions of these displays do a remarkable job in allowing the data to educate the analyst.

1.5 Numerical summary measures

While graphical displays are a key part of the relative distribution framework, summary measures remain an important tool for the comparison of distributional change. A good summary statistic makes it possible to provide a simple and precise answer to a substantive question, such as "has inequality in earnings grown significantly over the past 20 years?" or "has the upgrading in earnings been matched or exceeded by the downgrading?" Several summary measures are currently available for comparing aspects of distributional shape, e.g., the Gini index, the Theil index, and the coefficient of variation. The key challenge for such measures, however, is to summarize the right thing. As the "right thing" depends on the specific application, it would be useful to have a *framework* for developing summary measures, rather than a one-size-fits-all single statistic. The relative distribution provides such a framework and can be used as the basis for defining a wide and flexible range of summary measures. One of these measures – the mean absolute deviation of the relative distribution – captures the polarization or inequality that is the focus of the Gini index. It has the additional property of being easily decomposed into the contributions made by specific sections of the distribution (e.g., the upper and lower tails).

The generality of this framework for summary measure development is due to the fact that the relative distribution effectively captures *all* of the information that is necessary and sufficient for *strongly scale-invariant* comparison of distributions. Summary measures based on the relative distribution can be defined to capture the right thing, both from the theoretical and the statistical standpoint. By working with a measurement scale that preserves the detailed and nuanced properties of the distributions, the analyst is freed to focus on comparison or comparisons that are driven by theoretical interests, rather than methodological constraints.

Summary measures are no longer a luxury as the dimension of the analysis increases. Consider, for example, the gender wage gap data given above. Further analysis of these data naturally lead to decompositions that

(1) distinguish between location and shape changes in the two underlying earnings distributions; and (2) introduce explanatory covariates, such as education, work experience, and other workforce composition variables, each of which has a distribution of its own. The "education effect" in this context is a distributional effect: it captures the conditional *distribution* of wages at each level of education, rather than the conditional *mean*. These effects have both a composition component and a returns component, which parallel the traditional regression decomposition approach. The education composition effect compares the original distribution of earnings to the distribution obtained by reweighting the original education-specific conditional wage distributions by the new education profile. The education returns effect comprises the changes in the education-specific conditional wage distributions over time. Both components may induce a change in the location and/or shape of the earnings distribution. Graphical displays of the composition and returns components quickly proliferate, making summary measures a necessity. Again, the key issue is to ensure that these measures capture the features of substantive interest, revealing, rather than obscuring, the important structural features in the data.

Summary measures based on the relative distribution are robust to both outliers and to deviations from assumptions. This robustness follows from two properties of the relative distribution: the rescaling of the comparison distribution to the reference distribution and the absence of parametric assumptions. Outliers in either the reference or comparison distribution are not necessarily outliers in terms of the relative distribution. The rescaling maps the original units of both distributions to a rank measure (i.e., [0, 1]) moderating the influence of outliers. As a result, summary measures based on the relative distribution are less likely to be influenced by problem cases. The relative distribution, as well as the decomposition techniques, and natural summary measures in this framework are also fully nonparametric. They require minimal assumptions about the underlying distributions – either in terms of the individual distributions, or in terms of their relationship to one another. This actually distinguishes relative distribution methods from other nonparametric approaches, as most nonparametric approaches implicitly assume that the reference and comparison distributions have a well defined relationship to each other (e.g., are simply location shifted versions of each other) (Lehmann 1975).

1.6 Limitations

Relative distribution methods are not for small data sets. While the theory requires a minimum of 20 observations, realistically, the displays and methods are not well behaved with fewer than 200 observations, and the decomposition techniques become fully functional with 1000 or more. This is the tradeoff for the absence of parametric assumptions: full distributional

information requires data support for each quantile. With small to moderate data sets, the variation swamps the distributional information, so the uncertainty of the distributional estimates makes interpretation difficult. With more traditional parametric methods, we trade off uncertainty about the distribution for bias in the way the parametric distribution represents the distribution. For example, when we use means and variances to summarize the distribution, the implicit assumption is that these two parameters capture all of the information in the distribution. Parameter estimates based on small samples can be grossly misleading if the actual distribution is far from normal.

These methods are also not robust to the common data problem of "heaping." The heaping problem arises in the survey context when respondents report in round numbers rather than exact values. Classic examples can be found in self-reported data on income, age, and lifetime number of sex partners (Handcock, *et al* 1994; Heitjan and Rubin 1990; Morris 1993). Heaping can fundamentally change the quantile characteristics of a distribution, and the relative distribution graphical techniques in particular can be quite sensitive to this. Means and mean-based statistics are by contrast quite robust to heaping.

Full distributional information can also become overwhelming in the context of multivariate decomposition. This, again, is the price one pays for not assuming that the conditional mean and variance provide an adequate summary of the relationships of interest. As noted above, summary measures based on the relative distribution can be developed for multivariate analyses. These measures need not be used blindly, as the graphical displays of the relative distribution extend to all forms of the covariate decomposition.

The natural unit of analysis for relative distribution techniques is the population – not the individual. Some social scientists will find this natural; others will find it disconcerting. Measurement is clearly still anchored at the individual level, but virtually all of the displays and summaries reflect population attributes that have no analog at the individual level. The concepts represented, like inequality, are not properties of individuals. By making the group the unit of analysis, this approach takes the concept of a distribution as fundamental, rather than residual.

1.7 Organization of book

This book presents the techniques of relative distribution analysis in alternating chapters of statistical development and application. Interested readers with minimal statistical training should be able to work through the applications chapters independently to gain an understanding of the methods and their potential uses. Those interested in the statistical theory will find the chapters on measure development, estimation, and inference

contain all that is required to understand and apply these techniques. Exercises are provided at the end of these chapters to reinforce key theoretical points and provide an introduction to data analysis using relative distribution methods. Computer programs and data extracts used in the book are available via the Internet. Information on the website for this book is found in the Preface.

Chapter 2 provides a technical introduction to the relative distribution. It begins with a review of basic distributional concepts: probability mass functions for discrete populations, probability density functions (PDF) for continuous functions, cumulative distribution functions (CDF), quantiles (including percentiles and deciles), the quantile function, and transformations. These concepts are then used to define the relative distribution and its associated graphical representation. The chapter concludes with a review of the history and literature that contributes to the development and understanding of these methods.

Chapter 3 develops the technical basis for the decomposition of the relative distribution into location, scale, and shape shifts.

Chapter 4 applies the basic relative distribution methods to an analysis of the changes in the annual earnings distribution for full-time, full-year white male workers from 1967 to 1997. A decomposition of these changes into location and shape shifts shows both a decline in median real earnings, and a dramatic polarization in earnings over this period. This polarization is a key concept in the debates over growing inequality, and motivates the summary measures developed in the next chapter.

Chapter 5 discusses summary measures for the divergence between two distributions. It develops a decomposition of these measures into their location and shape components. It also develops a set of summary measures for capturing polarization in the distribution: the median relative polarization index, and its component upper and lower polarization indices.

Chapter 6 applies the divergence and polarization indices in an analysis of the changes in annual earnings for full-time, full-year workers by race and sex from 1967 to 1997. The analysis here focuses on the shape and location shifts that have taken place in the earnings distributions within each group. The relative distribution graphs, entropy summaries, and polarization indices provide a detailed picture of the earnings trends by group.

Chapter 7 extends the methods to the situation where covariates are measured on the individuals within the groups, and the comparisons need to be adjusted to take into account any differences in the distributions of these covariates.

Chapter 8 presents an application of the covariate adjustment procedures to the study of wage mobility, using data from two longitudinal panels of the National Longitudinal Survey (NLS). The location/shape decomposition is used to identify how wage growth has changed for the two cohorts. Covariate decomposition is then used to isolate the impact of differences in educational attainment between the two groups, and to contrast the trends

in mobility between more and less educated workers.

Chapter 9 develops the estimation and inference for the relative distribution, with emphasis on the relative CDF and PDF. The development begins with the case where the reference distribution is known, and then generalizes to the case where this distribution is estimated from the data.

Chapter 10 considers inference for summary measures based on the relative distribution. In addition to the measures developed in Chapter 5, measures are motivated by considering alternative hypotheses in testing situations.

Chapter 11 defines the relative distribution for discrete distributions and connects its properties to those of the relative distribution in the continuous situation. Estimation based on grouped data is discussed.

Chapter 12 applies the discrete data methods to an analysis of the changes in weekly hours worked for white male workers from 1980 to 1997. A significant polarization in work schedules is observed in the data. The covariate adjustment techniques are then applied to identify the role this work schedule polarization plays in growing wage inequality over the period.

Chapter 13 describes quantile estimation, focusing on quantile regression techniques. The most common model assumes the quantiles are a linear regression function of the covariates. The nonparametric quantile regression model is also considered.

Background material

Section 1.1

The analysis of the gender wage gap data is adapted from Bernhardt, *et al* (1995) and Morris (1996).

Section 1.2

Tukey (1977) emphasized the importance of looking at the data in all statistical analysis.

Simonoff (1996) discusses the value of methods motivated by prior beliefs in smoothness as a bridge between strict parametric methods and "purely" nonparametric methods.

Section 1.3

Du Toit, Steyn, and Stumpf (1986) give an overview of basic methods for analyzing and portraying data graphically, and their paper provides a useful set of historical references. They emphasize the goals of statistical communication and data exploration, noting the dangers of presenting only certain aspects of a data set in isolation.

Tufte (1983; 1990) presents many sophisticated and creative examples of graphical displays from diverse cultures and eras.

Section 1.5

Freedman, Pisani, Purves, and Adhikari (1991, Part II) give a conceptually clear and accessible development of the art of describing and summarizing univariate data. In particular, they look at histograms as a means of describing distributions, and motivate the use of means and standard deviations as summaries of distributional characteristics.

Computational issues

This section describes the availability of computer software to use the methods discussed in each of the following chapters. The software includes both commercial and free (shareware) resources. An important resource is the `statlib` archive at Carnegie–Mellon University; information on using `statlib` can be obtained by sending the message `send index` to the electronic mail address `statlib@lib.stat.cmu.edu`. In many instances, authors of the referenced papers will provide code of some sort upon request.

Exercises

Exercise 1.1. In Section 1.1 we considered the wages of full-time workers. The data for 1987 is in the file `cpswge87`. Calculate the usual summary statistics for the distribution of women's wages (e.g., mean, median, standard deviation and interquartile range). Repeat the process for men. Using these summaries, write a brief comparison of the two distributions.

Exercise 1.2. Calculate the median ratio of women's to men's wages based on the results of Exercise 1.1. Give an interpretation of it. Repeat the process using the mean ratio of women's to men's wages. Can you think of other numerical summaries that compare the wages of the two groups? Describe the value of each of these summaries, and the circumstances in which one or another may be preferable for use.

Exercise 1.3. Construct separate histograms of the women's and men's wages considered in Exercise 1.1. How does the histogram look if the default number of classes is used? Now create histograms with at least 50 classes. Compare the information provided by the two pairs of graphs.

Exercise 1.4. Repeat Exercise 1.3 using the logarithm of wages, instead of the wages themselves. Compare the descriptions of wages provided by the

graphs. Describe situations in which one or the other graph might be more appropriate.

Exercise 1.5. On the histograms constructed in Exercise 1.3, plot the normal distributions with the means and standard deviations calculated in Exercise 1.1. These are the best fitting normal distributions to the wage distributions. Are the normal approximations close to the true distributions? In which regions of the distributions are the approximations poor? Comment on the degree to which the numerical summary measures are appropriate summaries of the distributions.

Exercise 1.6. Using the graphical representations of the distributions of wages in Exercises 1.3–1.5, revise the comparison of the wage distributions given in Exercise 1.1. Do the histograms provide additional information about the distributions? Do they confirm the claims made about the distributions made in Exercises 1.1 and 1.2?

Exercise 1.7. If your software is capable, construct separate nonparametric density estimates of the women's and men's log-wages considered in Exercise 1.1. Compare the information provided by the graphs to the histograms in Exercise 1.4. For what purposes would you prefer the histogram estimates of the distribution to the other nonparametric density estimates? Do these nonparametric density estimates alter your descriptions of the distributions?

Exercise 1.8. Calculate the Lorenz curve for the distribution of women's wages in 1987. Does this curve change if the wages are expressed in 1967 dollars?

Chapter 2

The Relative Distribution

This book is mainly intended for quantitative researchers in the social sciences, so the prerequisite background in mathematical statistics has been kept to a minimum. For this chapter, social scientists familiar with statistical theory at the level of Rice (1995) should be able to follow the development with no difficulty. The more detailed results and proofs are given in Chapters 9-13 and the Appendices.

2.1 Basic distributional concepts

In this section we review fundamental concepts concerning distributions that underlie many of the ideas in the book. The objective is to present the requisite distributional theory as a coherent whole and to fix a standard notation.

Consider a measurement made on each member of a population of finite size. Unless otherwise noted, we will assume that the observation is a real number, as distinct from a *nominal* value such as race. The set of all possible values the measurement takes in the population is called the *outcome set.* We will assume the *population distribution* can then be described by listing each value in the outcome set along with the frequency with which members of the population take that value. For example, consider the hourly wages of full-time white women workers in the U.S. in 1998, measured to the closest penny. The distribution is the list of each value the wage takes (e.g., \$0.00, \$0.01, \$0.02, ...) along with the number of women with that wage. The *relative frequency distribution* replaces the frequency with the relative frequency (i.e., proportion) of members taking the value.

Probability Mass Function

Let X denote the value for a member of the population selected at random from the population. Then X is a random variable taking on values from the outcome set with probability given by the corresponding relative frequency. In this case X is a *discrete* random variable as it takes on only a

finite number of possible values. The *probability mass function* of X is then a listing of each value x, say, in the outcome set along with the probability that X takes on the value. We will denote this number by $P(X = x)$ for each x (in words, "the probability that $X = x$"). Note that we will always have:

$$0 \leq P(X = x) \leq 1 \quad \text{for any } x$$

with the function strictly positive for values in the outcome set, and

$$\sum_x P(X = x) \;=\; 1$$

where the sum is over the outcome set.

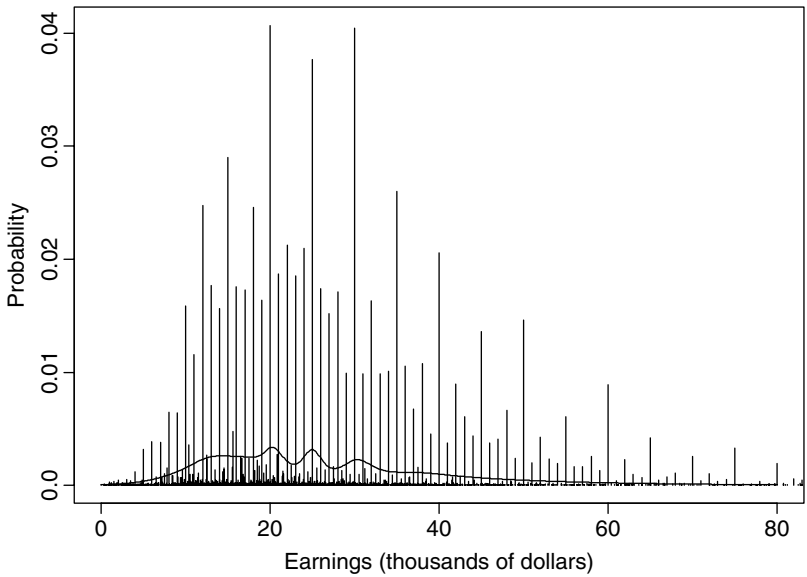

Fig. 2.1. The probability mass function for the distribution of women's earnings in 1998.

Figure 2.1 is a graph of the probability mass function for women's annual earnings, based on 1998 March Current Population Survey (CPS). The earnings scale is on the horizontal axis and the probabilities for each earnings value are represented by the vertical lines. As there is one line per earnings value the graph is very busy. A "smoothed" version of the probability mass function is plotted by the continuous line. From this display, we can see several features of the earnings distribution. It is quite unsymmetric in shape, with a long right-hand tail. It is also not a smooth function of

the earnings value. People tend to report earnings in round numbers (e.g., to the closest hundred or thousand dollar). This leads to a "heaping" of probability mass at these values.

In many situations it may be desirable to approximate the probability mass function of X by using a mathematically tractable or conceptually simpler form. For example, in the above graph we have placed a smooth curve through $P(X = x)$ and could use it to describe the distribution of earnings. Such approximations allow us to summarize the main features of the distribution using a continuous function even when the underlying probability mass function is discrete. Other examples are *histograms* and the *normal probability curve*. The latter is a parametric approximation that leads to great parsimony if it is accurate.

Probability Density Function

In some contexts it is necessary to consider infinitely many outcomes and probability mass functions become inappropriate. While we can assign probabilities to the individual outcomes for discrete random variables using relative frequencies, we need to consider outcome sets that consist of a continuum of possible values. For this we employ the continuous analog of the probability mass function – a *probability density function* (PDF) – to describe the distribution of probability over the outcome set. The PDF is a function $f(x)$ where x is in the outcome set, such that:

$$f(x) \ \geq \ 0 \quad \text{for all } x$$

$$\int_{-\infty}^{\infty} f(x)dx \ = \ 1.$$

The PDF enables probabilities to be calculated using the relationship:

$$P(a \leq X \leq b) \ = \ \int_a^b f(x)dx \qquad a \leq b.$$

Thus $f(x)$ serves the same role as the probability mass function. The smooth curve on Figure 2.1 is an example of a PDF. We do not have to assume that $f(x)$ is a continuous function of x, but we do need to assume that it is smooth enough for the above probabilities to exist. This property is called *absolute continuity* of the distribution (Kelly, 1994). Note that the probability assigned to any specific value is zero – we can only assign positive probabilities to sets of values that contain intervals.

Two continuous distributions are worth noting here, as they will play important roles in the rest of this book. The first is the *uniform distribution* on the outcome space the interval $[0, 1]$, and is defined by the PDF:

$$f(x) \ = \ \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases} .$$

For this distribution the probability that a randomly chosen value from the outcome space falls in the interval $[a, b]$, $0 \leq a \leq b \leq 1$ is just $b - a$. That is, no part of the interval is more likely to contain the value than any other part of the interval – hence the name. The second is the *standard normal distribution*, which has outcome space the set of all real numbers on the interval $(-\infty, \infty)$, and is defined by the PDF:

$$f(x) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} \quad -\infty < x < \infty.$$

The graph of this PDF is often called the "bell curve," and is the most common distributional approximation used in statistical methods.

Cumulative Distribution Function

A distribution, whether continuous or discrete, can also be characterized by its cumulative distribution function (CDF):

$$F(x) \;=\; P(X \leq x) \quad \text{for each } x \text{ in the outcome space.}$$

That is $F(x)$ gives the probability that a randomly chosen value is less than or equal to x. If X is discrete we have

$$F(x) \;=\; \sum_{y \leq x} P(X = y) \quad \text{for each } x \text{ in the outcome space,}$$

and if X is continuous

$$F(x) \;=\; \int_{-\infty}^{x} f(y)dy \quad \text{for each } x \text{ in the outcome space}$$

These relationships can be inverted to express the PDF in terms of the CDF. In the discrete case, this is

$$P(X = x) = F(x) - F(x-),$$

where x is in the outcome space and $x-$ is the largest value in the outcome space smaller than x. In the continuous case, the relationship is:

$$f(x) \equiv \frac{d}{dx} F(x) \equiv \lim_{h \to 0} \frac{F(x+h) - F(x)}{h}. \tag{2.1}$$

Thus $F(x)$ can be derived from either the probability mass function or the PDF. Note, also, that we can determine the probability mass function or the PDF from $F(x)$, so that we can characterize the distribution by either representation. Indeed, if $f(x)$ is continuous at x, $f(x)$ is the derivative of $F(x)$.

Quantile Function

A useful quantity related to the CDF is the *inverse cumulative distribution function*, also called the *quantile function*. This function will play a critical role in the development of the relative distribution. It is defined to be:

$$Q(p) = F^{-1}(p) = \inf_x\{x \mid F(x) \geq p\}.$$

The quantile $Q(p)$ can be thought of as the value of y below which a proportion p of the values fall. The reason for the infimum is that there could be many values y for which $F(y) = p$ if $F(y)$ is constant in the neighborhood to the right of $Q(p)$. If $F(x)$ is continuous, $Q(p) = \inf_x\{x \mid F(x) = p\}$ and $F(Q(p)) = p$ for $0 \leq p \leq 1$. Thus if the distribution is continuous and the CDF is strictly increasing when it is not zero or one, a quantile represents the exact value below which a proportion p of the values fall. One can also say that this value defines the pth quantile of the population (or equivalently, of the probability distribution of X). Special cases are the *median* ($p = 0.5$) and the lower and upper *quartiles* ($p = 0.25, p = 0.75$, respectively). If the distribution is discrete, then the definition of a quantile may be ambiguous, so the smaller value is chosen by convention. This choice ensures that the quantile function is left continuous. Two common ways to express the quantile function are through *deciles* (i.e., the quantiles corresponding to 0.0, 0.1, ..., 0.9, 1.0) and *percentiles* (i.e., the quantiles corresponding to 0.00, 0.01, ..., 0.99, 1.00). For example, the bottom decile is the quantile corresponding to $p = 0.10$. In the earnings distribution from Figure 2.1, the bottom decile is $Q(0.1) = \$11,500$. The median and upper quartiles are $Q(0.5) = \$24,000$ and $Q(0.75) = \$34,000$, respectively.

Often we will need to determine the probability distribution of some function of X. For example, if we know the distribution of earnings, we can determine the distribution of log-earnings. In general if the random variable Y is defined to be some function h, say, of X (i.e., $Y = h(X)$) then the CDF of Y is $F_Y(y) = P(Y \leq y) = P(h(X) \leq y)$. The outcome space of Y is the outcome space of X transformed by h. We usually can reexpress the last form in terms of the CDF of X. We call $h(x)$ a *monotone function* of x, if either $h(x) < h(y)$ whenever $x < y$ or $h(x) > h(y)$ whenever $x < y$. If $h(x)$ is a monotone function of x, we can always find $h^{-1}(x)$, the inverse of $h(x)$. If $u = h(x)$, the value of $h^{-1}(x)$ is just u. In this case

$$F_Y(y) = P(X \leq h^{-1}(y)) = F(h^{-1}(y)).$$

The *uniform distribution* plays a role for distributions similar to the role played by unity for arithmetic. Suppose we have a continuous probability distribution for X and the corresponding CDF is strictly increasing when it is not zero or one (i.e., the density does not have intervals where it is zero). Consider transforming X by the function $F(x)$, leading to the new random variable $Z = F(X)$. One can think of $F(x)$ as giving the percentile

in the distribution of x. Hence $F(X)$ is the percentile of a value randomly selected from the distribution. Intuitively, Z has a uniform distribution on the outcome space $[0, 1]$. For example, suppose F represents the distribution of grades from an exam in a large class. Then X represents the grade of a randomly chosen person in the class, and $Z = F(X)$ represents the percentile in the class that the person appeared. As the person is equally likely to be any class member, the percentile is equally likely to be any value from 0% to 100%. It is in this sense that the percentile of the person is uniform, even though the actual grade is not. Furthermore, let U be a random variable with a uniform distribution. Then transforming U by the quantile function $Q(x)$ leads to the new random variable $Q(U)$, which has the same probability distribution as X. We can think of U as a percentile chosen equally likely to be any value from 0% to 100%. Each percentile can also be associated with a person in the class, so randomly choosing a percentile is the same as randomly choosing a class member. Then $Q(U)$ gives the exam grade corresponding to the percentile, and hence the randomly chosen class member. We will use these properties in the next sections.

Numerical Summary Measures

Throughout this book we will summarize properties of population distributions through numerical measures. The overall level of a population is often summarized by the mean, or average value. The value of the mean can be expressed as the sum of each value in the outcome set weighted by the relative frequency distribution. For a discrete random variable X, the corresponding concept is that of an *expectation* or *expected value*. This can be formally defined as the weighted sum:

$$\mathrm{E}[X] = \sum_x xP(X = x)$$

where the sum is over the outcome set. We can also think about expectations of functions of random variables. Let $h(x)$ be a real-valued function for x in the outcome space. Then

$$\mathrm{E}[h(X)] = \sum_x h(x)P(X = x)$$

For example, consider $h(x) = |x|$ so $\mathrm{E}[|X|]$ is the mean absolute value taken by X.

Other summary measures for probability distributions can be defined in correspondence with their population counterparts. For example, the spread of a distribution is often summarized by its *variance*, defined as

$$\mathrm{Var}[X] = \mathrm{E}\left[|X - \mathrm{E}[X]|^2\right] = \sum_x |x - \mathrm{E}[x]|^2 P(X = x)$$

where the sum is over the outcome set.

For continuous random variables the definitions of expectation and variance can be based on their probability density functions. In particular,

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f(x) dx,$$

$$\mathrm{E}[h(X)] = \int_{-\infty}^{\infty} h(x) f(x) dx,$$

and

$$\mathrm{Var}[X] = \mathrm{E}\big[|X - \mathrm{E}[X]|^2\big] = \int_{-\infty}^{\infty} |x - \mathrm{E}[x]|^2 f(x) dx.$$

We shall return to these ideas in Chapter 5.

2.2 The relative distribution

Let Y_0 be a random variable representing a measurement for a population (e.g., hourly wages). We will call the population that generated Y_0 the *reference* population. Denote the CDF of Y_0 by $F_0(y)$ and the density by $f_0(y)$ (when the latter is defined). We do not place restrictions on the outcome space of the reference measurement, although in many applications it will only take on non-negative values.

Suppose we also observe another measurement of Y from a different population. We will call the population that generated Y the *comparison* population. It is assumed that Y has CDF $F(y)$ and density $f(y)$ (when the latter is defined). Typically Y is the measurement for a separate group or the same group during a later time period. The objective is to study the differences between the comparison distribution and the reference distribution.

Unless explicitly mentioned, we will assume that both F and F_0 are absolutely continuous with continuous densities and common support. The case where the distributions are discrete is treated in Chapter 11.

The *relative distribution* of Y to Y_0 is defined as the distribution of the random variable:

$$R = F_0(Y). \tag{2.2}$$

R is obtained from Y by transforming it by the CDF for Y_0, F_0. This has also been called the *grade transformation* (Cwik and Mielniczuk 1989). While this transformation is not widely used or understood in the social sciences, it is a very useful one, because R measures the relative rank of Y compared to Y_0. It is continuous on the outcome space $[0, 1]$, and we will call a realization of R, r, the *relative data*. We will sometimes use the abbreviation RD for relative distribution.

As a random variable, R has both a CDF and a PDF. Using the method described at the end of the previous section, we can reexpress the CDF of R as

$$G(r) \;=\; F(F_0^{-1}(r)) \;=\; F(Q_0(r)) \qquad 0 \le r \le 1,$$

where $Q_0(r)$ is the quantile function of F_0, and r represents the proportion of values.

The PDF of R (which we will call the *relative density*) can be obtained as the derivative of $G(r)$:

$$g(r) = \frac{f\big(Q_0(r)\big)}{f_0\big(Q_0(r)\big)} \qquad 0 \le r \le 1. \tag{2.3}$$

The relative density can be interpreted as a density ratio. This can be seen more easily by expressing $g(r)$ explicitly in terms of the original measurement scale, y. Let the rth quantile of R be denoted by the value y_r on the original measurement scale, so the y_r corresponding to r is $Q_0(r)$. The relative PDF is then:

$$g(r) = \frac{f(y_r)}{f_0(y_r)} \qquad y_r = Q_0(r) \ge 0.$$

Note, however, that while the relative density can be interpreted as a density ratio, it is a proper PDF in the sense that it integrates to 1 over the unit interval. A density ratio over the original measurement scale would not, in general, have this property. The rescaling imposed by the quantile function – the argument to the functions in the numerator and denominator of equation 2.3 – is what ensures that the relative density will integrate to 1. Because PDFs are one of the basic building blocks of statistical theory, the fact that the relative density is a proper PDF provides the relative distribution with a firm basis for estimation, inference, and interpretation, and a general framework for methodological development.

To understand the different components that together define the relative distribution, consider the PDFs of hypothetical reference and comparison groups shown in the top panel of Figure 2.2. The reference group distribution is approximately normal, while the comparison group distribution has a lower median and is right-skewed. The vertical and horizontal reference lines on the plot identify the components of the relative distribution. A solid vertical line is drawn at the quantile corresponding to $r = 0.4$, the value of y at the 40th percentile of Y_0. Here $y(r) = Q_0(r) = 6.37$. The density of observations at this value is given by the intersection of this line and the PDF for each group. This is shown by the two horizontal lines: $f_0\big(Q_0(r)\big)$ and $f\big(Q_0(r)\big)$ for the reference and comparison group respectively. Note that $f\big(Q_0(r)\big)$ is about half of $f_0\big(Q_0(r)\big)$. The relative density is defined as the ratio of these two quantities (see equation 2.3) for every value r in $[0, 1]$, and this density is plotted in the bottom panel of Figure 2.2. Note that at $r = 0.4$, the relative density is about 0.5, as the top graph suggests. For values in the lower two deciles of Y_0 ($r < 0.2$), the relative density is greater than 1, indicating a greater frequency of observations in the comparison distribution Y, and in the remaining deciles the value is

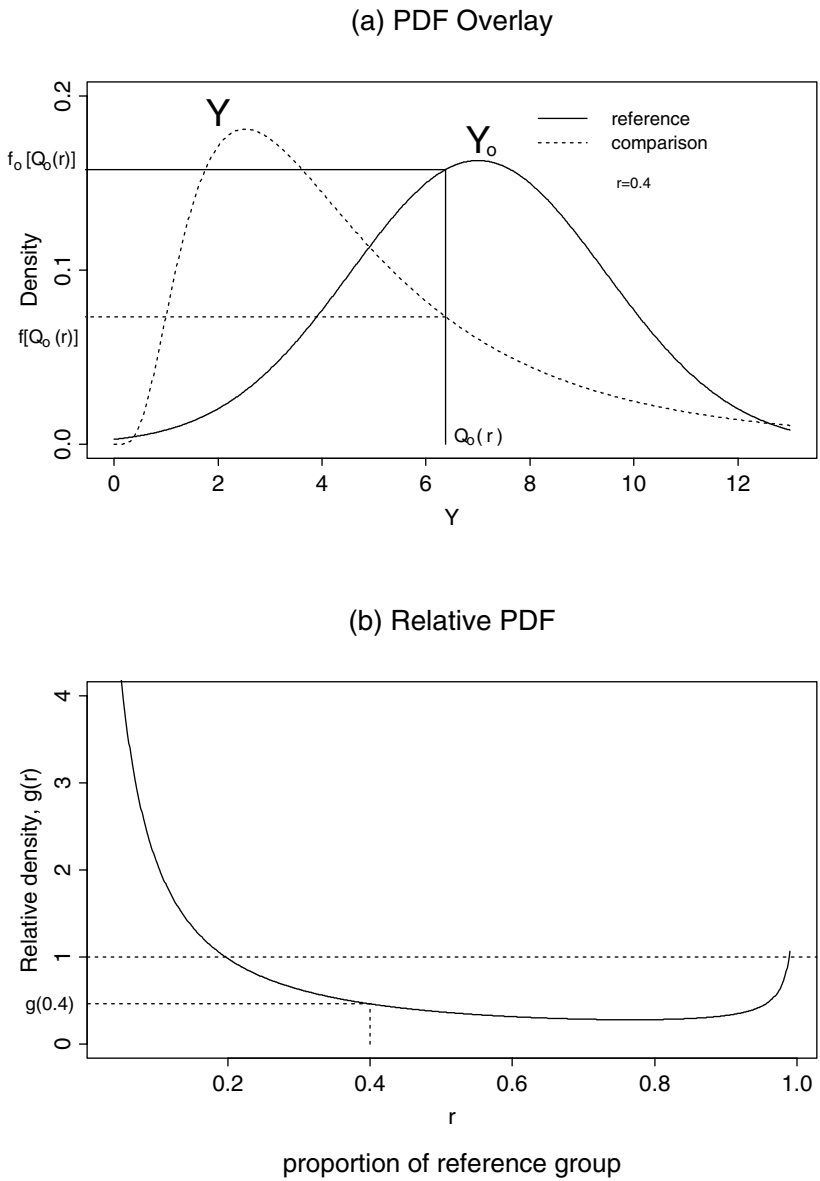

Fig. 2.2. PDFs for hypothetical reference and comparison groups (top panel) and their relative density (bottom panel).

less than 1, indicating a lower frequency of observations in Y. We present a more detailed discussion of the relative density plot elements with real data below.

The smoothness of F and F_0 ensure that $g(r)$ is continuous on $[0, 1]$. If the two distributions are identical, then the relative density is the uniform probability distribution on $[0, 1]$ and the CDF of the relative distribution is a 45^o line from $(0, 0)$ to $(1, 1)$.

The relative distribution is an intuitively appealing approach to the comparison problem because the relative data, PDF and CDF have clear, simple interpretations. The relative data can be interpreted as the percentile rank that the original comparison value would have in the reference population. The relative PDF $g(r)$ can be interpreted as a density ratio: the ratio of the fraction of respondents in the comparison group to the fraction in the reference group at a given level of the outcome attribute Y $(Q_0(r))$. The relative CDF, $G(r)$, can be interpreted as the proportion of the comparison group whose attribute lies below the rth quantile of the reference group. Note that even though the relative CDF is explicitly scaled in terms of quantiles, the implicit unit of comparison is the value of the attribute on the original measurement scale, with $y_r = Q_0(r) = F_0^{-1}(G(r))$ representing the cut-point.

For an example using real data, consider the distributions of men's and women's earnings in 1987. The PDF overlay for these distributions is shown in the top panel of Figure 2.3, and the relative density of women's to men's earnings in shown in the bottom panel. The relative density at the 20th percentile of men's wages is about equal to 2. This means women are about twice as likely as men to fall at this point of the earnings scale in 1987; or, equivalently, that the proportion of women with this level of earnings is about twice the proportion of men. The dollar value at this quantile, $Q_0(0.2)$, can be obtained from the labels on the upper axis, about $\$15,000$. The *dollar* amount is the same for both women and men $(y_r = Q_0(r) = F^{-1}(G(r)))$. Thus, each point on the relative PDF represents a specific earnings level, and as you travel along the relative PDF curve, you can read off the x and y axes the proportion of men and *relative* proportion of women who earned that level of income.

The relative density simplifies comparison in several ways. In contrast to the direct PDF overlay in Figure 2.3, which requires the viewer to construct the differences between the two curves at each point on the scale, the relative density codes this comparison directly in terms of a ratio. It provides a simple visual (and numerical) signal for information that exists but is not easy to process in the original PDF overlay (Chambers, *et al* 1983; Cleveland and McGill 1984).

The relative CDF for these two distributions is shown in Figure 2.4. At the median of the male earnings distribution, $r = 0.5$, $G(r) = 0.83$. This means that approximately 83% of women earn less than the median male. The upper and right axes are labeled in thousands of dollars, representing

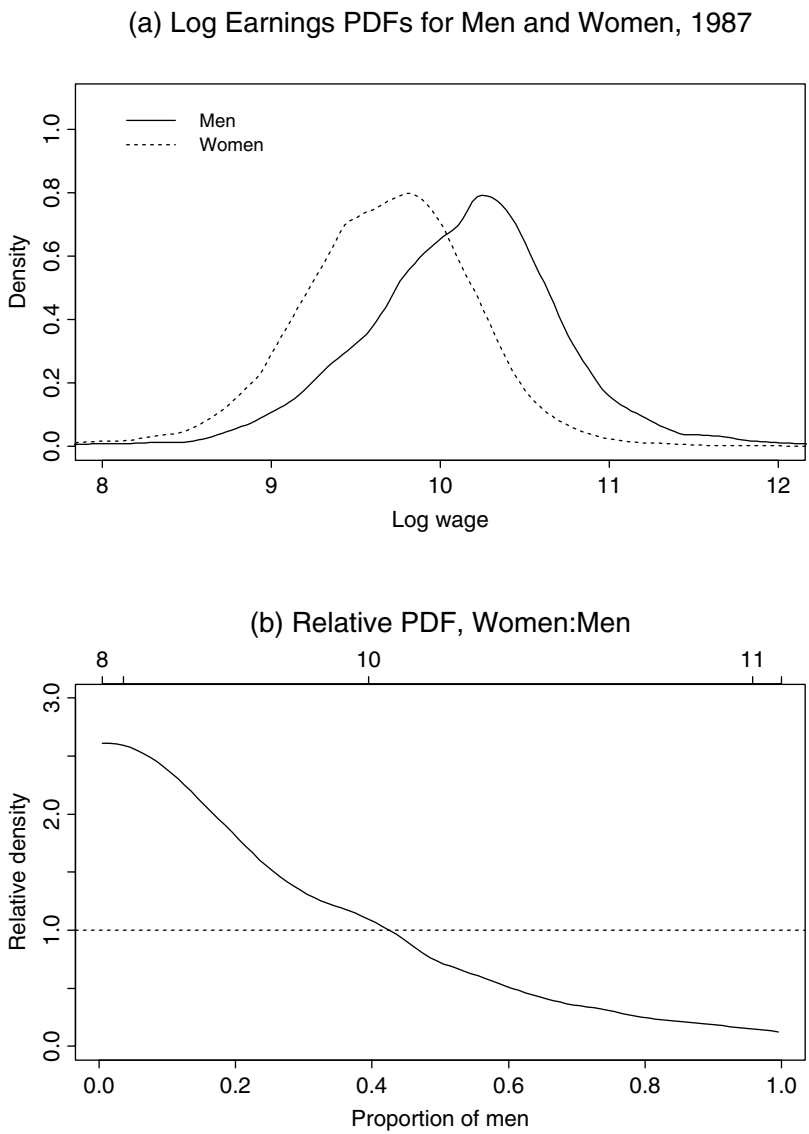

Fig. 2.3. The distribution of women's to men's earnings in 1987.

the quantiles for men ($Q_0(r)$) and women ($Q(r)$) respectively. Again the *dollar* amount along the CDF of R is constant in both distributions (you can see that explicitly here because both the upper and right axes are labeled). Each point on the relative CDF represents a specific earnings level, and as you travel along the relative CDF curve, you can read off the x and y axes the proportion of men and women who earned at or below that level of income. Note that the rescaling imposed by the quantile function is evident here, especially in the women's distribution. The distance between dollar values on the right hand scale is measured in units of persons rather than dollars. The distance between \$15 and \$20 is much larger than the difference between \$25 and \$30, because there are more people in the former range than the latter.

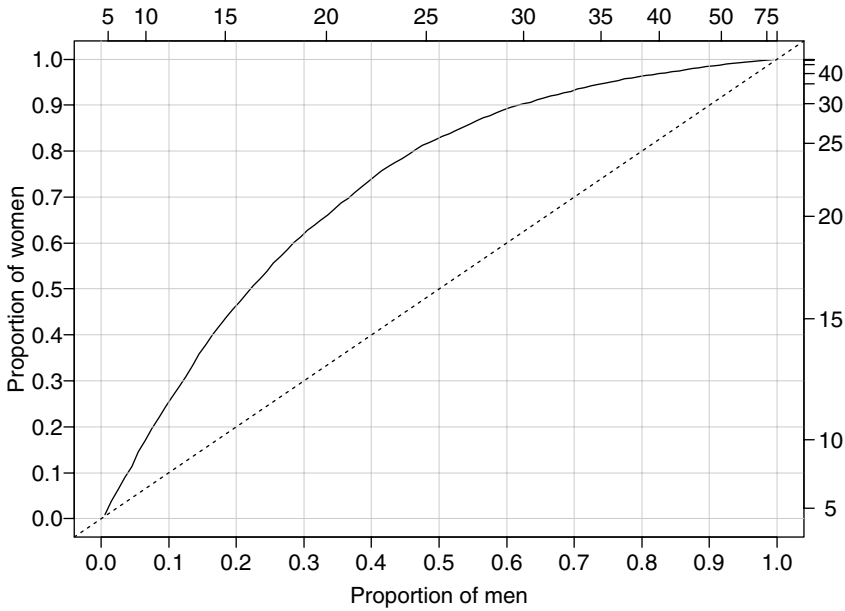

Fig. 2.4. The relative CDF of women's to men's earnings in 1987.

In general, the relative distribution is invariant to the scale of the distributions (up to a monotone transformation). For example, one obtains the same relative distribution from a comparison of log-attributes as from the comparison of the attributes. In our application, for example, we would obtain the same relative distribution from the ratio of earnings as we do from the difference in log-earnings. We discuss the scale invariance in more detail in Section 1.2.

Designating which distribution will serve as the reference distribution is a decision that must be made by the analyst, but it is often straightforward in application settings. Natural choices are suggested by time ordering, a well – understood standard reference group (like men in the example above), or an experimental control group. If the designation is reversed, the relative PDF and CDF will be symmetric around the distributional equivalence axis ($g(r) = 1$ for the PDF, and the 45^o line for the CDF) net of the rescaling. This changes the displays in predictable ways, and the substantive findings will be equivalent.

The relative density graph remains close to the original data, allowing the researcher to identify detailed differences between two distributions. The result is a more accessible, intuitively meaningful, and informative description of the data than that afforded by standard summary statistics or PDF overlays. When the graphical displays above are linked to the decomposition techniques developed in subsequent chapters, relative distribution methods become a powerful tool for analysis.

2.3 Using a known reference distribution

In many contexts the relative distribution can be formed using a theoretical or known distribution as the reference distribution. The typical context is when the reference distribution is derived from social or statistical theory and empirical observations are hypothesized to conform to this theory. In this case we usually have a random sample from some population to use for comparison, and the primary question is "goodness-of-fit": How well do the population observations conform to the theory?

This approach can serve as a useful diagnostic tool in any setting where statistical assumptions are made for univariate distributions. The classic example is residual diagnostics in the regression setting, where distributional assumptions are key to statistical inference. If the assumptions of linearity, constant variance, independence, and normality are satisfied then the standardized residuals should be approximately an independent sample from a normal distribution with mean zero and variance one. Regression diagnostic plots are often used to assess departures from these assumptions. To check the normality assumption, we compare the calculated standardized residuals to a standard normal reference distribution. This graphical test is traditionally done using a "normal scores" (Q-Q) plot, a P-P plot or a histogram overlaid with a normal curve. These plots are described in more detail in Section 2.4.2. While the P-P plot contains the same statistical information as the relative density, it does not provide an easy graphical read, largely because the information is coded into the display in a way that is difficult to extract. A plot of the relative density, by contrast, will code the relevant information into deviations from a straight horizontal

line that represents distributional equivalence. In this way, deviations from normality may be easily observed and interpreted.

Figure 2.5 gives an example of residual plots for a linear regression model designed to verify the principle of purchasing power parity (see the Background material to this chapter). Panel (a) is the standard normal scores (Q-Q) plot. It suggests some non-normality as the curve is bent in the middle and not globally linear, but the nature of the deviations from the normal distribution are not easily read from this display. Panel (b) is the relative distribution of the standardized residuals to the standard normal reference distribution. Here too one can immediately observe the non-normality of the residuals, but in this case the nature and magnitude of the deviations from the normal distribution are clearly visible. There are too few residuals in the lower tail of the distribution and about twice as many as expected in the second decile. At the median we have only half as many residuals as we would expect from a normal distribution. The upper quartile of the data is somewhat more consistent with the normal, but there is some evidence of a denser top tail. The patterns of polarization in the tails suggests nonhomogeneity of variance. This is confirmed by a plot of the standardized residuals versus the fitted values, which shows larger variability at larger fitted values.

The reference distribution can also be derived from social theory. Dagum (1977), for example, develops a model for personal income based on a theoretical specification of a differential equation to represent the regularity and permanence of income. The model posits that income follows a distribution with CDF depending on four parameters (cf., Singh and Maddala 1976). Many other types of CDFs have been proposed because they provide close fits to empirical income distributions. Examples include the Pareto distribution (Pareto 1897), the gamma distribution (Salem and Mount 1974), the beta distribution (McDonald 1984; Slottje 1984; Slottje 1987), and the income share elasticity models of Esteban (1986) and Majumder and Chakravarty (1990). These distributions are often chosen more for descriptive than theoretical reasons, motivated by a balance of parsimony, generality, and ease of use. For both the theoretical and descriptive models, the CDF of the income distribution can typically be expressed in a functional form that depends on a small number of parameters. For example, Majumder and Chakravarty (1990) propose:

$$F_0(y) = B_w\Big(\frac{1}{d} - \frac{a}{b}, \frac{a}{b}\Big),$$

with

$$w = \frac{(cy)^b}{(cy)^b + d},$$

where $B_w(\cdot)$ denotes the incomplete beta function (Abramowitz and Stegun 1965). The parameters a, b, c, and d are each interpretable in terms of in-

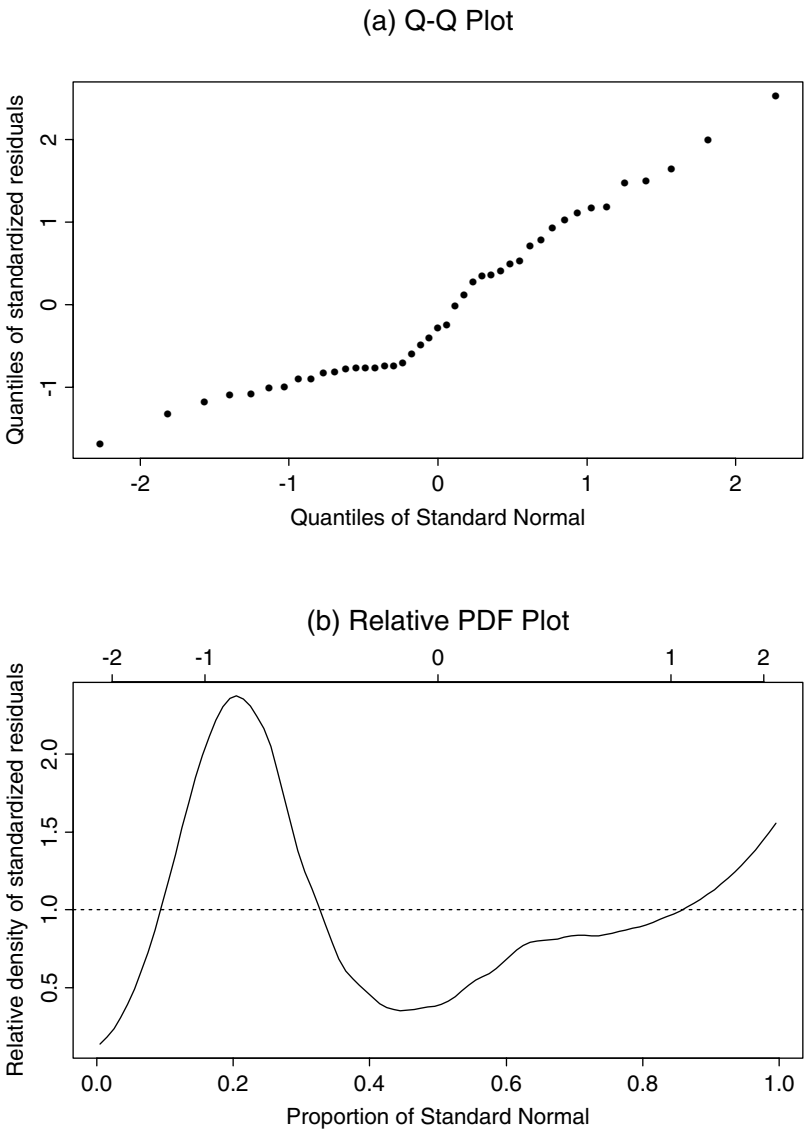

Fig. 2.5. The relative distribution of standardized residuals from the Purchasing Power Parity model.

come characteristics. This model includes many of the previously referenced models as special cases.

To complete the specification, the parameter values can be determined by theoretical considerations or estimated to produce a close fit to an observed income distribution. The reference distributions determined in this way are combinations of theoretically motivated and empirically representative forms. The relative distribution of the observed data to this reference distribution can be used to diagnose how well the data fit the model, and where in the distribution the discrepancies occur.

There are many other settings in which relative distributions could also be used as a diagnostic tool. One example is in survival analysis. Nonparametric (e.g., Kaplan-Meier) estimates of survival curves can be compared to a reference exponential distribution. The relative density can be used to identify exactly where the survival function departs from the exponential process. Relative distributions can also be used to check the assumption of proportional hazards in survival modeling by comparing the approximate nonparametric estimates. Another area in which these methods could be applied is test statistics. Relative distributions can be used to compare a bootstrap distribution to its theoretically derived asymptotic approximations. In the Bayesian framework, the relative distribution can be used to compare posterior to prior distributions, as a distributional analog to Bayes factors. We return to these issues in Chapter 9, in the context of estimation and inference. As distributional assumptions and comparisons lie at the heart of many statistical methods, relative distributions have a wide range of potential applications.

2.4 History and literature

2.4.1 Statistical origins

The ideas underlying the relative distribution framework have been recognized in statistics for decades, but explicit study of the relative data, PDF, and CDF has been uncommon. There are at least three directly relevant threads in the statistical literature. They are motivated by separate substantive research questions and rarely reference each other.

Parzen (1977; 1992) appears to be the first to systematically study aspects of the relative PDF as a basis for interdistributional comparison. He prefers to construct the reference distribution by pooling all groups. If λ is the proportion of the comparison group in the pooled reference group, then the CDF of the pooled reference group is

$$H(y) = \lambda F(y) + (1 - \lambda)F_0(y).$$

Parzen focuses on the relative distribution of F to H as part of *comparison change analysis* and refers to the corresponding relative CDF and PDF as

the *(pooled) comparison distribution* and *(pooled) comparison density*, respectively. He refers to the relative distribution as the *unpooled comparison distribution*.

Almost all the work on the pooled comparison distribution is relevant to the relative distribution, and vice versa. Let GP denote the CDF of the relative distribution of the comparison group to the pooled reference group. The interrelationship between the pooled and unpooled relative distributions can be expressed through the relation:

$$\mathrm{GP}^{-1}(p) = \lambda p + (1 - \lambda)G^{-1}(p).$$

Thus there is a 1:1 correspondence between the two relative distributions, and relationships for one can be expressed in terms of the other. An exception is some of the material on inference, which relies on the independence of the comparison and reference samples. We consider inference using a pooled reference sample in Section 9.3.

Parzen's students discuss the role of the comparison distribution (Prihoda 1981) and develop kernel density estimation for the comparison density (Alexander 1989). Alexander is very comprehensive and gives a broad review of the literature at that time. Eubank and LaRiccia (1987; 1995) propose a framework for developing summary measures for comparison and hypotheses testing based on the comparison density. These summary measures rely on generic features of Hermite and Legendre polynomials to test for location and scale effects on the oscillation patterns in the relative density. We discuss these at length in Chapter 10.

Throughout this book we use an unpooled reference group to form the relative distribution, but we note that a pooled reference group can be used in almost all cases. The issue is analogous to the choice of reference category in ANOVA and regression. Using a pooled reference distribution is equivalent to using the mean as the reference category in ANOVA, while using a specific comparison group is equivalent to indicator (or "dummy variable") specification in regression. We use a distinct reference group, because we believe this usually leads to more interpretable quantities. For example, consider the plots in Figures 2.3 and 2.4. The direct comparison of women to men makes these plots easy to interpret because it presents an unambiguous contrast. We could have, instead, compared women to the pooled population of men and women. This would have been more difficult to interpret, however, because the comparison is then confounded by the sex composition of the population.

Yet there are situations in which the use of the pooled reference might be preferable. In some contexts, the comparison of a specific group to a total population may be of direct interest. Another case is where an individual group was too small to provide adequate information for estimating distributional information. One solution is then to compare the complement to the total, to identify the contribution of the smaller group. A final case is when the comparison distribution is so different that the reference

distribution is effectively disjoint. In that case, the support of the comparison distribution may not be contained in that of the reference distribution. This would never occur if the reference distribution is made from the pooled samples. However, note that if the support of the reference distribution is not a subset of the support of the comparison distribution, then using the pooled reference distribution will probably not represent what we are trying to capture. For example, suppose there were no men with incomes below $500, but among women, 20% of the incomes fell in that range. If the sex ratio in the population was 1:1, then the relative density for earnings below $500 based on the pooled reference would be about 2 (20% of women vs. 10% of the total population). One might be tempted to interpret this as meaning that women were twice as likely as members of the general population to have earnings below $500. While technically true, this completely obscures the fact that there is not one man in this income range. In this case, the relative density may be more effectively estimated using the pooled reference, but the tradeoff is that we lose key aspects of the comparison.

In separate literature, Cwik and Mielniczuk (1989; 1993) have investigated nonparametric density estimation for the relative PDF. They refer to equation (2.2) as the *grade transformation*, because F_0 can be thought of as a grading function. In their terminology, the relative PDF is a *grade density*. They develop a kernel estimate for the relative PDF that has uniform almost sure convergence, and a method for choosing an estimate which is appropriately smooth. Gijbels and Mielniczuk (1995) generalize these results (to the Radon-Nikodym derivative) and determine the rates of uniform almost sure convergence.

In other literature, Li, *et al* (1996) develop the statistical properties of the relative CDF under the name of *two-sample vertical quantile comparison function.*

2.4.2 Relationship to probability plots

The relative CDF $G(r)$ is implicitly a theoretical *probability-probability* (P-P) plot of F against F_0, an empirical version of which was considered by Wilk and Gnanadesikan (1968). It is the plot $\{(F(x), F_0(x)) : x \in \mathbb{R}\}$ which can be represented in the functional form $\{(r, G(r)) : 0 \leq r \leq 1\}$ (cf., Chambers, *et al* 1983). Another allied probability plot is the *quantile-quantile* (Q-Q) plot: $\{(Q(r), Q_0(r)) : 0 \leq r \leq 1\}$. This represents a comparison of the quantiles and so is intrinsically measurement scale dependent. In particular the Q-Q plot will be a straight line when the comparison and reference distributions differ only in location or scale. See Wilk and Gnanadesikan (1968) for a description of the strength and uses of probability plots. The Q-Q plot can be motivated by a shift-function representation of the distributions (Doksum 1974):

$$F_0(y) = F(y + \Delta(y))$$

where

$$\Delta(y) = F^{-1}(F_0(y)) - y.$$

Here $Y_0 + \Delta(Y_0)$ has the same distribution as Y so that $\Delta(y)$ represents the shift needed to bring Y_0 up to Y. The shift-function $\Delta(y)$ have also been studied by Doksum (1976), and Switzer (1976) as measures of treatment effects. Despite its great value as a tool for comparing distributions, the Q-Q plot has deficiencies for use in social science applications because it is not scale invariant. While both Q-Q plots and the relative CDF form a complete summary of the information necessary for comparisons, the relative distributions are comparable across different time points and under different economic conditions, while the Q-Q plots are not. The primary reason for this is that the relative distribution is invariant to monotone transformations of the measurement scale, while the Q-Q plots are designed to reflect the measurement scale.

Holmgren (1995) gives a nice discussion of the merits of the relative CDF (P-P plots) compared to Q-Q plots for comparing the results of independent studies, each comparing a treatment to a control group. He also shows that under appropriate technical conditions the relative distribution is the maximal invariant – loosely speaking, that it summarizes all the information in the comparison between the two distributions that is independent of the measurement scale.

2.4.3 Relationship to Lorenz curves

One of the primary application areas for relative distribution methods is the study of inequality. Lorenz curves (Lorenz 1905) and the associated Gini index summary statistic are the standard method used for inequality comparisons, so it is natural to examine their relationship to the relative distribution. The Lorenz curve can be shown to be a particular kind of relative distribution, but relative distributions are more general in three key ways.

Lorenz curves are effectively CDFs: the cumulative fraction of Y that is held by the $p\%$ of the population with lowest values of the attribute. When everyone in the population has the same value of Y, the Lorenz curve is a $45°$ line joining $(0, 0)$ to $(1, 1)$. All other distributions curve below this.

To understand the link between Lorenz curves and relative distributions, it is necessary to understand what the Lorenz PDF represents. The PDF of a Lorenz curve is the rescaled density ratio of two distributions. The denominator is the income distribution: the fraction of earners at each level Y. The numerator can be called the "dollar" distribution. Just as each person is associated with an income level, each dollar in the economy is associated with the income level that produced it. The dollar distribution represents the likelihood that a given dollar came from a person with that income level, or equivalently, the fraction of the total dollars that are allocated to each income level (leaving aside point mass issues). The Lorenz

PDF is the rescaled density ratio of dollars to people at each level Y of income, and the Lorenz curve is therefore the CDF of a particular form of relative distribution. Using the notation from above, R_L is the grade transformation of the dollar distribution by the income CDF (we will call it the *Lorenz grade transformation*), and $G_L(r)$ is the corresponding relative CDF, the Lorenz curve.

The fundamental difference between the two approaches is that a Lorenz curve is defined by comparing two attributes within a single population (e.g., dollars and people), while relative distribution techniques compare attributes across two populations. To compare two groups using Lorenz curves, the Lorenz curve for each is constructed, graphically overlaid, and examined. The relative distribution, by contrast, compares one distribution directly to the other, using a single curve to encode the differences.

Another key difference is the unit of measurement. Lorenz curves map cumulative shares of Y to cumulative shares of population. Absolute equality is represented by a 45° line, and inequality is measured by the deviation of the Lorenz curve from this line. For example, the Gini index is defined to be twice the area between $G_L(r)$ and the 45^o line:

$$\text{Gini}(F) = 2\text{E}\big(R_L\big) - 1.$$

a rescaled mean of the Lorenz grade transformation. If the Lorenz curves for different groups do not cross, the groups can be ordered in terms of inequality by the curves or Gini indices, otherwise ordering is ambiguous. Relative distributions, by contrast, map population quantiles to population quantiles, for each *level* of Y. Distributional *equivalence*, rather than absolute equality, is represented for the relative CDF by the 45° line. Comparative (rather than absolute) inequality is represented by summary measures based on the relative PDF or CDF (see Chapter 5). While inequality is a natural application area for relative distribution methods, they are applicable to a much wider range of topics.

A final key difference concerns the level of scale invariance. Lorenz curves (and the summary measures based on them) are multiplicatively scale invariant, i.e., two distributions will have the same Lorenz curves if, and only if, they differ by a simple multiplicative constant. The relative distribution, by contrast, is invariant to *all monotonic* transformations of the original measurement scale. Relative distributions of the raw attribute, the log-attribute, or any other monotonic transformation of the attribute are equivalent. This means that the relative distribution makes less restrictive assumptions about the underlying utility functions in the inequality context, requiring only that they be monotonic. We shall call this the principle of *strong* scale invariance. Whenever this principle holds, the relative distribution plays the primary role in comparisons, in the sense that it contains all the information necessary for comparing distributions, making the minimal assumptions necessary for valid comparison. Holmgren (1995) shows that under appropriate technical conditions the relative distribution is the

maximal invariant. This means, loosely speaking, that any other quantity that contains the same information does not satisfy the principle of strong invariance (cf.,Lehmann 1983). It does not mean that the relative distribution is inappropriate when the assumptions are not known to hold, only that other comparisons may exist that can not be exclusively expressed in terms of the relative distribution.

2.4.4 Relationship to other econometric methods

With the dramatic changes in earnings distributions over the past three decades (for a review cf., Danziger and Gottschalk (1996)), and the limitations of the traditional Lorenz-based measures, the development of alternative methods for measuring distributional differences has become somewhat of a growth industry. Some of the methods are simple descriptive measures based on quantiles, for example, the 90:10 ratio, $Q(90)/Q(10)$, and its cousins the 90:50 and 50:10 ratios. Plotted as a time series, these are quite effective at representing changes in the tails of the earnings distribution (e.g., Smeeding and Gottschalk 1996). In addition, they are convenient to calculate. More detailed versions of quantile-based plots can be found in Karoly (1993), where ratios for each decile are constructed over time (e.g., $Q_p(t)/Q_p(t-i); p = 10, 20, ..., 90; i = 1, ..., t$). We discuss quantile-based methods in in Chapter 13.

In a spirit more similar to the relative density, Picot, *et al* (1990) work with a decile-based density ratio of earnings over time. The result is a histogram-like display, where the height of the bars represents the relative fraction of the comparison group in each decile of the reference group distribution. These methods provided the initial inspiration for the development of the relative distribution framework. Other authors have used density ratios based on a breakdown of the earnings scale into upper, middle, and lower class (or more accurately, income). Like the quantile ratio plots, these density ratio histograms provide a simple and convenient visual display for tracking the broad forms of distributional difference. The drawback to these methods is the absence of a general theoretical basis. Without this basis, the methods are limited to simple descriptive tasks. They provide no framework for decomposition, or for statistical estimation and inference. We discuss decile-based versions of the relative density in Chapter 9.

In the regression setting, Juhn, *et al* (1991) develop a method for isolating the impact of distributional changes in location (mean) and scale (variance) on mean wage differences between two groups. They apply their method to investigate the race-gap in wages, and Blau and Kahn (1994) apply it to the gender-gap. Their method is derived from the classic regression decomposition that separates changes in covariates (the $X's$) from changes in the "returns" to the covariates (the regression coefficients, or βs). The method partials out a series of terms representing changes in covariate

means, estimated returns to the covariates, changes in the mean residual earnings gap between the groups, and changes in the standard deviation of the men's residual earnings variation:

$$\begin{aligned} D_t - D_0 \;=\; (\delta X_t - \delta X_0)\beta_t \;&+\; (\beta_t - \beta_0)\delta X_t \\ &+\; (\delta\theta_t - \delta\theta_0)\sigma_t \;+\; (\sigma_t - \sigma_0)\delta\theta_t. \end{aligned}$$

Here, $D_t - D_0$ is the change in the mean wage gap between two groups from year t to year 0, δX represents the group difference in covariate means (e.g., education), β represents the estimate of the regression coefficient for the reference groups (the "returns" to the covariate), $\delta\theta$ represents the group difference in the average standardized residual (effectively the intercept difference for the group-specific regression estimates), and σ represents the standard deviation in the reference group's residual wage distribution. Interested readers are referred to the papers cited above for a more detailed explication.

While this method has the benefit of continuity with previous decomposition techniques, it also suffers from some of their drawbacks. First, it works only with average differences. Even though the residual wage distribution and the reference group's standard deviation are included, both distributions are collapsed into single number summaries: the average residual wage gap, and the standard deviation in the reference group residual wage distribution. This makes it impossible to examine how changes in the two distributions affect the relative density of each group at different levels of the earnings scale. Second, this method does not separately identify and estimate the effects of the changes in distributional shape for each group. Instead, the two are summarized and combined in the third term, which reflects simply the changes in the mean residual wage gap multiplied by the reference group's standard deviation. This has the effect of confounding the two shape shifts, again removing the level of detail needed to answer the most interesting questions – such as whether women's gains were due mainly to upgrading in their own wage distribution or to downgrading of the men's earnings. Finally, there is no graphical display to provide an intuitive feel for what the estimates imply. A fully distributional approach to location, shape, and covariate decomposition is possible in the relative distribution framework, and this is developed in Chapters 5 through 8.

DiNardo, *et al* (1996) use the general approach of forming compositionally adjusted distributions in order to isolate the marginal effects of changes in the covariate distribution on changes in the distribution of earnings. They apply this method to investigate the role of the minimum wage freeze and declining union density on the growth in earnings inequality over the 1980s (see also DiNardo and Lemieux 1996). These methods are largely subsumed by the relative distribution framework, and we take up the question of covariate adjustment in Chapter 7.

A final important method for tracking distributional change that has emerged from the econometric literature is quantile regression (Buchinsky

1995). We consider this in detail in Chapter 13.

2.4.5 Relationship to receiver operating characteristics curves

The relative CDF is also an *ordinal dominance curve* (ODC) used in the evaluation of the performance of medical tests for separating two groups (Bamber 1975). It is directly related to the *receiver operating characteristic* curve (ROC) through the relationship $\mathrm{ROC}(r) = 1 - G(1 - r)$. In this context, Hsieh (1995), Li, *et al* (1996), and Hsieh and Turnbull (1996) use an empirical process approach to describe the properties of ODC and ROC curves. In the guise of ROC methods the relative CDF is used extensively in a variety of fields (Begg 1991; Campbell 1994; Swets and Pickett 1982).

As noted above, a closely related quantity to the relative PDF $g(r)$ is that of the *density ratio*: $h(x) = f(x)/f_0(x), x \in \mathbb{R}$, considered by Silverman (1978). It is a key element of discriminant analysis (Hand 1982) and likelihood-ratio methods. Note that $h(x) = g\big(F_0(x)\big)$ and $g(r) = h\big(Q_0(r)\big)$. Absava and Nadareishvili (1985) study nonparametric estimation of the density ratio.

Background material

Section 2.1

Kelly (1994) is a careful and readable introduction to the probability theory underlying the methods in this book. He goes into much greater depth than is required here and either that book or one similar should be consulted if the brief descriptions given in the text require reinforcement. Appendix A in Kelly's book provides a good review of ideas from calculus and discrete mathematics useful for a complete understanding of the technical material here and in later chapters. Rice (1995) provides an introduction to mathematical statistics that gives special attention to data analysis and graphical displays. The level is appropriate for the methodology presented here, and the book can also be used as a reference for concepts not fully covered here.

Section 2.2

Simonoff (1996) describes how to estimate the population density from a sample when at least weak prior beliefs about the density (e.g., smoothness) are held. Conceptually these methods allow much more information to be extracted from data than is possible when no formal assumptions are made about the population. We consider these issues in depth in Chapter 9.

Section 2.3

The principle of purchasing power parity states that over long periods of time exchange rate changes will tend to offset the differences in inflation rate between the two countries whose currencies comprise the exchange rate. To verify this principle, Chatterjee, Handcock, and Simonoff (1995, page 153) consider a sample of 44 countries for the years 1975–1990. The target variable represents the estimated average annual rate of change of exchange rates from 1985 to 1990. The predicting variable represents the estimated average annual rate of change of the differences in wholesale price index values for the country versus the United States. The data were originally derived and supplied by Professor Tom Pugel of New York University's Stern School of Business, based on information given in the *International Financial Statistics Yearbook*, which is published by the International Monetary Fund. The residuals are from the regression model, excluding Iran, and are given on page 161.

Computational issues

Almost all standard statistical packages contain facilities for representing probability mass functions (bar charts), probability densities (histograms, frequency polygons), and fixed bin-width histograms. At the heart of relative distribution methods is the need for nonparametric density estimation. This issue is considered in depth in Chapter 9. Increasingly, many of these packages also contain nonparametric CDF and density estimation routines. Code for the relative density and CDF for standard packages such as SAS and S-PLUS is directly available from the website for this book. Many additional references are given by Simonoff (1996).

There are many sources for software to perform ROC-type analyses. Elizabeth J. Atkinson contributed S-PLUS code to the S-news electronic mailing list, which can be found in the `S-news` directory of `statlib.` The `Hmisc` library of S-PLUS functions written by Frank E. Harrell contains code for ROC estimation, and also for summary statistics useful for distributional comparison. The library is available from `statlib.` Harrell's website contains much information and software for distributional analysis, in addition to information on using S-PLUS.

Exercises

Exercise 2.1. Describe and graphically sketch some examples from real application contexts where one might expect the principle of strong scale invariance not to hold. What do the deviations from the principle represent? Are these important substantively?

Exercise 2.2. Suppose that X is a discrete random variable giving the total number of successes from n independent experiments where the probability of success in each experiment is p. The distribution of X is referred to as the *binomial* distribution. Let $x!$ be the number of possible permutations of x distinct objects. By convention 0! is defined to be 1. Show that the probability mass function of X is

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, 1, \ldots, n.$$

Here

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

is the number of possible combinations of n objects taken x at a time (ignoring the order of selection).

Exercise 2.3. Suppose that Y is a discrete random variable giving the *proportion* of successes from n independent experiments where the probability of success in each experiment is p. Note that the support of Y is a subset of $[0, 1]$. Using the result in Exercise 2.2, derive the probability mass function of Y.

Exercise 2.4. Suppose we have $n = 5$ experiments and $p = 0.25$. For the random variable in Exercise 2.3, determine $P(Y > \frac{1}{2})$. Is it greater than $P(0.3 < Y < 0.7)$?

Exercise 2.5. Consider the random variable in Exercise 2.3 with $n = 5$ experiments and $p = 0.25$. Plot the probability mass function of X. Determine the CDF of X. Graph the CDF separately from the probability mass function.

Exercise 2.6. Calculate the expectation of the random variable in Exercise 2.3. Give a heuristic reason for the value it takes. Calculate the variance and standard deviation of the random variable.

Exercise 2.7. Consider a discrete random variable that takes the values $0, 1/n, 2/n, \ldots, n$ with equal probability. Calculate the expectation of this random variable. Give a heuristic reason for the value it takes. Calculate the variance and standard deviation of the random variable. How do these numbers compare to those for the distribution in Exercise 2.3?

Exercise 2.8. Suppose that X is a continuous random variable with probability density function

$$f(x) = \begin{cases} Cx(1-x) & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}.$$

What is the value of C? Determine $P(X > \frac{1}{2})$. Is it greater than $P(0.3 < X < 0.7)$?

Exercise 2.9. Plot the PDF of the random variable in Exercise 2.8. Determine the CDF of the random variable. Graph both the PDF and CDF on the same plot.

Exercise 2.10. Calculate the expectation of the random variable in Exercise 2.8. Give a heuristic reason for the value it takes. Calculate the variance and standard deviation of the random variable. How do these numbers compare to those for the uniform distribution on $[0, 1]$?

Exercise 2.11. Suppose that X is a discrete random variable. Let $Y = aX+b$ be a function of X where a and b are constants. Show that

$$\mathrm{E}[Y] = a\mathrm{E}[X] + b,$$

and

$$\mathrm{Var}[Y] = a^2\mathrm{Var}[X].$$

Exercise 2.12. Answer Exercise 2.11 when X is a continuous random variable.

Exercise 2.13. The conditions that F and F_0 be absolutely continuous with continuous densities are stronger than is necessary for most of the properties of the relative distribution to apply. Show that if F is only continuous then R has the uniform distribution on $[0, 1]$. In general, that is, for F not necessarily continuous, show that $G(r) \leq r, \quad 0 \leq r \leq 1$, with equality failing if and only if r is not in the closure of the range of F.

Chapter 3

Location, Scale and Shape Decomposition

Differences between distributions can be divided into two basic components: changes in location and changes in shape. If the comparative distribution is a simple location-shifted version of the reference distribution, that is, $F(x) = F(x - c)$ or $F(x) = F(x \times c)$ for some constant c, then the difference between the two distributions can be parsimoniously summarized by this shift. Differences that remain after a location adjustment are differences in "shape" – a general concept that comprises scale, skew, and other distributional characteristics. In this chapter, we develop a general approach to decomposing the overall relative distribution into component relative distributions that represent differences in location and shape.

Location and shape shifts have substantive, as well as technical, meaning. In the earnings context, for example, a pure location shift would occur if every income were multiplied by the same factor, e.g., if every earner received the same cost of living adjustment. The entire earnings distribution would then be moved up (or down) on the dollar scale, but the underlying shape of the distribution would remain constant. The median earner's percentage increase (or decrease) in this case would summarize the experience of the entire workforce. A shape shift, by contrast, would occur if earners were redistributed along the earnings scale, keeping the location constant. The "declining middle class" scenario provides one example of such a redistribution, with earners moving from the middle of the distribution into the upper and lower tails. But other scenarios are also possible, with growth occurring in the upper tail (a pattern consistent with job upgrading), the lower tail (a pattern consistent with a declining real minimum wage), or the middle of the distribution (a pattern consistent with a more egalitarian restructuring of wages). In all of these cases, the change in the median earner's income would not necessarily represent the experience of earners in other sections of the distribution.

Simple changes in location and shape are easy to identify in the relative density display. Two examples using simulated data are shown in the panels of Figure 3.1. The top two panels show the impact of a location shift, first as a PDF overlay, then as the relative density. The distribution for the comparison group is left-shifted relative to the reference group, but it

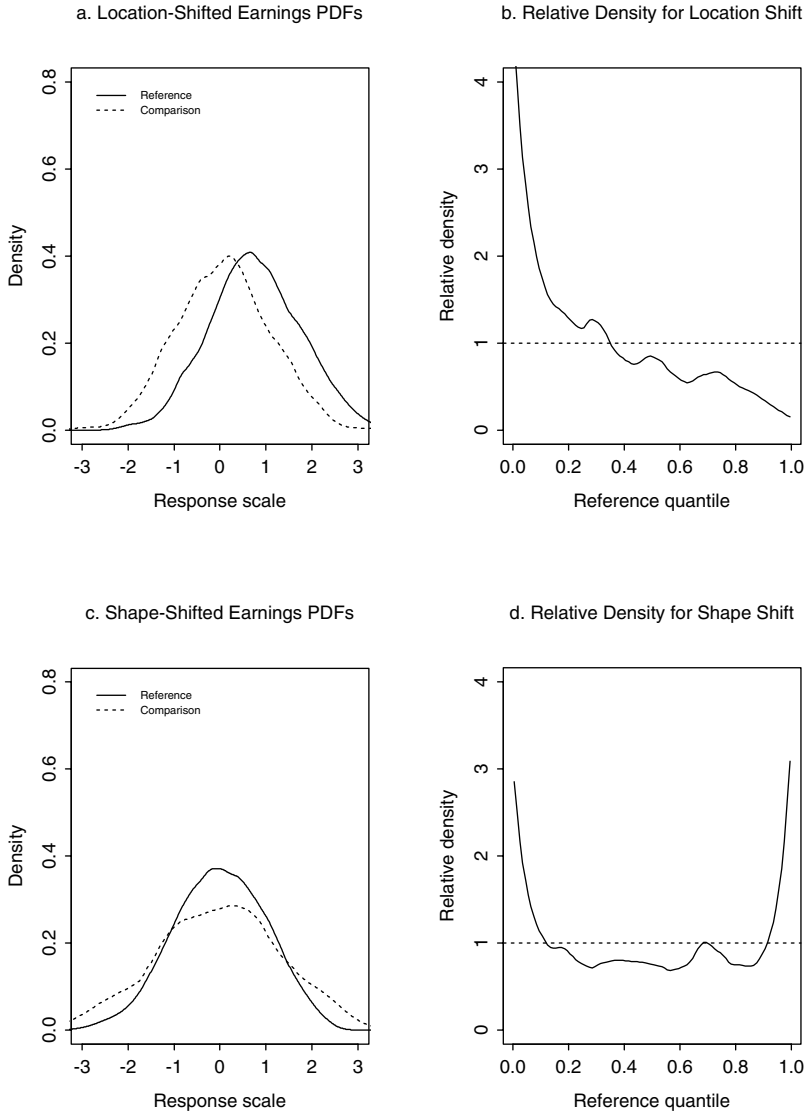

Fig. 3.1. Location and shape shifts in a hypothetical distribution. Panels a and b are the effect of a location shift on the PDF and relative PDF, respectively. Panels c and d are the effect of shape shifts on the PDF and relative PDF, respectively.

has the same shape. The relative distribution of comparison to reference observations in the right panel thus displays a simple monotonic decline. At the bottom of the response scale, there are relatively more observations in the comparison group than the reference group. The relative distribution shows this, as it is well above 1, and the value it takes can be interpreted to mean that comparison observations are about 4 times more likely than reference observations to be at the bottom of the scale. At about 0.5 on the response scale, the PDFs of the comparison and reference groups are about equal. This is where the distributions cross on the left panel, and where the relative distribution takes the value 1 on the right panel. From this point on, the 40th percentile of the reference distribution, there are more reference than comparison observations, and the relative distribution drops below 1. The value it takes near the top of the scale, about 0.4, indicates that comparison observations are about 60% $(100(1 - 0.4))$ less likely to be at the top than reference observations. Simple location shifts will always show a monotonically declining or inclining (if the comparison group is right-shifted) relative density like this.

In the bottom two panels, a scale shift is depicted. Here, the comparison group has a more "polarized" distribution than the reference group: there are more comparison observations at the top and bottom of the scale, and fewer in the middle. The relative distribution for this kind of shift takes a simple U-shape. At the top and bottom of the scale, comparison observations are about 2.5 times more likely than reference. In the middle, comparison observations are about 30% less likely than reference. Simple scale shifts will always take a parabolic shape in the relative density, U-shaped if the comparison distribution has relatively more spread than the reference, and inverted-U if the comparison distribution has relatively less spread.

In both cases, the relative density provides a simple, intuitively accessible picture of the distributional difference. Location or shape shifts operating in isolation can be quickly identified, and the impact can be observed and quantified across the whole response scale. In most applications, however, such simple shifts are unlikely to be observed. When both types of shifts are operating, or when factors other than scale are changing in the shape component, we need a way to separate out the various effects. The methods for this are developed below.

It should be noted that location and shape as defined in this chapter are atheoretic from the social science standpoint. They represent changes in the first two moments of the distribution rather than the underlying social process: the effect rather than the cause. The decomposition technique is therefore a descriptive rather than an explanatory tool. Description is an important first step in analysis, however, as it is necessary to understand what has changed before attempting to explain why. In later chapters, we will present methods for covariate adjustment that take the next step towards explanatory modeling.

3.1 Decomposing the relative distribution

Overall the relative distribution and its associated summary statistics are scale-invariant, but in general this is not true for the location, scale and other components identified through decomposition. Any form of location shift can be used for decomposition, but the results will depend on the scale adopted. This is because the concept of location shift is inherently scale dependent: a multiplicative shift changes the distribution in a different way than an additive shift. If we want to identify the effect of a location shift and separate it from other changes in the distribution, it is necessary to specify what scale this shift operates on. The choice of scale is application specific, and the analyst should choose the scale according to the nature of the data.

In the discussion below, we use an additive mean shift. This choice highlights the similarities and differences with the decomposition of variance in the linear model context. In many of the empirical chapters of this book, however, we use an additive median location shift instead. The median because population quantiles are a natural, robust and scale invariant unit of measurement, and an additive shift because the data are transformed to the log-scale. These and many other methods for location shifts can be used in the decomposition approach presented below. In each case the development is the same, with the alternative transformations replacing the additive mean shift used in the exposition.

Reversing the reference and comparison group designations will change the estimated effects of the location and shape shifts, and associated summary statistics, because these are defined in terms of the reference group scale. In general, however, this will not change the ranking or nature of the effects, and will therefore have little impact on the substantive findings. The choice is analogous to the selection of a reference category in dummy variable regression.

We start here by decomposing the relative distribution into location and shape components. The approach is easily generalized to further decompose the shape component into pieces that represent higher moments of the distribution. We consider these in the next section.

Let Y_{0L} denote a random variable describing the reference group *location-adjusted* to have the same mean as the comparison group. For an additive mean shift, we define Y_{0L} as the random variable $Y_0 + \rho$ where $\rho = \mu_Y - \mu_{Y_0}$. The CDF of Y_{0L} can be written as $F_{0L}(y) = F_0(y - \rho)$. The density corresponding to F_{0L} is $f_{0L}(y) = f_0(y - \rho)$. Y_{0L} defines a hypothetical group which has the location (here the mean) of the comparison group, but the shape of the reference group.

From these three distributions – Y_0, Y_{0L} and Y – we can construct two RDs that represent the effects of the location and shape changes. Generalizing the notation of Chapter 2, let $R \equiv R_0 = F_0(Y)$ be the relative distribution of Y to Y_0. To isolate the location shift we take the RD of

Y_{0L} to Y_0, denoted $R_0^{0L} = F_0(Y_{0L}) = F_0(Y_0 + \rho)$. R_0^{0L} will have a uniform distribution when the comparison and reference groups have the same location. To isolate the shape shift we take the RD of Y to Y_{0L}, denoted $R_{0L} = F_{0L}(Y) = F_0(Y - \rho)$. R_{0L} will have a uniform distribution when, net of location shifts, the two distributions have the same shape.

This can be represented in terms of the density ratios from (2.2):

$$\frac{f(y_r)}{f_0(y_r)} = \frac{f_{0L}(y_r)}{f_0(y_r)} \times \frac{f(y_r)}{f_{0L}(y_r)} \tag{3.1}$$

or, in more heuristic terms:

$$\begin{array}{c}\text{overall relative}\\ \text{density}\end{array} = \begin{array}{c}\text{density ratio for}\\ \text{the location difference}\end{array} \times \begin{array}{c}\text{density ratio for}\\ \text{the shape difference}\end{array} \tag{3.2}$$

The graphical display of the decomposition RD densities, which we will denote by g_0, g_0^{0L}, and g_{0L}, respectively, provides a useful visual summary of the relative size and nature of the components.

Technically, these two effects form an exact decomposition of the relative distribution of Y to Y_0 in the sense that R_{0L} is the relative distribution of R_0 to R_0^{0L}. The density ratio for the location effect is a proper density (i.e., it integrates to 1). The density ratio for the shape effect in general is not, because of the scale change imposed by using f_{0L} rather than f_0 as the reference distribution for R_{0L}. The shape density ratio in (3.1) instead preserves the cut-points, y_r, so that the location and shape effects are applied at the same value of y_r.

To make the rescaling explicit, we can express the relationship between the densities as:

$$g_0(r) = g_0^{0L}(r) \times g_{0L}(p) \qquad 0 \le r \le 1, \tag{3.3}$$

where $p = F_0^{0L}(r)$, the CDF of R_0^{0L}. Note that r is the percentile in the reference group for a given value of the attribute, y_r, and p is the percentile in the location-adjusted group at that same value.

3.2 Further decomposition of shape

The decomposition in the previous section defined shape as the residual differences that remain after an adjustment is made to match the locations of the two distributions. In this section we further decompose the shape component to pull out the difference in scale between the two distributions. If the comparative distribution is a simple location-scale shifted version of the reference distribution, that is, $F(x) = F(\frac{x-c}{s})$ for some constants c and s, then the difference between the two distributions can be parsimoniously summarized by these two characteristics (cf., the location-scale quantile regression models in Chapter 13). Differences that remain after a location

and scale adjustment are now residual differences in shape. To distinguish this definition of shape, we refer to it as residual shape. Our measure of scale in the example here is the standard deviation. As with location, other measures of scale can and should be used where appropriate.

Let Y_{0LS} denote a random variable describing the reference group *location-scale adjusted* to have the same mean and standard deviation as the comparison group. Let $\sigma(Y_0)$ and $\sigma(Y)$ be the standard deviations of the reference groups respectively, and define $\nu = \sigma(Y)/\sigma(Y_0)$. For an additive location and scale shift, we define Y_{0LS} as the random variable $\nu(Y_0 - \mu_{Y_0}) + \mu_Y$ The CDF of Y_{0LS} can be written as $F_{0LS}(y) = F_0((y-\gamma)/\nu)$, where $\gamma = \mu_Y - \nu\mu_{Y_0}$. The density corresponding to F_{0LS} is $f_{0LS}(y) = f_0((y-\gamma)/\nu)/\nu$. Y_{0LS} defines a hypothetical group which has the location (here the mean) and the scale (here the standard deviation) of the comparison group, but the residual shape of the reference group.

Let $R_0^{0LS} = F_0(Y_{0LS})$ be the relative distribution of Y_{0LS} to Y_0, $R_{0L}^{0LS} = F_{0L}(Y_{0LS})$ be the relative distribution of Y_{0LS} to Y_{0L} and $R_{0LS} = F_{0LS}(Y)$ be the relative distribution of Y to Y_{0LS}. The first level is the location-scale adjustment and the second measures the additional effect of the scale above and beyond location.

The decomposition can again be represented in terms of the density ratios:

$$\begin{aligned} \frac{f(y_r)}{f_0(y_r)} &= \frac{f_{0L}(y_r)}{f_0(y_r)} \times \frac{f(y_r)}{f_{0L}(y_r)} \\ &= \frac{f_{0L}(y_r)}{f_0(y_r)} \times \frac{f_{0LS}(y_r)}{f_{0L}(y_r)} \times \frac{f(y_r)}{f_{0LS}(y_r)} \end{aligned} \tag{3.4}$$

or, in more heuristic terms:

$$\begin{aligned} \text{overall relative density} = {} & \text{density ratio for difference due to location} \times \\ & \text{density ratio for difference due to scale after adjusting for location} \\ & \times \text{density ratio for residual shape difference} \end{aligned} \tag{3.5}$$

As before, these three effects form an exact sequential decomposition of the relative distribution of Y to Y_0 in the sense that R_{0LS} is the relative distribution of R_0 to R_0^{0LS}, and that R_{0L}^{0LS} is the relative distribution of R_0^{0LS} to R_0^{0L}.

The density ratio for the effect of the location difference is again a proper density ratio because it uses f_0 as the reference distribution. The other components are in general not proper densities, but they do represent the multiplicative increment at the right point of the outcome scale, y_r. The relative densities for all the components can again be graphically displayed. Mathematically, the relationship between the densities is:

$$g_0(r) \;=\; g_0^{0L}(r) \times g_{0L}^{0LS}(p) \times g_{0LS}(q) \qquad 0 \le r \le 1,$$

where $p = F_{0L}(r)$ and $q = F_{0LS}(r)$. Note that r is the percentile in the reference group for a given value of the attribute, y_r, while p and q are the percentiles in the location and location-scale adjusted group at that same value, respectively.

This sequential approach can be extended to additional parametric effects. If, for example, we wished to extract that component of residual shape that was normally distributed we would define a hypothetical group with CDF $F_G(y) = \Phi\big((y - \mu(F))/\sigma(F)\big)$ where $\Phi(\cdot)$ is the CDF of a normal distribution with mean zero and standard deviation one. The residual shape term would be decomposed into a term that measured the deviation of reference group from normality and a final residual shape term that measured the deviation of the comparison group from the normalized reference group. Generally, if it was believed that the relative CDF of the comparison group to the location-scale adjusted reference group group was $V(p),\ 0 \le p \le 1$, then we would define a hypothetical group with CDF $F_G(y) = V\big(F_0((y-\gamma)/\nu\big)$ and decompose the residual shape term relative to this distribution. Each parametric effect measures the additional impact of the parametric term in the sequence while the final term measures the residual shape effect. Altering the order of the parametric terms in the sequential decomposition will, in general, change the size of their effects, as in any sequential decomposition. This can be informative about the joint and individual impacts of the different terms.

Exercises

Exercise 3.1. Recall that a distribution is defined as symmetric if $F_0(\theta - x) = 1 - F_0(\theta + x)$ for all x where θ is the median of F_0. Suppose that F_0 is a symmetric distribution Suppose the $F(x)$ is a location-scale adjusted version of F_0, that is, $F(x) = F_0((x-\theta)/\theta)$. Will the RD be symmetric? If so, give a proof of the result. If not, give a counter example.

Exercise 3.2. Using the CPS earnings data for men and women in 1997, decompose the relative distribution into location, scale and residual shape shifts. Use men as the reference distribution, the raw (untransformed) earnings, an additive mean location adjustment, and a standard deviation scale adjustment. Interpret each component. Do the patterns in the residual shape component suggest any substantive hypotheses?

Exercise 3.3. Repeat Exercise 3.2, using women as the reference distribution. What changes and what remains the same? Do the substantive findings change? Are patterns now visible in the display that were less visible when men were used as the reference population?

Exercise 3.4. Repeat Exercise 3.2, using log-earnings, an additive median location adjustment, and an additive standard deviation scale adjustment. Describe any differences in the substantive findings.

Exercise 3.5. Repeat Exercise 3.4, using an additive IQR scale adjustment. Describe any differences in the substantive findings.

Exercise 3.6. Is one or the other of the location and scale adjustment alternatives more appropriate in this context? Explain why.

Chapter 4

Application: White Men's Earnings 1967–1997

4.1 Background

Previous research has shown that the earnings of American workers went through a series of dramatic changes over the last three decades. After years of strong growth, real wages began to stagnate in the 1970s (Bell and Freeman 1986), especially for workers with low education levels (Juhn and Murphy 1993). Poverty rates began to rise, after decades of steady contraction (Sawhill 1988), and the convergence of black to white earnings slowed noticeably (Juhn, *et al* 1991). The trend that attracted most attention, however, was the unprecedented growth in wage and earnings inequality during the 1980s. Using standard measures like the Gini index, researchers in the mid-80s documented increases on the order of 20–30%. Good reviews of this literature can be found in Levy and Murnane (1992) and Danziger and Gottschalk (1996). Much research has since been done to identify the origins of these large and rapid shifts. A number of factors appear to have contributed, including demographic changes (for contrasting views, see Dooley and Gottschalk 1982; Schrammel 1998; Welch 1979), industrial shifts (Danziger and Gottschalk 1993; Harrison and Bluestone 1988; Kosters and Ross 1987; Rosenthal 1985), technological changes that penalize workers with less education (there is more theory than evidence to support this, cf., Howell, *et al* 1998 for a critical review of the literature), changes in international trade and the "globalization' of access to a low-wage workforce (Sassen 1988; Wood 1994), the decline of worker's institutional protections like unions and the minimum wage (Card and Krueger 1995; DiNardo, *et al* 1996), and the reorganization of work and production at the firm level (Belous 1989; Cappelli 1995; Harrison 1994).

While much insight has been gained from this research, a fundamental question remains unanswered: Are the many changes in labor market structure leading toward a brighter, higher wage future for American workers (albeit through a bumpy transition), or are these changes producing a permanent rise in inequality?

Those who see a brighter future typically argue that the driving force behind these trends is a disparity between the high skill requirements of

postindustrial jobs, and the mediocre education and training which certain groups of workers bring to the labor market. The service economy shift, by upgrading rewards to the educated, is increasingly leaving behind the uneducated – especially poor women and minorities. Rising inequality thus results from insufficient human capital, and proponents of this view call for the institution of supply – side solutions, such as education and training programs, to remedy the problem (Berlin and Sum 1988; Johnston and Packer 1987). For these theorists, inequality will decline once supply catches up with demand.

Those who are less optimistic about the future argue that the shift to a service-based economy has produced an increasingly polarized job distribution: an upper tier of jobs with high wages, security, and mobility opportunities; a bottom tier of dead-end, low skill, often temporary and part-time jobs with low pay and security; and a dissipating middle range. Ultimately, they argue, the problem lies with the type of jobs being generated by industrial restructuring. Supply-side policies that emphasize education and training cannot overcome the increasingly polarized structure of demand, though they may alleviate the plight of those at the very bottom, i.e., the "underclass" (Auletta 1982; Harrison and Bluestone 1988; Sassen 1988). From this perspective, the increases in inequality are seen as relatively permanent, and unlikely to change unless the course of industrial restructuring is changed.

The two theses imply quite different trends in empirical inequality. The first implies an upgrading of the wage distribution: growth of jobs in the upper tail of the distribution, initially leaving behind a stagnant segment of unskilled jobs and workers in the lower tail, but eventually resulting in better wages and jobs for all. The second implies a steady, increasing polarization of the wage distribution: workers moving toward high- and low-wage jobs, away from the middle, generating a U-shaped distribution relative to the baseline.

These two patterns are quite distinct, and it should be a straightforward task to identify which is supported by the data. Empirical investigation, however, has been handicapped by methods that do not provide access to full distributional information. As a result, it has been difficult at times to gain consensus even on the most basic descriptions of the process: whether inequality has increased, the timing of distributional changes, which groups in the population are experiencing changes, and whether the trends are statistically significant. Relative distribution methods make this a simple and interesting task.

4.2 Data

The data are drawn from the annual March Supplement of the U.S. Current Population Survey (CPS) 1967 through 1997. The sample examined here

consists of white males, aged 16-66, and excludes the self-employed, full-time students, as well as those in the military and in farming (trends for other race and sex groups are examined in Chapter 6). We take real annual earnings as our income variable, defined as the income respondents reported receiving in wage and salary before deductions during the previous year, and deflated using the "Personal Consumption Expenditure" (PCE) deflator (United States Department of Commerce 1997) The PCE deflator tends to register lower levels of inflation than the "Consumer Product Index" or (CPI), thus real wages will rise more when the PCE deflator is used. The reported earnings were top-coded at varying levels through the years, starting at $50,000 in 1967 and rising to $200,000 by 1997. We have imputed values for these topcoded earnings in each year (about 0.5% of the cases) using a Pareto distribution. The mean of these imputed values is about 1.45 times the topcode; the value traditionally assigned to topcoded earnings.

It is worth noting that several conventional proxies are used in operationalizing the hypotheses of this analysis. First, while much of the debate is about changes in the number of jobs at different wage levels, we analyze the number of workers at each level. This proxy is used out of necessity since the major reliable data sets take the individual, not the job, as the unit of analysis. Second, while theories often talk about jobs in terms of their skill content, the analysis here is limited to earnings. This is because the first task of any analysis is descriptive rather than explanatory. Here we wish to provide a simple documentation of the trends in inequality. In Chapter 8 we will examine some of the skill-based claims using statistical extensions of the relative distribution techniques developed in the following chapters, but direct measurement of skills is a thorny problem (see Spenner 1985). This is especially true when studying trends over time, since the most common source of job skill measures, the *Dictionary of Occupational Titles*, has made substantial changes in skill and occupational definitions over the past two decades. Clearly, investigating the relations among skills, education, and wages, and differences in trends for each, remains a critical issue at both conceptual and empirical levels (see Howell and Wolff 1991).

The CPS March earnings series has been used in many studies of wage inequality. Note, however, that annual earnings reflect both the wage of a job or jobs (a demand side indicator) and the hours worked by the respondent (a supply side indicator). Earnings provide a good picture of the net impact of labor market changes on the living standard of workers, but there are other measures that are better suited to other questions. A better measure of changes in the structure of jobs would be hourly wages, which removes the confounding effect of hours worked (Juhn and Murphy 1993; Murphy and Welch 1992). We turn our attention to wages in Chapter 8. An alternative measure of living standard is household earnings; as many workers pool income with other household members. Household earnings can also be used to investigate the changing contribution of husbands, wives, and others to the family income pool (Cancian 1998). Finally, the longer

term impacts are perhaps best reflected by measures of wealth, rather than income (Wolff 1995).

To better understand the advantages of the relative distribution framework, we will compare it to more traditional methods for analyzing these data.

4.3 Findings

Using traditional descriptive techniques, the simple task of presenting the earnings trends over a 30-year period is more difficult than it might seem. Capturing the key distributional changes in a parsimonious and interpretable way over this long of a time series is a challenge. While PDF overlays are a good tool for comparing two or three distributions, 30 overlaid PDFs would be virtually impossible to decode.

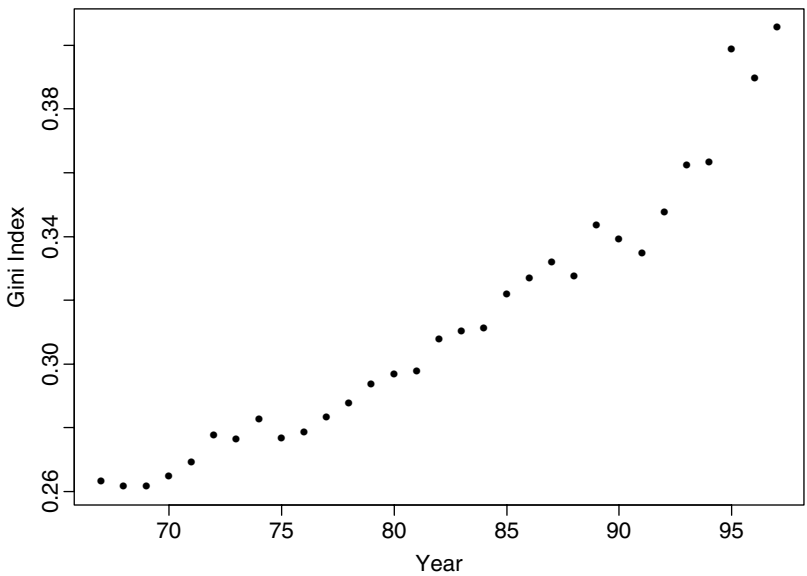

Fig. 4.1. The Gini index for annual earnings: 1967–1997.

One standard approach would be to summarize the distributional information into the Gini index and plot the Gini series over time. This plot is shown in Figure 4.1. The Gini series conveys the clear message that inequality has risen over the last 30 years, and that the rates of change have varied. In particular there seems to be a dramatic rise in inequality in the

last three years, 1995–1997. The plot conveys little information, however, on the questions of upgrading versus downgrading that are central to the alternative hypotheses we wish to examine. In addition, the interpretation of the Gini index series is complicated by the possibility that the underlying Lorenz curves may be crossing. Plotting all 30 Lorenz curves would solve this in principle, but in practice such a plot, like the overlaid PDFs, would be nearly impossible to read.

Perhaps the best traditional display in this context is the running boxplot, shown in Figure 4.2. This plot provides a compact summary of the yearly earnings distributions on a single scale, and a quick scan of the display permits a relatively accurate comparison of level, scale, and skewness. The boxes do a good job of presenting the information in the interquartile range of the distribution, but provide less useable detail in the tails of the distribution. Note that, in contrast to the Gini series, the boxplots do not suggest a dramatic rise in inequality in the last three years. The Gini index, like the Lorenz curve, is more likely to be affected by outliers in the tail of the distribution. The boxplot, and the relative distribution graphs, because they are based on quantiles, are less sensitive to outliers.

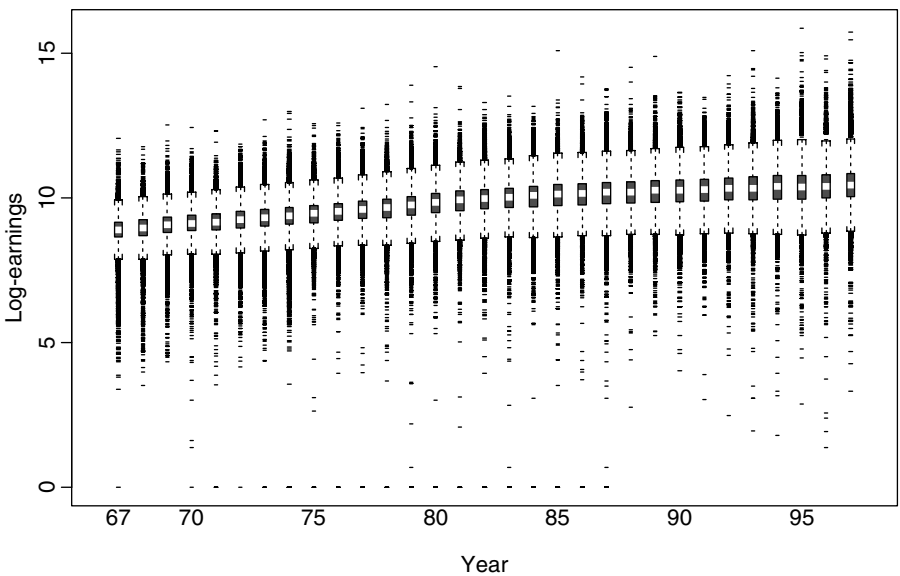

Fig. 4.2. Running boxplot of the annual earnings distribution for white men: 1967–1997.

In the relative distribution framework, we can use the decile time series plot to display these data. This plot is presented in Figure 4.3. The display

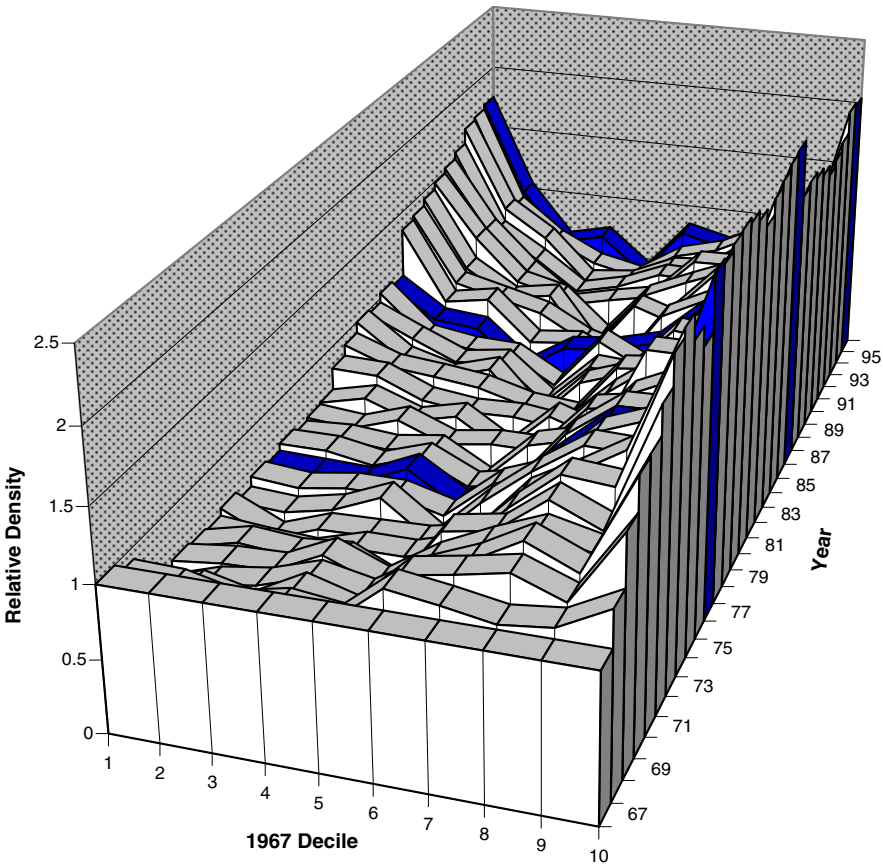

Fig. 4.3. The relative deciles for white men's earnings: 1967–1997.

has many advantages. Like the running boxplots, the relative deciles provide a visual image that is quickly scanned and easily understood, with much distributional detail preserved. Like the Gini series, the level of inequality is represented directly, so that key substantive information is visually accessible. In contrast to both of these plots, however, the relative deciles code *comparative*, rather than raw, distributional information. Here, the series displays the change from the baseline year, 1967. Every 10th year is shaded to give a sense of the progression of changes over time. One could instead highlight recession years or other years of interest. While both the boxplots and the relative distribution are based on quantiles, the decile RDs are bet-

ter at revealing the detail in the tails of the distribution. In this application, the tails are where the action is.

Several trends are apparent from the decile series in Figure 4.3. The early changes in the earnings distribution are marked by a general upshifting in wages: the density of earners in the top decile doubled over the first decade or so, while the density of earners in the bottom decile fell by nearly 50%. This is the image of a "rising tide that lifts all boats." By the 1980s, however, the changes begin to be driven by growth in the lower tail: the relative density in the bottom deciles rises back to its original level, while the relative density in the upper deciles stagnates. Lower tail growth continues as the dominant trend through the 1990s. The net result is a strong polarization in real earnings by the end of the 30-year period. As the series represents changes relative to 1967, the interpretation of the trend is quite striking: the growing density in the bottom deciles during the later years not only wiped out all of the gains these low-earning men made during the 1970s, it actually reversed them.

To get a more compact picture of the timing and nature of these changing trends, we can break the 30-year period into 3 decades, and compare the changes across the decades. Using traditional tools, one might plot the PDF overlay for 1967, 1977, 1987, and 1997, or the Lorenz curves for these years. These two displays are shown in the two panels of Figure 4.4. Several aspects of the earnings trend are apparent from these figures: real wages grew, then declined over this period, and the 1997 earnings distribution is more dispersed than any of the earlier years (the fatter tails and smaller peak are quite marked in the PDF overlay). The Lorenz curves are perfectly ordered, indicating a consistent trend of rising inequality over the three decades. The largest growth appears to come in the last decade.

Neither of these displays provides much information on the relative impact of location and shape changes over each decade. They also do not convey whether the upper and lower tails of the distibution are growing at the same rate, or for the same reasons (i.e., location or shape driven). This is what relative distribution methods are particularly good at pulling out of the data.

In the relative distribution framework, we can plot the relative PDFs for each decade. This plot is shown in Figure 4.5. In contrast to the 30-year decile series, which takes 1967 as the reference distribution for all years, each panel here takes the beginning year of the decade for the reference distribution and the end year of the decade for the comparison. This display therefore highlights the changes that took place *within* each decade. The differences in these changes are striking. The early 1970s were clearly marked by a strong upshifting in earnings, the 1980s by nearly symmetric polarization, and the 1990s by an earnings downshift, with only the top decile escaping the trend. While the shape of the RDs clearly points to the dominant trend for each decade – location shifts in the 1970s and 1990s, and shape shifts in the 1980s – the dominant trend may be masking some

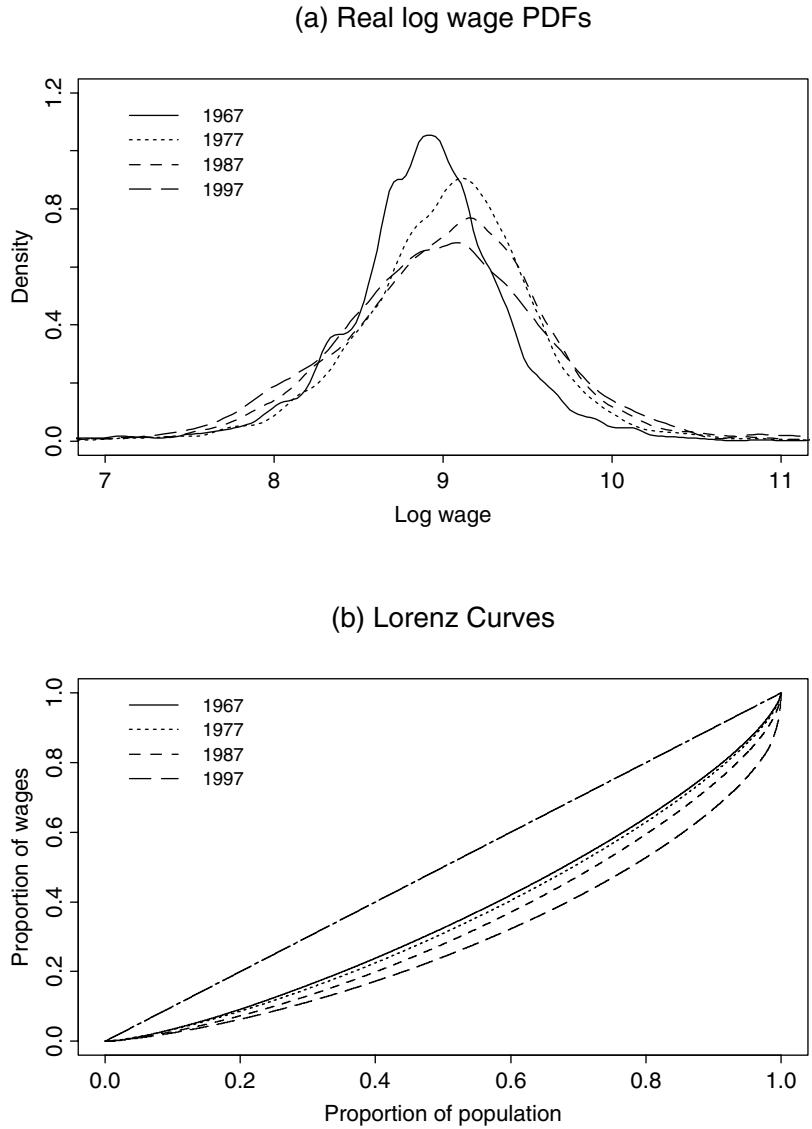

Fig. 4.4. PDFs and Lorenz Curves for the 1967, 1977, 1987, and 1997 earnings distributions for white men.

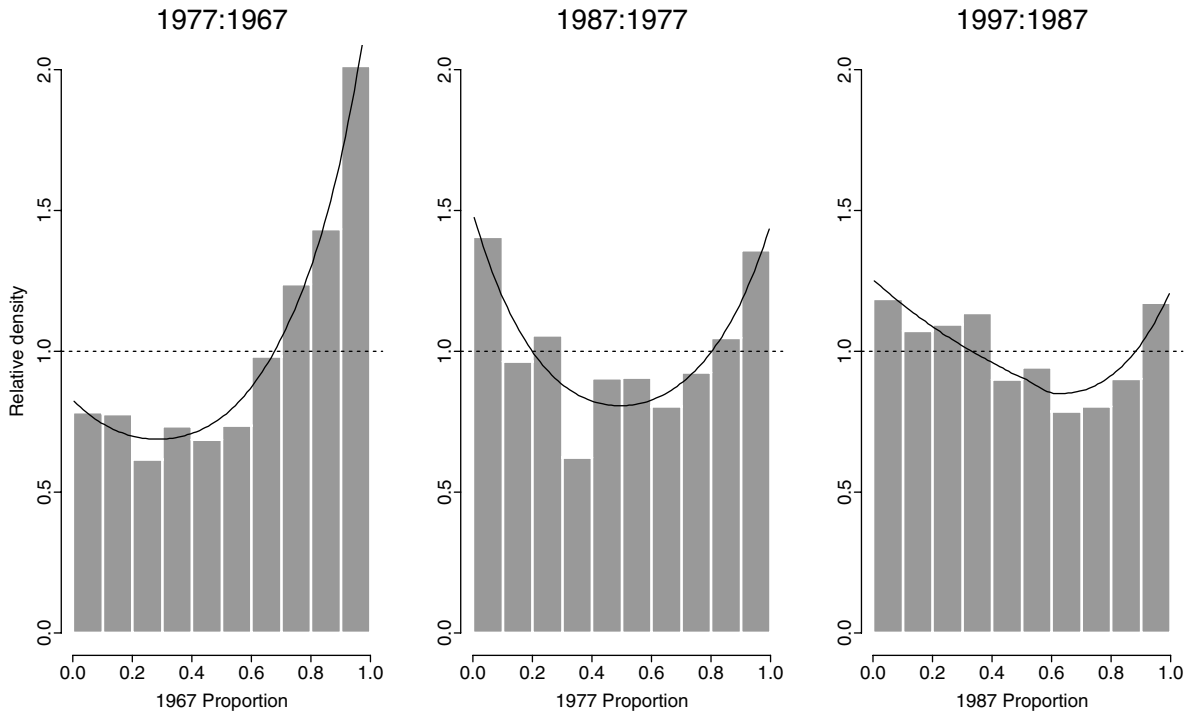

Fig. 4.5. Relative distributions of white men's earnings by decade, 1967–1997.

of the more subtle changes. To see this, one can plot the location- and shape-adjusted RDs.

The relative impact of location and shape shifts to the overall changes in each decade can be seen in Figure 4.6. We have used an additive median shift here for location adjustment because we are working with log-wages. Each row in this display represents a component of the change. The top row shows the overall change by decade (the same as that shown in Figure 4.5), the middle shows the effect of the median shift (the shape-adjusted RD), and the bottom shows the effect of the shape shift (the median–adjusted RD). The displays again highlight the distinctiveness of the earnings trends in each decade. The median upshift in real earnings during the 1970s was clearly the dominant factor during that decade, as expected. But there was also a small polarization trend that was not evident in the overall RD. This suggests that while the great majority of white men experienced growth in their real earnings during this period, some were already beginning to fall behind. It was not just those at the very bottom either; relative growth can be seen in each of the three lowest deciles. By the 1980s, the growth in real earnings came to a complete halt – the RD for the effect of the location shift is flat, indicating that if there had been no change in the shape of the earnings distribution during this decade, there would have been no change at all. Earnings polarization, by contrast, picked up speed during this decade. The strongest effects were in the top and bottom deciles, indicating that more men were now being left farther behind, wiping out any gains they might have seen in the previous decade. At the same time, a nearly equal fraction had joined the ranks of those whose earnings put them in the top 10% at the decade's beginning. In the 1990s, median real wages deteriorated. One can see the downshift clearly in the location panel. It is not as strong an effect as the upshift in the 1970s, but it is clearly in the opposite direction. The fraction of men whose earnings would have placed them in the bottom reference decile rose again, but this was now largely due to the general earnings downshift, rather than to polarization. Polarization did play a role during this decade, however, making it possible for those who joined the top decile to hold on to their gains while everyone else was losing. In contrast to the message conveyed by the Lorenz curves in Figure 4.5, the largest growth in polarization is shown here to have occurred during the middle decade, not the final one.

4.4 Discussion

The story told by these graphs is the "Great U-Turn" predicted by Harrison and Bluestone in 1988: real wages rose, stagnated, and then fell, while earnings inequality grew throughout the period. Those at the bottom of the distribution were hit the hardest, with both falling real earnings and growing inequality combining to push them farther behind at the end of

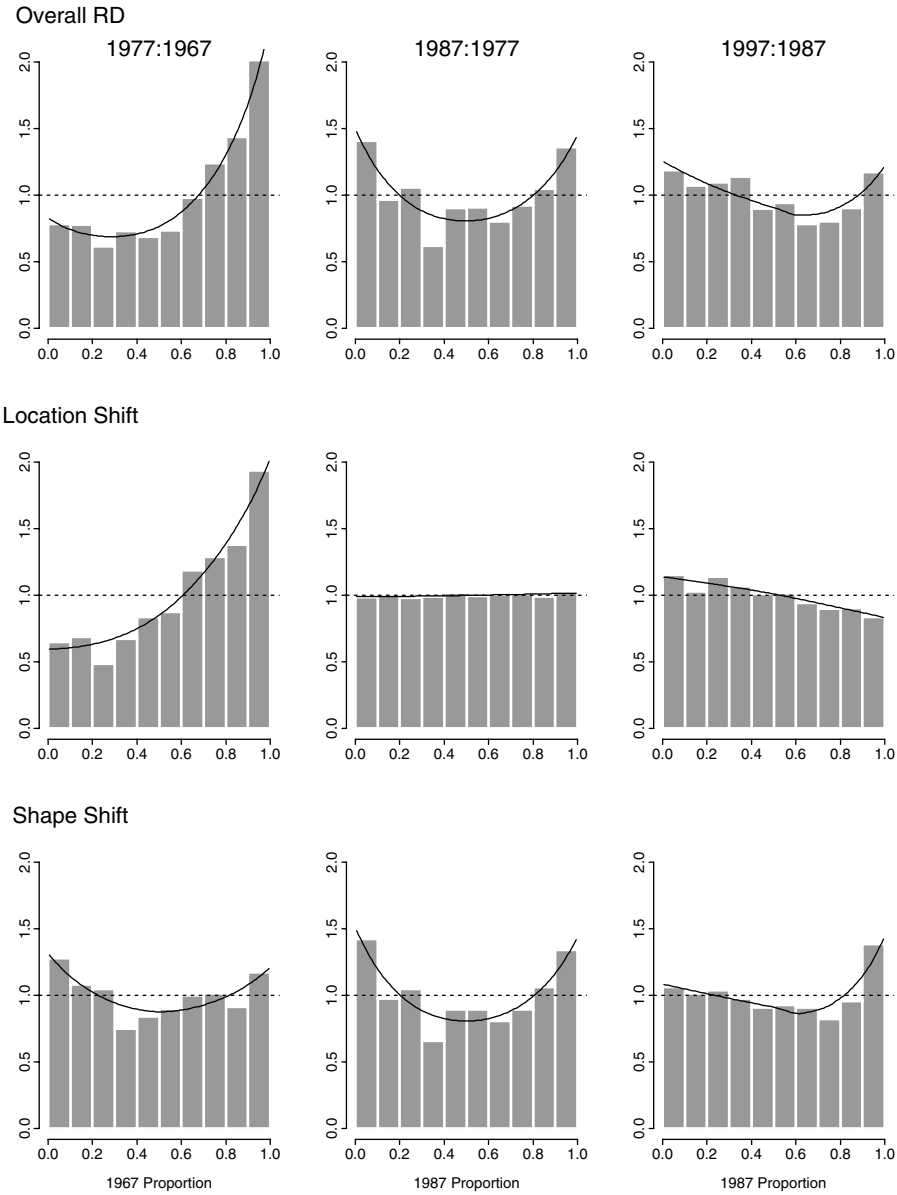

Fig. 4.6. Location and shape decomposition of white men's earnings by decade, 1977–1997.

the 1990s than they had been 30 years earlier. Those at the top made great gains in the 1970s by riding the egalitarian tide, but held on to these gains over the next two decades by staking out the top of an increasingly unequal distribution. The evidence gives little support to the supply-side advocates. Their best decade was the 1970s, when real wages grew, suggesting a shift to the higher wage postindustrial job structure of the future, and the small echo of polarization suggested a segment of the workforce whose rising earnings were not keeping pace with the others. These, perhaps, were the candidates for human capital investment. But it is hard to reconcile the supply-side argument with the stagnant – and then falling – real wages of the next two decades. The earners in the top decile were the only group that even held steady during this period. Everyone else lost ground. Other studies have estimated that three-quarters of the increase in inequality during this period was due to the fall in real earnings at the bottom of the distribution (Gottschalk 1997; Topel 1997). The one constant trend over all three decades was the growth in inequality, small at first, but later the dominant trend. This empirical pattern is more consistent with the alternative thesis: that a polarization in the structure of demand is leaving a permanent legacy of inequality.

Exercises

Exercise 4.1. Calculate the usual summary statistics for the distribution of earnings in 1967 (e.g., mean, median, standard deviation and interquartile range). Repeat the process for earnings in 1997. Based on these summaries, write a brief comparison of the two distributions.

Exercise 4.2. Calculate the Gini indices for the log-earnings in 1967. Are they the same as the values in Figure 4.1? If they are the same, explain why. If they differ, which of the two scales is more appropriate for this application?

Exercise 4.3. Calculate Theil's index for the earnings for each year from 1967 to 1997. Create a plot similar to Figure 4.1. How correlated are the indices of inequality? Construct a linear regression of the Gini index based on Theil's index. Are the usual assumptions of linear regression satisfied? If appropriate, modify the model. Identify years where the relationship between the two indices deviates from the norm. By looking at the distributions for those years, explain why they differ.

Exercise 4.4. A third index of inequality is the ratio between the 90th and 10th percentiles of the distribution. Repeat Exercise 4.2 using this measure. As an additional tool for addressing the questions, build a regression model for the Gini index based on both the percentile distance and Theil's index.

Exercise 4.5. Why are the PDFs in Figure 4.4 so rough? While increasing the degree of smoothing will reduce the roughness, is it the right thing to do in this situation? Discuss.

Exercise 4.6. Describe the pattern in mean earnings in Figure 4.2. Describe the pattern in scale and any apparent pattern in skewness. Are the distributions becoming less symmetric over time? What do the numbers and locations of the outliers indicate about the symmetry in the tails of the distributions?

Exercise 4.7. The reference distribution in Figure 4.3 is 1967 earnings. How different will Figure 4.3 appear if 1968 is used instead? How will it look if the final year (1997) is used. Discuss how any changes effect the substantive conclusions.

Exercise 4.8. Calculate the RD of 1967 to 1977. How does it compare to the first panel in Figure 4.5? Calculate the RD of 1977 to 1987 and compare it to the second panel. Do these "reverse" RDs provide the same information as the originals?

Exercise 4.9. Suppose that location was defined as a multiplicative median shift in Figure 4.6 rather than an additive median shift. Would the figure differ? When would the multiplicative median shift be more appropriate?

Exercise 4.10. Recreate Figure 4.6, pulling out both scale and location in the decomposition. How much of the shape effect is due to the scale difference? Do the residual shape differences suggest any substantive hypotheses?

This page intentionally left blank

Chapter 5

Summary Measures

5.1 Motivation

While graphical displays are a key part of the relative distribution framework, summary measures remain an important tool for the comparison of distributional change. A good summary statistic makes it possible to provide a simple and precise answer to a substantive question such as "has inequality in wage profiles grown significantly over the past 20 years?" or "has the upgrading in wage-gains been matched or exceeded by the downgrading?" The relative distribution provides a general framework for defining a wide and flexible range of summary measures. The generality of this framework is due to the fact that the relative distribution captures *all* of the information that is necessary and sufficient for strongly scale-invariant comparison.

Summary measures based on the relative distribution are robust to both outliers and to deviations from assumptions. This robustness follows from two properties of the relative distribution: the rescaling of the comparison distribution to the reference distribution, and the absence of parametric assumptions. Rescaling limits the impact of outliers. Outliers in either the reference or comparison distribution are not necessarily outliers in terms of the relative distribution, and the rescaling maps the original units of both distributions to a rank measure (i.e., [0, 1]) moderating the influence of abnormal values. As a result, summary measures based on the relative distribution are less likely to be influenced by problem cases. The absence of parametric assumptions means there are fewer assumptions to violate. The relative distribution, as well as the decomposition techniques, and natural summary measures in this framework are fully nonparametric. They require minimal assumptions about the underlying distributions – either in terms of the individual distributions or in terms of their relationship to one another. This actually distinguishes the relative distribution methods from other nonparametric approaches, most of which implicitly assume that the reference and comparison distributions have a well defined relationship to each other (e.g., are simply location shifted versions of each other) (Lehmann 1975).

In the next section we discuss the construction of measures of distributional divergence based on the relative distribution. We then discuss the decomposition of these measures to determine the contributions of location, scale and shape.

5.2 Measuring distributional divergence

The study of measures of the "closeness" of two distributions has a long history (Adhikari and Joshi 1956; Ali and Silvey 1966). There are a multitude of different measures, each motivated by the the application for which it was developed. Some of these are based on appeals to underlying statistical principles and some are *ad hoc.* Each has the common property of increasing as the distributions become more dissimilar. The measures differ in their scale and the type of deviation between the distributions that has the greatest influence. Ali and Silvey (1966) argue that it is reasonable to restrict measures of divergence to functions of the relative density, at least when the objective is to distinguish the two distributions on the basis of observation. This argument is based on the fact that the relative CDF is a sufficient statistic for the problem of comparison. We shall also examine measures that can be expressed purely in terms of the relative density and without separate reference to the underlying distributions.

Ali and Silvey define four basic properties that any measure of divergence should have. First, the measure should be well defined for all distributions. Second, the measure should be minimized when the two distributions are equal. Third, the measure should not increase when the data are aggregated into groups. And fourth, changes in a parameter should affect the measure of divergence in the same direction as the likelihood.

Based on these considerations they propose the following class of measures of divergence:

$$D_\phi(F;F_0) = \int_0^1 \phi\Big(g(r)\Big)\, dr \ ,$$

where ϕ is any continuous convex function on $(0,\infty)$. This class has also been proposed by Csiszár (1978). He refers to it as *directed divergence* measures or f–divergences. This is a very wide class that contains most commonly used measures, each corresponding to a particular choice of the weight function ϕ. Many of these are summarized in Table 5.1.

Table 5.1. Measures of divergence and their corresponding ϕ-functions

$\phi(p)$	Divergence Measure
$(p-1)\log(p)$	Jeffrey's or J-divergence
$-\log(p)$	Kullback's directed divergence
$p\log(p)$	Kullback-Leibler divergence
$\frac{1}{2}(\sqrt{p}-1)^2$	Kolmogorov's measure of distance Hellinger divergence
$\frac{1}{2}\lvert p-1\rvert$	Kolmogorov's variation distance L_1 divergence
$-p^{1-\lambda}$	Chernoff's measure of discriminatory information, $0<\lambda<1$
$1-p^{\lambda}$	Generalized Bhattacharya measure, $0<\lambda<1$
$[p^{\lambda}+p^{1-\lambda}]/(\lambda-1)$	Divergence of degree $\lambda, \lambda\neq 1$
$(p-1)^2$	Chi-squared divergence. Kagan's measure
$[p^{\lambda}-1]/\lambda(\lambda+1)$	Power weighted divergence $\lambda\neq 0,-1$

Rao (1982) studies a number of these measures and their applications in statistics.

While this is a wide class, it does not contain every reasonable measure of divergence. The median of the relative distribution and its interquartile range are two examples that would not be included. In the final section of this chapter we will consider a measure of polarization that is also not in this class. The justification for such measures is that they seek to capture special features of the divergence, rather than the overall dissimilarity of the distributions.

An additional property of measures that would be useful for our purposes is that they be decomposable in a way consistent with the decompositions we have developed for the relative density. This, unfortunately, is not generally possible. None of the measures in Table 5.1 is directly additively decomposable in the sense that

$$D_{\phi}(F;F_0) = D_{\phi}(F_{0L};F_0) + D_{\phi}(F;F_{0L}).$$

The problem arises in the rescaling of the second component, a point that was discussed in Section 3.1. Recall the relative density decomposition in (3.3):

$$g_0(r) \;=\; g_0^{0L}(r)\times g_{0L}(p) \qquad 0\leq r\leq 1,$$

where $p = F_0^{0L}(r)$. Had the argument for the final term been r instead of p, the divergence for $g_0(r)$ would have a simple decomposition into the two components for some measures. With the rescaling, however, there is no direct way to separate the integral for the total divergence into the two components for any measure.

Another way to see this is to think of the triangle formed by the three distributions F, F_0, and F_{0L} in the space of distributions. The deviance from F to F_0 will in general be less than the sum of the deviances of F_{0L} to F_0 and F to F_{0L}. Substantively, this means that the two component shifts may be operating in such a way as to counteract each other, so the effect of each separately may be more than the joint effect. While the components may be summable for particular distributions, there is a set of distributions for each directed divergence for which a direct decomposition is not possible.

As with the relative density, we will need to rescale one divergence component to make decomposition possible. This will generally mean that the components do not sum to the value for the overall divergence. The Kullback-Leibler measure (Kullback 1968), however, preserves the interpretation of the divergence for each component, and this gives it an advantage over the other measures.

5.3 Two measures of distributional divergence

The choice of measure, and its interpretation, should be application specific. The choice depends on the character of the difference between the distributions that the measure should be sensitive to. Here we consider two omnibus choices that are in wide use: chi-squared divergence and Kullback-Leibler divergence.

The *chi-squared divergence* is defined by

$$D_\chi(F;F_0) = \int_{-\infty}^{\infty} \left(\frac{(f(x) - f_0(x))^2}{f_0(x)} \right) dx = \int_0^1 \left(g(r) - 1 \right)^2 dr.$$

It is also called Pearson's ϕ^2 measure. For a detailed description, see Lancaster (1969). The measure represents the squared distance between the two densities normed by the prevalence of the reference group. This is clearer in the second expression, where it is simply the squared deviation of the relative density from the uniform density. Thus it weights deviations from uniformity by the square of their magnitude, similar to the weighting used by the variance and least-squares linear regression estimates.

The chi-squared divergence has been studied by Eubank, *et al* (1987). They use it as the basis for a test for the equality of the two distributions ($D_\chi(F;F_0) = 0$). It can also used to test other hypotheses of interest such as symmetry of a distribution and goodness-of-fit to data. We treat this topic in Section 10.2.2.

Perhaps the most commonly used measure of the divergence between two distributions is the *Kullback-Leibler divergence* defined by:

$$D(F;F_0) = \int_{-\infty}^{\infty} \log\left(\frac{f(x)}{f_0(x)} \right) dF(x) = \int_0^1 \log\big(g(r)\big) g(r)\, dr.$$

$D(F; F_0)$ is also known as the information number, discrimination function, and "distance."

The links between the Kullback-Leibler divergence and the relative distribution has been studied by Mielniczuk (1992) and Parzen (1994). The expression on the right-hand side of the equation is just the (differential) *negative entropy* of the relative density (Shannon 1948). Entropy is also a widely used measure of the dispersion of a distribution (Theil and Laitinen 1980). In this context we can interpret $D(F; F_0)$ as the expected information for discriminating g from a uniform distribution based on a single observation from R. For a detailed discussion of this measure, see Soofi (1994). We will use this measure as the basis for the decomposition summaries below.

5.4 Effect summary statistics

Recall from Chapter 3 that F_{0L} is the CDF of the location-adjusted reference group. In this section, we develop a summary measure that plays a role similar to the partial R^2 in traditional linear modeling, and provides an answer to the question: "How much does the location shift contribute to the difference between the two distributions?"

Summarizing the effect of location and shape changes on the overall relative distribution requires that we choose one of the measures of the overall distributional difference. The Kullback-Leibler divergence has a simple interpretation in terms of the relative distribution, and it is decomposable into the location, shape, and other components of interest, so we will work with it here.

The most direct way to summarize the contributions of location and shape is to compare the entropies of the three components of the decomposition in (3.3): $D(F; F_0)$, $D(F_{0L}; F_0)$, and $D(F; F_{0L})$. The first measures the overall divergence between the comparison and reference groups. The second measures the divergence between the the location-adjusted reference group and the reference group. This divergence directly summarizes the effect of a location shift on the distributional divergence. If the overall divergence is entirely due to a location shift then this component will equal the overall divergence $D(F; F_0)$. The third measure is the divergence between the comparison group and the location-adjusted reference group. This measures the divergence due to shape differences. If the overall divergence is entirely due to a location shift, then this component will be zero. If there is no deviation due to location, then this component will equal the overall divergence $D(F; F_0)$.

As discussed above, these entropies will not decompose directly. We can, however, use (3.1) to show

$$D(F; F_0) \quad = \quad D_Y(F_{0L}; F_0) \quad + \quad D(F; F_{0L}) \tag{5.1}$$

where

$$D_Y(F_{0L}; F_0) = \int_0^1 \log\Big(g_0^L(r)\Big) g(r)\ dr,$$

is a cross-entropy interpretable as the expected information for discriminating g_0^L from a uniform distribution based on a single observation from R. The relative sizes of these terms directly indicate the relative contributions of location and shape to the overall difference between the distributions.

This divergence decomposition can be extended to the location, spread and shape relative density decomposition in Chapter 3. Here, as there, the order in which the successive components are applied in the decomposition matters. In particular the divergence effect of spread will depend on the definition of the location adjustment. If the spread adjustment is applied before the location adjustment, a completely different decomposition of the overall divergence will result.

5.5 Measures motivated by hypothesis testing

Consider testing the hypothesis of equality between the reference and comparison distributions. Formally we consider the null hypothesis $H_0 : F(y) = F_0(y)\ \ \forall y$, where both F and F_0 are unknown. Parametric tests assume that both distributions belong to a family of distributions indexed by a (usually finite dimensional) parameter. Here we focus on nonparametric tests that assume the distributions are members of a class of distributions that cannot be indexed by a finite dimensional parameter. The properties of tests depend on the assumptions made about how the distributions differ. Alternatives to the null hypothesis can be expressed in the general form: $H_1 : F(y) = G(F_0(y))\ \ \forall y$, where G is the relative CDF. If the null hypothesis is true, then the relative distribution will be uniform, and so this test can be thought of as testing uniformity of g. The alternative hypothesis can be made specific by placing restrictions on g. For example, we can consider the *location alternatives* that assume $F(y) = F_0(y+\rho)$ for some unspecified ρ and continuous F_0. This specifies that $G(p) = F_0(F_0^{-1}(p) + \rho)$ for some unspecified ρ and continuous F_0. Another commonly used alternative family is the *scale alternatives*, which assume $F(y) = F_0(y/\rho)$ for some unspecified $\rho > 0$ and continuous F_0. This specifies that $G(p) = F_0(F_0^{-1}(p)/\rho)$.

Each specification of the alternatives can be thought of as a specification of the relative distribution, and each test can be thought of as choosing a divergence measure sensitive to the deviations specified by the alternative. In this sense, each test defines an implicit divergence measure. We give some commonly used measures below and consider inference for them in Chapter 10. Lehmann (1986) reviews much of the extensive literature on this topic.

Chernoff and Savage (1958) defined a useful class of divergence measures motivated by this testing situation. To be consistent with the standard

notation we will use the reference group formed by pooling the comparison and usual reference group (See Section 2.3). Let $H(y) = \lambda F(y)+(1-\lambda)F_0(y)$ be the CDF of the pooled reference group and GP be the relative CDF of F to H. Denote the PDF of $\tilde{R} = H(Y)$ by gp. Consider the class of measures:

$$D_{CS}(F; F_0) = \int J\Big(H(y)\Big)dF(y) = \int_0^1 J(r)\mathrm{gp}(r)dr = \mathrm{E}\Big[J(\tilde{R})\Big],$$

where $J(r)$ is called a *score function* on $[0, 1]$. The role of the score function is to weight different deviations from the uniform distribution. It can emphasize the tails or the slope of the relative distribution. The choice clearly depends on the alternative hypothesis, and different choices lead to different measures. One of the reasons this divergence measure is commonly used is because it is a function only of the relative data and has the form of a simple expectation. In practice, the score function is assumed to be non-constant and to be reasonably smooth. Specifically it must have derivatives that satisfy:

$$|J^{(k)}(r)| \leq K|r(1-r)|^{-k-\frac{1}{2}+\delta}$$

for some $\delta > 0, K > 0$, and $k = 0, 1, 2$.

We can also consider measures motivated by well known goodness-of-fit tests. For example, the Cramer-von Mises test:

$$\int_0^1 |\mathrm{GP}(p) - p|^2 dp,$$

the Anderson-Darling test:

$$\int_0^1 \frac{|\mathrm{GP}(p) - p|^2}{p(1-p)}\, dp,$$

and the Kolmogorov-Smirnov test:

$$\sup_{0\leq p\leq 1} |\mathrm{GP}(p) - p|.$$

In Section 10.2 we consider inference for $D_{CS}(F; F_0)$ as a means of conducting these tests.

5.6 Measuring distributional polarization

An important question in economic applications is whether one distribution is more unequal than another. This is a more specific shape-related question than those considered above: "To what extent does the shape difference between the two distributions take the form of rising (or declining) polarization?" Distributional polarization is of particular interest in the study

of inequality, because it captures a discrepancy in outcomes that is hidden when only trends in location are examined.

Most research on the distribution of wages and income uses one or more of the four scale-invariant measures of inequality: the coefficient of variation, the Gini index, Theil's index, and the variance of logarithms. These measures differ in a number of respects (cf., Kakwani 1980, pp. 63–95), e.g., the weight given to transfers in different parts of the distribution, but can be interpreted within a unified framework (Firebaugh 1999). For data grouped into income categories there are simple expressions for lower- and upper-bound estimates of the population values and more complicated expressions for more precise estimates (Kakwani 1980, pp. 96–125). Under fairly strong parametric assumptions about the income distribution, statistical inference for the ungrouped data can be based on the asymptotic distribution of the maximum likelihood estimator.

None of these measures, however, is designed to distinguish between growth in the upper and lower tails. Even if the measures register increasing inequality over time, one cannot distinguish a polarization of the distribution (increases in both tails) from upgrading (increases in the upper tail), downgrading (increases in lower tail). Since much of the substantive debate often turns on this level of detail, rather than on the extent of overall increases, these standard summary indices are of limited use.

The polarization index defined here and its decomposition provide a flexible and sensitive method for measuring the relative density in the center and the tails of the distribution. It plays the same role as the difference in Gini indices, coefficients of variation, or variances of log-values in measuring interdistributional inequality (cf., Grove and Hannum 1986), but it can be decomposed to compare the growth in the upper and lower tails. Because the polarization index is based on the relative distribution, it provides a simple link between what is observed in the graphical display and what is measured by the numerical summary.

5.6.1 The median relative polarization index

Isolating differences in distributional shape requires that differences in location be removed. To do this, we will focus on the location matched component of the relative distribution (g_{0L}). Measures based on this component isolate aspects of interdistributional inequality that are not due to location shifts. For example, if one distribution is more polarized than another, we would expect a U-shaped, location-adjusted relative distribution. If the distributions differ only in their level, then the location-adjusted relative distribution would be approximately uniform. As discussed in Chapter 3, location can be adjusted to equalize the medians, means, or other measures of central tendency between the two distributions. The index we develop here is based on median adjustment.

Ideally, we would like a statistic that measures the deviations of the relative distribution from the uniform distribution – as the uniform relative density represents distributional equivalence – and one that emphasizes the deviations in both the upper and lower tails. This is not unlike a standard variance measure, $\mathrm{E}(X-\mu)^2$. In the nonparametric context, it is more natural to consider more robust measures of spread, such as the median absolute deviation. The measure we develop is closely related to this. Recall from Chapter 3 that the median-matched relative distribution of Y to Y_0 is given by $R_{0L} = F_0(Y-\rho)$, where $\rho = Q(\frac{1}{2}) - Q_0(\frac{1}{2})$, the difference between the median of Y and the median of Y_0. Q is the quantile function, defined in Section 2.2. Because the medians of the two distributions have been matched, the median of R_{0L} is $\frac{1}{2}$, and our measure of polarization will reflect this. Define the *median relative polarization index* (MRP) of Y relative to Y_0 as:

$$\mathrm{MRP}(F;\ F_0) \quad = \quad 4\mathrm{E}\left[\mid R_{0L} - \frac{1}{2} \mid\right] \quad - \quad 1.$$

The MRP is then the mean absolute deviation around the median of the location-adjusted relative distribution, scaled to produce an index that varies between -1 and 1. If R_{0L} has a density g_{0L} (we suppress ρ), then $\mathrm{MRP}(F;\ F_0)$ can be reexpressed as:

$$\mathrm{MRP}(F;\ F_0) \quad = \quad 4\int_0^1 \mid r - \frac{1}{2} \mid g_{0L}(r)\ dr \quad - \quad 1$$

This expression makes it more clear how the measure weights the value of the relative distribution, $g_{0L}(r)$, by the distance from the center, $\mid r - \frac{1}{2} \mid$, thereby emphasizing the mass in the upper and lower tails more strongly than the mass in the center. Given the scaling, a value of zero represents no differences in distributional shape; positive values represent more polarization (increases in the tails of the distribution); and negative values represent less polarization (convergence towards the center of the distribution). If the only difference between F and F_0 is location (that is, $F_0(y) = F(y+\rho)$ for some ρ), then g_{0L} is the uniform distribution, and $\mathrm{MRP}(F;\ F_0)$ is zero, indicating that none of the differences between F and F_0 are due to differences in distributional shape.

The MRP has several useful characteristics. First, it is symmetric in the sense that $\mathrm{MRP}(F;\ F_0) = -\mathrm{MRP}(F_0;\ F)$. This means that the index is effectively invariant to whether F or F_0 is chosen as the reference distribution. Second, if the two distributions have the same median it is invariant to monotone transformations of the distributions: if $h(\cdot)$ is a monotone function on the support of Y_0, then the MRP of $h(Y)$ to $h(Y_0)$ is equal to the MRP of Y to Y_0. This means that the index will take the same value when applied, for example, to the logged data as when applied to the raw data (when a multiplicative location shift is used for the latter). For a more detailed discussion of scale invariance, see Section 2.1. Third, the MRP can

be interpreted in terms of a proportional shift of mass in the distribution from more central to less central values. A net change in mass, δp, a distance d towards the tails of the distribution (measured on the unit interval), will produce an MRP of $4d\delta p$. A value of 0.1, for example, is equivalent to a 10% population shift from the center of the distribution to the upper and lower quartiles. Finally, the MRP is decomposable along the scale of y. This makes it possible to compare the contribution of each section of the distribution to the overall polarization. A natural decomposition is the contributions made by components above and below the median of $g(r)$, and we define this decomposition in the next section.

5.6.2 Decomposing the median relative polarization index

Often we would also like to decompose the overall polarization into the contributions from the lower and upper tails of the distributions.

We define the lower (upper) polarization index (LRP) (URP) by:

$$\begin{aligned} \text{LRP}(F;\ F_0) &= 4\text{E}\left[\mid R_{0L} - \frac{1}{2} \mid \;\middle|\; R_{0L} \leq \frac{1}{2}\right] - 1, \\ \text{URP}(F;\ F_0) &= 4\text{E}\left[\mid R_{0L} - \frac{1}{2} \mid \;\middle|\; R_{0L} > \frac{1}{2}\right] - 1. \end{aligned}$$

These indices decompose the overall polarization index in the sense that:

$$\text{MRP}(F;\ F_0) = \frac{1}{2}\text{LRP}(F;\ F_0) + \frac{1}{2}\text{URP}(F;\ F_0). \tag{5.2}$$

The lower (upper) index is the contribution to the median index of the relative distribution below (above) its median. The two components can be reexpressed as:

$$\begin{aligned} \text{LRP}(F;\ F_0) &= 8\int_0^{\frac{1}{2}} \mid r - \frac{1}{2} \mid g_{0L}(r)\ dr - 1 \\ \text{URP}(F;\ F_0) &= 8\int_{\frac{1}{2}}^{1} \mid r - \frac{1}{2} \mid g_{0L}(r)\ dr - 1. \end{aligned}$$

The upper and lower indices have properties similar to the MRP: they vary between -1 and 1, have similar interpretations, and are symmetric and invariant to monotonic transformations. Positive values represent more polarization, i.e., increases in the tail of the distribution; negative values represent less polarization, i.e., convergence towards the center of the distribution. If the shape component of the relative distribution is uniform below (above) the median, then the lower (upper) index will be zero.

The decomposition (5.2) makes it possible to answer questions that commonly arise in distributional comparison. For example, is wage upgrading more pronounced than wage downgrading? Note that by matching location, the overall density of relative earnings is made equal above and

below the median. Thus the measure here is not addressing the question of whether median wages have risen or fallen (thus, whether wage upgrading is more *prevalent* than wage downgrading). Such location shifts have already been removed. The measures instead capture whether the residual changes have been more extreme above or below the median. Using the upper and lower indices, one can observe whether the increase is coming from symmetric growth in both tails of the distribution, or whether one tail is denser than the other.

Background material

A framework for the construction of summary measures based on orthogonal series expansions of the relative density has been developed by Eubank, *et al* (1987). This work presents a powerful and unifying framework for testing hypotheses about the differences between distributions using the coefficients of the expansion, but the summary measures provide indirect, rather than direct, measures of location and shape effects. These measures are discussed in depth in Chapter 10. The summary measures we present in this chapter differ both in motivation and substance.

Exercises

Exercise 5.1. Take an example where two distributions differ by a location shift and simple shape change (say, URP=LRP). Draw the PDF overlays for the two distributions, and sketch out the relative density $g(r)$. Explain why the estimates of the effect summary statistics in Section 5.4 will depend on the order of the location and shape adjustments.

Exercise 5.2. Compare the standard measures of distributional spread, the variance and the mean absolute deviation, with the MRP measure. What are the similarities? What are the differences? What does the MRP capture that comparing the standard measures from two distributions does not?

Exercise 5.3. Sketch up the following examples using PDF overlays and the relative density $g(r)$: (a) a positive location shift plus URP > LRP; (b) no location shift plus LRP < URP; (c) a negative location shift plus URP=LRP. Identify the areas on the relative density graph that should be equal.

Exercise 5.4. Show that the MRP can be reexpressed as:

$$\mathrm{MRP}(F;\ F_0) \quad = \quad 1 + 8 \int_0^{\frac{1}{2}} G_{0L}(r)\ dr \quad - \quad 4 \int_0^1 G_{0L}(r)\ dr\,,$$

where G_{0L} is the CDF of the location-adjusted relative distribution. This form can be used as the definition in the situation that the densities of the distributions do not exist. This will be very useful for discrete distributions (e.g., job tenure data, or grouped earnings data) – See Chapter 13.

Exercise 5.5. Suppose $Y_0 \sim \text{lognormal}(\mu_0, \sigma_0^2)$ and $Y \sim \text{lognormal}(\mu, \sigma^2)$. Show that

$$\text{LRP}(F_0;\ F) \ = \ 1 - \frac{4}{\pi}\arctan\left[\frac{\sigma_0}{\sigma}\right].$$

Exercise 5.6. Show that the LRP polarization can be expressed in a form useful for calculation:

$$\begin{aligned}\text{LRP}(F;\ F_0) \ &= \ 8\int_0^{\frac{1}{2}} G_{0L}(r)\ dr \ - \ 1 \\ &= \ 8\int_0^{\xi^0_{\frac{1}{2}}} F(y)f_0(y+\rho)\ dy \ - \ 1,\end{aligned}$$

where G_{LO} is the CDF of X_{LO}. Show that the URP index can be similarly reexpressed as:

$$\begin{aligned}\text{URP}(F;\ F_0) \ &= \ 3 \ - \ 8\int_{\frac{1}{2}}^1 G_{0L}(r)\ dr \\ &= \ 3 \ - \ 8\int_{\xi^0_{\frac{1}{2}}}^{\infty} F(y)f_0(y+\rho)\ dy.\end{aligned}$$

Chapter 6

Application: Earnings by Race and Sex: 1967–1997

6.1 Background

In Chapter 4, we saw that white men's earnings displayed several distinct shifts over the course of the last three decades. While upgrading was the dominant earnings trend during the 1970s, wages stagnated during the 1980s, and began to fall during the 1990s. During all three decades, earnings grew more polarized. The net impact was a dramatic growth in earnings inequality, with the fraction of men in the top and bottom relative deciles nearly doubling. A sequence of location and shape changes were taking place in the earnings distribution for white men, and the summary measures introduced in Chapter 5 will be used in this Chapter to provide a succinct picture of these changes.

In addition, we will compare these changes to the changes in earnings experienced by other groups of workers. If one thought that all workers shared the same shifting economic profile as white men, the analysis could instead turn to more detailed studies of the processes that generated these patterns. But the cumulative body of work in stratification and labor economics has made clear that labor market processes, in particular earnings determination, differ significantly by race and sex (Blau 1998; Marini 1989; Smith, *et al* 1989). Given the strong shape shifts that we observed in white men's wages during this period, it is likely that mean wage gaps do not tell the whole story. Economic changes have probably affected the distribution of earnings within as well as between groups, with different levels of upgrading, downgrading and polarization.

To compare earnings distributions across groups, the relative distribution can be defined in a number of ways. The most important choice is which reference distribution to use, e.g., the overall earnings distribution, the distribution for a particular group such as white males, or the group-specific 1967 distributions. Each of these answers a different question about trends in inequality; the first two choices emphasize the between-group changes, while the latter emphasizes the within-group changes. Understanding the within-group changes turns out to be a necessary first step. To begin directly by comparing each group's earnings distribution to either the entire

labor force or to a specific reference group makes interpretation difficult. An observed upgrading in the earnings of workers in the comparison group could stem from shifts in their own earnings, shifts in earnings of the comparison group(s), or a combination of changes in both groups. Without the prior step of performing separate within-group analyses, these processes are confounded and the sources of change cannot be identified. We therefore start by examining the within-group changes, forming the relative distribution time series for each group using its own 1967 distribution as the reference. This makes it possible to get a clear picture of the distributional shifts within each group, and to compare the pattern of shifts between them.

6.2 Data

The data used here are again the March Uniform Series of the Current Population Survey, for the years 1967 to 1997, and the sample restrictions are the same as those used in Chapter 4: full-time, year-round workers aged 16-65, not in school, the military, or farming. We now include white women, black men, and black women, as well as white men. We have chosen to use the term "black" throughout the Chapter rather than African American, because the data set used in the analysis, the CPS, requested respondents to identify their race using this label. The size of the resulting sample analyzed here for the 30-year series is over a million; the number in each year is on the order of 30,000. For the analyses broken down by race and sex, the typical sample sizes are about 20,000 for white males, 10,000 for white females, and 1500–2,000 for black males and females.

6.3 Findings

The entropy and polarization summaries for the changes in white men's earnings are plotted in Figure 6.1. The summaries are calculated on a yearly basis here, so we can now observe the changes within the decades, as well as between them. Like the relative decile graph in Figure 4.3, the measures in Figure 6.1 use 1967 as the reference year for calculating the whole series. This permits the total change from 1967 to be represented. The decade-specific changes, such as those displayed by Figure 4.5, can still be recovered by comparing the summary values at the beginning and end of each decade.

The first panel of Figure 6.1 shows the annual value of the location and shape entropy summaries. These represent the amount of the overall change in the earnings distribution since 1967 that is generated by the location and shape shifts respectively in each year. They do not show the direction of the change – e.g., whether the median is shifting up or down – but if we know

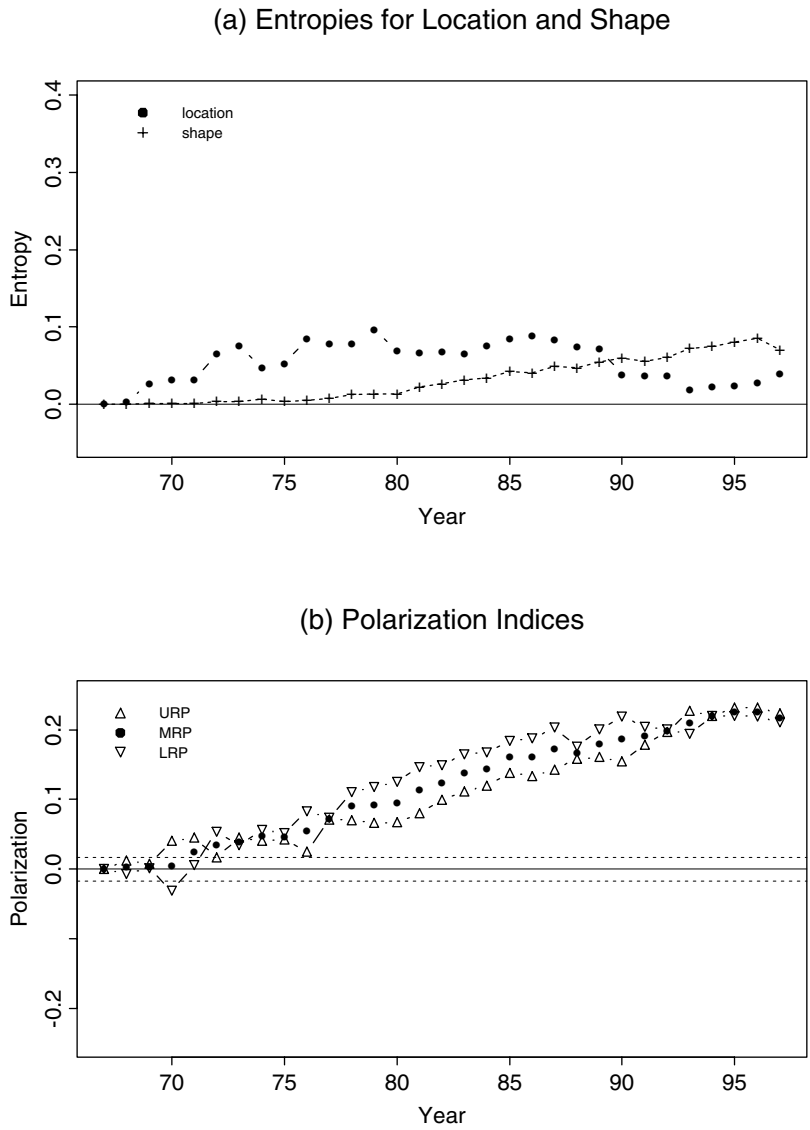

Fig. 6.1. Entropy and polarization summaries by year. Panel (a) shows the entropy summaries by year. Panel (b) shows the polarization summaries by year.

the initial direction from the relative density plots, then we can infer the direction here. The range of the y-axis in both panels has been chosen to be consistent with the later figures for the other groups. Both the location and shape entropies have approximately the same scale, so their relative size is directly interpretable.

The magnitude of the upshifting, stagnation, and downshifting of the median wage is clear in the location series. The direction – up, then down – we know from looking at the relative density displays in Chapter 4. Because 1967 is the reference throughout the series, the magnitude of the observed changes have a direct interpretation: by 1995, the median real wage had fallen nearly back to where it started in 1967, reversing the gains of the 1970s. There is some evidence that median wages started to climb again in the final two years, but given the volatility shown throughout the series, this may simply be noise. In contrast to the flat location shift density for the middle decade shown in Figure 4.6, we can now observe the volatility in the year-to-year changes over this decade. The location entropies for 1977 and 1987 are virtually identical, however, so the net effect for the decade is zero.

In contrast to the location series, the shape series shows a steady upward rise. As with the location entropy, the shape entropy does not represent the direction of the shape change from 1967, just the magnitude. The monotonic rise in the series indicates that the direction of change was consistent throughout the period, and we know from the RD displays in Chapter 4 that the trend in polarization was positive. In this case, the net growth in polarization by decade shown in the bottom row of Figure 4.5 apparently does not mask any volatility in the year-to-year changes. By 1990, changes in shape of the earnings distribution had become the major contributer to the overall change since 1967. The downward trend observed in the last years here may be more meaningful, as there is little evidence of volatility in the preceding years of the series. So perhaps the trends observed from 1995–1997 herald the beginnings of a real change.

The second panel of Figure 6.1 shows the annual relative polarization indices. These indices track changes in the shape of the distribution only, and they code the direction as well as the magnitude of the change. The MRP index represents the overall growth or decline in inequality from 1967, and the rising trend here indicates a strong increase in inequality that becomes significant by the early 1970s. The lower and upper indices represent the portion of the median index that is generated by polarization in earnings below and above the median respectively. For most of the period, the lower index is the larger of the two, indicating that downgrading in earnings was more pronounced than upgrading. But it is worth looking at the trends as well as the levels in these two indices. Polarization in the upper tail of the distribution displays no real trend before 1980, while polarization in the lower tail rises steeply. During the 1980s, the *growth* in each index is nearly the same, even though the level is still higher for the lower index given its

rise in the decade before. By the late 1980s, however, polarization in the lower tail begins to stabilize, while polarization in the upper tail begins to rise more steeply. The net result by the early 1990s is a distribution with nearly perfectly symmetric polarization in each tail. Because this period was also characterized by downshifting in the median wage, the net effect of the two trends – growing polarization in the upper tail and a falling median wage – is that the earners at the top of the distribution experienced no real gain, they just held on to their position while everyone else lost ground.

In sum, nearly all of the earnings gains made by white men during the 1970s had been erased by the 1990s. The only consistent trend over this time was the growth in inequality which, while often less noticeable in any single period when compared to the large swings in median earnings, quietly became the major component of change in the earnings distribution. Behind these net shifts, however, was a volatile set of competing trends. Growing inequality during the 1970s was largely driven by losses experienced among workers earning below the median, but the plight of these workers was mostly alleviated by strong gains in median real wages during this period. During the 1980s stagnation in median earnings, coupled with growing polarization both above and below the median, hit the least advantaged workers the hardest. By the 1990s, the falling median earnings and growing polarization in the upper tail of the earnings distribution combined to hit all but the most advantaged workers. They experienced no gains, but were at least protected from the losses. There is some evidence that this trend may be changing in the final two years, 1995–1997.

We now turn to the other demographic groups to see how their distributions changed over this period. Just for calibration, we start by plotting the median earnings ratios for each group in Figure 6.2. The lines represent $Q_t(\frac{1}{2})/Q_{67}(\frac{1}{2})$ within each group over time. They can therefore be used to compare the relative progress of each group. Note that the groups did not start at the same level. The common starting value of 1.0 reflects the fact that each group is being compared to their own initial value in 1967. Bearing this in mind, the trends in the figure are quite interesting. White men are at the bottom of the pile, at least in terms of median wage gains. They start out nearly 40% higher than the other groups, so their advantage is not eliminated by the end of the period, but their relative advantage has clearly eroded. The trends for black men are fairly similar, though their period of wage growth during the 1970s produces relatively more gains and lasts somewhat longer. Still, both groups of men face stagnant wages from about 1980 on. The two groups of women, by contrast, both continue to experience robust wage growth over most of the years of the series. The gains for black women are particularly striking, though this is largely because their initial wage deficit was so severe.

In Figure 6.3, we turn to the decade-specific overall relative densities for each group (compare to Figure 4.5 for white men). The 1970s display a

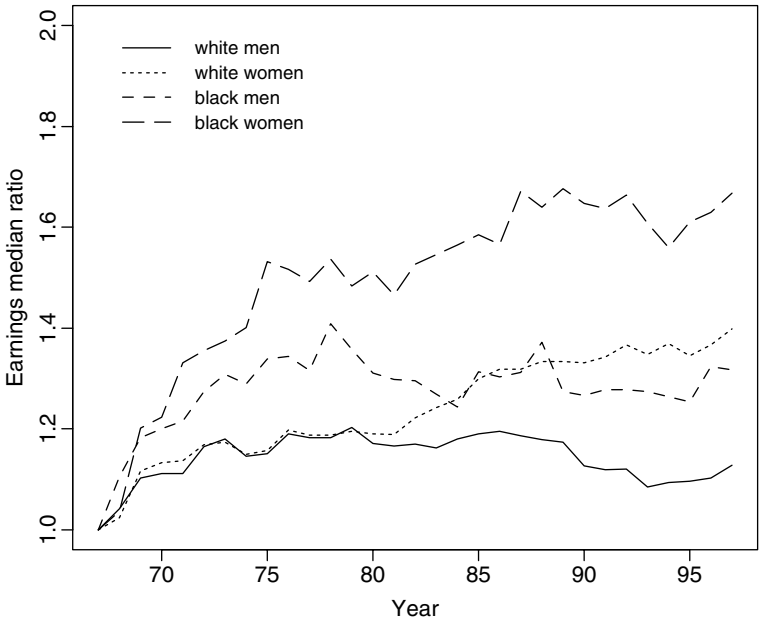

Fig. 6.2. Median earnings ratio series by group.

trend for all groups similar to that seen among white men, dominated by upshifting in median earnings. While the general trend is similar across the groups, there are also some interesting differences to note. For white women, the upward shift is fairly monotonic, net of a little sawtoothing across the deciles. For black men, the extremes are where we find the biggest differences. About 80% of those in the bottom decile of the earnings distribution had climbed above it by the end of the decade, and the fraction whose earnings would have put them in the original top decile more than doubled. For black women, earnings in the bottom original decile have virtually disappeared by the end of the decade. Their RD curve is steep and monotonic, suggesting that this group made the largest gains of all – at least relative to their original position. Whether black women's gains are larger or smaller than the gains made by other groups is not possible to gauge from this set of RDs. For that, it would be necessary to construct the between-group RD.

In the 1980s there is again an overall similarity with the pattern observed for white men: the strong gains in median earnings have disappeared and polarization is now quite visible. But some differences from the white male pattern can also be observed, as both groups of women continued to experience some limited upshifting during this decade. None of the three groups experienced the downshifting we saw among white men in the 1990s.

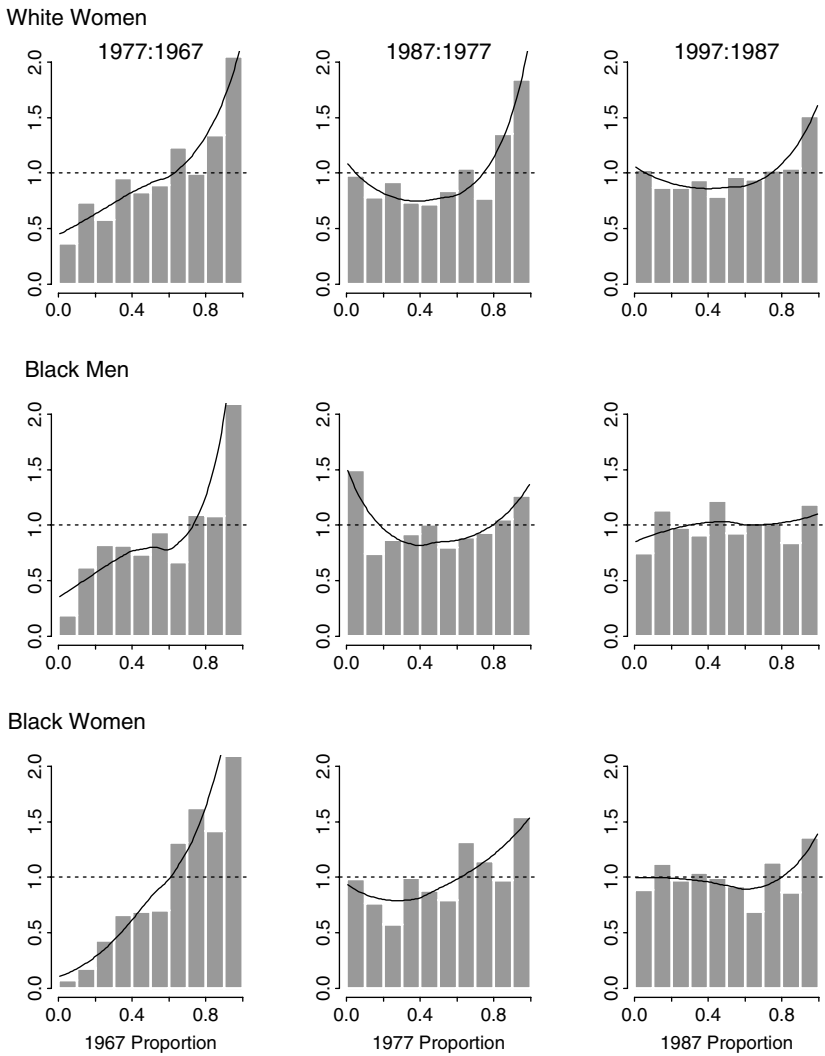

Fig. 6.3. Decade-specific overall relative densities by group.

Pulling out the impact of location and shape shifts in these distributions from a visual inspection of the overall RDs is somewhat tricky. Recall how completely the location shift masked the polarization of white men's earnings in the 1970s. In that case, we examined the RD decomposition graphs to help understand the contribution of each component (see Figure 4.6). A similar approach here would require 27 plots. We could fit them on three pages, so this is not an impossible display, but we will take this opportunity to see if the summary statistics provide a the necessary information to tease apart the location and shape changes in a more succinct fashion. The reader can judge in this case whether the summaries provide enough information to tell the story.

The location and shape entropy summaries for each group are plotted in Figure 6.4. As we might have expected, the location entropies are dramatically larger for all of these groups than they were for white men, and they remain the dominant component of the overall change throughout the period. While the general magnitude of the location shifts is comparable for these groups, the timing and sequence differs. For white women, there are two periods of growth: 1967–1975 and 1983–1990. The latter is the stronger one, a picture that doesn't quite square with the image obtained from the decade-specific overall RDs in Figure 6.3. There, the 1980s appeared to be a period of limited upshifting, at least relative to the 1970s. The difference between these two images is largely due to the role played by the shape shift during this period. This is partly observable in the shape entropy series in Figure 6.4, but will become much clearer in the polarization indices graphed in Figure 6.5. What can be seen in Figure 6.4 is a steadily rising shape entropy series, much like that seen for white men. Here, as there, this is indicative of growing inequality within the group. At the very end of the series, it would appear that another burst of wage growth has begun.

For black men, there are two distinct patterns: a period of strong gains in median wages from 1967–1974, and a period of volatile, but effectively stagnant, earnings after that time. This is consistent with the picture established by the overall RD graphs, which suggest that the 1970s was the strongest period of earnings growth. As the location shift stabilizes in the early 1980s, the shape shift begins.

For black women, the early 1970s are also a period of strong median gains, but the growth slows during the late 1970s, picks up again during the early 1980s and then stagnates during the 1990s. The shape changes in the black women's earnings distribution are larger than those seen in any other group, but it is not clear from this figure whether they point to growing or declining inequality. What can be seen is that the trend is largely monotonic until about 1980, and then begins to reverse direction. As with white women, however, the final years of the series again suggest strong wage growth, and stable polarization.

The polarization indices for each group are plotted in Figure 6.5. The differences among the groups are quite dramatic. For the first time, we see

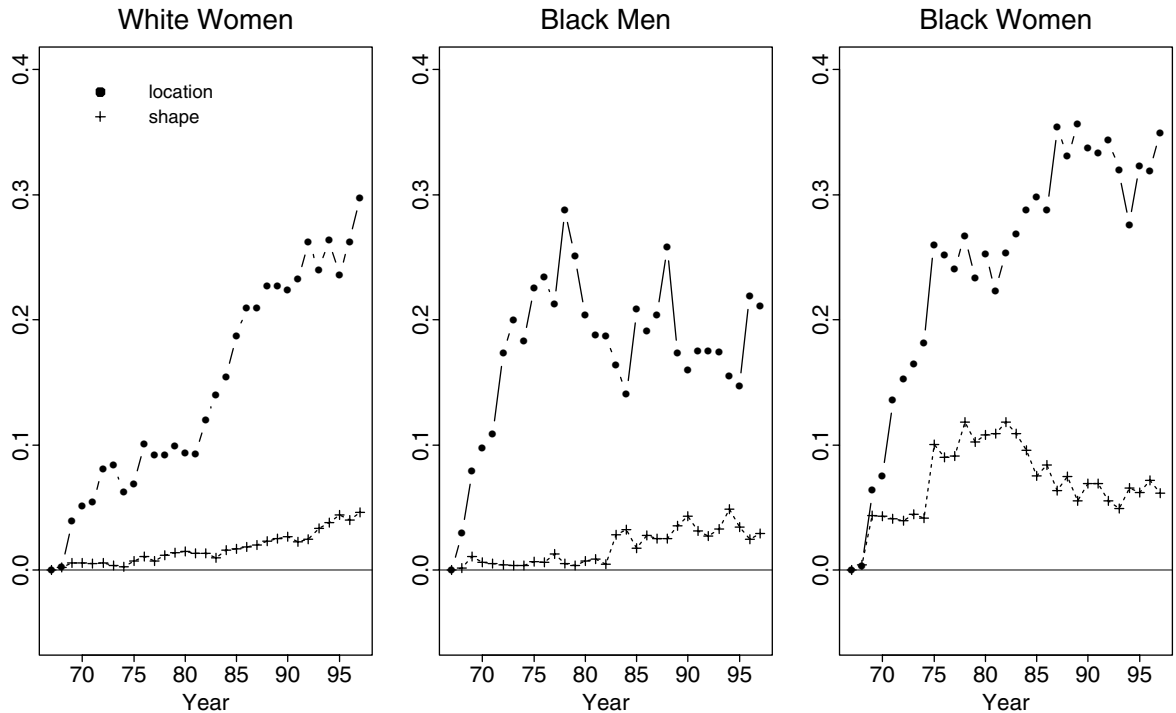

Fig. 6.4. Entropy summaries for location and shape changes in earnings: 1967–1997.

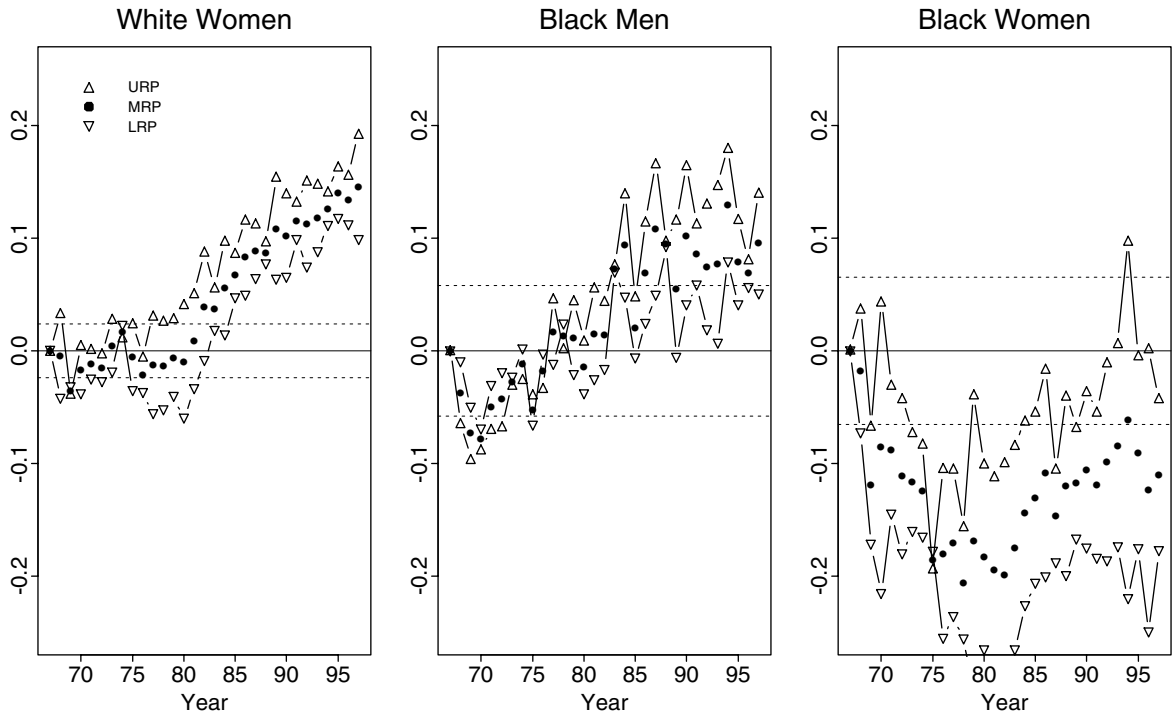

Fig. 6.5. Polarization indices by race/sex group.

the indices become negative, indicating that for some periods the earnings distribution for each of these groups is actually becoming more equal than it was in 1967 – there is convergence, rather than polarization in the relative earnings.

The story for white women is perhaps the simplest: a short period of equalization, followed by a steady growth in inequality after 1980. In the late 1970s, the LRP becomes significantly negative, indicating that the lower tail of the earnings distribution is growing closer to the median. The upper tail is stable during this period. In the early years of the 1980s, the lower tail quickly repolarizes, and from the mid to late 1980s both the lower and upper index are growing at the same rate. By the early 1990s polarization stabilizes, but in the final years the upper index rises again while the lower index falls slightly.

For black men, the story is somewhat similar. The convergence in the 1970s is apparent in both tails, however, and the lower tail begins to polarize by the mid 1970s, earlier than for white women. From 1980 on, there is steady growth in the URP, indicating strong polarization in the upper tail of the distribution. The LRP, by contrast, stabilizes around 1985. This is quite different than the trend we observed for white men, where the lower tail polarization began early, and led from 1980 on.

For black women, whose shape entropy was the strongest of all the groups, the pattern is dramatically different. Here, all of the indices show negative polarization for most of the observation period. The steepest drops are in the first decade and are led by contractions in the lower tail of the distribution. This is largely the effect of black women's movement out of domestic labor. At the turn of the first decade, the level is fairly stable for all three indices, though the lower tail contracts and then repolarizes. By the mid 1980s, however, polarization begins to rise, with equal contributions in *trend* from both tails. By the 90s, the lower tail has stabilized, while the upper tail continues to polarize, indicating that those at the bottom of the distribution are keeping pace with the middle, while those at the top are drawing away. If we had begun the series in 1980, the MRP would have become significant by about 1985.

In summary, all three groups experienced greater earnings gains during this period than did white men, particularly the two groups of women. The strongest period of wage growth came in the 1970s, and in contrast to white men, some of the largest gains were made by those at the bottom of the distribution. This was more than a rising tide lifting all boats; it was a period in which some of the severest earnings inequities were reduced, particularly for black men and women. The 1980s continued to be a period of strong earnings growth for both groups of women, but during this decade inequality also began to rise. This meant that those near the bottom made no progress, but they also did not lose their earlier gains. For black men, however, median earnings stagnated and inequality grew, so those at the bottom actually lost some of the gains from the previous decade. By the

1990s all of the groups faced sluggish wage growth and rising inequality. But faster growth in the upper tail than the lower meant that this was still mostly a "good news" story.

6.4 Discussion

Direct comparisons between the race and sex groups would be the logical next step, as we have now established how the earnings distributions have shifted within these groups. For example, using white men's earnings to define the original reference distribution in each year, one could form the relative distribution using women's or black workers' earnings. This would provide information on the dispersion of other worker's earnings relative to white men's for a particular year, as well as over time. An example can be found in Bernhardt, *et al* (1995). Alternatively, the distribution of the entire workforce could be taken as the starting reference point. This would be the "pooled comparison density" defined by Parzen (1992) and discussed in Section 2.4.1. The picture one gains is then of different types of workers moving around the overall distribution of earnings. Either approach would provide distributional comparisons that convey substantially more information than the usual differences in average wages between blacks and whites or men and women.

Taking a more explanatory tack, the baseline and relative distributions could be defined in terms of industries. This would provide a direct method of investigating how the shift from manufacturing to service industries, and the reorganization of work and production within industries, has affected earnings inequality within and between groups. This approach could be used to look at some of key controversies in the restructuring debate, such as whether service industries have generated more high-wage jobs or "McJobs", whether the trends in earnings inequality vary across sectors and over time, and whether the impact of restructuring has varied across groups. Relative distribution methods offer a more sensitive method for identifying which sections of the earnings distribution are changing, and as such, provide an indirect test of the underlying causal processes and a guide for future research.

While the tools provided by relative distribution methods are more sensitive and informative than other current methods for analyzing sample survey data, these tools can not replace qualitative studies. In the example above, qualitative work will ultimately be needed to establish the causal importance of industrial restructuring for earnings inequality. We need a better understanding of how firms have changed their production processes and therefore the skill levels they require. We need to identify the consequences of these changes for wage hierarchies and segmentation. And we need to know if the survival of firms in the postindustrial economy is dependent upon the continued creation of low-wage jobs. In answering these

questions, one must directly confront the problem of independently measuring the skill content of jobs (demand) and workers' skills (supply). Neither of these tasks is trivial.

Exercises

Exercise 6.1. Verify the median earnings ratios by race/sex given in Figure 6.2. Plot the ratio of the median earnings ratio of each group to the white men's ratio. Describe the patterns you see.

Exercise 6.2. Calculate the Gini indices for earnings in 1967, 1977, 1987 and 1997 for for each race/sex group. Summarize the image of inequality provided by the differences in Gini indices across time. How does it compare to the conclusions given in the chapter?

Exercise 6.3. Construct the relative distribution of white women's to white men's wages in 1967, 1977 1987 and 1997. Describe the changes you observe in the relative density graph. Apply the location/shape decomposition to 1967 and 1997, and calculate the entropy and polarization statistics for each year. Summarize your findings.

Exercise 6.4. Continuing on with Exercise 6.3, calculate and plot the yearly entropy summaries and polarization indices for the 1967–1997 period. Summarize the changes in the relative distribution of women's to men's wages over the period. Are there any aspects of the change that are more visible in the full index series than in the decade comparisons of the previous exercise?

Exercise 6.5. Repeat Exercises 6.3 and 6.4, using black men as the comparison population.

Exercise 6.6. Repeat Exercises 6.3 and 6.4, using white women as the reference population, and black women as the comparison population.

Exercise 6.7. Repeat Exercises 6.3 and 6.4, using the pooled population of white men and white women as the reference population. How does this change the interpretation of the plots and summary statistics? Does it change your understanding of the kinds of changes that have taken place in the earnings distributions for these two groups? Is the pooled comparison easier or more difficult to understand than the direct contrast in Exercise 6.3? reference population, and black women as the comparison population.

This page intentionally left blank

Chapter 7

Adjustment for Covariates

To this point we have focused on comparing the distributions of a single variable between two populations. Often there are covariates measured on the individuals which vary systematically by population, and the impact of these covariates is of interest. In the regression setting, it is natural to explore how the outcome for an individual depends on these covariates. In the relative distribution setting, there is an added dimension because distributional impacts are of interest, and these can take two forms. The first is a compositional shift in the covariates from one population to the other. For example, we might be interested in comparing the distribution of earnings for the population of workers in 1967 to that in 1987. We know that the sex composition of the working population changed over this period, and we want to quantify the impact of this change on the distribution of earnings. The second kind of effect is a change in the relationship between the covariate and the response variable. From our example, we also know that the *conditional* distributions of earnings by sex changed over this period (from Chapter 6), so that even if the sex composition of the working population had been stable, the overall earnings distribution would have changed.

If only the covariate composition changed, or only the covariate-response relation changed, then the source of the changes in the relative distribution could be immediately identified, but this rarely happens in practice. As with location and shape shifts, both types of covariate effects often change simultaneously, and we need a way to separate out the effects. The approach developed below is similar in principle to the location and shape decomposition, in that it constructs a counter-factual distribution to represent each shift in isolation. By adjusting the reference population to have the same covariate composition as the comparison population, we can answer questions like: "How would the earnings distribution have looked if there had been no changes in the sex composition of workers?" And we can interpret the residual differences in the relative distribution in terms of a change in the covariate-response relationship. In addition, we can go on to apply the location and shape decomposition to each of the covariate components. This makes it possible to nest the question we asked earlier

– "How did median and shape changes combine to produce the changing earnings for women?" – into the overall analysis.

The construction of the counter-factual distribution and its use for unconditional comparison is the topic of this chapter.

7.1 Compositional adjustment

In this section, we discuss the construction of a counter-factual distribution for the response variable in the reference population that is *composition-adjusted* to have the same distribution of the covariates as comparison population. For simplicity we will first discuss the situation where we have a single covariate that is categorical. The extensions to continuous covariates is give at the end of the section and the extension to multivariate covariates is considered in Section 7.4.

The basic approach is quite intuitive, and relies on the simple rule for relating conditional and unconditional probabilities:

$$P(Y = y) = \sum_z P(Y = y|Z = z)P(Z = z),$$

where the sum is over the outcome space of Z. The two components we seek in our decomposition are essentially the two terms on the right-hand side of this equation.

Let (Y_0, Z_0) and (Y, Z) denote random vectors describing the reference and comparison populations. As before, Y_0 and Y represent the variable we wish to compare across the two populations. We will call it the *response variable* here. Z_0 and Z represent the values of the *covariate*. We assume that the supports of Z_0 and Z are both $\{1, 2, \ldots K\}$. Let $\{\pi_k^0\}_{k=1}^K$ be the probability mass function of Z_0 and $\{\pi_k\}_{k=1}^K$ be the probability mass function of Z. These probability mass functions represent the population composition with respect to the covariate. For conditional comparisons of the response we can consider the densities of Y_0 given that $Z_0 = k$:

$$f_{Y_0|Z_0}(y\ |k) \qquad k = 1, \ldots, K$$

and the densities of Y given that $Z = k$:

$$f_{Y|Z}(y\ |k) \qquad k = 1, \ldots, K.$$

These densities represent the covariate-response relationship. The marginal densities of Y_0 and Y can be written as

$$f_0(y) \equiv \sum_{k=1}^K \pi_k^0 f_{Y_0|Z_0}(y\ |k) = \mathrm{E}_{\pi^0}\left[f_{Y_0|Z_0}(y\ |Z_0)\right] \tag{7.1}$$

and

$$f(y) \equiv \sum_{k=1}^{K} \pi_k f_{Y|Z}(y\ |k) = \mathrm{E}_\pi \Big[f_{Y|Z}(y\ |Z) \Big], \tag{7.2}$$

respectively. These formulas express the overall distribution of the response as a weighted average of the distributions given the covariate where the weights are the proportions of the population with that value of the covariate. This representation makes it clear how the covariate affects the overall distribution.

Consider the situation where the conditional distributions of the response are the same for each value of the covariate, that is, $f_{Y_0|Z_0}(y\ |k) = f_{Y|Z}(y\ |k), k = 1, \ldots, K$. Thus the subgroups defined by the covariate have identical distributions of the response and the relative distributions for each group across the two populations will each be uniform. The two populations are thus equivalent given the covariate value. The marginal density of Y_0 can be written as

$$f_0(y) = \sum_{k=1}^{K} \pi_k^0 f_{Y|Z}(y\ |k).$$

Comparing this to (7.2), we can see that any differences between $f(y)$ and $f_0(y)$ are now due to π_k^0 and π_k, the compositions of the covariate in each population.

Consider now the alternative situation where the probability mass function of the covariate is the same in each population, that is, $\pi_k^0 = \pi_k, k = 1, \ldots, K$. Then the marginal density of Y_0 can be written as

$$f_0(y) = \sum_{k=1}^{K} \pi_k f_{Y_0|Z_0}(y\ |k).$$

Now any differences between $f(y)$ and $f_0(y)$ in (7.2) are a result of the differences in the conditional densities $f_{Y_0|Z_0}(y\ |k)$ and $f_{Y|Z}(y\ |k), k = 1, \ldots, K$. These represent differences in the covariate-response relationship between the two populations.

We can construct a counter-factual distribution for the compositional difference using these ideas. We define the distribution of Y_0 *composition-adjusted* to Y to be:

$$f_{0C}(y) \equiv \sum_{k=1}^{K} \pi_k f_{Y_0|Z_0}(y\ |k) \tag{7.3}$$

Comparing (7.3) to (7.1) and (7.2) we see that the density $f_{0C}(y)$ corresponds to a counter-factual population with the covariate composition of the comparison population and the covariate-response relationship of the reference population. Comparisons of $f_{0C}(y)$ to $f(y)$ hold the population composition constant, and therefore isolate differences in the covariate-response relationship. By contrast, $f_0(y)$ and $f_{0C}(y)$ have the same

covariate-response relationship and comparisons between them isolate the impact of the compositional shifts.

Note that we could instead define f_{0C} by adjusting the comparison population to have the same marginal covariate composition as the reference population. The choice is unlikely to affect how we interpret the basic trends, but it does have some subtle implications for interpretation when we get to the stage of constructing the relative distribution components. We will return to this issue then.

The extension to continuous covariates is straightforward. Suppose that the covariate Z is continuous with density $f_Z(z)$ $z \in \mathbb{R}$. The composition-adjusted f_{0C} can be defined similarly to (7.3):

$$f_{0C}(y) \equiv \mathrm{E}_{f_Z}\left[f_{Y_0|Z_0}(y\ |Z)\right] = \int f_Z(z) f_{Y_0|Z_0}(y\ |z) dz. \tag{7.4}$$

The composition-adjusted distribution has the same interpretation as in the discrete case. In both situations we will denote the CDF corresponding to f_{0C} by F_{0C} and use Y_{0C} to denote a random variable randomly sampled from F_{0C}.

The composition-adjusted f_{0C} can be reexpressed as:

$$f_{0C}(y) = \int_0^1 g_Z(r) f_{Y_0|Z_0}\big(y\ |Q_{Z_0}(r)\big) dr, \tag{7.5}$$

where $g_Z(r)$ is the relative PDF of Z to Z_0 and $Q_{Z_0}(r)$ is the quantile function of Z_0. Recalling that

$$f_0(y) = \int_0^1 f_{Y_0|Z_0}\big(y\ |Q_{Z_0}(r)\big) dr,$$

this formulation makes it clear that the composition-adjusted distribution is a weighted version of the original distribution where the weighting is precisely the relative density of the covariate. A similar formulation exists when the covariate is discrete (Exercise 7.6).

Equation (7.5) suggests how a sample from the composition-adjusted distribution can be manufactured in practice. The values from the reference sample can be reweighted based directly on the relative distribution of Z to Z_0 to produce a synthetic sample with the correct properties (see Exercise 7.7).

7.2 Comparison of composition-adjusted distributions

Using the composition-adjusted response distribution, we can decompose the overall relative distribution into a component that represents the effect of changes in the marginal distribution of the covariate (the composition effect), and a component that represents the residual changes. The

method is similar to that used in Chapter 3: relative distributions of the composition-adjusted to the reference and comparison populations isolate the composition and residual effects respectively.

From the three distributions – Y_0, Y_{0C}, and Y – we can construct two relative distributions that represent the effect of the covariate composition and effect of residual changes. To isolate the composition effect, we take the relative distribution of Y_{0C} to Y_0, denoted $R_0^{0C} = F_{0C}(Y)$. R_0^{0C} will have a uniform distribution when the comparison and reference populations have the same marginal covariate distribution. To isolate the residual effect we take the relative distribution of Y to Y_{0C}, denoted $R_{0C} = F(Y_{0C})$. R_{0C} will have a uniform distribution when the conditional response distributions are the same in both populations.

The decomposition can be represented in terms of the density ratios:

$$\frac{f(y_r)}{f_0(y_r)} = \frac{f_{0C}(y_r)}{f_0(y_r)} \times \frac{f(y_r)}{f_{0C}(y_r)} \tag{7.6}$$

or, in more heuristic terms:

$$\begin{array}{c}\text{overall relative}\\ \text{density}\end{array} = \begin{array}{c}\text{density ratio for}\\ \text{the compositional effect}\end{array} \times \begin{array}{c}\text{density ratio for}\\ \text{the residual effect}\end{array} \tag{7.7}$$

These two effects form a decomposition of the relative distribution of Y to Y_0 in the same sense as the median/shape decomposition from Chapter 3. If R_0 is the relative distribution of Y to Y_0, then R_{0C} can be defined as the relative distribution of R_0^{0C} to R_0.

As before, the first density ratio is a proper density, while the second in general is not due to the scale change (see the discussion in Chapter 3 for clarification). The graphical display of the three relative densities, which we will denote by g_0, g_0^{0C}, and g_{0C}, respectively, provide a useful visual summary of the relative size and nature of the components.

It is interesting to consider how the interpretation of these relative densities would change if we define f_{0C} by adjusting the comparison population back to have the same marginal covariate composition as the reference population. For example, take the case when the the comparison population is formed by observing the reference population at a later time point. Under the definition in (7.3), the composition effect in (7.6) is obtained by taking the old distribution and moving the population composition forward in time, while the alternative definition obtains it by taking the new distribution and moving the composition backward in time. With respect to the covariate-response relationship, the definition in (7.3) identifies this by comparing the conditional response using the new population composition, while the alternative compares it using the old composition. Thus, if we really want to use the language, "if only this component had changed" from the reference to the comparison distribution, we can do so to refer to the composition effect if we use (7.3), as the residual effect reflects only additional changes in the response *given* that the composition changed. By the

same token, if we use the alternative definition, the residual effect represents what would have happened if only the conditional response distribution had changed, and the composition effect reflects any additional changes given that the conditional response distribution changed.

This suggests an alternative approach to the decomposition that would define the composition and residual effects as shifts from the reference distribution, and a third effect that represents the interaction between them. This will be left as an exercise for the reader.

While the decomposition above identifies the net effect of a change in covariate composition, holding all other factors constant, this may not always be the appropriate substantive choice. For example, an increase in the number of women in the workforce might affect the conditional distribution of earnings for men. This could happen if women were perfect but less expensive substitutes for men, and their increased supply drove down the wages employers were willing to pay. In this case, the "composition effect" as identified here would underestimate the true effect of the increasing share of women in the workforce.

7.3 Further decomposition by location/shape

Once the relative density has been composition-adjusted, one can examine both the composition and residual components for location and shape changes. As the residual component represents the changes in the conditional response distribution, the location shift in this component captures traditional changes in the "returns" to the covariate, and plays a role similar to the change in a regression coefficient for that covariate. The shape shift in this component captures additional changes in the covariate-response relationship that are often hidden when other methods are used.

Continuing with our hypothetical example, suppose the sex composition effect on the change in overall earnings is found to be large. A location/shape analysis might then go on to show that the composition effect was primarily a location shift – with the rising proportion of women earners dragging the median earnings down – while the residual was both location-shifted and more polarized. The location shift in this context would represent changes in the gender wage gap – the covariate-response relationship. Polarization would suggest that once changes in the median wage gap were netted out, there was a U-shaped relative distribution of women's to men's earnings, indicating that women's earnings were polarizing more rapidly than men's. This kind of analysis can provide a rich description of the interrelated distributional changes, which in turn can help to inform and focus a theoretical debate by clearly identifying what needs to be explained. Combining composition adjustment with location/shape decomposition is a straight forward sequential application of the preceding techniques.

7.4 Adjusting for multiple covariates

The construction of the counter-factual distribution for a single covariate was considered in Section 7.1. In this section we extend the approach to multiple covariates, both continuous and discrete. The principle is similar, and relies on decomposing the chain of conditional and marginal probabilities. Now, however, the dependence among the covariates will make it necessary to address the question of how to decompose the total difference into unique effects for each covariate. As in the regression setting, there is more than one way to do this.

Let Z_0 be the multivariate values of the covariates for the reference population and Z be the corresponding values for the comparison population. Let (Y_0, Z_0) and (Y, Z) denote random vectors describing the reference and comparison populations. As before, Y_0 and Y are the attributes we wish to compare. Let $f_Z(z)$ and $f_{Z_0}(z)$ be the joint densities of Z_0 and Z, respectively. If some of the components are discrete then the corresponding components of the densities are probability mass function. These distributions represent the population composition with respect to the covariates. For conditional comparisons of the response we can again consider the densities of Y_0 given that $Z_0 = z$:

$$f_{Y_0}(y \mid Z_0 = z)$$

and the densities of Y given that $Z = z$:

$$f_{Y|Z}(y \mid Z_0 = z).$$

These (univariate) densities represent the covariate-response relationship. These distributions have the same roles as their single covariate counterparts do in Section 7.1

We may still define the distribution of Y_0 *composition-adjusted* to Y to be the expected covariate-response relationship over the comparison covariate distribution:

$$f_{0C}(y) \equiv \mathrm{E}_{f_Z}\left[f_{Y_0|Z_0}(y \mid Z)\right].$$

If all the components of the covariate are continuous, then the expectation is a multivariate integral similar to (7.4). If some of the components of the covariate are discrete then the corresponding components of the expectation are sums similar to (7.3). The composition-adjusted distribution is, of course, continuous and univariate. It can be interpreted and used in the same way as in the single covariate case.

As an example, consider the situation where we have two continuous covariates Z^1 and Z^2. Let $Z_0 = (Z_0^1, Z_0^2)$ and $Z = (Z^1, Z^2)$. The joint effect of Z^1 and Z^2 is:

$$
\begin{aligned}
f_{0C}^{12}(y) &\equiv \mathrm{E}_{f_Z}\left[f_{Y_0|Z_0}(y\mid Z)\right] \\
&= \int\int f_Z(z^1,z^2) f_{Y_0|Z_0}(y\mid z^1,z^2)\; dz^1 dz^2 \qquad (7.8)\\
&= \int\int f_{Z^2|Z^1}(z^2|z^1) f_{Z^1}(z^1) f_{Y_0|Z_0}(y\mid z^1,z^2)\; dz^1 dz^2,
\end{aligned}
$$

where $f_Z(z^1,z^2)$ is the marginal distribution of Z; $f_{Z^2|Z^1}(z^2|z^1)$ is the distribution of Z^2 given Z^1; and $f_{Z^1}(z^1)$ is the marginal distribution of Z^1 all in the comparison population. For the reference population, $f_{Y_0|Z_0}(y\mid z^1,z^2)$ is the distribution of Y_0 given $Z_0=(z^1,z^2)$.

Sequential Adjustment

The composition-adjusted distribution adjusts the response jointly for the differences in all the covariates. The individual compositional effects of each variable can be investigated by analyzing the composition-adjusted distribution for that variable alone. It may be of interest to determine the contribution of each variable to the joint composition effect. Conceptually this is analogous to the situation in multiple linear regression where we wish to determine the effect of each predictor variable on the target variable. There, and here, there will not necessarily be a unique decomposition of the joint effect, unless one is willing to specify the order in which the covariates are applied. It is possible to define sequential effects that uniquely decompose the sum, but, in general, the order of the sequence will matter.

For the above two continuous covariate situation, the distribution of Y_0 Z^1-composition-adjusted to Y is

$$
\begin{aligned}
f_{0C}^{1}(y) &\equiv \mathrm{E}_{f_{Z^1}}\left[f_{Y_0|Z_0^1}(y\mid Z^1)\right] \\
&= \int f_{Z^1}(z^1) f_{Y_0|Z_1^0}(y\mid z^1)\; dz^1 \qquad (7.9)\\
&= \int\int f_{Z_0^2|Z_0^1}(z^2|z^1) f_{Z^1}(z^1) f_{Y_0|Z_0}(y\mid z^1,z^2)\; dz^1 dz^2,
\end{aligned}
$$

where $f_{Z_0^2|Z_0^1}(z^2|z^1)$ is the distribution of Z_0^2 given Z_0^1 in the reference population.

Suppose we wish to determine the compositional effect of Z^2 after the compositional effect of Z^1 has been taken into account. The compositional effect of Z^2 will not be Y_0 Z^2-composition-adjusted to Y due to the dependencies between the two covariates. Comparing this expression for Y_0 Z^1-composition-adjusted to Y to (7.8) we can see that the effect of the composition adjustment for the second covariate, above and beyond the first, is to replace $f_{Z_0^2|Z_0^1}(z^2|z^1)$ with $f_{Z^2|Z^1}(z^2|z^1)$. If the conditional distributions of the second covariate given the first are the same in both

populations, the second variable will have no compositional effect in addition to the first. This will hold even if the marginal distributions of the second covariate differ between the populations.

The decomposition in (7.2) can be extended to this situation. Let Y_{0C}^{12} be a random variable with density f_{0C}^{12} in (7.8) and CDF F_{0C}^{12}, while Y_{0C}^{1} be a random variable with density f_{0C}^{1} in (7.9) and CDF F_{0C}^{1}. The compositional effect of Z^1 can be represented by the relative distribution of Y_{0C}^{1} to Y_0, denoted $R_0^1 = F_0(Y_{0C}^1)$. R_0^1 will have a uniform distribution when the comparison and reference populations have the same marginal distribution of Z^1. The compositional effect of Z^2 in addition to Z^1 can be represented by the relative distribution of Y_{0C}^{12} to Y_{0C}^{1}, denoted $R_1^{12} = F_{0C}^1(Y_{0C}^{12})$. R_1^{12} will have a uniform distribution when the comparison and reference populations have the same conditional distribution of Z^2 given Z^1. Finally, to isolate the residual change we take the relative distribution of Y to Y_{0C}^{12}, denoted $R_{12} = F_{0C}^{12}(Y)$. R_{12} will have a uniform distribution when the conditional response distributions are the same in both populations.

Let R_0 be distributed as the relative distribution of Y to Y_0. The decomposition can be represented in terms of the density ratios:

$$\begin{aligned} \frac{f(y_r)}{f_0(y_r)} &= \frac{f_{0C}^1(y_r)}{f_0(y_r)} \times \frac{f(y_r)}{f_{0C}^1(y_r)} \\ &= \frac{f_{0C}^1(y_r)}{f_0(y_r)} \times \frac{f_{0C}^{12}(y_r)}{f_{0C}^1(y_r)} \times \frac{f(y_r)}{f_{0C}^{12}(y_r)} \end{aligned} \tag{7.10}$$

or, in more heuristic terms:

$$\begin{aligned} \text{overall relative density} = {} & \text{density ratio for compositional effect of } Z^1 \\ & \times \text{density ratio for compositional effect of } Z^2 \text{ given } Z^1 \\ & \times \text{density ratio for the residual effect} \end{aligned} \tag{7.11}$$

These three effects form a sequential decomposition of the relative distribution of Y to Y_0 in the sense that R_{12} is the relative distribution of R_0 to R_0^{12}, and that R_1^{12} is the relative distribution of R_0^{12} to R_0^1. The first level is the joint adjustment and the second measures the additional effect of the second covariate.

The density ratio for the compositional effect of Z^1 is a proper density while the others in general are not (see the discussion in Chapter 3).

Mathematically, the relationship between the densities is:

$$g_0(r) = g_0^1(r) \times g_1^{12}(p) \times g_{12}(q) \quad \text{where} \quad p = F_{0C}^1(r) \quad q = F_{0C}^{12}(r), \quad 0 \le r \le 1,$$

where r is the percentile in the reference population for a given value of the attribute, y_r, while p and q are the percentiles in the first and joint covariate composition-adjusted population at that same value, respectively.

This sequential approach can be extended to an arbitrary number of covariates in a straightforward way. Each compositional effect term measures the additional compositional effect of the covariate in the sequence while the final term measures the residual effect. Altering the order of the variables in the sequence can be informative about their relative effects, and in particular, one can examine the "unique" effect of each covariate by placing it last in the sequence.

Block Adjustment

In many applications the covariates will form a hierarchy where variables at each level of the hierarchy are grouped together as a block. One simple example of this is where a number of the covariates are control variables and the others are covariates of primary substantive interest. We would like to interpret the effects of the second group after the compositional effects of the variables in the first set have been adjusted for. This can be achieved in the same manner as the sequential decomposition in the previous section. We again denote the complete set of covariates by $Z_0 = (Z_0^1, Z_0^2)$ and $Z = (Z^1, Z^2)$. The set of covariates is split into two subsets (Z^1 and Z^2), representing the primary and secondary level of variables. The compositional effects can be applied in the sequential manner of the previous section, with the set of variables (Z^1) being adjusted for first and the second set (Z^2) being applied subsequently. The joint and sequential expressions (7.8) and (7.9) respectively still hold, although now they are multivariate adjustments and not necessarily single variable adjustments. The decomposition formulas also hold. For example in the example of control and interest variables the heuristic decomposition is:

$$\begin{array}{c}\text{overall relative}\\ \text{density}\end{array} = \begin{array}{c}\text{density ratio for}\\ \text{compositional effect}\\ \text{of control variables}\end{array} \times \begin{array}{c}\text{density ratio for}\\ \text{compositional effect of}\\ \text{variables of interest}\\ \text{adjusted for the controls}\end{array} \times \begin{array}{c}\text{density ratio for}\\ \text{the residual effect}\end{array}$$

If there are multiple hierarchies of variables, this approach can be applied in a sequential manner.

7.5 Categorical contrasts

For categorical covariates adjustment can also proceed as described above. In this context, however, it is often of interest to compare the groups defined by the covariate directly, rather than treating the covariate as a control

variable and adjusting to eliminate its compositional effects. For example, consider dividing a sample into two groups, those with a high school degree or less and those with one or more years of college. To analyze the impact of a change in the distribution of education on the distribution of wages, one can use either the adjustment approach described above, or a contrast approach. Using the compositional adjustment approach, the impact of the educational change on the distribution of wages is isolated, but the differences in the wage distributions between the two groups are not directly observed. Using the categorical contrast approach, the two wage distributions are instead explicitly compared; forming the relative wage distribution for the two groups and using all of the methods described in previous chapters (e.g., location/shape decomposition, entropy and polarization indices, etc.). The contrast approach, however, does not explicitly identify a composition "effect". Used together, adjustment and contrast methods can provide a detailed picture of the role of a categorical covariate in distributional change.

Exercises

Exercise 7.1. Create a decomposition along the lines of (7.6) that would define the composition and residual effects as shifts from the reference distribution, and a third effect that represents the interaction between them.

Exercise 7.2. The composition adjustment given in Section 7.1 need not be the reference distribution composition-adjusted to the comparison distribution, but could be the comparison distribution composition-adjusted to the reference distribution. Describe circumstances where this would be more appropriate than the definition given in (7.3).

Exercise 7.3. The composition adjustment given in Section 7.1 need not be made relative to either the comparison or reference distributions, but could be made relative to a third standard. For example, the population distribution of the covariate could be know from census or registry data. Alternatively, the covariate distribution could be the result of a population projection. Under these circumstances both reference and comparison distributions could be composition-adjusted to the standard. Discuss the advantages and disadvantages of this choice.

Exercise 7.4. The composition-adjusted distribution can be estimated using Monte-Carlo methods. Consider resampling with replacement from the reference sample stratified by the covariate with weights proportional to the relative PDF of the *covariate* distribution. Show that this is equivalent to resampling directly from the composition-adjusted distribution given in (7.3). As the size of the resample can be made arbitrarily large, the composition-adjusted distribution can be reconstructed with any desired accuracy.

Exercise 7.5. Consider again the distribution of earnings in 1967 and 1997 analyzed in Chapter 6. While the analysis there disaggregated by race/gender groups, it is of interest to see how changes in the proportions within race/gender groups have effected the population earnings distribution. Use the resampling algorithm described in Exercise 7.3 to composition adjust the 1967 earnings distribution to the 1997 distribution. Calculate the relative distribution of of 1997 earnings to the composition adjust 1967 earnings. Calculate the relative distribution of of 1997 to 1967 earnings, and plot the two relative distribution on the same graph. Did population composition changes have a big effect?

Exercise 7.6. This exercise uses the definition of a discrete relative distribution given in Chapter 11. As in Section 7.1, let (Y_0, Z_0) and (Y, Z) denote random vectors describing the reference and comparison populations. Here Y_0 and Y represent the response variables, and Z_0 and Z represent the values of a discrete covariate. Derive an expression for the distribution of Y_0 composition-adjusted to Y involving the discrete relative density of Z to Z_0.

Exercise 7.7. Let $Y_{01}, Y_{02}, \ldots, Y_{0n}$ be a sample from the reference population, with sampling weights $w_{01}, w_{02}, \ldots, w_{0n}$. Let $g_Z(r)$ be the relative density of a covariate in the comparison population to the reference population. Consider the new weights $v_{0i} = w_{0i} g_Z(Q_{Z_0}(Y_{0i}))$ where $Q_{Z_0}(r)$ is the quantile function of Z_0. Show that $Y_{01}, Y_{02}, \ldots, Y_{0n}$ with weights $v_{01}, v_{02}, \ldots, v_{0n}$ is a sample from the reference population composition-adjusted to the comparison population for the covariate.

Chapter 8

Application: Comparing Wage Mobility in Two Eras

8.1 Background

Much research has been done on the trend of growing wage inequality in the United States (for reviews see Karoly (1993), Levy and Murnane (1992), and Danziger and Gottschalk (1993; 1996). One of the important questions to emerge in this research concerns the issue of lifetime wage mobility: To what extent does the observed cross-sectional growth in wage inequality translate into growing polarization in the distribution of lifetime wage trajectories? If workers' lifetime wage mobility is high, then the cross-sectional trends are less worrisome Some have argued, for example, that there is simply more volatility in wages now, perhaps due to more frequent job changes, but that the long-term trajectories of workers' wages are no more polarized than before (Gottschalk and Moffit 1994; Stevens 1996). On the other hand, there is evidence that restructuring strategies at the firm level are dismantling internal labor markets and may be permanently changing the distribution of economic opportunities (Cappelli 1995; Harrison 1994). When firms replace on-the-job training and promotion with external hiring, or substitute temporary workers at the bottom of the job hierarchy, the traditional routes to career mobility are disrupted, especially for low-skill workers. If these are the forces driving the cross-sectional wage polarization, wage trajectories may become more polarized, with some workers increasingly stuck in a series of low wage marginalized jobs, while others experience "winner take all" wage gains. This scenario suggests permanent changes in the distribution of mobility and the emergence of a more rigidly segmented labor market.

8.2 Data

The longitudinal panels of the National Longitudinal Survey (NLS) data provide an opportunity to investigate these questions by comparing the wage growth profiles over time. The data we use here come from two cohorts

of the NLS, one initiated in 1966, the other in 1979. We will refer to these as the original and recent cohorts respectively. Both cohorts are 14–21 years old at the start of the survey and are followed for 16 years (through 1981 for the original cohort, and 1994 for the recent). For exposition here we restrict the examples to white males.

To examine the question of wage mobility, we analyze the growth profile of "permanent wages." Wages can be thought of as having a permanent and a transitory component, where the permanent component represents a smooth underlying age-earnings profile, net of the transitory shocks caused by such things as school-to-work transitions and job changes. Permanent wages are generally estimated using a mixed effects model: the effects of age on earnings are specified with fixed effects to capture the population average and random effects to capture population profile heterogeneity. The residual from the predicted profiles is defined as the transitory wage variance (for examples, cf., Gottschalk and Moffitt (1994), Haider (1997) Bernhardt, *et al* (1999), and Baker (1997)). This specification posits a unique age-earnings profile for each person. We will be working with the distributions of estimated permanent log-wage *gains* for each cohort. The gains are defined as the difference between the (constant dollar) estimated permanent log hourly wage at the beginning and end of each respondent's age profile. We will refer to these more simply as "wage gains" below. Given the age range in the cohort and the years of observation, the wage gain is specified over an 18-year period, as respondents age from 16 to 34 years old. Further details on the model and estimation are presented in Appendix C.

8.3 Findings

Table 8.1 presents the usual summary statistics, and Figure 8.1 shows the PDF overlays (panel a) and Lorenz curves (panel b). For the Gini and Lorenz measures, we have had to code negative wage gains to 0. There are a handful of negative values in the original cohort, and a much larger number (7% of all cases) in the recent cohort. These values indicate a loss in real wages over the observation period. Neither the Gini nor the Lorenz curve can handle negative values – something that limits their usefulness in this context. Relative distribution methods do not share this limitation.

Several aspects of the relative wage growth in the two cohorts are apparent from these figures and tables: the recent cohort experienced smaller average wage gains and these gains were more variable. While the frequency of high wage gains was comparable for the two cohorts, the frequency of low wage gains was much greater for the recent cohort (this is visible primarily in the PDF overlay). The Lorenz curve for the recent cohort lies uniformly below that of the original cohort, indicating that there is more inequality in the distribution of recent wage gains.

Table 8.1. Summary statistics for the permanent wages gains in the two cohorts.

Summary Statistic	Original Cohort	Recent Cohort
Sample size	1834	2103
Mean	1.085	0.878
Standard deviation	0.483	0.618
Interquartile range	0.549	0.789
Gini index	0.236	0.364

The key theoretical questions are hinted at but not easily quantified using the standard measures here. How well is the difference captured by a simple location shift? Is there evidence of growing polarization? Are the upper and lower tails of the distribution changing in similar ways? What is the role played by covariates like education? Relative distribution methods are well suited to these questions.

Figure 8.2 shows the relative CDF and PDF for the distribution of wage growth in the two NLS cohorts. The bottom axis is labeled in population quantiles and the top axis in (rescaled) log-wage gains. The differences in wage growth experienced by the two cohorts in the relative CDF panel are easily described using the horizontal and vertical gridlines. At the median of the original cohort wage growth, $r = 0.5$, the wage gain can be read from upper axis $F_0^{-1}(0.5) = 1.1$, about \$3.00. The relative CDF at this point is $G(r) = 0.63$, which means that 63% of the recent cohort experienced lower gains than this. The median wage gain ($F^{-1}(0.5)$) for the recent cohort can be read off of the right axis at $G(r) = 0.5$, and it is roughly 0.85, about \$2.30. About 70% of the original cohort had higher gains than this. We can also see that 27% of the recent cohorts wage gains are in the bottom decile of the original cohort distribution, the divergence between the two cohorts is greatest in the second and third deciles, and the proportions for the two cohorts converge above the 90th percentile.

The second panel shows the relative PDF of the wage growth, recent to original cohort. Compared to the relative CDF, the relative PDF provides a more intuitive display for many people: values above 1 represent more density in the recent distribution; values below 1 represent less, and the actual value is the multiplicative factor more (or less). This graph makes it clear that the biggest difference between the two cohorts is at the very bottom of the distribution: nearly three times as many recent wage gains fell into the bottom decile defined by the original cohort. By contrast, the frequency of wage gains in the middle of the distribution has fallen by 30%-40% for the recent cohort. The smallest discrepancy is at the highest levels of wage growth, although the recent cohort is still somewhat less likely to achieve such gains. For example, the relative density at the 85th percentile of original cohort wage growth ($F_0^{-1}(r) = 1.5$, a wage gain of about \$4.50)

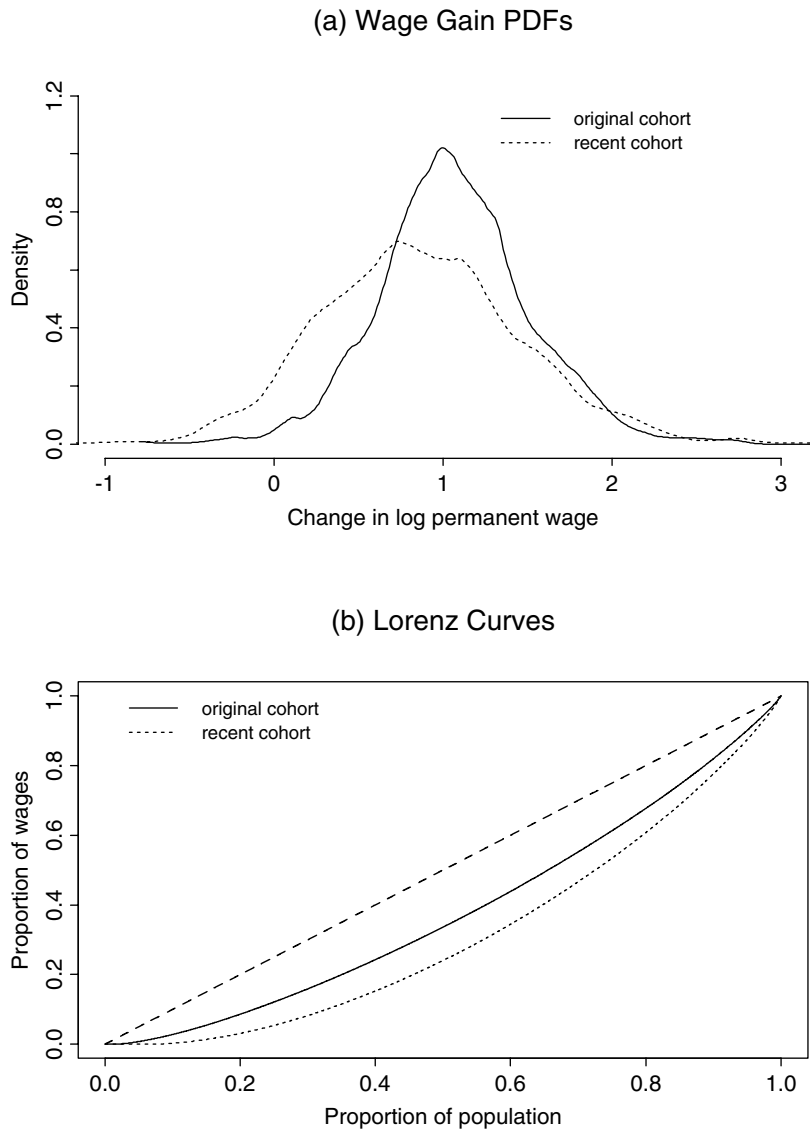

Fig. 8.1. The distributions of permanent wage growth in the original and recent NLS cohorts. (a) The PDFs for each cohort; (b) The Lorenz curves of these PDFs.

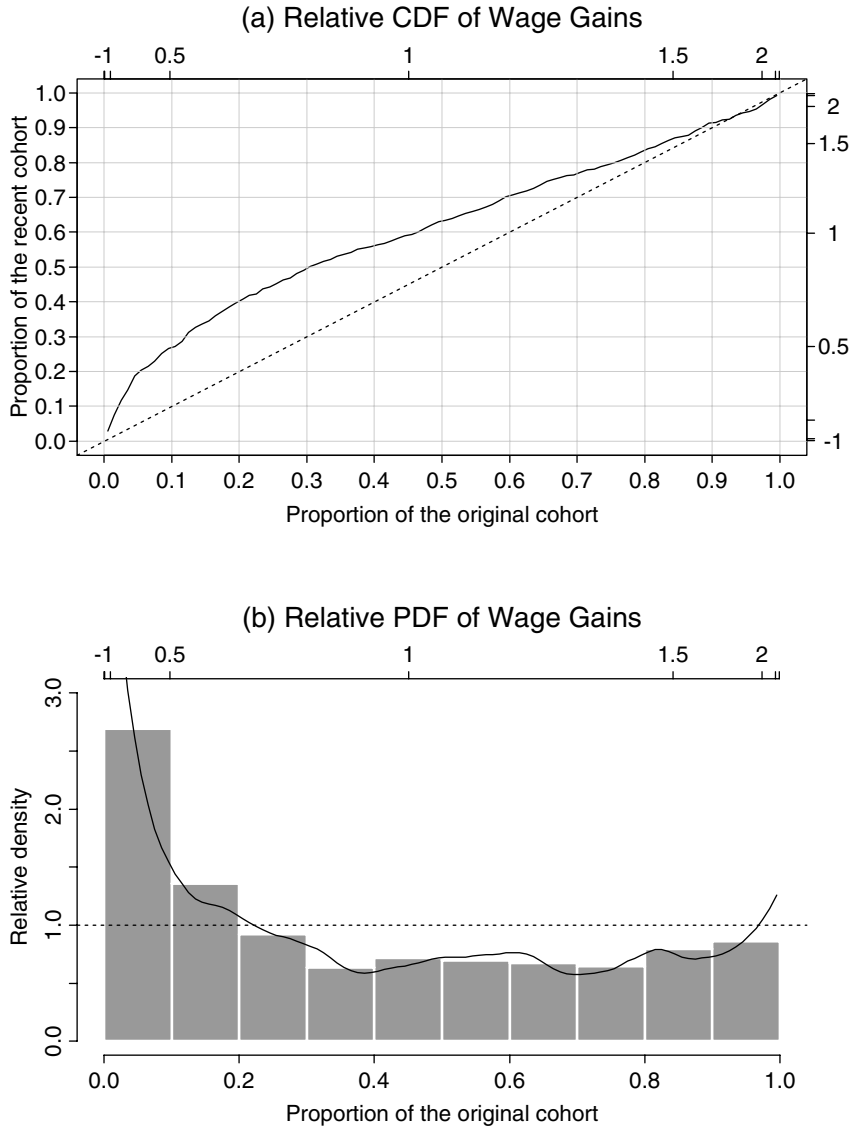

Fig. 8.2. The relative distribution of permanent wage growth in the original and recent NLS cohorts: (a) the relative CDF; (b) the relative PDF. The relative deciles are superimposed on the smooth density estimate. The upper and right axes are labeled in permanent log-wage gains.

is about 0.8. So 20% fewer recent cohort earners attained this level of wage growth.

8.2.1 Location and shape decompositions

Figure 8.3 presents the median and shape decomposition of the relative distribution of wage gains. The first panel represents the overall relative density (and is the same as Figure 8.3b). The second panel represents the effect of the median shift in the wage gains between the two cohorts – representing what the relative density would have looked like if there had been no change in distributional shape. The effects of the median shift are quite large. This alone would have placed nearly 70% of the recent wage gains in the bottom half of the original cohort distribution and virtually eliminated the gains in the top decile. Note, however, that neither tail of the observed RD is well reproduced by the median shift. The bottom decile of panel (b) is about 2.0, well below the value of 2.7 observed in the actual data, and the upper deciles are also substantially lower than observed. These differences are explained by the shape effects presented in panel (c), which are also quite large. Even without the lower median, the greater dispersion of wage gains in the recent cohort would have led to relatively more low-growth earners, and this effect is concentrated in the bottom decile. The polarization hollows out the middle of the wage gain distribution, with a cumulative loss of nearly a third of recent earners in deciles 3 through 8. At the top of the distribution, however, the growing spread in recent wage gains works in the opposite direction from the location shift: operating by itself, it would have increased the number of wage gains in the upper decile by more than 50%. In sum, the losses experienced by the recent cohort are produced by both lower median gains and polarization, while the higher wage gains are exclusively due to polarization.

The entropy summary is given on top of each figure, and the full set of summary statistics is presented in Table 8.2. The overall entropy for the change in wage growth between the two cohorts is 0.159. The location shift accounts for 65% of the total change, and the shape shift for 37%. The two numbers do not sum exactly to 100 because of the rescaling in the location-adjusted density ratio (see Section 3.1). The MRP index for the shape change displayed in panel (c) is 0.183 (95% CI 0.148–0.218). For comparison, two normal distributions with the same MRP would have a standard deviation ratio of 1.34. The size and sign of the estimate confirm the impression left by the graphical display: there has been a significant growth in permanent wage inequality between the two cohorts. The lower and upper polarization estimates indicate that both tails of the distribution are significantly positively polarized. The lower index is slightly larger, indicating greater polarization in the lower tail of the distribution than in the upper tail.

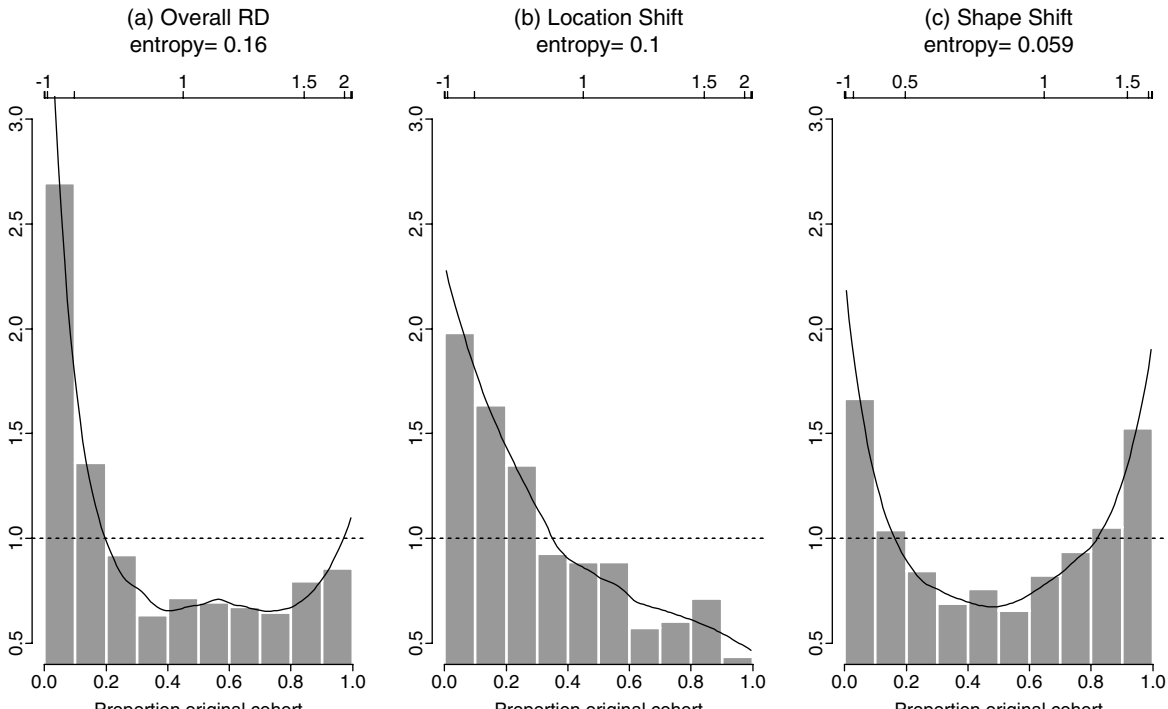

Fig. 8.3. Location/Shape decomposition of the relative distribution of permanent wage growth in the recent and original NLS cohorts. (a) the (unadjusted) relative density of wage growth; (b) the effect of the median difference in wage growth between the cohorts; (c) The median-adjusted relative density of wage growth (the effect of changes in distributional shape).

Table 8.2. Summary statistics for the location/shape decomposition of the relative distribution of wage growth: recent to original NLS cohort

Statistic	Estimate	95% CI	p–value
Entropy: overall change	0.159	0.119 – 0.198	0.000
median shift effect	0.104	0.062 – 0.146	0.000
shape shift effect	0.059	0.033 – 0.084	0.000
percent due to median	65.4%	46.6 – 84.2	0.000
percent due to shape	37.1	24.0 – 50.2	0.000
Polarization (MRP)	0.183	0.148 – 0.219	0.000
lower tail (LRP)	0.190	0.120 – 0.261	0.000
upper tail (URP)	0.176	0.105 – 0.247	0.000

The educational composition of the two NLS cohorts may have changed, and education is an influential covariate for wages. We can use the covariate adjustment technique to determine whether differences in the education profile between the two cohorts explain some of the observed changes in relative wage gains.

Figure 8.4 displays the relative distribution of final observed education in the two cohorts. Final education is measured as the number of years of schooling achieved in the last panel of the study, and is bottom-coded at 8 and top-coded at 18%. Note that this measurment scale is shown on the top axis. Neither cohort has many respondents with less than 12 years of education, so the RD is quite variable at this end of the scale. For example, the fraction of the original cohort with 10 years of education is about 2%, while the fraction in the recent cohort is about 3%, so the relative distribution is about 1.5. The fraction reporting a terminal high school degree can be seen in the first long horizontal section of the graph. Reading across the bottom axis one can see that this represents about 30% of the original cohort, while the RD value of 1.4 signifies that the relative fraction in the recent cohort is about 40% larger. There are relatively fewer respondents in the recent cohort reporting more than 12 years of schooling, with the fraction reporting a college degree is down by about 15%. Overall, the recent cohort appears to have a slightly downshifted education distribution. This may seem somewhat counterintuitive, but there are two reasons for the lower education levels in the recent cohort. The first reflects a real population trend. The rate of college attendance and completion peaked in the early 1970s, coinciding with the years of the Vietnam War draft, and raising the educational attainment of the original cohort. The second is an artifact of the sample. The original cohort had a higher rate of attrition than the recent cohort (26% and 8% respectively), and attrition was more likely to occur among the less educated.

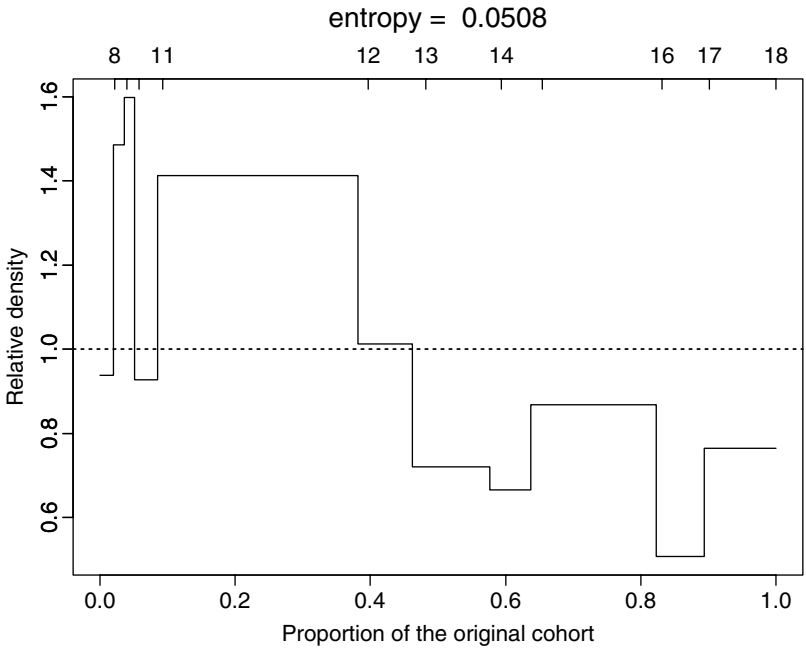

Fig. 8.4. The relative distribution of education for the recent to the original cohort. The upper axis is labeled by final education.

8.2.2 Covariate decompositions

Figure 8.5 graphically represents the adjustment of the relative distribution for education composition changes. Panel (a) shows the (unadjusted) relative density of wage gains (same as Figure 8.2b); panel (b) represents the education composition effects, and panel (c) represents the education-adjusted relative density of wage gains – that is, the expected relative density of wage gains had the education profiles of the two cohorts been identical.

Figure 8.5(b) is very close to a uniform distribution. The implication is that the difference in education composition between the two cohorts had little effect on the observed relative distribution of wage growth. The reduction in high wage gains seen in the first panel is associated with this compositional change, but the massive observed growth in the bottom decile is not. Figure 8.5(c) represents the education-adjusted relative wage gain distribution. In the absence of major compositional effects, the adjusted distribution is not much different than the original distribution. The graphical perception is confirmed by the entropy statistics: the entropy of the residual RD is 0.150, 94% of the total. Confidence intervals and p-values for the entropy statistics are given in Table 8.3.

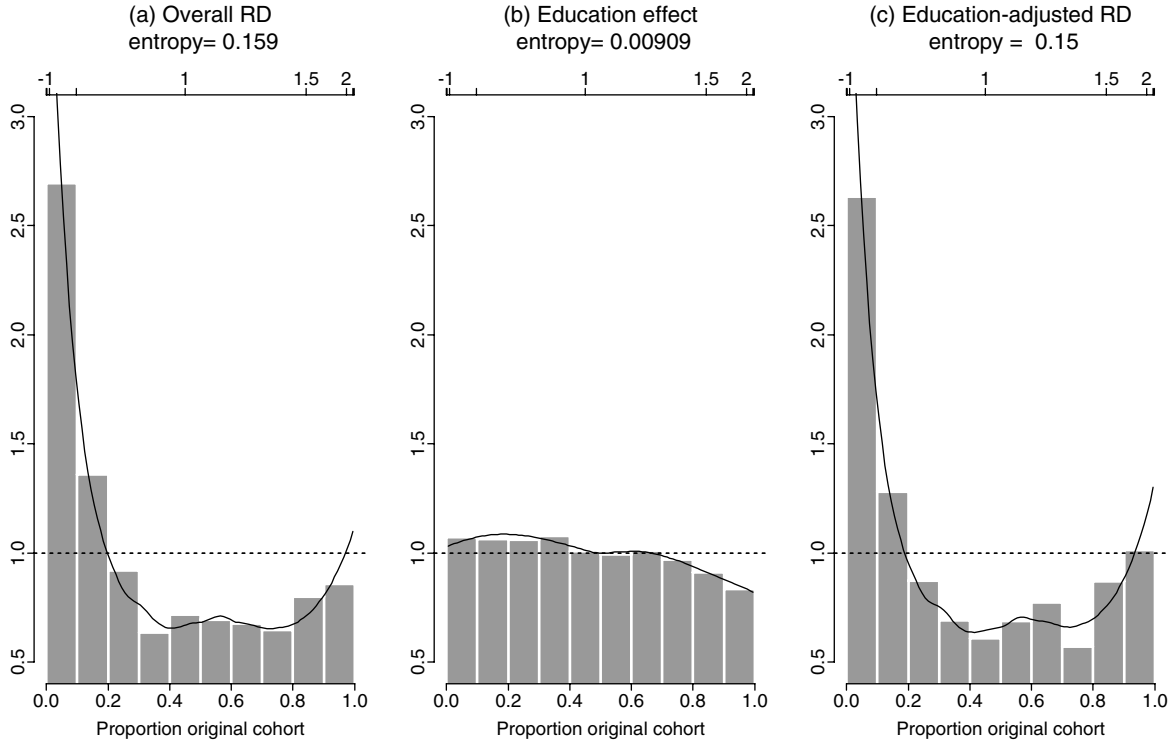

Fig. 8.5. Adjusting the relative distribution of permanent wage growth for changes in the education composition between the two cohorts. (a) The (unadjusted) relative density of wage growth; (b) the effect of changes in the education profile between the cohorts; (c) the education-adjusted relative density of wage growth.

Table 8.3. Entropy summaries for the education composition adjustment of wage gains: recent to original NLS cohort

Entropy	Estimate	95% CI	p-value
Overall change	0.159	0.119–0.198	0.000
education composition effect	0.009	0.003–0.005	0.004
composition–adjusted effect	0.150	0.110–0.189	0.000
Education composition percent	5.7%	4.5–6.9	0.004
Composition–adjusted percent	94.3	86.3–102.3	0.000

Once the relative density has been composition adjusted, one can examine both the composition and residual components for location and shape changes. As a hypothetical example, suppose the education composition effect in Figure 8.5 had been large. A location/shape analysis might then go on to show that the composition effect was primarily a location shift, while the residual was both location shifted and more polarized. Location shifts in the residual component capture the impact of the changing "returns" to education that are often the focus of regression-based wage analyses. Shape shifts go beyond this to represent changes in the dispersion of conditional returns that are typically ignored by regression-based models. This kind of analysis can provide a rich description of the interrelated distributional changes, which in turn can help to inform and focus a theoretical debate by clearly identifying what needs to be explained. Combining compositional adjustment with location/shape decomposition is straightforward, and given the lack of compositional effect found in the analysis above, we will not pursue this example further.

8.2.3 Categorical contrasts

In the literature on the growth in cross-sectional inequality, a consistent finding is that the wage premium for a college education has risen substantially (Juhn and Murphy 1993; Katz and Murphy 1992). While the college-educated are not doing uniformly better than they had in previous cohorts – those with low education are doing relatively worse on almost all measures: wages, job stability (Farber 1997), benefits (Farber 1996), and employment (DiPrete 1993). A natural question is whether this penalty can also be found in wage growth profiles, and what kinds of location and shape shifts are at work.

Figure 8.6 compares the distributions of wage gains for the two education groups, as density overlays (a and c) and as relative densities, recent to original cohort (b and d). Panels (a) and (b) compare the wage gains for the high school-educated across the two cohorts. The downshifting of wage gains for the recent cohort is quite apparent. Three times as many earners in the recent cohort experience wage gains in the bottom decile of

the original cohort, and there are 20–50% fewer wage gains in any of the deciles above the median. The relative distribution for this group is dominated by the location shift, as the relative density is nearly monotonic in its decrease. Panels (c) and (d) show the corresponding distributions for the college-educated. For this group, the change between the two cohorts is less pronounced, and takes a different form. The relative frequency of both low and high wage gains increases for the recent cohort, though low wage gains still predominate. About twice as many recent wage gains fell in the bottom decile of the original distribution, but the fraction falling in the highest decile also rose by nearly 20%. Overall, the relative distribution for the college-educated exhibits modest polarization and little evidence of a location shift.

Table 8.4. Summary statistics for cohort relative distributions by education

Measure	High School	College
Median ratio (unlogged)	0.77	1.00
Entropy	0.295	0.077
median shift effect	0.254	0.000
shape shift effect	0.041	0.077
Polarization (MRP)	0.17	0.19
lower tail (LRP)	0.22	0.29
upper tail (URP)	0.12	0.09

The summary statistics for these patterns are presented in Table 8.4. The median ratio (based on the unlogged wage gains) shows a 23% loss in real wage gains for the high school group, while the college group held steady. The entropy summary suggests that the overall change experienced by the high school group was three to four times as large as that experienced by the college group. While the median shift explained about 86% of the total change for the high school group, all of the change for the college group was due to changes in distributional shape. For both groups, the shape change took the form of growing inequality – as the MRP is significantly greater than 0 – with greater polarization in the lower tail of the distribution. The similarity in the magnitude of overall polarization was not evident from the graphical displays. For the college group, however, the polarization in the lower tail was much more extreme, because the index for the lower tail is over twice as large as that for the upper. This pattern is visible in the relative density panel in panel (d) of Figure 8.6: in the absence of a median shift, the panel is effectively displaying the shape shift.

To compare the two groups directly, we can compare the two RDs in Figure 8.6. Note that by first comparing the two groups *within* each cohort, we are effectively controlling for the compositional differences between cohorts, but the composition *effect* remains implicit. The patterns are

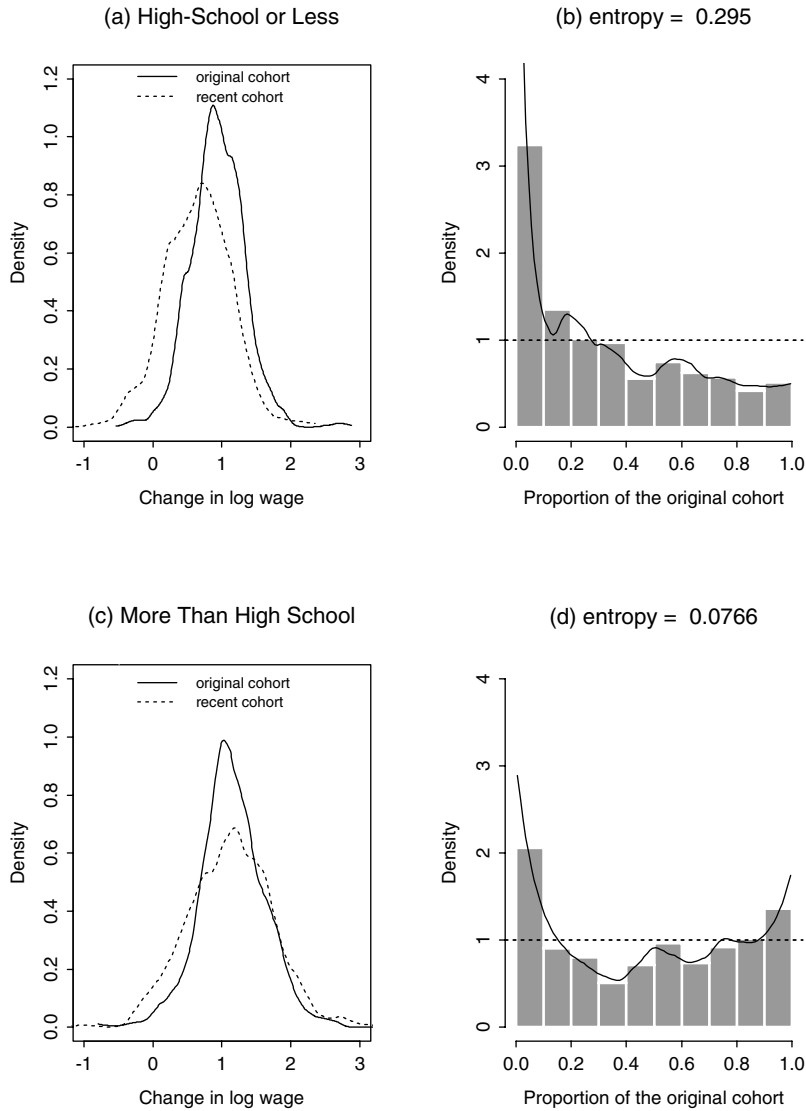

Fig. 8.6. The PDF overlays and cohort relative distributions of permanent wage growth for high school- and college-educated workers in the NLS. (a) Wage gain PDFs for workers with high school or less education in each cohort; (b) cohort relative distribution (R:O) for those with high school or less education; (c) wage gain PDFs for workers with some college education in each cohort; (d) cohort relative distribution (R:O) for those with some college education. A decile bar chart is superimposed on the relative density estimates. The upper and right axes are labeled in permanent log-wage gains.

summarized in Table 8.5 The first column indicates the decile of the college level wage gains. The second and third columns of the table represent the relative deciles (high school:college) for the original and recent cohorts (Figures 8.6, panels (b) and (d) respectively). The last column shows the difference, and represents how the high school group fared relative to the college group from one cohort to the next. In the bottom decile of the college wage gain distribution, for example, the fraction of high school earners rose from 19% to 25% between the two cohorts. Glancing down the last column of this table, we can see that high school earners fared worse in general: their relative fraction increased in almost every decile below the median and decreased in every decile above. These changes were produced by the combination of median and shape shifts in the high school and college RDs, and the relative impact of each shift can be identified.

Table 8.5. Decile relative distributions of high school to college educated: recent and original NLS cohorts

Decile	Original Cohort	Recent Cohort	Change in Decile
1	19.3	24.8	5.5
2	14.3	18.8	4.5
3	16.8	16.5	-0.3
4	8.7	13.0	4.3
5	10.2	11.5	1.2
6	10.8	6.5	-4.3
7	8.4	4.3	-4.0
8	5.1	2.4	-2.6
9	3.9	1.2	-2.7
10	2.5	1.0	-1.5

How much did the location and shape shifts in each groups' distribution contribute to the overall change in their relative positions? A natural way to answer this question would be to compare the observed changes in column 4 to the changes that would have occurred if only the medians (or shapes) had changed. This suggests a decomposition into the "marginal effects" of each change.

Let H_s^m and C_s^m denote the distribution of wage gains for the high school and college groups respectively, with the median adjusted to the mth cohort, and the shape (or conditional distribution of returns) from the sth cohort. For example, H_o^r is H_o^o, the original high school wage gain distribution, adjusted to have the same median as the recent cohort. Let $g(H : C)$ denote the relative density of wage gains for high school to college educated workers. Then the marginal effects of the median shift from the original relative density can be defined as

$$g(H_o^r : C_o^r) \; - \; g(H_o^o : C_o^o).$$

As with the location shift in the residual component of the compositional adjustment, the relative median shift here captures the effects of the change in median returns to education between the two cohorts. If both groups had experienced the same median gains, then their relative positions would be unchanged, and the quantity above would take the uniform value of zero. The marginal effect of the shape change in the high school distribution can be defined similarly as:

$$g(H_r^o : C_o^o) \; - \; g(H_o^o : C_o^o),$$

and the marginal effect of the shape change in the college distribution as:

$$g(H_o^o : C_r^o) \; - \; g(H_o^o : C_o^o). \tag{8.1}$$

Each of these effects compares the original relative density, $g(H_o^o : C_o^o)$, to the hypothetical density that would have been produced by a change in the specific distributional component alone.

The effects do not sum to the total difference shown in the last column of Table 8.5 because they do not occur independently. The difference between the sum of these effects and the total change can be interpreted as an interaction effect. The effect of each distributional change depends on the others: if the median shift has moved a substantial fraction of workers out of a decile, then the shape shifts will be operating on a smaller base and will move a correspondingly smaller fraction of workers out of that decile (or into another) than they would have in the absence of a median shift. This interaction among the effects makes a unique decomposition of the total change into the three components ambiguous, unless one is willing to specify the order in which the effects are applied. The principle is the same as that involved in decomposing the explained variance in the linear model context when the covariates are correlated. Here, as there, it is possible to define sequential effects that uniquely decompose the sum, but, in general, the order of the sequence will matter.

For comparison, we will obtain an exhaustive decomposition by defining the effects sequentially, specifying first the relative median shift:

$$g(H_o^r : C_o^r) \; - \; g(H_o^o : C_o^o),$$

then the shape change in the college distribution:

$$g(H_o^r : C_r^r) \; - \; g(H_o^r : C_o^r),$$

and the shape change in the high school distribution:

$$g(H_r^r : C_r^r) \; - \; g(H_o^r : C_r^r). \tag{8.2}$$

These effects do sum to the total change shown in Table 8.5. The sequential order seems reasonable in this case: median effects are in some sense more fundamental than shape effects, and changes in the reference distribution

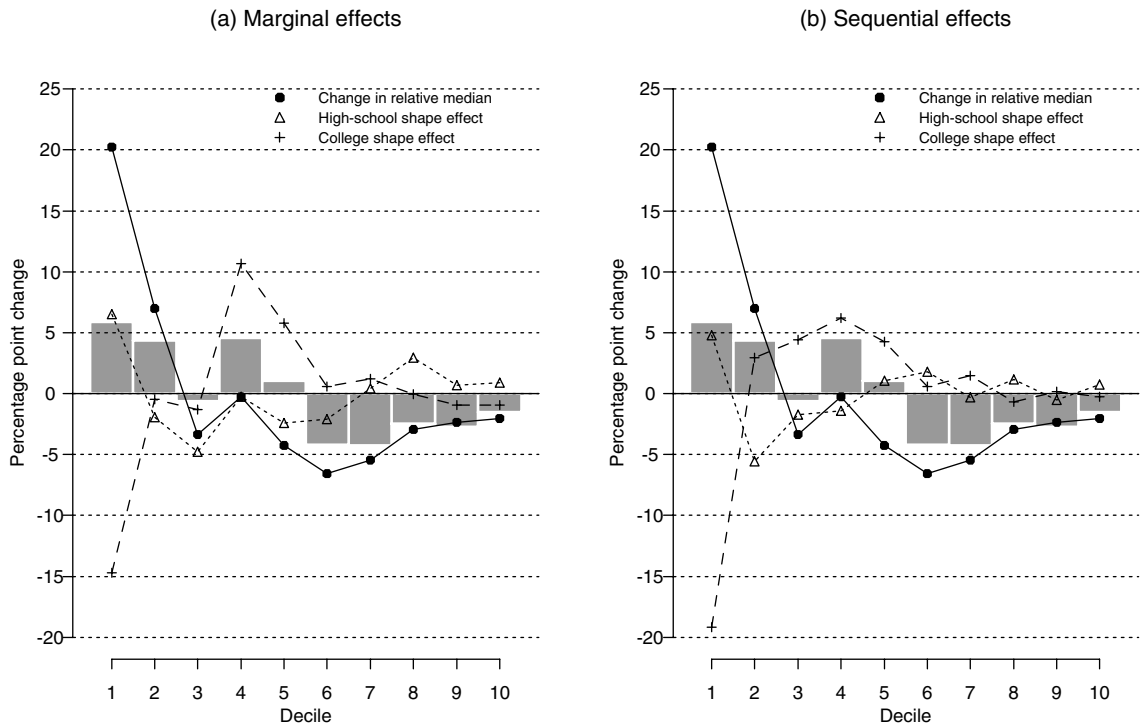

Fig. 8.7. Sources of the change in the cohort relative distribution of wage gains by education level. (a) The marginal effects as defined by (8.1); (b) the sequential effects as defined by (8.2).

(the college-educated here) can be seen as more fundamental than changes in the comparison distribution. The relative median shift is the same in (8.1) and (8.2).

Figure 8.7 presents the two decompositions side by side. Panel (a) represents the marginal effects as defined by (8.1), and panel (b) the sequential effects as defined by (8.2) The solid bars show the total change from Table 8.5. Each of the lines represents one of the three components in the decomposition. The interaction effect is not plotted in the first panel, but it is quite large in some deciles: {-4.7, -0.9, 4.2, -3.9, 2.4, 2.0, 1.5, -2.1, -1.5, and 0.9}

Qualitatively, the two decompositions show the same picture: the relative median shift tends to be the most influential contributor to the overall pattern. This reproduces the key finding from recent regression-based analyses that the education premium is having a large effect on wage changes. As we have seen here, the premium is not rising for the college-educated, but falling for the high school-educated.

The median effect is clearly modified by the shape changes, however, and this is something that would be missed by regression-based analyses. Particularly in the bottom decile, the relative median downshift alone would have produced nearly double the number of low wage gains for the high school-educated, but the strong polarization in the lower tail of the college group's distribution virtually nullified that shift. The net result looks much like the effect due to the modest polarization in the high school distribution: 5–7% more high school-educated workers in the lowest decile of the college earners' wage gains. Much of the growth in the remaining deciles below the median appears to have been generated by the polarization in the college wage gains: as college earners moved out of this section of the distribution, the relative fraction of high school earners increased. For the upper deciles the losses appear to be driven by the median shift, but dampened by the shape change in the high school distribution. Recalling the summary polarization indices shown in Table 8.4, we can see that the modest polarization in the upper tail of the high school educated distribution, and the lack of polarization in the upper tail of the college-educated distribution, helped to offset the high school group's losses in the upper deciles.

Exercises

Exercise 8.1. The model for the permanent wages described in Appendix C has individual-specific quadratic equations. That is, for each respondent, the profile is represented as:

$$y_{it} = b_{0i} + b_{1i}\text{age}_t + b_{2i}\text{age}_t^2,$$

where y_{it} is the log of real (PCE-deflated) permanent wages for respondent i at time t. Each of the coefficients b_{0i}, b_{1i}, and b_{2i} represent a combination of the fixed and random effects for the lifecycle growth in wages. This

is sometimes called the "full-profile heterogeneity" model. Consider an alternative model for the permanent wages where only the intercept b_{0i} is individual-specific (i.e., random). This assumes that the individual profiles only differ in overall level, but not in trend or shape. Show that the difference in permanent wage gains between two individuals is equal to their difference in intercepts.

Exercise 8.2. Fit the model for permanent wages with only the intercept random (See Exercise 8.1). Use it to estimate the permanent wage gains for individuals. Compare the results of the intercept-only model to the full-profile heterogeneity model for both cohorts.

Exercise 8.3. Repeat the analyses in the chapter using permanent wages estimated under the model with only the intercept random. Do the conclusions of the analysis change?

Exercise 8.4. The permanent wage model with individual-specific intercept and slope is intermediate between the intercept only model and the quadratic model used in the chapter. Fit the model for permanent wages with intercept and slope random (See Exercise 8.1). Use it to estimate the permanent wage gains for individuals. Compare the results from this model to the full-profile heterogeneity results.

Exercise 8.5. Repeat the analysis in the chapter using permanent wages estimated under the model with both the intercept and slope random. Compare the results to those obtained from Exercise 8.3 and the full-profile heterogeneity model.

Exercise 8.6. The complement of the permanent wages are the transient wage residuals. These are defined as ϵ_{it}, the difference between the observed log-wages and the permanent wages for respondent i at time t. Under the model, they are independent of transient wage effects at other time points and other respondents. They can be estimated by the difference between the observed log-wages and the estimated permanent wages. Estimate the transient wage effects for the original and recent cohorts. Give a statistical summary of their properties separately for each cohort.

Exercise 8.7. Calculate the relative distribution of transient wage effects from the recent cohort to those in the original cohort. Will there be a location effect? Compare the transient wage effects from the two cohorts using the relative distribution and other numerical summaries.

Exercise 8.8. The covariate decomposition in Section 8.3 adjusts for compositional differences in education. Age is another important determanent of wage. Would you suggest compositionally adjusting for age differences between the two cohorts? Give reasons why it should, or should not, be an important factor in this analysis. Complete a covariate decomposition using age similar to that in Section 8.3. Does it support your intuition about the effect of age?

Exercise 8.9. Another important determinant of wage is industry. Using national data (obtainable from the Bureau of Labor Statistics, series `EES00000001`, at `http://www.bls.gov/top20.html`) compare the change in the industrial composition of jobs over the period that these two cohorts were followed. The NLS respondents work in a range of industries, and may change industries over the observation period. Discuss how you would adjust for the changing mix of industries in each cohort, given the longitudinal nature of the surveys.

Exercise 8.10. What other covariates might be important to control for? Describe how you would interpret a compostional adjustment for these variables. Describe how you would interpret a categorical contrast.

This page intentionally left blank

Chapter 9

Inference for the Relative Distribution

In this chapter we address the estimation of the relative CDF, PDF, and Lorenz PDF based on survey data. The technical level of this chapter is higher than others in the book, but the ideas are quite intuitive. For those needing an introduction to the more technical concepts, Chapters 1-3 in Simonoff (1996) provide the necessary background. The more detailed results and proofs are given in Appendix D.

The estimation of univariate CDFs and PDFs has been extensively studied. See Silverman (1986), Scott (1992), and Simonoff (1996) for comprehensive descriptions of both the underlying theory and practice. This literature focuses almost exclusively on the situation where a single random sample from the distribution of interest is available. Results from this literature are directly applicable to the relative distribution context if the reference distribution is assumed to be known (see Section 2.3). In the more general case, where we have random samples from both the reference and comparison distribution (and the reference distribution is not known), we must turn to the literature on two-sample estimation, and this is much less extensive.

We first consider in Section 9.1 the situation where the reference distribution is known and the comparison distribution is estimated. This situation is closely allied to the usual one-sample situation. We consider a number of approaches to density estimation that also can be applied fruitfully to the two-sample situation. In Section 9.2 we consider the situation where both the reference and the comparison distributions are unknown. Insight can be gained into the process by considering the two-sample rank statistics. Estimators for both the relative CDF and PDF are developed that parallel those in the one-sample situation. In Section 9.3 inference for a reference group formed by pooling the comparison and reference groups is considered. In Section 9.4 the important situation where the observations are censored is studied as a generalization of the case where the values are completely observed. In Section 9.5 we consider the case, common in survey samples, where the observations have associated weights. In Section 9.6, we use the results in the previous sections to derive confidence intervals and confidence bands for the relative CDF and PDF.

In this chapter we do not explicitly consider the situation where both the comparison and reference distributions are known to be members of parametric families of densities. The estimation of parametric densities has been extensively studied - see Rice (1995). In Section 9.2.3.4, however, we consider the situation where the relative distribution is known to be a member of a parametric family.

9.1 Estimation when the reference distribution is known

As we noted in Section 2.4, there are some situations when the reference distribution is known or prespecified. In this section we assume that the reference CDF, F_0, is known and the data on the comparison population arises from a sample survey. That is, we assume that we have a sample $Y_1, Y_2, \ldots, Y_m$ that is independently and identically distributed from the population distribution F. In reality the sample is drawn from a large but finite population. The approach taken here assumes that there is some underlying process that generates the finite population, and that the process has distribution F. There are a number of perspectives on this issue that lead to distinct inference about the relative distribution. In particular we could consider random sampling from a finite and fixed population. A discussion of these issues in the context of inequality measures is given in Nygård and Sandström (1989). The situation where the observations have sample weights is considered in Section 9.5.

One naive approach to estimating the relative distribution is to estimate F and f and use the formulae in Section 2 to determine estimates of the relative CDF and PDF. In this two step process, one first estimates the distribution of the comparison population and then uses this to estimate the relative distribution. As the support and behavior of the comparison population will vary from problem to problem, here we will focus on the direct estimation of the relative distribution (See Exercise 9.1).

Based on the sample, we can define the *relative data* (Section 2.2):

$$R_j = F_0(Y_j) \qquad j = 1, \ldots, m. \tag{9.1}$$

We will assume in this section that F_0 is absolutely continuous (i.e., is continuous and not extremely rough – see Section 2.1). As the sample is independently and identically distributed drawn from the CDF F, the relative data are independently and identically distributed drawn from the CDF G. Thus we can directly apply CDF and PDF estimation methods to the relative data on the support $[0, 1]$.

9.1.1 Estimation of the relative cumulative distribution function

The natural estimator of the relative CDF is the *empirical cumulative distribution function*, denoted by $G_m(r)$ and defined to be the proportion of the relative data that do not exceed the value $r \in [0, 1]$. This function is also called the *empirical distribution function* and the *sample distribution function*. It can be represented mathematically as

$$G_m(r) = \frac{1}{m} \sum_{j=1}^{m} \mathcal{I}(R_j \leq r)$$

where

$$\mathcal{I}(S) = \begin{cases} 1 & \text{if the event } S \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

is the indicator function. Note that $G_m(r)$ is a step function of r with jumps of $1/m$ at the ordered values of the relative data. In addition, for a fixed but otherwise arbitrary value of r, $G_m(r)$ is itself a random variable. Our main interest in $G_m(r)$ is as an estimator of $G(r)$. The exact distribution of $mG_m(r)$ is binomial with m trials and probability of success $G(r)$ (Exercise 9.2).

This result makes it easy to determine the behavior of $G_m(r)$ for large samples. From the central limit theorem (Kelly 1994) the *asymptotic* behavior of $G_m(r)$ can be described: as the sample size increases, the distribution of $G_m(r)$ approaches that of a normal distribution with expected value $G(r)$ and variance $G(r)[1 - G(r)]/m$. Mathematically this can be written:

Theorem. *For each value of* $0 < r < 1$, $G_m(r)$ *is a consistent estimator of* $G(r)$. *The sequence* $G_m(r), m = 1, 2, \ldots,$ *is also asymptotically normal:*

$$G_m(r) \sim AN\left\{G(r), \frac{G(r)(1 - G(r))}{m}\right\} \qquad 0 < r < 1 \qquad (9.2)$$

as $m \to \infty$.

The notation is described more formally in the Background material. This result shows that there is convergence for each value of r individually. One commonly used measure of the global closeness of $G_m(r)$ to $G(r)$ is the Kolmogorov-Smirnov distance

$$D_m = \sup_{0<r<1} |G_m(r) - G(r)|.$$

The convergence of $G_m(r)$ to $G(r)$ occurs simultaneously for all r in the sense that D_m converges to zero with probability one, that is, $P[\lim_{m\to\infty} D_m = 0] = 1$. This result suggests that for large sample sizes the deviation between $G_m(r)$ and $G(r)$ will be small for all r. There is

much known about the distribution of D_m and related quantities – see, for example, Serfling (1980).

The relative distribution of Y_0 to Y is $G^{-1}(p)$, the quantile function of G. The natural estimator $G^{-1}(p)$ is the pth quantile of the sample distribution function $G_m(r)$ defined by

$$G_m^{-1}(p) = \inf\{r : G_m(r) \geq p\}.$$

The properties of $G_m^{-1}(p)$ as an estimator of $G^{-1}(p)$ are similar to those of $G_m(r)$ as an estimator of $G(r)$. In particular, similar to (9.2), we have:

Theorem. *Assume that* $0 < p < 1$ *and let* $\lambda_p = F_0^{-1}(p)$. *Suppose both* $F_0(y)$ *and* $F(y)$ *possess densities* ($f_0(y)$ *and* $f(y)$, *respectively) in a neighborhood of* λ_p, *and* $f_0(y), f(y)$ *are positive and continuous at* λ_p. *Then the density* $g(r)$ *of* $G(r)$ *exists at* $r = p$ *and,*

$$G_m^{-1}(p) \sim AN\left\{G^{-1}(p), \frac{p(1-p)}{mg^2(G^{-1}(p))}\right\} \tag{9.3}$$

as $m \to \infty$.

Finally consider the complementary situation when the comparison distribution is known, while the reference distribution must be estimated from a sample. Using the above inverse relationship, the natural estimators of $G^{-1}(p)$ and $G(r)$ are

$$\tilde{G}_n^{-1}(p) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{I}\Big[F(Y_{0i}) \leq p\Big]$$

and

$$\tilde{G}_n(r) = \inf\{p : \tilde{G}_n^{-1}(p) \geq r\},$$

respectively. The above result can then be reformulated:

Theorem. *Assume that* $0 < r < 1$ *and let* $\lambda_r = F^{-1}(r)$. *Suppose both* $F_0(y)$ *and* $F(y)$ *possess densities* ($f_0(y)$ *and* $f(y)$, *respectively) in a neighborhood of* λ_r, *and* $f_0(y), f(y)$ *are positive and continuous at* λ_r. *Then*

$$\tilde{G}_n(r) \sim AN\left\{G(r), \frac{r(1-r)g^2(r)}{n}\right\} \tag{9.4}$$

as $n \to \infty$.

Although the estimators described in this section are well studied they do have the drawback that they are step functions, while $G(r)$ is usually continuous and much smoother. This suggests that alternative estimators exist that may better reflect the properties of $G(r)$. In particular, if we had a smooth estimator of $g(r)$, $\hat{g}(r)$ say, we could estimate $G(r)$ by $\int_0^r \hat{g}(p)dp$. We shall briefly review such estimators for the PDF in the next section.

9.1.2 Estimation of the relative probability density function

9.1.2.1 Estimation using a histogram.

As the relative PDF is the derivative of the relative CDF, it will tend to be less smooth than the CDF and often more difficult to estimate. In this section we briefly review some approaches. For a detailed discussion see Simonoff (1996), Chapter 2, whom we follow. The relationship between the PDF and CDF given in equation (2.1) suggests splitting up $[0, 1]$ into K equisized intervals each with width $h = 1/K$. If the number of intervals is large enough, then each interval will be small and we can consider an estimator of the form:

$$\hat{g}(r) = \frac{G_m(b_{j+1}) - G_m(b_j)}{h}, \qquad x \in (b_j, b_{j+1}], \tag{9.5}$$

where $b_j = (j-1)/K$, $j = 1, \ldots, K+1$ and $(b_j, b_{j+1}]$ defines the boundaries of the jth interval. This is the familiar *histogram* estimator of $g(r)$. The advantages of the histogram in this setting are ease of interpretability and convenient construction with most statistical packages. For example, if $K = 10$, then each interval on the horizontal axis corresponds to a decile of the reference distribution. A graph of the decile histogram estimator for the standardized residuals from the purchasing power parity model (discussed in Section 2.3) is given in Figure 9.1. We can see that the estimate for the third decile is about 2.4, indicating there are 2.4 times as many standardized residuals in the third decile of the standard normal distribution as we would expect by chance if the regression assumptions were satisfied. The fifth decile has only 30% of the expected number of residuals. This estimate should be compared to that of Figure 2.5.

How can we evaluate $\hat{g}(r)$ as an estimator of $g(r)$? It is natural to consider the squared error, $\mathrm{SE}(r) = [\hat{g}(r) - g(r)]^2$, and its expected value (mean squared error), $\mathrm{MSE}(r) = E_g\,[\hat{g}(r) - g(r)]^2$. If we wish to measure global accuracy over the full interval $[0, 1]$, we can consider the integrated squared error,

$$\mathrm{ISE} = \int_0^1 [\hat{g}(u) - g(u)]^2\, du,$$

and its expected value, mean integrated squared error (MISE). We can measure these for any sample size and K, but here will consider the asymptotic behavior when the sample size $m \to \infty$. As we get more data we should increase the number of intervals to capture the detailed structure of $g(r)$ but do so slower than the sample size increases to reduce the variability of $\hat{g}(r)$ within each interval. Based on the binomial distribution for $mG_m(r)$ and (9.5), the distribution of $m\hat{g}(r)$ is binomial with m trials and probability of success $G(b_{j+1}) - G(b_j)$ where $x \in (b_j, b_{j+1})$. Simonoff (1996) shows that if the interval width $h \to 0$, and $mh \to \infty$, as $m \to \infty$ and $g(r)$ is smooth enough ($g'(r)$ is absolutely continuous and square integrable), then

$$\begin{aligned}\text{Bias}\left[\hat{g}(r)\right] &\equiv E_g\left[\hat{g}(r)\right] - g(r) \\ &= \frac{1}{2}g'(r)\left[h - 2(r - b_j)\right] + O\left(h^2\right), \qquad r \in (b_j, b_{j+1}], \end{aligned} \tag{9.6}$$

while the variance is

$$\text{Var}\left[\hat{g}(r)\right] = \frac{g(r)}{mh} + O\left(m^{-1}\right). \tag{9.7}$$

If the relative distribution is uniform then the bias is exactly zero and the variance is $(1 - h)/(hm)$.

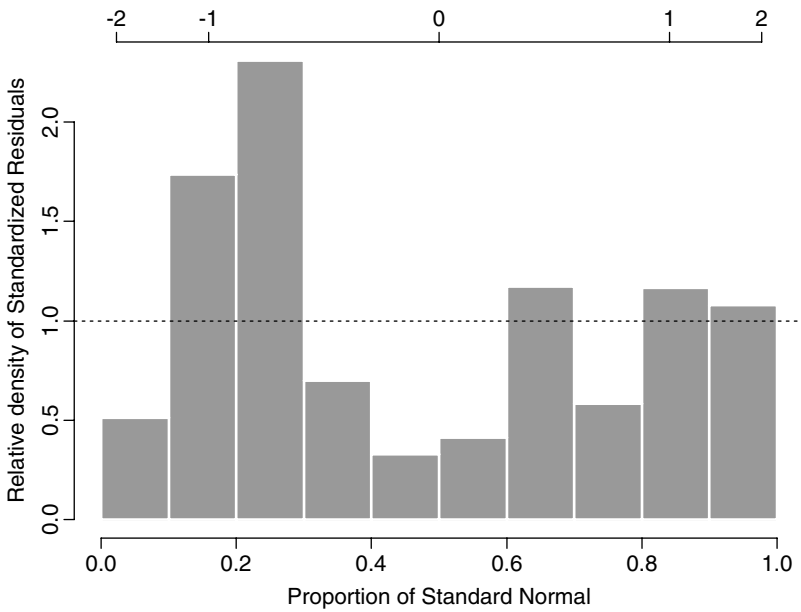

Fig. 9.1. Histogram of the relative distribution of standardized residuals for the Purchasing Power Parity data.

Combining the squared bias and variance yields the mean squared error,

$$\begin{aligned}\text{MSE}\left[\hat{g}(r)\right] &= \text{Var}\left[\hat{g}(r)\right] + \text{Bias}^2\left[\hat{g}(r)\right] \\ &= \frac{g(r)}{mh} + \frac{g'(r)^2}{4}\left[h - 2(r - b_j)\right]^2 \\ &\quad + O\left(m^{-1}\right) + O\left(h^3\right). \end{aligned}$$

Finally, integrating over each interval, and then summing interval by interval, gives

$$\text{MISE} = \frac{1}{mh} + \frac{h^2 R(g')}{12} + O\left(m^{-1}\right) + O\left(h^3\right), \tag{9.8}$$

where

$$R(v) = \int_{-\infty}^{\infty} [v(x)]^2 dx.$$

If the relative distribution is uniform then the MISE is $(1-h)/(hm)$. We will write AMISE to represent the asymptotic MISE (that is, the leading two terms in the expansion of MISE). The minimization of MISE requires explicitly balancing bias and variance through the choice of the number of bins, or equivalently $h = 1/K$. For the uniform distribution the choice $h = K = 1$ minimizes the MISE. This choice, however, is unrealistically smooth and so not useful in general. For non-uniform relative distributions, the minimizer of AMISE is easily determined as

$$h_0 = \left[\frac{6}{R(g')}\right]^{1/3} m^{-1/3}.$$

In practice, we need to specify a particular non-uniform g to operationalize this rule. A reasonable candidate is a beta distribution with shape parameters (2,2) which is normal-looking and leads to the rule:

$$h_0 = 0.7937 m^{-1/3}.$$

Many rules-of-thumb have been suggested that are similar to:

$$h_0 = 2IQR m^{-1/3},$$

where IQR is an estimate of the interquartile range of the distribution. If we use a distribution of normal shape but with the spread of the uniform we get the rule $h_0 = m^{-1/3}$. See Simonoff (1996) for a discussion of these issues.

9.1.2.2 Kernel density estimation.

The histogram estimator has the drawback in that it is a step-function and does not adapt to the shape of the relative density. This means that it may not be as close to the actual relative density as is practically possible, undersmoothing or oversmoothing by some nontrivial degree. We can improve on this by considering estimators of the form:

$$g_m(r) = \frac{1}{mh} \sum_{j=1}^{m} K\left(\frac{r - R_j}{h}\right), \tag{9.9}$$

where $K(\cdot)$ is a function satisfying

$$\int_{-1}^{1} K(u)du = 1, \quad \int_{-1}^{1} uK(u)du = 0, \quad \int_{-1}^{1} u^2 K(u)du = \sigma_K^2 > 0. \tag{9.10}$$

This is called a *kernel density estimator* and $K(\cdot)$ is called a (bounded) *kernel function*. The estimator can be thought of as a generalization of the histogram estimator with each point r being estimated separately as the center of its own interval and the kernel being used to place more weight closer to the observations R_j. To reflect our belief that the underlying relative density is smooth, we will assume that g is uniformly continuous and g''' is square integrable. As for the histogram we can consider the properties of the estimator as the sample size increases and the "bandwidth" h decreases. If $h \to 0$ with $mh \to \infty$ as $m \to \infty$, then by Taylor Series expansions (Silverman 1986):

$$\text{Bias}[g_m(r)] = \frac{1}{2}h^2\sigma_K^2 g''(r) + O(h_m^4)$$

and

$$\text{Var}[g_m(r)] = \frac{g(r)R(K)}{mh} + O(m^{-1}).$$

The advantage of the kernel density estimator can be seen in the bias term. It is now of size h^2 rather than h.

Theorem. *Assume that* $0 < r < 1$, *and suppose both* $F_0(y)$ *and* $F(y)$ *possess densities* ($f_0(y)$ *and* $f(y)$, *respectively) that are smooth (enough so that* g *is uniformly continuous). Let* $K(\cdot)$ *be a twice continuously differentiable kernel function (satisfying (9.10)) and vanishing outside some bounded interval. For each bandwidth sequence* $\{h_m\}$ *with* $h_m \to 0$ *with* $mh_m \to \infty, m/n \to \kappa^2 < \infty$ *we then have*

$$g_m(r) \sim AN\left\{g(r) + \frac{1}{2}h^2\sigma_K^2 g''(r), \frac{g(r)R(K)}{mh_m}.\right\}$$

If the bandwidth approaches zero quickly the bias in $g_m(r)$ gets small relative to the standard deviation of $g_m(r)$. We can then (asymptotically) ignore the bias term:

Theorem. *Under the same conditions as the previous result, suppose the bandwidth sequence* $\{h_m\}$ *with* $h_m \to 0$ *with* $mh_m^3 \to \infty$, $mh_m^5 \to 0, m/n \to \kappa^2 < \infty$. *Then*

$$g_m(r) \sim AN\left\{g(r), \frac{g(r)R(K)}{mh_m}.\right\}$$

To operationalize the estimator, we need to choose a kernel function and an estimator for the bandwidth. The choice for the kernel function depends on properties of the (unknown) g. Fortunately there are many choices that appear to work similarly well (Simonoff 1996). In our applications we use the biweight, which has the mathematical form

$$\frac{15}{16}(1-u^2)^2 \qquad -1 \le u \le 1$$

and zero otherwise. To choose the bandwidth, we again balance the average mean squared error over the interval. The asymptotic MISE can be shown to be:

$$\frac{R(K)}{mh} + \frac{1}{4}h^4\sigma_K^4 R(g'').$$

The form of the bandwidth h to minimize the asymptotic MISE is

$$h_{0R} = \left[\frac{R(K)}{\sigma_K^4 R(g'')}\right]^{1/5} m^{-1/5}. \tag{9.11}$$

Note that h_{0R} does not satisfy the conditions of the previous theorem, as it approaches zero too slowly. If h_{0R} or another $O(m^{-1/5})$ bandwidth sequence is used the bias of $g_m(r)$ will need to be accounted for in confidence intervals based on these asymptotic approximations to the distribution of $g_m(r)$.

Choosing a bandwidth in practice requires an estimate of $R(g'')$. The most popular approach is the one by Sheather and Jones (1991), which estimates $R(g'')$ by $R(\hat{g}'')$ where $\hat{g}''$ is a smoothed estimator of g''. Simonoff (1996) discusses this estimator and the many other choices.

While kernel density estimators are very common, they have a significant weakness for estimating the relative density: edge (or boundary) effects. The form of the estimator generates downward bias for $r < h$ and values greater than $1 - h$, unless the relative density is zero there. Intuitively this is because the estimator does not recognize the boundaries of the interval. It acts as if the relative density is zero outside the interval $[0, 1]$ while it is in fact undefined there. There are ways to remove this bias by using a special kernel function called a *boundary kernel* (Gasser and Müller 1979). We will discuss these in Section 9.2.3.2 and will also consider other estimators that do not have this drawback.

Figure 9.2 graphs the kernel density estimator of the Purchasing Power Parity standardized residuals using the Sheather-Jones bandwidth choice of $h = 0.816$. Also on the figure is an estimator that does not suffer from boundary bias. We will consider it in the next section. Although the estimators differ close to the boundaries their behavior in the interior of the region is similar.

9.1.2.3 Regression based density estimation.

In this section we consider an estimator for the relative density based on a regression view point. For a detailed discussion, see Fan and Gijbels (1996), whom we follow. From (9.6) and (9.7) we can see that the histogram estimator satisfies:

$$E_g\left[\hat{g}(r)\right] \approx g(r) \qquad \mathrm{Var}\left[\hat{g}(r)\right] \approx \frac{g(r)}{mh}.$$

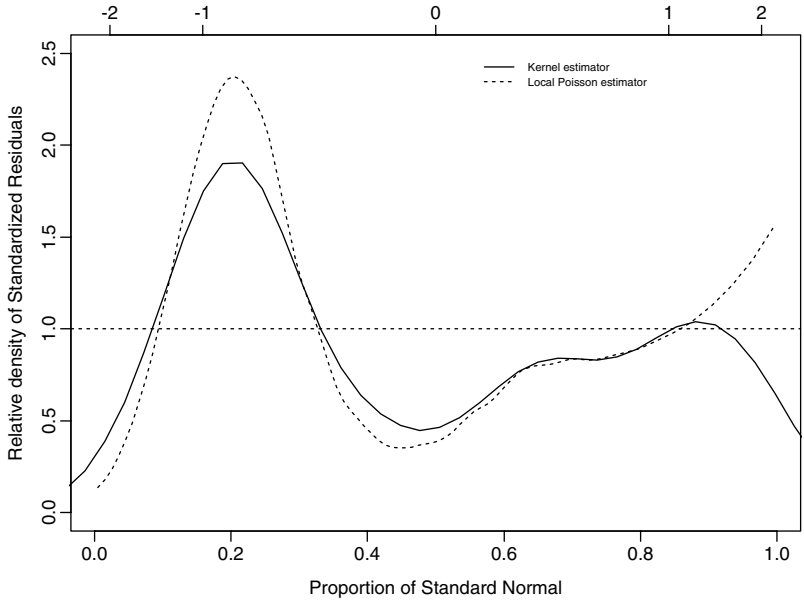

Fig. 9.2. Nonparametric estimates of the relative density of standardized residuals for the Purchasing Power Parity data.

We see that the histogram estimator has mean $g(r)$, nonhomogeneous variance $g(r)/(mh)$. Thus we can think of the histogram estimator as supplying the raw "data" for a regression of $\{\hat{g}(r_j)\}_{j=1}^K$ on $\{r_j\}_{j=1}^K$, where $r_j = (j - \frac{1}{2})/K$ are the centers of the intervals. We note that the joint distribution of the $\{m\hat{g}(r_j)\}_{j=1}^K$ is multinomial, so that the individual values are approximately independent for large K. We can model each value as a Poisson random variable with a nonparametric mean. This approach is described in Simonoff (1998) and Loader (1999). These models produce a smooth and non-negative mean function estimate $\hat{m}_P(r), 0 < r < 1$. A density estimator can be obtained by normalization:

$$\hat{g}_P(r) = \frac{\hat{m}_P(r)}{\int_0^1 \hat{m}_P(r)dr}.$$

If software for nonparametric Poisson regression is not available a cruder approach can be based on nonparametric least-squares regression. The idea is to approximately homogenize the variances by transforming the histogram estimator using a variance-stabilizing transformation (Anscombe 1948):

$$y(r_j) = 2\sqrt{m\hat{g}(r_j) + \frac{3}{8}} \qquad j = 1, \ldots, K.$$

We can then fit a nonparametric regression curve to the transformed data $\{y(r_j), r_j\}_{j=1}^{K}$ and produce a smooth mean function estimate $\hat{m}(r), 0 < r < 1$. First the mean function estimate is back-transformed:

$$\hat{g}_T(r) = \left[\frac{\hat{m}(r)^2}{4} - \frac{3}{8}\right]$$

to get to the original scale. The back-transformed mean curve cannot be used directly as an estimator of the relative density as it may be negative for some values of r and may not integrate to unity. However we can obtain a density estimator by truncation and normalization:

$$\hat{g}_R(r) = \begin{cases} \frac{\hat{g}_T(r)}{\int_0^1 \hat{g}_T(r)\mathcal{I}(\hat{g}_T(r)>0)dr} & \hat{g}_T(r) > 0 \\ 0 & \text{otherwise} \end{cases}.$$

To apply these approaches in practice we need to choose a nonparametric regression estimator. Two popular approaches are smoothing splines and local polynomial estimators. The smoothing spline approach is to fit a function through the scatter plot that balances goodness-of-fit to the points with the smoothness of the fit. The approach can also be viewed as breaking up the unit interval into segments and fitting low degree polynomials piecewise to each segment. The polynomials are chosen to match derivatives where they intersect at the boundaries of the segments so as to produce a function that is smooth. The local polynomial estimators can be regarded as a generalization of linear regression to allow for local nonlinearity in the mean function. For a discussion of these, see Simonoff (1996) and Fan and Gijbels (1996). For most of the application in this book we use Poisson regression based on a local quadratic estimator. This can be regarded as an extension of generalized linear regression to allow for local nonlinearity in the mean function. It is easy to interpret and is adaptable to more sophisticated situations.

Figure 9.2 graphs the Poisson local-quadratic density estimator of the Purchasing Power Parity standardized residuals. As with the kernel estimator, a bandwidth is required to specify the interval width over which the mean curve is approximately linear. We choose the value that minimizes the corrected Akaike Information Criterion proposed by Hurvich, *et al* (1998). They show that the criterion works well in practice and is easy to implement. The value chosen here is $h = 0.55$. Note that this should not be directly compared with the kernel bandwidth because the local-quadratic smoother is already using the smooth histogram estimates, and hence should be smaller. An advantage of the local-quadratic density estimator is that it does not suffer from the boundary bias of the kernel estimator. However the variance of the estimator near the boundary is much larger

than in the interior as there is less data close to the boundary that can be used. See Simonoff (1996), Section 5.3 for a way to choose the bandwidth for the variance-stabilized approach.

9.1.2.4 Exponential family based density estimation.

The final approach we will consider for estimating the relative density will be based on approximating it by using a member of a family of densities. The idea is to specify a large and flexible family of densities and use the maximum likelihood estimate within that family as the density estimator. While this introduces a parametric approach to the estimation process, the form of the parametric assumptions used here are substantially more flexible than those typically encountered in parametric methods. In particular, our approach makes use of basis functions, rather than a single functional form. In the resulting estimation process, the standard tradeoff between parametric assumptions and flexibility sacrifices less flexibility than usual.

Suppose we believed that the relative density was a member of a family of densities that were indexed by a parameter θ. For example we could consider the beta family:

$$g_\theta(r) = \frac{1}{B(\theta)} r^{\theta_1 - 1}(1-r)^{\theta_2 - 1} \qquad 0 < r < 1.$$

where $\theta = (\theta_1, \theta_2) > 0$ and $B(\theta)$ is the beta function:

$$B(\theta) = \frac{\Gamma(\theta_1)\Gamma(\theta_2)}{\Gamma(\theta_1 + \theta_2)}.$$

Based on the relative data, $R_1, \ldots, R_m$ the log-likelihood for θ is

$$\begin{aligned} \mathcal{L}(\theta; R_1 = r_1, \ldots, R_m = r_m) &\equiv \log\Big(P(R_1 = r_1, \ldots, R_m = r_m)\Big) \\ &= \sum_{j=1}^{m} \log\Big(P(R_j = r_j)\Big) \\ &= \sum_{j=1}^{m} \log g_\theta(r_j) \end{aligned} .$$

This can then be maximized as a function of θ for given values of $R_1 = r_1, \ldots, R_m = r_m$. If $\hat{\theta}$ are the values that maximize the log-likelihood, and hence likelihood, then the corresponding density $g_{\hat{\theta}}(r)$ is the maximum likelihood estimator of $g(r)$ within this family. The maximum value can often be found explicitly or as a solution to a simple expression. However in many cases, such as this one, the solution must be found using a numerical optimization routine.

This likelihood based approach has many advantages. It is often possible to prove that the estimate almost always exists and is unique. Because the estimator is obtained within a regular maximum likelihood framework,

much is known about its properties. For example, standard errors, convergence rates, and asymptotic distributions are well known (Brown 1986). A disadvantage of this approach is the assumption that the relative density is actually a member of this family. While the beta family is commonly used in statistical applications it does not capture the wide range of possible shapes that the relative density might have. What is needed is a much wider family that retains the computational and interpretative advantages of the beta family.

Consider an *exponential family* of densities of the form:

$$g_\theta(r) = g_0(r) \exp\left\{ \sum_{k=1}^{K} \theta_k \phi_k(r) - \Psi_K(\theta) \right\} \qquad 0 < r < 1. \tag{9.12}$$

where $\theta = (\theta_1, \ldots, \theta_K) \in \Theta = \{\theta \in \mathbb{R}^K : \Psi_K(\theta) < \infty\}$. Here $g_0(r)$ is a reference distribution, often taken to be the uniform density on $[0, 1]$. The function $\Psi_K(\theta)$ is the normalizing value so that each density integrates to one:

$$\Psi_K(\theta) = \log\left\{ \int_0^1 g_0(r) \exp\left\{ \sum_{k=1}^{K} \theta_k \phi_k(r) dr \right\} \right\}.$$

At the heart of the family are the *basis functions* $\{\phi_k(r)\}_{k=1}^K$, which are any bounded and linear independent functions such that

$$S_K \equiv \text{span}\{1, \{\phi_k(r)\}_{k=1}^K\}$$

is a linear space. See Brown (1986) for an extended description of the properties of such families. Three common choices for the basis functions are polynomials, trigonometric series, and spline bases.

The beta family is an exponential family with $K = 2$. It can be represented in the above form by writing $\phi_1(r) = \log(r), \phi_2(r) = \log(1 - r)$, and $\Psi_K(\theta) = -\log B(\theta)$. However, note that these two basis functions are not bounded nor independent and so are not formally in this family.

We can approximate $g(r)$ within this class by the maximum likelihood estimate. Based on the relative data, $R_1, \ldots, R_m$ the log-likelihood for θ is

$$\begin{aligned} &\mathcal{L}(\theta; R_1 = r_1, \ldots, R_m = r_m) \\ &= \sum_{j=1}^{m} \log g_\theta(r_j) \\ &= \sum_{j=1}^{m} \log g_0(r_j) + m \sum_{k=1}^{K} \theta_k \bar{\phi}_k(r_1, \ldots, r_m) - m\Psi_K(\theta), \end{aligned}$$

where

$$\bar{\phi}_k(r_1, \ldots, r_m) = \frac{1}{m} \sum_{j=1}^{m} \phi_k(r_j) \qquad k = 1, \ldots, K.$$

The relative data appear in the likelihood only through the statistics $\{\bar{\phi}_k\}_{k=1}^K$ and so they are sufficient statistics for θ. The log-likelihood is strictly concave and so the MLE is unique if it exists (Brown 1986). The gradient of the log-likelihood is

$$\begin{aligned} S(\theta; R_1 = r_1, \ldots, R_m = r_m) &\equiv \left[\frac{\partial \mathcal{L}(\theta; R_1 = r_1, \ldots, R_m = r_m)}{\partial \theta_k}\right] \\ &= \left[m\bar{\phi}_k(r_1, \ldots, r_m) - m\mathrm{E}_\theta\big(\phi_k(R_1)\big)\right]. \end{aligned}$$

where

$$\mathrm{E}_\theta\big[\phi_k(R_1)\big] \equiv \int_0^1 \phi_k(r) g_\theta(r) dr.$$

If $\hat{\theta}$ are the values that maximize the log-likelihood, then $g_{\hat{\theta}}(r)$ is the unique density within the family that satisfies:

$$\mathrm{E}_{\hat{\theta}}\big[\phi_k(R_1)\big] = \bar{\phi}_k(R_1, \ldots, R_m) \quad k = 1, \ldots, K.$$

These K "moment" conditions can often be solved explicitly or can produce greatly simplified equations. Let

$$\begin{aligned} H(\theta) &= \left[\frac{\partial^2 \Psi_K(\theta)}{\partial \theta_i \partial \theta_j}\right] \\ &= \left[\int_0^1 \phi_i(r)\phi_j(r) g_\theta(r) dr - \int_0^1 \phi_i(r) g_\theta(r) dr \int_0^1 \phi_j(r) g_\theta(r) dr\right] \end{aligned}$$

be the hessian matrix of $\Psi_K(\theta)$. The information matrix for the relative data is then $I(\theta) = -mH(\theta)$. Thus a Newton-Raphson algorithm can be used to find $\hat{\theta}$ in any case.

The statistical properties of the estimator $\hat{\theta}$ depend on what we assume about the relative density. In the simplest case the relative density is a member of the exponential family, so that the goal is to identify the correct member. The properties of the MLE can be summarized as:

Theorem. *Suppose that $g(r) = g_{\theta_0}(r)$ for some $\theta_0 \in \Theta$. Then*

$$\hat{\theta} \sim AN\big\{\theta_0, I^{-1}(\theta_0)\big\} \qquad 0 < r < 1$$

as $m \to \infty$.

This relation can be used to determine confidence intervals and bands for $g(r)$ as well as quantities derived from it (see Section 10.3 for the application to polarization statistics).

Interestingly there is much that can be said about the estimator if it is *not* in the exponential family. First consider the expected log-likelihood for a single observation R_1 :

$$\begin{aligned}\mathrm{E}_{R_1}\big[\mathcal{L}(\theta; R_1)\big] &\equiv \int_0^1 \log[g_\theta(r)]g(r)dr \\ &= \int_0^1 \log[\frac{g(r)}{g_\theta(r)}]g(r)dr + \int_0^1 \log[g(r)]g(r)dr \\ &= -D(g; g_\theta) + I(g).\end{aligned}$$

where $D(g; g_\theta)$ is the Kullback-Leibler divergence between g and g_θ, and $I(g) = D(F; F_0)$ is the entropy of the relative density (See Section 5.3). Suppose θ^* is the value in Θ that maximizes the expected log-likelihood. Then θ^* can be thought of as the MLE for an infinite sample size. Based on the above equation, we see that θ^* also minimizes the Kullback-Leibler divergence, so that we can also think of it as the member of the exponential family closest to the relative density. Of course, if the relative density is actually a member of the exponential family then $g_{\theta^*}(r) = g(r)$ and the two densities coincide. Furthermore we can interpret the difference between the two densities as the *model misspecification,* that is, the degree to which the exponential family model fails to capture the actual relative density. This relationship can be further quantified by:

$$D(g; g_{\hat{\theta}}) = D(g; g_{\theta^*}) + D(g_{\theta^*}; g_{\hat{\theta}})$$

This indicates that the divergence between the true relative density and the MLE can be decomposed into two components. The first measures the model misspecification and the second the *model uncertainty.* As the sample size increases the divergence between the MLE and the best the MLE can be (θ^*) decreases to zero. Stone (1989) and Barron and Sheu (1991) discuss the rates of convergence of these and related quantities. Statistically, we can be more specific about the size of the model uncertainty via:

Theorem. *Suppose θ^* is the value in Θ that maximizes the expected log-likelihood. Then the MLE satisfies*

$$\hat{\theta} \sim AN\big\{\theta^*, I^{-1}(\theta^*)J(\theta^*)I^{-1}(\theta^*)\big\} \qquad 0 < r < 1$$

as $m \to \infty$.

Thus $\hat{\theta}$ approaches the closest member of the exponential family to the relative density. $J(\theta^*) \equiv \mathrm{Var}[S(\theta^*; R_1 = r_1, \ldots, R_m = r_m)]$ can be estimated by m times the sample covariance matrix of $G(\hat{\theta}; R_1), \ldots, G(\hat{\theta}; R_m)$ where

$$G(\theta; R = r) \equiv \left[\frac{\partial \log g_\theta(r)}{\partial \theta_k}\right] = \Big[\phi_k(r) - \mathrm{E}_\theta\big(\phi_k(R)\big)\Big].$$

Based on the expected likelihood, $g_{\theta^*}(r)$ is the unique density within the family that satisfies:

$$\mathrm{E}_{\theta^*}\big[\phi_k(R_1)\big] = \mathrm{E}_g\big[\phi_k(R_1)\big] \qquad k = 1, \ldots, K.$$

Thus we can view the MLE as the solution to these equations when the expectation on the right-hand side is replaced by the average over the relative data.

The issue of the assessment of model misspecification is a difficult one as it is tied to the issue of model uncertainty. The total uncertainty in an estimator is a combination of model uncertainty and model misspecification. Most statistical approaches typically only take into account the model uncertainty when they assess the overall uncertainty. In general complicated model families (e.g., semiparametric, local polynomial) have smaller model misspecification while simple model families (e.g., beta) have smaller model uncertainty. One approach to measuring model uncertainty is to nest a given model family in a more complicated one. Then the model uncertainty of the simple models can be assessed as the divergence of the closest member of the simple family from the closest member of the more complicated families. Of course, this is a measure relative to the more complicated family rather than to the actual relative density.

The analysis of model misspecification described in this section assumes that the basis functions are prespecified. However most exponential family models choose the basis functions in a data-adaptive manner. This effectively increases the size of the modeling family in a way that improves its flexibility, but makes it difficult to quantify its complexity. When approaches to density estimation involve flexible parametric models, exploratory data analysis, model building, and diagnostics, labels such as "parametric," "nonparametric," and "data adaptive" are less meaningful.

The asymptotic normality results (9.2) and (9.3) can be extended to the estimates of the relative CDF ($G_{\hat{\theta}}$) and quantile function ($G^{-1}_{\hat{\theta}}$). See Stone and Koo (1986).

Like the local regression estimates, exponential family estimates do not suffer from boundary bias. In addition, if the choice of basis is data-driven, then the approach adapts locally to the behavior of the density. Figure 9.3 graphs two exponential family density estimates of the Purchasing Power Parity standardized residuals. The first uses $K = 6$ basis functions. It is much more peaked than the local-linear and kernel density estimates of Figure 9.2, and smoother in the upper tail. The second uses $K = 10$ basis functions and is much more wiggly in the upper tail, but retains the sharp peak in the second and third deciles.

As a practical matter both the form and number of basis functions must be chosen. A basis that results in a cubic spline has been advocated (Kooperberg and Stone 1991, and the references therein; Stone, *et al* 1997). Stone (1990) refers to the corresponding family of densities as the *log-spline exponential family*. A cubic spline is a function that is a piecewise cubic polynomial on any subinterval defined by adjacent "knots"; has continuous second derivatives; and has a third derivative that is a step function with

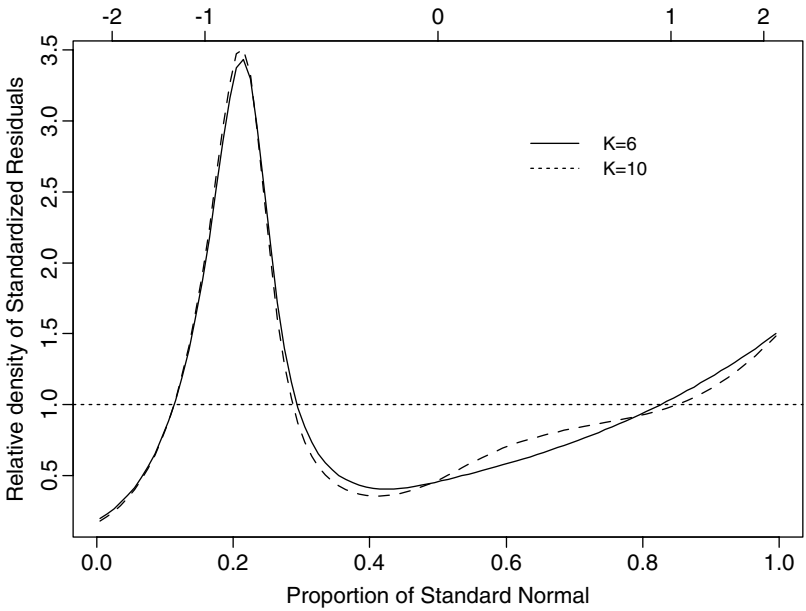

Fig. 9.3. Exponential family estimates of the relative density of standardized residuals for the Purchasing Power Parity data.

jumps at the knots. If the knots are taken to be the location of the relative data these these are called *cubic B-splines*. In general the number and location of the knots determine the basis, and there may be more appropriate choices than the data values. These choices are roughly equivalent to the choices of kernel function and bandwidth in kernel density estimation.

Stone, *et al* (1997) describe an algorithm to search through a wide range of bases and select the basis that maximizes a modification of the likelihood that penalizes more complicated models. However the statistical properties of this search process are unclear, and claims about the optimality of the final choice are unwise. Although the procedure uses reasonable statistical criteria at each stage of the process, the multiple stages obscure their overall characteristics and make them intractable to statistical analysis. In particular, the standard measures of model uncertainty (e.g., standard errors) derived above are not valid because they do not incorporate the uncertainty in the model search process that will tend to underestimate the true uncertainties.

Alternative approaches exist. Smith and Kohn (1996) develop an approach to basis selection using Bayesian ideas. George and Foster (1997) also

develop and use an empirical Bayes approach. A fuller Bayesian approach is developed by Denison, *et al* (1998). They define prior distributions for the components of the model, including the number and form of the basis functions. These papers focus on a nonparametric regression setting. The ideas need to be further developed to be directly applicable in this setting.

Another popular way to choose the number of basis functions is to modify the likelihood for fixed K to penalize more complicated models. Consider the penalized log-likelihood, which is now a function of K and $\theta_1, \ldots, \theta_K$:

$$\mathcal{L}(\theta, K; R_1 = r_1, \ldots, R_m = r_m) - \frac{1}{2} K \log(n)/n.$$

This criterion was suggested by Schwarz (1978) and can be motivated as an approximation to a Bayes procedure for model choice under a special set of priors. An alternative is to use versions of the Akaike Information Criterion (AIC) corrected for its tendency to undersmooth (Hurvich, *et al* 1998). Many other alternatives exist (see Haughton 1988 for a discussion).

The statistical efficiency of these model selection procedures and the flexibility of the log-spline model relative to approaches of the same complexity are unknown. Prime alternatives are local likelihood methods (Loader 1999), and generalizations of the local polynomial model of the previous section where the bandwidth is allowed to vary with the location in the unit interval (Simonoff 1996).

In our applications we have chosen log-spline exponential families where the number of basis functions is chosen subjectively. The procedure of Kooperberg and Stone (1992) is used to select the form of the basis functions (that is, the placement of the knots for the cubic spline within the unit interval).

For the example in Figure 9.3, their procedure chooses $K = 6$ basis functions and four knots at 0.16, 0.22, 0.28, and 0.50. The location of these knots near the peak reflect that the density is judged to be changing more quickly there. The other estimator with $K = 10$ was chosen because the automatic choice seems to oversmooth and lose much of the detail in the upper tail. Such simple sensitivity analysis is a useful tool for investigating the uncertainty due to the choice of the number of basis functions.

9.1.2.5 Orthogonal series density estimation.

Orthogonal series density estimators were suggested by Čencov (1962) and are allied with the exponential family approach. Instead of the family (9.12) consider the *orthogonal series* family of functions:

$$g_\theta(r) = g_0(r) + \sum_{k=1}^{\infty} \theta_k \phi_k(r) \qquad 0 < r < 1 \tag{9.13}$$

where $\theta_k \in \mathbb{R}$. Again $g_0(r)$ is a reference distribution, often taken to be the uniform density on $[0,1]$. Here $\{\phi_k(r)\}_{k=1}^{\infty}$ form a complete orthonormal basis for the space of all square integrable functions on $[0,1]$. By orthonormal we mean that

$$\int_0^1 \phi_i(r)\phi_j(r)dr = \mathcal{I}(j=k).$$

Suppose the $h(r)$ is a square integrable function on $[0,1]$, that is,

$$|h|^2 \equiv \int_0^1 h^2(r)dr < \infty.$$

The basis is complete if for all $h(r)$ there exists a sequence of constants $\{\theta_k\}_{k=1}^{\infty}$ such that

$$|h(r) - \sum_{k=1}^{K} \theta_k\phi_k(r)|^2 \to 0 \quad \text{as} \quad K \to \infty.$$

Thus a square integrable density can be represented precisely as an element of this family. We can write

$$g(r) = g_0(r) + \sum_{k=1}^{\infty} \theta_k\phi_k(r) \qquad 0 < r < 1$$

so that

$$\theta_k = \int_0^1 \phi_k(r)g(r)dr \quad = \quad \mathrm{E}\big(\phi_k(R)\big) \qquad k = 1,2,\ldots. \tag{9.15}$$

The natural estimates of θ_k use G_m in place of G :

$$\hat{\theta}_k = \frac{1}{m}\sum_{j=1}^{m} \phi_k(R_j) \qquad k = 1,2,\ldots.$$

Many possible choices for the basis exist – see Tapia and Thompson (1978). In practice the relative density is usually estimated by

$$\hat{g}(r) = g_0(r) + \sum_{k=1}^{K} \hat{\theta}_k\phi_k(r) \qquad 0 < r < 1,$$

where the number of terms K plays the role of the smoothing parameter. The choice of K has been studied by Wahba (1981) and Hart (1985). As m increases, the number of terms should also increase, but at a slower rate. Results in these papers suggest that $K = O(m^{1/4})$ is the optimal rate for a square integrable density.

Note that the estimate may be negative for some values of r. However a truncation and normalization process such as that applied to the regression estimator in Section 9.1.2.3 can be used. On the positive side, the estimator can be reexpressed as a kernel estimator, and produces asymptotically optimal convergence (Hall 1986).

9.2 Estimation when both distributions are unknown

In most application contexts, the CDF of the reference distribution is also unknown and must be estimated from sample data. As for the comparison population, we will assume that we have a sample $Y_{01}, Y_{02}, \ldots, Y_{0n}$ that are independently and identically distributed from the population distribution F_0. We also assume that the two samples are independent. In Section 9.3 we consider the situation where the reference distribution is formed by pooling the reference and comparison groups, and we provide further comments in the Background material on situations where the two samples are dependent.

As in Section 9.1, it is natural to estimate $F_0(y)$ by the empirical distribution function of the reference sample:

$$F_{n0}(y) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{I}(Y_{0i} \leq y) \qquad -\infty < y < \infty.$$

The results of that section can be restated in terms of F_{n0} and F_0 rather than G_m and G: $F_{n0}(y)$ converges uniformly to $F_0(y)$ with probability one and is asymptotically normal. We can then consider using $F_{n0}(y)$ as a surrogate for $F_0(y)$ in the definition of the relative data and so define the *quasirelative data*

$$Q_j = F_{n0}(Y_j) \qquad j = 1, \ldots, m.$$

As $F_{n0}(y)$ should be close to $F_0(y)$, we expect $\{Q_j\}_{j=1}^m$ to be close to $\{R_j\}_{j=1}^m$. Note that the $\{Q_j\}_{j=1}^m$ are not independent as they depend on the $\{Y_{0i}\}_{i=1}^n$. However they will be close to uncorrelated as their pairwise correlation is $O(n^{-1})$. The quasirelative data have been extensively studied (Lehmann 1953; Lin and Sukhatme 1993).

9.2.1 Properties of the quasirelative data

In this section we consider the distribution of the quasirelative data. A key role is played by the rank transformation. Let

$$T_j = \sum_{i=1}^{n} \mathcal{I}(Y_{0i} \leq Y_j) + \sum_{l=1}^{m} \mathcal{I}(Y_l \leq Y_j)$$

be the rank of Y_j in the combined vector $\{Y_1, Y_2, \ldots, Y_m, Y_{01}, \ldots, Y_{0n}\}$, where for the first sum ties are broken in favor of the $\{Y_{0i}\}_{i=1}^n$. Let

$$S_j = \sum_{l=1}^{m} \mathcal{I}(Y_l \leq Y_j)$$

be the rank of Y_j in $\{Y_1, Y_2, \ldots, Y_m\}$, where ties are broken randomly. The quasirelative data can then be expressed as

$$Q_j = \frac{1}{n}(T_j \ - \ S_j) \qquad\qquad j = 1, \ldots, m.$$

By using the concepts of invariance and sufficiency it is possible to argue that all the statistical information in the two samples relevant to inference for the relative distribution is in the ordered values of the quasirelative data. For an argument in the context of nonparametric hypothesis testing, see Fraser (1957), Section 5.3. If $T_{(1)} < T_{(2)} < \cdots < T_{(m)}$ represent the ordered $T_1, T_2, \ldots, T_m$ and $Q_{(1)} \leq Q_{(2)} \leq \cdots \leq Q_{(m)}$ the ordered quasirelative data then

$$Q_{(j)} = \frac{1}{n}(T_{(j)} \ - \ j) \qquad\qquad j = 1, \ldots, m.$$

The properties of the joint distribution of $\{Q_{(j)}\}_{j=1}^{m}$ can be determined from the relative distribution alone:

$$P[Q_{(1)} = q_1, \ldots, Q_{(m)} = q_m] = \frac{1}{\binom{n+m}{m}} \mathrm{E}\left[\prod_{j=1}^{m} g(U_{(s_j)})\right], \tag{9.14}$$

where $q_1 \leq \ldots \leq q_m \in \{0, 1/n, \ldots, 1\}$. The $\{U_{(s)}\}_{s=1}^{n+m}$ are the order statistics in a sample of $n+m$ uniform $[0,1]$ variables and

$$q_j = \frac{1}{n}(s_j - j),$$

where $s_j < s_2 < \cdots < s_m$. It is important to note that the right-hand side of (9.14) only depends on the relative PDF g and not on F or F_0 separately. If the relative distribution is uniform, then all $\binom{n+m}{m}$ possible combinations of $q_1 \leq \ldots \leq q_m \in \{0, 1/n, \ldots, 1\}$ are equally likely. In particular the marginal distribution of Q_j is uniform on $\{0, 1/n, \ldots, 1\}$.

The distribution of the quasirelative data has been determined for many relative densities, most notably those corresponding to *Lehmann's alternatives* $g(r) = r^k/(k+1)$ (Lehmann 1953). Lin and Sukhatme (1993) consider many other choices for g that are important in power calculation for many nonparametric two-sample tests.

While these exact results can be used to determine the characteristics of estimates of the relative distribution, these expressions are intractable in all but the simplest choices for the relative distribution. The developments in the next section focus on the asymptotic distributions of estimates for general choices.

9.2.2 Estimation of the relative cumulative distribution function

Let $F_m(y) = \frac{1}{m}\sum_{j=1}^{m} \mathcal{I}(Y_j \leq y)$ be the empirical distribution function of Y. In this section we will derive properties of the natural estimator of $G(r)$ in (2.1):

$$G_{n,m}(r) \quad = \quad F_m(F_{n0}^{-1}(r)) \qquad\qquad 0 < r < 1.$$

Note that this can be reexpressed in terms of the empirical distribution function of the quasirelative data:

$$G_{n,m}(r) = \frac{1}{m}\sum_{j=1}^{m}\mathcal{I}(Q_j \le r) \qquad 0 < r < 1. \tag{9.16}$$

Hsieh and Turnbull (1996) show that when $g(r)$ is bounded on any subinterval of $[0,1]$, $G_{n,m}(r)$ converges to G almost surely uniformly for $0 \le r \le 1$. That is, the Kolmogorov-Smirnov distance

$$D_{n,m} = \sup_{0<r<1} |G_{n,m}(r) - G(r)|$$

converges to zero with probability one. Thus $G_{n,m}(r)$ replicates the properties of the empirical estimator when the reference distribution is known.

We now turn to the convergence of $G_{n,m}(r)$ to G in distribution. In the context of ROC curve estimation Gastwirth (1968, Theorem 3.2) proved the following regarding the asymptotic distribution of $G_{n,m}(r)$:

Theorem. *Assume that* $0 < r < 1$, *and let* $\lambda_r = F_0^{-1}(r)$. *Suppose both* $F_0(x)$ *and* $F(x)$ *possess densities* ($f_0(x)$ *and* $f(x)$, *respectively) in a neighborhood of* λ_r *and* $f_0(x), f(x)$ *are positive and continuous at* λ_r *then*

$$G_{n,m}(r) \sim AN\left\{G(r), \frac{G(r)(1-G(r))}{m} + \frac{r(1-r)g^2(r)}{n}\right\} \tag{9.17}$$

as $m \to \infty, m/n \to \kappa^2 < \infty$.

There are now two sample sizes to consider, and we assume that they increase at the same rate. If either sample size is fixed, this result does not hold. Note, however, if the reference sample size increases at a slower rate (9.4) holds, and if the comparison sample size increases at a slower rate (9.2) holds. Thus we can view the first term as the uncertainty about G when F_0 is known and the second as the uncertainty about G when F is known.

Hence we can interpret the additional term in the asymptotic variance for $G_{n,m}(r)$ compared to $G_m(r)$ as the price we pay for using F_{n0} as a surrogate for the unknown F_0. Note that the increase in uncertainty is the uncertainty for estimating G via a sample from the reference distribution when F is known (that is, equation (9.4)). It is proportional to the square of the relative PDF, and depends only on the size of the reference sample and the relative distribution. It does not depend on the reference or comparison distributions individually.

Cwik and Mielniczuk (1990) develop an estimator of the relative CDF based on integrating an estimate of the relative density. They show the uniform strong consistency of their estimate.

It is possible to develop the statistical properties of $G_{n,m}(r)$ by viewing it as an estimated empirical process (Shorack and Wellner 1986, Section 5.5) and generalizing the results of Durbin (1973). In this context, the above result follows under slightly stronger conditions from a result of Hsieh (1995) based on a strong approximation of the empirical ROC curve (his Lemma 3). Hsieh and Turnbull (1996) generalize these results, using an empirical process approach to show that

$$\sqrt{m}\{G_{n,m}(r)-G(r)\} = B_1^{(m)}(G(r))+\lambda g(r)B_2^{(m)}(r)+o(m^{\frac{1}{2}}\log^2 m) \quad (9.18)$$

almost surely uniformly on any subinterval of $[0,1]$. Here $\{B_1^{(m)}(r) : 0 \leq r \leq 1\}$ and $\{B_2^{(m)}(r) : 0 \leq r \leq 1\}$ are two independent Brownian bridges. The result (9.17) follows from this stronger result.

The precise relationship between the relative CDF and ROC curves is discussed in Li, (1996). They *et al* extend the above result to the situation where the data from both comparison and references samples are possibly censored.

In Appendix D we give a direct proof of (9.17) based on classical U-statistic methodology.

9.2.2.1 The Asymptotic joint distribution of $G_{n,m}(r)$ and $G_{n,m}(s)$.

In most applications, including those in this book, we need to estimate the relative CDF at more than one quantile. As the estimates are based on the same data, they will be correlated. It is important to model the joint distribution at different quantiles. The result of Hsieh and Turnbull (1996) can be used to extend (9.17) to cover the joint distribution at a fixed number of quantiles. The result for two points is:

Theorem. *Assume that* $0 < r \leq s < 1$, *and let* $\lambda_\nu = F_0^{-1}(\nu)$ *for* $0 < \nu < 1$. *Suppose both* $F_0(x)$ *and* $F(x)$ *possess densities* ($f_0(x)$ *and* $f(x)$, *respectively) in neighborhoods of* λ_r *and* λ_s. *Assume that the densities are positive and continuous at* λ_r *and* λ_s *then the joint distribution of* $G_{n,m}(r)$ *and* $G_{n,m}(s)$ *is asymptotically normal:*

$$\begin{pmatrix} G_{n,m}(r) \\ G_{n,m}(s) \end{pmatrix} \quad \sim \quad AN\left\{ \begin{pmatrix} G(r) \\ G(s) \end{pmatrix}, \Sigma \right\}$$

where

$$\Sigma = \begin{pmatrix} \frac{G(r)(1-G(r))}{m} + \frac{r(1-r)g^2(r)}{n} & \frac{G(r)(1-G(s))}{m} + \frac{r(1-s)g(r)g(s)}{n} \\ \frac{G(r)(1-G(s))}{m} + \frac{r(1-s)g(r)g(s)}{n} & \frac{G(s)(1-G(s))}{m} + \frac{s(1-s)g^2(s)}{n} \end{pmatrix}$$

as $m \to \infty, m/n \to \kappa^2 < \infty$.

In the special case that F and F_0 are identical, $G_{n,m}(r), 0 < r < 1$, is a Brownian Bridge with drift r and scale $1/n + 1/m$. The results in this section will be used to calculate simultaneous confidence bands for $G(r)$ based on $G_{n,m}(r)$ (Section 9.6).

9.2.3 Estimation of the relative probability density function

In this section we will consider estimating the relative density $g(r)$ based on samples from both the reference and comparison distributions. The situation is similar to that for the relative CDF. Our approach will be to apply the estimators developed for the relative data in Section 9.1.2 to the quasirelative data. The motivation and application of these estimators will be the same, but their statistical properties will be different as the reference distribution is estimated, rather than known. Here we will describe the estimators and their statistical properties.

9.2.3.1 Estimation using a histogram.

The histogram estimator with K equisized intervals is

$$\hat{g}_{n,m}(r) = \frac{G_{n,m}(b_{j+1}) - G_m(b_j)}{h}, \qquad x \in (b_j, b_{j+1}],$$

where $b_j = (j-1)/K,\ j = 1, \ldots, K+1$ and $(b_j, b_{j+1}]$ defines the boundaries of the jth interval.

Mielniczuk (1992) shows this estimator is close to the estimator in (9.5) by developing an inequality for the supremum of the distance between them. He then observes that when $m = n$ and relative density is bounded and differentiable

$$\sup_{0 \le r \le 1} |\hat{g}_{n,m}(r) - g(r)| = O_p(\frac{1}{h}\sqrt{\log n/n} + h).$$

See Silverman (1986) or Wand (1995) for details.

9.2.3.2 Kernel density estimation.

Motivated by (9.9), we consider the following estimator of $g(r)$

$$g_{n,m}(r) = \frac{1}{mh_m} \sum_{j=1}^{m} K\left(\frac{r - Q_j}{h_m}\right) \tag{9.19}$$

Cwik and Mielniczuk (1989) prove that if the bandwidth $h_m \to 0$ in the right way as $m, n \to \infty$, $g_{n,m}(r)$ converges to $g(r)$ almost surely uniformly for $0 \le r \le 1$. That is, the global deviation $\sup_{0 \le r \le 1} |g_{n,m}(r) - g(r)|$ converges to zero with probability one. Li, *et al* (1996) prove a similar result.

The basic asymptotic distributional property of the estimator is described in the following result:

Theorem. *Assume that* $0 < r < 1$, *and suppose both* $F_0(x)$ *and* $F(x)$ *possess densities* ($f_0(x)$ *and* $f(x)$, *respectively) that are smooth (enough so that* g *is uniformly continuous). Let* $K(\cdot)$ *be a twice continuously differentiable kernel function (satisfying (9.10)) and vanishing outside some bounded interval.*

For each bandwidth sequence $\{h_m\}$ *with* $h_m \to 0$ *with* $mh_m^3 \to \infty$, $mh_m^5 \to 0, m/n \to \kappa^2 < \infty$ *we then have*

$$g_{n,m}(r) \sim AN\left\{g(r), \frac{g(r)R(K)}{mh_m} + \frac{g^2(r)R(K)}{nh_m}\right\}. \tag{9.20}$$

We remark that the result holds if $f_0(x)$ and $f(x)$ satisfy the conditions of the theorem in Section 9.1.2.2 and are bounded and uniformly continuous. It is informative to compare the properties of this estimator to those of the estimator (9.9) based on direct observation of $R_1, R_2, \ldots, R_m$. We can interpret the additional term in the asymptotic variance for $g_{n,m}(r)$ compared to $g_m(r)$ as the price we pay for using F_{n0} as a surrogate for the unknown F_0.

Alexander (1989) develops a kernel density estimator for the relative density of Y to a pooled reference group formed by merging the comparison and reference groups. If λ is the proportion of the pooled reference group in the comparison group then the CDF of the pooled reference group is $H(y) = \lambda F(y) + (1-\lambda)F_0(y)$. He proposes a Gasser-Müller boundary kernel be used to overcome the bias of the estimator close to 0 and 1. He shows that it is a consistent estimator and is asymptotically normal when the relative distribution is uniform.

The above result only holds for r in the interior of $[0, 1]$. Cwik and Mielniczuk (1993) develop a boundary kernel estimator that uses the method of reflection to overcome the bias the above estimator will have at 0 and at 1 when the relative density is nonzero there. Their estimator is

$$gb_{n,m}(r) = \frac{1}{mh_m}\sum_{j=1}^{m} K\left(\frac{r - Q_j}{h_m} + \frac{r + Q_j}{h_m} + \frac{r + Q_j - 2}{h_m}\right)$$

The additional two terms "reflect" the data outside the two boundaries to ensure smoothness there. They give a sketch of the proof of this result.

Theorem. *Assume that* $0 < r < 1$ *and* $g(r)$ *is differentiable on* $[0, 1]$. *Let* $K(\cdot)$ *be a three times continuously differentiable kernel function satisfying* (9.10) *and vanishing outside some bounded interval. For each bandwidth sequence* $\{h_m\}$ *with* $h_m^{1/2}\log(m) \to 0$ *with* $mh_m^{5/2} \to \infty$, $m/n \to \kappa^2 < \infty$ *we then have*

$$g_{n,m}(r) \sim AN\left\{\mathrm{E}(g_m(r)), \frac{g(r)R(K)}{mh_m} + \frac{g^2(r)R(K)}{nh_m}\right\}. \tag{9.21}$$

Here $g_m(r)$ is the kernel density estimator in Section 9.1.2.2 based on the relative data so that

$$\text{Bias}[g_{n,m}(r)] = \frac{1}{2}h_m^2\sigma_K^2 g''(r) + O(h_m^4).$$

Thus the estimator has the same order of bias as the estimator based on the relative data.

In Appendix D, we give a proof for the result (9.20) that exploits the structural properties of the relative density. This allows us to nicely use theory for U-statistics with estimated parameters and empirical process ideas.

Cwik and Mielniczuk (1993) also consider how to choose the bandwidth h_m. If we knew the reference distribution then (9.12) would be an appropriate choice to minimize the AMISE. However the MISE is inflated by the second term on the variance (9.20). Cwik and Mielniczuk (1993) show that a better choice of bandwidth is

$$h_{0Q} = [1 + R(g)]^{1/5} h_{0R}.$$

To estimate this, they suggest using the plug-in method of Silverman (1986) or the method of Sheather and Jones (1991). We use the latter in this book. Also note that this bandwidth choice is not consistent with the bandwidth chosen in result (9.12). There a nonasymptotically optimal bandwidth is deliberately chosen so that the squared bias of the estimator is of smaller order than the variance term.

Simulation results indicate that the asymptotic variance expression used in this result is a poor approximation to the finite-sample variance of the estimator when $g(r)$ is not smooth (i.e., $g'''(r)$ has large magnitude). In this case the other terms in the expansion (A.3) for $g_{n,m}(r)$ play a significant role (even though they are asymptotically negligible). By working through the proof of (9.20) it is possible to refine the variance estimate to give the following:

Theorem. *An expression for the variance of $g_{n,m}(r)$ that is more accurate when the sample sizes are small is:*

$$\text{Var}[g_{n,m}(r)] \approx \frac{g(r)R(K) - h_m g^2(r)}{mh_m} + \frac{\left(g(r)\sqrt{R(K) - h_m} + g'(r)R(K)\sqrt{h_m r(1-r)}\right)^2}{nh_m}$$

as $h_m \to 0$ with $mh_m \to \infty, m/n \to \kappa^2 < \infty$ as $m \to \infty$.

Simulation results indicate that this estimate is quite accurate, unless $n, m \leq 30$ or the relative density is very rapidly changing. This expression can be used with the normal approximation in the calculation of (pointwise) confidence intervals for $g(r)$ based on $g_{n,m}(r)$.

9.2.3.3 Regression-based density estimation

The regression based estimators discussed in Section 9.1.2.3 can be applied to the quasirelative data. If a local-constant version is used then it will have properties similar to those of the kernel density estimators discussed above. Local-linear versions should have better performance and not require adjustments close to the boundary. However, results about the statistical properties of regression based estimators for the quasirelative data have yet to be proven. The arguments in Section 9.1.2.3 and simulation results suggest that their real-world performance is at least as good as the kernel estimators, and they are just as easy to apply. Most of the density estimates in this book use a Poisson local-quadratic regression estimator applied to the quasirelative data with a bandwidth chose guided by the correct AIC (Hurvich, *et al* 1998).

9.2.3.4 Exponential family based density estimation

When the reference distribution is unknown, we need to consider a model for the joint distribution of both samples. If we retain model (9.12) for the relative distribution then the log-likelihood is:

$$\begin{aligned}
&\mathcal{L}(\theta, F_0; \{Y_j\}_{j=1}^m, \{Y_{0i}\}_{i=1}^n) \\
\equiv &\log\Big(P(Y_1 = y_1, \ldots, Y_m = Y_m, iY_{01} = y_{01}, \ldots, Y_{0n} = y_{0n},) \Big) \\
= &\log\Big(P(R_1 = F_0(y_1), \ldots, R_m = F_0(Y_m), Y_{01} = y_{01}, \ldots, Y_{0n} = y_{0n},) \Big) \\
= &\sum_{j=1}^{m} \log g_\theta(F_0(y_j)) + \sum_{i=1}^{n} \log f_0(y_{0i}).
\end{aligned}$$

In this form, F_0 is a nuisance parameter, as our focus is on the estimation of g_θ. One approach is to specify an exponential family model for F_0 similar to (9.12) but with the real line for support. Then the above log-likelihood could be maximized with respect to the parameters from both models. As an alternative, note that F_{n0} is, technically, a nonparametric maximum likelihood estimator for F_0. This will remove the dependence on F_0 at the cost of ignoring some information. Ignoring constants that do not influence θ, the log-likelihood reduces to

$$\sum_{j=1}^{m} \log g_\theta(F_{n0}(y_j)) \sum_{j=1}^{m} \log g_\theta(Q_j).$$

This can be seen as the likelihood that would result if we use the quasirelative data in place of the relative data in the exponential family model of Section 9.1.2.4. We can also motivate it as a pseudolikelihood estimator based on the quasirelative data that ignores the dependence within the

quasirelative data induced by the reference sample. The estimates can then be determined using the procedures outlined in Section 9.1.2.4. Note, however, that the standard errors given in that section will not be exact as they presume a model which is not exactly correct.

A third alternative is to consider the exact likelihood for the quasirelative data based on (9.14). This is computationally more intensive, but avoids the direct approximations of the other methods.

The effect of the dependence on the quasirelative data decreases as the reference sample size increases. In practice, the estimator based on the quasirelative data performs quite well. Note however that issues such as the choice of the exponential family class have not been theoretically explored.

9.2.3.5 Orthogonal series density estimation

Orthogonal series density estimates can be easily applied to the two-sample situation. As for other methods, the natural estimator is obtained by replacing the relative data by the quasirelative data. If this is applied to (9.15) the estimator of θ_k is:

$$\hat{\theta}_k = \frac{1}{m}\sum_{j=1}^{m}\phi_k(Q_j) \qquad k = 1, 2, \ldots .$$

Eubank *et al* (1987) consider estimation of the relative density using an orthogonal series estimator when the reference distribution is unknown. They note that the coefficients can be reexpressed as

$$\theta_k = \int_{-\infty}^{\infty}\phi_i(F_0(y))f(x)dx \qquad k = 1, 2, \ldots .$$

Thus if estimators of F and F_0 exist they can be substituted in to estimate θ_k and hence $g(r)$. Under general conditions Eubank, *et al* prove that the estimators are asymptotically unbiased and normal. They also give expressions for the asymptotic variance. This work is primarily interested in estimating the relative density for use in testing hypotheses about F and F_0. We return to this topic in Chapter 10.

9.3 Estimation for a pooled reference group

In some application contexts it is better to form the reference distribution by pooling the reference and comparison groups. As discussed in Section 2.4, this may be the best approach when the group to total comparison is of specific interest, when one of the groups is too small to support distributional estimates, or when the individual distributions of the two groups

are nearly disjoint. This approach has been studied extensively by Parzen and his students.

If λ is the proportion of the pooled reference group in the comparison group then the CDF of the pooled reference group is $H(y) = \lambda F(y) + (1 - \lambda)F_0(y)$. Interest then focuses on the relative distribution of F to H. Parzen (1977; 1992) refers to the corresponding relative distribution as the *(pooled) comparison distribution*. Denote the relative CDF of F to H by $\mathrm{GP}(r)$ and that of F_0 to H by $\mathrm{GQ}(r)$. Denote the corresponding densities by $\mathrm{gp}(r)$ and $\mathrm{gq}(r)$, respectively. If the comparison and reference distributions coincide both $\mathrm{gp}(r)$ and $\mathrm{gq}(r)$ will correspond to uniform distributions.

We usually assume that λ is known. A typical source would be census data. Alternatively we may consider the situation where the sample sizes m and n reflect the population sizes, so that $\lambda_{n,m} = m/(m+n)$ approaches λ as the samples sizes increase.

Based on independent samples from both reference and comparison groups, the natural estimator of $H(y)$ is

$$H_{n,m}(y) = \lambda_{n,m} F_m(y) + (1 - \lambda_{n,m}) F_{n0}(y).$$

Note, however, that this estimator is clearly correlated with $F_m(y)$, the estimator for the comparison group. Thus although we can use $F_n(y)$ and $H_{n,m}(y)$ in each of the estimators discussed in this chapter, the results describing their statistical properties will need to be reevaluated.

Define the *pooled quasirelative data*

$$P_j = H_{n,m}(Y_j) = \frac{1}{n+m} T_j \qquad j = 1, \ldots, m.$$

where T_j is the rank of Y_j in $\{Y_1, Y_2, \ldots, Y_m, Y_{01}, \ldots, Y_{0n}\}$, the combined vector where for the first sum ties are broken in favor of the $\{Y_{0i}\}_{i=1}^n$ (Section 9.2.1). Thus the pooled quasirelative data are directly related to the quasirelative data through the relationship

$$P_{(j)} = \frac{1}{n+m}(nQ_{(j)} - j) \qquad j = 1, \ldots, m.$$

The joint distribution of $\{P_{(j)}\}_{j=1}^m$ can be derived from (9.14) :

$$P[P_{(1)} = p_1, \ldots, P_{(m)} = p_m] = \frac{1}{\binom{n+m}{m}} \mathrm{E}\left[\prod_{j=1}^m g(U_{(s_j)})\right].$$

where $p_j = s_j/(n+m)$ and the $\{U_{(s)}\}_{s=1}^{n+m}$ are the order statistics in a sample of $n+m$ uniform $[0,1]$ variables with $s_j < s_2 < \cdots < s_m$. If the relative distribution is uniform then all $\binom{n+m}{m}$ possible combinations of $p_1 < \ldots < p_m \in \{1/(n+m), \ldots, 1\}$ are equally likely. In particular the marginal distribution of $P_{(j)}$ is uniform on $\{1/(n+m), \ldots, 1\}$.

Based on this relationship most of the statistical results given for the (unpooled) relative distribution can be reformulated to cover the pooled

situation. Estimators for $\mathrm{GP}(r)$ and $\mathrm{gp}(r)$ can be based on the pooled quasirelative data:

$$\mathrm{GP}_{n,m}(r) = \frac{1}{m}\sum_{j=1}^{m}\mathcal{I}(P_j \le r) \qquad 0 < r < 1$$

$$\mathrm{gp}_{n,m}(r) = \frac{1}{mh_m}\sum_{j=1}^{m} K\left(\frac{r - P_j}{h_m}\right)$$

In particular, the result (9.18) can be used to show that

$$\sqrt{m}\{\mathrm{GP}_{n,m}(r) - \mathrm{GP}(r)\} = \mathrm{gp}(r)B_1^{(m)}(\mathrm{GP}(r)) + \lambda \mathrm{gq}(r)B_2^{(m)}(\mathrm{GQ}(r)) + o(m^{\frac{1}{2}}\log^2 m)$$

almost surely uniformly on any subinterval of $[0, 1]$.

Theorem. *Under the same conditions as* (9.17)*,*

$$GP_{n,m}(r) \sim AN\left\{ GP(r), \frac{\lambda gp^2(r)\,GP(r)(1 - GP(r))}{m\kappa^2} + \frac{gq^2(r)\,GQ(r)(1 - GQ(r))}{n} \right\}$$

as $m \to \infty, m/n \to \kappa^2 < \infty$*.*

Theorem. *Under the same conditions as* (9.20)*,*

$$gp_{n,m}(r) \sim AN\left\{ gp(r), \frac{R(K)gp(r)(1 - gp(r))}{mh_m} \right\}$$

as $m \to \infty, m/n \to \kappa^2 < \infty$*.*

The first result is proved in Aly, *et al* (1987) and a sketch of the proof of the second is in Parzen (1983). Many related results concerning pooled relative distributions are given in Alexander (1989) and Parzen (1999).

9.4 Estimation when the data are censored

Suppose we wish to compare the job stability of young men and women at the start of their working lives. We conduct a survey and ask respondents the length of time that they spent working for their first full-time employer. In principle we can compare the early job stability distributions of the two groups using the relative distribution of their measured first employment spells. Some respondents may still be on the job at the time of the survey, *censoring* their employment spells. These spells are called *right censored*, because we know the start but not the end of the spell. Censoring can also take other forms: respondents may not have had a job, so the entire spell

is not observed; they may not recall the start date, or may only recall that the job started in a particular year. Here we will focus on right censoring, although the results can be extended to the other forms. See Kalbfleisch and Prentice (1980) for an introduction to these ideas.

Let S_j and E_j be the start and end dates of the first job for respondent j in the comparison group. Under right censoring S_j is always observed, but E_j will not be if it occurs after the date the respondent took the survey SD_j. The employment spell is $Y_j = E_j - S_j$ and $C_j = SD_j - S_j$ is the observed censoring time. The employment spell was observed if $\delta_j = \mathcal{I}(Y_j \leq C_j)$ is 1. The observed data for the comparison group is $\{Y_j^o, \delta_j\}_{j=1}^m$ where $Y_j^o = \min(Y_j, C_j)$ is the observed time. In general the C_j and Y_j can be dependent, but here we assume that they are independent and identically distributed from CDFs $R(y)$ and $F(y)$, respectively. Let D_i be the censoring times, Y_{0i} be the employment spells, γ_i be the censoring variable, and Y_{0i}^o be the observed times for the $i = 1, \ldots n$ person from the reference sample. The Y_{0i} and D_i are independent with CDFs $F_0(y)$ and $H^o(y)$, respectively. As before the objective is estimate the relative distribution of F to F_0.

Kaplan and Meier (1958) proposed the *product-limit estimator* of F_0 :

$$F_{n0}(y) = 1 - \prod_{Y_{0(i)}^o \leq y} \left[\frac{n-i}{n-i+1} \right]^{\gamma_{(i)}},$$

where $0 \leq Y_{0(1)}^o \leq Y_{0(2)}^o \leq \cdots Y_{0(n)}^o$ are the ordered survival times, and $\gamma_{(i)}$ is the corresponding censoring indicator. This is the generalization of the empirical CDF to the censored case. When there is no censoring this estimator reduces to the empirical CDF. If the largest observation is censored ($\gamma_{(n)} = 1$) the estimator is not a CDF, and so is redefined to be one for $y \geq Y_{0(n)}^o$. A similar product-limit estimator, $F_m(y)$, can be constructed for F.

The natural estimator of $G(r)$ is

$$G_{n,m}(r) \quad = \quad F_m(F_{n0}^{-1}(r)) \qquad\qquad 0 < r < 1,$$

Cao, *et al* (1999) note that this is just the product-limit estimator of $G(r)$ based on the *censored quasirelative data:*

$$\{F_{n0}(Y_j), \delta_j\} \qquad j = 1, \ldots, m.$$

Gastwirth and Wang (1988) obtained the generalization of the result (9.17) to this setting:

Theorem. *Under the same conditions as* (9.17),

$$G_{n,m}(r) \sim AN\left\{ G(r), \frac{\sigma^2(r; G; H)}{m} + \frac{\sigma^2(r; I; H_0) g^2(r)}{n} \right\}$$

as $m \to \infty, m/n \to \kappa^2 < \infty$. *Here* $H(r) = R(F_0^{-1}(r))$, $H_0(r) = R_0(F_0^{-1}(r))$, $I(r) = r$ *and*

$$\sigma^2(r;G;H) = (1-G(r))^2 \int_0^r \frac{g(p)dp}{(1-G(p))^2(1-H(p))}.$$

Li, *et al* (1996) prove a similar result.

Estimates of the relative density $g(r)$ can be based on the the censored quasirelative data and censored versions of the density estimates discussed in Sections 9.4. Cao, *et al* (1999) propose the kernel density estimator (9.19) and show that:

Theorem. *Under the same conditions as* (9.20)*,*

$$g_{n,m}(r) \sim AN\left\{ gp(r), \frac{R(K)g(r)}{mh_m(1-H(r))} + \frac{R(K)g^2(r)h_u(r)}{nh_mP(\delta_0=1)(1-H_0(r))^2} \right\}$$

as $m \to \infty, m/n \to \kappa^2 < \infty$*. Here* $h(r)$ *is the relative density of an uncensored to actual reference group employment spell.*

Note that when there is no censoring this reduces to (9.20).

9.5 Estimation when the data are weighted

In many circumstances the samples are obtained by probability sampling, so that observations are drawn with unequal probabilities rather than as a simple random sample. One common procedure is *stratified sampling*, used when accurate estimates are desired for small population subgroups, or when the variation of an attribute is much smaller within identifiable subgroups of a group than it is between the subgroups. For example, the employment spells of young workers tend to be shorter than those of older workers. We could sample older workers in greater proportion than their population proportion to obtain better estimates of the overall distribution. Suppose the group can be partitioned into K subgroups of size $N_1, \ldots, N_K$. The population CDF is then

$$F_0(y) = \frac{1}{N}\sum_{k=1}^{K} N_k F_k(y)$$

where $F_{0k}(y)$ is the CDF of the kth subgroup and $N = N_1 + \ldots + N_k$ is the population size. If we sample n_k individuals from the kth subgroup then the natural estimate of the population CDF is

$$F_n(y) = \frac{1}{N}\sum_{k=1}^{K} N_k F_{kn}(y) = \frac{1}{n}\sum_{j=1}^{n} w_j \mathcal{I}(Y_{0j} \le y). \tag{9.22}$$

where $F_{kn}(y)$ is the empirical CDF of the kth subgroup, Y_{0j} is the observation from the j individual, $n = n_1 + \ldots + n_k$, and $w_j = (N_k/N)/(n_k/n)$

if j is in subgroup k. Each observation has associated with it w_j, a *sample weight* representing the population prevalence relative to the sample prevalence. As long as an appropriate sample frame has been established, w_j can be controlled by survey design. Under simple random sampling $w_j = 1$. See Thompson (1992) for an extensive treatment of approaches to sampling.

To estimate the relative distribution based on sampling with weights the weighted empirical distribution function can be used in place of the usual empirical distribution function. In particular, the *weighted quasirelative data* becomes

$$Q_j = F_{n0}(Y_j) \qquad j = 1, \ldots, m,$$

where F_{n0} is the weighted empirical distribution function in (9.22). Note that the weighted quasirelative data has the sample weights of the comparison sample, if they were used. Hence estimates of the relative CDF and PDF can be determined using versions of each of the methods described in Section 9.6 for weighted data. These are straightforward and described in the references given for each method.

9.6 Confidence intervals and confidence bands

In this section we will consider pointwise confidence intervals and simultaneous confidence sets for $G(r)$ and $g(r)$ for $0 \le r \le 1$. These sets can be constructed directly from the results of the previous sections.

If the sample size is not small (i.e., $n, m > 30$), we can use the normal approximation to the exact distributions of the estimator as the basis for the intervals. If the sample sizes are small we can use the bootstrap to determine the sampling distribution of the estimate and the corresponding critical values. Here we will discuss approximations to those critical values based on the normal approximation in moderate to large samples. The sample sizes for the referenced applications and the one considered in this book tend to be large (e.g., 1300–3000), and the approximations will be very close to the exact values.

9.6.1 Confidence intervals and confidence sets for $G(r)$

If the sample size is not small, we can use the normal approximation to produce the distribution of the estimate as the basis for a test for a given significance level α :

$$P\left(|G_{n,m}(r) - G(r)| \le z_{\alpha/2} \times \sqrt{\widehat{\mathrm{Var}}\left[G_{n,m}(r)\right]} \right) \to 1 - \alpha$$

as $m \to \infty, m/n \to \kappa^2 < \infty$. Here $\sqrt{\widehat{\mathrm{Var}}\left[G_{n,m}(r)\right]}$ is the variance estimate obtained by replacing the relative CDF and density by their estimates in the variance expression of (9.17).

It is also of interest to determine confidence bands for $G(r)$ for all $0 \leq r \leq 1$. While Kolmogorov-Smirnov type bounds are possible, it is more appealing to use a band width at a given point proportional to the estimated standard deviation at that point. These have been called "equal-precision" bands by Nair (1984), who applied them to survival functions and Burr and Doss (1993) who applied them to the median survival in the Cox proportional hazards model. To this end, consider the process:

$$G^s(r) = \frac{G(r)}{\sqrt{\frac{G(r)(1-G)(r))}{m} + \frac{r(1-r)g^2(r)}{n}}}.$$

We then have:

Theorem. *Under the same conditions as* (9.17)*, and with* $0 < r \leq t < 1$ *then*

$$\begin{pmatrix} G^s_{n,m}(r) \\ G^s_{n,m}(t) \end{pmatrix} \quad \sim \quad AN\left\{ \begin{pmatrix} G^s(r) \\ G^s(t) \end{pmatrix}, \Sigma^s \right\}$$

where

$$G^s_{n,m}(r) = \frac{G_{n,m}(r)}{\sqrt{\frac{G_{n,m}(r)(1-G_{n,m}(r))}{m} + \frac{r(1-r)g^2_{n,m}(r)}{n}}}$$

and Σ^s *is the correlation matrix corresponding to* Σ *in* (9.17).

The confidence bands can then be calculated by simulating the stochastic process in the theorem where the parameters in the covariance structure are replaced by their estimates. A value $L^{(n)}_\alpha$ can be estimated from multiple simulations that satisfies:

Theorem. *Under the same conditions as* (9.17)*, there exists a* $L^{(n)}_\alpha$ *such that*

$$P\left(\left| \frac{G_{n,m}(r) - G(r)}{\sqrt{\frac{G_{n,m}(r)(1-G_{n,m}(r))}{m} + \frac{r(1-r)g^2_{n,m}(r)}{n}}} \right| \leq L^{(n)}_\alpha, \quad 0 < r < 1 \right) \to 1 - \alpha$$

as $m \to \infty, m/n \to \kappa^2 < \infty$*. Thus the band*

$$G_{n,m}(r) \pm L^{(n)}_\alpha \sqrt{\frac{G_{n,m}(r)(1 - G_{n,m}(r))}{m} + \frac{r(1-r)g^2_{n,m}(r)}{n}}$$

has asymptotic coverage probability $1 - \alpha$.

9.6.2 Confidence intervals and confidence sets for $g(r)$

The approach of the previous section can be applied to $g(r)$ in a very similar fashion. The pointwise bands can be based on (9.20). The confidence intervals in this book use this approach. The simultaneous bands result requires the extension of (9.20) to the joint distribution, which will not be given here.

Background material

As noted in Chapter 2, Kelly (1994) provides a readable introduction to the probability theory underlying the methods in this book. Rice (1995) is a useful source for the mathematical statistics required. Serfling (1980) goes into much greater depth than these two references.

The notation $N(\mu, \sigma^2)$ is used to denote a normal (or Gaussian) distribution with mean μ and variance σ^2. The standard normal distribution discussed in Section 2.1 is $N(0, 1)$ and the corresponding CDF is often denoted by $\Phi(x), -\infty < x < \infty$.

The description of asymptotic properties in Section 9.1.1 was very brief. More formally, consider a sequence of random variables $X_1, X_2, \ldots$ where the mth random variable has CDF $F_m(x)$. Suppose X has CDF $H(x)$. We say that the X_m *converges in distribution* to X if, for each continuity point of $H(x)$,

$$\lim_{m\to\infty} F_m(x) = H(x).$$

This concept measures a sense in which the X_m are "cross-sectionally" close to X when the sample size is large. It does not focus on how close a particular sequence of X_m is to X, only the aggregate. We say that the X_m *converges with probability one* to X if,

$$P\left(\lim_{m\to\infty} X_m = X\right) = 1.$$

This concept measures a sense in which the X_m are "longitudinally" close to X when the sample size is large. If a sequence converges with probability one, then it also converges in distribution.

We say the sequence is *asymptotically normal* with "mean" μ_m and "variance" $\sigma_m^2 > 0$ if

$$\frac{X_m - \mu_m}{\sigma_m}$$

converges in distribution to a standard normal distribution. In this situation $H(x) = \Phi(x)$ and so is continuous for each $-\infty < x < \infty$. For additional information see Kelly (1994) or Serfling (1980).

Standard notation concerning the convergence properties of sequences is as follows:

(1) *Deterministic sequences:* Let x_n and y_n be two real-valued deterministic (nonrandom) sequences. Then, as $n \to \infty$,
   (a) $x_n = O(y_n)$ if and only if $\limsup_{n\to\infty} |x_n/y_n| < \infty$,
   (b) $x_n = o(y_n)$ if and only if $\lim_{n\to\infty} |x_n/y_n| = 0$.

(2) *Random sequences:* Let X_n and Y_n be two real-valued random sequences. Then, as $n \to \infty$,
   (a) $X_n = O_p(Y_n)$ if and only if for all $\epsilon > 0$, there exist δ and N such that $P(|X_n/Y_n| > \delta) < \epsilon$, for all $n > N$,
   (b) $X_n = o_p(Y_n)$ if and only if for all $\epsilon > 0$, $\lim_{n\to\infty} P(|X_n/Y_n| > \epsilon) = 0$.

Simonoff (1996) has a discussion of the use of this notation. Throughout most of this chapter we have assumed that the information on the reference and comparison distributions are independent. A natural situation where they are dependent is where they are both observations on the same entity. For example, they could be earnings of the same individual at two points in time where interest focuses on the relative earnings between the two time points.

One viewpoint is to try to understand the bivariate distribution of the comparison and reference value. In the spirit of the relative distribution approach, the copula can be used as a summary of the dependence between the two components in a manner independent of the marginal distributions of the comparison and reference values. Joe (1997) gives a book length treatment of this approach, and generalizes it to arbitrary multivariate distributions. For approaches to testing independence, see also Kallenberg and Ledwina (1999). Another viewpoint, closer to that of this book, is that the relative distribution is of primary interest and the measuring of the dependence is of secondary interest. Most of the methods developed in this book can still be applied in this situation, although their statistical properties will be modestly altered. In general the effect will be to decrease the variance of the estimates and lead to underestimation of uncertainty of the estimates. This area has received little attention in the existing research literature.

Computational issues

Most of the estimation methods described in this chapter applied standard univariate distributional estimators to the quasirelative data. The quasirelative data can be determined from the rank statistics described in Section 9.2.1. We have used S-PLUS for all of the computations and figures presented in this book. SAS code is also available, from the WWW site for this book.

Virtually any statistical package can produce fixed bin-width histograms. Increasingly many contain other nonparametric CDF and density

estimation routines. In particular, SAS Version 7.0 contains routines for kernel density estimation and local polynomial regression that can be used to implement the estimators in Sections 9.2.3.2 and 9.2.3.3, respectively.

Venables and Ripley (1997) gave S-PLUS code to calculate the Sheather-Jones (1991) bandwidth for kernel density estimators.

The collection `log-spline` in the `S` directory of `statlib` contains S-PLUS functions that calculate log-spline density estimates based on methods described in Kooperberg and Stone (1991).

Code for the relative density and CDF for standard packages such as SPSS, MINITAB, SAS, and S-PLUS that uses these facilities is directly available from the WWW site for this book. Many additional references are given by Simonoff (1996).

Exercises

Exercise 9.1. When the reference distribution is known, the formulae in Section 2.2 can be used to convert any estimator of F and f into an estimator of G or g. From an estimation viewpoint only, give three reasons why the direct estimation of G or g is preferred. As an example, consider the case where both the reference and comparison distributions are very right-skewed in a similar manner.

Exercise 9.2. Show that the distribution of $G_m(r)$ is given by

$$P\left[G_m(r) = \frac{k}{m}\right] = \binom{m}{k}[G(r)]^k[1-G(r)]^{m-k} \quad k = 1, \ldots, m.$$

Hence show that $\mathrm{E}[G_m(r)] = G(r)$ and $\mathrm{Var}(G_m(r)) = G(r)[1-G(r)]/m$.

Exercise 9.3. Show that the distribution of the histogram estimator of the density is given by

$$P\left[\hat{g}(r) = \frac{k}{m}\right] = \binom{m}{k}[H(r)]^k[1-H(r)]^{m-k} \quad k = 1, \ldots, m,$$

where $r \in (b_j, b_{j+1})$ and $H(r) = G(b_{j+1}) - G(b_j)$. Hence derive the expressions for the bias and variance of the estimator given in (9.6) and (9.7).

Exercise 9.4. Use the definition of the quasirelative data in Section 9.2.1 to show that

$$G_{n,m}(r) = \frac{1}{m}\sum_{j=1}^{m} \mathcal{I}(Q_j \le r) \qquad 0 < r < 1.$$

Exercise 9.5. Give two examples of distributions that have supports on $[0, 1]$ and are exponential families of the form (9.12).

Exercise 9.6. Consider the beta family of distributions given in Section 9.1.2.4. Suppose g is the member with $\theta = (1, 1)$. Consider the subfamily with $\Theta_S = \{(\theta_1, \theta_2) > 0 : \theta_2 = \theta_1 + 1\}$. Calculate the Kullback-Leibler divergence between g and g_θ, $\theta \in \Theta_S$. Suppose $\hat{\theta} = (0.5, 1.5)$. Calculate the model misspecification and model uncertainty for this case. Determine θ^*, the value in Θ_S that maximizes the expected log-likelihood.

Exercise 9.7. What are the advantages of the exponential family formulation in Section 9.1.2.4 over the orthogonal series formulation in Section 9.1.2.5? Can you think of some disadvantages?

Exercise 9.8. Prove the result (9.14).

Exercise 9.9. Derive the expression (9.14) for Lehmann's alternatives. You may find Lehmann (1953) useful.

Exercise 9.10. Discuss the statistical properties of the weighted CDF estimator $F_n(y)$ given in (9.22). In doing so, derive a result similar to (9.2) for this estimator.

Chapter 10

Inference for Summary Measures

In this chapter we develop estimators for summary measures based on the relative distribution. The motivation for these measures is given in Chapter 5.

The first two sections present general results for summary measures of distributional difference, based on properties of estimators of the relative density. In Section 10.3 we introduce estimators of median relative polarization index and give results on its asymptotic statistical properties. We show that the estimator is asymptotically normal, and give refined results when the comparison and reference distributions are equal. The computation of these asymptotic distributions requires estimates of the covariance matrices. These are given in Section 10.4. In Section 10.5, we turn to the lower and upper polarization indices, and give properties of the natural estimators of these indices. In practice, researchers are often concerned with testing whether the polarization indices are zero against a complementary alternative. In Section 10.6, we address this issue by developing confidence intervals and confidence bands for the indices. In Section 10.7, we give a general algorithm to determine confidence intervals for summary measures based on the bootstrap. The more detailed results and proofs are given in Appendix E.

10.1 Inference for two measures of distributional difference

In Chapter 9 we considered estimators for the relative CDF and PDF. As the summary measures are functions of these, one procedure to generate the estimators is to replace occurrences of the relative CDF and PDF with their estimators. For example, consider the directed divergence measures given in Section 5.2. If $\hat{g}(r)$ is an estimator of the relative PDF then we can estimate these measures by:

$$\hat{D}_{\phi}(F;F_0) = \int_0^1 \phi\big(\hat{g}(r)\big)\ dr.$$

However, the statistical properties of the estimators need to be determined. Results exist for the two measures of distributional divergence considered in Section 5.3. Mielniczuk (1990) estimates the chi-squared divergence based on $gb_{n,m}(r)$ given in Section 9.2.3.2 and $G_{n,m}$ in Section 9.2.2. He shows that if g has a bounded second derivative the estimator is asymptotically normal with mean $D_\phi(F;F_0)$.

Many estimators have been proposed for the Kullback-Leibler divergence. As the Kullback-Leibler divergence is the entropy of the relative distribution, most estimators can be written in the form:

$$D(F;F_0) \equiv D(g) = \int_0^1 \log\big(\hat{g}(r)\big)\hat{g}(r)\ dr,$$

where $\hat{g}(r)$ is an estimate of the relative density (See Section 9.2.3). Examples of estimators based on estimating g directly include kernel based estimators (Joe 1989), log-density based estimators (Barron, *et al* 1992) and other nonparametric estimators (Ebrahimi, *et al* 1994). In particular, Mielniczuk (1992) proves that if $D(F;F_0)$ is finite and g is bounded then $D(\hat{g})$ is strongly consistent for $D(g)$ when $\hat{g}$ is the histogram based estimator of g given in Section 9.2.3.1. In Section 10.7 we discuss bootstrap measures of uncertainty and confidence intervals applicable to general measures, and $D(F;F_0)$ in particular.

10.2 Measures motivated by hypothesis testing

In Section 5.5 we introduced some divergence measures motivated by testing the null hypothesis of equality of the comparison and reference distributions when both distributions were unknown, and samples from each were available.

One approach to testing is to consider a measure of divergence between the two distributions $D(F;F_0)$ that is sensitive to the deviations from the null hypothesis specified by the alternative hypothesis. The test can be conducted by estimating the measure based on sample information and comparing the estimate to the values we would expect from it if the null hypothesis of equality were true. This approach is advocated by Parzen (1979), Eubank, *et al* (1987), and Eubank and LaRicca (1992). They proceed by estimating the relative density and comparing a measure of its divergence from the uniform distribution as a test statistic.

We will examine this and a number of variants of this approach in the next section.

10.2.1 Measures based on linear rank statistics

Consider the divergence measures proposed by Chernoff and Savage (1958) given in Section 5.5. The natural estimator of $D_{CS}(F; F_0)$ replaces the CDF of the pooled group relative distribution with the empirical version based on the pooled relative data $\{\tilde{R}_j\}_{j=1}^m$:

$$\hat{D}_{CS}(F; F_0) = \sum_{j=1}^{m} J(\tilde{R}_j).$$

We can use the estimator for the pooled reference group because a measure based on the pooled distribution is implicitly a measure of the differences between the two separate distributions. The estimator above is an example of a *linear rank statistic* (Serfling 1980). Many test statistics can be expressed in this form. Alexander (1989) gives the following table:

Table 10.1. Common test statistics that can be expressed as linear rank statistics in Chernoff-Savage form.

Name	Score Function
Tests for Location Alternatives	
Wilcoxon	$J(p) = p$
Normal scores	$J(p) = \Phi^{-1}(p)$
Median test	$J(p) = \mathcal{I}(p < \frac{1}{2})$
Tests for Scale Alternatives	
Mood test	$J(p) = (p - \frac{1}{2})^2$
Normal scores	$J(p) = \Phi^{-1}(p)^2$
Ansari-Bradley	$J(p) = \lvert p - \frac{1}{2} \rvert$

As in Section 5.5, let gp be the relative density of F to H (the pooled CDF) and gq be the relative density of F_0 to H. Then $\tilde{R} = H(Y)$ and $\tilde{S} = H(Y_0)$ have PDFs gp and gq, respectively. Chernoff and Savage (1958) proved that

Theorem. *Suppose $m, n \to \infty$ such that $\lambda_m = n/(m+n)$ is bounded away from 0 and 1. Then*

$$\hat{D}_{CS}(F; F_0) \sim AN\left\{ D_{CS}(F; F_0),\ \lambda_m^2 \left[\frac{1}{m} \mathrm{Var}[B_1(\tilde{R})] + \frac{1}{n} \mathrm{Var}[B_2(\tilde{S})] \right] \right\}$$

where

$$B_1(r) = \int_0^r J'(p) gp(p) dp$$

$$B_2(r) = \int_0^r J'(p) gq(p) dp.$$

Alexander (1989) derives the score functions that maximize the asymptotic relative efficiency for location and scale alternatives. He also shows that the Wilcoxon test is optimal if the pooled reference distribution is logistic, and the Normal scores test is optimal for location and scale alternatives if the pooled reference distribution is normal.

10.2.2 Measures based on chi-squared divergence

Eubank, *et al* (1987), henceforth ELR, show that many hypothesis tests can be placed in a general framework based on decomposing the chi-squared divergence. Their approach uses the orthogonal series representation of g given in Section 9.2.3.5. Let $\{\phi_k(r)\}_{k=1}^{\infty}$ form a complete orthonormal basis for the space of all square integrable functions on $[0,1]$. They show that the chi-squared divergence can be expressed as

$$D_\phi(F;F_0) \quad = \quad \sum_{k=1}^{\infty} \theta_k^2,$$

where θ_k are the coefficients given in (9.15). Thus the chi-squared divergence can be decomposed into additive contributions from each of the functions in the basis. The coefficients are therefore referred to as the *components* of $\hat{D}_\phi(F;F_0)$. Each of these components measures the divergence of the comparison distribution from the reference distribution in a direction defined by the corresponding basis function. By choosing difference sets of basis functions we can define alternative decompositions of the overall chi-squared divergence.

If any of the components is nonzero the null hypothesis of equality is false. Hence ELR propose tests of subhypotheses that the individual θ_k are zero. The θ_k can be estimated by the $\hat{\theta}_k$ given in Section 9.2.3.5. ELR show that these estimates are asymptotically unbiased and normal. The hypothesis test is conducted by comparing $\hat{\theta}_k$ to its asymptotic distribution under the null hypothesis.

The power of the ELR framework is in the range of hypothesis tests that are specific cases. Let $\eta_k(x)$ be the Legendre polynomials on $[-1,1]$ and let $\phi_k(p)$ be the normalized $\eta_k(2p-1)$, $0 \le p \le 1$. The first component θ_1 is a measure of the divergence of g in the direction of the location alternatives given in Section 5.5. The estimate $\hat{\theta}_1$ given in Section 9.2.3.5 is the Wilcoxon rank sum. The second component θ_2 is a measure of the divergence of g in the direction of the scale alternatives given in Section 5.5. The estimate $\hat{\theta}_2$ given in Section 9.2.3.5 is the Mood statistic. If the basis is the Hermite polynomials then the estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ correspond to the Normal scores and Klotz statistics, respectively.

These identifications can be given an intuitive explanation. If the comparison distribution is a location shifted version of the reference distribution, then the relative density will tend to be monotone (i.e., $g'(p)$ will not

change sign). Hence if the first basis function is monotone, θ_1 will tend to capture location shifts. If the comparison distribution is a scale shifted version of the reference distribution then $g'(p)$ will tend to change sign once. Hence if the second basis function is bowl-shaped, θ_2 will tend to capture scale shifts. For example, consider location and scale decomposition in Figure 3.1. The two relative densities strongly reflect this pattern. In general higher oscillating basis functions will capture higher frequency departures of the relative density from uniformity. These in turn will tend to correspond to higher moment differences between the underlying distributions. For additional discussion of this argument see Eubank, *et al* (1987), Section 2.3 and Alexander (1989), Section 2.3.4.

ELR focus on testing if a given set of m components are zero. For example, tests of $k = 1$ and $k = 2$ correspond approximately to location and scale differences between the distributions. An alternative is to consider weighted sums of the components: $\sum_{k=0}^{\infty} \alpha_k \theta_k$ estimated by $\sum_{k=0}^{\infty} \alpha_k \hat{\theta}_k$ where $\sum_{k=0}^{\infty} \alpha_k \text{Var}(\hat{\theta}_k) < \infty$. If $\alpha_k = 1/k(k+1)$ and the Legendre polynomial basis is used, the estimate corresponds to the Anderson-Darling statistic. If $\alpha_k = 1/k^2\pi$ and the sine basis $\phi_k(p) = \sin(2\pi p)$ is used, the estimate corresponds to the Cramer-von Mises statistic. These statistics involve all components but successively down weight the higher order components. The advantage is that the power of the statistic for any given alternative approaches one as $m, n \to \infty$. The disadvantage is that the power for a given component decreases rapidly with k. Thus large sample sizes are required to reject the null if the alternative hypothesis effects only a higher order component. If we test individual components then we will have good small sample power for alternatives that effect those components, but have an inconsistent test for alternatives that do not effect those components. If we can choose a basis and the components that closely reflect the alternative hypothesis we most wish to protect against, then the compromise is a good one. Thus the ELR framework sheds light into the nature of omnibus test statistics versus specific test statistics. For an indepth analysis of this issue, and an alternative framework, see Alexander (1989).

From a practical standpoint, it is sometimes difficult to choose bases and components that are appropriate for the specific application. The ELR approach sidesteps this issue by relying instead on generic features of the chosen basis functions. In many contexts, however, it will be preferable to use decompositions directly motivated by scientific theory and substantive questions. The decompositions we have given in Chapters 3 and 7 are motivated more along these lines. They reflect emerging hypotheses about the nature and origins of the changes in the earnings distribution, and they take advantage of what these methods are very good at answering: the detailed effects of location and shape changes, and the impacts of other covariates.

10.2.3 Measures based on data-driven Neyman's tests

An approach closely allied with the approach of Eubank, *et al* (1987) is the data-driven Neyman's test developed in Ledwina (1994). For simplicity we consider the goodness-of-fit situation where F_0 is known. To test the null hypothesis that F is F_0 she approximates the relative distribution as a member of an exponential family given in Section 9.2.3.4. As a divergence measure she proposed:

$$D_K(F; F_0) = \sum_{k=1}^{K} \theta_k^2,$$

where θ_k is given in (9.12). The test statistic is the natural estimator of $D_N(F; F_0)$

$$\hat{D}_K(F; F_0) = \sum_{k=1}^{K} \bar{\theta}_k^2.$$

Neyman (1937) proposed the test using the Legendre polynomial basis given above and K fixed. Ledwina (1994) proposed to choose K in a "data-driven" manner. Instead of a single exponential family she considers a sequence of exponential families with the $k-1$th family nested in the kth, $k = 1, 2, \ldots, K$. The estimator she chooses maximizes the penalized likelihood based on Schwarz's criterion given that Section 9.2.3.4. The test statistic is then $\hat{D}_S(F; F_0)$ where S is the number of basis functions that minimizes the penalized likelihood.

Ingot and Ledwina (1996) show that this test is asymptotically optimal in the sense of intermediate efficiency (Kallenberg 1983) for a wide range of alternatives. They also show that non-data-driven tests such as the Kolmogorov-Smirnov and Cramer-von Mises tests are optimal for only a narrow class of alternatives. Ingot, *et al* (1998) show that the test is asymptotically as efficient as the most powerful Neyman-Pearson test if the level of significance tends to zero also. While these statements depend on the choice of asymptotic framework, the practical value of the approach is supported by extensive simulations studies reported in Ledwina (1994) and Ledwina (1996).

10.3 Inference for the median relative polarization

The Median Relative Polarization (MRP) is defined in Section 5.6.1. Here we focus on the estimation based on sample survey data. Let $Y_1, Y_2, \ldots, Y_m$ be independently and identically distributed from the distribution F and let $Y_{01}, Y_{02}, \ldots, Y_{0n}$ be independently and identically distributed from the distribution F_0. Assume the two samples are mutually independent. Denote the empirical distribution function of Y and Y_0 by $F_m(y) =$

$m^{-1}\sum_{j=1}^{m}\mathcal{I}(Y_j \le y)$ and $F_{n0}(y) = n^{-1}\sum_{i=1}^{n}\mathcal{I}(Y_{0i} \le y)$, respectively. Here $\mathcal{I}(\cdot)$ is the indicator function. In this section we develop the properties of $\widehat{\mathrm{MRP}}(F; F_0) \equiv \mathrm{MRP}(F_m; F_{n0})$, the natural estimator of $\mathrm{MRP}(F; F_0)$. This estimator requires an estimator of the location shift between Y and Y_0 $\rho = Q(\frac{1}{2}) - Q_0(\frac{1}{2})$. The natural estimator of ρ is $\hat{\rho} = \hat{Q}(\frac{1}{2}) - \hat{Q}_0(\frac{1}{2})$, the difference between the empirical quantiles. Some insight into the estimator can be gained by reexpressing it as:

$$\widehat{\mathrm{MRP}}(F;\ F_0) \quad = \quad \frac{4}{m}\sum_{j=1}^{m}\left|\hat{Q}_j - \frac{1}{2}\right| \quad - \quad 1. \tag{10.1}$$

where $\{\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_m\}$ are the *location matched quasirelative data*:

$$\hat{Q}_j = F_{n0}(Y_j - \hat{\rho}) \qquad j = 1, \ldots, m.$$

This definition modifies that given for the quasirelative data in Section 9.2 so that the distributions are location matched. The expression (10.1) follows directly from the definitions of the empirical distribution function and some algebra. Note that the $\{Q_j\}_{j=1}^{m}$ are not independent as they depend on the $\{Y_{0i}\}_{i=1}^{n}$. However, as we shall see in the next section, they will be close to uncorrelated (their pairwise correlation is $O(n^{-1})$).

10.3.1 The asymptotic distribution of $\widehat{\mathrm{MRP}}(F;\ F_0)$

Recall from Chapter 3 that the median-matched relative distribution of Y to Y_0 is given by $R_{0L} = F_0(Y - \rho)$, where $\rho = Q(\frac{1}{2}) - Q_0(\frac{1}{2})$, the difference between the median of Y and the median of Y_0. Q is the quantile function, defined in Section 2.2. The MRP is the expectation of $4|\ R_{0L} - \frac{1}{2}\ | - 1$, and so it is natural to expect the average based on the location matched quasirelative data to be (asymptotically) unbiased. We have the following description of the asymptotic distribution of $\widehat{\mathrm{MRP}}(F;\ F_0)$.

Theorem. *If $F_0(x)$ and $F(x)$ possess continuous densities satisfying $f(\xi_{\frac{1}{2}}) > 0$ and $f_0(\xi^0_{\frac{1}{2}}) > 0$, then*

$$\widehat{MRP}(F;\ F_0) \sim AN\left\{MRP(F;\ F_0), \sigma^2_{MRP}\right\}$$

as $m/n \to \kappa^2 < \infty, m \to \infty$. The asymptotic variance is:

$$\begin{aligned}\sigma^2_{MRP} = &\frac{16}{m}\mathrm{Var}(|\ R_{0L} - \frac{1}{2}\ |) + \frac{16}{n}\mathrm{Var}(|\tilde{R}_{0L} - \frac{1}{2}\ |) \\ &+ \frac{1}{m}\sigma^2(\delta\ (\xi)) + \frac{1}{n}\sigma_0^2(\delta_0(\xi)),\end{aligned} \tag{10.2}$$

where

$$\sigma^2(\delta\ (\xi)) = 4\delta^2(\xi) + 2\delta\ (\xi)\big[LRP(F;F_0) - URP(F;F_0)\big]$$
$$\sigma_0^2(\delta_0(\xi)) = 4\delta_0^2(\xi) - 2\delta_0(\xi)\big[LRP(F;F_0) - URP(F;F_0)\big] \qquad (10.3)$$

$LRP(F;\ F_0)$, $URP(F;\ F_0)$ are the lower and upper polarization indices given in Section 5.6.2, and

$$\delta_0(\xi) = \eta(\xi_{\frac{1}{2}}, \xi^0_{\frac{1}{2}})/f_0(\xi^0_{\frac{1}{2}})$$
$$\delta\ (\xi) = \eta(\xi_{\frac{1}{2}}, \xi^0_{\frac{1}{2}})/f(\xi_{\frac{1}{2}})$$
$$\eta(\xi_{\frac{1}{2}}, \xi^0_{\frac{1}{2}}) = \int_{-\infty}^{\infty} f(y) f_0(y - \xi_{\frac{1}{2}} + \xi^0_{\frac{1}{2}})\ dy - 2\int_{-\infty}^{\xi_{\frac{1}{2}}} f(y) f_0(y - \xi_{\frac{1}{2}} + \xi^0_{\frac{1}{2}})\ dy.$$

Here $\tilde{R}_{0L} = F(Y_0 + \rho)$ is the location matched distribution of Y_0 with respect to Y. It can be regarded as the inverse of R_{0L} as it has CDF $\tilde{G}_{0L}(r) = G^{-1}_{0L}(r),\ 0 \le r \le 1$. This result was proven in Handcock and Janssen (1998b).

The second term is the additional uncertainty incurred by using F_{n0} rather than F_0 in the definition of the quasirelative data. The last two terms are the contributions to the variance due to the estimation of the nuisance parameter ρ. If ρ were known then the correct variance is given by the first two terms. This will typically be the case when the comparison and reference distribution are known from the substantive theory to have the same median.

Note that there are many circumstances when the estimation of ρ does not contribute to the asymptotic variance. This will be the case when $\eta(\xi_{\frac{1}{2}}, \xi^0_{\frac{1}{2}})$ is zero. A sufficient condition for this is that $f(y)f_0(y - \xi_{\frac{1}{2}} + \xi^0_{\frac{1}{2}})$ is symmetric and, in particular, when both the target and comparison distributions are symmetric. The more skewed the two distributions are, the larger the variance contribution of ρ. If the contributions to the polarization from each tail are equal, the final term is zero. This can occur even if the comparison and target distributions are not symmetric. This general property of zero contribution is related to the idea of orthogonal tangent spaces in adaptive estimation theory (Bickel, *et al* 1993).

10.3.2 The asymptotic distribution of $\widehat{\mathbf{MRP}}(F;\ F_0)$ under equality

In many applications, researchers will test the hypothesis that the MRP is zero against the complementary alternative. To construct a test of this hypothesis it is interesting to consider the distribution of the estimate under the nonparametric null hypothesis that the reference and comparison distributions are identical ($F \equiv F_0$). For this purpose the following result is more accurate than the general result given in Section 10.3.1.

Theorem. *Under the hypothesis* $H_0 : F = F_0$,

$$\mathrm{E}\left[\widehat{MRP}(F;\ F_0)\right] = \begin{cases} \frac{1}{n+1} & n \text{ even} \\ \frac{1}{n} & n \text{ odd} \end{cases}.$$

$$\mathrm{Var}\left[\widehat{MRP}(F;\ F_0)\right] = \frac{\sigma^2_{MRP}}{nm}(m+n+1),$$

where

$$\sigma^2_{MRP} = \begin{cases} \frac{1}{3} + \frac{1}{(n+1)^2} & n \text{ even} \\ \frac{1}{3} - \frac{1}{n(n+1)} & n \text{ odd} \end{cases}$$

In addition, $\widehat{MRP}(F;\ F_0)$ *is asymptotically normal as* $n \to \infty$ *or* $m \to \infty$ *or both.*

These results indicate that, under the null hypothesis, $\widehat{\mathrm{MRP}}(F;\ F_0)$ is asymptotically unbiased (and can easily be made unbiased for fixed n). In addition it is asymptotically normal with variance very close to $(1/3)[1/n + 1/m]$, when n and m are moderate to large. For fixed m, as n increases so that our knowledge of F_0 increases, $\widehat{\mathrm{MRP}}(F;\ F_0)$ is unbiased with variance exactly $1/3m$. The additional inflationary factors involving n are the price we pay for not knowing F_0 and having to estimate it from the data by F_{n0}. The proof of this result uses Lemma F.2 given in Appendix F. Clearly

$$m\mathrm{Var}\left[\widehat{\mathrm{MRP}}(F;\ F_0)\right] = 16\sigma^2_{|Q_i - \frac{1}{2}|}\left[(m-1)\phi + 1\right],$$

so the variance follows easily. As the expectation and covariance matrix of the $|Q_i - \frac{1}{2}|$ are known and finite, the asymptotic normal result follows from the central limit theorem (Serfling 1980).

10.3.3 The joint distribution of the median relative polarization indices

In many situations the MRP is calculated for multiple comparison distributions relative to a fixed reference distribution. For example, the reference distribution could be earnings for workers in 1967 and the comparison distributions would be earnings series for workers in subsequent years. The objective would be to see how the polarization changed over time relative to a fixed reference year.

Generalizing our notation, let $Y_t = (Y_{t1}, \ldots, Y_{tm_t})$ denote the independent sample from the tth comparison distribution F_t, $t = 1, 2, \ldots, T$. Denote the vector of median polarization indices by $\mathbf{MRP} = \{\mathrm{MRP}(F_1;\ F_0), \mathrm{MRP}(F_2;\ F_0), \ldots, \mathrm{MRP}(F_T;\ F_0)\}'$. Let $\widehat{\mathbf{MRP}} = \{\widehat{\mathrm{MRP}}(F_1;\ F_0), \widehat{\mathrm{MRP}}(F_2;\ F_0), \ldots, \widehat{\mathrm{MRP}}(F_T;\ F_0)\}'$ be the vector of estimates of the MRPs.

The theorem in Section 10.3.1 can then be extended to cover the joint distribution of $\widehat{\mathbf{MRP}}$:

Theorem. *If the conditions of the theorem given in Section 10.3.1 are satisfied by* $F_0, F_t, t = 1, 2, \ldots, T$, *then*

$$\widehat{\mathbf{MRP}} \quad \sim \quad AN\{\mathbf{MRP}, \Sigma_{MRP}\} \tag{10.4}$$

as $m/n \to \kappa^2 < \infty, m \to \infty$. *An explicit expression for* Σ_{MRP} *is given in Section 10.4.2.*

As in previous subsection, the result can be refined to represent the joint distribution under the null hypothesis that the reference and comparison distributions are identical ($F_t \equiv F_0, t = 1, 2, \ldots, T$).

Theorem. *Under the hypothesis* $H_0 : F_t = F_0,\ t = 1, \ldots, T$, $\widehat{\mathbf{MRP}}$ *is asymptotically normal:*

$$\widehat{\mathbf{MRP}} \qquad \sim \qquad AN\left\{\mathbf{0}, \frac{\sigma^2_{MRP}}{n}\left(diag(\gamma_t) + \mathbf{1}\mathbf{1}' - \mathbf{I}\right)\right\} \tag{10.5}$$

as $n, m_1, \ldots, m_T \to \infty$. *Here* $\mathbf{I}$ *is the* $T \times T$ *identity matrix,* $\mathbf{1}$ *is the* T *unit vector and* $\gamma_t = m_t + n + 1/m_t, t = 1, \ldots, T$.

The proof is similar to that for the theorem in Section 10.3.2.

10.4 Computing standard errors

In this section we give the formula for the computation and estimation of the asymptotic distributions given in the theorem in Section 10.3.1 and Section 10.3.3.

10.4.1 The asymptotic variance of the estimate of MRP

For computing (10.2), it is convenient to have an expression directly in terms of the relative CDF:

$$16\text{Var}(\mid R_{0L} - \frac{1}{2} \mid) \;=\; 4 - 32\int_0^1 (r - \frac{1}{2}) G_{0L}(r)\, dr - [\text{MRP}(F;\ F_0) + 1]^2.$$

An alternative estimator can be based on the sample variance of $\{\mid Q_i - \frac{1}{2} \mid\}_{i=1}^m$. For use in practice, an estimate of $\tilde{G}0L(r)$ is needed and the natural choice is

$$\widehat{G}_{n,m}(r) \;=\; F_m(F_{n0}^{-1}(r)+\hat{\rho}) \quad 0<r<1.$$

The properties of this estimator with ρ known are developed in Section 9.2.2. Based on a simple extension of the results when the nuisance parameter ρ is estimated, it can be shown that $\widehat{G}_{n,m}$ is a $\sqrt{n}-$consistent and asymptotically normal estimator of $\tilde{G}0L(r)$ as $m/n \to \kappa^2 < \infty, m \to \infty$. Alternate estimators exist that can also be extended to this situation. See Li, *et al* (1996). The terms in (10.3) require estimates of the reference and comparison densities. These can be estimated using any of the approaches described in Section 9.1. In this book we have used a Poisson regression based estimator with the same bins and smoothing parameter for each distribution.

10.4.2 The asymptotic variance of the joint distribution

In this section we give an explicit formula for the asymptotic covariance matrix of $\widehat{\mathbf{MRP}}$ given in (10.4). Define $\rho_t = \xi_{\frac{1}{2}t} - \xi^0_{\frac{1}{2}}$ and $\xi_{\frac{1}{2}t} = F_t^{-1}(\frac{1}{2})$. Let $R_{t0L} = F_0(Y_t - \rho_t)$ be the relative distribution of F_t location matched to F_0 and $\tilde{R}_{0tL} = F_t(Y_0 + \rho_t)$ be the relative distribution of F_0 location matched to F_t. Define G_t to be the CDF of R_{t0L}. Based on (10.2), the (ts)th element of Σ_{MRP} is

$$\begin{aligned}
&\frac{1}{m}\mathcal{I}\{t=s\}\left[16\mathrm{Var}\Big(\mid R_{t0L}-\frac{1}{2})\mid\Big)+\sigma_t^2(\delta_t(\xi_t))\right]\\
&+\frac{1}{n}16\mathrm{Cov}\Big(|\tilde{R}_{0tL}-\frac{1}{2}\mid,|\tilde{R}_{0sL}-\frac{1}{2}\mid\Big)\\
&+\frac{1}{n}\Big[\quad \delta_{t0}(\xi_t)\left[\mathrm{LRP}(F_t;\ F_0)-\mathrm{URP}(F_t;\ F_0)\right]\\
&\quad+\ \delta_{s0}(\xi_s)\left[\mathrm{LRP}(F_s;\ F_0)-\mathrm{URP}(F_s;\ F_0)\right]\\
&\quad+4\delta_{t0}(\xi_t)\delta_{s0}(\xi_s)\Big],
\end{aligned}$$

where $\sigma_t^2(\cdot)$, $\delta_t(\cdot)$, and $\delta_{t0}(\cdot)$ are versions of $\sigma^2(\cdot)$, $\delta(\cdot)$ and $\delta_0(\cdot)$, respectively, given in (10.3) with F_t as the comparison distribution. As for (10.2), we can express the covariance directly in terms of the relative CDF:

$$\begin{aligned}
&16\mathrm{Cov}\Big(|\tilde{R}_{0tL}-\frac{1}{2}\mid,|\tilde{R}_{0sL}-\frac{1}{2}\mid\Big)\;=\\
&16\int_0^1|G_t(r)-\frac{1}{2}||G_s(r)-\frac{1}{2}|\,dr-[\mathrm{MRP}(F_t;\ F_0)+1][\mathrm{MRP}(F_s;\ F_0)+1].
\end{aligned}$$

These terms can be estimated using the approach given in Section 10.4.1.

10.5 Statistical properties of estimates of the upper and lower indices

As in Section 10.3 we will consider the situation where we have independent samples from both distributions. The natural estimators of the lower and upper indices are $\widehat{\mathrm{LRP}}(F;\ F_0) \equiv \mathrm{LRP}(F_m;\ F_{n0})$ and $\widehat{\mathrm{URP}}(F;\ F_0) \equiv \mathrm{URP}(F_m;\ F_{n0})$, respectively. These estimates can be reexpressed as

$$\widehat{\mathrm{LRP}}(F;\ F_0) \quad = \quad \frac{8}{m}\sum_{j=1}^{m}\left|Q_j - \frac{1}{2}\right|\mathcal{I}(Q_j \le \frac{1}{2}) \quad - \quad 1$$

$$\widehat{\mathrm{URP}}(F;\ F_0) \quad = \quad \frac{8}{m}\sum_{j=1}^{m}\left|Q_j - \frac{1}{2}\right|\mathcal{I}(Q_j > \frac{1}{2}) \quad - \quad 1$$

The vectors $\mathbf{LRP}, \mathbf{URP}$ and $\widehat{\mathbf{LRP}}, \widehat{\mathbf{URP}}$ are defined analogously to $\mathbf{MRP}$ and $\widehat{\mathbf{MRP}}$. We state the following result:

Theorem. *If the conditions of the theorem in Section 10.3.1 are satisfied by $F, F_t, t = 1, 2, \ldots, T$, then $\widehat{\mathbf{LRP}}$ and $\widehat{\mathbf{URP}}$ are asymptotically normal:*

$$\begin{pmatrix} \widehat{\mathbf{LRP}} \\ \\ \widehat{\mathbf{URP}} \end{pmatrix} \quad \sim \quad AN\left\{ \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \Sigma_{LRP} & \Sigma_{UL} \\ \Sigma_{UL} & \Sigma_{URP} \end{pmatrix} \right\} \tag{10.7}$$

as $m/n \to \kappa^2 < \infty, m \to \infty$.

The result for the case of equality of the distributions is given in the next section. There we also give explicit expressions for $\Sigma_{\mathrm{UL}}, \Sigma_{\mathrm{LRP}}$ and Σ_{URP} under equality of the distributions. Formulae for the general case are given in Handcock and Janssen (1998a).

These results indicate that $\widehat{\mathrm{LRP}}_t(F;\ F_0)$ and $\widehat{\mathrm{URP}}_t(F;\ F_0)$ are asymptotically unbiased. In addition they are asymptotically jointly normal. Under the null hypothesis that the reference and comparison distributions are identical ($F_t \equiv F_0, t = 1, 2, \ldots, T$) the variance is very close to $(5/3)[1/n + 1/m]$, when n and m are moderate to large. It is also important to note that the correlation between $\widehat{\mathrm{LRP}}_t(F;\ F_0)$ and $\widehat{\mathrm{URP}}_t(F;\ F_0)$ is about 3/5, and should not be ignored when interpreting the values. For fixed m, as n increases so that our knowledge of F_0 increases, $\widehat{\mathrm{LRP}}(F;\ F_0)$ and $\widehat{\mathrm{URP}}_t(F;\ F_0)$ are approximately unbiased with variance $5/3m$.

The proof of (10.7) is similar to that of (10.4) (given in Handcock and Janssen (1998a)) using the properties of two-sample U-statistics.

Computational formulae for the covariance terms in (10.7) can be derived in the same manner as those in Section 10.4. They are similar with terms of the form $\mid R_{t0L} - \frac{1}{2} \mid$ replaced by terms of the form $\mid R_{t0L} - \frac{1}{2} \mid \mathcal{I}\{ R_{t0L} \leq \frac{1}{2}\}$ for LRP and terms of the form $\mid R_{t0L} - \frac{1}{2} \mid \mathcal{I}\{ R_{t0L} > \frac{1}{2}\}$ for URP. Exact formula are available in Handcock and Janssen (1998a).

10.5.1 Distribution of the upper and lower indices under equality

In this section we consider the joint distributions of LRP(F; F_0) and URP(F; F_0) under the null hypothesis that the reference and comparison distributions are identical ($F_t \equiv F_0, t = 1, 2, \ldots, T$).

Theorem. *Under the hypothesis* $H_0 : F_t = F_0,\ t = 1, \ldots, T$, $\widehat{\mathbf{LRP}}$ *and* $\widehat{\mathbf{URP}}$ *are asymptotically normal:*

$$\begin{pmatrix} \widehat{\mathbf{LRP}} \\ \\ \widehat{\mathbf{URP}} \end{pmatrix} \quad \sim \quad AN\left\{ \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \Sigma^0_{LRP} & \Sigma^0_{UL} \\ \Sigma^0_{UL} & \Sigma^0_{LRP} \end{pmatrix} \right\}$$

as $n, m_1, \ldots, m_T \to \infty$. *The* (t,s)th *element of the* $T \times T$ *matrix* Σ^0_{LRP} *is*

$$\begin{cases} \frac{m_t+n+1}{nm_t}\left(5 \ - \ \frac{2}{(n+1)^2}\right) & t = s \\ \frac{1}{n}\left(5 \ + \ \frac{2}{(n+1)^2}\right) & t \neq s. \end{cases}$$

The (t,s)th *element of the* $T \times T$ *matrix* Σ^0_{UL} *is*

$$\begin{cases} \frac{m_t+n+1}{nm_t}\left(-3 \ + \ \frac{2}{(n+1)^2}\right) & t = s \\ \frac{1}{n}\left(-3 \ + \ \frac{2}{(n+1)^2}\right) & t \neq s. \end{cases}$$

The expressions are given for n even. The expressions for n odd are similar. The derivation is similar to that of the result in Section 10.3.2.

10.6 Tests of significance and multiple comparisons

In this section we use the results of Section 10.3 to construct confidence intervals and bands for the MRP. Results for the LRP and URP are similar. Often we would like to test if the MRP in a given situation has a statistically significant difference from zero.

If the sample sizes are not small, we can use the normal approximation to the exact distribution of the estimate as the basis for a test for a given significance level α :

$$P\left(\ |\widehat{\mathrm{MRP}}(F;\ F_0)| \leq z_{\alpha/2} \times \sqrt{\mathrm{Var}\left[\widehat{\mathrm{MRP}}(F;\ F_0)\right]} \ \right) \approx 1-\alpha. \qquad (10.6)$$

For most applications, the data sets will be large survey samples for which the sample sizes tend to be large (e.g., 1300–3000). As a result the approximations will be very close to the exact values. Consider the situation in Section 10.3 where we have many comparison distributions and wish to test if the MRP of *any* of them is significantly different from zero. Using the notation of that section, we wish to choose a critical value $L_{\alpha/2}$ such that for a given significance level α :

$$P\left(\ -L_{\alpha/2} \leq \widehat{\mathrm{MRP}}(F_t;\ F_0) \leq L_{\alpha/2},\ t=1,\ldots,T \ \right) \approx 1-\alpha.$$

If we wish to test the hypothesis based on individual level data, we can use the normal approximation to the exact (multivariate) distribution of $\widehat{\mathbf{MRP}}$ as the basis for a test (result (10.4) or (10.5)). The correlations between the components are non-negligible, so the usual Bonferroni inequality will lead to an unduly conservative critical value. An alternative approach is to use the Dunn-Šidák inequality (Hochberg and Tamhane 1987):

$$P\left(\ -L_{\alpha/2} \leq \widehat{\mathrm{MRP}}(F_t;\ F_0) \leq L_{\alpha/2},\ t=1,\ldots,T \ \right) \geq$$
$$\prod_{t=1}^{T} P\left(\ -L_{\alpha/2} \leq \widehat{\mathrm{MRP}}(F_t;\ F_0) \leq L_{\alpha/2} \ \right).$$

A conservative significance level can then be attained by using (10.6) and choosing the critical value to be $z_{\tilde{\alpha}/2}$, where $\tilde{\alpha} = 1 - (1-\alpha)^{1/T}$ for a given overall significance level α.

A better choice of critical value can be determined by using a method that explicitly takes into account the multivariate structure of $\widehat{\mathbf{MRP}}$. Dunnett (1989) considers the case where the distribution has a product correlation structure (i.e., the correlation can be written as $\beta_t \beta_s,\ t \neq s$. In our situation $\beta_t = 1/\sqrt{\gamma_t}\ t = 1,\ldots,T$). Using this approach, $L_{\alpha/2}$ can be calculated exactly (his algorithm is freely available as AS 251). The critical values for simultaneous comparisons made in this book are based on this method.

10.7 Bootstrap confidence intervals and achieved significance level

The bootstrap can be used to determine confidence intervals for general divergence measures. Confidence intervals for many measures can be derived from the asymptotic approximations to the distributions of their estimators. Examples are given in Sections 10.2 thru 10.5. However for many measures these distributions are not known, nor do simple approximations exist. A good example is the the Kullback-Leibler divergence. It can be expressed as a functional of g

$$D_{KL}(F;F_0) \equiv D_{KL}(g) = \int_0^1 \log\big(g(r)\big)dG(r) = \mathrm{E}_G\big[\log\big(g(X)\big)\big].$$

Some of the estimators proposed for $D(F;F_0)$ are described in Section 10.1. The bootstrap can be used to estimate the distributions of such divergence estimates even under these circumstances. For general information about the bootstrap see Efron (1993). A more detailed technical description is given in Shao and Tu (1995). Here we will describe one approach to implementing the bootstrap.

Let $D(F;F_0) \equiv D(g)$ be a divergence measure that can be expressed a functional of g which is smooth under appropriate assumptions on the density g. Let $Q_j = F_{n0}(Y_j), j = 1,\ldots,m$ be the quasirelative data from Section 9.2 and let $g_{n,m}(r)$ be an estimate of g based on $\{Q_j\}_{j=1}^m$ (Section 9.2.3). The procedures described here aim to determine measures of uncertainty for $D(g)$ based on generic features of $g_{n,m}(r)$. However the properties of the procedure will depend on the specific properties of the estimator.

A bootstrap algorithm for confidence intervals can be described as:

1) Select B independent bootstrap samples based on samples from the reference and comparison groups. Denote the reference group samples by $\{Y_{0i}^{*1}\}_{i=1}^n,\ldots,\{Y_{0i}^{*B}\}_{i=1}^n$, where each is a sample from $Y_{01}, Y_{02},\ldots,Y_{0n}$ consisting of n data values drawn with replacement. Denote the comparison group samples by $\{Y_j^{*1}\}_{j=1}^m,\ldots,\{Y_j^{*B}\}_{j=1}^m$, where each is a sample from $Y_1, Y_2,\ldots,Y_m$ consisting of m data values drawn with replacement.
2) Let $F_b(y) = \frac{1}{n}\sum_{i=1}^n \mathcal{I}(Y_{0i}^{*b} \le y)$ be the EDF of $\{Y_{0i}^{*b}\}_{i=1}^n$. Form the bootstrapped approximate relative data:
$$X_j^{*b} = F_b(Y_j^{*b}) \qquad j = 1,\ldots,m$$
and the bootstrapped relative density estimate $g_{n,m}(X^{*b})$.
3) Evaluate the bootstrapped estimate of the divergence $D(g_{n,m}(X^{*b}))$.
4) Define
$$K_{\mathrm{BOOT}}(r) = P[D(g_{n,m}^{*b}) \le r] \qquad 0 < r < 1.$$

An approximately $100(1-\alpha)\%$ confidence interval for $D(g)$ is the bootstrap percentile interval:

$$\left(K^{-1}_{\mathrm{BOOT}}(\alpha/2), K^{-1}_{\mathrm{BOOT}}(1-\alpha/2)\right).$$

The accuracy of the confidence set may be improved using more sophisticated methods – see Shao and Tu (1995), Chapter 4.

For the p-value for the hypothesis test $H_0 : F = F_0$, we seek to estimate the achieved significance level, $\mathrm{ASL} = P_{H_0}[D(\hat{U}) \geq D(\hat{g})]$ where $\hat{U}$ is the distributional estimate of g if the null hypothesis is correct. Let R_j be the rank of Y_j in the combined vector $\{Y_1, Y_2, \ldots, Y_m, Y_{01}, \ldots, Y_{0n}\}$, where the ties are broken in favor of the $\{Y_{0i}\}_{i=1}^n$, and S_j be the rank of Y_j in $\{Y_1, Y_2, \ldots, Y_m\}$, where ties are broken arbitrarily. The approximate relative data can then be expressed as $X_j = \frac{1}{n}(R_j - S_j), j = 1, \ldots, m$. Under H_0, $\{R_j\}_{j=1}^m$ is a random sample of size m from the integers 1 thru $n+m$ drawn without replacement and $\{S_j\}_{j=1}^m$ is determined by the $\{R_j\}_{j=1}^m$. Thus $\{X_j\}_{j=1}^m$ can be simulated directly from the known distribution of the ranks under the null hypothesis. Note that $\{X_j\}_{j=1}^m$ retains the dependence among the $\{X_j\}_{j=1}^m$ even though the marginal distribution of X_j is uniform on $\{0, \frac{1}{n}, \frac{2}{n}, \frac{n-1}{n}, 1\}$.

The most direct bootstrap algorithm to estimate ASL is:

1) Select B independent bootstrap samples: $\{X_j^{*1}\}_{j=1}^m, \ldots, \{X_j^{*B}\}_{j=1}^m$ from the above distribution of the ranks under H_0.
2) Evaluate the estimate of the divergence $D(g_{n,m}^{*b})$ based on $\{X_j^{*b}\}_{j=1}^m$.
3) Approximate ASL by

$$\widehat{\mathrm{ASL}} = \frac{1}{B}\#\{D(g_{n,m}^{*b}) \geq D(g_{n,m})\}.$$

The availability of a distribution under the null hypothesis for generating the bootstrap data that does not depend on characteristics of the unknown F and F_0 is key to the approach. For small sample sizes a test can be based on the permutation distribution of the ranks. See Efron and Tibshirani (1993), Chapter 15 for details.

As the bootstrap is based on a nonparametric estimator, the convergence rates of the bootstrap estimate may be slow. The quality of the bootstrap approximation can be improved in a number of ways. First, by drawing observations not from F_{n0} and F_m directly, but from smoothed versions of them (Shao and Tu 1995). This may help if the sample sizes ($n, m < 100$) are small, or if a nonsmooth estimate of g is used (e.g., a histogram estimator). Second, a smaller smoothing parameter can be used to reduce the bias in the estimator for g. Third, an estimator for the density adapted to the divergence functional may be used. In particular, estimators based on Poisson regression, log-density estimation such as log-spline density estimators (Kooperberg and Stone 1992) or other basis function expansions (Barron, *et al* 1992) can be considered. These estimators work well on $[0, 1]$ and have good analytical and computational properties.

Background material

The Kullback-Leibler divergence is central to many statistical endeavors – see Kullback (1968) and Soofi (1994). A fundamental reason for this is the connection to measures of information. Recall that the Kullback-Leibler divergence is just the entropy of the corresponding relative distribution.

There is an almost endless array of measures that have been proposed based on linear rank statistics and their variants. Other examples are the distance measures given in Table 4.5.1 of Titterington (1985).

Expressions for the covariance terms in (10.7) can be easily derived from the expressions in Section 10.4.2. The significance bands in Section 10.6 based on the assumption of equality of the distributions is the most useful in practice.

Exercises

Exercise 10.1. Show that the score function $J(p) = p$ with $\hat{D}_{CS}(F; F_0)$ results in the Wilcoxon statistic.

Exercise 10.2. Show that the score function $J(p) = \mathcal{I}(p < \frac{1}{2})$ with $\hat{D}_{CS}(F; F_0)$ results in the median test.

Exercise 10.3. Show that the score function $J(p) = (p - \frac{1}{2})^2$ with $\hat{D}_{CS}(F; F_0)$ results in the Mood test.

Exercise 10.4. Show that the score function $J(p) = |p - \frac{1}{2}|$ with $\hat{D}_{CS}(F; F_0)$ results in the Ansari-Bradley test.

Exercise 10.5. Use the theorem in Section 10.2.1 to derive the asymptotic distribution for the Wilcoxon statistic.

Exercise 10.6. Use the theorem in Section 10.2.1 to derive the asymptotic distribution for the median test statistic.

Exercise 10.7. Use the theorem in Section 10.2.1 to derive the asymptotic distribution for the Mood's test statistic.

Exercise 10.8. Use the theorem in Section 10.2.1 to derive the asymptotic distribution for the Ansari-Bradley test statistic.

Exercise 10.9. Suppose that the Legendre polynomials on $[-1, 1]$ are used to form the basis for the orthogonal series expansion given in Section 10.2.2. Show that the estimate of θ_1 given in Section 9.2.3.5 corresponds to the Wilcoxon rank sum.

Exercise 10.10. Suppose that the Legendre polynomials on $[-1, 1]$ are used to form the basis for the orthogonal series expansion given in Section 10.2.2.

Show that the estimate of θ_2 given in Section 9.2.3.5 corresponds to the Mood statistic.

Exercise 10.11. Suppose that the Hermite polynomials on $[-1, 1]$ are used to form the basis for the orthogonal series expansion given in Section 10.2.2. Show that the estimate of θ_1 given in Section 9.2.3.5 corresponds to the Normal scores test statistics.

Exercise 10.12. Suppose that the Hermite polynomials on $[-1, 1]$ are used to form the basis for the orthogonal series expansion Section 10.2.2. Show that the estimate of θ_2 given in Section 9.2.3.5 corresponds to the Klotz statistic.

Exercise 10.13. Justify the statement that if the comparison distribution is a location shifted version of the reference distribution, then the relative density will tend to be monotone (i.e., $g'(p)$ will not change sign). Give a counter example to the general claim.

Exercise 10.14. Suppose the comparison distribution is a scale shifted version of the reference distribution. Prove that $g'(p)$ will change sign exactly once, or give a counter example.

Exercise 10.15. Suppose that the Legendre polynomials on $[-1, 1]$ are used to form the basis for the orthogonal series expansion Section 10.2.2. Suppose the divergence measure is the weighted sum of the components: $\sum_{k=0}^{\infty} \alpha_k \theta_k$ where $\alpha_k = 1/k(k+1)$. Show that the estimator $\sum_{k=0}^{\infty} \alpha_k \hat{\theta}_k$ corresponds to the Anderson-Darling statistic.

Exercise 10.16. Suppose that the sine basis $\phi_k(p) = \sin(2\pi p)$ is used to form the basis for the orthogonal series expansion Section 10.2.2. Suppose the divergence measure is the weighted sum of the components: $\sum_{k=0}^{\infty} \alpha_k \theta_k$ where $\alpha_k = 1/k^2\pi$. Show that the estimator $\sum_{k=0}^{\infty} \alpha_k \hat{\theta}_k$ corresponds to the Cramer-von Mises statistic. .

Exercise 10.17. Look up the definition of intermediate efficiency given in Kallenberg (1983). What is it "intermediate" between? Give a critique of it in relation to the usual concepts of efficiency.

Exercise 10.18. What are the advantages of using a smoother estimator of F_0 in place of the empirical CDF F_{n0} in the definition of the location matched quasirelative data (Section 10.3)? What are the disadvantages?

Exercise 10.19. Give examples of comparison and reference distributions which are both nonsymmetric and for which the estimate of ρ does not inflate the asymptotic variance of $\widehat{\text{MRP}}(F; F_0)$ Give an intuitive explanation why this is so.

Exercise 10.20. Use the results in Appendix F to prove the theorem in Section 10.3.2.

Exercise 10.21. Verify the expression for the asymptotic distribution of $\widehat{\mathbf{MRP}}$, under the null hypothesis that the reference and comparison distributions are identical, given in Section 10.3.3.

Exercise 10.22. Give a bootstrap algorithm for $\hat{D}_{CS}(F; F_0)$ with the score function $J(p) = p$. Implement the algorithm and use it to compare a $\mathrm{N}(0, 2)$ distribution to a standard normal reference. Use $n = m = 5$ as the sample sizes. How does the bootstrap distribution of Step 3 in Section 10.7 compare to that given by the result in Section 10.2.1

This page intentionally left blank

Chapter 11

The Relative Distribution for Discrete Data

In this chapter we modify the definition of the relative distribution for continuous data given in Chapter 2 to cover discrete distributions and group-level data. We do so by introducing the idea of a random grade transformation. This approach ensures that the discrete relative distribution is continuous even though the source distributions are discrete. Extending the fundamental concept to the discrete data context ensures that the analysis of discrete distributions retains the tractability and interpretability of their continuous cousins.

The initial sections are concerned with the definition and interpretability of the discrete relative distribution, PDF and CDF. In the remainder of the chapter we address the inferential issues for the discrete relative distribution that have been addressed for the continuous version in Chapter 9 and 10. In Sections 11.3, we study the statistical properties of an estimator of the discrete relative CDF and PDF when the reference distribution is known. In Section 11.4 we extend this to the situation where both reference and comparison distributions are unknown. In many circumstances the sample information is discretized into the proportions falling into categorical bins formed by cut points. In Section 11.5, we define the group-level relative density appropriate for this situation and study estimators for it. The definition and estimation of the relative polarization indices based on both discrete and group-level indices is considered in Section 11.6. Much of this development utilizes the closeness of the definitions between the discrete and continuous versions. This parsimony enables both forms of information to be understood within the same framework.

11.1 The discrete relative distribution

In this section we will define a version of the relative distribution useful when the comparison or reference distributions are discrete or grouped.

Consider first the situation where both Y and Y_0 are discrete with outcome set $\{x_i\}_{i=1}^{Q}$, where Q can be infinity. Let the two probability mass functions be

$$p_i \;=\; P(Y = x_i) \qquad\qquad p_{0i} \;=\; P(Y_0 = x_i) \qquad\qquad i \;=\; 1, \ldots, Q$$

The CDF of Y_0,

$$F_0(x) \;=\; \sum_{i:x_i \le x} p_{0i} \qquad x \in \mathbb{R},$$

is a step-function with jumps of size p_{0i} at each x_i. Hence, for discrete data the grade transformation (2.2) does not produce a satisfactory scale for comparison of the two distributions. Consider the random transformation

$$F_0^d(x) \;=\; U\Big[F_0(x_{i-1}),\; F_0(x_i)\Big] \qquad \text{for} \;\; x_{i-1} < x \le x_i, \;\; i \;=\; 1, \ldots, Q.$$

where x is an element of the outcome space. Here $U[a,\ b]$ is the uniform distribution on the interval $(a,\ b]$. We define $F(x_0) = F_0(x_0) = 0$ for any $x_0 < x_1$. We call $F_0^d(x)$ a random transformation as it maps the value x to a random value in the interval from $F_0(x_{i-1})$ to $F_0(x_i)$. We can think of F_0^d as an extension of F_0 that has a continuous range, and is more in the spirit of the grade transformation. Note that $F_0^d(x)$ approaches $F_0(x)$ as the outcome space becomes more dense, and the two coincide for continuous outcome spaces.

The *discrete grade transformation* of Y to Y_0 is defined to be the random variable

$$R = F_0^d(Y). \tag{11.1}$$

R is obtained from Y by transforming it by using the function F_0^d, and so it is absolutely continuous with outcome space $[0, 1]$. Note that like the (continuous) grade transformation (2.2), R measures the relative rank of Y compared to Y_0, we shall also refer to the distribution of R as the relative distribution. The CDF of R is

$$G(r) \;=\; \Big(r - F_0(x_{i-1})\Big)g(i) \;+\; F(x_{i-1}), \tag{11.2}$$

where

$$F_0(x_{i-1}) < r \le F_0(x_i) \;\; i \;=\; 1, \ldots, Q.$$

Here

$$g(i) = \frac{p_i}{p_{0i}} \qquad\qquad i \;=\; 1, \ldots, Q. \tag{11.3}$$

This distribution is a natural generalization of the discrete CDF. First, its CDF matches the actual CDF on the outcome set: $G(r) = F(F_0^{-1}(r))$, for $r = F_0(x_i), i = 1, \ldots, Q$. Second, $G(r)$ is the linear interpolant between these values for $0 \le r \le 1$.

The *discrete relative density* $g(r)$ is defined to be the (right-continuous) derivative of $G(r)$, that is, the step function with values

$$g(i) \qquad\qquad \text{for} \qquad F_0(x_{i-1}) < r \le F_0(x_i)$$

for $0 \leq r \leq 1$. Thus although the CDF $G(r)$ is continuous and the PDF $g(r)$ exists, the latter is not continuous. However if the two distributions are identical, then the discrete relative distribution is the uniform probability distribution on $[0, 1]$, and the CDF of the relative distribution is a 45^o line from $(0, 0)$ to $(1, 1)$. Thus this definition retains the interpretability of the continuous version, and reduces to that version when the outcome space is continuous.

In particular, graphical displays similar to the plot of the relative CDF and density in the continuous case can be created by plotting $G(r)$ and $g(r), 0 \leq p \leq 1$. Note that the discrete PDF is defined by the values $\{g(i), F_0(x_i)\}_{i=1}^{Q}$. The relative CDF has support $\{F_0(x_i)\}_{i=1}^{Q}$ and values $G(i) = F(x_i)$.

11.2 Application: men's and women's hours worked

In Section 2.2, we used the distribution of earnings for women and men in 1987 to illustrate the definition of the relative distribution in the continuous case. In this section we consider the distributions of total hours worked for the same groups.

The data are drawn from the U.S. Current Population Survey (CPS) in its annual March supplement for 1987. The selected sample consists of white males and females, aged 16–66, and excludes the self-employed, full-time students, and those in the military and in farming. In each survey, respondents were asked, "In the weeks that ... worked, how many hours did ... usually work per week?" and "During 19XX in how many weeks did ... work even for a few hours? Include paid vacation and sick leave as work." Hours worked last year is derived by multiplying the reported hours worked per week last year by the reported weeks worked last year. The resultant samples sizes here are quite large; $n = 25,047$ for the men, and $m = 22,030$ for the women

Figure 11.1 is back-to-back display of their empirical probability mass functions. The women's distribution is plotted on top and the men's distribution on the bottom. This type of graph is a very good way to compare two discrete distributions that are on a metric scale. The distributions have a natural discreteness due to the tendency for respondents to report hours around standard work week schedules (e.g., 35, 37.5, or 40 hours per week).

The center of the men's distribution appears to be slightly above the center of the women's distribution. The spread of the distributions looks about the same. Both these observations (and other comparisons) can be verified using numerical summary measures of location, scale, and skewness, for example. Note, however, that the discreteness and coarseness of the distribution makes direct comparison of the distributions difficult.

Figure 11.2 is the discrete relative density of the women's to the men's distribution. A description of the estimators and their statistical properties

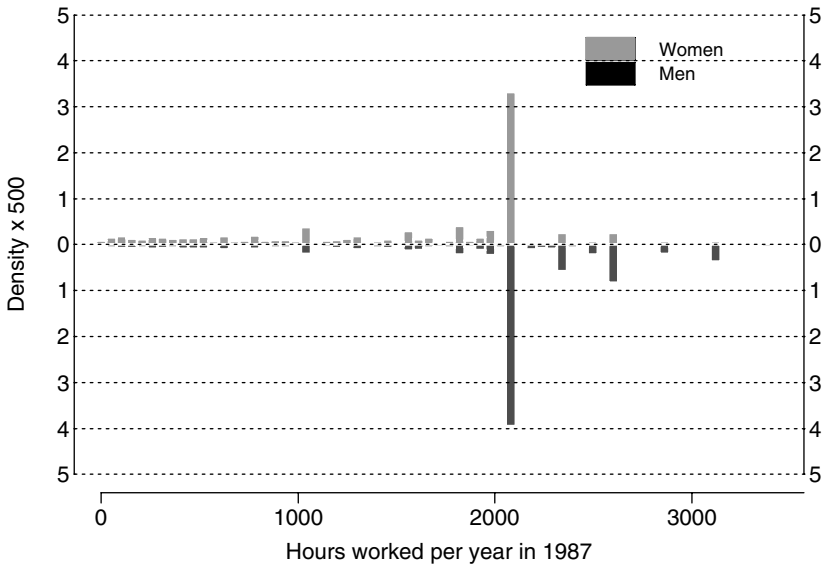

Fig. 11.1. The distributions of the hours worked per year for women and men in 1987 from the 1988 CPS.

is postponed until the next section. The horizontal line at 1 represents the relative density if the two distributions were identical. The upper axis is labeled in equivalent weekly hours worked (annual hours/52), and can be used for both men and women (see Section 2.2 and the discussion of Figure 2.2 there). We can see that women are much more likely to fall in the lower quantiles of the men's distribution ("part-time" workers). The plateau between the 20% quantile and the 65% quantile of the men's hours worked represents the men and women working 40 hours per week and 52 weeks a year (2,080 hours per year). By reading across the x-axis, we can see that $65\% - 20\% = 45\%$ of male workers in 1987 were working the equivalent of a standard 40-hour week. The relative density for this group is 0.9, indicating that women were 90% as likely as men to be working this schedule, that is, $90\% \times 45\% = 40\%$ were working the equivalent of a standard 40-hour week. There appears to be a spike in the women's hours just below the standard 40 hour week (about the 19% quantile of the men's distribution). The upper axis shows that this corresponds to working 37.5 hours per 52-week year. This may indicate that women are more likely to subtract a half-hour lunch break from their total 8-hour workday (or not be paid for it). Women are much less likely than men to report working

more than the standard 40-hour week ("overtime" workers), a pattern that strengthens as the number of hours worked increases.

For discrete data again, the relative distribution enhances direct comparison between the distribution. While the histograms in Figure 11.1 give a general sense of the differences in the two distributions, the details are hard to see, in part because the graphs are dominated by the modes. By contrast, the relative density in Figure 11.2 provides more precise information about *both* the modal differences and the detailed differences in the upper and lower tails.

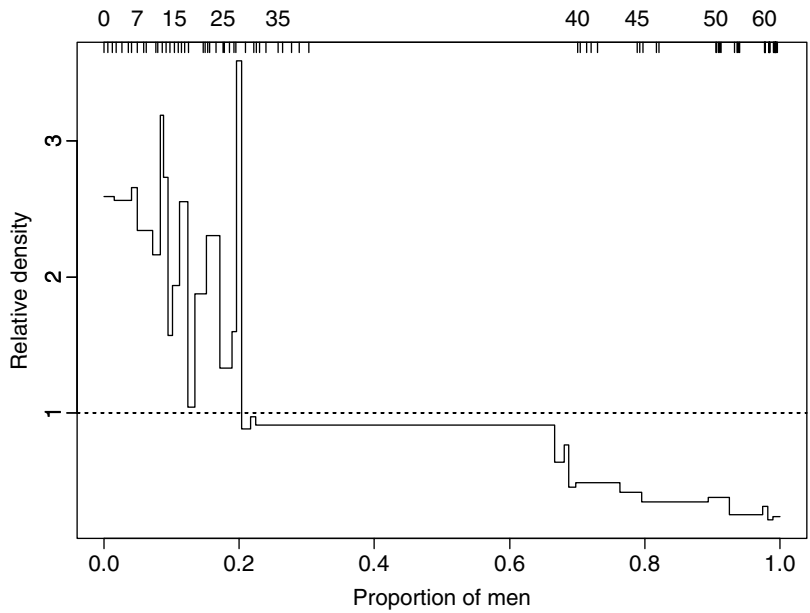

Fig. 11.2. The relative density of hours worked per year for women to men in 1987 from the 1988 CPS. The upper and right axes are labeled in average hours worked per 52-week year.

Figure 11.3 shows the relative CDF of women's to men's hours worked. The concave shape reflects the left-shifted location of the women's distribution relative to the men's. The number of hours worked by the median woman is below the 25% quantile for the men; 90% of women work fewer hours than the 75% quantile for the men. Conversely, only 60% of the men work less than the 80% quantile for the women. Reading off the top axis labels, we can also see that the fraction of female workers reporting less than the standard 40-hour work week is about 45%, compared to 20% for

men. This is consistent with information from other sources on the prevalence of women in the temporary help industry and their predominance in part-time jobs (see, for example, Belous 1989). Substantially more of the male workers are putting in *longer* working hours than women – about 30% of men are working over 40 hours per week, compared to about 10% of women. Note that the discreteness and bumps in the individual densities have much less visible impact on the relative CDF than they do on the individual densities and the relative PDF.

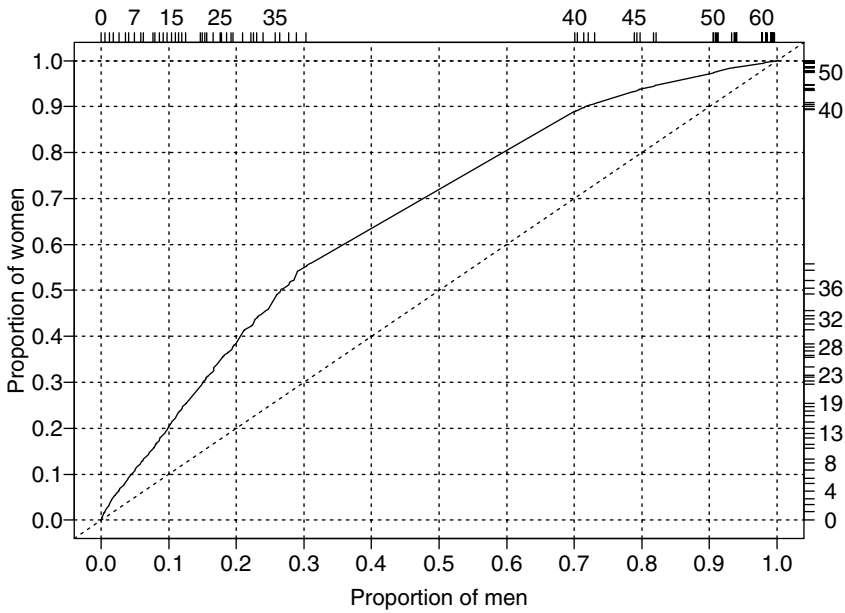

Fig. 11.3. The relative CDF of hours worked for women and men in 1987 from the 1988 CPS.

These figures suggest how the relative distribution can aid the comparison of distributions. This is not to suggest that they can replace the direct graphical overlay (as in Figure 11.1); their objective is to focus on those characteristics of the individual distributions important for comparing the two.

11.3 Inference when the reference distribution is known

In this section we assume that the reference CDF, F_0, is known and the data on the comparison group arises from a sample survey. That is, we assume that we have a sample $Y_1, Y_2, \ldots, Y_m$ that are independently and identically distributed from the population distribution F. We will assume throughout that the outcome space is finite (i.e., Q is finite).

Consider the relative data $R_j = F_0^d(Y_j),\ j = 1, \ldots, m$. As the sample is independently and identically drawn from the CDF F, the relative data are independently and identically drawn from the CDF G. Thus we could directly apply the CDF and PDF estimation methods of Sections 9.1 and 9.2 to the relative data on the support $[0, 1]$.

11.3.1 Estimation of the discrete relative CDF

The natural estimator of $G(r)$ is

$$G_m(r) \ = \ \frac{1}{m} \sum_{j=1}^{m} \mathcal{I}(F_0^d(Y_j) \leq r). \tag{11.4}$$

All the properties of this estimator described in Section 9.1 apply. However this estimator includes a small amount of unnecessary variation due to the random grade transformation. A direct estimator may do better. In addition most of the methods in Section 9.2 make various assumptions about the smoothness of the relative density which will not be true for the discrete relative density. We now consider more direct methods.

Note that $F_0(Y_j)$ has a multinomial distribution taking values $\{F_0(x_1), \ldots, F_0(x_Q)\}$ with probabilities $\mathbf{p} \ = \ \{p_1, \ldots, p_Q\}$. Motivated by (11.2), a direct estimator of $G(r)$ is

$$G_m(r) \ = \ \left(r - r_{i-1}\right)\widehat{g}_m(i) \ + \ \sum_{j=1}^{i-1} \widehat{p}_j \qquad r_{i-1} < r \leq r_i, \tag{11.5}$$

where $r_i = F_0(x_i), i \ = \ 1, \ldots, Q$, and

$$\widehat{p}_i \ = \ \frac{1}{m} \sum_{j=1}^{m} \mathcal{I}(Y_j = x_i) \qquad i \ = \ 1, \ldots, Q$$

and

$$\widehat{g}_m(i) = \frac{\widehat{p}_i}{p_{0i}} \qquad\qquad i \ = \ 1, \ldots, Q. \tag{11.6}$$

Much can be said about G_m as an estimator of $G(r)$. Perhaps the most useful result is

Theorem. *The estimator (11.5) satisfies:*

$$G_m(r) \sim AN\left\{G(r), \frac{\alpha\Sigma\alpha^T}{m}\right\} \qquad 0 < r < 1$$

as $m \to \infty$. *Here* $\alpha = (r - r_{i-1}, 1 - r + r_{i-1})$ *and*

$$\Sigma = \begin{pmatrix} G(r_i)(1 - G(r_i)) & G(r_{i-1})(1 - G(r_i)) \\ G(r_{i-1})(1 - G(r_i)) & G(r_{i-1})(1 - G(r_{i-1})) \end{pmatrix}.$$

Note that if $r = r_i$ then this reduces to the result (9.2) as this linear interpolating estimator and estimator (11.4) coincide. If r is not one of the r_i values, then this estimator can be shown to have smaller variance than (11.4) (Exercise 11.17).

11.3.2 Estimation of the discrete relative PDF

The properties of the estimator (11.5) suggest that we use (11.6) as an estimator of the relative PDF. Estimation of the relative PDF $g(r)$ reduces to the estimation of $g(i)$ as the cut points $r_i, i = 1, \ldots, Q$ are known. The natural estimator of $\mathbf{g} = \{g(1), \ldots, g(Q)\}$ is $\widehat{\mathbf{g}}_m = \{\widehat{g}_m(1), \ldots, \widehat{g}_m(Q)\}$ given in (11.6).

Note that the distribution of $\widehat{\mathbf{g}}_m$ is that of scaled multinomial proportions. Hence we have (see, e.g., Serfling 1980):

Theorem. *The estimator in (11.6) satisfies:*

$$\widehat{\mathbf{g}}_m \sim AN\left\{g(i), \frac{\Omega}{m}\right\} \tag{11.7}$$

as $m \to \infty$. *Here the covariance matrix is*

$$\Omega = \begin{cases} -g(i)g(j) & i \neq j \\ g(i)\left(\frac{1}{p_{0i}} - g(i)\right) & i = j \end{cases}.$$

11.4 Inference for the discrete relative distribution

In most application contexts, the CDF of the reference distribution is also unknown and must be estimated from sample data. We will assume that we have a sample $Y_{01}, Y_{02}, \ldots, Y_{0n}$ that is independently and identically distributed from the population distribution F_0. We assume that this sample

and the sample from the comparison group defined in the previous section are independent. As in Section 9.1, it is natural to estimate $F_0(y)$ by the empirical distribution function of the reference sample:

$$F_{n0}(y) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{I}(Y_{0i} \le y) \qquad -\infty < y < \infty.$$

In this section we discuss the estimation and inference issues for the relative density $g(r), 0 < r < 1$. Note that it is defined by $\{r_i, g(i)\}_{i=1}^{Q}$. Motivated by (11.4), consider the following estimator of $g(i)$

$$g_{n,m}(i) = \frac{\widehat{p}_i}{\widehat{p}_{0i}} \qquad i = 1, \ldots, Q, \tag{11.8}$$

where

$$\widehat{p}_{0i} = \frac{1}{n}\sum_{l=1}^{n}\mathcal{I}(Y_{0l} = x_i) \qquad l = 1, \ldots, Q.$$

We can estimate the cut points r_i by $\hat{r}_i = F_{n0}(x_i) = \widehat{p}_{0i}$. Let $\mathbf{r_{n,m}} = \{\hat{r}_1, \ldots, \hat{r}_Q\}$, $\mathbf{r} = \{r_1, \ldots, r_Q\}$, and $\mathbf{g_{n,m}} = \{g_{n,m}(1), \ldots, g_{n,m}(Q)\}$.

The asymptotic properties of the estimator are described in the following result:

Theorem. *The estimator* $\{\hat{r}_i, g_{n,m}(i)\}$ *of* $\{r_i, g(i)\}$ *satisfies:*

$$\mathbf{g_{n,m}} \sim AN\left\{\mathbf{g}, \frac{1}{m}\Omega + \frac{1}{n}\Omega_o\right\} \tag{11.9}$$

where

$$\Omega_o = \begin{cases} -g(i)g(j) & i \ne j \\ g^2(i)\left(\frac{1-p_i}{p_i}\right) & i = j \end{cases}$$

$$\mathbf{r_{n,m}} \sim AN\left\{\mathbf{r}, \frac{1}{n}\Omega_{p0}\right\}$$

where

$$\Omega_{p0} = \begin{cases} -p_{0i}p_{0j} & i \ne j \\ (1-p_{0i})p_{0i} & i = j \end{cases}$$

as $m \to \infty, m/n \to \kappa^2 < \infty$. *In addition, the two estimators are asymptotically independent.*

It is informative to compare the properties of this estimator to those of the estimator (11.5). We can interpret the additional term in the asymptotic variance for $g_{n,m}(i)$ compared to $g_m(i)$ as the price we pay for using F_{n0} as a surrogate for the unknown F_0.

In the special case that F and F_0 are identical, we have:

Theorem. *If* $F \equiv F_0$ *then*

$$\mathbf{g_{n,m}} \quad \sim \quad AN\left\{\mathbf{1}\ ,\ \left\{\frac{1}{m}+\frac{1}{n}\right\}\left\{\mathrm{diag}\left(\frac{1}{p_i}\right)-\mathbf{11}'\right\}\right\}$$

as $m \to \infty, m/n \to \kappa^2 < \infty$. *Here* $\mathbf{1}$ *is the* $Q \times 1$ *identity vector.*

In this situation, the inflation of the variance for the less frequent values of the reference distribution is clearly seen.

The results in this section can be used to calculate simultaneous confidence bands for the relative distribution based on $g_{n,m}$ (Section 9.6). The proof of these results is given in Appendix E.

The numerators in $\mathbf{g_{n,m}}$ are the sample proportions from a sample of size m from a multinomial distribution with probability parameter $\mathbf{p}$. Similarly, the denominators in $\mathbf{g_{n,m}}$ are the sample proportions from a sample of size n from a multinomial distribution with parameter $\mathbf{p_0} = \{p_{01}, \ldots, p_{0Q}\}$. Thus $\mathbf{g_{n,m}}$ can be consider to be the componentwise ratio of two independent multinomial distributions. The results then follow from applying the delta method, standard asymptotic results about the distributions of ratios (Hinkley 1969) and some algebraic manipulations.

We assume throughout that the supports of the reference and comparison distributions coincide. If there exists an x_i such that $p_i = P(Y = x_i) = 0, p_{0i} = P(Y_0 = x_i) > 0$ then $g(i) = 0$ and we interpret the diagonal element of Ω_0 as zero. This result holds with a degenerate normal distribution. If it is known that $p_i = 0$ then this component need not be estimated. If $p_i > 0$ and $p_{0i} = 0$ then the discrete relative distribution exists on a collapsed support. In this case $r_i = r_{i-1}$ and define $\hat{g}_m(i) = g(i) = 0$.

11.5 Grouped data

Consider the situation where both the comparison and reference distributions are continuous, but only group-level statistics of the sample information are reported. Let the common outcome set of Y and Y_0 be partitioned into a finite number of groups (Q) with ith cut point $c[i]$, defined by:

$$F_0(c[i]) \ = \ \frac{i}{Q} \quad \text{or} \quad c[i] \ = \ F_0^{-1}(\frac{i}{Q}), \quad i \ = \ 0, 1, \ldots, Q.$$

Instead of the individual level samples we only observe the sample proportion of each group whose values fall in the interval $[c[i-1], c[i])$ .The difference between this situation and the previous discrete situation is that in this case, the groups are equally probable (with respect to the reference distribution), and we have implicitly continuous underlying distributions. Note however that for most purposes we can think of the group-level distributions as discrete distributions with outcome space $c[1], \ldots c[Q]$ and probability mass functions:

$$p_i = F(c[i]) - F(c[i-1]) \qquad p_{0i} = F_0(c[i]) - F_0(c[i-1]) \quad i = 1, \ldots, Q.$$

We will implicitly assume that the values are equally distributed between cut points, so that the relative density is constant between cut points. The discrete relative density $g(r)$ is then given by the construction in Section 11.1 with $g(i)$ equal to the proportion of the comparison group whose values fall in the interval $[c[i-1], c[i])$, divided by the proportion in the reference distribution (See (11.3)).

11.5.1 Estimation of the grouped relative density

Let

$$\hat{p}_i = \frac{1}{m} \sum_{j=1}^{m} \mathcal{I}(c[i-1] \le Y_j < c[i]), \qquad i = 1, \ldots, Q$$

be the proportion of the comparison sample falling into each group and

$$\hat{p}_{0i} = \frac{1}{n} \sum_{i=1}^{n} \mathcal{I}(c[i-1] \le Y_{0i} < c[i]), \qquad i = 1, \ldots, Q$$

be the proportion of the reference sample falling into each group. The cut points for the groups are usually based on the reference sample, that is, $c[i] = F_{n0}(\frac{i}{Q}), \quad i = 0, 1, \ldots, Q$. Thus the proportion of the sample from the reference distribution falling in each group is exactly $1/Q$. The natural estimate of the $g(i)$ is $\widehat{g}(i) = \hat{p}_i / \hat{p}_{0i}$.

The behavior of the estimate depends on how the cut points are determined. If the cut points are known (rather than estimated from the data), then the distribution of $\widehat{\mathbf{g}} = \{\widehat{g}(1), \ldots, \widehat{g}(Q)\}$ is given by (11.9).

Suppose that the cut points are estimated as quantiles of the reference sample. That is, $c[i] = F_{n0}(r_i), \; i = 0, 1, \ldots, Q$, where $0 = r_0 < r_1 < \cdots < r_Q = 1$. Typically these are equally spaced: $r_i = i/Q$. For example, the *decile* version corresponds to $r_i = i/10$ and $Q = 10$. We then have

$$\begin{aligned} \widehat{g}(i) &= \frac{F_m(c[i]) - F_m(c[i-1])}{F_{n0}(c[i]) - F_{n0}(c[i-1])} \\ &= \frac{F_m(F_{n0}^{-1}(r_i)) - F_m(F_{n0}^{-1}(r_{i-1}))}{r_i - r_{i-1}} \\ &= \frac{1}{r_i - r_{i-1}} \Big[G_{n,m}(r_i) - G_{n,m}(r_{i-1}) \Big] \qquad i = 1, \ldots, Q. \end{aligned}$$

where $G_{n,m}(r)$ is the empirical CDF for the relative distribution of Y to Y_0 given in (9.16). Note that this is known for $r = r_i, i = 0, \ldots, Q$.

The asymptotic statistical properties of this estimator can then be derived from the asymptotic joint distribution of $\{G_{n,m}(r); 0 \le r \le 1\}$ given in Section 9.2.2.1.

Theorem. *Under the conditions of (9.17)*

$$\widehat{\mathbf{g}} \quad \sim \quad AN\left\{\mathbf{g}, \frac{1}{m}\Omega \quad + \quad \frac{1}{n}\Omega_\gamma\right\}$$

where the ijth element of Ω_γ *is*

$$\frac{\gamma_{ij} - \gamma_{i-1,j} - \gamma_{i,j-1} + \gamma_{i-1,j-1}}{(r_i - r_{i-1})(r_j - r_{j-1})}$$

and $\gamma_{ij} = (\min(r_i, r_j) - r_i r_j) g(i) g(j)$ *as* $m \to \infty, m/n \to \kappa^2 < \infty$.

11.6 Inference for the relative polarization indices

In this section we consider the definition and estimation of relative polarization indices for discrete and group-level data. The treatment of summary measures for discrete level data is greatly simplified by the definition given in Section 11.1. As the discrete relative distribution is continuous, the definitions and interpretations of the summary measures of Chapter 5 still apply. In addition the estimators described in Chapter 10 can be used. Here we will focus on group-level data from underlying continuous distributions. We will also assume that the discretization is based on an even number Q of equispaced classes with respect to the reference distribution.

The relative polarization for the group level data can be defined by:

$$\begin{aligned} MRP(F;\ F_0) = & \quad \frac{Q}{Q-2}\left[4\mathrm{E}(\mid R - \frac{1}{2} \mid) \quad - \quad 1\right] \\ = & \quad \frac{Q}{Q-2}\left[4\int_0^1 \mid r - \frac{1}{2} \mid g(r)\ dr \quad - \quad 1\right] \\ = & \quad \frac{4}{Q-2}\sum_{i=1}^{Q}\left| \frac{i - \frac{1}{2}}{Q} - \frac{1}{2} \right| g(i) \quad - \quad \frac{Q}{Q-2}. \end{aligned} \tag{11.10}$$

This expression is analogous to the definition for the continuous case given in Section 5.6. The two coincide if we assume that the underlying continuous relative density is constant between cut points. The group level version has been rescaled by a factor $Q/(Q-2)$ to ensure that it has range -1 to 1.

The natural estimate of the group level relative polarization index is

$$\widehat{MRP}(F;\ F_0) \quad = \quad \frac{4}{Q-2}\sum_{i=1}^{Q}\left| \frac{i - \frac{1}{2}}{Q} - \frac{1}{2} \right| \widehat{g}(i) \quad - \quad \frac{Q}{Q-2}. \tag{11.11}$$

As the estimate is a weighted average of the $\hat{g}(i)$, its properties can be derived from those of the previous section.

Theorem. *If the cut points are known, that is,* $c[i] = F_0^{-1}(r_i), r_i = i/Q, i = 0, 1, \ldots, Q$, *then we should adjust the estimator to reflect this (i.e., use* $\widehat{g}(i) = p_i/(1/Q)$. *This estimator satisfies:*

$$\widehat{MRP}(F;\ F_0) \quad \sim \quad AN\left\{MRP(F;\ F_0), \qquad \frac{1}{m}\Sigma_1\right\},$$

where Σ_r *is*

$$\frac{16}{(Q-2)^2}\left[(Q+1)^2 q(r, Q/2, 0) - 2(Q+1)q(r, Q/2, 1) + q(r, Q, 2) - \mu^2\right]$$

$$q(j,k,l) = \frac{1}{Q}\sum_{i=1}^{k} i^l g^j(i) \qquad j = 1, 2, \qquad k = 0, 1, 2, \qquad l = Q/2,\ Q$$

$$\mu \ = \ (Q+1)q(1, Q/2, 0) - 2q(1, Q/2, 1) + q(1, Q, 1)$$

as $m \to \infty, m/n \to \kappa^2 < \infty$.

The estimate can be reexpressed as

$$\frac{4}{Q-2}\mathbf{q}^T\widehat{\mathbf{g}} \quad - \quad \frac{3Q+2}{Q-2},$$

where the ith element of $\mathbf{q}$ is $(Q - 2i + 1)\mathcal{I}\{i \le Q/2\} + 1$. The result then follows from some algebraic manipulations.

Note that if we use the estimator based on the ratio or sample proportions instead, we have

Theorem. *If the estimator (11.8) is used for* $g(i)$ *in (11.10) then*

$$\widehat{MRP}(F;\ F_0) \quad \sim \quad AN\left\{MRP(F;\ F_0), \qquad \frac{1}{m}\Sigma_1 \ + \ \frac{1}{n}\Sigma_2\right\}$$

as $m \to \infty, m/n \to \kappa^2 < \infty$.

We can be more precise about the bias:

$$\mathrm{E}\left[\widehat{MRP}(F;\ F_0)\right] \ = \ MRP(F;\ F_0) + \frac{(Q-1)}{n}MRP(F;\ F_0) + o(\frac{1}{n}) + o(\frac{1}{m}).$$

The second term is a result of the uncertainty in the reference group proportions.

Typically the exact equispaced cut points are unknown and are usually estimated from the reference sample, that is, $c[i] = F_{n0}(\frac{i}{Q}), i = 0, 1, \ldots, Q$. We then have

Theorem. *If the reference group cut points are estimated by their sample values then*

$$\widehat{MRP}(F;\ F_0) \ \sim \ AN\left\{MRP(F;\ F_0), \quad \frac{1}{m}\Sigma_1 \ + \ \frac{1}{n}\frac{16}{(Q-2)^2}\left[\mathbf{q}^T\Omega_\gamma\mathbf{q}\right]\right\}$$

as $m \to \infty, m/n \to \kappa^2 < \infty$.

We can consider the refinement of these results to the situation when the reference and comparison distributions are equal.

Theorem. *Under the hypothesis* $H_0 : F = F_0$,

$$\widehat{MRP}(F;\ F_0) \ \sim \ AN\left\{0, \quad \frac{Q+2}{3(Q-2)} \cdot \left\{\frac{1}{m} + \frac{1}{n}\right\}\right\}$$

as $m \to \infty, m/n \to \kappa^2 < \infty$. *If the cut points are known so that the estimator* $\widehat{g}(i) = p_i/(1/Q)$ *is used:*

$$\widehat{MRP}(F;\ F_0) \ \sim \ AN\left\{0, \quad \frac{Q+2}{3(Q-2)} \cdot \frac{1}{m}\right\}.$$

11.6.1 The joint distribution of the group level relative polarization

The relative polarization index is often calculated for multiple comparison distributions relative to a fixed reference distribution.

Suppose $Y_t = (Y_{t1}, \ldots, Y_{tm_t})$ is an independent sample from the tth comparison distribution $F_t \quad t = 1, 2, \ldots, K$. The estimate of the group level relative density for the tth target distribution is denoted by $\widehat{g}_t(i)$. Denote by $\mathbf{MRP} = \{MRP(F_1;\ F_0), MRP(F_2;\ F_0), \ldots, MRP(F_K;\ F_0)\}^T$, and $\widehat{\mathbf{MRP}} = \{\widehat{MRP}(F_1;\ F_0), \widehat{MRP}(F_2;\ F_0), \ldots, \widehat{MRP}(F_K;\ F_0)\}^T$, the vectors of indices and estimates of the indices, respectively. In this situation the numerators in the estimates $\{\widehat{g}_t(i)\}_{t=1}^K$ are uncorrelated, and the dependence is introduced only by the common denominator and cut points.

Typically the exact equispaced cut points are unknown and are usually estimated from the reference sample, that is, $c[i] \ = \ F_{n0}(\frac{i}{Q})$, for $i \ = \ 0, 1, \ldots, Q$. We then have

Theorem. *If the reference group cut points are estimated by their sample values then*

$$\widehat{\mathbf{MRP}} \qquad \sim \qquad AN\{\mathbf{MRP}, \Sigma_{\mathbf{MRP}}\},$$

where the (t, s)th *element of the* $K \times K$ *matrix* $\Sigma_{\mathbf{MRP}}$ *is:*

$$\frac{1}{m_t}\Sigma_1^t\mathcal{I}\{t=s\} \;+\; \frac{1}{n}\frac{16}{(Q-2)^2}\left[\mathbf{q}^T\Omega_\gamma^t\mathbf{q}\right]$$

as $n, m_1, \ldots, m_K \to \infty$ *at the same rate. Here* Σ_1^t *and* Ω_γ^t *are the versions of* Σ_1 *and* Ω_γ *for the* tth *comparison distribution.*

We can specialize this result to the situation where all the distributions are equal. Using the approach in Section 10.6 these can be used to construct simultaneous significance bands for polarization index.

Theorem. *Under the hypothesis* $H_0 : F_t = F_0, \quad t = 1, \ldots, K,$ $\widehat{\mathbf{MRP}}$ *is asymptotically normal:*

$$\widehat{\mathbf{MRP}} \quad \sim \quad AN\left\{\mathbf{0}, \frac{Q+2}{3(Q-2)}\begin{pmatrix} \gamma_1 & 1 & & \ldots & & & 1 \\ 1 & \ldots & 1 & \gamma_2 & 1 & \ldots & 1 \\ 1 & & & \ldots & & 1 & \gamma_K \end{pmatrix}\right\}.$$

where

$$\gamma_t \;=\; \frac{1}{m_t} + \frac{1}{n} \quad t = 1, \ldots, K.$$

as $n, m_1, \ldots, m_K \to \infty$ *at the same rate.*

11.6.2 Indices of upper and lower polarization

As described in Section 5.7, the median relative polarization index can be decomposed into contributions from the lower tails and contributions from the upper tails of the distributions.

For group-level data the lower polarization index can be defined as

$$\begin{aligned} LRP(F;\, F_0) = \;& \frac{Q}{Q-2}\left[8\int_0^{\frac{1}{2}} \mid r - \frac{1}{2} \mid g(r)\, dr \;\; - \;\; 1\right] \\ = \;& \frac{8}{Q-2}\sum_{i=1}^{Q/2}\left| \frac{i-\frac{1}{2}}{Q} - \frac{1}{2}\right| g(i) \;\; - \;\; \frac{Q}{Q-2}. \end{aligned}$$

The natural estimate of the group level lower relative polarization index replaces $g(i)$ by its sample version described in the previous sections. The upper relative polarization can be defined and estimated in the same manner. Note that the comparison and reference distributions are median matched in this definition, that is, it is assumed that $g(1) + \cdots + g(Q/2) = \frac{1}{2}$. The asymptotic distributions of the estimators is very similar to that for the median relative polarization index and will not be detailed here.

Background material

The use of the density ratio is the natural way to define the relative probability mass function for discrete data. This was first suggested by Parzen (1983). Parzen (1993) proposed the distribution-type P-P plot, which is tantamount to the discrete relative CDF (11.2). The approach in Section 11.1 places each of these functions within a coherent framework centered around the concept of a continuous relative distribution. This ensures that both continuous and discrete distributions can be treated in a similar manner.

In this chapter we do not explicitly consider the situation where the comparison and reference distributions are known to be members of parametric families of discrete distributions. See Simonoff (1996) for a discussion of the estimation of discrete parametric distributions.

Exercises

Exercise 11.1. Suppose that Y_0 is a binary random variable taking the values -1 and 1 with equal probability. That is,

$$P(Y_0 = -1) = P(Y_0 = 1) = \frac{1}{2}.$$

Let Y have a standard normal distribution. What is the distribution of the grade transformation $R = F_0(Y)$, given in (2.2)?

Exercise 11.2. Suppose that Y is a binary random variable taking the values -1 and 1 with equal probability. Let Y_0 have a standard normal distribution. What is the distribution of the grade transformation $R = F_0(Y)$? Is the relative CDF of Y to Y_0 the inverse of the one in Exercise 11.1?

Exercise 11.3. Suppose that both Y and Y_0 are binary random variables taking the values -1 and 1 with equal probability. What is the distribution of the grade transformation $R = F_0(Y)$? Does the grade transformation represent a satisfactory definition for the relative distribution? Explain.

Exercise 11.4. In Exercise 11.3, suppose the support of both Y and Y_0 is $\{a, b\}$ for values $a < b$. What is the distribution of R?

Exercise 11.5. Use the definition of the random grade transformation to show that the CDF of the discrete relative distribution is given by (11.2).

Exercise 11.6. Use the definition of the random grade transformation to show that the discrete relative distribution is absolutely continuous. Then derive the formula for the PDF of the relative distribution given in (11.3).

Exercise 11.7. Determine the random grade transformation of Y to Y_0 for the distributions in Exercise 11.1 How does it differ from the (nonrandom) grade transformation?

Exercise 11.8. Determine the random grade transformation of Y to Y_0 for the distributions in Exercise 11.2 How does it differ from the (nonrandom) grade transformation?

Exercise 11.9. Answer Exercises 11.3 and 11.4 for the random grade transformation. How do they differ from the (nonrandom) grade transformations?

Exercise 11.10. Let Y and Y_0 be absolutely continuous distributions. Show that the random grade transformation and the grade transformation coincide.

Exercise 11.11. In Section 11.3 it is claimed that the linear interpolating estimator of $G(r)$ has variance that is less than or equal to the variance of the estimator (11.3). Prove that the variance is the same at the points $r = r_i, i = 1, \ldots, Q$. Show that the variance is strictly less at the other points. This result can be shown algebraically. Can you give a heuristic proof of the result?

Exercise 11.12. In the following sequence of questions we verify and expand on the analysis in Section 11.2 of the total hours worked for women and men in 1987.
What is the median women's hours worked? What is the median men's hours worked? Compare the IQR of the two distributions? Calculate parallel boxplots for women and men separately. Calculate the skewness of each distribution.
Write a summary comparing the two distributions based on these summary statistics.

Exercise 11.13. Calculate the probability mass function of the total hours worked for women in 1987. Do the same for the men.
Calculate the relative PDF (Figure 11.2) and CDF (Figure 11.3). Verify the numerical summaries given in Section 11.2.

Exercise 11.14. Calculate the probability mass function of the total hours worked for women in 1997. Do the same for the men. Are these distributions different from their 1987 counterparts?
Calculate the relative PDF of 1997 women to 1997 men. Is it similar to the one given in Figure 11.2 for 1987?

Exercise 11.15. Calculate the relative PDF of women in 1997 to women in 1987. Calculate the relative PDF of men in 1997 to men in 1987. Are the two similar? Complete a decomposition analysis of women to men similar to that in Section 8.4 for earnings. Summarize your findings and discuss the differences from those of the earnings analysis.

Exercise 11.16. Prove the theorem in Section 11.3.1.

Exercise 11.17. Prove the theorem in Section 11.3.2.

Exercise 11.18. Verify the formula (11.9)

Exercise 11.19. Prove the theorem in Section 11.5.1.

This page intentionally left blank

Chapter 12

Application: Changes in the Distribution of Hours Worked

12.1 Background

In this application, we will turn our attention to the distribution of hourly wages for all workers, rather than the subset of workers who are employed full-time, full-year. While the demographics of the labor force changed substantially during the period of growing earnings inequality, changes of similar magnitude were also occurring in the structure of the labor market. Restructuring took two forms: continuing decline in manufacturing employment leading to the emergence of a "service economy" (Fuchs 1968), and a rise in market-mediated employment relations such as outsourcing, subcontracting, and temporary, contingent, and part-time work contracts.

Deindustrialization is associated for many with the substitution of bad jobs for good ones. Service sector jobs have traditionally paid less, offered fewer benefits, and more part-time employment (Costrell 1988; Meisenheimer II 1998). While these changes have been found to be associated with some of the loss in middle-income jobs and subsequent polarization in earnings, the evidence suggests that other factors must also be playing a role. For one the decline in manufacturing employment is not a recent phenomenon. Over the past 50 years, manufacturing's share of employment has been falling steadily, almost linearly. In addition, earnings inequality has been growing within different industrial sectors, including manufacturing.

The "good jobs–bad jobs" debate has thus increasingly focused on the changing nature of employment relations within firms. In contrast to the idea that some industries provide good jobs with stable employment and high wages, while other industries provide bad jobs with low wages, insecurity and no mobility, the evidence suggests these employment strategies are being used together, not only within industries, but also within firms (Cappelli 1994; Harrison 1994). As a result, research has began to shift to firm-level analyses of employment restructuring.

The postwar years of earnings growth and equalization emerged during a unique period in American industrial history. The period was marked by the development of a system of employment relations often referred to as the "internal labor market" (Doeringer and Piore 1971). The key characteristic

of an internal labor market is a formal hierarchy of jobs within firms that are filled primarily by internal promotion rather than through external recruitment. The resulting system serves to buffer employment relations – including decisions about wages, job mobility, and training – from the volatility of external market pressures.

In stylized form, the internal labor market was characterized by the lifetime job. Workers started at one company, stayed with it, and were guaranteed job security and yearly raises. In return, employers obtained control over labor supply and a committed workforce, or at least a negotiated truce with labor. For jobs higher in the skill hierarchy, the system also provided customized training, since workers learned on the job and therefore brought firm-specific knowledge and tested skills to each new position (Kochan, *et al* 1986; Piore and Sabel 1984).

The terms of this trade-off deteriorated for American employers in the mid-70s. Cost reduction became an important basis of competition, and internal labor markets became a natural target. Cost reduction requires flexibility in who is hired, for how long, for how much, and for which tasks. To get this flexibility, some firms have adopted high-performance work systems that can benefit their employees as well as productivity (Pfeffer 1994; Piore and Sabel 1984). Other employers, however, are now more willing to rely instead on the external labor market, as the high-performance systems require significant initial investments in technology and training. With the changes in corporate financing and governance in the wake of banking deregulation, the "shareholder revolution" has skewed the incentives towards short term growth in dividends, rather than long term reinvestment of profits (for a review, see Applebaum and Berg 1996). The wave of "downsizing" that took place during the late 1980s and 1990s heralded this change. For the first time, employment losses finally reached deep into the white collar occupations (Cappelli 1992), though some question the extent of the change (Gordon 1996). There are many good reviews of this literature (Appelbaum 1987; Cappelli 1995; Colclough and Tolbert 1992; Harrison 1994; Osterman 1994; Pfeffer and Baron 1988).

In one of the first systematic studies of the growth in "market-mediated" employment relations, Belous (1989) documented a dramatic rise in the number of contingent workers during the 1980s. While the total labor force grew by 14% during this period, the number of agency temporary workers grew by 175%, part-time employment grew by 21%, employment in the business service sector – the primary provider of subcontracted human services – grew by 70%, and self-employment grew by 19%. Overall, Belous estimates that the contingent workforce grew from about 25-28% of the workforce to 30-37% of the workforce during the decade. Subsequent estimates, using more refined definitions and different data sources have generally been lower, from 5% (Abraham 1990; Polivka 1996) to 17% (National Center on Educational Quality of the Workforce 1995) of the workforce. At the same time, nearly 80% of firms reported making use of flexible

staffing arrangements, excluding the use of part-time workers, which is extremely widespread (Houseman 1997; Mishel and Bernstein 1994, p229). While agencies specializing in temporary clerical workers accounted for two-thirds of the total temporary employment in 1972, they comprised only 55% in 1982 (Abraham 1990). Increasingly, firms are turning to temporary workers to staff other specialized, nonclerical positions.

This change in business practices may help to explain the growth in the dispersion of hourly wages for the workforce as a whole: hourly wages for part-time workers are on average 70% of those of comparable full-time workers (Belous 1989, p104). We will use relative distribution methods to take an initial look at this question. To proxy the change in employment relations, we will use the distribution of weekly hours worked. Substitution of part-time for full-time workers should result in an increased dispersion in weekly hours worked, with growth in the lower tail of the distribution. Note that we could also have used weeks worked during the last year to proxy for contingent work status, as contingent workers are less likely to have job security, and may spend more weeks unemployed or out of the labor force when they lose their jobs. We have run the analysis with each variable and the findings are essentially the same, so we report here on the results from the part-time analysis.

The relative distribution can easily be used to test whether the distribution of weekly hours worked has changed over time, and the analysis will provide an example of an application of the methods to discrete data. In addition, the decomposition techniques discussed in Chapters 7 and 8 can be used here to determine whether the change in the distribution of hours worked is associated with the change in the distribution of hourly wages.

12.2 Data

The data are drawn from the March supplement of the Current Population Survey (CPS) for earnings years 1980 through 1997. For simplicity, the selected sample consists of white males, aged 16–66 and excludes the self-employed, full-time students, and those in the military and in farming. Women and minority workers have typically been over-represented in the contingent workforce, so some of the impact of this type of restructuring may play out through a widening of the gender and race wage gaps. But if firm-level restructuring is playing a role in growing earnings inequality more generally, then its effects should also be visible among white men. From previous chapters, we have seen that this group experienced a marked growth in annual earnings inequality, even among full-time full-year workers. On the one hand, this indicates that changes in hours worked will not explain all of the increase in wage inequality. On the other, it suggests that white men will not be immune to the trends.

Our measure of the distribution of hours worked is taken from the CPS question: "In the weeks that you worked, how many hours did you usually work per week?" We will refer to this as the work schedule distribution below.

For this analysis, we use hourly wages, rather than annual earnings, as our income measure. The choice is dictated by the nature of the research question. Annual earnings are a function of hours worked, and the correlation between these two variables is not the focus of interest here. We are not asking whether firms have reduced their overall utilization of labor (or workers their overall supply of hours), but instead whether firms are substituting less expensive contingent workers for more expensive full-time workers. If so, then the distribution of wage offers will have changed, and it is these wage offers that we analyze here. We derive the hourly wage by dividing the reported annual income from wages and salary by the annual hours worked last year. We construct annual hours worked as the product of weeks worked last year and usual hours per week. The number of weeks worked last year is taken from the CPS question, "During 19XX, how many weeks did you work either full-time or part-time, not counting work around the house? Include paid vacation and paid sick leave." They are deflated using the PCE deflator to represent 1997 real dollars.

12.3 Findings

To motivate the analysis here, we start by examining the polarization trends in hours and wages. Figure 12.1 displays the relative polarization index series for hours worked (panel (a)) and hourly wages (panel (b)). Both series show similar levels of polarization over the period, and both display greater polarization in the upper tail than in the lower. The trends are sufficiently similar to suggest that there may be a causal relation. These are marginal trends, however, so we proceed with this in mind.

12.3.1 Changes in the distribution of hours worked

Traditional summary statistics give some sense of the changes in the distributions of hours worked These are presented in Table 12.1. The median, not surprisingly, is 40 hours per week, the standard work week. The mean shows a slight upshift, and the standard deviation indicates greater dispersion. The IQR suggests that the dispersion may be greater in the upper tail than the lower.

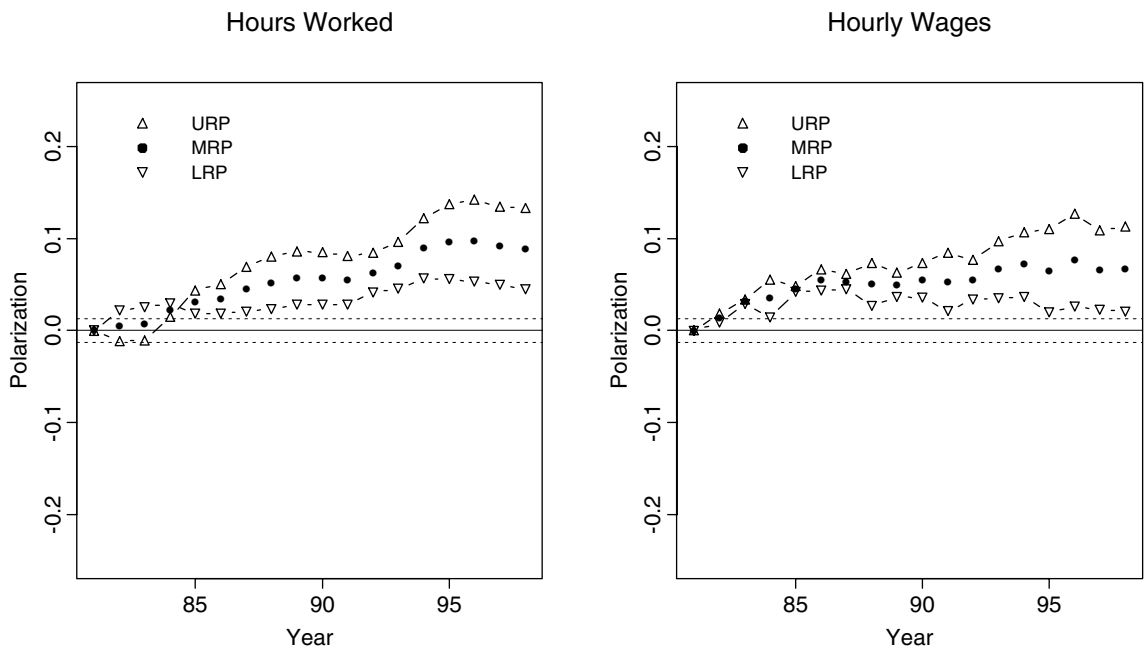

Fig. 12.1. The relative polarization indices (lower, median and upper) for changes in the distributions of hours worked and hourly wages.

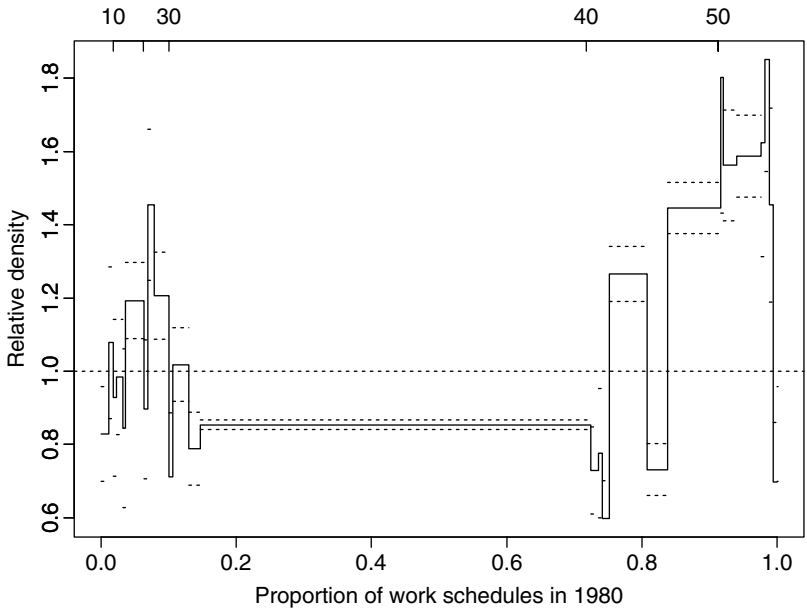

Fig. 12.2. The relative distribution of usual weekly hours worked in 1997 to that in 1980. The upper axis is labeled in 1980 weekly hours worked. The dotted lines are 95% pointwise confidence bounds.

Table 12.1. Summary statistics for the distributions of hours worked in 1980 and 1997.

Summary Statistic	1980	1997
Sample size	38,459	26,908
Mean	41.1	42.2
Standard deviation	10.7	11.4
Median	40	40
Interquartile range	40–45	40–48

To get a more complete picture of the changes, we can examine the relative distribution of work schedules in 1997 to that in 1980. This is shown in Figure 12.2. Conceptually, this relative density is similar to the one constructed for earnings in previous chapters, though there is no need for deflation here as the scale is the same in both time periods. The graph is not nearly as smooth because of natural discreteness in reported hours around standard work week schedules (e.g., 35–40 hours per week). The labels at the top show the usual weekly hours worked.

The bar from the 10% through the 70% quantiles in the figure represents individuals working 40 hours per week and 52 hours per year, indicating that $70\% - 10\% = 60\%$ of workers in 1980 were working the standard work week. The relative density for this group is about 0.85, indicating that such workers were about 14% less common in 1997 (so $86\% \times 60\% = 52\%$ are working the standard 40-hour week in 1997). The polarization in work schedules can be clearly seen, and the stronger polarization in the upper tail is also evident.

The proportion of workers in 1997 reporting less than the standard 40-hour week does not appear to have grown, except for certain schedules. The thin spike in the lower tail occurs at about 24 hours per week, and the value indicates that about 45% more workers reported 24-hour work weeks in 1997. There has also been a slight increase in those reporting 25–35 hours per week. Other part-time schedules, however, are less common in 1997. Overall, the fraction of workers reporting less than 35 hours per week is about the same in both years: 11%. This suggests that white men have not been affected by the growth of part-time jobs.

There has been growth in the upper tail of the hours distribution, however, with about 60% more workers reporting 50- to 60-hour work weeks. The fraction working more than a standard 40-hour work week grew from 28% in 1980 to 35% in 1997. White men appear to have increased rather than decreased the hours they spend working.

The estimated MRP for the relative distribution of work schedules in Figure 12.2 is 0.089 (95% CI 0.080–0.098). The estimated lower and upper relative polarization indices are 0.045 and 0.133 respectively, indicating relatively more growth in the upper tail. Both are significant at the 95% level.

Figure 12.3 displays the changes in the distribution of hourly wages over this period. The first panel shows the estimates of the two densities overlaid. As the sample sizes here are over 20,000 in each year, the sampling variability will have only a modest effect on the form of the kernel density estimates. The 1997 distribution is slightly downshifted from the 1980 distribution, and its tails are somewhat denser. This can be verified using summary measures of location, scale, and skewness. In particular, the mean of the 1997 values is larger than that of the 1980 values ($17.78 vs. $16.36), while the median is slightly smaller ($13.49 vs. $14.26). This suggests the importance of changes in distributional shape.

Panel (b) shows the shape shift in the wage distribution, that is, the location-matched relative density of 1997 to 1980 wages. We have used a multiplicative median adjustment here, as the wages are not logged. Median-matching has a small effect, as the medians of the two distributions are quite close. It does, however, reveal the marked polarization in the upper tail, which was not visible in the PDF overlay. In the absence of the small median wage downshift, the fraction of workers in the upper decile of the wage distribution would have risen by nearly 50%. But the density in

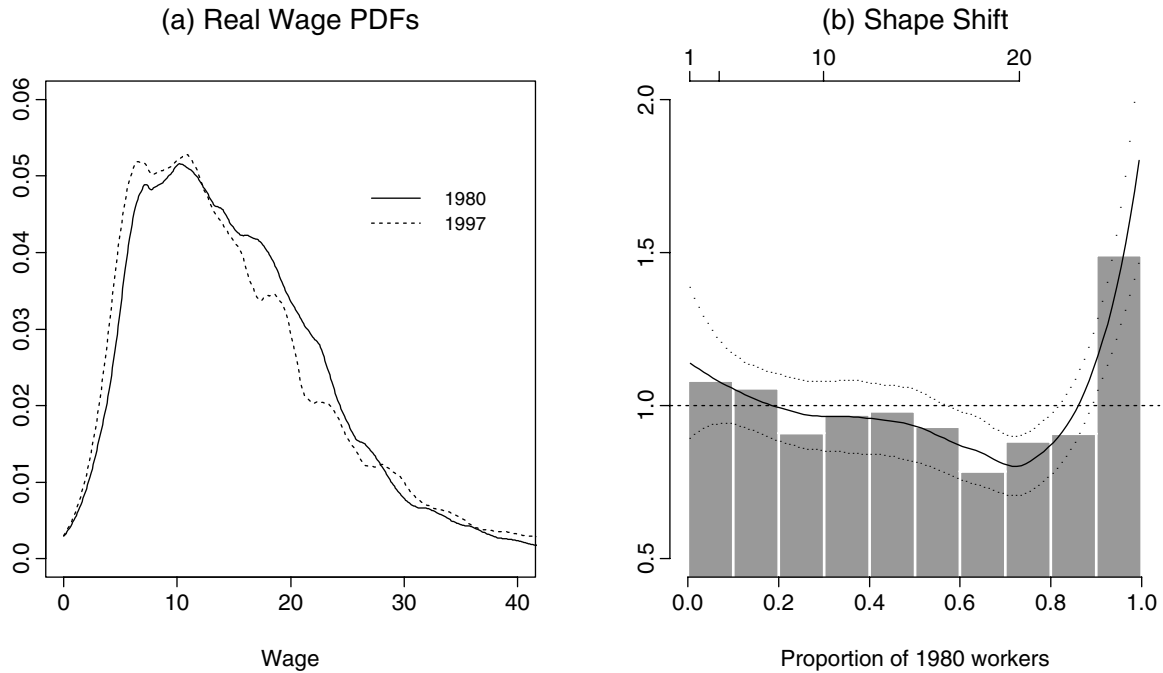

Fig. 12.3. (a) The distributions of hourly wages in 1980 and 1997 expressed in 1997 dollars. (b) The shape shift, shown by the location-matched relative density of 1997 to 1980 wages. The upper axis is labeled in 1997 dollars.

the lower tail would also have grown by about 10%. Without the median-matching, the growth in the upper and lower tails is more symmetric: 32% and 24% respectively.

Figure 12.3 shows that the hourly wage distribution of white men has polarized during this period. The estimate of the MRP of the relative distribution of hourly wages in Figure 12.3 is 0.063 (95% CI $0.054 - 0.072$). The upper and lower indices are 0.11 and 0.02 respectively. Only the upper index is significant at the 95% level.

12.3.2 Linking changes in hours worked to changes in wages

The fact that the upper tails of both the wage and work schedule distributions have polarized suggests that there may be a link between the two changes. In this section, we begin to explore this possibility. We disaggregate the overall wage distribution and polarization indices into subgroups of workers defined by the number of weekly hours worked. This approach will allow us to characterize within- and between-group changes in the wage distribution over time. We median-match the overall distribution before separating into the different groups to net out the aggregate change. Residual location shifts within groups will now capture upshifting and downshifting relative to the pooled population.

Table 12.2 presents the mean and median wages for three groups of workers: those working less than the standard work week (< 35 hours per week), those working a standard work week (35–40 hours per week), and those working more than the standard work week (> 40 hours per week).

Table 12.2. Mean and median hourly wages, 1980 and 1997, by work group (in 1997 real dollars).

	Mean		Median	
Work Group	1980	1997	1980	1997
Total workforce	\$16.48	\$18.09	\$14.51	\$13.75
part-time	11.05	15.60	7.17	6.86
standard	17.05	16.97	15.53	13.88
overtime	17.32	20.58	14.93	15.87

The differences in wage trends are quite pronounced. Real median earnings for both part-time and standard workers fell from 1980 to 1997; by about 5% for part-time workers, and about 10% for standard. By contrast, median earnings rose for the overtime group by about 6%. The net result was a change in relative position. In 1980, the wages for standard and overtime workers were very close, while the wages for part-time workers lagged behind. In 1997, part-time workers remained behind, but the median earnings for overtime workers had pulled ahead of standard workers by about

14%. The means tell a different story: they rise rather than fall over the period for most groups, and part-time workers appear to be doing relatively better using this measure. Given the strong right skew in the unlogged wage distribution, the median is probably the better location measure.

Figure 12.4 shows the 1997 to 1980 relative wage distributions for the three groups of workers. Because these wage distributions have not been median-matched *within* each group, the relative densities represent both group-specific median and overall shape shifts (see Chapters 3 and 4). The median upshift for the overtime workers is quite visible. The relative PDF shows that compared to their counterparts in 1980, about twice as many of these men were earning wages in the top decile in 1998. This upshift, however, masks a slight polarization in their earnings. The median relative polarization index for this group is the same as for the pooled population, with the upper tail contributing more than the lower. The set of polarization indices for each group are presented in Table 12.3. The loss in median earnings is visible for the standard full-time workers. We can also see a significant polarization in the upper tail of their earnings distribution ($URP = 0.06$), but the polarization in the lower tail is significantly negative, so the net overall polarization for this group is close to 0. Part-time workers show a milder and nearly symmetric polarization.

Table 12.3. Polarization indices for the relative distribution of hourly wages, 1980 to 1997, by work group. Indices significant at the 95% level are indicated by *.

Work Group	LRP	MRP	URP
Overall	0.02	0.06*	0.11*
part-time	0.05	0.05*	0.04
standard	-0.04*	0.01	0.06*
overtime	0.03	0.06*	0.09*

If each of the group specific polarization indices were close to 0, this would imply that after holding changes in work schedule constant there is no residual polarization in wages. The polarization we observe in the overall wage distribution must then be due entirely to the changing composition of work schedules. If, instead, all of the group specific polarization indices were about equal to the overall workforce indices, then holding the changes in work schedule constant does nothing to reduce the observed polarization in wages. This would suggest that the polarization in work schedules has contributed little to the polarization in wages.

Instead, we see a mix of these two scenarios. Workers on the standard work week show a wage distribution with upper tail polarization, and lower tail convergence. Those who worked part-time show low but symmetric positive polarization over this time period. And those who worked overtime show greater polarization. As the average hourly wage for the overtime

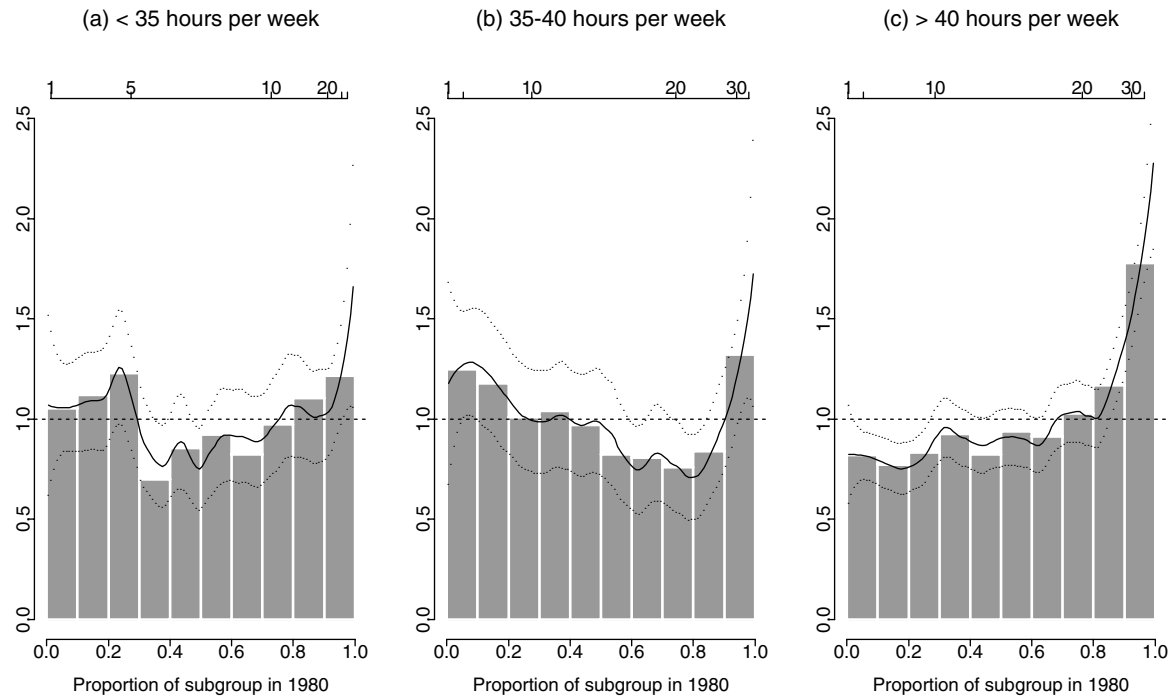

Fig. 12.4. Relative wage distributions for three groups of workers, defined by their usual work schedule: (a) less than 35 hours per week; (b) 35–40 hours per week; (c) more than 40 hours per week. The upper axis is labeled in 1997 dollars.

group is 15% higher than the overall workforce in 1997, their within-group distributional shifts will amplify the upper polarization in the overall wage distribution beyond that expected by the growth in their numbers.

At this point, we can make several initial conclusions. The shifts in work schedules that we observed do not completely account for the polarization in wages, because there is evidence of residual polarization within the main group of full-time workers. At the same time, however, the shifts in work schedules likely had *some* effect. This effect is difficult to establish from the three within-group graphs because the scales on the horizontal axis are standardized to within-group quantiles, so the group position on the overall scale (visible on the top axis) is not easy to decode.

To summarize the composition effect, we can decompose the overall relative distribution of wages by the distribution of hours worked, using the methods from Chapter 7. These methods allow us to to adjust the relative wage distribution so that we can examine the residual differences assuming there had been no changes in work schedules.

Figure 12.5 graphically represents the decomposition of the relative wage density by work schedule changes (see Chapter 7, Section 7.2). Panel (a) is the original (unadjusted) relative wage density 12.2. Panel (b) represents the part of (a) that is attributable to the effect of changes in the distribution of weekly hours worked. Panel (c) represents the *hours-adjusted* relative wage density – what the relative density would have looked like in the absence of any compositional changes in work schedules.

From panel (b), we can see that the shift in weekly hours worked had an extremely modest effect on the distribution of hourly wages. There appears to be a very slight downshift effect, and some polarization in the upper tail. The small size of the effect is somewhat surprising given the large increase in higher-paid overtime workers by 1997. What this panel shows, however, is that *absent any changes in the relative wages paid to overtime workers*, the impact of the change in work schedules would have been negligible.

The RD in the middle panel is formed by comparing the composition-adjusted 1980 wage distribution to the original 1980 wage distribution. As we noted above, in 1980 the overtime workers had little median wage advantage over the rest of the workforce: \$14.93 vs. \$14.51, about 3%. Absent an increase in this wage differential, a rising share of overtime workers could only affect the relative wage distribution if their 1980 wages were more or less polarized than the rest of the workforce. In fact, their wages were less polarized (the MRP is -0.09 and significant). The slight growth in the lower tail seen in this panel must therefore be due to the slight increase in the fraction of part-time workers, whose median wages – \$7.17 per hour – are substantially lower than the rest of the workforce.

The panel makes clear that the shift in work schedules did not drive the majority of the rising inequality in wages, and panel (c) shows a residual polarization virtually identical to the unadjusted RD in (a). The growth in overtime workers *indirectly* contributed to the growth in upper tail wage

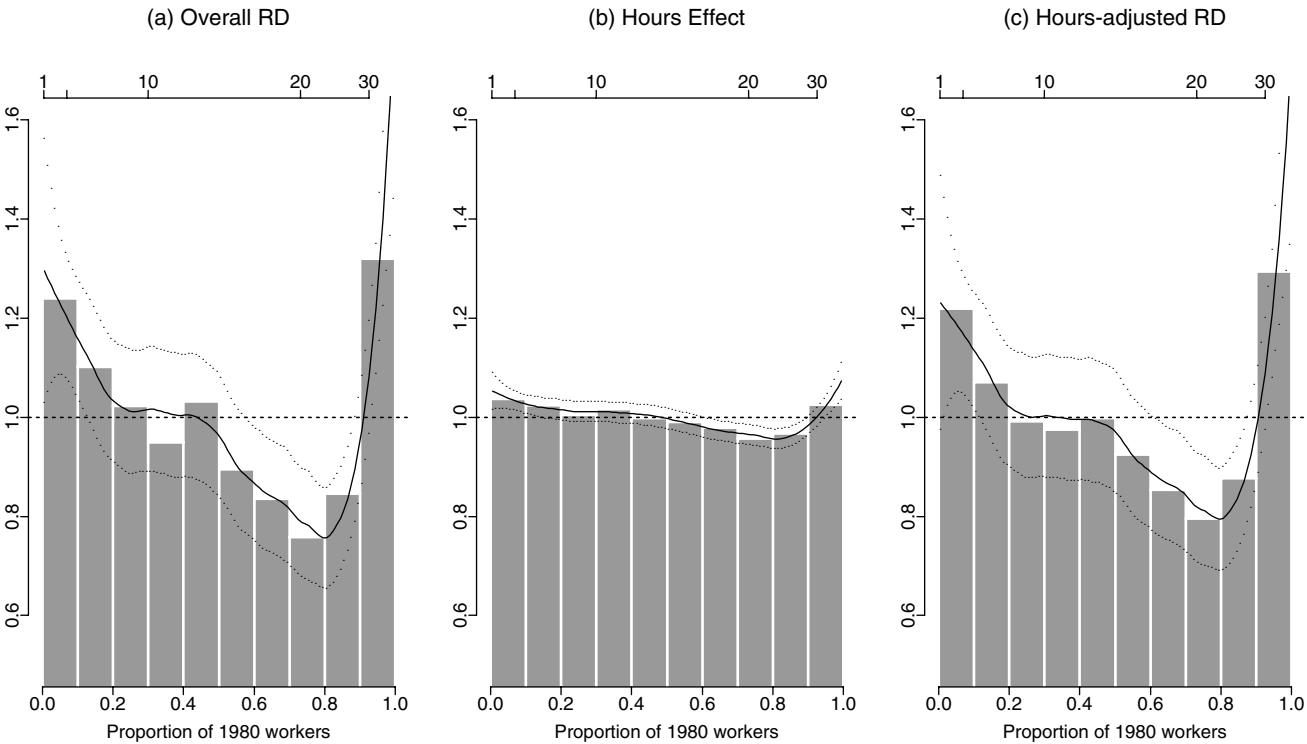

Fig. 12.5. Decomposition of the relative wage distribution, 1997 to 1980, by hours worked: (a) Original unadjusted relative wage density; (b) effect of changes in the distribution of hours worked; (c) the relative wage density adjusted for changes in hours worked. The upper axis is labeled in 1997 dollars. The entropy for each relative density is shown above this.

polarization, but only because their relative wage advantage increased. It was not the hours, but the wages paid for these hours, that were ultimately responsible for the growth in overall wage polarization.

12.4 Discussion

The work schedule distribution for white men has polarized in the last two decades, and many are now working longer hours. As shown by the RD for the distribution of weekly hours worked, the fraction working the standard 35–40 hour work week has fallen by 15%, and most of the corresponding increase is found among overtime workers. During this period a similar polarization occurred in hourly wages. While there was a modest decrease in the median wage, the fraction in the top wage decile grew by about 30%. This similarity in marginal wage and work schedule changes suggests that the change in the mix of work schedules could be driving a substantial part of the growth in wage inequality. Using stratified analyses and the decomposition technique, however, we find this is not the case. The exploratory graphics demonstrate significant residual wage polarization in the regular full-time work group (which continues to comprise about 50% of the workforce), and in the overtime group. The findings from the decomposition analysis show at best a modest polarizing effect of the compositional change in work schedules, and the residual polarization is nearly the same as the unadjusted. Unexpectedly, the composition effects were not found at the highest earnings level, but at the lowest. Thus the greater increase in upper tail of the work schedule distribution does not appear to account for the increase in the upper tail of the wage distribution. Instead, it was the relative median wage increase for overtime workers that provided the boost in the upper tail of the wage distribution. Had these overtime workers been paid the same relative wages in 1997 as they had in 1980, there would have been little impact of their growth in number.

Exercises

Exercise 12.1. The analysis in the chapter focused on white males. The case for women may be very different as the numbers and role of women in the workforce has changed over the time period 1980 to 1997. Follow through the analysis of Section 12.3 for white women. How do the relative polarization indices over time compare to that for white men given in Figure 12.1? How does the relative distribution of hours worked compare to that for white men given in Figure 12.2? Summarize the effect of changes in the work schedule distribution on the distribution of wages for white women. Discuss the similarities and differences with the case for white men.

Exercise 12.2. Minority workers have typically been over-represented in the contingent workforce. Follow through the analysis of Section 12.3 for black men and then for black women. How do the relative polarization indices over time compare to that for white men given in Figure 12.1? How does the relative distribution of hours worked compare to that for white men given in Figure 12.2? Summarize the effect of changes in the work schedule distribution on the distribution of wages for black men and black women. Discuss the similarities and differences with the case for white men.

Exercise 12.3. In the chapter, hours worked was used as a proxy for the in employment relations. Many other factors have changed over the time period under study. Follow through the analysis of Section 12.3 adjusting for the number of years of education rather than hours worked. Summarize the effect of changes in the distribution of this measure of educational attainment on the distribution of wages. How do these effects differ from the effects of changes in the work schedule distribution?

Exercise 12.4. Changes in the work schedule and educational attainment are related to each other and both will effect the distribution of wages. Follow through the analysis of Section 12.3 adjusting for both the number of years of education and hours worked. You will need to use a bivariate adjustment to do this. Use a block adjustment and then consider both sequential adjustments. Does the effect of changes in the work schedule distribution depend on changes in the distribution of educational attainment?

Exercise 12.5. A third factor that may be important is change in the age distribution of the workers. Repeat Exercise 12.4 adjusting for the number of years of education, hours worked, and age. Are changes in the work schedule a major factor after the other two factors have been adjusted for?

Exercise 12.6. Answer Exercise 12.3 for white women rather than white men.

Exercise 12.7. Answer Exercise 12.5 for black women rather than white men.

Exercise 12.8. Changes in the sex composition of the work force may be an important factor also. Consider the population of white workers (i.e., pooling the women's and men's samples). Follow through the analysis of Section 12.3 adjusting for changes in the sex distribution in addition to hours worked. Is the effect of changes in the work schedule distribution dependent on changes in the sex distribution?

Exercise 12.9. In addition to sex, changes in the race composition of the work force may be important. Consider the population of all workers (i.e., pooling the samples by race and sex). Follow through the analysis of Section 12.3 adjusting for changes in the race/sex distribution in addition to hours worked. Is the effect of changes in the work schedule distribution dependent on changes in the race/sex distribution?

Exercise 12.10. The analysis in the chapter focused on the net change from 1980 to 1997. Variations within this time period are not observed. Follow through the analysis of Section 12.3 comparing 1989 to 1980, and then comparing 1997 to 1989. How does the relative distribution of hours worked compare in these two periods? Summarize the effect of changes in the work schedule distribution on the distribution of wages for the two periods.

Chapter 13

Quantile Regression

In this chapter we consider regression models for the relationship between a primary variable of interest and measured covariates. By far the most common regression models used in practice are for the mean. That is, they focus on modeling the mean of the conditional distribution of the target variable given the values of the covariates as a function of the covariate values. However the mean is only one characteristic of the conditional distribution that is of interest. More generally, we wish to compare how other characteristics of the conditional distribution change with changing values of the covariates.

The conditional distributions can be characterized by their quantiles. By choosing particular quantiles, attention can be focused on other aspects of the conditional distributions, such as the upper or lower tail behavior. In this chapter we show how the regression model can be extended to cover modeling of these quantiles.

In the first section we consider the inference for quantiles based on a sample from the target distribution alone. In Section 13.2 we consider the general regression model for a target variable based on covariates. There, models for the quantile function are treated in the same framework as models for the mean function. The paper of Koenker and Bassett (1978) led to a surge in interest in parametric quantile regression. They explored the properties of linear regression models for the quantiles. This model is covered in Section 13.3. While the linear model for quantiles is as useful and interpretable, the conditional quantile functions are rarely linear in multiple quantiles. In Section 13.4 we discuss nonparametric quantile regression models that provide flexible models for more complicated situations.

13.1 Estimation of quantiles

In this section we consider estimation of the quantiles of a distribution based on a random sample from it.

In Section 2.1 the quantile function corresponding to a CDF $F(y)$ was defined to be:

$$Q(p) \equiv F^{-1}(p) = \inf_y \{y \mid F(y) \ \geq \ p \} \qquad 0 < p < 1.$$

We introduce $Q(p)$ to keep the notation simple, and will refer to it as the pth quantile of F. Note that the quantile function is nondecreasing and $Q(r)$ approaches $Q(p)$ as $r < p$ approaches p (i.e., the quantile function is left-continuous). Hence

$$Q(p) \leq y \ \text{ if and only if } \ F(y) \geq p.$$

Let $y_L = \sup\{y \mid F(y) = 0\} \geq -\infty$, and $y_U = \inf\{y \mid F(y) = 1\} \leq \infty$ so that the support of the distribution is $[y_L, y_U]$. We assume throughout that $F(y)$ is absolutely continuous with PDF $f(y) > 0$, $y_L < y < y_U$. In this case $F(y)$ is 1-1 and $Q(p)$ is differentiable for $0 \leq p \leq 1$. The analog of the PDF for the quantile function is the *quantile-density function* $q(p) \equiv Q'(p) = 1/f(Q(p))$ (Parzen 1979). Tukey (1965) calls $q(p)$ the *sparsity function.*

As in Chapter 9, we assume that we have a sample $Y_1, Y_2, \ldots, Y_m$ that are independently and identically distributed from the population distribution F. There we estimated the CDF by the empirical CDF F_m, so it is natural to estimate the quantile function by the inverse of F_m :

$$\hat{Q}_m(p) \equiv F_m^{-1}(p) = \inf_x \{x \mid F_m(x) \ \geq \ p \} \qquad 0 < p < 1$$

We will refer to $\hat{Q}_m(p)$ as the *sample* or *empirical* pth *quantile* of F. Note that the CDF of $\hat{Q}_m(p)$ is

$$P[\hat{Q}_m(p) \leq y] = P[F_m(y) \geq p] = \sum_{j=\lceil mp}^{m} \binom{m}{j} [F(y)]^j [1 - F(y)]^{m-j},$$

where $\lceil x$ is the least integer at least as large as x. If F has a density then $\hat{Q}_m(p)$ has a density that can be obtained by differentiating this function.

As these results suggest, the properties of $\hat{Q}_m(p)$ as an estimator of $Q(p)$ are closely tied to the properties of F_m as an estimator of F. We can asymptotically approximate the distribution of $\hat{Q}_m(p)$ for each individual p :

Theorem. *Assume that* $0 < p < 1$. *Suppose* $F(y)$ *possesses a density* $f(y)$ *in a neighborhood of* $Q(p)$ *that is positive and continuous at* $Q(p)$, *then*

$$\hat{Q}_m(p) \sim AN\left\{ Q(p), \frac{p(1-p)q(p)}{m} \right\} \tag{13.1}$$

as $m \to \infty$.

A similar result holds for the joint estimation of quantiles. For example, if $0 < r < p < 1$, the bivariate distribution of $(\hat{Q}_m(p), \hat{Q}_m(r))$ is asymptotically normal with correlation $\sqrt{r(1-p)}/\sqrt{p(1-r)}$.

This result shows that there is convergence for each value of r individually. To measure the global closeness of $\hat{Q}_m(p)$ to $Q(p)$, we can again use the Kolmogorov-Smirnov distance

$$D_m = \sup_{0<p<1} |\hat{Q}_m(p) - Q(p)|.$$

The convergence of $\hat{Q}_m(p)$ to $Q(p)$ occurs simultaneously for all p only if both y_L and y_U are finite. In this case D_m converges to zero with probability one if $\sup_{0\le p\le 1} q(p) < \infty$. This result suggests that for large sample sizes the deviation between $\hat{Q}_m(p)$ and $Q(p)$ will be small for all p. The empirical quantile function and other estimators of the quantile function have been extensively studied – see Csörgő (1983) and Serfling (1980).

While the empirical quantile function is simple to calculate, smoother estimators of the quantile function have more desirable statistical properties. A smooth estimator of the quantile function can lead to a smooth estimator of both the quantile and the quantile density functions. An estimate of the quantile density function would be necessary for estimating the asymptotic variance of the empirical quantile function. In addition it represents an interesting reexpression of the distributional shape. Cheng and Parzen (1997) propose a natural class of smooth estimators:

$$\hat{Q}_m(p) = \int_0^1 \hat{Q}_m(r) d_r K_m(p, r) \quad 0 < p < 1,$$

where for each $p, K_m(p, \cdot)$ is a CDF on $[0, 1]$. If $K_m(p, r)$ is a point mass at $r = p$, then the estimator is the empirical quantile function. Smoother choices of $K_m(p, r)$ result in smoother estimators for the quantile function. The quantile density function can then be estimated by the derivative of $\hat{Q}_m(p)$:

$$\hat{q}_m(p) = \frac{d}{dp}\hat{Q}_m(p) \quad 0 < p < 1.$$

The properties of these estimators depend on the choice of $K_m(p, r)$. One common choice is the difference kernel:

$$d_r K_m(p, r) = K\left(\frac{p - r}{h}\right) dr,$$

where $K(\cdot)$ is a kernel function given in (9.10) and h is a bandwidth which decreases to zero as the sample size increases. For appropriate choices of $K_m(p, r)$ these estimators outperform their empirical counterparts. See Parzen (1979) and Cheng (1998) for a discussion of the statistical properties of these estimators.

Confidence intervals for $Q(p)$ will be ambiguous if there are values $y > Q(p)$ for which $F(y) = p$. We will assume that $F(y)$ is strictly increasing in a neighborhood of $Q(p)$, that is, there exist $a < Q(p) < b$ such that $F(x) < F(y)$ if $a < x < y < b$ to the right of $Q(p)$. Based on (13.1) we can

construct a confidence interval for $Q(p)$ using the normal approximation to the empirical quantile function. However this requires an estimate of $q(p)$ to calculate the width. As an alternative consider a confidence interval of the form $Y_{(L)}$ to $Y_{(K)}$ where $Y_{(1)} \leq \ldots \leq Y_{(m)}$ are the ordered data values, and $1 \leq L < K \leq m$ are integers. If $U_j = F(Y_j)$ then $U_1, \ldots, U_m$ are independent and uniform on $[0, 1]$. Let $U_{(1)} \leq \ldots \leq U_{(m)}$ be the ordered values so that $U_{(j)} = F(Y_{(j)})$. Note that

$$P(Y_{(L)} \leq Q(p) \leq Y_{(K)}) = P(U_{(L)} \leq p \leq U_{(K)}),$$

which is independent of F. We can then choose K and L to have the correct coverage (e.g., 95%) and also to produce a small interval. One rule for constructing an approximate $100(1-\alpha)\%$ confidence interval is to choose

$$L = mp + Z_{\alpha/2}\sqrt{mp(1-p)} \qquad K = 2mp - L,$$

where $Z_{\alpha/2}$ is the $\alpha/2$th quantile of a standard normal distribution. The L should be rounded down to an integer. It can be shown that this interval has asymptotically the same properties as the interval based on (13.1) when $q(p)$ is known.

Figure 13.1 is a plot of the empirical quantile function for the log of women's earnings in 1997 (see Section 2.1).

The median log-earnings of $Q(\frac{1}{2})$ appears to be about 10.1. The median earnings is then just $\exp(10.10) = \$24,239$. The sample size is $m = 14,341$ and for a 95% confidence for the median $K = 7053, L = 7288$. The interval is then from 10.090 to 10.109. Figure 13.2 is a plot of the density quantile function based on the log-spline estimator in (9.12). The function increases dramatically in the tails because the quantile function is changing quickly there.

The empirical quantile function estimates of the 10% and 90% quantiles are 9.38 and 10.80, so the 90–10% range is 1.42. An approximate 95% confidence for this range can be based on (13.1) if we estimate $q(0.1)$ and $q(0.9)$. If we use the log-spline estimator in Figure 13.2, the interval runs from 1.41 to 1.45.

13.2 Motivation for quantile regression

Consider the situation where, in addition to the variable of interest, we observe covariates on the individuals and the impact of these covariates on the response is of interest. Let Y represent the variable of interest, which we call the *target variable,* and let X represent the values of the *covariates.* The covariates can be both discrete and continuous. In Chapter 7, we developed a distributional technique to adjust the distribution of the target variable for differences in the distribution of the covariates between populations. Here we focus on a single population, and wish to understand

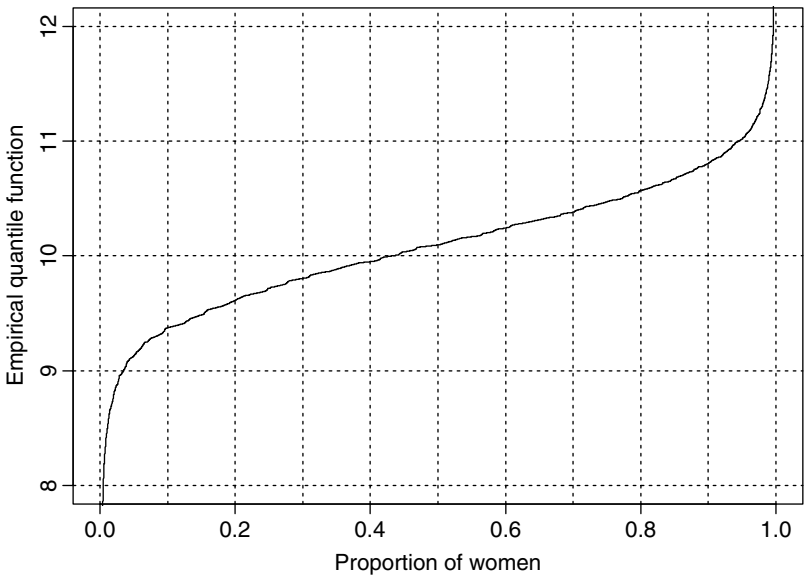

Fig. 13.1. The empirical quantile function for the log-earnings for women in 1997 from the 1998 CPS.

the relationship between the target variable and the covariates with the objective of predicting the target variable from the covariates.

Let Y_j and X_{jk} be the target value and kth covariate, respectively, for the jth individual. Let $X_j = (1, X_{j1}, \ldots, X_{jK})$ be the vector of K covariates for the jth individual augmented by a constant term. The data is then $\{Y_j, X_j\}_{j=1}^m$, where the target and the covariates are measures on the same member of the population.

Regression models form a framework for modeling Y as a function of the values of the covariates X. The objective in this framework is to predict the values of the target variable for given values of the covariates, and conduct inference about the parameters of the model.

Let $F(y|x)$ be the CDF of the conditional distribution of Y given $X = x$. The model for the relationship takes the form:

$$Y_j = \theta(X_j) + \epsilon_j \quad j = 1, \ldots, m \tag{13.2}$$

for some unknown function $\theta(x)$. The ϵ_j represent the random variation of Y from its functional relationship with X and are referred to, unpejoratively, as *errors*. We will assume throughout this chapter that the errors

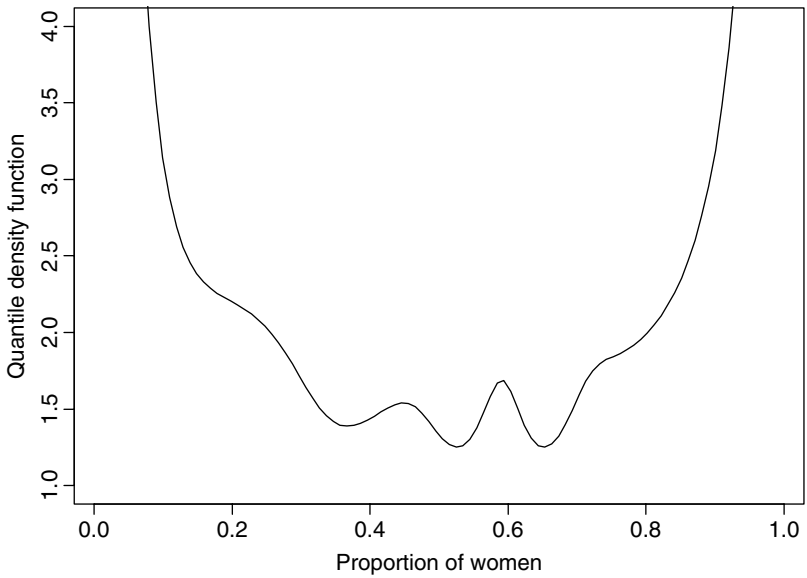

Fig. 13.2. The quantile density function for the log-earnings for women in 1997 from the 1998 CPS.

are conditionally independent and identically distributed given $X = x$. We also assume that the errors are independent of the covariates. Clearly, each of these assumptions can be weakened or altered to different circumstances. Denote the conditional distribution of ϵ_j given $X_j = x$ by $E(y|x)$. Formally,

$$P[\epsilon_j \leq \epsilon | X = x] = E(\epsilon|x) \quad j = 1, \ldots, m$$

These regression models specify that the conditional distributions satisfy the relationship:

$$F(y|x) = E(y - \theta(x)|x) \quad \text{for all } y \text{ and } x. \tag{13.3}$$

To further define $\theta(x)$ we need to clarify its relationship to the errors.

The most commonly studied model is based on representing the functional relationship between the covariates and the conditional mean of the target variable given the covariates. For this version of (13.2) the error distribution $E(y|x)$ is defined to have zero mean for each x. Hence $\theta(x)$ is the conditional expectation of Y given $X = x$, $m(x) = \mathrm{E}[Y|X = x]$. Usually, it is also assumed that $E(y|x)$ has variance $\sigma^2(x)$.

The regression model can be reexpressed in terms of the quantiles. The conditional quantile function of Y given $X = x$ is:

$$Q_p(x) \equiv \inf_y \{y \mid F(y|x) \ \geq \ p \} \qquad 0 < p < 1.$$

Denote the conditional quantile function of the errors by

$$U_p(x) \equiv \inf_y \{y \mid E(y|x) \ \geq \ p \} \qquad 0 < p < 1.$$

As we assume that $F(y|x)$ is absolutely continuous and strictly increasing on its support, the model (13.3) can be reexpressed as:

$$Q_p(x) = \theta(x) + U_p(x) \quad \text{for all } p \text{ and } x. \tag{13.4}$$

The error distributions can sometimes be assumed to differ only in terms of a multiplicative scale factor, so that the data follow a *location-scale model:*

$$Y_j = \theta(X_j) + \sigma(X_j)\upsilon_j \qquad j = 1, \ldots, m,$$

where the υ_j are independent and identically distributed with unknown quantile function $U(p)$ independent of X_j. The υ_j are assumed to have unit scale, and the scale factor $\sigma(x) > 0$. The location-scale model can be expressed as:

$$U_p(x) = \sigma(x)U(p) \quad \text{for all } p \text{ and } x.$$

A further simplification is the *location model* where $E(y|x)$ is completely independent of x, that is,

$$U_p(x) = \sigma U(p) \quad \text{for all } p \text{ and } x,$$

for some $\sigma > 0$.

For the regression model for the mean, the constraint on the errors can also be stated as $\int_0^1 U_p(x)\ dp = 0$. If it is also assumed that $E(y|x)$ has variance $\sigma^2(x)$, then $\int_0^1 U_p^2(x) dp = \sigma^2(x)$.

The most commonly used regression model is *linear regression* where

$$m(x) = \beta_0 + \beta_1 x_1 + \cdots \beta_K x_K = x\beta, \tag{13.5}$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_K)^T$ are the *regression coefficients* and $x = (1, x_0, x_1, \ldots, x_K)^T$. The classical linear regression model for the mean has independent and identically distributed normal errors. This is the location model with $U(p)$ as the quantile function of the standard normal distribution. If this model is specified correctly then inference for the parameters β, σ and prediction can be done within the likelihood or least-squares framework. We will not develop them here - see von Eye and Schuster (1998) for a book length treatment.

In general, the regression model for the mean can be viewed as inference for $m(x)$, the mean of the conditional distribution of the response, where the influence of the observations with values of the covariate not equal to x is

determined by the assumptions made about $m(x)$ and the error terms. The linear regression model assumes that $m(x)$ has a parametric linear form. If $m(x)$ is only assumed to be a smooth function of x the situation is referred to as *nonparametric regression*. In this case, observations with values of the covariate close to x should have similar values of the mean, and information can be borrowed from them to infer $m(x)$, and vice versa. In both the linear and nonparametric situations, the theory for the mean regression model has been well developed. See Simonoff (1996) for a discussion.

The mean is only one characteristic of the conditional distribution of the response. It is of interest to model other characteristics such as the quantiles. Hogg (1975) has argued for the use of conditional quantile functions as descriptions of distributional change within the regression context.

Let $p \in (0,1)$ be a given quantile of interest. Suppose that instead of defining the regression model for $m(x)$ by constraining the mean of $E(y|x)$ to be zero, the pth quantile of the error distribution is constrained to be zero: $E(0|x) = p$ for each x. Then $\theta(x)$ is $Q_p(x)$, the conditional pth quantile of Y given $X = x$. The model (13.1) for the relationship takes the form:

$$Y_j = Q_p(X_j) + \epsilon_j \quad j = 1, \ldots, m, \tag{13.6}$$

where the conditional quantile functions of the errors satisfy the constraint $U_p(x) = 0$ for each x. The relationship (13.6) along with the constraint on the errors define the general quantile regression model.

Interest in the quantiles can also be motivated by doing a search for location measures $\theta(x)$ that are robust to the distribution of the error terms. Under the location model, the mean function $m(x)$ is the value of $\theta(x)$ that minimizes the squared error loss $\mathrm{E}[L(Y - \theta(X))|X = x]$ where $L(y) = y^2$. Under the same model, $Q_p(x)$ minimizes the loss $\mathrm{E}[L_p(Y - \theta(X))|X = x]$ relative to the asymmetric absolute loss function $L_p(y) = |y| + (2p - 1)y$. If $p = \frac{1}{2}$ this is just the absolute loss function $L_{\frac{1}{2}}(y) = |y|$, so that $Q_{\frac{1}{2}}(x)$ is the conditional median function of Y. If the conditional distribution of Y given $X = x$ is very asymmetric or heavy-tailed, the median may be a better summary of location than the mean.

When the error terms are both heteroscedastic and asymmetric then a wide range of loss functions result in interesting forms for $\theta(x)$. Aigner, Amemiya, and Poirier (1976) and Newey and Powell (1987) use asymmetric squared error loss ($L(y) = |p - \mathcal{I}(y \leq 0)|y^2$) to define a location measure they call *expectiles*. Although they are less interpretable than the quantiles, inference for them is somewhat easier than for the asymmetric absolute loss. Efron (1991) extends this work by estimating the quantiles based on the expectiles. His estimators have the computational advantages of the expectiles and greater efficiency when the errors are normal. He advocates location measures based on the class of asymmetric power loss functions: $L_{p,\alpha}(y) = |p - \mathcal{I}(y \leq 0)|y^\alpha,\ 0 < p < 1,\ \alpha > 0$. This includes the above loss functions as special cases. He argues for values $0 \leq \alpha \leq 2$ with the smaller values being more robust and the higher values having greater efficiency

when the errors are normal-like. In the remainder of this chapter we focus on quantiles estimated via asymmetric absolute loss, although much of the development is applicable with other choices of loss function.

For fixed p, the conditional quantile function $Q_p(x)$ plays the same role as the conditional mean function $m(x)$ in the nonparametric regression model for the mean. Each choice of p focuses attention on a different characteristic of the conditional distribution of Y given $X = x$. The choice of p will vary from application to application. Often the median ($p = \frac{1}{2}$) is the most natural, or the quartiles ($p = 0.25, 0.75$). Some interesting guidance and examples are given by Efron (1991) and Buchinsky (1998).

In the next sections we consider models for the conditional quantile function that parallel those developed for the conditional mean function.

13.3 Linear quantile regression

The linear quantile regression model is linear regression for a given quantile of the conditional distribution of Y given $X = x$. In addition to the general model in (13.6), the quantile is assumed to be a linear function of the covariates:

$$Q_p(x) = x\beta_p \qquad x \in D_X, \tag{13.7}$$

where D_X is the range of applicability of the model. Note that the regression coefficients β_p depend on the quantile p. Expressed in this form, the model may not hold for multiple p, except under special conditions on the error terms.

For given p, the regression coefficients can be interpreted in much the same way as they are in regression for the mean function. We can interpret β_{pk} as the change in the pth quantile due to a unit change in the kth covariate, holding the values of the other covariates fixed.

Recall in linear regression for the mean that the regression coefficients are characterized by minimizing the squared-error loss. For quantile regression, $\theta(x) = Q_p(x) = x\beta_p$ minimizes the loss $\mathrm{E}[L_p(Y - \theta(X))|X = x]$ for each x. Thus the vector of coefficients β_p is the value of γ that minimizes the loss $\mathrm{E}_X\Big[\mathrm{E}[L_p(Y - X\gamma)|X]\Big]$, overall. If the covariates are nonrandom, then the outer expectation is redundant. There will be more on this later.

Following the analogy to least-squares regression, the natural estimator for β_p is $\hat{\beta}_p$, the value of γ that minimizes

$$\sum_{j=1}^{m} L_p(y_j - x_j\gamma). \tag{13.9}$$

This is also the method of moments estimator of β_p. The natural estimator of $Q_p(x)$ is then $x\hat{\beta}_p$.

The statistical properties of $\hat{\beta}_p$ have been studied under a variety of conditions. In almost all social science applications, the data arise from observational studies or sample surveys where some of the covariates are not under the control of the researcher. The most common situation is where the data $\{Y_j, X_j\}_{j=1}^m$ are a, possibly stratified, random sample from the population of interest. Under these conditions the randomness of the covariates should be taken into account when the statistical properties of $\hat{\beta}_p$ are evaluated. We are assuming throughout that the covariates are independent of the error terms. Even though the model is specified conditionally on the covariates, the statistical properties of the estimates will clearly be different if we condition on the values of the covariates actually observed rather than take their statistical variation in to account. For ease of exposition we will condition on the observed values of the covariates when describing the properties of the estimators. To avoid degeneracies, it is necessary to place mild restrictions on the values the covariates take as the sample size increases. In particular, we will assume that $\frac{1}{m}X^TX$ is a nonsingular matrix and approaches a positive definite matrix as the sample size increases. This assumption is standard in linear regression for the mean. In the Background material we discuss how these assumptions can be weakened.

Let $u_p(x)$ be the quantile density function corresponding to $U_p(x)$. Suppose first that the errors are identically distributed so that the data follow the location model and hence $u_p(x) = u_p$ independent of x. In this case the quantile functions for different p are all parallel with slopes $\beta_k(p) = \beta_k,\ k = 1, \ldots, K$ and the intercepts are $\beta_0(p) = U_p$. The simplest result is:

Theorem. *Suppose that* $0 < p < 1$ *and* $(Y_j, X_j)_{j=1}^m$ *follow the linear quantile regression model (13.7) and the distribution of the errors is independent of* x. *If* u_p *is positive and continuous in a neighborhood of* p *then, as* $m \to \infty$,

$$\hat{\beta}_p \sim AN\left\{\beta_p, \frac{\sigma_p^2}{m}[X^TX]^{-1}\right\} \tag{13.10}$$

where $\sigma_p^2 = p(1-p)u_p^2$.

The result is similar to that of (13.1) in the sense that σ_p^2 is the variance of the quantile estimate based on a sample of size m from the error distribution. This result is similar to that for least-squares regression with a redefined σ_p^2.

This result can be extended to the case where the error distribution depends on x:

Theorem. *Suppose that* $0 < p < 1$ *and* $(Y_j, X_j)_{j=1}^m$ *follow the linear quantile regression model (13.7). If* $u_p(x)$ *is positive and continuous in a neighborhood of* p *for each* x *then*

$$\hat{\beta}_p \sim AN\left\{\beta_p, \frac{p(1-p)}{m}[X^TDX]^{-1}[X^TX][X^TDX]^{-1}\right\} \tag{13.11}$$

as $m \to \infty$. *Here* D *is the diagonal matrix with* jjth *element* $1/u_p^2(x_j)$.

As $\hat{Q}_p(x)$ is a linear combination of the estimated regression coefficients, it is also asymptotically unbiased and normal.

Confidence intervals for the regression coefficients and regression quantiles can be based on these results. Applying them in practice requires an estimate of $u_p(x)$, the conditional quantile density function. The approaches in Section 13.1 can each be applied here. In particular, it is also possible to use the order statistics to define confidence intervals. A number of bootstrap estimators for the covariance matrix have also been developed. Buchinsky (1995) compares these and other approaches. He finds that a bootstrap estimator performs the best in the general setting, but requires extensive computational time.

If the model is assumed to hold simultaneously for multiple quantiles then either D_X, the plausible range of values the covariates can take, is limited or the quantile functions are parallel so that the errors are independent of x. Otherwise, if $0 < r < p < 1$ are two values for which the model hold there will be values of the covariates for which $Q_p(x) = x\beta(p)$ will be less than $Q_r(x) = x\beta(r)$. If the above estimation procedure is used for both r and p then the estimates will be correlated. The two above results can be generalized to show that $\hat{\beta}(p)$ and $\hat{\beta}(p)$ are jointly asymptotically normal with the natural covariance matrix. See Powell (1986) and Koenker and Bassett (1978) for details. The problem here is that the estimates for each quantile are computed separately. Thus even if the errors are identically distributed, the estimated curves may cross in the domain of the covariates. For particular classes of error distributions, Koenker (1984), Cole and Green (1992) and others consider alternative estimation schemes to ensure the estimated curves do not cross. He (1994) proposes a procedure based on the location-scale model called *restricted regression quantiles* that produces estimates that do not cross. that produces

The location model is very restrictive. It is natural to consider the class of linear heteroscedastic models:

$$U_p(x) = (x\gamma)U(p) \quad \text{for all } p \text{ and } x,$$

where $x\gamma > 0$ for $x \in D_X$. This is a location-scale model where the scale function $\sigma(x) = x\gamma$ is linear in the covariates. Such models have been studied by Gutenbrunner and Jureckova (1992), Koenker and Zhao (1994), and He (1997).

In Chapter 4 we studied the distributions of annual earnings of white males from 1967–1997 using relative distribution methods. Figure 4.2 is the running boxplot of the earnings by year. In Figure 13.3 the empirical quantiles (13.1) are plotted for $p = 5\%, 25\%, 50\%, 75\%$, and 95%. Superimposed over the empirical quantiles are the quantile estimates based on the linear model. The linear model appears to fit quite well for each of the quantiles. Note, however, the deviations between the empirical quantiles and the lines

are greater than expected from statistical variation, and so there is some lack of fit. As a general pattern, we can see the increase in spread of the distributions over time with greater changes for the extreme quantiles.

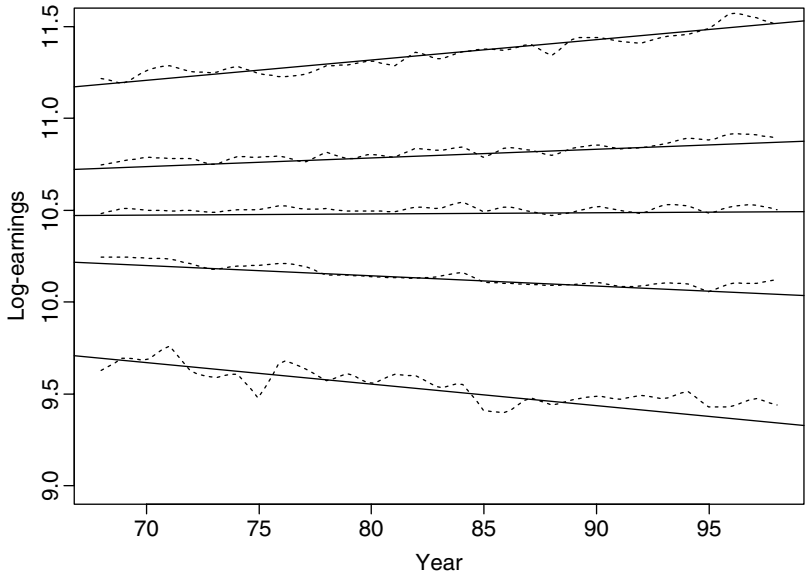

Fig. 13.3. Empirical quantiles and linear quantile estimates for the annual earnings distributions: 1967–1997. The quantiles shown are $p = 5\%$, 25%, 50%, 75%, and 95%.

13.4 Nonparametric quantile regression

The linear models of the previous section have many advantages, including ease of interpretation and computation. However they place very restrictive assumptions on the conditional quantile functions, especially if the model is assumed to hold for multiple quantiles. Given the success of nonparametric regression methods for the mean function, it seems natural to consider nonparametric estimation of $Q_p(x)$. The various *nonparametric quantile regression* models only assume that $Q_p(x)$ is a smooth function of x.

Cheng (1983; 1984) described the asymptotic properties of kernel density estimators (Section 9.3.2). Stute (1986) considered nearest neighbor kernel estimators when the covariate distribution is random. Janssen and

Veraverbeke (1987), and Lejeune and Sarda (1988) also proposed nonparametric estimators based on kernel or local polynomial ideas.

In the context of growth measurement for biomedical studies, Cole (1988) proposed an estimation method based on transforming the conditional distributions to an approximately normal form. In the discussion to that paper, Cox and Jones introduced a form of smoothing spline model for the conditional quantile functions. Koenker, *et al* (1992) proposed smoothing spline models based on L_1 and L_∞ smoothness norms. An indepth study of quantile smoothing spline models was undertaken in Koenker, Ng, and Portnoy (1994). He and Shi (1994) considered a related approach. Ng (1996) showed how monotonicity and convexity constraints could be imposed, and demonstrated a computationally efficient algorithm.

Non-parametric estimation using local polynomial methods have been extensively used for estimating the mean function. Their advantages extend to the quantile case. Chaudhuri (1991) developed the asymptotic theory for local polynomial estimators (Section 9.3.3). Fan, *et al* (1994) and Fan, Yao, and Tong (1996) further developed the theoretical ideas. Fan and Gijbels (1996) provide a book length treatment of local polynomial regression, and discuss quantile estimation. We follow Yu and Jones (1998) in the development below.

As in the situation of nonparametric estimation for the mean function, each of these methods have their strengths and weaknesses. Often theoretical advantages are lost in the sea of practical details. In this section we will focus on local-polynomial methods, both because of their theoretical strengths and the benefits of the extensive practical knowledge that has been built up about them.

The idea of the local linear fitting is to approximate $Q_p(x)$ by a linear function

$$\tilde{Q}_p(z) = Q_p(x) + q_p(x)(z - x)$$

for values, z, of the covariates close to x. We can then fit a line to the data values with covariates close to x, and estimate $Q_p(x)$ by $\tilde{Q}_p(x)$, the value on the regression line. Of course, data with covariates closer to x should receive more weight than data further away, and we should fit a different model for each value of x we are interested in.

Figure 13.4 displays the local-linear quantiles estimates for $p = 5\%$, $25\%, 50\%, 75\%$, and 95%. In this case they are quite close to the empirical quantiles as the sample sizes are very large – m is about 20,000. The pattern seen here, suggests that the linear model of the previous section is not a perfect fit to the data.

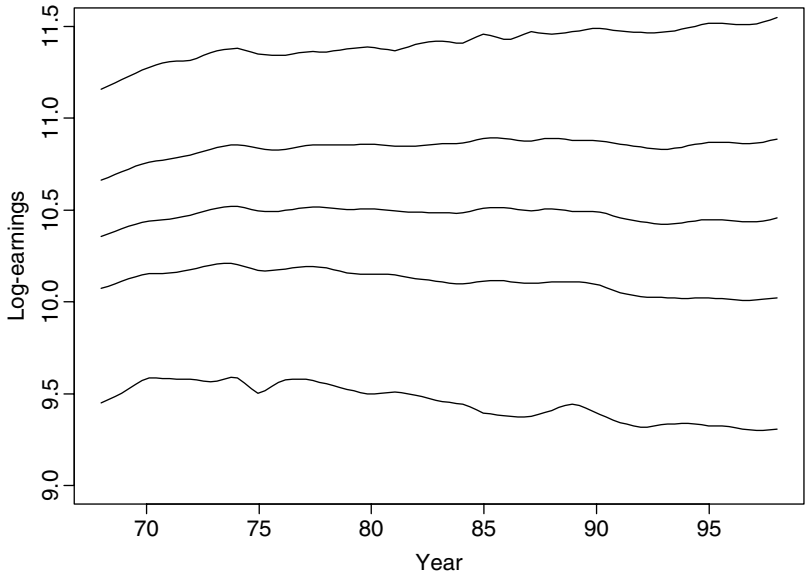

Fig. 13.4. The local linear quantile estimates for the annual earnings distributions: 1967–1997. The quantiles shown are $p = 5\%$, 25%, 50%, 75%, and 95%.

Background material

The study of quantiles has a long history, going back at least to Francis Galton in 1885. Estimation has traditionally been based on the order statistics. Quantile estimation in the regression setting was considered in the seminal paper of Koenker and Bassett (1978). They were motivated by issues of robust estimation in the regression context, and especially the detection of outliers. Koenker and Bassett focused on linear quantile regression. Subsequently Ruppert and Carroll (1980), Koenker and Portnoy (1987), and others further developed the theory. Powell (1986) considers the extension to censored data. Buchinsky (1998) is a good review paper on the parametric approach and its extensions. The nonparametric quantile regression model was considered by Janssen and Veraverbeke (1987), Lejeune and Sarda (1988) and Chaudhuri (1991). Yu and Jones (1998) provide a description of the local-linear approach to quantile estimation. Together with Buchinsky (1998), it provides a good review of the state-of-the-art of the practice of quantile regression.

It is not necessary to assume that the errors ϵ_j are independent of X_j.

In the general case, estimation can be based on the generalized method of moments (Powell 1986). Issues relating to endogeneity and longitudinality are similar to those in the mean regression situation. See Chamberlain (1984) for an introduction.

One of the insights of Koenker and Bassett (1978) was that the minimization problem (13.9) could in fact be phrased as a linear programming problem with an expanded parameter space. Using standard results and for a given data set, they showed the existence and optimality of solutions to the minimization problem. The estimates themselves could be obtained by efficient and numerically stable linear programming algorithms. The representation also made it clear that the solutions were robust to outliers in the target variable, in much the same way as the sample quantiles are robust to changes in data values. That is, the estimate is unchanged by any changes in y_j that do not change the sign of $y_j - x_j\hat{\beta}_p$. Gutenbrunner and Jureckova (1992) gave a statistical interpretation to the solutions to the dual of the linear programming problem that extended the duality of order statistics and ranks to the linear regression setting. For additional information about the linear programming representation, see Buchinsky (1998).

Exercises

Exercise 13.1. Derive the formula for the CDF of $\hat{Q}_m(p)$ given in Section 13.1. You can use the fact given in Section 9.2 that $mF_m(y)$ is a binomial random variable on m trials and with probability of success $F(y)$.

Exercise 13.2. Derive the formula for the PDF of $\hat{Q}_m(p)$ based on the result for the CDF given in Section 13.1. Give a heuristic rationale for its relationship to the binomial probability mass function.

Exercise 13.3. Derive the relationship (13.3) from the definition of the regression model.

Exercise 13.4. Show that the median is a value that minimizes the mean absolute deviation of a distribution around that value.

Exercise 13.5. Give an example where regression for the median function is more appropriate substantively than regression for the mean function. Why do you think the former is rarely used in practice?

Exercise 13.6. Use the definition of the regression model to show the equivalence of the formulations given in (13.3) and (13.4).

Exercise 13.7. Construct the empirical quantile function for men's log-earnings in 1997. How does it compare to that for women given in Figure 13.1.

Exercise 13.8. Fit a linear quantile regression model to the median of men's log-earnings from 1967-1997, using year as the independent variable. Fit

models for the $p = 5\%$, 25%, 75%, and 95% quantiles and graph them in a plot similar to Figure 13.3.

Exercise 13.9. Fit a linear quantile regression model to the median of the population log-earnings from 1967-1997 using year, and race/sex group as covariates. Compare the coefficients of the four groups. Repeat the analysis for the $p = 5\%$, 25%, 75%, and 95% quantiles.

Exercise 13.10. Refit the linear quantile regression model to the median of the population log-earnings from 1967-1997. Include an interaction term between race, sex and year. Which of these factors appears to have the greater effect? Do significant interactions exist? Answer the same questions for the $p = 5\%$, 25%, 75%, and 95% quantiles.

Appendices

A. Descriptions of the data sets

The data sets can be obtained electronically over the World Wide Web, by connecting to the Relative Distribution website. A link to the website is maintained by the publisher at:

`http://www.springer-ny.com/stats`

under the heading "Author/Editor Home Pages." The website contains descriptions of the variables and data file formats. The data files are S-PLUS and SAS system files, so it should be possible to import them into virtually any statistical, database management or spreadsheet package.

B. More on computational issues

Many of the computational issues relevant to relative distribution methods are generic to density estimation, and smoothing methods in general. Hence computer code for density estimation can be coopted for relative density estimation. Sources for computer code are given in the "Computational issues" sections of the chapters. However the transient nature of the sources and locations make a comprehensive listing untenable. As an alternative, the relative distribution website contains links to software that the authors found useful to implement relative distribution methods. These links will be updated to ensure they are active. We welcome additions, corrections and updates to this information by authors or readers.

S-PLUS was used to construct all the figures and numerical summaries in this book, so we have focused on S-PLUS-related software. Discussion of any software does not imply any endorsement of any kind about that software, and we provide no warranty of any kind on the correctness or usefulness of any software mentioned, or provided, or of the accuracy of our descriptions of the software. Users of any software should consider the software as being used at their own risk.

C. Estimation of permanent wages and wage growth

Permanent wages are unobserved, and estimated using a mixed effects model for the age-earnings profile.

For each respondent, the profile is represented as:

$$y_{it} = b_{0i} + b_{1i}\text{age}_t + b_{2i}\text{age}_t^2,$$

where y_{it} is the log of real (PCE-deflated) permanent wages for respondent i at time t. Each of the coefficients b_{0i}, b_{1i}, and b_{2i} represent a combination of the fixed and random effects for the lifecycle growth in wages:

$$b_{ji} = \beta_j + \tau_{ji}, \qquad \tau_{ji} \sim N(0, \sigma_j^2) \quad j = 0, 1, 2$$

where the β_j are the "fixed effects", and the τ_{ji} are independent random draws from normal distributions. The fixed-effects quadratic in age, $\beta_0 + \beta_1\text{age}_t + \beta_2\text{age}_t^2$, captures the mean growth in wages over the life-cycle, and the τ_{ji} capture the heterogeneity in individual profiles. For the motivation of these models and discussion of alternative specifications cf., Gottschalk and Moffitt (1994) and Haider (1997). Further details of the wage estimation procedure can be found in Bernhardt, *et al* (1999).

We use this mixed-effects specification to fit a wage profile for each respondent that covers the 16–34 year old age range. The wage gain for each respondent is then defined as

$$w_i = (y_{it} \mid \text{age}_t = 34) - (y_{it} \mid \text{age}_t = 16) \qquad i = 1, \ldots, n.$$

The observed log-wages, z_{it}, are modeled as:

$$z_{it} = y_{it} + \epsilon_{it},$$

where $\epsilon_{it}, i = 1, \ldots, n, t = 1, \ldots, 15$ are i.i.d. $N(0, \sigma^2)$ with unknown $\sigma > 0$. The values used throughout Chapter 8 to represent wage growth for an individual are the empirical Bayes estimates of w_i based on this random effects model. For a discussion see, Diggle, *et al* (1994), Section 5.6.

All respondents with two or more valid wage observations are used during the wage estimation procedure. For this analysis, we restrict the sample to respondents not lost to attrition.

D. Proof of results in Chapter 9

In this appendix we sketch proofs of results about the estimates of the relative CDF and the relative density given in Chapter 9. The proofs were originally given in Handcock and Janssen (1998b).

The asymptotic distribution of multivariate U-statistics with estimated parameters

To illustrate the statistical behavior of the estimator (9.17), first consider estimating the joint distribution of $F_m(\lambda_r)$ and $F_m(\lambda_s)$ with $\lambda_\nu = F_0^{-1}(\nu)$ for $0 < \nu < 1$, based on the bivariate comparison sample U-statistic

$$U_m(\gamma) \quad = \quad \frac{1}{m}\sum_{j=1}^{m} h(Y_j;\gamma) \quad = \quad \{F_m(\gamma_1), F_m(\gamma_2)\}$$

with kernel

$$h(Y;\gamma) \quad = \quad \{\mathcal{I}(Y \le \gamma_1), \mathcal{I}(Y \le \gamma_2)\}.$$

The expectation of $U_m(\gamma)$ is

$$\theta(\gamma) = \{F(\gamma_1), F(\gamma_2)\}.$$

Suppose that $F(x)$ is differentiable at $x = \lambda_r$ and $x = \lambda_s$ and $f(\lambda_r), f(\lambda_s) > 0$ then the joint asymptotic distribution of $m^{\frac{1}{2}}[U_m(\lambda) - \theta(\lambda)]$ is well known (see Serfling 1980). In our situation $\lambda = (\lambda_r, \lambda_s) = (F_0^{-1}(r), F_0^{-1}(s))$ is unknown and it is estimated by $\hat{\lambda}_n = (F_{n0}^{-1}(r), F_{n0}^{-1}(s))$. The standard results for U-statistics do not apply to $m^{\frac{1}{2}}[U_m(\hat{\lambda}_n) - \theta(\lambda)]$. However a multivariate Gaussian limit can be obtained from the general expansion

$$m^{\frac{1}{2}}[U_m(\hat{\lambda}_n) - \theta(\lambda)] \;=\; m^{\frac{1}{2}}[U_m(\lambda) - \theta(\lambda)] \;+\; m^{\frac{1}{2}}(\hat{\lambda}_n - \lambda)'\bigtriangledown\theta(\lambda) \;+\; o_p(1),$$

where $\bigtriangledown\theta(\lambda)$ is the 2×2 matrix:

$$\Big\{\frac{\partial\theta(\cdot)}{\partial\gamma_1}, \frac{\partial\theta(\cdot)}{\partial\gamma_2}\Big\}'\Big|_{\gamma=\lambda}.$$

This expansion is used by Randles (1982) in the one-sample situation (i.e., $\hat{\lambda}_n$ and the univariate U-statistic are based on the same sample). For our two-sample situation we will need a generalization of Randles' result which we state in full generality as Lemma D.1.

Following Randles (1982), let $h(Y_1, Y_2, \ldots, Y_r;\gamma)$ be a multivariate symmetric kernel of degree r based on the sample $Y_1, Y_2, \ldots, Y_m$ and with the expected value

$$\theta(\gamma) = \mathrm{E}_\lambda[h(Y_1, Y_2, \ldots, Y_r;\gamma)],$$

where λ denotes the true parameter value. Let $U_m(\gamma)$ be a q-variate U-statistic corresponding to $h(\cdot;\gamma)$. Let $\hat{\lambda}_n$ be an estimate of λ based on the first sample $\{Y_{01}, Y_{02}, \ldots, Y_{0n}\}$.

Condition D.1 Suppose there is a neighborhood of λ, call it $K(\lambda)$, and a constant $K_1 > 0$, such that if $\gamma \in K(\lambda)$ and $D(\gamma, d)$ is a sphere centered at γ with radius d satisfying $D(\gamma, d) \subset K(\lambda)$, then

$$\mathrm{E}[\sup_{\gamma' \in D(\gamma,d)} |h(Y_1, Y_2, \ldots, Y_r; \gamma') - h(Y_1, Y_2, \ldots, Y_r; \gamma)|] \leq K_1 d$$

and

$$\lim_{d \to 0} \mathrm{E}[\sup_{\gamma' \in D(\gamma,d)} |h(Y_1, Y_2, \ldots, Y_r; \gamma') - h(Y_1, Y_2, \ldots, Y_r; \gamma)|^2] = 0.$$

Condition D.2A Assume that $\theta(\gamma)$ has a zero differential at $\gamma = \lambda$, that

$$n^{\frac{1}{2}}[\hat{\lambda}_n - \lambda] = O_p(1)$$

and that

$$m^{\frac{1}{2}}[U_m(\lambda) - \theta(\lambda)] \xrightarrow{\mathcal{D}} N(0, \Omega),$$

where

$$\Omega = \mathrm{Var}\big(\mathrm{E}[h(Y_1, Y_2, \ldots, Y_r; \lambda)|Y_1]\big) \tag{A.1}$$

is positive definite, as $m \to \infty, m/n \to \kappa^2 < \infty$.

Condition D.2B: Assume that $\theta(\gamma)$ has a nonzero differential at $\gamma = \lambda$, that

$$m^{\frac{1}{2}}[U_m(\lambda) - \theta(\lambda), (\hat{\lambda}_n - \lambda)'] \xrightarrow{\mathcal{D}} N_{p+q}(0, \Sigma),$$

where

$$\Sigma = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \kappa^2 \Sigma_{22} \end{pmatrix}$$

and

$$\Omega = \mathbf{D}' \Sigma \mathbf{D} \tag{A.2}$$

is positive definite, where

$$\mathbf{D} = \begin{pmatrix} I_{q \times q} \\ \nabla\theta(\lambda) \end{pmatrix}$$

and $\nabla\theta(\lambda)$ is the $p \times q$ matrix:

$$\Big\{ \frac{\partial\theta(\cdot)}{\partial\gamma_1}, \ldots, \frac{\partial\theta(\cdot)}{\partial\gamma_p} \Big\}' \Big|_{\gamma=\lambda}$$

as $m \to \infty, m/n \to \kappa^2 < \infty$.

We then have the following multivariate version of Theorem 2.13 in Randles (1982):

Lemma D.1. If Condition D.1 holds and, in addition, one of Condition D.2A or D.2B holds, then

$$m^{\frac{1}{2}}[U_m(\hat{\lambda}_n) - \theta(\lambda)] \xrightarrow{\mathcal{D}} N(0, \Omega),$$

where Ω is given by (A.1) *or* (A.2)*, respectively.*

The proof of Lemma D.1 follows from the following extension of Theorem 2.8 in Randles (1982) to our two-sample situation: If $n^{\frac{1}{2}}\big(\hat{\lambda}_n - \lambda\big) = O_p(1)$ and Condition D.1 holds, then

$$N^{\frac{1}{2}}[U_m(\hat{\lambda}_n) - \theta(\hat{\lambda}_n) - U_m(\lambda) + \theta(\lambda)] \xrightarrow{P} 0,$$

where $N = n + m$. The proof of this result follows closely that of Theorem 2.8 in Randles (1982).

Sketch of the proof of the result (9.17)
From e.g., Serfling (Serfling 1980, Theorem B, p. 80), we have

$$n^{\frac{1}{2}}[\hat{\lambda}_{nr} - \lambda_r, \hat{\lambda}_{ns} - \lambda_s] \sim \quad \text{AN}\left\{\begin{pmatrix}0\\0\end{pmatrix}, \Sigma_{22}\right\}$$

with

$$\Sigma_{22} = \begin{pmatrix} \frac{r(1-r)}{(f_0(\lambda_r))^2} & \frac{r(1-s)}{f_0(\lambda_r)f_0(\lambda_s)} \\ \frac{r(1-s)}{f_0(\lambda_r)f_0(\lambda_s)} & \frac{s(1-s)}{(f_0(\lambda_s))^2} \end{pmatrix}$$

as $n \to \infty$. From standard results about the sample distribution function:

$$m^{\frac{1}{2}}[U_m(\lambda_r) - \theta(\lambda_r), U_m(\lambda_s) - \theta(\lambda_s)] \sim \quad \text{AN}\left\{\begin{pmatrix}0\\0\end{pmatrix}, \Sigma_{11}\right\}$$

with

$$\Sigma_{11} = \begin{pmatrix} G(r)(1-G(r) & G(r)(1-G(s) \\ G(r)(1-G(s) & G(s)(1-G(s) \end{pmatrix}$$

as $m \to \infty$. Thus Condition D.2B of Lemma D.1 follows with:

$$\mathbf{D} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ f(\lambda_r) & 0 \\ 0 & f(\lambda_s) \end{pmatrix}.$$

Condition D.1 follows because $F(x)$ is differentiable in neighborhoods of λ_r. Thus Lemma D.1 gives the required asymptotic Gaussian distribution and an easy calculation of $\mathbf{D}'\Sigma\mathbf{D}$ gives the covariance expression in (9.20).

∎

Sketch of the proof of result (9.20)
As K is twice differentiable we can expand the estimator:

$$\begin{aligned} g_{n,m}(r) &= \frac{1}{mh_m}\sum_{j=1}^{m} K\left(\frac{r - R_j}{h_m}\right) \\ &= \frac{1}{mh_m}\sum_{j=1}^{m} K\left(\frac{r - F_0(Y_j)}{h_m}\right) \\ &+ \frac{1}{mh_m}\sum_{j=1}^{m} \frac{F_0(Y_j) - F_{n0}(Y_j)}{h_m} K'\left(\frac{r - F_0(Y_j)}{h_m}\right) \qquad (A.3) \\ &+ \frac{1}{mh_m}\sum_{j=1}^{m} \frac{(F_0(Y_j) - F_{n0}(Y_j))^2}{2h_m^2} K''(\Delta_j) \\ &= g_m(r) + T_{n,m}(r) + R_{n,m}(r), \end{aligned}$$

where Δ_j is between $h_m^{-1}(r - F_{n0}(Y_j))$ and $h_m^{-1}(r - F_0(Y_j))$.

Lemma D.2. $\sqrt{mh_m}\ R_{n,m}(r) \xrightarrow{P} 0$ *as* $m \to \infty$.

Sketch of the proof of Lemma D.2

For simplicity assume the support of K is contained in $[-1, 1]$. We can bound $R_{n,m}(r)$ as

$$|R_{n,m}(r)| \leq \frac{1}{mh_m} \sum_{j=1}^{m} \frac{(F_0(Y_j) - F_{n0}(Y_j))^2}{2h_m^2} |K''(\Delta_j)|$$

and express Δ_j as

$$\Delta_j = \frac{r - F_0(Y_j)}{h_m} - \theta_j \frac{F_0(Y_j) - F_{n0}(Y_j)}{h_m},$$

where $0 \leq \theta_j \leq 1$. Note that for $\Delta_j \notin [-1, 1]$ we have $K''(\Delta_j) = 0$, therefore the terms in the sum that are different from zero are those for which $\Delta_j \in [-1, 1]$. Therefore, with $\Delta_{n0} = \sup_t |F_0(t) - F_{n0}(t)|$, we have

$$|R_{n,m}(r)| \leq \frac{1}{mh_m^3} \Delta_{n0}^2 C(K'') \sum_{j=1}^{m} \mathcal{I}\{-1 \leq \Delta_j \leq 1\},$$

where $C(K'')$ is the upper bound for K''. Now

$$-1 \leq \Delta_j \leq 1 \iff r - h_m \leq F_0(Y_j) + \theta_j(F_0(Y_j) - F_{n0}(Y_j)) \leq r + h_m.$$

Therefore

$$\mathcal{I}\{-1 \leq \Delta_j \leq 1\} \leq \mathcal{I}\{r - h_m - \Delta_{n0} \leq F_0(Y_j) \leq r + h_m + \Delta_{n0}\}$$

and

$$\begin{aligned}
|R_{n,m}(r)| &\leq C(K'') \frac{\Delta_{n0}^2}{h_m^3} \frac{1}{m} \sum_{j=1}^{m} \mathcal{I}\{r - h_m - \Delta_{n0} \leq F_0(Y_j) \leq r + h_m + \Delta_{n0}\} \\
&= C(K'') \frac{\Delta_{n0}^2}{h_m^3} \cdot \big\{ G_m\big(r + h_m + \Delta_{n0}\big) - G_m\big(r - h_m - \Delta_{n0}\big) \big\} \\
&= C(K'') \frac{\Delta_{n0}^2}{h_m^3} \cdot \big\{ \big[G_m\big(r + h_m + \Delta_{n0}\big) - G_m\big(r - h_m - \Delta_{n0}\big) \big] \\
&\qquad - \big[G\big(r + h_m + \Delta_{n0}\big) - G\big(r - h_m - \Delta_{n0}\big) \big] \big\} \\
&\quad + C(K'') \frac{\Delta_{n0}^2}{h_m^3} \cdot \big[G\big(r + h_m + \Delta_{n0}\big) - G\big(r - h_m - \Delta_{n0}\big) \big] \\
&= I_{n,m}^1 + I_{n,m}^2
\end{aligned}$$

where $G(s) = \frac{1}{m} \sum_{j=1}^{m} \mathcal{I}\{F_0(Y_j) \leq s\}$. As G is a Lipschitz function, we have

$$|G\big(r + h_m + \Delta_{n0}\big) - G\big(r - h_m - \Delta_{n0}\big)| \leq 2L_G(h_m + \Delta_{n0})$$

and

$$\sqrt{mh_m}I_{n,m}^2 \le 2L_G C(K'')\frac{\Delta_{n0}^2}{h_m^3}\sqrt{mh_m}\big(h_m+\Delta_{n0}\big) = o_p(1)$$

as $\Delta_{n0} = O_p(n^{-\frac{1}{2}})$ and $mh_m^3 \to \infty$. We now consider $I_{n,m}^1$:

$$|I_{n,m}^1| \le$$
$$2C(K'')\frac{\Delta_{n0}^2}{h_m^3}\cdot \sup_{|t|\le h_m+\Delta_{n0}} \big|\big[G_m\big(r+t\big)-G_m\big(r\big)\big]-\big[G\big(r+t\big)-G\big(r\big)\big]\big|.$$

The Dvoretzky–Kiefer–Wolfowitz (1956) bound for the tails of Δ_{n0} yields that for any given $\epsilon > 0$ there exists some finite C such that $\Delta_{n0} \le Cn^{-\frac{1}{2}}$ up to an event with probability less than or equal to ϵ. The inequality $|t| \le h_m + \Delta_{n0}$ on this set means that $|t| \le C_1 h_m$ for some constant C_1. Using (2.13) in Stute (1982) we see that

$$\sup_{|t|\le C_1 h_m} \big|\big[G_m\big(r+t\big)-G_m\big(r\big)\big]-\big[G\big(r+t\big)-G\big(r\big)\big]\big| = O_p\Big(\sqrt{\frac{-h_m\log h_m}{m}}\Big).$$

Therefore, as $mh_m^3 \to \infty$,

$$\sqrt{mh_m}I_{n,m}^1 = o_p(1).$$

Lemma D.2 follows as $\epsilon > 0$ is arbitrary. ▪

The second term in (A.3) can be expressed as a two-sample U-statistic:

$$T_{n,m}(r) = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} k_{h_m}(Y_{0i}, Y_j; r)$$

with two-sample kernel

$$k_{h_m}(x,y;r) = -\big(\mathcal{I}\{x\le y\}-F_0(y)\big)\frac{1}{h_m^2}K'\Big(\frac{r-F_0(y)}{h_m}\Big),$$

which is dependent on m via h_m. Note that $\mathrm{E}[k_{h_m}(Y_0, Y; r)] = 0$ and the projections are $\mathrm{E}[k_{h_m}(Y_0, y; r)] = 0$ and

$$\begin{aligned} g_{1h_m}(x;r) &= \mathrm{E}[k_{h_m}(x,Y;r)] \\ &= -\int_0^1 \big(\mathcal{I}\{F_0(x)\le s\}-s\big)\frac{1}{h_m^2}K'\Big(\frac{r-s}{h_m}\Big)g(s)\,ds. \end{aligned}$$

Jammalamadaka and Janson (1986) consider the asymptotic behavior of one-sample U-statistics with kernel depending on m. Based on an extension of their ideas to two-sample U-statistics with kernel depending on m, we can obtain:

$$T_{n,m}(r) = \frac{1}{n}\sum_{i=1}^{n} g_{1h_m}(Y_{0i};r) + o_p\big((nh_m)^{-\frac{1}{2}}\big) \tag{A.4}$$

as $m \to \infty, m/n \to \kappa^2 < \infty$. Now

$$\begin{aligned}
&\sqrt{mh_m}\frac{1}{n}\sum_{i=1}^{n} g_{1h_m}(Y_{0i};r) \\
&= -\sqrt{\frac{m}{h_m^3}}\int_0^1 \{\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}\{F_0(Y_{0i}) \le t\} - t\big\}K'\big(\frac{r-t}{h_m}\big)g(t)\ dt \\
&\stackrel{\mathcal{D}}{=} -\sqrt{\frac{m}{n}}\frac{1}{h_m^{3/2}}\int_0^1 U_n(t)K'\big(\frac{r-t}{h_m}\big)g(t)\ dt,
\end{aligned}$$

where $U_n(t)$ is the uniform empirical process (Shorack and Wellner, 1986). We then have

$$\begin{aligned}
&\sqrt{\frac{m}{n}}\frac{1}{h_m^{3/2}}\int_0^1 U_n(t)K'\big(\frac{r-t}{h_m}\big)g(t)\ t \\
&= \sqrt{\frac{m}{n}}\frac{1}{h_m^{1/2}}\int_0^1 U_n(t)g(t)\ dK\big(\frac{r-t}{h_m}\big) \\
&= \sqrt{\frac{m}{n}}\big[\frac{1}{h_m^{1/2}}K\big(\frac{r-t}{h_m}\big)U_n(t)g(t)\big]_0^1 \\
&- \sqrt{\frac{m}{nh_m}}\int_0^1 K\big(\frac{r-t}{h_m}\big)[g(t)U_n(dt) + U_n(t)g'(t)\ t] \\
&= -\sqrt{\frac{m}{n}}\frac{1}{h_m^{1/2}}\int_0^1 K\big(\frac{r-t}{h_m}\big)g(t)U_n(dt) \\
&- \sqrt{\frac{m}{n}}\frac{1}{h_m^{1/2}}\int_0^1 K\big(\frac{r-t}{h_m}\big)U_n(t)g'(t)\ dt \\
&= I^3_{n,h_m} + I^4_{n,h_m}.
\end{aligned}$$

The second term I^4_{n,h_m} converges in probability to zero:

$$|I^4_{n,h_m}| \le \sqrt{\frac{m}{n}}\sup_{0\le t\le 1}|U_n(t)|\sqrt{h_m}\int_0^1 \frac{1}{h_m}K\big(\frac{r-t}{h_m}\big)|g'(t)|\ dt = o_p(1)$$

as $m \to \infty, m/n \to \kappa^2 < \infty$, because $\sup_{0\le t\le 1}|U_n(t)| = O_p(1)$ and the integral converges to $|g'(r)|$ by Bochner's Theorem. Thus I^4_{n,h_m} is asymptotically negligible. Note, however, that if $g(r)$ is not smooth, this term can contribute variation in moderate sample sizes. In fact, we can write

$$\begin{aligned}
\mathrm{Var}[I^4_{n,h_m}] =& \frac{m}{n}\frac{1}{h_m}\int_0^1\int_0^1 K\big(\frac{r-t}{h_m}\big)K\big(\frac{r-u}{h_m}\big)\mathrm{Cov}(U_n(t),U_n(s))g'(s)g'(t)dsdt \\
&\to h_m\kappa^2 r(1-r)[g'(r)]^2R^2(K)
\end{aligned}$$

as $m \to \infty, m/n \to \kappa^2 < \infty$. While this variance is small it is highly correlated with I^3_{n,h_m} and $T_{n,m}(r)$ so that, in small sample sizes, its contribution matters. This expression will be used in Section 9.2.3.2 to obtain an expression for the variance that is more accurate in small samples than the asymptotic approximation.

Pulling together these results, (A.4), Lemma D.2 and (A.3) we obtain that

$$\begin{aligned}\sqrt{mh_m}\big[g_{n,m}(r) - \tilde{g}_{n,m}(r)\big] = & \sqrt{\frac{m}{n}}\frac{1}{h_m^{1/2}}\int_0^1 K\big(\frac{r-t}{h_m}\big)g(t)U_n(dt) \\ & + \frac{1}{h_m^{1/2}}\int_0^1 K\big(\frac{r-t}{h_m}\big)U_{1m}(dt) \\ & + o_p(1)\end{aligned}$$

where

$$\tilde{g}_{n,m}(r) = \frac{1}{h_m}\int_0^1 K\big(\frac{r-t}{h_m}\big)g(t)dt$$

and U_{1m} is the empirical process for G. To complete the proof of the theorem we need to show that $\tilde{g}_{n,m}(r)$ approaches $g(r)$ at a fast enough rate that $\sqrt{mh_m}\,\big[\,\tilde{g}_{n,m}(r) - g(r)\,\big] \to 0$. Note that $\tilde{g}_{n,m}(r)$ has the same distribution as $g_m(r)$ in (9.9) so we can use standard kernel density results to see that $mh_m^5 \to 0$ is a sufficient condition.

To calculate the variance terms explicitly, we can reexpress the above as

$$\begin{aligned}\sqrt{mh_m}\big[g_{n,m}(r) - \tilde{g}_{n,m}(r)\big] = & \\ \sqrt{\frac{m}{n}}\frac{1}{n^{1/2}}\sum_{i=1}^n \Big\{ & \frac{1}{h_m^{1/2}}K\big(\frac{r-U_i}{h_m}\big)g(U_i) - \frac{1}{h_m^{1/2}}\mathrm{E}\big[K\big(\frac{r-U_i}{h_m}\big)g(U_i)\big]\Big\} \\ + \frac{1}{m^{1/2}}\sum_{i=1}^m \Big\{ & \frac{1}{h_m^{1/2}}K\big(\frac{r-R_i}{h_m}\big) - \frac{1}{h_m^{1/2}}\mathrm{E}\big[K\big(\frac{r-R_i}{h_m}\big)\big]\Big\} \\ + o_p(1), & \end{aligned}$$

where $U_1, \ldots, U_n$ are i.i.d. uniform $[0,1]$ independent of the R_i's. The variance of the first term is then

$$\begin{aligned} & \frac{m}{n}\int_0^1 \frac{1}{h_m}K^2\big(\frac{r-t}{h_m}\big)g^2(t)\;dt - \frac{m}{n}h_m\Big[\int_0^1 \frac{1}{h_m}K\big(\frac{r-t}{h_m}\big)g(t)\;dt\Big]^2 \\ \to \quad & \kappa^2 g^2(r)R(K)\end{aligned}$$

as $m \to \infty, m/n \to \kappa^2 < \infty$. The expression for the second term follows similarly.

E. Proof of results in Chapter 10

Multivariate U-statistics with estimated parameters

In this appendix we use a U-statistics approach to decompose the MRP estimator into independent components. The decomposition forms the basis of the theorems in Sections 10.3.1 and 10.3.3.

We first give a result about two-sample U-statistics that is used in the proof, and has independent interest. Theorem 2.8 of Randles (1982) can be extended to our two-sample situation. Following Randles, let $h(Y_1, Y_2, \ldots, Y_{r_1}; Y_{01}, Y_{02}, \ldots, Y_{0r_2}; \gamma)$ be a bivariate symmetric kernel of degree (r_1, r_2) based on the independent samples $Y_1, Y_2, \ldots, Y_m$ and Y_{01}, $Y_{02}, \ldots, Y_{0n}$. Denote the expected value of $h(\cdot; \cdot, \gamma)$ by

$$\theta(\gamma) = \mathrm{E}_\lambda[h(Y_1, Y_2, \ldots, Y_{r_1}; Y_{01}, Y_{02}, \ldots, Y_{0r_2}; \gamma)],$$

where λ denotes a parameter value. Let $U_{m,n}(\gamma)$ be a U-statistic corresponding to $h(\cdot; \gamma)$. Lemma E.1 is an extension to the two-sample situation of Theorem 2.8 of Randles. The following conditions are useful (see Remark 2.14 and Conditions 2.2 and 2.3 in Randles 1982):
Condition E.1 Suppose there is a neighborhood of λ, call it $K(\lambda)$, and a constant $K_1 > 0$, such that if $\gamma \in K(\lambda)$ and $D(\gamma, d)$ is a sphere centered at γ with radius d satisfying $D(\gamma, d) \subset K(\lambda)$. If

$$S = \sup_{\gamma' \in D(\gamma,d)} |h(Y_1, \ldots, Y_{r_1}; Y_{01}, \ldots, Y_{0r_2}; \gamma') - h(Y_1, \ldots, Y_{r_1}; Y_{01}, \ldots, Y_{0r_2}; \gamma)|$$

then $\mathrm{E}[S] \leq K_1 d$ and $\lim_{d \to 0} \mathrm{E}[S^2] = 0$.
Condition E.2 Suppose

$$n^{-\frac{1}{2}} \left(\hat{\lambda} - \lambda \right) = O_p(1).$$

Lemma E.1. Suppose Conditions E.1 and E.2 are satisfied. Then

$$N^{\frac{1}{2}} [U_{n,m}(\hat{\lambda}) - \theta(\hat{\lambda}) - U_{n,m}(\lambda) + \theta(\lambda)] \xrightarrow{P} 0,$$

where $N = n + m$.

The proof of this result follows closely that of Theorem 2.8 in Randles.

Sketch of the proofs of results in Sections 10.3.1 and theorem 10.3.3

The MRP can be reexpressed as:

$$\mathrm{MRP}(F;\ F_0) \ = \ 1 + 8\theta_2(F,\ F_0; \xi) - 4\theta_1(F,\ F_0; \xi)$$

where

$$\theta_1(F,\ F_0;\lambda) = \int_{-\infty}^{\infty} F(y) f_0(y-\lambda_1+\lambda_2)\ dy = P\Big(Y_1 \le Y_{01}+\lambda_1-\lambda_2\Big)$$

$$\theta_2(F,\ F_0;\lambda) = \int_{-\infty}^{\lambda_1} F(y) f_0(y-\lambda_1+\lambda_2)\ dy,$$

$\lambda = (\lambda_1, \lambda_2)$ is the parameter, $\xi = (\xi_{\frac{1}{2}}, \xi^0_{\frac{1}{2}})$. Define the two-sample kernel $h_1(x, y; \lambda) = \mathcal{I}(y \le x + \lambda_1 - \lambda_2)$. The corresponding generalized U-statistic

$$U^{(1)}_{m,n}(\lambda) = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} h_1(Y_{01}, Y_j; \lambda)$$

(the Wilcoxon two–sample rank statistic) is an unbiased estimator of $\theta_1(F,\ F_0; \lambda)$. Similarly we can define the two-sample kernel $h_2(x, y; \lambda) = \mathcal{I}(y \le x + \lambda_1 - \lambda_2)\mathcal{I}(x \le \lambda_2)$. The corresponding two-sample U-statistic

$$U^{(2)}_{m,n}(\lambda) = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} h_2(Y_{0i}, Y_j; \lambda)$$

is an unbiased estimator of $\theta_2(F;\ F_0; \lambda)$. The estimator can then be expressed as:

$$\widehat{\mathrm{MRP}}(F;\ F_0) \ = \ 1 + 8U^{(2)}_{m,n}(\hat{\xi}) - 4U^{(1)}_{m,n}(\hat{\xi}),$$

where $\hat{\xi} = (\hat{\xi}_{\frac{1}{2}}, \hat{\xi}^0_{\frac{1}{2}})$, so that

$$\widehat{\mathrm{MRP}}(F;\ F_0) - \mathrm{MRP}(F;\ F_0) \ =$$
$$8[U^{(2)}_{m,n}(\hat{\xi}) - \theta_2(F;\ F_0; \xi)] - 4[U^{(1)}_{m,n}(\hat{\xi}) - \theta_1(F;\ F_0; \xi)]. \qquad (A.5)$$

To simplify the notation we will suppress reference to F and F_0 in $\theta_1(F,\ F_0; \lambda)$ and $\theta_2(F,\ F_0; \lambda)$. We would like to show the asymptotic Gaussianality of

$$U^{(2)}_{m,n}(\hat{\xi}) - \theta_2(\xi) = \Big(U^{(2)}_{m,n}(\xi) - \theta_2(\xi)\Big) \qquad (A.6)$$
$$+ \Big(\theta_2(\hat{\xi}) - \theta_2(\xi)\Big) \qquad (A.7)$$
$$+ \Big((U^{(2)}_{m,n}(\hat{\xi}) - U^{(2)}_{m,n}(\xi)) - (\theta_2(\hat{\xi}) - \theta_2(\xi))\Big).$$

Using Lemma D.1 it can be shown that the last term in the decomposition is of smaller order than the first two terms. For the term (A.7) we have

$$\theta_2(\hat{\xi}) - \theta_2(\xi) = \phi_2(\xi)\Big(\hat{\xi}_{\frac{1}{2}} - \xi_{\frac{1}{2}}\Big) + [F(\xi_{\frac{1}{2}}) f_0(\xi^0_{\frac{1}{2}}) - \phi_2(\xi)]\Big(\hat{\xi}^0_{\frac{1}{2}} - \xi^0_{\frac{1}{2}}\Big) + o_p(n^{-\frac{1}{2}}),$$

where

$$\phi_2(\lambda) = \int_{-\infty}^{\lambda_1} f(y) f_0(y - \lambda_1 + \lambda_2)\, dy.$$

We further have (see Serfling 1980, p.93):

$$n^{\frac{1}{2}}\left(\hat{\xi}^0_{\frac{1}{2}} - \xi^0_{\frac{1}{2}}\right) = n^{\frac{1}{2}} \frac{\left[F_0(\xi^0_{\frac{1}{2}}) - F_{n0}(\xi^0_{\frac{1}{2}})\right]}{f_0(\xi^0_{\frac{1}{2}})} + o_p(1).$$

Therefore

$$\begin{aligned}\theta_2(\hat{\xi}) - \theta_2(\xi) = &\frac{\phi_2(\xi)}{f(\xi_{\frac{1}{2}})}\left(F(\xi_{\frac{1}{2}}) - F_m(\xi_{\frac{1}{2}})\right) \\ &+ [\frac{1}{2} - \frac{\phi_2(\xi)}{f_0(\xi^0_{\frac{1}{2}})}]\left(F_0(\xi^0_{\frac{1}{2}}) - F_{n0}(\xi^0_{\frac{1}{2}})\right) + o_p(n^{-\frac{1}{2}}) \qquad (A.8)\end{aligned}$$

as $m/n \to \kappa^2 < \infty, m \to \infty$. Finally we consider the term (A.6). Based on the ideas of Randles and Wolfe (1979), we obtain the following:

$$\begin{aligned}U^{(2)}_{m,n}(\xi) - \theta_2(\xi) = &\frac{1}{m}\sum_{j=1}^{m}\left[g_2(Y_j;\xi) - \theta_2(\xi)\right] \\ &+\frac{1}{n}\sum_{i=1}^{n}\left[g_1(Y_{0i};\xi) - \theta_2(\xi)\right] \quad + o_p(n^{-\frac{1}{2}})\end{aligned}$$

as $m/n \to \kappa^2 < \infty, m \to \infty$. Here we have used the projections:

$$\begin{aligned}g_1(x;\xi) =& \mathrm{E}[h_2(x, Y_1;\xi)] = \mathcal{I}(x \le \xi^0_{\frac{1}{2}})F(x + \xi_{\frac{1}{2}} - \xi^0_{\frac{1}{2}}) \\ g_2(y;\xi) =& \mathrm{E}[h_2(Y_{01}, y;\xi)] = (\frac{1}{2} - F_0(y - \xi_{\frac{1}{2}} + \xi^0_{\frac{1}{2}}))\mathcal{I}(y \le \xi_{\frac{1}{2}}).\end{aligned}$$

A similar approach is valid for $U^{(1)}_{m,n}(\xi)$:

$$\begin{aligned}U^{(1)}_{m,n}(\xi) - \theta_1(\xi) = &\frac{1}{m}\sum_{j=1}^{m}\left[\tilde{g}_2(Y_j;\xi) - \theta_1(\xi)\right] \\ &+\frac{1}{n}\sum_{i=1}^{n}\left[\tilde{g}_1(Y_{0i};\xi) - \theta_1(\xi)\right] \quad + o_p(n^{-\frac{1}{2}})\end{aligned}$$

as $m/n \to \kappa^2 < \infty, m \to \infty$. Here the projections are:

$$\begin{aligned}\tilde{g}_1(x;\xi) =& \mathrm{E}[h_1(x, Y_1;\xi)] = F(x + \xi_{\frac{1}{2}} - \xi^0_{\frac{1}{2}}) \\ \tilde{g}_2(y;\xi) =& \mathrm{E}[h_1(Y_{01}, y;\xi)] = 1 - F_0(y - \xi_{\frac{1}{2}} + \xi^0_{\frac{1}{2}}).\end{aligned}$$

In addition

$$\theta_1(\hat{\xi}) - \theta_1(\xi) = \frac{\phi_1(\xi)}{f(\xi_{\frac{1}{2}})}\left(F(\xi_{\frac{1}{2}}) - F_m(\xi_{\frac{1}{2}})\right) - \frac{\phi_1(\xi)}{f_0(\xi^0_{\frac{1}{2}})}\left(F_0(\xi^0_{\frac{1}{2}}) - F_{n0}(\xi^0_{\frac{1}{2}})\right) + o_p(n^{-\frac{1}{2}}), \quad (A.9)$$

where

$$\phi_1(\lambda) = \int_{-\infty}^{\infty} f(y) f_0(y - \lambda_1 + \lambda_2)\, dy$$

as $m/n \to \kappa^2 < \infty, m \to \infty$.

Applying (A.6), (A.9) and some algebra, (A.5) can be reexpressed in terms of the sums of two independent random variables:

$$\widehat{\mathrm{MRP}}(F;\ F_0) - \mathrm{MRP}(F;\ F_0) \ = \ \frac{1}{n}\sum_{i=1}^{n} a(Y_{0i}) + \frac{1}{m}\sum_{j=1}^{m} b(Y_j) + o_p(n^{-\frac{1}{2}}) \quad (A.10)$$

as $m/n \to \kappa^2 < \infty, m \to \infty$, where

$$a(x) = 8\left[g_1(x;\xi) - \theta_2(\xi)\right] - 4\left[\tilde{g}_1(x;\xi) - \theta_1(\xi)\right] - 4\left[\mathcal{I}(x \le \xi^0_{\frac{1}{2}}) - \frac{1}{2}\right](1 + \delta_0(\xi)).$$

$$b(x) = 8\left[g_2(x;\xi) - \theta_2(\xi)\right] - 4\left[\tilde{g}_2(x;\xi) - \theta_1(\xi)\right] + 4\left[\mathcal{I}(x \le \xi_{\frac{1}{2}}) - \frac{1}{2}\right]\delta\ (\xi).$$

These expressions can be reduced with a little algebra to:

$$a(x) = -4|F(x + \xi_{\frac{1}{2}} - \xi^0_{\frac{1}{2}}) - \frac{1}{2}| + 2\mathrm{Sign}(x - \xi^0_{\frac{1}{2}})\delta_0(\xi) - \mathrm{MRP}(F;\ F_0) + 1$$

$$b(x) = \ 4|F_0(x - \xi_{\frac{1}{2}} + \xi^0_{\frac{1}{2}}) - \frac{1}{2}| - 2\mathrm{Sign}(x - \xi_{\frac{1}{2}})\delta\ (\xi) + \mathrm{MRP}(F;\ F_0) - 1.$$

Note that we have used Property (a) of Section 2.2 to introduce $\mathrm{MRP}(F;\ F_0)$ into the expression for $b(x)$. The theorems in Sections 10.3.1 and 10.3.3 then follow directly from the consideration of (A.10).

F. Properties of the quasirelative data under equality

In this appendix we give properties of the quasirelative data under the hypothesis of equality of the comparison and reference distributions. These properties form the basis of the results in Section 10.3 . First, let us consider the properties of the quasirelative data:

Lemma F.1. Under the hypothesis $H_0 : F = F_0$,

1) $\mathrm{E}(Q_j) = \frac{1}{2}$ and $\sigma^2 \ = \ \mathrm{Var}(Q_j) = \frac{n+2}{12n}$.
2) The correlation between Q_i and Q_j is $\frac{1}{n+2}$.

3)

$$\mathrm{E}(|Q_j - \frac{1}{2}|) = \begin{cases} \frac{n+2}{4(n+1)} & n \text{ even} \\ \frac{n+1}{4n} & n \text{ odd} \end{cases}$$

$$\sigma^2_{|Q_j - \frac{1}{2}|} = \mathrm{Var}(|Q_j - \frac{1}{2}|) = \begin{cases} \frac{(n+2)(n^2+2n+4)}{48n(n+1)^2} & n \text{ even} \\ \frac{(n^2+2n-3)}{48n^2} & n \text{ odd} \end{cases}$$

The covariance between $|Q_i - \frac{1}{2}|$ and $|Q_j - \frac{1}{2}|$ is $(1/(n+2))\sigma^2_{|Q_j - \frac{1}{2}|}, i \neq j$.

Proof of lemma F.1

By the symmetry in the ranks produced by the hypothesis, the marginal distribution of Q_j is uniform on $\{i/n : i = 0, \ldots n\}$. So part a) follows easily. Also

$$\mathrm{Var}(\sum_{j=1}^{m} Q_j) = \mathrm{Var}\left(\frac{1}{n}\sum_{j=1}^{m} T_j - \frac{m(m+1)}{2n}\right) = \frac{m(n+m+1)}{12n},$$

using standard results about ranks. By symmetry the pairwise correlation between Q_i and Q_j is the same for each $i \neq j$. If we denote the common value by ϕ then

$$\mathrm{Var}(\sum_{j=1}^{m} Q_j) = m\sigma^2 \left[(m-1)\phi + 1\right].$$

Comparing the last two expressions proves b). Part c) follows by applying a similar approach to the distribution of $|Q_i - \frac{1}{2}|$. ▪

References

Abraham, K. (1990) Restructuring the employment relationship: The growth of market-mediated work arrangements, in *New Developments in the Labor Market*. (ed.) The MIT Press, Cambridge, MA, pp. 85-118.

Abraham, K and McKersie, R. (1990) *New Developments in the Labor Market*. The MIT Press, Cambridge, MA.

Abramowitz, M and Stegun, I. (1965) *Handbook of Mathematical Functions*. Dover, New York, NY.

Absava, RM and Nadareishvili, MM. (1985) Nonparametric estimation of density distribution of two unknown dimensions in D-dimensional Euclidean space (Russian). *Sakharthvelos Ssr Mecnierebatha Akademis Moambe*, **117**, 257-260.

Adhikari, BP and Joshi, DD. (1956) Distance, discrimination et résumé exhaustif. *Publ Inst Statist Univ Paris*, **5**, 57-74.

Aerts, M, Augustyns, I and Janssen, P. (1994) Smoothing sparse multinomial data using local polynomial fitting. *Journal of Computational and Graphical Statistics*, **3**, 57-66.

Aigner, D, Amemiya, T and Poirer, D. (1976) On the estimation of production frontiers: Maximum likelihood estimation of the parameters of a discontinuous density function. *International Economic Review*, **17**, 372-396.

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle, in *2nd International Symposium on Information Theory*. BN Petrov and F Csaki (ed.) Akademia Kiado, Budapest, pp. 267-281.

Alexander, WP. (1989) *Boundary kernel estimation of the two sample comparison density function*. Unpublished Ph.D. thesis, Statistics, Texas A&M University.

Ali, SM and Silvey, SD. (1966) A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, **28**, 131-142.

Allison, P. (1984) *Event History Analysis: Regression for Longitudinal Event Data*. Sage Publications, Newbury Park, CA.

Aly, E-EAA, Csorgo, M and Horvath, L. (1987) P-P plots, rank processes and Chernoff-Savage theorems, in *New Perspectives in Theoretical and Applied Statistics*. ML Puri, JP Vilaplana and W Wertz (ed.) John Wiley & Sons, New York, NY.

Anscombe, FJ. (1948) The transformation of Poisson, binomial, and negative-binomial data. *Biometrika*, **35**, 246-254.

Appelbaum, E. (1987) Restructuring work: Temporary, part-time, and at-home employment, in *Computer Chips and Paper Clips: Technology and Women's Employment*. H Hartmann (ed.) National Academy Press., Washington, DC, pp. 268-310.

Applebaum, E and Berg, P. (1996) Financial market constraints and business strategy in the USA, in *Creating Industrial Capacity: Towards Full Employment*. J Michie and JG Smith (ed.) Oxford University Press, Oxford, pp. 192-221.

Atkinson, AB, Bourguignon, F and Morrison, C. (1992) *Empirical studies of earnings mobility*. Harwood Academic Publishers, Philadelphia.

Auletta, K. (1982) *The Underclass*. Random House, New York.

Baker, M. (1997) Growth rate heterogeneity and the covariance structure of life–cycle earnings. *Journal of Labor Economics*, **15**, 338-375.

Bamber. (1975) The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, **12**, 387-415.

Barron, AR, Györfi, L and van der Meulen, EC. (1992) Distribution estimation consistent in total variation and in two types of information divergence. *IEEE Transactions in Information Theory*, **38**, 1437-1454.

Barron, AR and Sheu, C-H. (1991) Approximation of density functions by sequences of exponential families. *Annals of Statistics*, **19**, 1347-1369.

Begg, CB. (1991) Advances in statistical methodology for diagnostic medicine in the 1980's. *Statistics in Medicine*, **10**, 1887-1895.

Bell, L and Freeman, R. (1986) *The facts about rising industrial wage dispersion in the United States*. Industrial Relations Research Association, Annual Meetings, pp. Proceedings, 331-337.

Belous, RS. (1989) *The contingent economy: The growth of the temporary, part-time, and sub-contracted workforce*. NPA Report 239, National Planning Association, Washington, D.C.

Berger, MC and Hirsch, BT. (1983) The civilian earnings experience of Vietnam-era veterans. *Journal of Human Resources*, **18**, 455-479.

Berlin, G and Sum, A. (1988) *Toward A More Perfect Union: Basic Skills, Poor Families, and Our Economic Future*. Project on Social Welfare and the American Future, Ford Foundation, New York, NY.

Bernhardt, A, Morris, M, Handcock, M and Scott, M. (1999) *Inequality and mobility: Trends in wage growth for young adults*. Working paper 99-03, Pennsylvania State University, Population Research Institute, University Park, PA.

Bernhardt, AD, Morris, M and Handcock, MS. (1995) Women's gains or men's losses? A closer look at the shrinking gender gap in earnings. *American Journal of Sociology*, **101**, 302-328.

Bernhardt, AD, Morris, M and Handcock, MS. (1999) Trends in job instability and wages for young adult men. *Journal of Labor Economics*, (forthcoming).

Berry, S, Gottschalk, P and Wissoker, D. (1995) An error components model of the impact of plant closings on earnings. *Review of Economics and Statistics*, **70**, 701-707.

Bickel, PJ, Klaassen, CAJ, Ritov, Y and Wellner, JA. (1993) *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore, MD.

Blau, FD. (1998) Trends in the well-being of American women, 1970-1995. *Journal of Economic Literature*, **36**, 112-165.

Blau, FD and Kahn, LM. (1994) Rising wage inequality and the U.S. gender gap. *American Economic Review*, **84**, 23-28.

Blau, PM. (1977) *Inequality and Heterogeneity*. Free Press, New York, NY.

Bogdan, M and Ledwina, T. (1996) Testing uniformity via log-spline modeling. *Statistics*, **28**, 131-157.

Brown, LD. (1986) *Fundamentals of Statistical Exponential Families With Applications in Statistical Decision Theory*. Institute of Mathematical Statistics, Hayward, CA.

Bryk, AS and Raudenbush, SW. (1992) *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications, Newbury Park, CA.

Buchinsky, M. (1994) Changes in the U.S. Wage Structure 1963-1987: An Application of Quantile Regression. *Econometrica*, **62**, 405-458.

Buchinsky, M. (1995) Estimating the asymptotic covariance matrix for quantile regression models: A Monte Carlo study. *Journal of Econometrics*, **68**, 303-338.

Buchinsky, M. (1995) Quantile regression, Box-Cox transformation model, and the U.S. wage structure, 1963-1987. *Journal of Econometrics*, **65**, 109-154.

Buchinsky, M. (1998) Recent advances in quantile regression models: A practical guideline for empirical research. *Journal of Human Resources*, **33**, 88-126.

Burr, D and Doss, H. (1993) Confidence bands for the median survival time as a function of covariates in the Cox model. *Journal of the American Statistical Association*, **88**, 1330-1340.

Butler, RJ and McDonald, JB. (1987) Interdistributional income inequality. *Journal of Business and Econometric Statistics*, **5**, 13-18.

Campbell, G. (1994) Advances in statistical methodology for evaluation of diagnostic and laboratory tests. *Statistics In Medicine*, **13**, 499-508.

Cancian, M. (1998) Assessing the effects of wives' earnings on family income inequality. *Review of Economics and Statistics*, **80**, 73.

Cao, R, Janssen, PJ and Veraverbeke, N. (1999) *Relative density estimation with censored data*. Technical Report, Limburgs Universitair Centrum, Hasselt.

Cappelli, P. (1992) Examining managerial displacement. *Academy of Management Journal*, **35**, 203-217.

Cappelli, P. (1994) The effect of restructuring on employees, in *Looking Ahead: The Restructuring of Employment*. (ed.) National Planning Association, Washington, DC.

Cappelli, P. (1995) Rethinking employment. *British Journal of Industrial Relations*, **33**, 563-602.

Card, D and Krueger, AB. (1995) *Myth and Measurement: The New Economics of the Minimum Wage*. Princeton University Press, Princeton.

Čencov, NN. (1962) Evaluation of an unknown distribution density from observations. *Soviet Mathematics*, **3**, 1559-1562.

Chamberlain, G. (1984) Panel data, in *Handbook of Econometrics, Vol II*. Z Griliches and MD Intriligator. (ed.) Elsevier Science, Amsterdam, pp. 1247-1318.

Chambers, JM, Cleveland, WS, Kleiner, B and Tukey, PA. (1983) *Graphical Methods For Data Analysis*. Wadsworth, Pacific Grove, CA.

Chatterjee, S, Handcock, MS and Simonoff, JS. (1995) *A Casebook for a First Course in Statistics and Data Analysis*. John Wiley & Sons, New York, NY.

Chaudhuri, P. (1991) Nonparametric estimates of regression quantiles and their local Bahadur representation. *Annals of Statistics*, **19**, 760-777.

Cheng, C and Parzen, E. (1997) Unified estimators of smooth quantile and quantile density functions. *Journal of Statistical Planning and Inference*, **59**, 291-307.

Cheng, KF. (1983) Nonparametric estimators for percentile regression functions. *Communications in Statistics, Series A: Theory and Methods*, **12**, 681-692.

Cheng, KF. (1984) Nonparametric estimation of regression function using linear combinations of sample quantile regression functions. *Sankhyā Series A*, **46**, 287-302.

Cheng, KF and Wu, JW. (1998) An optimal test for the mean function hypothesis. *Statistica Sinica*, **8**, 477-487.

Cheng, MY. (1994) A bandwidth selector for local linear density estimators. *Journal of Computational and Graphical Statistics*, **3**, 57-66.

Chernoff, H and Savage, IR. (1958) Asymptotic normality and efficiency of certain non-parametric test statistics. *Annals of Mathematical Statistics*, **29**, 972-994.

Cleveland, WS and McGill, R. (1984) Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, **79**, 531-554.

Colclough, G and Tolbert, CM. (1992) *Work in the Fast Lane: Flexibility, Divisions of Labor, and Inequality in High-Tech Industries*. SUNY Press, New York, NY.

Cole, TJ. (1988) Fitting smoothed centile curves to reference data. *Journal of the Royal Statistical Society, Series A*, **151**, 385-418.

Cole, TJ and Green, PJ. (1992) Smoothing reference centile curves: The LMS method and penalized likelihood. *Statistics in Medicine*, **11**, 1305-1319.

Commission on the Skills of the American Workforce. (1990) *America's Choice: High Skills or Low Wages!*. The National Center on Education and the Economy, Washington, D.C.

Costrell, RM. (1988) *The Effects of Industry Employment Shifts on Wage Growth, 1947-1987*. Report of testimony, U.S. Congress, Joint Economic Committee, Washington, DC.

Csörgő, M. (1983) *Quantile processes with statistical applications*. CBMS-NSF Regional Conference Series in Applied Mathematics, Philadelphia, SIAM.

Csiszár, I. (1978) *Information measures: A critical survey*. Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions and the Eighth European Meeting of Statisticians, Technical University of Prague, Prague (1974), Vol. B, pp. 73-86.

Cwik, J and Mielniczuk, J. (1989) Estimating density ratios with application to discriminant analysis. *Communications in Statistics*, **18**, 3057-3069.

Cwik, J and Mielniczuk, J. (1990) Some topics in estimation of Neyman–Pearson and performance curves, in *Cosmex*. W Kasprzak and A Weron (ed.) World Scientific, Singapore, pp. 114-129.

Cwik, J and Mielniczuk, J. (1993) Data-dependent bandwidth choice for a grade density kernel estimate. *Statistics and Probability Letters*, **16**, 397-405.

Dagum, C. (1977) A new model for personal income distribution: Specification and estimation. *Economie Appliquee*, **30**, 413-436.

Dagum, C. (1980) Inequality measures between income distributions with applications. *Econometrics*, **48**, 1791-1803.

Dalton, H. (1920) The measurement of the inequality of incomes. *Economic Journal*, **30**, 348-361.

Dalton, H. (1929) *Some aspects of the inequality of incomes in modern communities*. Routledge and Paul, London.

D'Amico, R. (1984) Does employment during high school impair academic progress? *Sociology of Education*, **57**, 152-164.

Danziger, S and Gottschalk, P. (1993) *Uneven Tides: Rising Inequality in America*. Russell Sage Foundation, New York, NY.

Danziger, S and Gottschalk, P. (1996) *America Unequal*. Russell Sage Foundation, New York, NY.

Denison, DGT, Mallick, BK and Smith, AFM. (1998) Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society Series B*, **60**, 333.

Diggle, P, Liang, KL and Zeger, SL. (1994) *Analysis of Longitudinal Data*. Oxford University Press, Oxford.

DiNardo, J, Fortin, N and Lemieux, T. (1996) Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrica*, **64**, 1001-1044.

DiNardo, J and Lemieux, T. (1996) Diverging male wage inequality in the United States and Canada, 1981-1988: Do Institutions explain the difference. *Industrial and Labor Relations Review*, **50**, 629-651.

DiNardo, J and Pischke, J. (1996) The returns to computer use revisited: Have pencils changed the wage structure too? *Quarterly Journal of Economics*, **112**, 291-303.

DiPrete, TA. (1993) Industrial restructuring and the mobility response of American workers in the 1980s. *American Sociological Review*, **58**, 74-96.

Doeringer, PB and Piore, MJ. (1971) The theory of internal labor markets, in *Internal Labor Markets and Manpower Analysis*. (ed.) Heath, Lexington, Mass, pp. 13-92.

Doksum, K. (1974) Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *Annals of Statistics*, **2**, 267-277.

Doksum, KA and Sievers, GL. (1976) Empirical probability plots and statistical inference for non-linear models in the the two sample case. *Biometrika*, **63**, 421-434.

Dooley, M and Gottschalk, P. (1982) Does a younger male labor force mean growing earnings inequality? *Monthly Labor Review*, **105**, 42-45.

Du Toit, SHC, Steyn, AGW and Stumpf, RH. (1986) *Graphical Exploratory Data Analysis*. Springer-Verlag, Berlin.

Duncan, G, Boisjoly, J and Smeeding, T. (1996) Economic mobility of young workers in the 1970s and 1980s. *Demography*, **33**, 497-509.

Dunnett, CW. (1989) Algorithm AS 251: Multivariate normal probability integrals with product correlation structure. *Applied Statistics*, **38**, 564-573.

Durbin, J. (1973) Weak convergence of the sample distribution function when parameters are estimated. *Annals of Statistics*, **1**, 279-290.

Dvoretzky, A, Kiefer, J and Wolfowitz, J. (1956) Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Annals of Mathematical Statistics*, **27**, 642-669.

Ebrahimi, N, Pflughoeft, K and Soofi, ES. (1994) Two measures of sample entropy. *Statistics and Probability Letters*, **20**, 225-234.

Efron, B. (1991) Regression percentiles. *Statistica Sinica*, **1**, 93-125.

Efron, B and Tibshirani, R. (1993) *An Introduction to the Bootstrap*. Chapman and Hall, New York, NY.

Esteban, J. (1986) Income share elasticity and the size distribution of income. *International Economic Review*, **27**, 439-444.

Eubank, RL and LaRiccia, VN. (1992) Asymptotic comparison of Cramer–von Mises and nonparametric function estimation techniques for testing goodness-of-fit. *Annals of Statistics*, **20**, 2071-2086.

Eubank, RL, LaRiccia, VN and Rosenstein, RB. (1987) Test statistics derived as components of Pearson's phi-squared distance measure. *Journal of the American Statistical Association*, **82**, 816-825.

Eubank, RL, LaRiccia, VN and Schuenemeyer, JL. (1995) Component type tests with estimated parameters. *Probability and Mathematical Statistics*, **15**, 275-289.

Fan, J and Gijbels, I. (1996) *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London.

Fan, J, Hu, TC and Truong, YK. (1994) Robust non-parametric function estimation. *Scandinavian Journal of Statistics*, **21**, 433-446.

Fan, J, Yao, Q and Tong, H. (1996) Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, **83**, 189-206.

Farber, HS. (1996) *Job creation in the United States: Good or bad?* Working paper, Industrial Relations, Princeton University, Department of Economics.

Farber, HS. (1997) The changing face of job loss in the United States: 1981-93. *Brookings Papers on Economic Activity, Microeconomics Supplement*, 55-142.

Feldstein, M. (1980) *The American Economy in Transition*. The University of Chicago Press, Chicago, IL.

Firebaugh, G. (1999) Empirics of world income inequality. *American Journal of Sociology*, (forthcoming).

Fraser, DAS. (1957) *Nonparametric Methods in Statistics*. John Wiley & Sons, New York, NY.

Freedman, D, Pisani, R, Purves, R and Adhikari, A. (1991) *Statistics*. Norton, New York, NY.

Freeman, R and Katz, L. (1994) Rising wage inequality: The United States vs. other advanced countries, in *Working Under Different Rules*. R Freeman (ed.) Russell Sage Foundation, New York, NY, pp. 29-62.

Fuchs, V. (1968) *The Service Economy*. Columbia University Press, New York, NY.

Gasser, T and Müller, H-G. (1979) Kernel estimation of regression functions, in *Smoothing Techniques for Curve Estimation*. T Gasser and M Rosenblatt (ed.) Springer-Verlag, Berlin, pp. 23-68.

Gastwirth, JL. (1968) The first-median test: A two-sided version of the control median test. *Journal of the American Statistical Association*, **63**, 692-706.

Gastwirth, JL and Wang, JL. (1988) Control percentile test procedures for censored data. *Journal of Statistical Planning and Inference*, **18**, 267–276.

George, EI and Foster, DP. (1997) *Calibration and empirical Bayes variable selection*. Technical report, University of Texas, Austin.

Gijbels, I and Mielniczuk, J. (1995) Asymptotic properties of kernel estimators of the Radon-Nikodym derivative with applications to discriminant analysis. *Statistica Sinica*, **5**, 261-278.

Ginther, DK. (1995) *A nonparametric analysis of the U.S. earnings distribution*. Discussion Paper 1067-95, Institute For Research On Poverty, University Of Wisconsin-Madison.

Gittleman, M, Horrigan, M and Joyce, M. (1996) *Have Family Income Mobility Patterns Changed?*. Manuscript, U.S. Department of Labor, Bureau of Labor Statistics, Washington, DC.

Gittleman, M and Joyce, M. (1996) Earnings mobility and long-run inequality: An analysis using matched CPS data. *Industrial Relations*, **35**, 180-196.

Glasbey, CA. (1989) Discussion of "Space-time modelling with long-memory dependence: Assessing Ireland's wind power resource." by Haslett and Raftery. *Journal of the Royal Statistical Society, Series C*, **38**, 1-50.

Good, IJ. (1950) *Probability and the Weighing of Evidence*. Griffin, London.

Gordon, DM. (1996) *Fat and mean: The Corporate Squeeze of Working Americans and the Myth of Managerial "Downsizing"*. Martin Kessler Books, New York, NY.

Gottschalk, P. (1997) Inequality, income growth, and mobility: The basic facts. *Journal of Economic Perspectives*, **11**, 21-40.

Gottschalk, P and Moffit, R. (1994) The growth of earnings instability in the U.S. labor market. *Brookings Papers on Economic Activity*, **2**, 217-272.

Grove, DJ and Hannum, R. (1986) On measuring intergroup inequality. *Sociological Methods and Research*, **15**, 142-159.

Gutenbrunner, C and Jureckova, J. (1992) Regression rank scores and regression quantiles. *Annals of Statistics*, **20**, 305-330.

Haider, S. (1997) *Earnings instability and earnings inequality of males in the United States: 1967–1991*. Unpublished manuscript, University of Michigan, Ann Arbor, MI.

Hall, P. (1986) On the rate of convergence of orthogonal series density estimators. *Journal of the Royal Statistical Society, Series B*, **48**, 115-122.

Hand, DJ. (1982) *Kernel Discriminant Analysis*. John Wiley & Sons, New York, NY.

Handcock, MS. (1996) *Statistical properties of the relative distribution and relative polarization indices for grouped or discrete data*. Technical report, Department of Statistics and Operations Research, New York University, New York, NY.

Handcock, MS and Janssen, P. (1998a) *Statistical properties of the relative polarization indices*. Technical Report, Department of Statistics, The Pennsylvania State University, University Park, PA.

Handcock, MS and Janssen, P. (1998b) *Statistical properties of the relative distribution and relative density*. Technical Report, Pennsylvania State University, Department of Statistics, University Park, PA.

Handcock, MS and Morris, M. (1998) Relative distribution methods. *Sociological Methodology*, **28**, 53-97.

Handcock, MS, Morris, M and Bernhardt, AD. (1994) *Imputation methods for coarse income data*. Technical report, Department of Statistics and Operations Research, New York University, New York, NY.

Handcock, MS, Morris, M and Bernhardt, AD. (1997) *A comparison of the dispersion in earnings in the CPS and NLSY 1979–1994*. Working paper 98-14, The Pennsylvania State University, Population Research Institute, University Park, PA.

Hannum, R and Longbotham, R. (1986) On measuring intergroup inequality. *Sociological Methods And Research*, **15**, 142-159.

Harrison, B. (1994) *Lean and Mean: The Changing Landscape of Corporate Power in the Age of Flexibility*. Basic Books, New York, NY.

Harrison, B and Bluestone, B. (1988) *The Great U-Turn: Corporate Restructuring and the Polarizing of America*. Basic Books, New York, NY.

Hart, J. (1985) On the choice of a truncation point in Fourier series density estimation. *Journal of Statistical Computation and Simulation*, **21**, 95-116.

Hastie, T and Loader, C. (1993) Local regression: Automatic kernel carpentry. *Statistical Science*, **8**, 120-129.

Haughton, DMA. (1988) On the choice of a model of fit data from an exponential family. *Annals of Statistics*, **16**, 342-355.

He, X. (1997) Quantile curves without crossing. *The American Statistician*, **51**, 186-192.

He, X and Shi, PD. (1994) Convergence rate of B-spline estimators of nonparametric conditional quantile functions. *Journal of Nonparametric Statistics*, **3**, 299-308.

Heckman, J and Singer, B. (1984) A model for minimizing the impact of distributional assumptions in econometric models for the analysis of duration data. *Econometrica*, **52**, 271-320.

Heitjan, DF and Rubin, DB. (1990) Inference from coarse data via multiple imputation with application to age heaping. *Journal of the American Statistical Association*, **85**, 304-314.

Herrmann, E, Gasser and Kneip, A. (1992) Choice of bandwidth for kernel regression when residuals are correlated. *Biometrika*, **79**, 783-795.

Hinkley, DV. (1969) On the ratio of two correlated normal random variables. *Biometrika*, **56**, 635.

Hochberg, Y and Tamhane, AC. (1987) *Multiple Comparisons Procedures*. John Wiley & Sons, New York, NY.

Hogg, RV. (1975) Estimates of percentile regression lines using salary data. *Journal of the American Statistical Association*, **70**, 56-59.

Holmgren, EB. (1995) The P-P plot as a method for comparing treatment effects. *Journal of the American Statistical Association*, **90**, 360-365.

Houseman, SN. (1997) Flexible staffing arrangements in the U.S. *Worklife Report*, **10**, 6.

Howell, D and Wolff, E. (1991) Trends in the growth and distribution of skills in the U.S. workplace, 1960-1985. *Industrial and Labor Relations Review*, **44**, 486-502.

Howell, DR, Duncan, M and Harrison, B. (1998) *Low wages in the U.S. and high unemployment in Europe: A critical assessment of the conventional wisdom*. Working Paper Series I 5, New School for Social Research, Center for Economic Policy Analysis, New York, NY.

Hsieh, F. (1995) The empirical process approach for semiparametric two-sample models with heterogeneous treatment effect. *Journal of the Royal Statistical Society, Series B*, **57**, 735-748.

Hsieh, F and Turnbull, BW. (1996) Nonparametric methods for evaluating diagnostic tests. *Statistica Sinica*, **6**, 47-62.

Hurvich, CM, Simonoff, JS and Tsai, C-L. (1998) Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Series B*, **60**, 271-293.

Ingot, T, Kallenberg, WCM and Ledwina, T. (1998) Vanishing shortcomings of data-driven Neyman's tests. *Annals of Statistics*, **9**, 1982-2019.

Ingot, T and Ledwina, T. (1996) Asymptotic optimality of data-driven Neyman's tests for uniformity. *Annals of Statistics*, **9**, 1982-2019.

Jammalamadaka, SR and Janson, S. (1986) Limit theorems for a triangular scheme of U-statistics with applications to inter-point distances. *Annals of Probability*, **14**, 1347-1358.

Janssen, P and Veraverbeke, N. (1987) On nonparametric regression estimators based on regression quantiles. *Communications in Statistics: Theory and Methods*, **16**, 383-396.

Joe, H. (1989) Estimation of entropy and other functionals of a multivariate density. *Annals of the Institute of Mathematical Statistics*, **41**, 683-697.

Joe, H. (1997) *Multivariate Models and Dependence Concepts*. Chapman and Hall, New York, NY.

Johnston, WB and Packer, AE. (1987) *Workforce 2000: Work and Workers for the Twenty-first Century*. Hudson Institute, Indianapolis, IN.

Juhn, C and Murphy, KM. (1993) Wage inequality and the rise in returns to skill. *The Journal of Political Economy*, **101**, 410-422.

Juhn, C, Murphy, KM and Pierce, B. (1991) Accounting for the slowdown in black–white wage convergence, in *Workers and Their Wages*. M Kosters (ed.) American Enterprise Institute Press, Washington, DC, pp. 107-143.

Kakwani, N. (1980) *Income Inequality and Poverty*. Oxford University Press, New York, NY.

Kalbfleisch, JD and Prentice, RL. (1980) *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, New York, NY.

Kallenberg, WCM. (1983) Intermediate efficiency, theory and examples. *Annals of Statistics*, **11**, 170-182.

Kallenberg, WCM and Ledwina, T. (1999) Data driven rank tests for independence. *Journal of the American Statistical Association*, **94**, 285.

Kaplan, EL and Meier, P. (1958) Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457-481.

Karoly, LA. (1993) The trend in inequality among families, individuals, and workers in the United States: A twenty-five year perspective, in *Uneven Tides: Rising Inequality In America*. S Danziger and P Gottschalk (ed.) Russell Sage., New York, NY, pp. 19-97.

Katz, LF and Murphy, KM. (1992) Changes in relative wages, 1963-1987: Supply and demand factors. *Quarterly Journal of Economics*, **107**, 35-78.

Kelly, DG. (1994) *Introduction to Probability*. Macmillan, New York, NY.

Klerman, JA and Karoly, LA. (1993) *The transition to stable employment: Milling around?*. Unpublished manuscript, Rand, Santa Monica, CA.

Kochan, TA, Katz, HC and McKersie, RB. (1986) *The Transformation of American Industrial Relations*. Basic Books, New York, NY.

Koenker, R. (1984) A note on L-estimates for linear models. *Statistics and Probability Letters*, **2**, 323-325.

Koenker, R and Bassett, G. (1978) Regression quantiles. *Econometrica*, **46**, 33-50.

Koenker, R, Ng, P and Portnoy, S. (1994) Quantile Smoothing Splines. *Biometrika*, **81**, 673-680.

Koenker, R and Portnoy, S. (1987) L-estimation for linear models. *Journal of the American Statistical Association*, **82**, 851-857.

Koenker, R, Portnoy, S and Ng, P. (1992) Nonparametric estimation of conditional quantile functions, in L_1 *Statistical Analysis and Related Methods*. Y Dodge (ed.) North-Holland, Amsterdam, pp. 217-229.

Koenker, R and Zhao, QS. (1994) L-estimation for linear heteroscedastic models. *Journal of Nonparametric Statistics*, **3**, 223-235.

Kooperberg, C and Stone, CJ. (1991) A study of logspline density estimation. *Computational Statistics and Data Analysis*, **12**, 327-347.

Kooperberg, C and Stone, CJ. (1992) Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics*, **1**, 301-328.

Kosters, MH and Ross, R. (1987) *The distribution of earnings and employment opportunities: A re-examination of the evidence*. Occasional paper, American Enterprise Institute.

Kullback, S. (1968) *Information Theory and Statistics*. Dover, New York, NY.

Kullback, S and Leibler, S. (1951) On information and sufficiency. *Annals of Mathematical Statistics*, **22**, 79-86.

Lancaster, GM. (1969) A characterization of certain conformally Euclidean spaces of class one. *Proceedings of the American Mathematical Society*, **21**, 623-628.

Lancaster, HO. (1969) *The Chi-squared Distribution*. John Wiley & Sons, New York, NY.

Lancaster, P. (1969) *Theory of Matrices*. Academic Press, New York, NY.

Ledwina, T. (1994) Data-Driven version of Neyman's smooth tests of fit. *Journal of the American Statistical Association*, **89**, 1000-1005.

Lehmann, EL. (1953) The power of rank tests. *Annals of Mathematical Statistics*, **24**, 23-43.

Lehmann, EL. (1975) *Nonparametrics: Statistical Methods Based On Ranks*. Holden Day, Oakland, CA.

Lehmann, EL. (1983) *The Theory of Point Estimation*. John Wiley & Sons, New York, NY.

Lehmann, EL. (1986) *Testing Statistical Hypotheses*. CRC Press, New York, NY.

Lejeune, MG and Sarda, P. (1988) Quantile regression: A nonparametric approach. *Computational Statistics and Data Analysis*, **6**, 229-239.

Lejeune, MG and Sarda, P. (1992) Smooth estimators of distribution and density functions. *Computational Statistics and Data Analysis*, **14**, 457-471.

Levy, F and Murnane, R. (1992) U.S. earnings levels and earnings inequality: A review of recent trends and proposed explanations. *Journal of Economic Literature*, **30**, 1333-1381.

Li, G, Tiwari, RC and Wells, MT. (1996) Quantile comparison functions in two-sample problems, with application to comparisons of diagnostic markers. *Journal of the American Statistical Association*, **91**, 689-698.

Lin, C-H and Sukhatme, S. (1993) Hoeffding type theorem and power comparisons of some two-sample rank tests. *Journal of the Indian Statistical Association*, **31**, 71-83.

Little, RJA and Rubin, DB. (1978) *Statistical Analysis with Missing Data*. John Wiley & Sons, New York, NY.

Loader, C. (1999) *Local Regression and Likelihood*. Springer-Verlag, New York, NY.

Longford, N. (1994) *Random Coefficient Models*. Chapman Hall, New York, NY.

Lorenz, MO. (1905) Methods of measuring the concentration of wealth. *Journal of the American Statistical Association*, **9**, 209-219.

Majumder, A and Chakravarty, S. (1990) Distribution of personal income: Development of a new model and its application to U.S. income data. *Journal of Applied Econometrics*, **5**, 189-196.

Marcotte, D. (1994) *The Declining Stability of Employment in the U.S.: 1976-1988*. manuscript, University of Maryland.

Marini, M. (1989) Sex differences in earnings in the United States. *American Sociological Review*, **15**, 343-382.

Marron, JS and Schmitz, HP. (1992) Simultaneous density estimation of several income distributions. *Econometric Theory*, **8**, 476-488.

Mayer, KU and Tuma, N. (1990) *Event History Analysis in Life Course Research*. University of Wisconsin Press, Madison, WI.

McCulloch, CE. (1997) Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, **92**, 162-170.

McCulloch, RE. (1989) Local model influence. *Journal of the American Statistical Association*, **84**, 473-478.

McDonald, J. (1984) Some generalized functions for the size distribution of income. *Econometrica*, **52**, 647-663.

Meisenheimer II, JR. (1998) The services industry in the 'good' versus 'bad' job debate. *Monthly Labor Review*, **121**, 22-47.

Mielniczuk, J. (1990) Remark concerning data-dependent bandwidth choice in density estimation. *Statistics and Probability Letters*, **9**, 27-33.

Mielniczuk, J. (1992) Grade estimation of Kullback–Leibler information number. *Probability and Mathematical Statistics*, **13**, 139-147.

Mincer, J and Jovanovic, B. (1981) Labor mobility and wages, in *Studies in Labor Markets*. S Rosen (ed.) University of Chicago Press, Chicago, IL, pp. 21-63.

Mishel, L and Bernstein, J. (1994) *The State of Working America, 1994–1995*. M.E. Sharpe, Armonk, NY.

Monks, J and Pizer, S. (1998) Trends in voluntary and involuntary job turnover. *Industrial Relations*, **37**, 440-459.

Morris, M. (1993) Telling tails explain the discrepancy in sexual partner reports. *Nature*, **365**, 437-440.

Morris, M. (1996) Vive la difference: Continuity and change in the gender wage gap, 1967–1987, in *Social Differentiation and Social Inequality*. J Baron, D Treiman and D Grusky (ed.) Westview Press, Boulder, pp. 211-240.

Morris, M, Bernhardt, AD and Handcock, MS. (1994) Economic inequality: New methods for new trends. *American Sociological Review*, **59**, 205-219.

Murphy, KM and Welch, F. (1992) The structure of wages. *Quarterly Journal of Economics*, **107**, 285-326.

Nahm, JW. (1989) *Nonparametric least absolute deviations estimation*. Unpublished Ph.D. thesis, Department of Economics, University of Wisconsin.

Nair, VN. (1984) Confidence bands for survival functions with censored data: A comparative study. *Technometrics*, **26**, 265-275.

Nasar, S. (1992) Women's progress stalled? Just not so. *New York Times*, October 18, page 1.

National Center on Educational Quality of the Workforce. (1995) *The EQW national employer survey: First findings*. Report, University of Pennsylvania, Philadelphia, PA.

Newey, WK and Powell, JL. (1987) Asymmetric Least Squares Estimation and Testing. *Econometrica*, **55**, 819-847.

Neyman, J. (1937) Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions, Series A*, **236**, 333-380.

Ng, PT. (1996) An algorithm for quantile smoothing splines. *Computational Statistics and Data Analysis*, **22**, 99-118.

Norleans, MX. (1995) *EMU V1.0: An SPLUS object for fitting a generalized mixed linear model for correlated responses with the (restricted) maximum likelihood technique*. Statlib Archive, available at http://lib.stt.cmu.edu/S/emu.v10.

Nygård, F and Sandström, A. (1989) Income inequality measures based on sample surveys. *Journal of Econometrics*, **42**, 81-95.

Osterman, P. (1994) Internal labor markets: Theory and change, in *Labor Economics and Industrial Relations*. C Kerr and P Staudohar (ed.) Harvard University Press, Cambridge, MA.

Pareto, V. (1897) *Course d'Economie Politique*. F. Pichon, Paris.

Parzen, E. (1977) *Nonparametric statistical data science: A unified approach based on density estimation and testing for 'white noise'*. Technical Report 47, Statistical Sciences Division, State University of New York at Buffalo, Buffalo, NY.

Parzen, E. (1979) Nonparametric statistical data modeling. *Journal of the American Statistical Association*, **74**, 105-131.

Parzen, E. (1983) *FUN.STAT: Quantile approach to two sample statistical data analysis*. Canadian Statistical Society Meeting, Vancouver, BC.

Parzen, E. (1992) Comparison change analysis, in *Nonparametric Statistics And Related Topics*. A Saleh (ed.) Elsevier, Holland, pp. 3-15.

Parzen, E. (1993) Change P-P plot and continuous sample quantile function. *Communications in Statistics, Series A*, **22**, 3287-3304.

Parzen, E. (1994) *From comparison density to two sample analysis*. First U.S./Japan Conference on the Frontiers of Statistical Modeling: An Information Approach, pp. 39-56, Netherlands: Kluwer.

Parzen, E. (1999) Statistical methods mining, two sample data analysis, comparison distributions, and quantile limit theorems, in *Asymptotic Methods in Probability and Statistics*. B Szyszkowicz (ed.) Elsevier, Amsterdam, pp. (in press).

Pergamit, M. (1995) *Assessing school to work transitions in the United States*. NLS Discussion Paper 96-32, U.S. Department of Labor, Bureau of Labor Statistics, Washington, DC.

Pfeffer, J. (1994) *Competitive Advantage Through People*. Harvard, Cambridge, MA.

Pfeffer, J and Baron, J. (1988) Taking the workers back out: Recent trends in the structuring of employment. *Research in Organizational Behavior*, **10**, 257-303.

Picot, G, Myles, J and Wannel, T. (1990) *Good jobs/Bad jobs and the declining middle class: 1967-86*. Research Paper 28, Statistics Canada, Business and Labor Market Analysis Group, Ottawa, Ontario.

Piore, MJ and Sabel, CF. (1984) *The Second Industrial Divide*. Basic Books, New York, NY.

Playfair, W. (1786) *The Commercial and Political Atlas; Representing, By Means of Stained Copper-late Charts, the Progress of the Commerce, Revenues, Expenditure, and debts of England, During the Whole of the Eighteenth Century*. T.Burton, for J.Wallis, London, England.

Polivka, AE. (1996) Contingent and alternative work arrangements, defined. *Monthly Labor Review*, **119**, 3-9.

Polivka, AE. (1996) A profile of contingent workers. *Monthly Labor Review*, **119**, 10-21.

Powell, J. (1986) Censored regression quantiles. *Journal of Econometrics*, **32**, 143-155.

Prihoda, TJ. (1981) *A generalized approach to the two sample problem: The quantile approach*. Unpublished Ph.D. thesis, Department Of Statistics, Texas A&M University.

Rae, DW. (1981) *Equalities*. Harvard University Press, Boston, MA.

Randles, RH. (1982) On the asymptotic normality of statistics with estimated parameters. *Annals of Statistics*, **10**, 462-474.

Randles, RH and Wolfe, DA. (1979) *Introduction to the Theory of Nonparametric Statistics*. John Wiley & Sons, New York, NY.

Rao, CR. (1982) Diversity: Its measurement, decomposition, apportionment and analysis. *Sankhyā Series A*, **44**, 1-22.

Rayner, JCW and Best, DJ. (1989) *Smooth Tests of Goodness of Fit*. Oxford University Press, Oxford.

Rice, JA. (1995) *Mathematical Statistics and Data Analysis*. Wadsworth, Pacific Grove, CA.

Rose, S. (1995) *The Decline of Employment Stability in the 1980s*. National Commission on Employment Policy, Washington, DC.

Rosenthal, N. (1985) The shrinking middle class: Myth or reality? *Monthly Labor Review*, **108**, 3-10.

Rubin, DB. (1987) *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, NY.

Ruppert, D and Carroll, RJ. (1980) Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association*, **75**, 828-838.

Salem, A and Mount, T. (1974) A convenient descriptive model of income distribution. *Econometrica*, **42**, 1115-1127.

Sassen, S. (1988) *The Mobility of Labor and Capital: A Study in International Investment and Labor Flow*. Cambridge University Press, New York, NY.

Sawhill, I. (1988) Poverty in the U.S.: Why is it so persistent? *Journal of Economic Literature*, **16**, 1073-1119.

Schrammel, K. (1998) Comparing the labor market success of young adults from two generations. *Monthly Labor Review*, **121**, 3-48.

Schwartz, J and Winship, C. (1980) The welfare approach to measuring inequality, in *Sociological Methodology*. P Holland (ed.) Jossey-Bass, San Francisco.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.

Scott, DW. (1992) *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, New York, NY.

Serfling, RJ. (1980) *Approximation Theorems in Mathematical Statistics*. John Wiley & Sons, New York, NY.

Shannon, CE. (1948) A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379-423.

Shao, J and Tu, D. (1995) *The Jackknife and Bootstrap*. Springer-Verlag, New York, NY.

Sheather, SJ and Jones, MC. (1991) A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, **53**, 683-690.

Shorack, GR and Wellner, JA. (1986) *Empirical Processes With Applications to Statistics*. John Wiley & Sons, New York, NY.

Silverman, BW. (1978) Density ratios, empirical likelihood and cot death. *Applied Statistics*, **X**, 26-33.

Silverman, BW. (1986) *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Simonoff, JS. (1994) The construction and properties of boundary kernels for smoothing sparse multinomials. *Journal of Computational and Graphical Statistics*, **3**, 57-66.

Simonoff, JS. (1996) *Smoothing Methods in Statistics*. Springer-Verlag, New York, NY.

Simonoff, JS. (1998) Three sides of smoothing: categorical data smoothing, nonparametric regression, and density estimation. *International Statistical Review*, **66**, 137-156.

Singh, S and Maddala, G. (1976) A function for size distribution of incomes. *Econometrica*, **44**, 963-970.

Slottje, D. (1984) A measure of income inequality based upon the beta distribution of the second kind. *Economics Letters*, **15**, 369-375.

Slottje, D. (1987) Relative price changes and inequality in the size distribution of various components of income. *Journal of Business and Economic Statistics*, **5**, 19-26.

Smeeding, T and Gottschalk, P. (1996) America's income inequality: Where do we stand? *Challenge*, **39**, 45-53.

Smith, JP, Badmann, RL and Niesswiadomy, M. (1989) Black economic progress after Myrdal. *Journal of Economic Literature*, **27**, 519-564.

Smith, M and Kohn, R. (1996) Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, **75**, 317-343.

Soofi, ES. (1994) Capturing the intangible concept of information. *Journal of the American Statistical Association*, **89**, 1243-1254.

Spenner, K. (1985) The upgrading and downgrading of occupations: Issues, evidence, and implication for education. *Review of Educational Research*, **55**, 125-154.

Stevens, AH. (1996) *Changes in earnings instability and job loss*. Unpublished manuscript, Rutgers University, New Brunswick.

Stone, CJ. (1989) Uniform error bounds involving logspline models. *Annals of Statistics*, **17**, 335-356.

Stone, CJ. (1990) Large-sample inference for log-spline models. *Annals of Statistics*, **18**, 717–741.

Stone, CJ, Hansen, MH and Truong, YK. (1997) Polynomial splines and their tensor products in extended linear modeling. *Annals of Statistics*, **25**, 1371.

Stone, CJ and Koo, C-Y. (1986) Logspline density estimation. *Contemporary Mathematics*, **59**, 1-15.

Stute, W. (1982) The oscillation behavior of empirical processes. *Annals of Probability*, **10**, 86-107.

Stute, W. (1986) Conditional empirical process. *Annals of Statistics*, **14**, 1180-1187.

Swets, JA and Pickett, RM. (1982) *Evaluation of Diagnostic Systems: Methods From Signal Detection Theory*. Academic Press, New York, NY.

Switzer, P. (1976) Confidence procedures for two-sample problems. *Biometrika*, **63**, 13-25.

Tapia, RA and Thompson, JR. (1978) *Nonparametric Probability Density Estimation*. Johns Hopkins University Press, Baltimore, MD.

Theil, H and Laitinen, K. (1980) Singular moment matrices in applied econometrics, in *Multivariate Analysis V*. PR Krishnaiah (ed.) Elsevier, North Holland, pp. 629-649.

Thompson, SK. (1992) *Sampling*. John Wiley & Sons, New York, NY.

Tilly, R. (1990) *Short Hours, Short Shrift: Cases and Consequences of Part-Time Work*. Economic Policy Institute, Washington, DC.

Titterington, DM, Smith, AFM and Makov, UE. (1985) *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York, NY.

Topel, RH. (1997) Factor proportions and relative wages: The supply-side determinants of wage inequality. *Journal of Economic Perspectives*, **11**, 55-74.

Tufte, ER. (1983) *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT.

Tufte, ER. (1990) *Envisioning Information*. Graphics Press, Cheshire, CT.

Tukey, JW. (1965) Which part of the sample contains the information? *Proceedings of the National Academy of Sciences*, **53**, 127-134.

Tukey, JW. (1977) *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

United States Department of Commerce. (1995) *Statistical Abstracts of the United States*. U.S. Government Printing Office, Washington, DC.

United States Department of Commerce. (1997) *Summary of latest NIPA Tables*. Bureau of Economic Analysis, Washington, DC, available at http://www.bea.doc.gov/bea/dn1.htm.

Useem, M and Capelli, P. (1997) The pressures to restructure employment, in *Change at Work*. P Cappelli, L Bassi, H Katz, D Knoke, P Osterman and M Useem (ed.) Oxford University Press, New York, NY, pp. 173-207.

Venables, W and Ripley, B. (1997) *Modern Applied Statistics with S-PLUS*. Springer-Verlag, New York, NY.

Vidakovic, B. (1999) *Statistical Modeling by Wavelets*. John Wiley & Sons, New York, NY.

Von Eye, A and Schuster, C. (1998) *Regression Analysis for Social Sciences*. Academic Press, New York, NY.

Wahba, G. (1981) Data-based optimal smoothing of orthogonal series density estimates. *Annals of the Institute of Statistical Mathematics*, **9**, 146–156.

Wand, M and Jones, M. (1995) *Kernel Smoothing*. Chapman and Hall, London.

Welch, F. (1979) Effects of cohort size on earnings: The baby boom babies' financial bust. *Journal of Political Economy*, **87**, S65-S97.

Wilk, MB and Gnanadesikan, R. (1968) Probability plotting methods for the analysis of data. *Biometrika*, **55**, 1-17.

Wolff, EN. (1995) *Top Heavy: A Study of the Increasing Inequality of Wealth in America.*. Twentieth Century Fund Press, New York, NY.

Wolpin, K. (1987) Estimating a structural search model: The transition from school to work. *Econometrica*, **55**, 801-817.

Wood, A. (1994) *North-South Trade, Employment, and Inequality: Changing Fortunes in a Skill-driven World*. Oxford University Press, New York, NY.

Yamaguchi, K. (1991) *Event History Analysis*. Sage Publications, Newbury Park, CA.

Yu, K and Jones, MC. (1998) Local linear quantile regression. *Journal of the American Statistical Association*, **93**, 228-237.

This page intentionally left blank

Subject Index