Statistical Inference for the Relative Density

MARK S. HANDCOCK and PAUL L. JANSSEN

Sociological Methods Research 2002; 30; 394
DOI: 10.1177/0049124102030003005

Social scientists are increasingly interested in techniques for comparing changes in distributional shape in addition to mean levels. One such technique is based on the relative distribution, a nonparametric summary of the information required for scale-invariant comparisons between two distributions. The relative distribution is being used by social scientists to represent and analyze distributional differences, enabling researchers to move well beyond comparisons of means and variances in a simple intuitive way. The authors develop a nonparametric estimator for the relative density function. They study its asymptotic properties, derive computable expressions for the asymptotic variance, and consider local bandwidth selection. They also illustrate how the relative density can be decomposed into a component due to location differences and a component due to shape differences. This decomposition identifies that component of interdistributional dissimilarity due to interdistributional inequality. The methods are illustrated by comparing the earnings distributions of working women to that of working men based on the 1990 census and to women from 1967 to 1996.

Statistical Inference for the Relative Density

MARK S. HANDCOCK
University of Washington

PAUL L. JANSSEN
Limburgs Universitair Centrum, Belgium

1. INTRODUCTION

In sociological research, differences among groups or changes over time are a common focus of study. While means and variances are typically the basis for the methods used in this research, the underlying social theory often implies properties of distributions that are not well captured by these summary measures.

AUTHORS' NOTE: *The authors acknowledge helpful comments by Annette D. Bernhardt, Jeffrey S. Simonoff, Martina Morris, and HenSiong Tan. Special thanks are given to Jack Hall for pointing out the connection with ROC curves. We would also like to thank an anonymous referee for extensive comments that have improved the presentation of this article. This research was supported by the Center for Statistics and the Social Sciences with funds from the University Initiatives Fund at the University of Washington.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 30 No. 3, February 2002 394-424

Consider some of the current controversies regarding growing inequality in earnings, the effects of job change on permanent wages, and racial differences in test scores. The distributional differences that animate the debates in these areas are complex. They comprise the usual mean-shifts and changes in variance but also more subtle comparisons of changes in the upper and lower tails of the distributions. Survey and census data on such attributes contain a wealth of distributional information, but traditional methods of data analysis leave much of this information untapped. A good example of the limitations of the traditional summary measures is the debate surrounding the gender wage gap. While the ratio of median earnings of women to men was stable in the middle half of past century, it began to narrow in the 1980s, and this led to predictions that gender inequality may be history. However, an analysis of the distributions of men's and women's earnings makes it clear that progress toward gender equality of earnings was largely limited to women in the bottom end of the earnings distribution (Bernhardt, Morris, and Handcock 1995). Readers interested in trends in the gender wage gap can find more information in the review article by Morris and Western (1999). Many other examples of the failure of the traditional numerical summary methods are given in Handcock and Morris (1998, 1999). Other illustrations using census and Current Population Survey information on the earnings distributions of women and men are given in this article.

The relative distribution is a tool for the comparison of distributions in terms of their differences in shape. It appears, explicitly and implicitly, in many independent research areas (Parzen 1992; Cwik and Mielniczuk 1993; Holmgren 1995; Li, Tiwari, and Wells 1996). Recently, it has been used to study changes in economic characteristics over time and between demographic groups. For example, Morris, Bernhardt, and Handcock (1994) study changes in yearly earnings by race and gender from 1967 to 1987. Bernhardt et al. (1995) used it and its extensions to take a closer look at the shrinking gender gap in earnings. Handcock and Morris (1998) used the relative distribution to study the changes in the distribution of yearly hours worked between 1975 and 1993. In each of these areas of study, the pattern of the changes has made it necessary to study differences beyond the usual differences in the summary measures of

location and variation (Butler and McDonald 1987; Karoly 1993). Additional applications are given in Handcock and Morris (1999).

1.1. REVIEW OF FUNDAMENTAL IDEAS: THE RELATIVE DISTRIBUTION

In this section, we briefly review the statistical formalism underlying the concept of a relative distribution. A more expansive development is given in Section 2.2 of Handcock and Morris (1999).

Let F_0 be the cumulative distribution function (CDF) of an outcome attribute measured on a reference group and F be the corresponding CDF for a comparison group. Typically, the comparison group 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 distributions of the outcome attribute in the reference and comparison groups.

Let $Y_0 \sim F_0$ and $Y \sim F$. We suppose that F_0 and F are absolutely continuous with common support. The *grade transformation* of Y to Y_0 is defined as the random variable (Cwik and Mielniczuk 1989):

$$R = F_0(Y). \tag{1}$$

R is obtained from Y by transforming it by the function F_0, and so it is continuous with outcome space $[0, 1]$. As R measures the relative rank of Y compared to Y_0, we refer to the distribution of R as the *relative distribution*. We can express the CDF of R as

$$G(r) \;=\; F(Q_0(r)), \quad 0 \le r \le 1 \tag{2}$$

where r represents the proportion of values, and $Q_0(r) = \inf_y\{y \mid F_o(y) \;\ge\; r\}$ is the quantile function of F_0. The probability density function (PDF) of R is

$$g(r) = \frac{f(Q_0(r))}{f_0(Q_0(r))}, \qquad 0 \le r \le 1. \tag{3}$$

Figure 1 presents the constituents of a relative distribution and their relationship. Panel (a) graphs the densities corresponding to reference and comparison distributions for hypothetical groups of income earners. The horizontal scale is in thousands of dollars. In this illustration, the reference group distribution is approximately Gaussian, while the comparison group distribution has a higher me-

dian and is left-skewed. A solid vertical line is drawn at the quantile corresponding to $r = 0.6$, the value of y at the 60th percentile of Y_0. Here $y(r) = Q_0(r) = 7.63$. 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: $f_0(Q_0(r))$ and $f(Q_0(r))$ for the reference and comparison groups, respectively. Note that $f(Q_0(r))$ is about half of $f_0(Q_0(r))$. The relative density is defined as the ratio of these two quantities (see equation (3)) for every value r in $[0, 1]$, and this density is plotted in the bottom panel of Figure 1. Note that at $r = 0.6$, the relative density is about 0.5, as the top graph suggests. For values in the lower eight deciles of Y_0 ($r < 0.8$), the relative density is less than 1, indicating a lower frequency of observations in the comparison distribution Y, and in the remaining two deciles the value is greater than 1, indicating a higher frequency of observations in the comparison group. Finally, the upper axis is labeled in thousands of dollars. This complements the lower axis labeling in terms of the proportion of the reference group: It gives the dollar value at the corresponding percentile of the reference group. For example, the 90th percentile (i.e., $r = 0.9$) corresponds to an earnings of \$10,000.

If the two distributions are identical, then the CDF of the relative distribution is a 45^o line, and the PDF of the relative distribution is the uniform PDF.

The relative distribution is an intuitively appealing approach to the comparison problem because both the density and the CDF have clear, simple interpretations. The relative density $g(r)$ can be interpreted as the ratio of the comparison population to the reference population at a given level ($Q_0(r)$). 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. More technically, a proportion $G(r)$ of the Y is below the values of a proportion r of Y_0.

The relative density has been explicitly studied in at least two areas. Parzen (1977, 1992) has studied the relative CDF and density as part of "comparison change analysis." He refers to them as the *comparison density* and *comparison distribution*. Along the same thread, Eubank, LaRiccia, and Rosenstein (1987) have developed statistics based on the relative density for comparing distributions. Separately, Cwik and Mielniczuk (1989, 1990, 1993) have

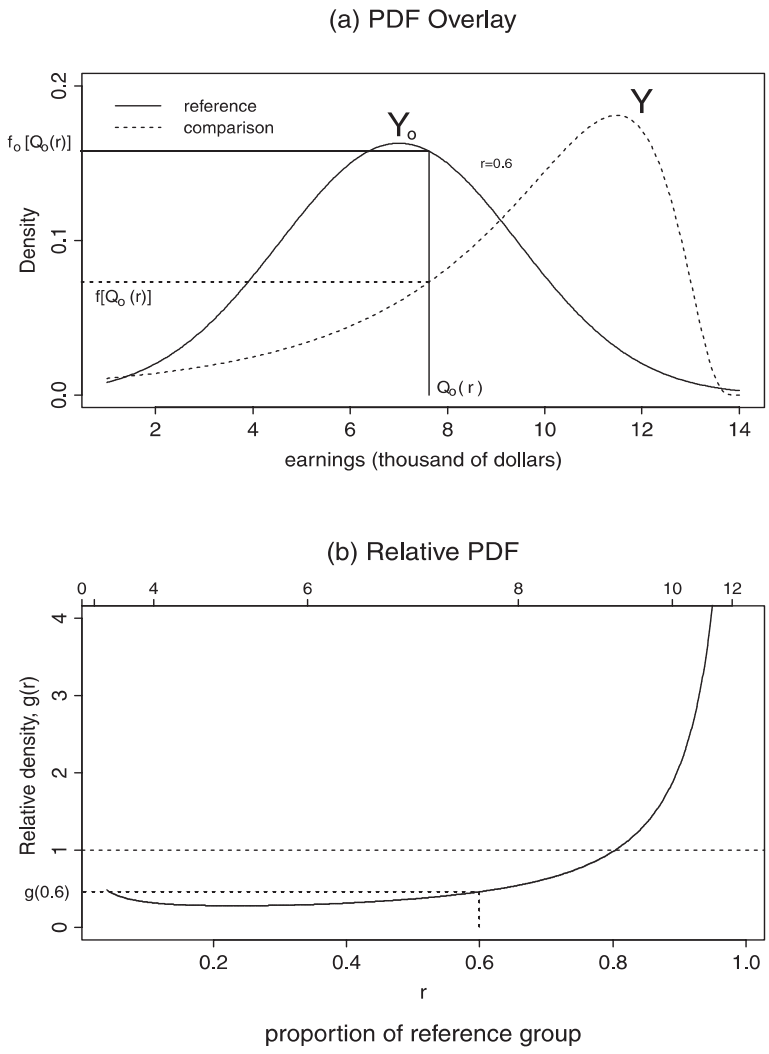


Figure 1: Components of the Relative Density for Hypothetical Reference and Comparison Groups

NOTE: The top panel has graphs of their densities, and the bottom panel is their relative density. PDF = probability density function.

considered nonparametric density estimation for the relative density. They have developed a method for choosing an estimator that is appropriately smooth. Based on (1), Cwik and Mielniczuk (1989) have referred to the relative density as the *grade density*. Cwik and Mielniczuk (1990) have considered related issues in their study of Neyman-Pearson curves.

The relative CDF $G(r)$ is implicitly a theoretical 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 $\{(r, G(r)) : 0 \leq r \leq 1\}$. Holmgren (1995) gives a nice discussion of the merits of the relative CDF (p-p plots) compared to q-q plots. His theoretical justification and results carry over to a relative distribution framework. He formally shows that, under appropriate assumptions, the relative distribution forms a complete summary of the information required for scale-invariant comparisons between the comparison and reference distributions. Our general framework and inferential results extend his methodology. The relative CDF is an ordinal dominance curve related to receiver operating characteristic (ROC) curves used in the evaluation of the performance of medical tests for separating two populations (Begg 1991; Campbell 1994). The precise relationship between the relative CDF and ROC curves is discussed in Li et al. (1996).

Inference for the relative CDF has been considered by Gastwirth (1968 [his Theorem 3.2]). Cwik and Mielniczuk (1990) developed an estimator of the relative CDF based on integrating an estimate of the relative density. They showed the uniform strong consistency of their estimate. Handcock and Janssen (1998) showed that the natural estimator of the CDF is a jointly asymptotic Gaussian by treating it as an generalized U-statistic with an estimated parameter. A statistic is classified as a U-statistic if it can be expressed as the sum of component statistics that satisfy certain symmetry conditions. The classification is important as the statistical properties of such aggregate statistics have been deeply explored (Hoeffding 1948). We explicitly construct U-statistics in the appendix that provide a window of insight into the statistical properties of the estimators considered in this article. For more information on U-statistics, see Serfling (1980). In the context of ROC curve esti-

mation, Hsieh (1995) also derives the result under slightly stronger conditions based on a strong approximation of the empirical ROC curve (his Lemma 3). Handcock and Janssen (1998) give a nice direct proof based on classical U-statistic methodology.

In this article, we develop the statistical estimation of, and inference for, the relative density. In the remainder of this section, we give two applications to illustrate the value of the approach. In Section 2, we propose a kernel-based estimator and show that it is asymptotically Gaussian. We derive expressions for the asymptotic variance and use this to determine confidence intervals. We also show how the bandwidth can be chosen using local bandwidth selection criteria developed in Cao et al. (2000). Section 3 develops a distributional decomposition of the relative density into a component representing the effect of the difference in location between two distributions and a component representing the relative distribution adjusted for this difference. This decomposition allows the location and shape difference between two distributions to be separated and graphically compared.

1.2. APPLICATION TO MEN'S AND WOMEN'S EARNINGS DISTRIBUTIONS

An example of two distributions and their relative density is given in Figures 2 and 3. Figure 2 presents and compares the earnings distributions for white men and white women based on the 1990 decennial Census of Population and Housing. The density estimates in Figure 2 are based on kernel smoothing techniques (Simonoff 1998). An interesting feature of census earnings data is the rounding of the values that leads to a density that is much more spiky than typically seen. A key attribute of the estimator is the so-called "bandwidth," which controls the degree of smoothness. Intuitively, a key feature of the underlying density is its smoothness—that is, how rapidly the density changes for small changes in earnings. For example, compare the densities in Figure 2 to the Gaussian density with the same mean and variance. The densities in the figure are much less smooth and have many local modes, even though their overall shapes are similar. We would like to allow the smoothness of our estimator of the density to be flexible. The data will tell us much about the smoothness of the underlying density, and we would also

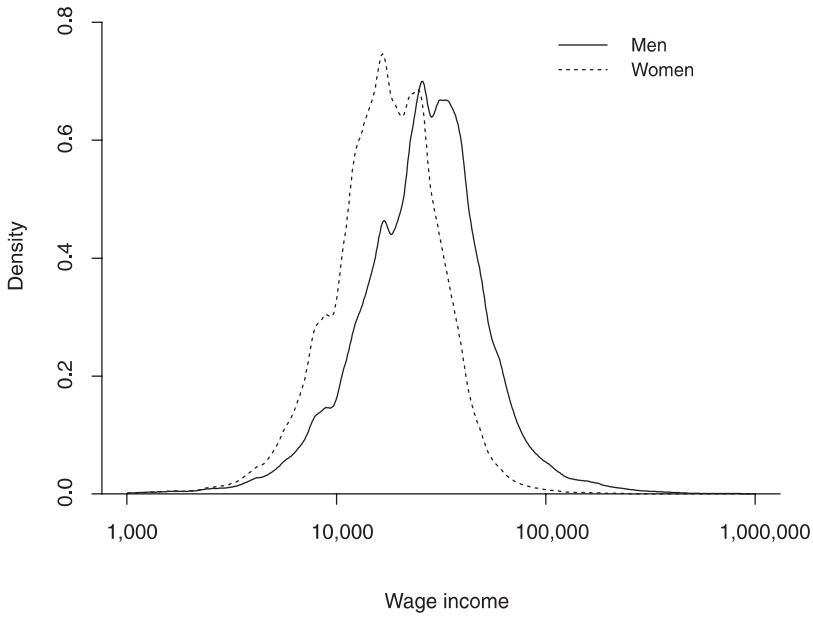


Figure 2: The Distributions of White Women's and White Men's Earnings from the 1990 Census

NOTE: The horizontal axis is on a logarithmic scale.

like the estimate to reflect our a priori sense that the density should be somewhat smooth but likely not as smooth as the Gaussian density. Most estimators of densities include a parameter that increases the smoothness of the estimate as it increases in value. The details of the mechanism for controlling the smoothness depend on the form of the estimator. We give the details for kernel estimators in Section 2, where the bandwidth is denoted by h (see Section 2 for details). In this case, the bandwidth for the two estimators was chosen by the Akaike Information Criterion (AIC), corrected for its tendency to undersmooth (Simonoff 1998). The sample sizes here are very large ($n = 428,902$ for the men and $n = 284,866$ for the women).

The center of the men's distribution appears to be at a higher earnings than the center of the women's distribution (medians of $26,625

and $17,570, respectively). In addition, the spread of the men's distribution appears to be slightly wider (standard deviations in log-dollars of 0.60 and 0.70, respectively, and $32,764 and $14,666, respectively).

Figure 3 is the density of the relative distribution of women's to men's earnings. A description of the estimator and a study of its statistical properties are given in the next section. The value of one represents the relative density if the two distributions were identical. We can see that women are markedly overrepresented in the lower quantiles of the men's distribution. The relative frequency does not balance out until about the 45th percentile of the men's distribution. The decline in the relative frequency of women is steady for the top nine-tenths of men's earnings. The bumps apparent in the density are due to uneven heaping in the reported earnings for both samples.

The relative density enhances comparison of the distributions in two ways. First, it expresses the relative frequency in terms of a ratio, which is easier to understand both visually and numerically. Second, it rescales the horizontal axis so that length is equivalent to the proportion of men's earnings. This facilitates direct comparisons between women's and men's earnings because the two axes are now in comparable units. For example, women are more than 1.7 times more frequent than men between the 5th and 20th percentiles of the men's distribution. The upper axis is labeled in thousands of dollars earned by men. As in Figure 1, this relates to the lower axis, which is labeled in terms of the proportion of men. Effectively, it gives the quantiles of the men's distribution. This enables the absolute dollar values of earnings to be referenced in the overall relative plot. For example, we see that the mode of the relative distribution occurs at about $7,500. At these earnings, the frequency of women is about 2.2 times that of men. We also see that for all earnings above $25,000 (the 40th percentile for men), men are relatively more prevalent than women.

These figures demonstrate 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 2); rather, they complement the overlay by focusing on those characteristics of the individual distributions essential for comparison alone. Figures 2 and 3 provide absolute and relative comparisons, respectively.

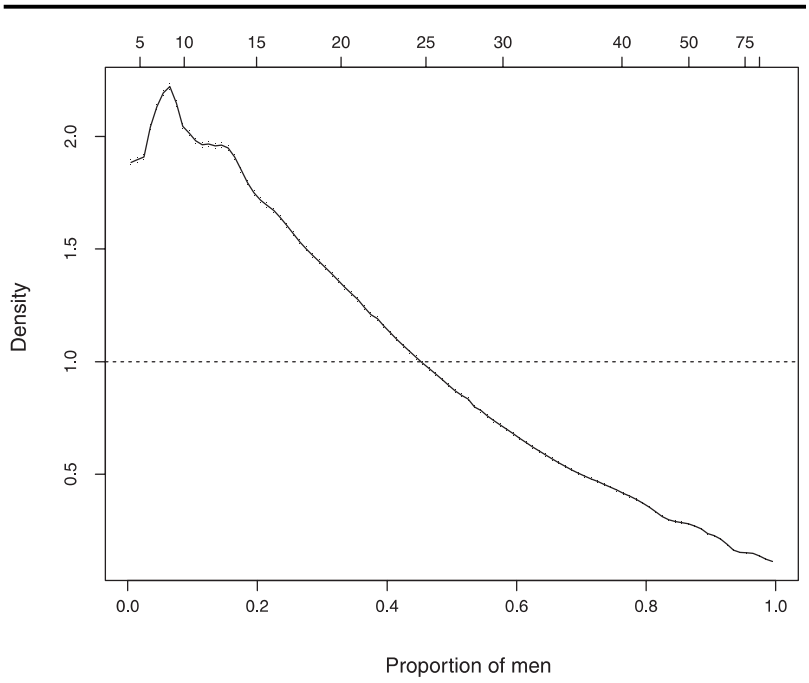


Figure 3: The Relative Density of White Women's to White Men's Earnings From the 1990 Census

NOTE: The upper axis is labeled in thousands of dollars. The dotted lines are 95 percent pointwise confidence bounds.

2. ESTIMATION OF THE RELATIVE DENSITY

In this section, we consider estimating the relative density $g(r)$. As in our applications, in practice, information about the reference and comparison distributions is often available in the form of independent samples from both distributions. Hence, suppose $Y_1, Y_2, \ldots, Y_m$ are i.i.d. F, and independently $Y_{01}, Y_{02}, \ldots, Y_{0n}$ are i.i.d. F_0. We do not consider sample weight here, but it can be incorporated in a straightforward manner. Denote the empirical distribution function of Y_0 by $F_{n0}(y) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{I}(Y_{0i} \leq y)$, where $\mathcal{I}(\cdot)$ is the indicator function.

2.1. ESTIMATION WHEN THE REFERENCE DISTRIBUTION IS KNOWN

Some insight into the estimation process can be gained by considering the hypothetical situation in which we directly observe independent values from the relative distribution (1): $R_1, R_2, \ldots, R_m$ i.i.d. from G (e.g., $R_j = F_0(Y_j)$). We refer to the R_j as the *relative data*. The estimation of $g(r)$ is then a standard density estimation problem in which the support is [0, 1]. In this article, we will consider kernel density estimates (Silverman 1986):

$$g_m(r) = \frac{1}{mh_m} \sum_{j=1}^{m} K\left(\frac{r - R_j}{h_m}\right), \tag{4}$$

where $K(\cdot)$ satisfies the conditions

$$\begin{aligned} \int_{-\infty}^{\infty} K(u)du &= 1, \\ \int_{-\infty}^{\infty} uK(u)du &= 0, \\ \int_{-\infty}^{\infty} u^2K(u)du &= \sigma_K^2 > 0. \end{aligned} \tag{5}$$

The basic logic of this estimator is that an estimate of the density at r can be thought of as a weighted average proportion of relative data values that are close to r. If there are many data values close to r, then the density will likely be higher there. The kernel is often chosen to have similar shape to the Gaussian density. This weights the number of data values closer to r more than those further away—in line with the notion that values further away should have less influence on the estimate at r than the values that are close. Wand and Jones (1995) give a book-length treatment of the motivations and properties of kernel density estimators such as these. We assume that the underlying relative density is sufficiently smooth (g being uniformly continuous and g''' being square integrable). If $h_m \to 0$ with $mh_m \to \infty$ as $m \to \infty$, then by Taylor series expansions (Silverman 1986),

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

and

$$\mathbb{V}[g_m(r)] = \frac{g(r)R(K)}{mh_m} + O(m^{-1}),$$

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

2.2. ESTIMATION WHEN THE REFERENCE DISTRIBUTION IS UNKNOWN

In most situations, we do not observe the relative data directly but only have access to independent information on the comparison and reference distributions. In this case, we can consider the *quasirelative data* generated by the grade transformation of the Y_i by F_n rather than F_0 :

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

We use this term to distinguish the Q_j from the closely related R_j. As $F_{n0}(\cdot)$ estimates $F_0(\cdot)$, we might expect $\{Q_j\}_{j=1}^m$ to act as a surrogate for $\{R_j\}_{j=1}^m$. 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 (their pairwise correlation is $O(n^{-1})$). In any case, the behavior of estimates based on the quasi-relative data must be determined separately from those for estimates based directly on relative data.

Motivated by (4), we consider the following estimator of $g(r)$:

$$g_{n,m}(r) = \frac{1}{mh_m} \sum_{j=1}^{m} K\left(\frac{r - Q_j}{h_m}\right). \tag{6}$$

The asymptotic properties of the estimator are described in the following result:

Theorem 1. 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 (5)) 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

$$\sqrt{mh_m}\left[\, g_{n,m}(r) - g(r) \,\right] \xrightarrow{\mathcal{D}}$$

$$N\left(0, \quad g(r)R(K) + \kappa^2 g^2(r)R(K) \right),$$

where $R(K) = \int_{-1}^{1} K^2(z)dz$.

This result suggests that the sampling distribution of $g_{n,m}(r)$ can be approximated by a Gaussian distribution with mean $g(r)$ and variance:

$$\mathbb{V}[g_{n,m}(r)] = \frac{g(r)R(K)}{mh_m} + \frac{g^2(r)R(K)}{nh_m} \tag{7}$$

when the sample sizes are large. It is sufficient that $f_0(x)$ and $f(x)$ are positive, bounded, and uniformly continuous in some neighborhood of $x = Q_0(r)$. It is informative to compare the properties of this estimator to those of the estimator (4). 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. In the appendix, we give a proof for the result 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. The methodology has some independent interest. For example, it can be used in the extension of this result to cover nonparametric estimators of the relative density based on local-polynomial fitting.

Simulation results (not shown here) 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.1) for $g_{n,m}(r)$ play a significant role (even though they are asymptotically negligible). By working through the proof of the theorem, it is possible to refine the variance estimate to give the following:

Corollary 1. An expression for the variance of $g_{n,m}(r)$ that is more accurate when the sample sizes are small is

$$\mathbb{V}[g_{n,m}(r)] \approx \frac{g(r)R(K) - h_m g^2(r)}{mh_m}$$

$$+\frac{\left(g(r)\sqrt{R(K)-h_m}+g'(r)R(K)\sqrt{h_m r(1-r)}\right)^2}{nh_m}$$

as $h_m \to 0$ with $mh_m \to \infty, m/n \to \kappa^2 < \infty$ as $m \to \infty$.

Simulation results (not shown here) 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 Gaussian approximation in the calculation of (pointwise) confidence intervals for $g(r)$ based on $g_{n,m}(r)$.

If the sample sizes are small, the bootstrap can be used 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 Gaussian approximation in moderate to large samples (i.e., $n, m > 30$). The sample sizes for the referenced applications and the one considered in this study tend to be large (e.g., greater than 1,000), and the approximations will be very close to the exact values.

If the sample size is not small, we can use the Gaussian approximation to the distribution of the estimate as the basis for a test for a given significance level α:

$$P\Bigg(g_{n,m}(r) - z_{\alpha/2} \times \sqrt{\widehat{\mathbb{V}}\left[g_{n,m}(r)\right]} \;\leq\; g(r)$$
$$\leq g_{n,m}(r) + z_{\alpha/2} \times \sqrt{\widehat{\mathbb{V}}\left[g_{n,m}(r)\right]} \Bigg) \to 1-\alpha$$

as $m \to \infty, m/n \to \kappa^2 < \infty$. Here $\sqrt{\widehat{\mathbb{V}}\left[g_{n,m}(r)\right]}$ is the variance estimate obtained by replacing the relative density by its estimates in the variance expression of Corollary 1.

2.3. SELECTION OF THE KERNEL BANDWIDTH

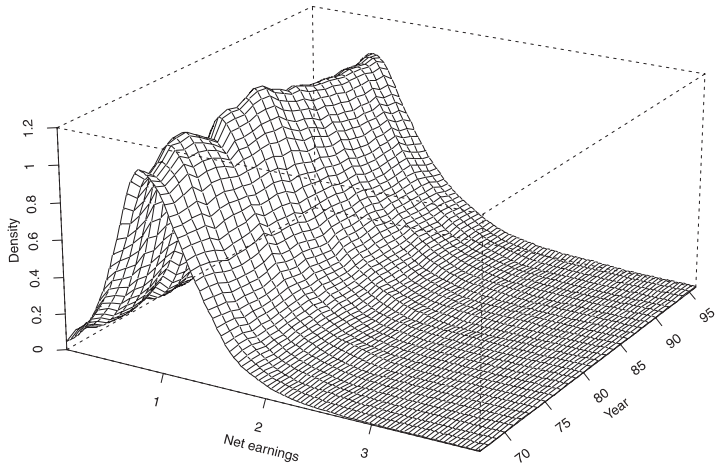
To apply the estimator (6), a kernel function K and the bandwidth h must be selected. The choice for the kernel function depends on properties of the (unknown) g. Fortunately, there are many choices that appear to work similarly well. To choose the bandwidth, it is common to select values that approximately minimize the average mean squared error of the estimator over the entire unit interval. In

the case where the reference distribution is known, the choice of h has been well studied (see, e.g., Simonoff 1998). Cao et al. (2000) have developed an approach that allows the bandwidth for estimation of $g(r)$ to vary with r. This has the advantage of allowing the estimator to adapt to the local properties of the underlying relative density. That is, for values of r where $g(r)$ is smooth, the bandwidth can be chosen to be large (reducing variability), and for values of r where $g(r)$ is rapidly changing, the bandwidth can be chosen to be small (reducing bias). Cao et al. propose a method based on a two-stage smoothed plug-in approach with a beta distribution as the reference. They show that the resulting estimator has good properties when applied to simulated data (see Cao et al. 2000 for details). All the relative density estimates in this study use local bandwidths chosen according to their method. We have tried alternative fixed-bandwidth estimators (e.g., Cwik and Mielniczuk 1993) and found the local bandwidth approach to be preferable for samples of this size. We will not give explicit details of the implementation here but refer the reader to the development given in Cao et al. However, we have written code to implement this method and have made it freely available on the relative distribution Web site (see Section 5).

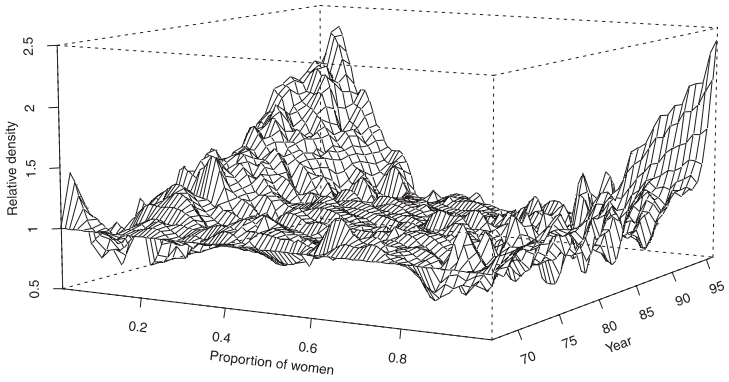
3. APPLICATION TO THE COMPARISON OF WOMEN'S NET EARNINGS OVER TIME

In many applications, we have individual-level data available from economic surveys and collected over time. In this case, a common reference distribution can be used for each time point. The resulting series of relative distributions facilitates comparison among the time series of distributions to produce a clearer image of the relative changes over time.

Figure 4 represents the net earning distributions for working women from 1967 through 1996. 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 females, ages 16 to 66, and excludes the self-employed, full-time students, and those in the military and in farming. We take real annual earnings as our income variable, defined as the income re-



(a) Net earnings densities over time



(b) Net earnings relative densities over time

Figure 4: Net Earnings Distributions Over Time for Working Women From 1967-1996

SOURCE: Based on the Current Population Survey.

spondents reported receiving in wage and salary before deductions during the previous year. Net earnings is defined as the earnings divided by mean earnings for that year. The resulting sample size is of the order of 10,000 per year.

Panel (a) shows nonparametric estimates of the densities for each year. Distributions of this type are considered in Härdle (1990) and Engel and Kneip (1999). The graph illuminates the basic log-Gaussian structure of the distributions and suggests the variation in absolute shape over time. Panel (b) presents the relative density estimates using the 1967 distribution as the common reference distribution. Hence, the distribution for 1967 is uniform, while the others represent changes relative to that in 1967. These distributions complement those in the first panel by emphasizing how the distributions have changed relative to each other. From them the dramatic increase in variability in net earnings is apparent over time. The earlier relative densities are uniform and stay approximately so through the early 1970s. They confirm and visualize one of the "stylized facts" of changes in economic conditions for working women over this period: From the early 1980s, there was a marked trend of increase in the proportion of women in both the upper and lower tails of the net earnings distribution, as evidenced by the U-shaped densities.

4. DECOMPOSING THE RELATIVE DISTRIBUTION INTO LOCATION AND SHAPE COMPONENTS

The similarity in shape of the two densities in Figure 1 suggests that the differences between the distributions could be largely explained in terms of a location shift. To investigate this possibility, we can compare the multiplicatively scaled version of the women's earnings that has the same median as the men's earnings (i.e., $Y^m = Y \times \text{median}(Y_0)/\text{median}(Y)$) to the men's earnings. Figure 5 presents the relative density of the scaled women's earnings to those of the original men's distribution. Note that the concept of location is closely tied to the measurement scale (here dollars). For example, equating locations additively will give different results than equating them multiplicatively. Which is better depends on the shape

of the distributions. In this case, the two distributions have similar shape on the log-dollar scale, and so it makes sense to shift the distributions additively on the log-scale to equate them. This corresponds to a multiplicative shift on the original dollar scale. Note that the relative density is the same if the men are shifted to the women or the women are shifted to the men.

The median matching reduces the discrepancy between the two distributions in the sense that the relative density is closer to the uniform density. The sinking tails indicate that the women's earnings are less frequent in the extremes than the men's earnings (consistent with less variability in the women's distribution compared to the men's distribution). The relative greater density of women around the 40 percent quantile and 80 percent quantile is also apparent. These relative densities are based on samples that have been median matched using the ratio of sample medians (rather than the unknown ratio of population medians). This introduces an additional small amount of variability that has not been reflected in the confidence bands given in the figure.

We would like to quantify the degree to which the overall difference between the earnings of men and women apparent in Figure 5 is attributable to this difference in median value and how much is attributable to other differences between the two distributions. We would like, then, to decompose the overall difference between the two populations (as expressed through their relative distribution) into a component due to the difference in location between the two populations and a component due to the difference in shape between the two populations. The probabilistic formulation of this idea is as follows. Let Y_A denote a random variable describing the reference population multiplicatively, scaled to have the same median as the comparison population (i.e., the random variable $\rho \times Y_0$ where $\rho = \text{median}(Y)/\text{median}(Y_0)$). We shall say that Y_A is Y_0 *median adjusted to* Y. The CDF of Y_A can be written as $FA(y) = F_0(y/\rho)$. Y_A defines a hypothetical reference population that has the same location (as represented via the median) as the comparison population and retains the shape of the reference population. We remark that alternative choices of location matching (e.g., additive median shift, mean shifting) can also be used. In that case, the development given below is similar, with the alternative Y_A replacing the median-

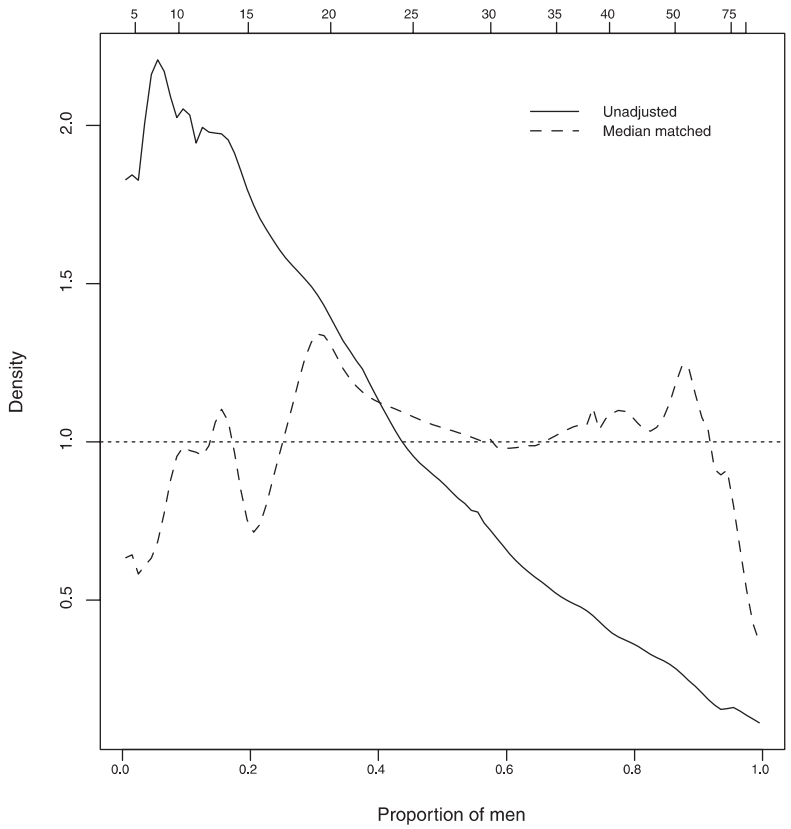


Figure 5: The Relative Density of Median-Matched Women's to Men's Earnings

NOTE: The upper axis is labeled in thousands of dollars. The dotted lines are 95 percent pointwise confidence bounds.

adjusted version given here. We choose the median-adjusted definition because the distribution of earnings is approximately log-normal. For example, if we consider the earnings distributions of men and women, the hypothetical distribution for men's earnings would have the same median as the women's distribution and the same shape as the men's distribution.

4.1. DECOMPOSING THE RELATIVE DISTRIBUTION

From these three distributions—Y, Y_A and Y_0—we can construct two relative distributions that represent the effects of the location and shape changes. In the notation of Section 1.1, let $R_{1,0} = F_0(Y)$, $R_0^A = F_0(Y_A)$, and $R_A^1 = F_A(Y)$. Note that the relative distribution R_0^A will have a uniform distribution when the comparison and reference populations have the same median. In general, R_0^A represents the location difference between Y_0 and Y. The relative distribution R_A^1 will have a uniform distribution when the only difference between the two groups is the median adjustment. We can interpret R_A^1 as representing the differences between Y and Y_0 not directly due to location differences (as measured by the median adjustment). In this sense, it would be the relative distribution had the two populations had the same median.

These two components form a decomposition of $R_{1,0}$ in the sense that R_A^1 is the relative distribution of $R_{1,0}$ to R_0^A. The decomposition can be graphically represented in terms of the corresponding densities. Denote the densities of $R_{1,0}$, R_0^A, and R_A^1 by g_0^1, g_0^A, and g_A^1, respectively. Denote the CDF of R_0^A by $F_0^A(r) = F_A(Q_0(r))$, $0 \leq r \leq 1$. Mathematically, the relationship between the densities is

$$g_0^1(r) = g_0^A(r) \times g_A^1(p) \quad \text{where} \quad p = F_0^A(r), \quad 0 \leq r \leq 1.$$

Note that r is the percentile in the reference population for a given value of the measurement, and p is the percentile in the hypothetical reference population of that same value. Thus, we can interpret the relationship in terms of the value of the measurement. Heuristically, we can represent the decomposition of the relative densities by

$$\begin{array}{c}\text{Overall relative} \\ \text{density}\end{array} = \begin{array}{c}\text{density ratio for} \\ \text{the location difference}\end{array} \times \begin{array}{c}\text{density ratio for} \\ \text{the shape difference.}\end{array}$$

In this sense, the overall relative density between the populations can be thought of as the product of a relative density representing the effect of the difference in location and the location-adjusted relative density. The latter component may be thought of as the discrepancy due to the different shapes of the two distributions. By comparing plots of g_0^1, g_0^A, and g_A^1 side-by-side, we can gauge the relative size and nature of the components.

4.2. EXAMPLE: THE IMPACT OF LOCATION ON RELATIVE EARNINGS

In the beginning of this section, we compared women's earnings to a version of men's earnings with the same median. Figure 5 graphically displays the degree to which the discrepancy between men's and women's earnings is a reflection of a location shift. We can now formalize this comparison in terms of the above decomposition.

Figure 6 graphically represents the decomposition of the relative density of women's to men's earnings in terms of the effects of location differences. Panel (a) is the (unadjusted) relative density of Figure 6 that is decomposed into the two components. Panel (b) represents the component of (a) that is attributable to differences in the location (as expressed by the median) between the two gender populations. We see that the difference in location describes the majority of the discrepancy between men's and women's earnings. A location shift alone would have resulted in a larger discrepancy in the lower tail and a somewhat decreased discrepancy in the upper tail. Panel (c) represents the relative density in (a) adjusted for the difference in the location. In terms of the notation of the previous section, panel (a) is $g_0^1(r)$, panel (b) is $g_0^A(r)$, and panel (c) is $g_A^1(r)$ each for $0 \leq r \leq 1$.

The differences in shape between the two distributions have a moderate (and statistically significant) effect on the differences in earnings compared to the overall effect of the difference in location between men's and women's earnings. The shape effects are most apparent in the tails of the distribution, where they lower the relative density by 20 to 40 percent. In addition, the shape changes tend to ameliorate the discrepancy between men and women for low earners and exacerbate them for high earners. This can be seen because

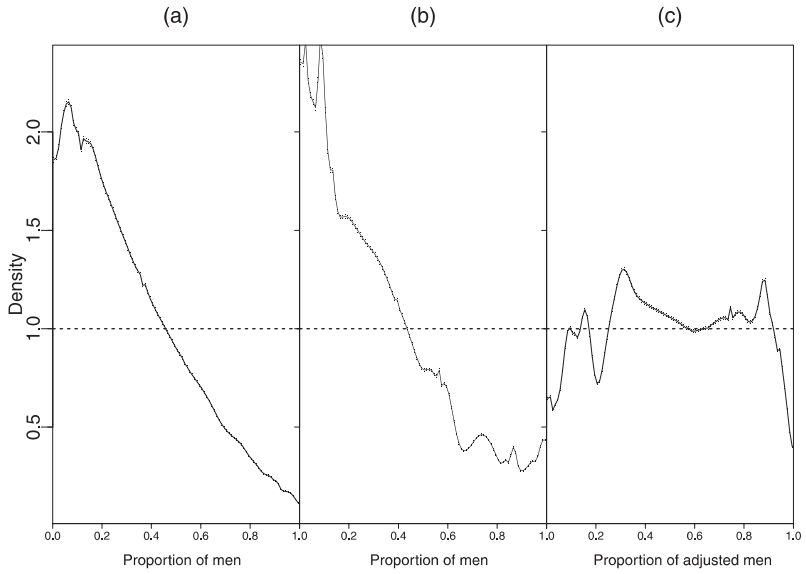


Figure 6: Decomposing the Differences Between Women's to Men's Earnings Into the Impact of Their Difference in Location and Differences in Shape

NOTE: (a) The (unadjusted) relative density of women's to men's earnings from the 1990 census. (b) The effect of the location difference between women's to men's earnings on (c). (c) The relative density of women's to men's earnings, adjusted for the difference in location.

the relative density is below unity for the lower 25 and the upper 10 percent of the men's distribution. This means that—net of the higher median earnings for men—there are fewer women in the extremes of the earnings distribution than men. Thus, the shape differences reduce the relative frequency of both very low- and high-earning women. In contrast, the higher median earnings for men result in higher relative frequency of women in the lower earnings and decrease them in the higher earnings.

5. DISCUSSION

The relative density estimator discussed in Section 2 has intuitive appeal and is simple to calculate. The distributional results also allow hypotheses about the relationship between distributions to be tested. However, kernel estimators such as $g_{n,m}(r)$ have well-known drawbacks such as the substantial underestimation of $g(r)$ for values of r close to 0 or 1 (relative to the size of h). However, this problem can be overcome in a number of ways, including the use of a boundary kernel, as considered in Cwik and Mielniczuk (1993) (see also Wand and Jones 1995). A number of alternative estimators to the kernel estimator can be considered. The most natural are local likelihood smoothers (Loader 1999) based on the quasi-relative data. We are in the process of investigating the properties of this approach, and our results will appear elsewhere.

The location-shape decomposition discussed in Section 3 increases the utility of the relative distribution approach by allowing for comparisons to be made, standardizing for summary characteristics of the distributions. The relative and absolute sizes of the location and shapes effects can then be compared.

The software and data necessary for the applications of relative distribution methods, including those described here, are available at http://www.stat.washington.edu/handcock/RelDist. Additional information and applications of interest can also be found there.

PROOFS

A constructive proof of Theorem 1 can be based on the theory for U-statistics with estimated parameters and empirical process ideas. As K is twice differentiable, we can expand the estimator (6):

$$g_{n,m}(r) = \frac{1}{mh_m}\sum_{j=1}^{m} K\left(\frac{r - Q_j}{h_m}\right)$$

$$= \frac{1}{mh_m}\sum_{j=1}^{m} K\left(\frac{r - F_0(Y_j)}{h_m}\right)$$

$$+\frac{1}{mh_m}\sum_{j=1}^{m}\frac{F_0(Y_j)-F_{n0}(Y_j)}{h_m}K'\left(\frac{r-F_0(Y_j)}{h_m}\right)$$

$$+\frac{1}{mh_m}\sum_{j=1}^{m}\frac{(F_0(Y_j)-F_{n0}(Y_j))^2}{2h_m^2}K''\left(\Delta_j\right)$$

$$=g_m(r)+T_{n,m}(r)+R_{n,m}(r), \tag{A.1}$$

where Δ_j is between $h_m^{-1}(r-F_{n0}(Y_j))$ and $h_m^{-1}(r-F_0(Y_j))$. The last term is a remainder that is of smaller order in probability than the other terms. We prove this in the lemma below. The first term is the one-sample estimator (4), and the second represents the penalty for using the quasi-relative data in place of data from the exact relative distribution in the first term.

The second term 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

$$k_{h_m}(x,y;r)=-\left(\mathcal{I}\{x\le y\}-F_0(y)\right)\frac{1}{h_m^2}K'\left(\frac{r-F_0(y)}{h_m}\right),$$

which is dependent on m via h_m. Note that $[k_{h_m}(Y_0,Y;r)]=0$ and the projections are $\mathbb{E}[k_{h_m}(Y_0,y;r)]=0$ and

$$\begin{aligned}g_{1h_m}(x;r)&=\mathbb{E}[k_{h_m}(x,Y;r)]\\&=-\int_0^1\left(\mathcal{I}\{F_0(x)\le s\}-s\right)\frac{1}{h_m^2}K'\left(\frac{r-s}{h_m}\right)g(s)ds.\end{aligned}$$

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

$$T_{n,m}(r)=\frac{1}{n}\sum_{i=1}^{n}g_{1h_m}(Y_{0i};r)+o_p\left((nh_m)^{-\frac{1}{2}}\right) \tag{A.2}$$

as $m\to\infty, m/n\to\kappa^2<\infty$. Now,

$$\sqrt{mh_m}\frac{1}{n}\sum_{i=1}^{n} g_{1h_m}(Y_{0i};r)$$

$$= -\sqrt{\frac{m}{h_m^3}}\int_0^1 \left\{\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}\{F_0(Y_{0i}) \le t\} - t\right\} K'\left(\frac{r-t}{h_m}\right) g(t)dt$$

$$\stackrel{\mathcal{D}}{=} -\sqrt{\frac{m}{n}}\frac{1}{h_m^{3/2}}\int_0^1 U_n(t)K'\left(\frac{r-t}{h_m}\right) g(t)dt,$$

where $U_n(t)$ is the uniform empirical process (Shorack and Wellner 1986). We then have

$$\sqrt{\frac{m}{n}}\frac{1}{h_m^{3/2}}\int_0^1 U_n(t)K'\left(\frac{r-t}{h_m}\right) g(t)dt$$

$$= \sqrt{\frac{m}{n}}\frac{1}{h_m^{1/2}}\int_0^1 U_n(t)g(t)dK\left(\frac{r-t}{h_m}\right)$$

$$= \sqrt{\frac{m}{n}}\left[\frac{1}{h_m^{1/2}}K\left(\frac{r-t}{h_m}\right) U_n(t)g(t)\right]_0^1$$

$$-\sqrt{\frac{m}{nh_m}}\int_0^1 K\left(\frac{r-t}{h_m}\right)[g(t)U_n(dt) + U_n(t)g'(t)dt]$$

$$= -\sqrt{\frac{m}{n}}\frac{1}{h_m^{1/2}}\int_0^1 K\left(\frac{r-t}{h_m}\right) g(t)U_n(dt)$$

$$-\sqrt{\frac{m}{n}}\frac{1}{h_m^{1/2}}\int_0^1 K\left(\frac{r-t}{h_m}\right) U_n(t)g'(t)dt$$

$$= I^3_{n,h_m} + I^4_{n,h_m}.$$

The second term I^4_{n,h_m} converges in probability to zero:

$$|I^4_{n,h_m}| \le \sqrt{\frac{m}{n}}\sup_{0\le t\le 1}|U_n(t)|\sqrt{h_m}\int_0^1 \frac{1}{h_m}K\left(\frac{r-t}{h_m}\right)|g'(t)|dt = o_p(1)$$

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

$$\begin{aligned}\mathbb{V}[I^4_{n,h_m}] &= \frac{m}{n}\frac{1}{h_m}\int_0^1\int_0^1 K\left(\frac{r-t}{h_m}\right)K\left(\frac{r-u}{h_m}\right)\\ &\quad\times Cov(U_n(t),U_n(s))g'(s)g'(t)dsdt\\ &\quad\to h_m\kappa^2 r(1-r)[g'(r)]^2R^2(K)\end{aligned}$$

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 Corollary 3 to obtain an expression for the variance that is more accurate in small samples than the asymptotic approximation.

Pulling together these results, (A.2), the lemma, and (A.1), we obtain that

$$\begin{aligned}&\sqrt{mh_m}\left[g_{n,m}(r)-\tilde{g}_{n,m}(r)\right]\\ &=\sqrt{\frac{m}{n}}\frac{1}{h_m^{1/2}}\int_0^1 K\left(\frac{r-t}{h_m}\right)g(t)U_n(dt)\\ &+\frac{1}{h_m^{1/2}}\int_0^1 K\left(\frac{r-t}{h_m}\right)U_{1m}(dt)\\ &+o_p(1),\end{aligned}$$

where

$$\tilde{g}_{n,m}(r)=\frac{1}{h_m}\int_0^1 K\left(\frac{r-t}{h_m}\right)g(t)dt,$$

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 $\sqrt{mh_m}\left[\tilde{g}_{n,m}(r)-g(r)\right] \to 0$. Note that $\tilde{g}_{n,m}(r)$ has the same distribution as $g_m(r)$ in (4), 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

$$\sqrt{mh_m}\left[g_{n,m}(r)-\tilde{g}_{n,m}(r)\right]=\sqrt{\frac{m}{n}}\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left\{\frac{1}{h_m^{1/2}}K\left(\frac{r-U_i}{h_m}\right)g(U_i)\right.$$
$$\left.-\frac{1}{h_m^{1/2}}\mathbb{E}\left[K\left(\frac{r-U_i}{h_m}\right)g(U_i)\right]\right\}$$
$$+\frac{1}{m^{1/2}}\sum_{i=1}^{m}\left\{\frac{1}{h_m^{1/2}}K\left(\frac{r-Q_i}{h_m}\right)\right.$$
$$\left.-\frac{1}{h_m^{1/2}}\mathbb{E}\left[K\left(\frac{r-Q_i}{h_m}\right)\right]\right\}+o_p(1),$$

where $U_1,\ldots,U_n$ are i.i.d. uniform $[0,1]$ independent of the Q_is. The variance of the first term is then

$$\frac{m}{n}\int_0^1\frac{1}{h_m}K^2\left(\frac{r-t}{h_m}\right)g^2(t)dt-\frac{m}{n}h_m$$
$$\times\left[\int_0^1\frac{1}{h_m}K\left(\frac{r-t}{h_m}\right)g(t)dt\right]^2\ \rightarrow\ \kappa^2g^2(r)R(K)$$

as $m\rightarrow\infty, m/n\rightarrow\kappa^2<\infty$. The expression for the second term follows similarly.

Finally, we consider the third term in (A.1). The following lemma indicates that it does not affect the estimator asymptotically.

Lemma. $\sqrt{mh_m}\,R_{n,m}(r)\xrightarrow{P}0$ *as* $m\rightarrow\infty$.

Proof of the lemma. For simplicity, assume the support of K is contained in $[-1,1]$. We can bound $R_{n,m}(r)$ as

$$|R_{n,m}(r)|\leq\frac{1}{mh_m}\sum_{j=1}^{m}\frac{(F_0(Y_j)-F_n(Y_j))^2}{2h_m^2}|K''\left(\Delta_j\right)|$$

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 $0\leq\theta_j\leq1$. Note that for $\Delta_j\notin[-1,1]$, we have $K''\left(\Delta_j\right)=0$. 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,

$$-1 \leq \Delta_j \leq 1 \Leftrightarrow r - h_m \leq F_0(Y_j) + \theta_j(F_0(Y_j) - F_{n0}(Y_j)) \leq r + h_m.$$

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

$$\begin{aligned}
|R_{n,m}(r)| &\leq C(K'')\frac{\Delta_{n0}^2}{h_m^3} \cdot \frac{1}{m} \\
&\quad \times \sum_{j=1}^{m} \mathcal{I}\{r - h_m - \Delta_{n0} \leq F_0(Y_j) \leq r + h_m + \Delta_{n0}\} \\
&= C(K'')\frac{\Delta_{n0}^2}{h_m^3} \cdot \left\{ G_m\left(r + h_m + \Delta_{n0}\right) - G_m\left(r - h_m - \Delta_{n0}\right) \right\} \\
&= C(K'')\frac{\Delta_{n0}^2}{h_m^3} \cdot \left\{ \left[G_m\left(r + h_m + \Delta_{n0}\right) - G_m\left(r - h_m - \Delta_{n0}\right) \right] \right. \\
&\quad \left. - \left[G\left(r + h_m + \Delta_{n0}\right) - G\left(r - h_m - \Delta_{n0}\right) \right] \right\} \\
&\quad + C(K'')\frac{\Delta_{n0}^2}{h_m^3} \cdot \left[G\left(r + h_m + \Delta_{n0}\right) - G\left(r - h_m - \Delta_{n0}\right) \right] \\
&= I_{n,m}^1 + I_{n,m}^2,
\end{aligned}$$

where $G(s) = \frac{1}{m}\sum_{j=1}^{m} \mathcal{I}\{F_0(Y_j) \leq s\}$. As G is Lipschitz, we have

$$|G\left(r + h_m + \Delta_{n0}\right) - G\left(r - h_m - \Delta_{n0}\right)| \leq 2L_G(h_m + \Delta_{n0})$$

and

$$\sqrt{mh_m}I^2_{n,m} \leq 2L_G C(K'')\frac{\Delta^2_{n0}}{h^3_m}\sqrt{mh_m}\left(h_m + \Delta_{n0}\right) = o_p(1)$$

as $\Delta_{n0} = O_p(n^{-\frac{1}{2}})$ and $mh^3_m \to \infty$. We now consider $I^1_{n,m}$:

$$|I^1_{n,m}| \leq 2C(K'')\frac{\Delta^2_{n0}}{h^3_m} \times \sup_{|t| \leq h_m + \Delta_{n0}} \left|\left[G_m\left(r+t\right) - G_m\left(r\right)\right] - \left[G\left(r+t\right) - G\left(r\right)\right]\right|$$

The Dvoretzky, Kiefer, and Wolfowitz (1956) bound for the tails of Δ_{n0} yields that for any given $\epsilon > 0$, there exists some finite C such that $\Delta_{n0} \leq Cn^{-\frac{1}{2}}$ up to an event with probability less than or equal to ϵ. The inequality $|t| \leq h_m + \Delta_{n0}$ on this set means that $|t| \leq C_1 h_m$ for some constant C_1. Using (2.13) in Stute (1982), we see that

$$\sup_{|t| \leq C_1 h_m} \left|\left[G_m\left(r+t\right) - G_m\left(r\right)\right] - \left[G\left(r+t\right) - G\left(r\right)\right]\right| = O_p\left(\sqrt{\frac{-h_m \log h_m}{m}}\right).$$

Therefore, as $mh^3_m \to \infty$,

$$\sqrt{mh_m}I^1_{n,m} = o_p(1).$$

The lemma follows as $\epsilon > 0$ is arbitrary.

REFERENCES

Begg, Colin B. 1991. "Advances in Statistical Methodology for Diagnostic Medicine in the 1980's." *Statistics in Medicine* 10:1887-95.

Bernhardt, Annette D., Martina Morris, and Mark S. Handcock. 1995. "Women's Gains or Men's Losses? A Closer Look at the Shrinking Gender Gap in Earnings." *American Journal of Sociology* 101:302-28.

Butler, Richard J. and James B. McDonald. 1987. "Interdistributional Income Inequality." *Journal of Business and Economic Statistics* 5:13-18.

Campbell, Gregory. 1994. "Advances in Statistical Methodology for Evaluation of Diagnostic and Laboratory Tests." *Statistics in Medicine* 13:499-508.

Cwik, Jan and Jan Mielniczuk. 1989. "Estimating Density Ratios With Application to Discriminant Analysis." *Communications in Statistics* 18:3057-69.

———. 1990. "Some Topics in Estimation of Neyman-Pearson and Performance Curves." Pp. 114-29 in *Cosmex*, edited by W. Kasprzak and A. Weron. Singapore: World Scientific.

———. 1993. "Data-Dependent Bandwidth Choice for a Grade Density Kernel Estimate." *Statistics and Probability Letters* 16:397-405.

Dvoretzky, Aryeh, Jack Kiefer, and Jack Wolfowitz. 1956. "Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator." *Annals of Mathematical Statistics* 27:642-69.

Engel, Joachim and Alois Kneip. 1999. "Recent Approaches to Estimating Income Distributions, Engel Curves and Related Functions." Discussion paper, University of Bonn.

Eubank, Randy L., Victor N. LaRiccia, and Rebecca B. Rosenstein. 1987. "Test Statistics Derived as Components of Pearson's Phi-Squared Distance Measure. *Journal of the American Statistical Association* 82:816-25.

Gastwirth, John L. 1968. "The First-Median Test: A Two-Sided Version of the Control Median Test." *Journal of the American Statistical Association* 63:692-706.

Handcock, Mark S. and Paul Janssen. 1998. "Statistical Properties of the Relative Distribution and Relative Density." Technical report, Pennsylvania State University, Department of Statistics, University Park, PA.

Handcock, Mark S. and Martina Morris. 1998. "Relative Distribution Methods." *Sociological Methodology* 28:53-97.

———. 1999. *Relative Distribution Methods in the Social Sciences*. New York: Springer-Verlag.

Härdle, Wolfgang. 1990. *Applied Nonparametric Regression*. Cambridge, UK: Cambridge University Press.

Hoeffding, Wassily. 1948. "A Class of Statistics With Asymptotically Normal Distribution." *Annals of Mathematical Statistics* 19:293-325.

Holmgren, Eric B. 1995. "The P-P Plot as a Method for Comparing Treatment Effects." *Journal of the American Statistical Association* 90:360-65.

Hsieh, Fushing. 1995. "The Empirical Process Approach for Semiparametric Two-Sample Models With Heterogeneous Treatment Effect." *Journal of the Royal Statistical Society, Series B* 57:735-48.

Jammalamadaka, Sreenivasa R. and Svante Janson. 1986. "Limit Theorems for a Triangular Scheme of *U*-Statistics With Applications to Inter-Point Distances." *Annals of Probability* 14:1347-58.

Karoly, Lynne A. 1993. "The Trend in Inequality Among Families, Individuals, and Workers in the United States: A Twenty-Five Year Perspective." Pp. 19-97 in *Uneven Tides: Rising Inequality in America*, edited by S. Danziger and P. Gottschalk. New York: Russell Sage.

Li, Gang, Ram C. Tiwari, and Martin T. Wells. 1996. "Quantile Comparison Functions in Two-Sample Problems, With Application to Comparisons of Diagnostic Markers." *Journal of the American Statistical Association* 91:689-98.

Loader, Clive. 1999. *Local Regression and Likelihood*. New York: Springer-Verlag.

Morris, Martina, Annette D. Bernhardt, and Mark S. Handcock. 1994. "Economic Inequality: New Methods for New Trends." *American Sociological Review* 59:205-19.

Morris, Martina and Bruce Western. 1999. "Inequality in Earnings at the Close of the Twentieth Century." *Annual Review of Sociology* 25:627-57.

Morris, Martina and Mark S. Handcock. 1999. *Relative Distribution Methods in the Social Sciences*. New York: Springer-Verlag.

Parzen, Emmanuel. 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.

———. 1992. "Comparison Change Analysis." Pp. 3-15 in *Nonparametric Statistics and Related Topics*, edited by A. Saleh. Holland: Elsevier.

Serfling, Robert J. 1980. *Approximation Theorems in Mathematical Statistics*. New York: John Wiley.

Shorack, Galen R. and Jon A. Wellner. 1986. *Empirical Processes With Applications to Statistics*. New York: John Wiley.

Silverman, Bernard W. 1986. *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall.

Simonoff, Jeffrey S. 1998. "Three Sides of Smoothing: Categorical Data Smoothing, Nonparametric Regression, and Density Estimation." *International Statistical Review* 66:137-56.

Stute, Winfried. 1982. "The Oscillation Behavior of Empirical Processes." *Annals of Probability* 10:86-107.

Wand, Matt and M. Chris Jones. 1995. *Kernel Smoothing*. London: Chapman & Hall.

Wilk, Martin B. and Ram Gnanadesikan. 1968. "Probability Plotting Methods for the Analysis of Data." *Biometrika* 55:1-17.

Mark S. Handcock is professor of statistics and sociology at the University of Washington. His research involves methodological development, and is based largely on motivation from questions in the social sciences. In this area he focuses on the development of statistical models for the analysis of longitudinal data on job mobility, wage trajectories and distributional changes in earnings. He also works in the fields of random graph modeling, spatial statistics and inference for stochastic processes.

Paul L. Janssen is a full professor at the Center for Statistics, Limburgs Universitair Centrum, Diepenbeek, Belgium. His current research interests are nonparametric estimation, resampling methodology, and statistical modeling in survival analysis.