An Introduction to Relative Distribution Methods

by

Mark S. Handcock

Professor of Statistics and Sociology
Center for Statistics and the Social Sciences

CSDE Seminar Series

March 2, 2001

In collaboration with:

Martina Morris

0.0 Introduction and Motivations

Motivating issue:

To compare the economic status of groups across the full distribution

- relative to each other
- over time

Tracking the earnings of men and women over time.

Focus on yearly earnings of

- individual full–time year–round workers, 16–65 years old, not in school, the military or farming.
- non-Hispanic white men and women, non-Hispanic black women.
- Consider 1967 to 1987 (i.e, 21 years of data)
- Based on the U.S. Current Population Survey (annual March Supplement).
- Sample sizes are large 20,000 per year for males, 10,000 per year for females.

How can/should we compare the earnings?

One approach is to consider a summary measure of level (e.g., mean, median)

- What can we learn from such summaries?
  - where the "centers" of the distributions comparatively lie.

- What is lost from such summaries?
  - How are the men's earnings distributed about their median value?
  - How are the women's earnings distributed about their median value?
  - Are they distributed in a similar fashion?

- We do not know if substantively important differences between (the distributions of) men's and women's earnings are adequately summarized.

- What other differences can matter?

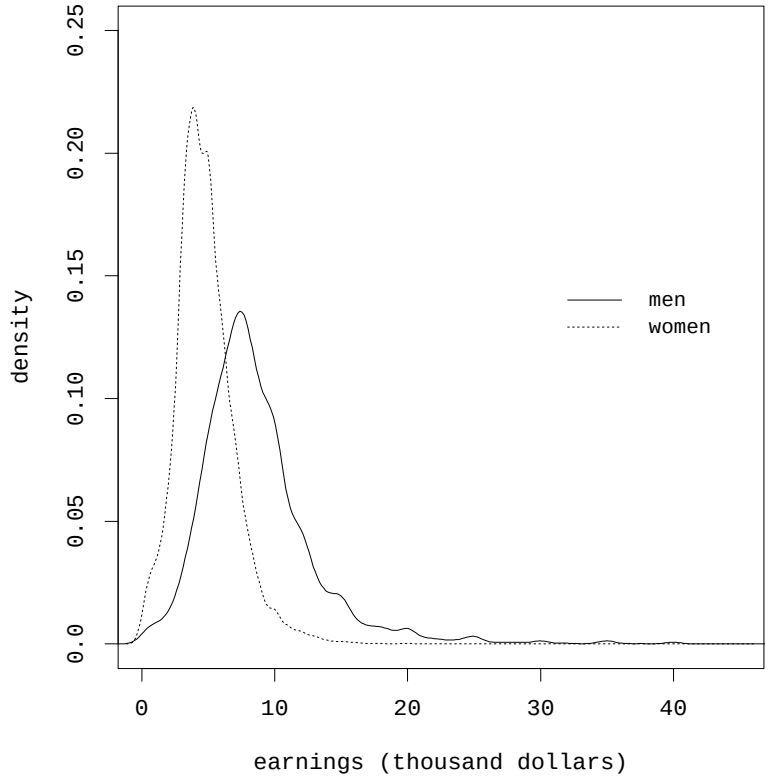


Fig. 1. The distributions of earnings in 1967.

- Spread – variation about the median within each group.
- Quantiles – proportions above or below fixed values (e.g., poverty line).
- Shape – the full distributional features of each curve.

- We require a framework that captures the differences between distributions:

  - in a parsimonious way

  - flexible representation of patterns of differences in the data that are not preconceived.

  - builds on easily interpretable numerical summaries such as location and spread.

Idea: Rescale the women's earnings *relative* to the men's so that both are on the same interpretable scale.

e.g. Break up the men's distribution into deciles:

0 - 4420

4421 - 5500

5501 - 6500

. . .

13500 -

What proportion of women fall within each of these groups:

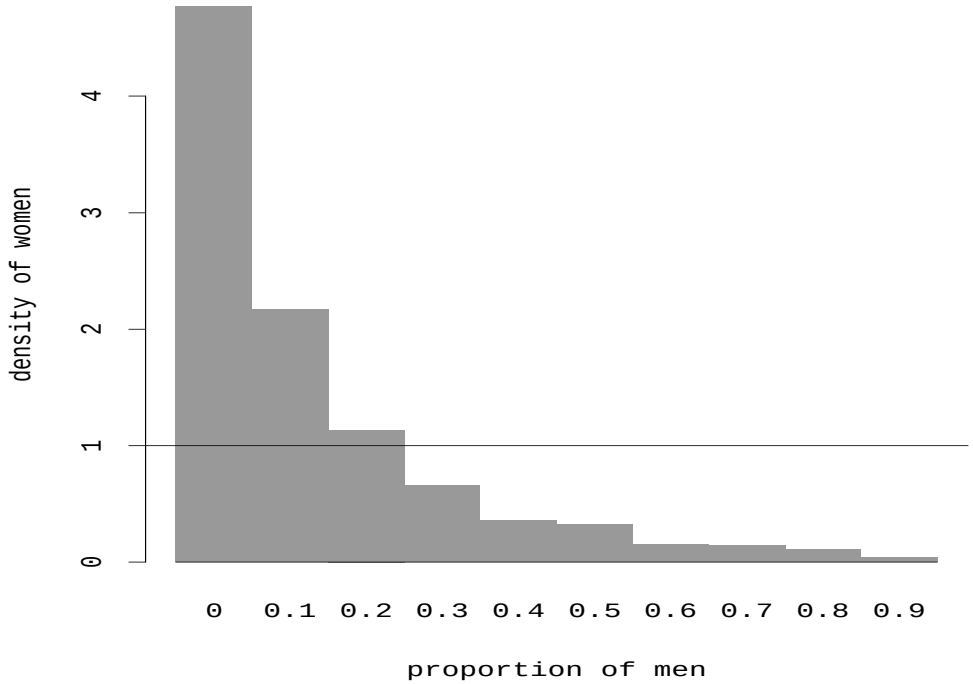


Fig. 2. The relative decile distribution of women's to men's earnings in 1967.

We can see how frequent the women are for groups of men of equal size.

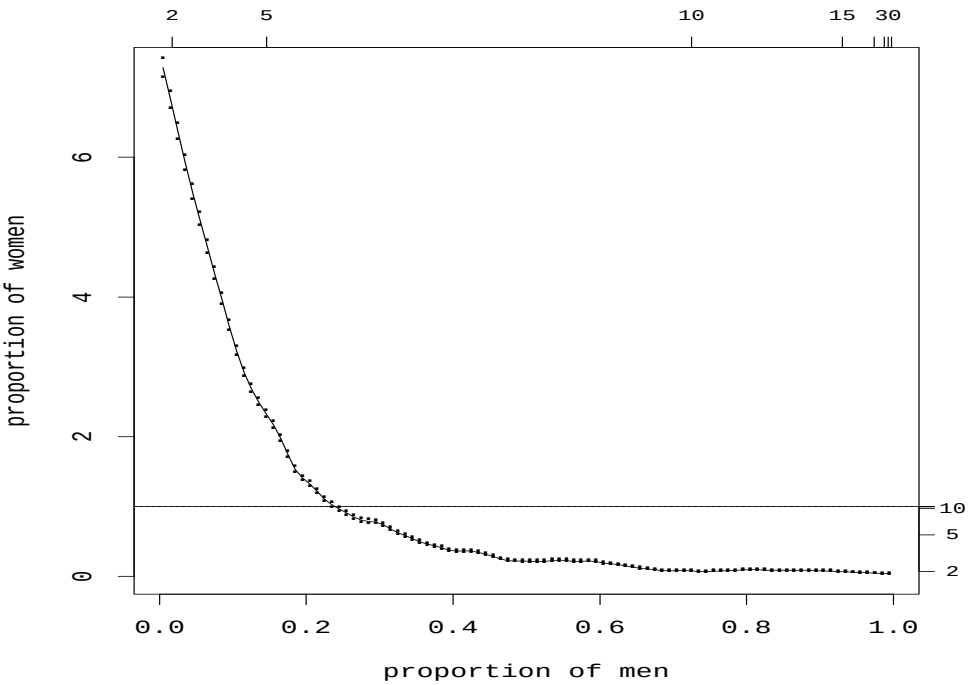


Fig. 3. The relative distribution of women's to men's earnings in 1967.

1.0 Technical Stuff: The Relative Distribution

- Does the idea have depth?

Y_0 a measurement for a reference population

$F_0(y)$ the CDF and $f_0(y)$ the density

Y the measurement on a comparison population

$F(y)$ CDF and density $f(y)$

The objective is to study the differences between the distributions of Y and Y_0 using Y_0 as the reference.

Consider the *grade transformation* of Y to Y_0 :

$$R = F_0(Y)$$

Cwik and Mielniczuk (1989)

Refer to R as the *relative distribution* of Y to Y_0

The cumulative distribution function (CDF) of R is

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

The corresponding density,

$$g(r) = \frac{f\Big(F_0^{-1}(r)\Big)}{f_0\Big(F_0^{-1}(r)\Big)} \qquad\qquad 0 \le r \le 1$$

where r represents the proportion of values
f and f_0 are the densities.

- Interpretations

$G(r)$ the *relative CDF*: a proportion $G(r)$ of the target population are below the level of a proportion r of the reference population

$g(r)$ the *relative density*, represents the ratio of the frequency of the target population to the frequency of the reference population at the r^{th} quantile of the reference population level $[\, F_0^{-1}(r)\,]$

- The relative distribution focuses directly on the comparison rather than individual distributions.

- The scale is in terms of the ranks rather than levels: we count "people" rather than "dollars".

- The relative distribution is invariant to monotone transformation of each of the variable (e.g., wages vs. log–wages).

- If the two distributions are identical then the relative distribution is uniform on $[0, 1]$.

- Many indices and measures can be defined based on $g(p)$ alone, emphasizing interesting aspects of their relative shape.

  For example, Kullback–Leibler information number:

$$KL(Y, Y_0) = \int_0^1 \log[g(p)]g(p)dp$$

χ^2 divergence:

$$\chi^2(Y, Y_0) = \int_0^1 [g(p) - 1]^2 dp$$

L_1 distance:

$$Ł_1(Y, Y_0) \equiv \int |f_0(x) - f(x)|f_0(x)dx = \int_0^1 |g(p) - 1|dp$$

1.1 Application: Changes in Men's Earnings 1975–93

Consider distributions of real hourly wages for white men between 1975 and 1993 (based on CPS March Supplement).

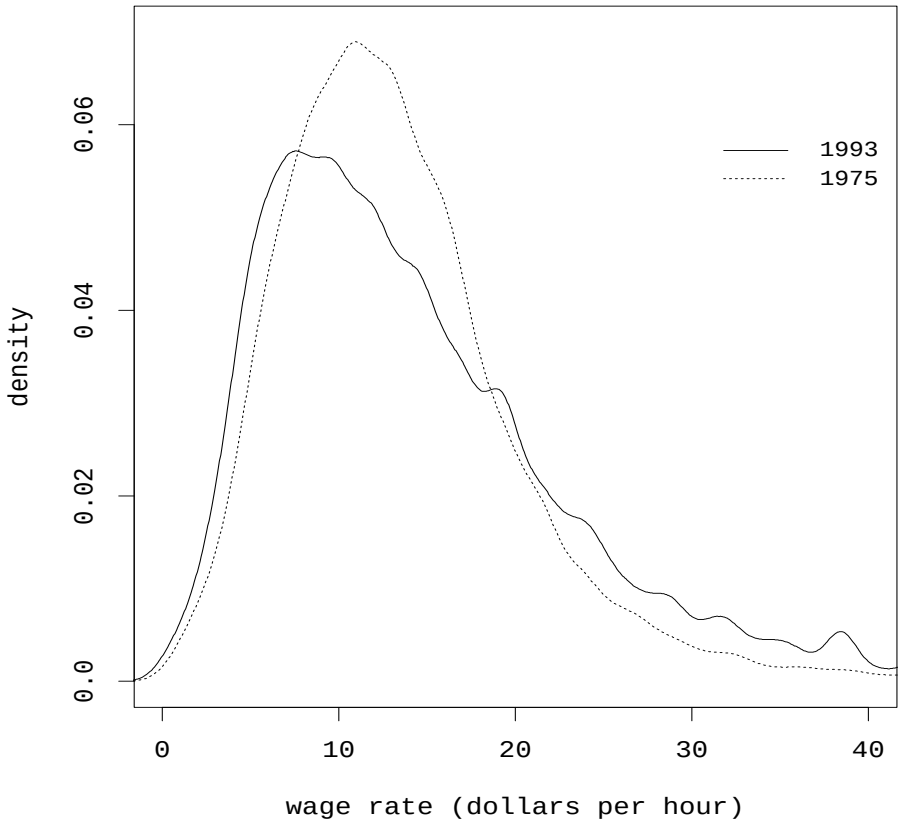


Fig. 4. The distributions of real hourly wages in 1975 and 1993 expressed in 1993 dollars.

- To explain the rise in wage inequality, researchers have started to look at the restructuring and reorganization of the American firm and its effects on workers (Cappelli 1994, Harrison 1994).

  - a dramatic increase in the use of (so called "contingent" workers) vs. (so called "core" workers).

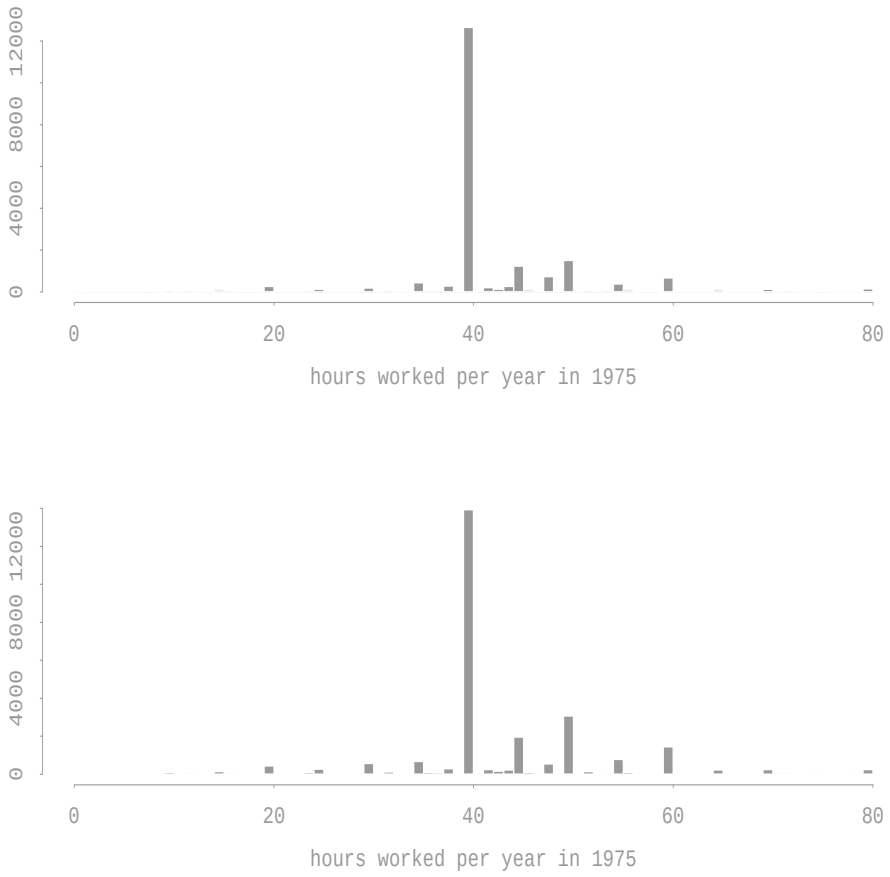
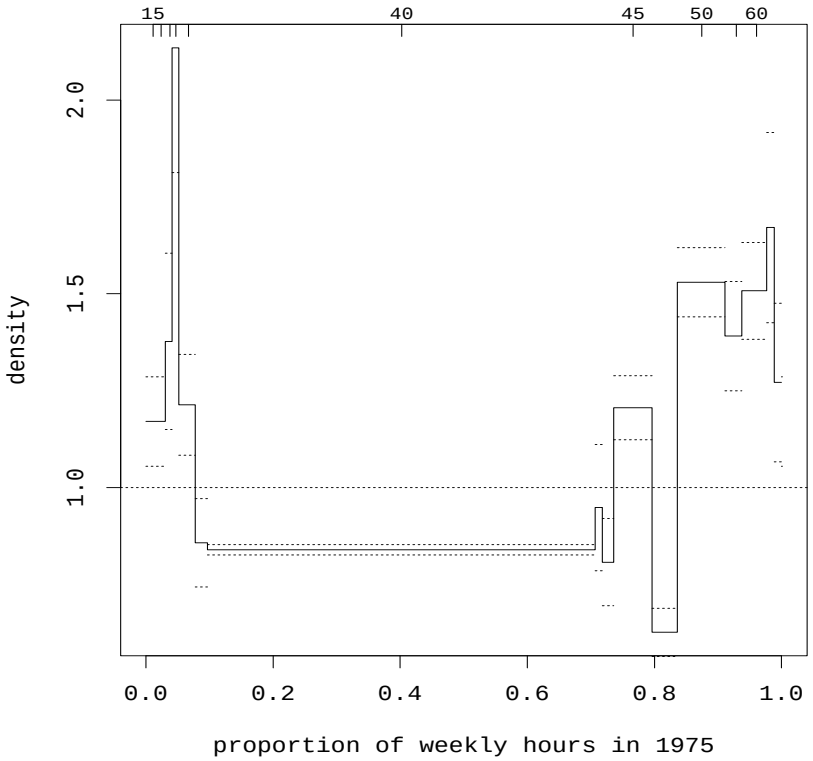


Figure 5: The distributions of usual weekly hours worked for 1975 and 1993.

The Relative Distribution of Work Schedules



- The polarization in work schedules is quite apparent.

- Could this explain the growth in the dispersion of hourly wages?

2.0 Adjusting the Distributions for Differences in Covariates

Suppose the two groups differ in terms of a covariate Z

For example, we can adjust the relative distribution of wages for the changing distribution of work schedules.

(Y_0, Z_0) reference wages and work schedules

(Y, Z) comparison wages and work schedules

Idea: Construct a "synthetic population" for the reference group that has the same composition of the covariate as the comparison wage.

For example, what would the 1975 wages have looked like if the 1975 group had as many working the same hours as the 1993 group?

The marginal density of Y_0 is

$$f_0(y) \equiv f_{Y_0}(y) = \int f_{Y_0|z_0}(y \mid z)\ f_{Z_0}(z)\ dz.$$

Let Y_A be the random variable that gives the measurement Y_0 for the (hypothetical) reference population with

- a marginal distribution of Z is the same as in the comparison population and

- the conditional distributions of the measurement given Z (i.e., $f_{Y_0|Z_0}(y\,|z)\,$) the same as in the reference population.

The density of Y_A can be written as:

$$f_A(y) = \int f_{Y_0|Z_0}(y\,|z)\; f_Z(z)\; dz.$$

Call Y_A the random variable describing

Y_0 *compositionally adjusted to* Z

- We can now determine the compositional effects of the covariate by comparing Y_m to Y_0.

  Let $X_{m,0} = F_0(Y_m)$ be the relative distribution of Y_m to Y_0.
  - describes the differences between Y_0 and Y due to compositional effects.

  Let $X_{1,m} = F_m(Y)$ be the relative distribution of Y to Y_m.
  - describes the differences between Y and Y_0 not due to the compositional differences.

- Let $X \equiv X_{1,0} = F_0(Y)$ be the relative distribution of Y to Y_0.

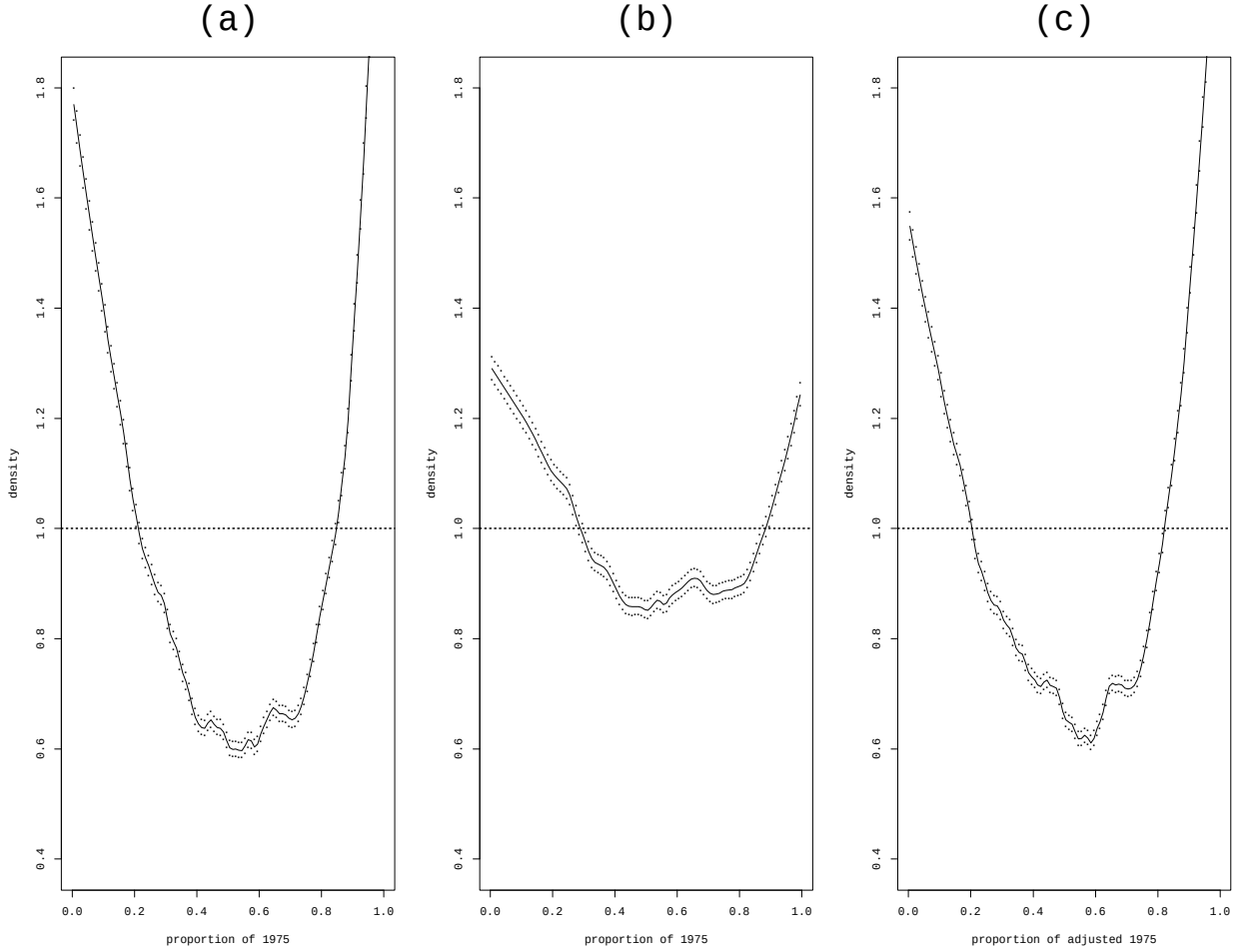
- $X_{1,m}$ is the relative distribution of $X_{1,0}$ to $X_{m,0}$.

- Heuristically, we can represent the decomposition of the relative densities by:

$$\begin{array}{c}\text{overall relative}\\ \text{density}\end{array} = \begin{array}{c}\text{relative density representing}\\ \text{compositional effect of the covariate}\end{array} \times \begin{array}{c}\text{covariate–adjusted}\\ \text{relative density}\end{array}$$

- 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.

2.1 Decomposition of Changes in Wages due to Changes in Work Schedules



- The polarization in work schedules had a modest polarizing effect on the wages.

- In terms of the magnitude of the effect, this shift did not drive the majority of the rising inequality in wages

3.0 Conclusions

- Relative distribution ideas represent a general framework for the comparative analysis of distributional difference and change.
  - can be used as the basis for exploratory, descriptive and analytical techniques.

- Summary measures based on the relative density can be used to test hypotheses about distributional differences.

- Decomposition techniques enable one to isolate location, shape and compositional effects.
  - enables one to distinguish the impact of changes in population mix (a demographic process) from changes in attribute allocation (a social or economic process).