The Stata Journal (2021)
21, Number 4, pp. 885–951 DOI: 10.1177/1536867X211063147

Relative distribution analysis in Stata

Ben Jann
University of Bern
Bern, Switzerland
ben.jann@soz.unibe.ch

Abstract. In this article, I discuss the method of relative distribution analysis and present Stata software implementing various elements of the methodology. The relative distribution is the distribution of the relative ranks that the outcomes from one distribution take on in another distribution. The methodology can be used, for example, to compare the distribution of wages between men and women. The presented software, `reldist`, estimates the relative cumulative distribution and the relative density, as well as the relative polarization, divergence, and other summary measures of the relative ranks. It also provides functionality such as location and shape decompositions or covariate balancing. Statistical inference is implemented in terms of influence functions and supports estimation for complex samples.

Keywords: st0656, reldist, relative distribution, relative ranks, relative density, median relative polarization, divergence, location and shape decomposition, covariate balancing, Gastwirth index, reweighting, influence function

1 Introduction

Although earlier work on relative distributions and related approaches can be found in the statistical literature (for example, Ćwik and Mielniczuk [1989, 1993]), the methodology has not been popular in applied work before Mark S. Handcock, Martina Morris, and coauthors introduced it to the social sciences in some influential applied (Morris, Bernhardt, and Handcock 1994; Bernhardt, Morris, and Handcock 1995; and Bernhardt et al. 2001) and methodological contributions (Handcock and Morris 1998, 1999; Handcock and Janssen 2002) in the mid 1990s and early 2000s. Even today, however, relative distribution methods do not seem to experience very widespread use, which might partly be because of lack of user-friendly statistical software supporting such analyses (apart from an R package by Handcock and Aldrich [2002]; see Handcock [2016]).

Nevertheless, I believe that relative distribution analysis is a valuable complement to other approaches for distributional comparisons, which typically look at differences in (counterfactual) density, distribution, or quantile functions (for example, DiNardo, Fortin, and Lemieux [1996] and Chernozhukov, Fernández-Val, and Melly [2013]). A key feature of relative distribution analysis is that it focuses on positions within distributions rather than on absolute outcome values. The methodology can be used, for example, to study how wage distributions differ by gender or ethnic groups or how income polarization changed over time. A few examples from the literature illustrate the scope of potential applications: Alderson, Beckfield, and Nielsen (2005) studied

changes in income inequality in several countries; Bliege Bird et al. (2008) analyzed the anthropogenic influence on vegetational diversity in Australia; Del Giudice (2011) looked at gender differences in adult romantic attachment; Eggers and Spirling (2016) studied cohesive party voting in the British House of Commons between 1836 and 1910; Clementi, Molini, and Schettino (2018) analyzed changes in the consumption distribution over time in Ghana; and Panek and Zwierzchowski (2020) studied changes in household income polarization in Poland.

In an attempt to improve the accessibility of the methodology to applied researchers, I provide an overview of relative distribution methods in this article, and I present software that makes the methodology available in Stata. The software, called `reldist`, can be used to estimate and plot the relative density function (relative PDF), a histogram of the relative distribution, or the relative distribution function (relative CDF). Furthermore, it computes relative polarization indices and distributional divergence measures, as well as descriptive statistics of the relative data, and it supports the decomposition of the relative distribution by adjusting for location, scale, and shape differences or by adjusting for differences in covariate distributions. Estimation of standard errors and confidence intervals is provided for all quantities, including support for complex samples. I tried to make the software as versatile as possible while also maintaining user friendliness, for example, by following official Stata standards in terms of syntax, output, and stored results.

The article is structured as follows. In the next section, I give an overview of the main concepts of relative distribution analysis, including definitions of relative ranks and the relative distribution, as well as elements such as location and shape decompositions, distributional divergence and relative polarization summary measures, and covariate adjustment approaches. Most of the discussed material is also covered in Handcock and Morris (1999), but I focus on elements I consider most relevant from an applied perspective, and I use a somewhat different notation. Furthermore, I introduce reweighting as an additional strategy for covariate adjustment. In section 3, I then discuss the computational details involved in the estimation of the quantities presented in section 2. I cover different variants of how to compute relative ranks, the relative cumulative distribution, the relative density, the relative histogram, summary measures, and covariate balancing, and I distinguish between continuous and categorical outcomes when relevant. Again, many of the relevant issues are also addressed by Handcock and Morris (1999), but my exposition is more focused on specific implementation. Section 4 then introduces the software and its options, and section 5 provides several worked examples.

The article further contains an appendix covering the estimation of sampling variances by means of influence functions (IFs). The appendix is rather technical and can safely be ignored by readers who are only interested in the practical application of the methods; it is not needed for obtaining an understanding of relative distribution methods and for being able to correctly apply the software and interpret the results. Nonetheless, I consider the appendix an important and original contribution providing results that cannot be found elsewhere in the literature. I first illustrate how IFs can be obtained by analogy to the method of moments and then derive specific expressions for

all relative distribution quantities of interest, including possible covariate adjustment. One virtue of an IF-based approach is that it leads to expressions that are compatible with complex survey estimation.

2 Theory

In this section, I summarize the main statistical concepts that are relevant for relative distribution analysis. For an in-depth treatment of the topic, see Handcock and Morris (1999). For a more recent introduction, also see chapter 5 in Hao and Naiman (2010).

2.1 CDF and density

Let Y be a continuous outcome variable of interest. Y is assumed to be a random variable with CDF

$$F_Y(y) = P(Y \leq y), \quad y \in \mathbb{R}$$

That is, for any value y, the CDF provides the probability of Y taking on a value that is smaller than or equal to y. The PDF of Y is then defined as the first derivative of the CDF, that is,

$$f_Y(y) = F'_Y(y) = \frac{\mathrm{d}F_Y(y)}{\mathrm{d}y}$$

Hence, the integral of the density from $-\infty$ to y is equal to the value of the CDF at value y:

$$F_Y(y) = \int_{-\infty}^{y} f_Y(t)\,\mathrm{d}t$$

Likewise, the integral of the density between values a and b provides the probability that Y falls into interval $(a, b]$:

$$P(a < Y \leq b) = F_Y(b) - F_Y(a) = \int_a^b f_Y(y)\,\mathrm{d}y$$

Finally, let $q_Y(p) = F_Y^{-1}(p)$ be the inverse of F_Y, that is, the quantile function of Y, such that

$$y = q_Y\{F_Y(y)\} = F_Y^{-1}\{F_Y(y)\}$$

2.2 Relative ranks

Define

$$r_Y(y) = F_Y(y)$$

as the "relative rank" of outcome y in distribution F_Y. Because F_Y is a CDF, r lies between 0 and 1. Handcock and Morris (1999) call r the "relative data", and Ćwik and Mielniczuk (1989) speak of the "grade transformation".

Relative ranks have a distribution themselves that depends on the distribution of the y values at which $r_Y(y)$ is evaluated. For example, if the y values are distributed according to F_Y, then r has a uniform distribution.

2.3 The relative distribution

Let F_X be a comparison distribution and F_Y be a reference distribution. In relative distribution analysis, we are interested in how F_X is distributed relative to F_Y. The relative CDF of F_X with respect to F_Y is defined as the distribution of the relative ranks that outcome values distributed according to F_X take on in distribution F_Y. That is, we are interested in the distribution of $r_Y(y)$ for y values distributed according to F_X, which can be obtained by inverting r to y using F_Y^{-1} and then applying F_X. Hence, the relative CDF is given as

$$G(r) = F_X\{F_Y^{-1}(r)\}, \quad r \in [0,1] \tag{1}$$

Stated differently, for each value of $r = F_Y(y)$, the relative CDF obtains the corresponding value of $F_X(y)$, keeping y fixed, which leads to the tuples

$$\{F_X(y), F_Y(y)\}, \quad y \in \mathbb{R}$$

Plotted in a diagram with r $[= F_Y(y)]$ on the horizontal axis and $G(r)$ $[= F_X(y)]$ on the vertical axis, all points will lie on the diagonal if the two distributions are identical [that is, $G(r) = r$ in this case, as can easily be seen in (1)].[1] If the outcome values in the comparison distribution tend to be lower than the outcome values in the reference distribution, the points will lie above the diagonal (and vice versa). The relative distribution might also cross the diagonal, for example, if one of the distributions is more polarized than the other. Figure 1 provides an illustration. On the left, three examples of the density functions of two distributions are shown. In the middle panel, the corresponding relative distribution functions are displayed.

1. The diagram of F_X by F_Y is also known as "probability–probability plot" (P–P plot; for a Stata implementation, see Cox [2004]).

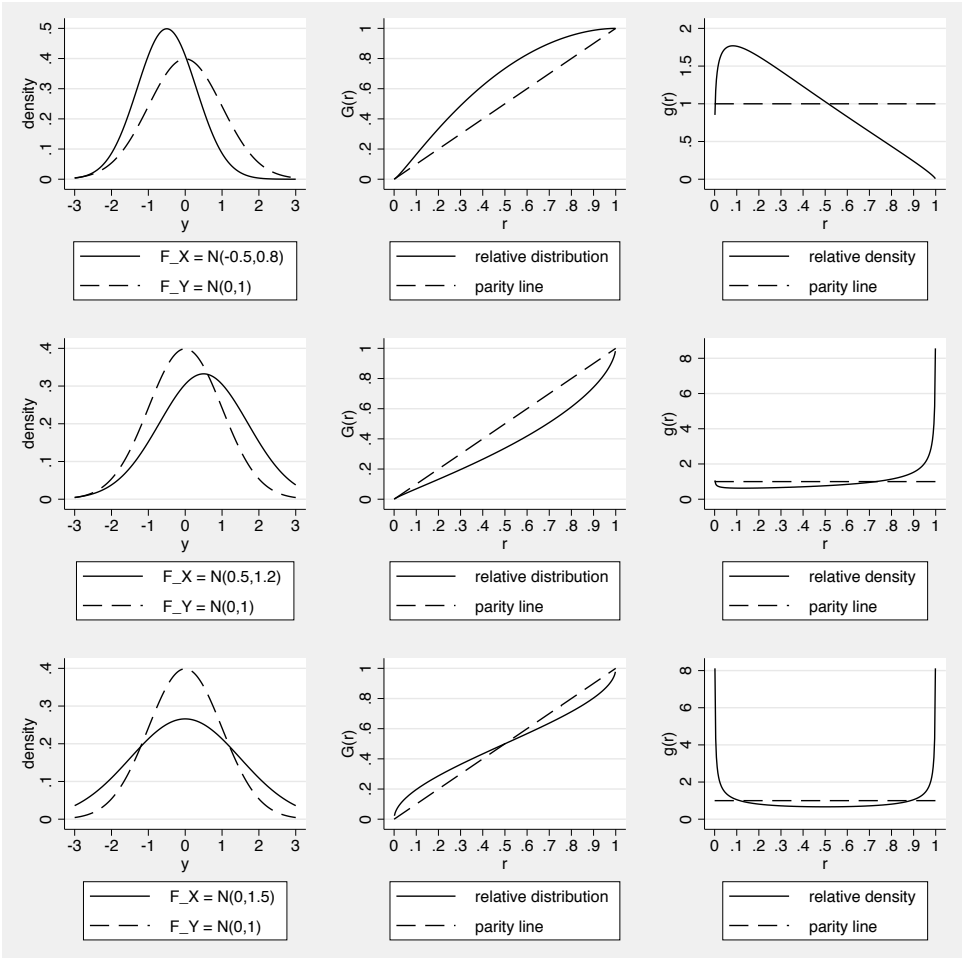

Figure 1. Illustration of the relative distribution

2.4 The relative density

Because $G(r)$ is a CDF, we can take the first derivative to obtain the density. Employing the chain rule, the relative PDF of F_X with respect to F_Y can be written as

$$g(r) = \frac{\mathrm{d}G(r)}{\mathrm{d}r} = \frac{f_X\{F_Y^{-1}(r)\}}{f_Y\{F_Y^{-1}(r)\}}, \quad r \in [0,1] \tag{2}$$

As can be seen, the relative density is equal to the ratio of the densities of the two distributions at a specific y value [that is, $g(r)$ is equal to the ratio of the two densities at the y value equal to quantile r of F_Y]. Nonetheless, $g(r)$ is a proper PDF because it is positive and integrates to 1.

If the two compared distributions are identical, $g(r)$ will be equal to 1 for all r, as is easy to see in (2). If the comparison distribution tends to have lower values than the reference distribution, the relative density will be larger than 1 at low values of r and smaller than 1 for large r (and vice versa). Likewise, assuming similar locations of the two distributions, if the comparison distribution is more polarized than the reference distribution, the relative density will be larger than 1 at small and large values of r and below 1 in between (and vice versa). An illustration of different situations is provided in the right panel of figure 1.

2.5 Location and shape decomposition

Distributions can have different "locations," meaning that they differ, say, in their mean or median. If a large location difference exists, the relative CDF and PDF will be dominated by this difference. In many applications, it may thus be informative to distinguish between a "location effect" and the difference in distributional shape, net of location.

As shown by Handcock and Morris (1999), the overall relative density can be decomposed into a "location effect" and a "shape effect" by constructing a location-adjusted distribution and then using this counterfactual distribution in place of either F_X or F_Y. For example, let

$$\widetilde{Y} = Y - \mu_Y + \mu_X \tag{3}$$

be a location-adjusted variant of Y, where μ is a location measure such as the median or the mean. In general, if $\widetilde{Y} = t(Y)$, the distribution of $\widetilde{Y}$ is equal to $F_Y\{t^{-1}(y)\}$. This means that

$$F_{\widetilde{Y}}(y) = P(Y - \mu_Y + \mu_X \leq y) = P(Y \leq y + \mu_Y - \mu_X) = F_Y(y + \mu_Y - \mu_X)$$

is a location-adjusted reference distribution that has the same location as the comparison distribution. The overall relative density can then be written as

$$g(r) = \frac{f_X\{F_Y^{-1}(r)\}}{f_Y\{F_Y^{-1}(r)\}} = \underbrace{\frac{f_{\widetilde{Y}}\{F_Y^{-1}(r)\}}{f_Y\{F_Y^{-1}(r)\}}}_{\text{location effect}} \times \underbrace{\frac{f_X\{F_Y^{-1}(r)\}}{f_{\widetilde{Y}}\{F_Y^{-1}(r)\}}}_{\text{shape effect}} \tag{4}$$

The first factor, the location effect, is equal to the ratio between the density of the location-adjusted reference distribution and the unadjusted reference distribution. The second factor, the shape effect, is the ratio between the density of the (unadjusted) comparison distribution and the location-adjusted reference distribution. However, note that

$$\frac{f_X\{F_Y^{-1}(r)\}}{f_{\widetilde{Y}}\{F_Y^{-1}(r)\}}, \quad r \in [0,1]$$

is not a proper density, because it is evaluated over y values distributed according to F_Y instead of $F_{\widetilde{Y}}$. It may therefore be more useful to characterize the shape effect by the adjusted relative PDF

$$g_{X\widetilde{Y}}(r) = \frac{f_X\{F_{\widetilde{Y}}^{-1}(r)\}}{f_{\widetilde{Y}}\{F_{\widetilde{Y}}^{-1}(r)\}}$$

or the corresponding adjusted relative CDF

$$G_{X\widetilde{Y}}(r) = F_X\{F_{\widetilde{Y}}^{-1}(r)\}$$

Instead of adjusting F_Y, the decomposition could also be defined by adjusting the comparison distribution. That is, we could use

$$\widetilde{X} = X - \mu_X + \mu_Y \qquad \text{with} \qquad F_{\widetilde{X}}(y) = F_X(y + \mu_X - \mu_Y)$$

such that

$$g(r) = \frac{f_X\{F_Y^{-1}(r)\}}{f_Y\{F_Y^{-1}(r)\}} = \underbrace{\frac{f_X\{F_Y^{-1}(r)\}}{f_{\widetilde{X}}\{F_Y^{-1}(r)\}}}_{\text{location effect}} \times \underbrace{\frac{f_{\widetilde{X}}\{F_Y^{-1}(r)\}}{f_Y\{F_Y^{-1}(r)\}}}_{\text{shape effect}} \tag{5}$$

As above, one of the components is not a proper density. To describe the location effect, we may thus prefer

$$g_{X\widetilde{X}}(r) = \frac{f_X\{F_{\widetilde{X}}^{-1}(r)\}}{f_{\widetilde{X}}\{F_{\widetilde{X}}^{-1}(r)\}} \quad \text{and} \quad G_{X\widetilde{X}}(r) = F_X\{F_{\widetilde{X}}^{-1}(r)\}$$

instead of $f_X\{F_Y^{-1}(r)\}/f_{\widetilde{X}}\{F_Y^{-1}(r)\}$. Results from (4) and (5) will generally not be the same, although for some of the measures discussed below, it does not matter whether we adjust F_X or F_Y.

So far, an additive location shift has been used to adjust the comparison or reference distribution. For variables that can only be positive (for example, wages), it may be more natural to use a multiplicative shift and hence rescale the data proportionally. A multiplicative location adjustment of the reference distribution is given by $\widetilde{Y} = Y \times \mu_X/\mu_Y$, and hence

$$F_{\widetilde{Y}}(y) = F_Y(y \times \mu_Y/\mu_X)$$

The comparison distribution could be adjusted analogously. Furthermore, besides the location, we could also adjust the scale of the distributions. An (additive) location and scale adjustment of the reference distribution could be accomplished using $\widetilde{Y} = (Y - \mu_Y) \times s_X/s_Y + \mu_X$, such that

$$F_{\widetilde{Y}}(y) = F_Y\{(y - \mu_X) \times s_Y/s_X + \mu_Y\}$$

where s is a scale measure such as the interquartile range (IQR) or the standard deviation. For the multiplicative adjustment, there is no natural way to take account of the scale. However, using logarithms we can implement a proportional location and scale adjustment as $\widetilde{Y} = \exp[\{\ln(Y) - \mu_{\ln(Y)}\} \times s_{\ln(X)}/s_{\ln(Y)} + \mu_{\ln(X)}]$, such that

$$F_{\widetilde{Y}}(y) = F_Y(\exp[\{\ln(y) - \mu_{\ln(X)}\} \times s_{\ln(Y)}/s_{\ln(X)} + \mu_{\ln(Y)}])$$

2.6 Summary measures

2.6.1 Divergence

Handcock and Morris (1999) suggest Pearson's χ^2 divergence and the Kullback–Leibler divergence (relative entropy) as measures for distributional divergence, that is, as summary measures for the overall difference between the comparison distribution and the reference distribution. The χ^2 divergence between F_X and F_Y is defined as

$$\chi^2 = \int_{-\infty}^{\infty} \frac{\{f_X(y) - f_Y(y)\}^2}{f_Y(y)}\,\mathrm{d}y = \int_0^1 \{g(r) - 1\}^2\,\mathrm{d}r$$

The equality between the first and second expressions follows from the substitution rule for integrals, noting that $y = F_Y^{-1}(r)$ and $\mathrm{d}F_Y^{-1}(r)/\,\mathrm{d}r = 1/f_Y\{F_Y^{-1}(r)\}$. Likewise, the Kullback–Leibler divergence, which has an information-theoretic interpretation (negative entropy of the relative density), is defined as

$$\mathrm{KL} = \int_{-\infty}^{\infty} \ln\left\{\frac{f_X(y)}{f_Y(y)}\right\} f_X(y)\,\mathrm{d}y = \int_0^1 \ln\{g(r)\}g(r)\,\mathrm{d}r$$

For both measures, the divergence of F_X with respect to F_Y is not generally equal to the divergence of F_Y with respect to F_X. That is, the direction from which we look at the problem matters. An example for a symmetric divergence measure[2] is the total variation distance (TVD)

$$\mathrm{TVD} = \int_{-\infty}^{\infty} \frac{1}{2}\left|\frac{f_X(y)}{f_Y(y)} - 1\right| f_Y(y)\,\mathrm{d}y = \int_0^1 \frac{1}{2}\,|g(r) - 1|\,\mathrm{d}r$$

which is equal to half the area between the relative density curve and the parity line. Besides being symmetric, the TVD has an intuitive interpretation: it quantifies the proportion of data mass that would have to be redistributed in one of the distributions to make it equal to the other distribution. In the case of categorical data, the TVD is equal to the dissimilarity index by Duncan and Davis (1953), which is often used in analyses of segregation (for Stata implementations, see, for example, Jann [2004] or Reardon and Townsend [1999]).

For all three measures, in a location and shape decomposition, the location-effect divergence and the shape-effect divergence do not add up to the overall divergence. For example, we could location-adjust the reference distribution as in (3) and then obtain the location-effect divergence from $g_{\widetilde{Y}Y}(r)$ and the shape-effect divergence from $g_{X\widetilde{Y}}(r)$. Unfortunately, these two divergences do not add up to the overall divergence. For the Kullback–Leibler divergence, however, as pointed out by Handcock and Morris (1999), the following equality holds:

$$\mathrm{KL} = \mathrm{KL}_{X\widetilde{Y}Y} + \mathrm{KL}_{X\widetilde{Y}}$$

2. That is, the comparison and reference distribution can be swapped without changing the measure. The equality holds in theory; in an empirical application, the agreement will only be approximate because of the smoothing involved in density estimation.

$\text{KL}_{X\widetilde{Y}Y}$ is a (negative) cross-entropy defined as

$$\text{KL}_{X\widetilde{Y}Y} = \int_{-\infty}^{\infty} \ln\left\{\frac{f_{\widetilde{Y}}(y)}{f_Y(y)}\right\} f_X(y)\,\mathrm{d}y = \int_0^1 \ln\{g_{\widetilde{Y}Y}(r)\}\, g(r)\,\mathrm{d}r$$

This suggests that, in practice, it may make sense to identify the location-effect divergence as the difference between the overall divergence and the shape-effect divergence. An advantage of such an approach is also that results will not depend on whether we adjust the reference distribution or the comparison distribution.

2.6.2 Polarization

To compare the degree of inequality between the comparison distribution and the reference distribution, Handcock and Morris (1999) suggest the median relative polarization index (MRP). The index is positive if the comparison distribution is more unequal than the reference distribution; if the reference distribution is more unequal than the comparison distribution, the index will be negative. The MRP is defined as

$$\text{MRP} = 4 \times E_X\{|r_{\widetilde{Y}}(y) - 0.5|\} - 1 \qquad \text{MRP} \in [-1, 1]$$

where E_X is the expectation over the comparison distribution and $r_{\widetilde{Y}}(y)$ is the relative rank of y in the location-adjusted reference distribution (using the median as the location measure). The justification for the MRP is that the median of the location-adjusted relative ranks is 0.5 and the location-adjusted relative ranks will have a uniform distribution if the two distributions have the same shape. In this case, $E_X\{|r_{\widetilde{Y}}(y) - 0.5|\}$ is equal to $1/4$, such that MRP is 0. In the extreme case that all data mass of the comparison distribution is located in regions below and above the range of the location-adjusted reference distribution, $r_{\widetilde{Y}}(y)$ will be 0 or 1 for all y with positive density in F_X, such that $E_X\{|r_{\widetilde{Y}}(y) - 0.5|\} = 0.5$ and hence MRP $= 1$. In the opposite extreme, $r_{\widetilde{Y}}(y)$ will always be 0.5, leading to an MRP of -1.

The MRP can be decomposed into a lower polarization index (LRP) and an upper polarization index (URP) that quantify the relative polarization in the lower or upper half of the distribution, respectively:

$$\text{LRP} = 4 \times E_X[\text{abs}\{r_{\widetilde{Y}}(y) - 0.5\} | r_{\widetilde{Y}}(y) \leq 0.5] - 1$$
$$\text{URP} = 4 \times E_X[\text{abs}\{r_{\widetilde{Y}}(y) - 0.5\} | r_{\widetilde{Y}}(y) > 0.5] - 1$$

Because the conditional expectations in the definitions of LRP and URP each cover half the distribution of the location-adjusted relative ranks, the total polarization index is equal to the average of the lower and upper indices, that is,

$$\text{MRP} = 0.5 \times \text{LRP} + 0.5 \times \text{URP}$$

2.6.3 Other summary measures

Descriptive statistics of the relative ranks compose a further class of relative distribution summary measures. Quantities of interest may be, for example, the mean or median of the relative ranks, their standard deviation, or their IQR.

Note that the mean of the relative ranks is equivalent to the Gastwirth index, which measures the "probability that a randomly selected woman earns at least as much as a randomly chosen man" (Gastwirth 1975, 33).[3]

2.7 Covariate balancing

2.7.1 Integrating over conditional distributions

Handcock and Morris (1999) discuss covariate adjustment in terms of conditional distributions integrated over covariates. I will slightly change notation for the following exposition. Let $D \in \{0, 1\}$ be an indicator distinguishing between a comparison group ($D = 1$) and a reference group ($D = 0$), and let Y be an outcome variable available in both groups. The comparison distribution is $F_{Y|D=1}$, that is, the distribution of Y in group $D = 1$; the reference distribution is $F_{Y|D=0}$. Furthermore, let Z be a continuous covariate. Our goal is to obtain the relative distribution of $F_{Y|D=1}$ with respect to $F_{Y|D=0}$ while adjusting for possible differences in the distribution of Z between the two groups.

The marginal distribution of Y in group d can be written as

$$F_{Y|D=d}(y) = \int_{-\infty}^{\infty} f_{Z|D=d}(z) F_{Y|D=d,Z}(y|z)\,\mathrm{d}z$$

where $f_Z(z)$ is the density of Z and $F_{Y|Z}(y|z)$ is the conditional distribution of Y given Z. A counterfactual distribution can now be constructed by replacing one of the components. For example,

$$F^C_{Y|D=0}(y) = \int_{-\infty}^{\infty} f_{Z|D=1}(z) F_{Y|D=0,Z}(y|z)\,\mathrm{d}z \tag{6}$$

is the marginal distribution of Y that we would expect in the reference group if it had the same covariate distribution as the comparison group. That is, we can obtain the counterfactual distribution by integrating the conditional distribution of Y in the reference group over the covariate distribution of the comparison group. The covariate-adjusted relative distribution can then be obtained by comparing $F_{Y|D=1}$ with $F^C_{Y|D=0}$.[4]

The approach can be generalized to multiple covariates by integrating over the joint distribution of all covariates or to discrete covariates by taking weighted sums instead of integrals.

3. To be precise, according to the formal definition given by Gastwirth (1975), the index is equal to 1 minus the average relative rank of men's earnings in the distribution of women's earnings. However, because of the symmetry of the problem, this is equivalent to the average relative rank of women's earnings in the distribution of men's earnings. For a discussion of the Gastwirth index, also see Le Breton, Michelangeli, and Peluso (2012).

4. Naturally, we might as well adjust the comparison distribution and then compare the covariate-adjusted comparison distribution with the reference distribution. The two perspectives address the same question (that is, how the relative distribution of Y would look like if the two groups had the same distribution of Z) but give somewhat different answers. In the decomposition literature, this is discussed as the "index problem" (see, for example, Jann [2008]).

2.7.2 Reweighting

An equivalent but more attractive approach from an applied perspective is to conceptualize covariate adjustment as reweighting in the spirit of DiNardo, Fortin, and Lemieux (1996). Define

$$P(D=1|Z=z) = 1 - P(D=0|Z=z)$$

as the conditional probability of belonging to the comparison group given Z. Furthermore, define

$$\Psi(z) = \frac{P(D=1|Z=z)/P(D=1)}{P(D=0|Z=z)/P(D=0)}$$

We can then write the counterfactual distribution of Y in the reference group as

$$F^{C}_{Y|D=0}(y) = \int_{-\infty}^{\infty} f_{Z|D=0}(z) F_{Y|D=0,Z}(y|z) \Psi(z)\,\mathrm{d}z \tag{7}$$

This indicates that the counterfactual distribution can be estimated by simply reweighting the data by an estimate of $\Psi(z)$.[5] Mathematically, (7) is equivalent to (6) because

$$\Psi(z) = \frac{P(D=1|Z=z)/P(D=1)}{P(D=0|Z=z)/P(D=0)} = \frac{P(D=1|Z=z) \times \frac{f_Z(z)}{P(D=1)}}{P(D=0|Z=z) \times \frac{f_Z(z)}{P(D=0)}} = \frac{f_{Z|D=1}(z)}{f_{Z|D=0}(z)}$$

(using Bayes' theorem in the last step). The practical advantage of reweighting over integrating is that $\Pr(D=1|Z=z)$, and therefore, $\Psi(z)$ is relatively easy to estimate using binary choice models even if Z is a vector of several covariates (for example, logistic regression).[6]

In any case, whether we integrate over conditional distributions or we use reweighting, constructing counterfactual distributions in such a way assumes that the conditional distribution of Y is "stable", that is, that the covariate distribution can be modified without changing the conditional distribution. However, even if such an exogeneity assumption is unrealistic in a given application, the "as if" scenarios based on counterfactual distributions can still be informative.

Furthermore, note that reweighting can be used as an alternative approach to identify location and shape effects (instead of applying adjustments as described in section 2.5) by modeling Ψ as a function of Y. This is particularly useful if the analyzed outcome is categorical.

3 Estimation

For the following discussion, assume that there is a random sample of size n for which we observe two variables, X and Y. Furthermore, there is information on sampling

5. To reweight the comparison group, we would use factor $1/\Psi(z)$ instead of $\Psi(z)$.
6. In their description of the implementation of relative distribution methods in R, Handcock and Aldrich (2002) conduct covariate adjustment by resampling observations based on relative frequencies of covariate values. This is equivalent, in expectation, to reweighting the data by $\Psi(z)$.

weights w as well as a (possibly empty) vector of covariates Z. That is, the data are (Y_i, X_i, w_i, Z_i), $i = 1, \ldots, n$. Set $w_i = 1$ for all i in case there are no sampling weights.

We intend to analyze the relative distribution of X with respect to Y between two subsamples. Let D be an indicator for the comparison subsample ($D_i = 1$ if observation i belongs to the comparison subsample, 0 if it does not), and let $\mathcal{D} = \{i|D_i = 1\}$ be the set of indices for which $D_i = 1$. Likewise, let R be an indicator for the reference subsample ($R_i = 1$ if observation i belongs to the reference subsample, 0 if it does not), and let $\mathcal{R} = \{i|R_i = 1\}$ be the set of indices for which $R_i = 1$. That is, we want to compare the distribution of X in subsample $\mathcal{D}$ with the distribution of Y in subsample $\mathcal{R}$.

We will use $F_{X|D}$ to denote the former, that is, the conditional distribution of X given $D = 1$, and $F_{Y|R}$ to denote the latter. In general, we will use letter "D" for quantities related to $\mathcal{D}$ and letter "R" for quantities related to $\mathcal{R}$. For example, $W_D = \sum D_i w_i$ and $W_R = \sum R_i w_i$ will be the sum of weights in the comparison sample and the reference sample, respectively. Furthermore, define $W = \sum w_i$ as the total sum of weights.

Note that Y and X may be the same and that $\mathcal{D}$ and $\mathcal{R}$ do not have to be distinct nor exhaustive. I use such a general setup to cover all possible cases. For example, if the subsamples are distinct and $Y = X$, then we are in a setting in which a single variable is compared between two groups (for example, a comparison of wages from a sample of females to wages from a sample of males). Likewise, if $D = R$ and $Y \neq X$, we compare two variables within the same sample (for example, a comparison of data on wages for the same individuals between two time points). Furthermore, if $X = Y$ and $\mathcal{D}$ is included in $\mathcal{R}$, then we compare the distribution of a variable in a subsample with the pooled distribution of that variable. Finally, if the union of $\mathcal{D}$ and $\mathcal{R}$ does not cover the whole sample (that is, if there are observations for which $D = R = 0$), we are in a subpopulation estimation setting. Taking account of the observations that do not belong to the subpopulation may be important for standard error estimation.

3.1 The relative CDF

To obtain an estimate for the relative CDF

$$G(r) = F_{X|D}\{F^{-1}_{Y|R}(r)\}, \quad r \in [0, 1]$$

one can compute the relative rank of X_i in distribution $F_{Y|R}$ for each $i \in \mathcal{D}$ and then take the value of the empirical CDF of these relative ranks at value r. That is, first compute

$$\widehat{r}_i = \frac{1}{W_R} \sum_{j \in \mathcal{R}} w_j \mathbb{1}(Y_j \leq X_i) \qquad \text{for all } i \in \mathcal{D}$$

where $\mathbb{1}(a)$ is the indicator function (1 if a is true, 0 if false). Then obtain the CDF as

$$\widehat{G}(r) = \frac{1}{W_D} \sum_{i \in \mathcal{D}} w_i \mathbb{1}(\widehat{r}_i \leq r) \tag{8}$$

An issue with this simple computation is that it leads to a step function with jumps at distinct values of $\widehat{r}$. Let (i) refer to observations in $\mathcal{D}$ ordered by $\widehat{r}$, such that $\widehat{r}_{(1)} \leq \widehat{r}_{(2)} \leq \cdots \leq \widehat{r}_{(n_D)}$. If $\widehat{r}_{(i)} < r < \widehat{r}_{(i+1)}$, that is, if evaluation point r falls between two values of $\widehat{r}$, then $\widehat{G}(r)$ will be equal to the CDF corresponding with the lower value of $\widehat{r}$. Such behavior makes sense in case of an ordinary CDF. However, in the context of the relative distribution, it appears more appropriate to linearly interpolate between the two points because this is equivalent to breaking ties proportionally between the comparison distribution and the reference distribution. Hence, determine $\widehat{G}(r)$ as

$$\widehat{G}(r) = \widehat{G}_{(i')} + \left\{\widehat{G}_{(i'+1)} - \widehat{G}_{(i')}\right\} \frac{r - \widehat{r}_{(i')}}{\widehat{r}_{(i'+1)} - \widehat{r}_{(i')}} \tag{9}$$

with

$$\widehat{G}_{(i)} = \frac{1}{W_D} \sum_{j \in \mathcal{D}} w_j \mathbb{1}\big(\widehat{r}_j \leq \widehat{r}_{(i)}\big)$$

where i' is selected such that $\widehat{r}_{(i')} < r \leq \widehat{r}_{(i'+1)}$ [with $\widehat{r}_{(0)} = \widehat{G}_{(0)} = 0$ if $\widehat{r}_{(1)} > 0$ and $\widehat{r}_{(n_D+1)} = \widehat{G}_{(n_D+1)} = 1$ if $\widehat{r}_{(n_D)} < 1$]. For values of r that have an exact match in $\widehat{r}_i$, $i \in \mathcal{D}$, this leads to the same result as (8). For r values without an exact match, (9) is equivalent to picking the result from a linear segmented curve connecting the points given by $\{\widehat{G}_{(i)}, \widehat{r}_{(i)}\}$, $i = 1, \ldots, n_D$.

Equation (9) improves on (8) in that it uses interpolation in regions where (8) is flat. It does not, however, take into account that flat regions in (8) may include outcome values that only exist in $F_{Y|R}$, nor does it take into account that there might be regions where the true $G(r)$ is upright because of outcome values that only occur in $F_{X|D}$. To handle these issues and obtain an estimate that exactly traces the observed data pattern, we can compute the empirical CDF for $F_{X|D}$ and $F_{Y|R}$ at each observed value in the data and then use linear interpolation to obtain $\widehat{G}(r)$. Let $\mathcal{Y} = \{y_{(1)}, \ldots, y_{(J)}\}$ be the ordered set of all distinct outcome values observed for $F_{X|D}$ and $F_{Y|R}$. We then compute

$$\widehat{r}^D_{(j)} = \frac{1}{W_D} \sum_{i \in \mathcal{D}} w_i \mathbb{1}\big(X_i \leq y_{(j)}\big) \quad \text{and} \quad \widehat{r}^R_{(j)} = \frac{1}{W_R} \sum_{i \in \mathcal{R}} w_i \mathbb{1}\big(Y_i \leq y_{(j)}\big)$$

for all $j = 1, \ldots, J$, add origin $\widehat{r}^D_{(0)} = \widehat{r}^R_{(0)} = 0$, and obtain the relative CDF as

$$\widehat{G}(r) = \begin{cases} \widehat{r}^D_{(j^r)} & \text{if } r = 0 \\ \widehat{r}^D_{(j_r)} & \text{if } r = 1 \\ 0.5\left\{\widehat{r}^D_{(j_r)} + \widehat{r}^D_{(j^r)}\right\} & \text{if } r = \widehat{r}^R_{(j)} \text{ for any } j \\ \widehat{r}^D_{(j')} + \left\{\widehat{r}^D_{(j'+1)} - \widehat{r}^D_{(j')}\right\} \frac{r - \widehat{r}^R_{(j')}}{\widehat{r}^R_{(j'+1)} - \widehat{r}^R_{(j')}} & \text{else} \end{cases} \tag{10}$$

where j_r and j^r denote the smallest and largest value of j, respectively, for which $\widehat{r}^R_{(j)} = r$, and where j' is chosen such that $\widehat{r}^R_{(j')} < r < \widehat{r}^R_{(j'+1)}$. For graphical display, we

may also directly plot $\widehat{r}^D_{(j)}$ against $\widehat{r}^R_{(j)}$ and linearly connect the points. All estimates for $\widehat{G}(r)$ obtained using (10) will lie on that curve.

If all values in $\mathcal{Y}$ exist in both distributions, (10) will lead to the same results as (9). Furthermore, for continuous data, at least if the dataset is not very small, results from the two approaches will be very similar. Equation (10), however, leads to more appropriate results than (9) if the data are discrete.

3.2 Computing relative ranks

Relative density estimation and the estimation of summary measures of the relative distribution are typically implemented by analyzing the relative ranks of X_i, $i \in \mathcal{D}$ in distribution $F_{Y|R}$. A naïve approach is to compute the relative ranks using the values of the empirical CDF of $F_{Y|R}$, that is,

$$\widehat{r}_i = \frac{1}{W_R} \sum_{j \in \mathcal{R}} w_j \mathbb{1}(Y_j \leq X_i) \tag{11}$$

A problem with this approach is that the empirical CDF is a step function. This is particularly troublesome if there is heaping in the data such that there are large steps in the CDF, as is often the case with discrete data. One improvement is to use the so-called middistribution function instead of the regular CDF that deducts half a step size from the ranks in regions where the CDF is upright.[7] Let

$$\widehat{P}_R(Y = y) = \frac{1}{W_R} \sum_{j \in \mathcal{R}} w_j \mathbb{1}(Y_j = y)$$

be the relative frequency of outcome y in $F_{Y|R}$ (that is, the step size in the CDF at value y). The relative ranks computed according to the middistribution function then are

$$\widehat{r}_i = \frac{1}{W_R} \sum_{j \in \mathcal{R}} w_j \mathbb{1}(Y_j \leq X_i) - \frac{1}{2} \widehat{P}_R(Y = X_i) \tag{12}$$

Note that (12) differs from (11) only for observations that have ties in $F_{Y|R}$ (that is, observations that hit a step). For all other observations, $\widehat{P}_R$ is 0, and hence the two computations lead to the same result. The relative midranks are preferable over the naïve relative ranks because their average is exactly 0.5 if the two empirical distributions are identical. For the naïve relative ranks, this does not hold; their average will be larger than 0.5 in this situation. That is, the naïve relative ranks have an upward bias. The size of the bias depends on how much heaping there is in the data. The more heaping there is, the larger the bias.

Using the midrank adjustment removes the bias in the relative ranks. Heaping, however, will still lead to undesirable results such as arbitrary spikes in the relative

7. The term "middistribution function" has been coined by Emanuel Parzen (see, for example, Parzen [2004]), but the same concept also appears in various other sources under different names. For background information and references, see the help file of user command `distplot` (Cox 1998).

density estimate. A solution to this second issue is to break ties randomly and hence smooth out the step sizes of the CDF across tied observations. These broken relative ranks (including midrank adjustment) can be written as

$$\widehat{r}_i = \frac{1}{W_R}\sum_{j\in\mathcal{R}} w_j \mathbb{1}(Y_j \leq X_i) - \widehat{P}_R(Y = X_i)\frac{\widehat{P}_D(X = X_i) + 0.5w_i - \delta_i}{\widehat{P}_D(X = X_i)} \tag{13}$$

where $\widehat{P}_D(X = y)$ is the relative frequency of outcome y in $F_{X|D}$ and δ_i is the relative rank of X_i among all ties of X_i in $F_{X|D}$ when ties are broken randomly. Let $w_1^{(i)}, \ldots, w_{K_i}^{(i)}$ be the randomly ordered set of weights from the observations in $F_{X|D}$ that are equal to X_i (including observation i), where K_i is the size of the set (the order is kept stable across observations, that is, $w_k^{(i)} = w_k^{(j)}$ if $X_i = X_j$). Let k_i be the position of observation i in this set. The expression for δ_i then is

$$\delta_i = \frac{1}{\sum_{k=1}^{K_i} w_k^{(i)}} \sum_{k=1}^{k_i} w_k^{(i)}$$

which simplifies to $\delta_i = k_i/K_i$ if the weights are constant.[8]

To obtain broken relative ranks without midrank adjustment, set $0.5w_i$ in (13) to 0. Whereas the midrank adjustment can have a strong effect on results if relative ranks are computed without breaking ties [(11) versus (12)], the adjustment is only of minor importance in (13) because breaking ties makes the individual step sizes small (unless there is large variation in weights).

For location-adjusted relative ranks, the same equations can be applied to appropriately transformed input variables. For example, to compute the relative ranks based on a location-adjusted reference distribution, use

$$\widetilde{Y} = Y - \widehat{\mu}_{Y|R} + \widehat{\mu}_{X|D}$$

instead of Y in the above equations, where $\widehat{\mu}_{Y|R}$ is the median or mean of Y in subsample $\mathcal{R}$ and $\widehat{\mu}_{X|D}$ is the median or mean of X in subsample $\mathcal{D}$. Location, scale, multiplicative, or logarithmic adjustments can be handled analogously.

In contrast, for shape adjustment, one of the distributions has to be swapped. For example, to compute the relative ranks based on a shape-adjusted comparison distribution (that is, a comparison distribution that has the same shape as the reference distribution but a different location), use

$$\widetilde{X} = Y - \widehat{\mu}_{Y|R} + \widehat{\mu}_{X|D}$$

instead of X, and then set the comparison sample to $\widetilde{\mathcal{D}} = \mathcal{R}$ instead of $\mathcal{D}$.

8. Because of the random ordering, repeated computation of (13) will lead to slightly different results for the relative density and other estimates unless the weights are constant (or unless there are no ties). One (arbitrary) solution to enforce stable results is to sort the observations within ties in (ascending or descending) order of the weights.

3.3 The relative PDF

3.3.1 Kernel density estimation for continuous data

Estimation of the relative density can be implemented by applying a univariate density estimator to the relative ranks [preferably as defined in (13)]. Compared with a standard density estimation problem, there are two specific complications that should be accounted for. First, the support of the relative density is bounded at 0 and 1. Standard density estimators, however, are designed such that they smoothly approach 0 outside the support of the observed data, which leads to an underestimation of the density at the boundaries. Second, automatic bandwidth selection should be adapted to take account of the specific nature of relative data.

Given an evaluation point $r \in [0, 1]$, a kernel density estimate of the relative density can be written as

$$\widehat{g}(r) = \frac{1}{W_D} \sum_{i \in \mathcal{D}} w_i K_c(r, \widehat{r}_i, h) \tag{14}$$

where $K_c(r, \widehat{r}_i, h)$ is a boundary-corrected kernel function with bandwidth h. For example, the renormalization technique uses

$$K_c(r, \widehat{r}_i, h) = \frac{1}{h} K\left(\frac{r - \widehat{r}_i}{h}\right) c(r, h) \qquad \text{with} \qquad c(r, h) = \left\{ \int_{(0-r)/h}^{(1-r)/h} K(x)\, \mathrm{d}x \right\}^{-1}$$

where $K(x)$ is a standard kernel function such as the Gaussian kernel. The logic of the procedure is to rescale the density estimate by the inverse of the area of the kernel function that lies within the support of r. For some alternative boundary correction techniques, see Jann (2007).

The bandwidth h that determines the degree of smoothing (larger values for h lead to a smoother PDF) can either be set manually or be determined automatically from the data. Various suggestions for automatic bandwidth selectors exist in the literature, some based on crude rules of thumb and some employing more sophisticated procedures (see Jann [2007] for an overview of some of the suggestions). For relative density estimation, these standard bandwidth selectors should be adapted to take account of the specific nature of relative data. Suggestions for appropriate modifications are given by Ćwik and Mielniczuk (1993). The `reldist` command below supports several automatic bandwidth selectors, but we refrain from discussing their details here.[9]

9. Estimator (14) uses a global bandwidth that is constant across observations. A popular alternative is the adaptive estimator based on a varying bandwidth depending on the local density of the data (Abramson 1982). For the adaptive estimator, replace h by $h_i = h \times \sqrt{\widetilde{g}^0 / \widehat{g}^0(\widehat{r}_i)}$ where $\widehat{g}^0(\widehat{r}_i)$ is an initial (constant-bandwidth) density estimate and $\widetilde{g}^0$ is the geometric mean of $\widehat{g}^0(\widehat{r}_i)$ over all observations in $\mathcal{D}$. The procedure may be iterated several times (each time using the density estimate from the last step to determine the new h_i), but typically additional iterations do not change the estimate much. The adaptive estimator is attractive for regular density estimation because there is a one-to-one relation between the density and the local sample size. For the relative density, however, the local sample size is constant for one of the groups (the reference group), such that the adaptive estimator appears less convincing.

3.3.2 Histogram density estimation

A complement to kernel density estimation is to obtain a histogram of the relative density. Let $(a, b]$ be an interval on the support of r. The histogram density estimate for that interval can then be obtained as

$$\widehat{g}(a,b) = \frac{\widehat{P}_D(a < r \leq b)}{b-a} = \frac{1}{W_D} \sum_{i \in \mathcal{D}} w_i \frac{\mathbb{1}(a < \widehat{r}_i \leq b)}{b-a}$$

(with a modification in the case of $a = 0$ such that the interval includes the lower bound). A convenient setup is to split the support of r into K evenly sized bins defining the intervals $(0, \frac{1}{K}], (\frac{1}{K}, \frac{2}{K}], \ldots, (\frac{k-1}{K}, \frac{k}{K}], \ldots, (\frac{K-1}{K}, 1]$, such that each bin covers $\frac{1}{K}$th of the reference distribution.

The histogram density has an intuitive interpretation. For example, a value of 2 means that the fraction of the comparison distribution that falls into the bin is twice as large as the fraction of the reference distribution. In other words, the comparison distribution is overrepresented in the bin by a factor of 2. A value of 0.5 means that the proportion of the comparison distribution is only half the proportion of the reference distribution. A kernel density estimate of the relative ranks has, in principle, the same meaning (it shows the relative overrepresentation or underrepresentation multiplier at each level of r), but the explicit binning may make the histogram more easy to interpret.

3.3.3 Discrete relative density for categorical data

For categorical data, the relative density can be computed directly from the relative probabilities across the levels of the data. Without loss of generality, let $k = 1, \ldots, K$ be these levels. The relative density for level k is then estimated as

$$\widehat{g}_k = \frac{\widehat{P}_D(X = k)}{\widehat{P}_R(Y = k)}$$

with

$$\widehat{P}_D(X = k) = \frac{1}{W_D} \sum_{i \in \mathcal{D}} w_i \mathbb{1}(X_i = k) \quad \text{and} \quad \widehat{P}_R(Y = k) = \frac{1}{W_R} \sum_{i \in \mathcal{R}} w_i \mathbb{1}(Y_i = k)$$

Discrete relative density $\widehat{g}_k$ is well defined only for levels k that exist in the reference distribution.

When plotting the relative density for categorical data, $\widehat{g}_k$ can be plotted against $\widehat{P}_R(Y \leq k)$ using a step function, including an additional point at coordinate $(\widehat{g}_1, 0)$ for the first step. Alternatively, the density can be plotted using a histogram with bar widths equal to $\widehat{P}_R(Y = k)$ and bar midpoints equal to $\widehat{P}_R(Y \leq k) - \widehat{P}_R(Y = k)/2$.

3.4 Divergence

3.4.1 Continuous data

To estimate the χ^2, Kullback–Leibler, and dissimilarity measures, obtain an estimate of the relative density over a grid of evaluation points and then "integrate" the result. For example, let $r_k = k/K - 1/(2K)$, $k = 1, \ldots, K$, be a regular grid of evaluation points spanning the support of r from $1/(2K)$ to $1 - 1/(2K)$. The divergence measures can then be estimated as

$$\widehat{\chi^2} = \frac{1}{K}\sum_{k=1}^{K}\{\widehat{g}(r_k) - 1\}^2, \quad \widehat{\mathrm{KL}} = \frac{1}{K}\sum_{k=1}^{K}\widehat{g}(r_k)\ln\{\widehat{g}(r_k)\}, \quad \widehat{\mathrm{TVD}} = \frac{1}{2K}\sum_{k=1}^{K}|\widehat{g}(r_k) - 1| \tag{15}$$

where $\widehat{g}(r_k)$ is the density estimate at evaluation point r_k (that is, the integral is approximated by using a rectangle of width $1/K$ around each evaluation point). The size of the evaluation grid should not matter too much for the results, as long as it is sufficiently dense. However, results may strongly depend on the bandwidth used for density estimation. Divergence measures will typically increase with a decrease in the bandwidth. Stated differently, more smoothing leads to lower divergence. In general, TVD is less sensitive in this regard than the other two measures.

An alternative is to obtain the divergence measures from a histogram of the relative density. Assuming K evenly sized bins covering the whole range of r, the histogram-based estimates of the divergence measures can be obtained using (15) with $\widehat{g}(r_k)$ replaced by the histogram estimate of the relative density in bin k. Results may strongly depend on the number of bins.

3.4.2 Categorical data

Divergence measures for categorical data can be defined in terms of the categorical relative density as introduced above. Let $k = 1, \ldots, K$ be the levels of the data. The divergence estimates then are

$$\widehat{\chi^2} = \sum_{k=1}^{K}\frac{\left(\widehat{p}_k^D - \widehat{p}_k^R\right)^2}{\widehat{p}_k^R}, \qquad \widehat{\mathrm{KL}} = \sum_{k=1}^{K}\widehat{p}_k^D\ln\left(\frac{\widehat{p}_k^D}{\widehat{p}_k^R}\right), \qquad \widehat{\mathrm{TVD}} = \sum_{k=1}^{K}\frac{1}{2}\left|\widehat{p}_k^D - \widehat{p}_k^R\right|$$

where $\widehat{p}_k^D = \widehat{P}_D(X = k)$ and $\widehat{p}_k^R = \widehat{P}_R(Y = k)$.

3.5 MRP

For the polarization indices, first compute location-adjusted relative ranks using one of the above methods, where the median is used as the location measure. Let $\widehat{\widetilde{r}}_i$ be these location-adjusted ranks. Whether we transform the reference data or the comparison data does not matter. An estimate for MRP can then be obtained as

$$\widehat{\text{MRP}} = \left(\frac{4}{W_D} \sum_{i \in \mathcal{D}} w_i \left| \widehat{\widetilde{r}}_i - 0.5 \right| \right) - 1$$

Furthermore, using

$$\widehat{\text{LRP}} = \left\{ \frac{8}{W_D} \sum_{i \in \mathcal{D}} w_i \left| \widehat{\widetilde{r}}_i - 0.5 \right| \mathbb{1}(\widehat{\widetilde{r}}_i < 0.5) \right\} - 1$$

$$\widehat{\text{URP}} = \left\{ \frac{8}{W_D} \sum_{i \in \mathcal{D}} w_i \left| \widehat{\widetilde{r}}_i - 0.5 \right| \mathbb{1}(\widehat{\widetilde{r}}_i > 0.5) \right\} - 1$$

as estimates for LRP and URP ensures that

$$\widehat{\text{MRP}} = \frac{\widehat{\text{LRP}} + \widehat{\text{URP}}}{2}$$

Note that, in theory, the MRP of $F_{X|D}$ with respect to $F_{Y|R}$ is equal to $-$MRP of $F_{Y|R}$ with respect to $F_{X|D}$. In practice, however, heaping in the data may cause the median of the location-adjusted relative ranks to differ from 0.5 and hence cause this relation to be violated. Applying middistribution correction and breaking ties when computing the ranks typically reduces the discrepancy but may not entirely remove it.

3.6 Covariate balancing

Assume that $\mathcal{D}$ and $\mathcal{R}$ are distinct and exhaustive, such that D is an indicator for the comparison group ($D = 1$) versus the reference group ($D = 0$). A simple approach for covariate adjustment by reweighting is to run a logistic regression of D on Z and obtain predictions $\widehat{p}_i = \widehat{P}(D = 1|Z = Z_i)$ from the model. To reweight the reference group, define adjusted weights

$$\widetilde{w}_i = \begin{cases} w_i \frac{\widehat{p}_i}{1-\widehat{p}_i} c_R & \text{if } i \in \mathcal{R} \\ w_i & \text{else} \end{cases}$$

where $c_R = W_R / \sum_{i \in \mathcal{R}} w_i \frac{\widehat{p}_i}{1-\widehat{p}_i}$ is a scaling factor ensuring that the group size (that is, its sum of weights) remains constant, and use these weights in all computations instead of the original weights. Likewise, to reweight the comparison group, define the adjusted weights as

$$\widetilde{w}_i = \begin{cases} w_i \frac{1-\widehat{p}_i}{\widehat{p}_i} c_D & \text{if } i \in \mathcal{D} \\ w_i & \text{else} \end{cases}$$

with $c_D = W_D / \sum_{i \in \mathcal{D}} w_i \frac{1-\widehat{p}_i}{\widehat{p}_i}$. The described procedure is equivalent to what is known as "inverse probability weighing" (IPW) in the causal inference literature (see [TE] **teffects ipw**). Any other approach to obtain balancing weights may do as well. See, for example, `kmatch` (Jann 2017) for techniques such as entropy balancing or matching.

4 The reldist command

The command `reldist` implements the methods discussed above. The `moremata` (Jann 2005) package is required. For installation, type

```
. ssc install reldist, replace
. ssc install moremata, replace
```

4.1 Syntax

4.1.1 Estimation

The command `reldist` has two syntaxes. Use syntax 1 if you want to analyze the relative distribution of a single variable between two groups or subpopulations. Syntax 2 is for comparing two variables within a single sample.

Syntax 1 (two-sample relative distribution):

`reldist` *subcmd varname* [*if*] [*in*] [*weight*], `by(`*groupvar*`)` [*options*]

where *groupvar* identifies two groups to be compared.

Syntax 2 (paired relative distribution):

`reldist` *subcmd varname refvar* [*if*] [*in*] [*weight*] [, *options*]

where *varname* and *refvar* specify two variables to be compared.

In both cases, `fweights`, `iweights`, and `pweights` are allowed (see [U] **11.1.6 weight**), and *subcmd* can be

`pdf`	to estimate the PDF of the relative distribution, possibly including a histogram of the relative density
`histogram`	to estimate a histogram of the relative density
`cdf`	to estimate the relative CDF (equivalent to a so-called probability–probability plot)
`divergence`	to estimate the Kullback–Leibler divergence (entropy), the χ^2 divergence, or the dissimilarity index (TVD) of the relative distribution
`mrp`	to estimate the MRP as well as its decomposition into an LRP and URP
`summarize`	to estimate summary statistics such as the mean or median of the relative ranks and, optionally, store the relative ranks in a new variable

4.1.2 Creating a graph after estimation

After applying `reldist pdf`, `reldist histogram`, or `reldist cdf`, the command `reldist graph` can be used to draw a graph of the results. The syntax is

`reldist graph` [`,` *graph_options*]

An alternative is to generate the graph directly using option `graph()` with `reldist pdf`, `reldist histogram`, or `reldist cdf`.

4.1.3 Storing IFs after estimation

The command `predict` can be applied after `reldist` to generate the IFs of the estimated parameters (one variable per parameter). The syntax is

`predict` { *stub*`*` | *newvarlist* } [*if*] [*in*] [`, scores` *density_options*]

where *stub* specifies a common prefix for the names of the generated variables; alternatively, *newvarlist* specifies an explicit list of variable names to be used. Option `scores` is allowed for compatibility reasons; it does not do anything. *density_options* can be used to modify how auxiliary densities are estimated during the computation of the IFs; see page 909 for a description of available *density_options* (option `boundary()` will have no effect because unbounded support is assumed for auxiliary densities).

The command `total` (see [R] **total**) can be applied to the stored IFs to replicate the standard errors reported by `reldist`.

4.2 Options for reldist

4.2.1 Main options

`by(`*groupvar*`)` specifies a binary variable that identifies the two groups to be compared. By default, the group with the lower value will be used as the reference group. `by()` is required in syntax 1 and not allowed in syntax 2.

`swap` reverses the order of the groups identified by `by()`. `swap` is only allowed in syntax 1.

`pooled` uses the pooled distribution across both groups as the reference distribution. `pooled` is only allowed in syntax 1.

`balance(`*spec*`)` balances covariate distributions between the comparison group and the reference group using reweighting. `balance()` is only allowed in syntax 1. The syntax of *spec* is

[*method*`:`] *varlist* [`,` *options*]

where *method* is either `ipw` for inverse probability weighting based on logistic regression (the default) or `eb` for entropy balancing (using `mm_ebal()` from `moremata`), *varlist* specifies the list of covariates to be balanced, and *options* are as follows:

`reference` reweights the reference group. The default is to reweight the comparison group. Option `pooled` is not allowed with `balance(`*varlist*`, reference)`.

`contrast` compares the balanced distribution with the unbalanced distribution. Use this option to see how the balancing changes the distribution. If `contrast` is specified together with `reference`, the balanced reference distribution will be used as the comparison distribution. If `contrast` is specified without `reference`, the balanced comparison distribution will be used as the reference distribution.

logit_options are options to be passed through to `logit` (see [R] **logit**). *logit_options* are only allowed if *method* is `ipw`.

`btolerance(`#`)`, where # $\geq$ 0, specifies the tolerance for the entropy balancing algorithm. The default is `btolerance(1e-5)`. A warning message is displayed if a balancing solution is not within the specified tolerance. `btolerance()` is only allowed if *method* is `eb`.

`noisily` displays the output of the balancing procedure.

`generate(`*newvar*`)` stores the balancing weights in variable *newvar*. This is useful if you want to check whether covariates have been successfully balanced.

`adjust(`*spec*`)` applies location, scale, and shape adjustments to the comparison and reference distributions. `adjust()` is not allowed with `reldist mrp`. The syntax of *spec* is

adjust [`,` *options*]

where *adjust* specifies the desired adjustments. *adjust* may contain any combination of at most two of the following keywords:

`location`	adjust location
`scale`	adjust scale
`shape`	adjust shape

By default, the specified adjustments are applied to the comparison distribution. However, a colon may be included in *adjust* to distinguish between distributions: Keywords before the colon affect the comparison distribution; keywords after the colon affect the reference distribution. For example, type `adjust(location scale)` to adjust the location and scale of the comparison distribution. Likewise, you could type `adjust(:location scale)` to adjust the reference distribution. Furthermore, `adjust(location : shape)` would adjust the location of the comparison distribution and the shape of the reference distribution. *options* are as follows:

`mean` uses the mean for the location adjustment. The default is to use the median.

`sd` uses the standard deviation for the scale adjustment. The default is to use the IQR.

`multiplicative` uses a multiplicative adjustment instead of an additive adjustment. *adjust* may only contain one keyword in this case, either `location` or `shape`. An error will be returned if the location ratio between the comparison distribution and the reference distribution is not strictly positive.

`logarithmic` performs the adjustments on logarithmically transformed data. An error will be returned if the data are not strictly positive.

rank_options specify the details about the computation of relative ranks. These options are irrelevant for `reldist histogram`, `reldist cdf`, and `reldist divergence` unless option `pdf` is specified and for `reldist pdf` if `discrete` or `categorical` is specified. The options are as follows:

`nobreak` changes how the relative ranks are computed in case of ties. By default, `reldist` breaks ties randomly for comparison values that have ties in the reference distribution (in ascending order of weights if weights have been specified). This leads to improved results if there is heaping in the data. Specify `nobreak` to omit breaking ties.

`nomid` changes how the relative ranks are computed in case of ties. By default, `reldist` uses midpoints of the steps in the cumulative distribution for comparison values that have ties in the reference distribution. This ensures that the average relative rank is equal to 0.5 if the comparison and reference distributions are identical. Specify `nomid` to assign relative ranks based on full steps in the CDF.

`descending` sorts tied observations in descending order of weights. The default is to use ascending sort order. Option `descending` has no effect if `nobreak` is specified or if there are no weights.

nostable breaks ties randomly (within unique values of weights). The default is to break the ties in the sort order of the data (within unique values of weights). Option nostable has no effect on the results reported by reldist. It may, however, affect the ranks stored by option generate() or the IFs stored by predict (unless option nobreak is specified).

replace allows the user to replace existing variables. This is relevant for generate() with reldist summarize and generate() in balance().

4.2.2 Additional options for reldist pdf

n(#) sets the number of evaluation points for which the PDF is to be computed. A regular grid of # evaluation points between 0 and 1 will be used. The default is n(101) (unless option discrete or categorical is specified, in which case n() has no default). Only one of n(), at(), and atx is allowed.

at(*numlist* | *matname*) specifies a custom grid of evaluation points between 0 and 1 by providing either a *numlist* (see [U] **11.1.8 numlist**) or the name of a matrix containing the values (the values will be taken from the first row or the first column of the matrix, depending on which is larger). Only one of n(), at(), and atx is allowed.

atx[(comparison | reference | *numlist* | *matname*)] specified without argument causes the relative PDF to be evaluated at each distinct outcome value that exists in the data (possibly after applying adjust()), instead of using a regular evaluation grid on the probability scale. All outcome values across both distributions will be considered. To restrict the evaluation points to outcome values from the comparison distribution or the reference distribution, specify atx(comparison) or atx(reference), respectively. Alternatively, specify a grid of custom values by providing either a *numlist* (see [U] **11.1.8 numlist**) or the name of a matrix containing the values (the values will be taken from the first row or the first column of the matrix, depending on which is larger). Only one of n(), at(), and atx is allowed.

discrete causes the data to be treated as discrete. The relative PDF will then be evaluated at each level of the data as the ratio of the level's frequency between the comparison distribution and the reference distribution instead of using kernel density estimation, and the result will be displayed as a step function. If option n() or at() is specified, the step function will be evaluated at the points of the corresponding probability grid instead of returning the relative density for each outcome level. Options nobreak, nomid, descending, and *density_options* have no effect if discrete is specified. Furthermore, options histogram and adjust() are not allowed.

categorical is like discrete but additionally requests that the data only contain positive integers. Factor-variable notation will be used to label the coefficient in the output table.

histogram[(#)] computes a histogram in addition to the PDF, where # is the number of bins. If # is omitted, 10 bins will be used.

`alt` uses an alternative estimation method for the histogram. See the histogram options below.

density_options set the details of kernel density estimation. The options are as follows:

`bwidth(`*#* | *method*[`, nord`]`)` determines the bandwidth of the kernel, the halfwidth of the estimation window around each evaluation point. Use `bwidth(`*#*`)`, *# >* 0, to set the bandwidth to a specific value. Alternatively, type `bwidth(`*method*`)` to choose an automatic bandwidth selection method. Choices are `silverman` (optimal of Silverman), `normalscale` (normal scale rule), `oversmoothed` (oversmoothed rule), `sjpi` (Sheather–Jones solve-the-equation plugin), `dpi`[`(`*#*`)`] (Sheather–Jones direct plugin estimate, where *#* specifies the number of stages of functional estimation; the default is `2`), or `isj` (diffusion estimator bandwidth). The default is `bw(sjpi)`. See Jann (2007) for information on `silverman`, `normalscale`, `oversmoothed`, `sjpi`, and `dpi`. For `isj`, see Botev, Grotowski, and Kroese (2010).

By default, if estimating the density of the relative data, all bandwidth selectors include a correction for relative data based on Ćwik and Mielniczuk (1993). Specify suboption `nord` to omit the correction.

`bwadjust(`*#*`)` multiplies the bandwidth by *#*, where *# >* 0. The default is `bwadjust(1)`.

`boundary(`*method*`)` sets the type of boundary correction method. Choices are `renorm` (renormalization method; the default), `reflect` (reflection method), or `lc` (linear combination technique). See Jann (2007) for details on boundary correction methods.

`adaptive(`*#*`)` specifies the number of iterations used by the adaptive kernel density estimator. The default is `adaptive(0)` (nonadaptive density estimator).

`kernel(`*kernel*`)` specifies the kernel function to be used. *kernel* may be one of `epanechnikov` (Epanechnikov kernel function), `epan2` (alternative Epanechnikov kernel function), `biweight` (biweight kernel function), `triweight` (triweight kernel function), `cosine` (cosine trace), `gaussian` (Gaussian kernel function), `parzen` (Parzen kernel function), `rectangle` (rectangle kernel function), or `triangle` (triangle kernel function). The default is `kernel(gaussian)`.

`napprox(`*#*`)` specifies the grid size used by the binned approximation density estimator (and by the data-driven bandwidth selectors). The default is `napprox(512)`.

`exact` causes the exact kernel density estimator to be used instead of the binned approximation estimator. The exact estimator can be slow in large datasets if the density is to be evaluated at many points.

`graph`[`(`*graph_options*`)`] displays the results in a graph. The coefficients table will be suppressed in this case (unless option `table` is specified). Alternatively, use `reldist graph` to display the graph after estimation.

`ogrid(#)` sets the size of the approximation grid for outcome labels. The default is `ogrid(401)`. The grid is stored in `e(ogrid)` and will be used by graph option `olabel()` to determine the positions of outcome labels. Type `noogrid` to omit the computation of the grid (no outcome labels will then be available for the graph). Option `ogrid()` is only allowed if the relative density is computed with respect to an evaluation grid on the probability scale. If the relative density is evaluated with respect to specific outcome values (for example, if `atx` is specified), the outcome labels will be obtained from the information stored in `e(at)`.

4.2.3 Additional options for reldist histogram

`n(#)` specifies the number of histogram bars. The reference distribution will be divided into # bins of equal width. That is, each bin will cover 1/#th of the reference distribution. The default is `n(10)`.

`alt` uses an alternative estimation method. The default method obtains the relative histogram by computing the empirical CDF of both distributions at all values that exist in the data (across both distributions). The alternative method obtains the relative histogram based on the empirical CDF of the relative ranks. In both cases, if necessary, linear interpolation will be used to map the relative CDF to the evaluation points.

`discrete` causes the data to be treated as discrete. The relative density will then be evaluated at each level of the data as the ratio of the level's frequency between the two distributions, and the width of bars will be proportional to the reference distribution. Option `alt` has no effect and options `n()` and `adjust()` are not allowed if `discrete` is specified.

`categorical` is like `discrete` but additionally requests that the data only contain positive integers. Factor-variables notation will be used to label the coefficient in the output table.

`graph`[(*graph_options*)] displays the results in a graph. The coefficients table will be suppressed in this case (unless option `table` is specified). Alternatively, use `reldist graph` to display the graph after estimation.

`ogrid(#)` sets the size of the approximation grid for outcome labels. The default is `ogrid(401)`. The grid is stored in `e(ogrid)` and will be used by graph option `olabel()` to determine the positions of outcome labels. Type `noogrid` to omit the computation of the grid (no outcome labels will then be available for the graph). `ogrid()` is not allowed together with `discrete` or `categorical`.

4.2.4 Additional options for reldist cdf

`n(#)` sets the number of evaluation points for which the CDF is to be computed. A regular grid of # evaluation points between 0 and 1 will be used. The default is `n(101)` (unless option `discrete` or `categorical` is specified, in which case `n()` has no default). Only one of `n()`, `at()`, and `atx` is allowed.

`at(`*numlist* | *matname*`)` specifies a custom grid of evaluation points between 0 and 1 by providing either a *numlist* (see [U] **11.1.8 numlist**) or the name of a matrix containing the values (the values will be taken from the first row or the first column of the matrix, depending on which is larger). Only one of `n()`, `at()`, and `atx` is allowed.

`atx`[`(`<u>`comp`</u>`arison` | <u>`ref`</u>`erence` | *numlist* | *matname*`)`] specified without argument causes the relative CDF to be evaluated at each distinct outcome value that exists in the data (possibly after applying `adjust()`), instead of using a regular evaluation grid on the probability scale. All outcome values across both distributions will be considered. To restrict the evaluation points to outcome values from the comparison distribution or from the reference distribution, specify `atx(comparison)` or `atx(reference)`, respectively. Alternatively, specify a grid of custom values by providing either a *numlist* (see [U] **11.1.8 numlist**) or the name of a matrix containing the values (the values will be taken from the first row or the first column of the matrix, depending on which is larger). Only one of `n()`, `at()`, and `atx` is allowed.

`alt` uses an alternative estimation method. The default method obtains the relative CDF by computing the empirical CDF of both distributions at all values that exist in the data (across both distributions). The alternative method obtains the relative CDF based on the empirical CDF of the relative ranks. In both cases, if necessary, linear interpolation will be used to map the relative CDF to the evaluation points.

<u>`discr`</u>`ete` causes the data to be treated as discrete. The relative CDF will then be evaluated at each observed outcome value instead of using an evaluation grid on the probability scale. Option `discrete` leads to the same result as specifying `atx`. Option `adjust()` is not allowed if `discrete` is specified.

<u>`cat`</u>`egorical` is like `discrete` but additionally requests that the data only contain positive integers. Factor-variables notation will be used to label the coefficient in the output table.

<u>`graph`</u>[`(`*graph_options*`)`] displays the results in a graph. The coefficients table will be suppressed in this case (unless option `table` is specified). Alternatively, use `reldist graph` to display the graph after estimation.

`ogrid(`#`)` sets the size of the approximation grid for outcome labels. The default is `ogrid(401)`. The grid is stored in `e(ogrid)` and will be used by graph option `olabel()` to determine the positions of outcome labels. Type `noogrid` to omit the computation of the grid (no outcome labels will then be available for the graph). Option `ogrid()` is only allowed if the relative CDF is computed with respect to an evaluation grid on the probability scale. If the relative CDF is evaluated with respect to specific outcome values (for example, if `atx` is specified), the outcome labels will be obtained from the information stored in `e(at)`.

4.2.5 Additional options for reldist divergence

`over(`*overvar*`)` computes results for each subpopulation defined by the values of *overvar*.

`entropy` or `kl` computes the Kullback–Leibler divergence (entropy) of the relative distribution. This is the default.

`chi2` or `chisquared` computes the χ^2 divergence of the relative distribution.

`tvd` or `dissimilarity` computes the dissimilarity index (TVD) of the relative distribution.

`all` computes all supported divergence measures. `all` is equivalent to `entropy chi2 tvd`.

`n(#)` specifies the number of histogram bars or, if option `pdf` is specified, the number of kernel density evaluation points used to estimate the relative distribution. The default is `n(20)` or, if option `pdf` is specified, `n(100)`.

`alt` uses an alternative estimation method for the histogram. See the histogram options above.

`pdf` computes the divergence measures based on a kernel density estimate instead of a histogram estimate.

density_options set the details of the kernel density estimation. This is only relevant if option `pdf` is specified. See page 909 for available options.

`discrete` causes the data to be treated as discrete. The relative density will then be evaluated at each level of the data as the ratio of the level's frequency between the two distributions. Option `alt` has no effect, and options `n()`, `pdf`, and `adjust()` are not allowed if `discrete` is specified.

`categorical` is like `discrete` but additionally requests that the data only contain positive integers.

`compare`[`(`*options*`)`] estimates divergence measures for two models of the relative distribution, a main model and an alternative model, and also reports the difference between the two variants. *options* are `balance()` and `adjust()` as described above. `balance()` and `adjust()` specified as main options are applied to the main model; `balance()` and `adjust()` specified within `compare()` are applied to the alternative model.

4.2.6 Additional options for reldist mrp

`over(`*overvar*`)` computes results for each subpopulation defined by the values of *overvar*.

`multiplicative` applies multiplicative location adjustment. The default is to use additive adjustment. Only one of `logarithmic` and `multiplicative` is allowed.

`logarithmic` causes the location (and, optionally, scale) adjustment to be performed on the logarithmic scale. Only one of `logarithmic` and `multiplicative` is allowed.

scale[(sd)] adjusts the scale of the data before computing the polarization indices. If scale is specified without argument, the IQR is used; that is, the scale of the data will be adjusted such that the IQR is the same in both distributions. Specify scale(sd) to use the standard deviation instead of the IQR. scale is not allowed if multiplicative is specified.

4.2.7 Additional options for reldist summarize

over(*overvar*) computes results for each subpopulation defined by the values of *overvar*.

statistics(*statnames*) specifies a space-separated list of summary statistics to be reported. The default is statistics(mean). The following summary statistics are supported:

mean	mean
variance	variance
sd	standard deviation
median	median; equivalent to p50
p#	#th percentile, where # is an integer between 1 and 99
iqr	interquartile range (p75–p25)

generate(*newvar*) stores the relative ranks (based on adjusted data) in variable *newvar*. Depending on adjust(), different observations may be filled in.

4.2.8 Variance estimation options

level(#) specifies the confidence level, as a percentage, for confidence intervals. The default is level(95) or as set by set level (see [R] **level**).

vce(*vcetype*) determines how standard errors and confidence intervals are computed. *vcetype* may be

analytic [, *density_options*]

cluster *clustvar* [, *density_options*]

svy [*svy_vcetype*] [, *svy_options* *density_options*]

bootstrap [, *bootstrap_options*]

jackknife [, *jackknife_options*]

The default is vce(analytic), which computes the standard errors based on IFs. Likewise, vce(cluster *clustvar*) computes IF-based standard errors allowing for intragroup correlation, where *clustvar* specifies to which group each observation belongs. In both cases, *density_options* specify how auxiliary densities are estimated during the computation of the IFs (see page 909 for details; option boundary() will have no effect because unbounded support is assumed for auxiliary densities).

`vce(svy)` computes standard errors, taking the survey design as set by `svyset` (see [SVY] **svyset**) into account. The syntax is equivalent to the syntax of the `svy` prefix command (see [SVY] **svy**); that is, `vce(svy)` is `reldist`'s way to support the `svy` prefix. If *svy_vcetype* is set to `linearized`, the standard errors are estimated based on IFs; use *density_options* to specify the details of auxiliary density estimation in this case. For a *svy_vcetype* other than `linearized`, *density_options* are not allowed.

`vce(bootstrap)` and `vce(jackknife)` compute standard errors using `bootstrap` or `jackknife`, respectively (see [R] **bootstrap** or [R] **jackknife**); see [R] ***vce_option***.

If a replication technique is used for standard error estimation (that is, `vce(bootstrap)`, `vce(jackknife)`, or `vce(svy)` with *svy_vcetype* other than `linearized`), the bandwidth used by `reldist pdf` will be held fixed across replications (that is, if relevant, the bandwidth will be determined upfront and then held constant). If you want to repeat the bandwidth search in each replication, use `bootstrap`, `jackknife`, or `svy` as a prefix command.

Simulation results suggest that the IF-based standard errors work well in most situations. They may be severely biased, however, if there is heaping in the data. Replication-based techniques may yield more valid results in this case.

`nose` prevents `reldist` from computing standard errors. This saves computer time.

4.2.9 Reporting options

`citransform` reports transformed confidence intervals depending on the type of reported statistics (log transform for PDF and histogram density, logit transform for CDF and descriptive statistics, and inverse hyperbolic tangent transform for polarization indices). `citransform` only has an effect in Stata 15 or newer.

`noheader` suppresses the output header.

[`no`]`table` controls the output table containing the estimated coefficients. `notable` suppresses display of the table; `table` enforces display of the table if option `graph` has been specified.

display_options are standard reporting options such as `cformat()` or `coeflegend`; see the *Reporting options* in [R] **Estimation options**.

4.3 Options for reldist graph

4.3.1 Main graph options

`refline(`*line_options*`)` specifies options to affect the rendition of the parity line. See [G] ***line_options***.

`norefline` suppresses the parity line.

4.3.2 Additional options after reldist pdf

cline_options affect the rendition of the PDF line. See [G] ***cline_options***.

histopts(*options*) specifies options to affect the rendition of the histogram bars (if a histogram was computed) and the corresponding confidence spikes. *options* are as follows:

barlook_options affect the rendition of the histogram bars. See [G] ***barlook_options***.

ciopts(*rcap_options*) specifies options to affect the rendition of the confidence spikes of the histogram bars. See [G] ***rcap_options***.

noci omits the confidence spikes of the histogram bars.

nohistogram omits the histogram bars.

4.3.3 Additional options after reldist histogram

barlook_options affect the rendition of the histogram bars. See [G] ***barlook_options***.

4.3.4 Additional options after reldist cdf

noorigin prevents adding a $(0, 0)$ coordinate to the plotted line. If the first X coordinate of the CDF is larger than 0 and the range of the CDF has not been restricted by at() or atx, reldist graph will automatically add a $(0, 0)$ coordinate to the plot. Type noorigin to override this behavior.

cline_options affect the rendition of the CDF line. See [G] ***cline_options***.

4.3.5 Confidence intervals

level(#) specifies the confidence level, as a percentage, for confidence intervals. level() and ci() are not allowed together.

citransform plots transformed confidence intervals depending on the type of reported statistic (log transform for PDF and histogram density, and logit transform for CDF).

ci(*name*) obtains the confidence intervals from e(*name*) instead of computing them from e(V). e(*name*) must contain two rows and the same number of columns as e(b). For example, after bootstrap estimation, you could type ci(ci_percentile) to plot percentile confidence intervals. ci() and level() are not allowed together.

ciopts(*options*) specifies options to affect the rendition of the confidence intervals. See [G] ***area_options***, or after reldist histogram, see [G] ***rcap_options***. Use option recast() to change the plot type used for confidence intervals. For example, type ciopts(recast(rline)) to use two lines instead of an area.

noci omits the confidence intervals.

4.3.6 Outcome labels

[y]olabel[(*spec*)] adds outcome labels on a secondary axis. olabel() adds outcome labels for the reference distribution; yolabel() adds outcome labels for the comparison distribution (only allowed after reldist cdf). The syntax of *spec* is

[#\# | *numlist*] [, [noprune | prune(*mindist*)] at format(%*fmt*) *suboptions*]

#\# requests that (approximately) # outcome labels be added at (approximately) evenly spaced positions; the default is #6. Alternatively, specify *numlist* to generate labels for given outcome values.

prune(*mindist*) requests that an outcome label (but not its tick) be omitted if its distance to the preceding label is less than *mindist* (an exception are labels that have the same position; in such a case, the largest label will be printed). The default is prune(0.1); type prune(0) or noprune to print labels at all positions. The difference between prune(0) and noprune is that prune(0) will only print one label per position, whereas noprune prints all labels, including labels that have the same position.

at causes *numlist* to be interpreted as a list of probabilities for which outcome labels are to be determined. Labels obtained this way will not be pruned.

format(%*fmt*) specifies the display format for the outcome labels. The default is format(%6.0g). See [D] **format** for available formats.

suboptions are as described in [G] ***axis_label_options***.

Option [y]olabel may be repeated. Use suboptions add and custom to generate multiple sets of labels with different rendering; see [G] ***axis_label_options***.

[y]otick(*spec*) adds outcome ticks on a secondary axis. otick() adds outcome ticks for the reference distribution; yotick() adds outcome ticks for the comparison distribution (only allowed after reldist cdf). The syntax of *spec* is

numlist [, *suboptions*]

where *numlist* specifies the outcome values for which ticks are to be generated and *suboptions* are as described in [G] ***axis_label_options***. Option [y]otick() may be repeated. Use suboptions add and custom to generate multiple sets of ticks with different rendering; see [G] ***axis_label_options***.

[y]oline(*spec*) draws added lines at the positions of the specified outcome values on a secondary axis. oline() adds outcome lines for the reference distribution; yoline() adds outcome lines for the comparison distribution (only allowed after reldist cdf). The syntax of *spec* is

numlist [, *suboptions*]

where *numlist* specifies the outcome values for which added lines be generated and *suboptions* are as described in [G] ***added_line_options***. Option [y]`oline()` may be repeated to draw multiple sets of lines with different rendering.

[y]`otitle(`*tinfo*`)` provides a title for the outcome scale axis; see [G] ***title_options***. `otitle()` is for the reference distribution; `yotitle()` is for the comparison distribution (only allowed after `reldist cdf`).

Technical note: The positions of the outcome labels, ticks, or lines are computed from information stored by `reldist` in `e()`, either from the quantiles stored in `e(ogrid)` or from the values stored in `e(at)`, depending on context. There is an undocumented command called `reldist olabel` that can be used to compute the positions after the relative distribution has been estimated. Use this command, for example, if you want to draw a custom graph from the stored results without applying `reldist graph`. The syntax is

`reldist olabel` [`##` | *numlist*] [`,` [`noprune`|`prune(`*mindist*`)`] `at format(%`*fmt*`) tick(`*numlist*`) line(`*numlist*`) y`]

where `#`# or *numlist* specifies the (number of) values for which labels are to be generated, `prune()` determines the pruning (see above), `at` changes the meaning of the main *numlist* (see above), `format()` specifies the display format for the labels, `tick()` specifies values for which ticks are to be generated, `line()` specifies values for which added lines are to be generated, and `y` requests outcome labels for the Y axis of the relative CDF (only allowed after `reldist cdf`). `reldist olabel` stores the following in `r()`:

Macros

`r(label)`	label specification for use in an `xlabel()` option
`r(label_x)`	expanded and sorted *numlist*
`r(tick)`	tick specification for use in an `xtick()` option
`r(tick_x)`	expanded and sorted *numlist* from `tick()`
`r(line)`	line specification for use in an `xline()` option
`r(line_x)`	expanded and sorted *numlist* from `line()`

4.3.7 General graph options

`addplot(`*plot*`)` provides a way to add other plots to the generated graph. See [G] ***addplot_option***.

twoway_options are any options other than `by()` documented in [G] ***twoway_options***.

4.4 Stored results

`reldist` stores its results in `e()`, similar to official Stata's estimation commands. See the online documentation of `reldist` for details.

5 Examples

5.1 Wage mobility in two eras

I illustrate some of the features of `reldist` by replicating an analysis of permanent wage growth from Handcock and Morris (1999, chap. 8). The data cover wages of white males from two cohorts of the National Longitudinal Survey, an "original" cohort started in 1966 and a "recent" cohort started in 1979. The variable of interest is the estimated growth in permanent wages between age 16 and age 34 (see appendix C in Handcock and Morris [1999]). The data further contain information on the achieved educational level, and there is a variable providing sampling weights.[10]

```
. use nls
(NLS data from Handcock and Morris (1999))
. describe
Contains data from nls.dta
 Observations:         3,937                  NLS data from Handcock and
                                                Morris (1999)
    Variables:             4                  15 Jun 2021 10:17
                                              (_dta has notes)
-------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
-------------------------------------------------------------------------------
cohort          byte    %15.0g     cohort     Cohort
chpermwage      double  %9.0g                 Estimated permanent log-wage gain
                                                over 18-year period (age 16 to
                                                34)
endeduc         byte    %9.0g                 Number of years of schooling
                                                achieved in last wave
wgt             double  %9.0g                 Sampling weight
-------------------------------------------------------------------------------
Sorted by:
. tabstat chpermwage [aw=wgt], by(cohort) stat(count mean sd med iqr) nototal
Summary for variables: chpermwage
Group variable: cohort (Cohort)
         cohort |         N      Mean        SD       p50       IQR
----------------+--------------------------------------------------
original (1966) |      1834  1.085075  .4831473  1.063587  .5812791
  recent (1979) |      2103  .8782476  .6182544  .8535296  .8001999
----------------------------------------------------------------------
```

Wage growth has been somewhat larger in the original cohort than in the recent cohort. The outcome variable is defined as the difference in (constant dollar) log hourly wages, so a value of 1.085 for the original cohort corresponds to a real wage growth of $\{\exp(1.085)-1\}\times 100 = 196\%$. For the recent cohort, the average is only 0.878 (141%). We can also see that inequality in wage growth has been more pronounced in the recent cohort than in the original cohort because the standard deviation of log wage gains is larger. Looking at the median and IQR instead of the mean and standard deviation leads to qualitatively similar findings.

10. The data have been obtained from http: // www.stat.ucla.edu / ~handcock / RelDist / Data / R / RDBnls.RData.

5.1.1 The relative CDF

The relative CDF of log wage gains between the recent cohort and the original cohort can be obtained as follows, with the graph displayed in figure 0.2:

```
. reldist cdf chpermwage [pw = wgt], by(cohort) notable
Cumulative relative distribution              Number of obs     =      3,937
  F1: cohort = recent (1979)                  Comparison obs    =      2,103
  F0: cohort = original (1966)                Reference obs     =      1,834

. reldist graph, ciopts(fcolor(%50) lcolor(%0))
> xlabel(0(.1)1, grid) xtitle(Proportion of original cohort)
> ylabel(0(.1)1, grid angle(0)) ytitle(Proportion of recent cohort)
> olabel(-.5(.5)3) olabel(.2, at add custom tstyle(minor))
> yolabel(-.5(.5)3, angle(0))
```

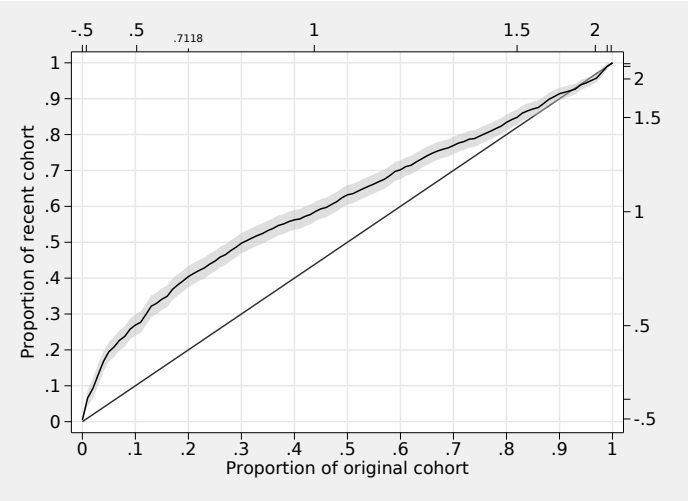

Figure 2

The horizontal axis of the graph corresponds to cumulative proportions of the original cohort, and the vertical axis corresponds to cumulative proportions of the recent cohort; both are ordered by the size of wage growth. Each point on the curve maps quantiles of the two distributions. For example, the value of the 20% quantile in the original cohort is equal to the 40% quantile in the recent cohort because the curve crosses point $(0.2, 0.4)$. The 20% quantile in the original cohort corresponds to a log wage growth of 0.7118, that is, a wage growth of about 104%. In the original cohort, 20% experienced a wage growth of at most 104%; in the recent cohort, this proportion increased to 40%. That is, relative to the original cohort, wage growth of 104% or less is overrepresented by factor 2 in the recent cohort.

Comments on the used commands: Option `notable` has been applied to `reldist cdf` to suppress the output table containing the CDF estimate. By default, the CDF is evaluated at 101 points so that the table would fill a whole page. Here is an example of

how the table looks if we use a reduced set of evaluation points; option `at(0.1(0.1)0.9)` requests 9 evaluation points located at original cohort cumulative proportions 0.1, 0.2, . . . , 0.9:

```
. reldist cdf chpermwage [pw = wgt], by(cohort) at(.1(.1).9)
Cumulative relative distribution                Number of obs     =      3,937
  F1: cohort = recent (1979)                    Comparison obs    =      2,103
  F0: cohort = original (1966)                  Reference obs     =      1,834
------------------------------------------------------------------------------
  chpermwage | Coefficient  Std. err.      [95% conf. interval]
-------------+----------------------------------------------------------------
          p1 |   .2692422   .0152101       .2394219    .2990626
          p2 |     .40432     .01508       .3747547    .4338853
          p3 |   .4973859   .0144863       .4689846    .5257871
          p4 |   .5624279   .0140866       .5348102    .5900456
          p5 |   .6321856   .0138188        .605093    .6592782
          p6 |   .7017939   .0133607       .6755994    .7279883
          p7 |    .769657   .0122928       .7455562    .7937579
          p8 |   .8339943   .0112497       .8119385    .8560501
          p9 |   .9139871   .0086089       .8971088    .9308653
------------------------------------------------------------------------------
(evaluation grid stored in e(at))
```

Coefficient `p2` corresponds to cumulative proportion 0.2; as already discussed, the value of the relative CDF is about 0.4 at this point.

Furthermore, the graph has been produced by first estimating the CDF using `reldist cdf` and then plotting the result using `reldist graph`. We could also have drawn the graph in a single step by including option `graph()` in the call to `reldist cdf` (see examples farther down). Options `olabel()` and `yolabel()` have been applied to `reldist graph` so that additional labels are included in the graph indicating the approximate positions of specific outcome values. Labels are only printed if they are not too close together; the suppressed labels are indicated by additional ticks (this can be changed; see the description of the `olabel()` option above). By default, the values provided in `olabel()` and `yolabel()` are interpreted as outcome values to be included in the graph. However, if suboption `at` is specified, the provided values are interpreted as cumulative proportions; in this case, `reldist graph` will include labels for the corresponding quantiles in the graph. A second `olabel()` option has been used in this way in the command above to print the outcome value of the 20% quantile of the original cohort.[11] Finally, option `ciopts()` has been added to make the confidence area transparent. The options specified within `ciopts()` are standard options for area plots; see [G] ***area_options***.

11. Suboption `add` has been specified in the second `olabel()` option so that the labels from both `olabel()` options are printed, suboption `custom` has been specified to apply custom styling to the second set of labels, and suboption `tstyle(minor)` selects the style. These are standard axis-labeling suboptions; see [G] ***axis_label_options***.

5.1.2 The relative PDF

Relative overrepresentation and underrepresentation of the recent cohort with respect to the distribution of wage growth in the original cohort can be seen more directly in the relative PDF. The relative PDF can be obtained as follows, with the graph displayed in figure 3:

```
. reldist pdf chpermwage [pw = wgt], by(cohort) histogram notable
Relative density                                Number of obs     =      3,937
  F1: cohort = recent (1979)                    Comparison obs    =      2,103
  F0: cohort = original (1966)                  Reference obs     =      1,834
                                                Bandwidth         =  .02710796

. reldist graph, ciopts(fcolor(%50) lcolor(%0))
> olabel(-.5(.5)3, grid) olabel(.2, at add custom tstyle(minor))
> xlabel(0(.1)1) xtitle(Proportion of original cohort)
> ylabel(0(.5)5, angle(0) grid) ytitle(Relative density)
```

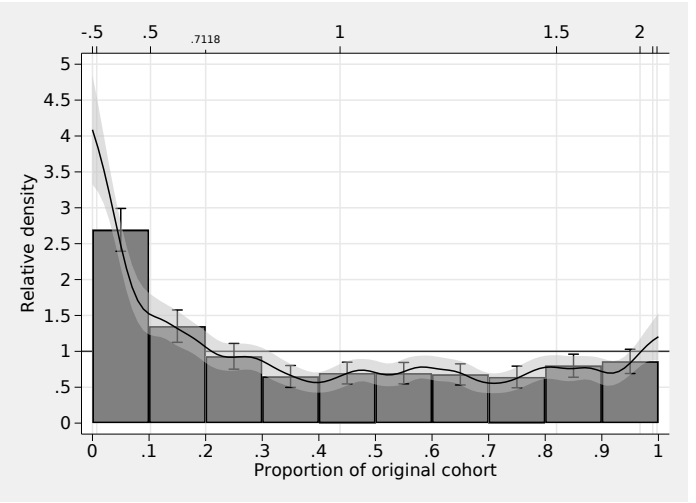

Figure 3

A relative density larger than 1 means the recent cohort is overrepresented at the corresponding level of wage gains, and values lower than 1 mean the recent cohort is underrepresented relative to the original cohort. We can now directly see that the largest distributional differences are at the bottom of the distribution. The recent cohort has a much larger density than the original cohort in regions below the 10% quantile of the original cohort (overrepresentation factor of 1.5 to 4) and generally a larger density below about the 20% quantile. At quantiles above that, the recent cohort is underrepresented, although there is some evidence for a reduced discrepancy at the top of the distribution (above the 80% quantile) or even a reversal at the very top (above, say, the 97% quantile; although the confidence interval includes the parity line in this region, which means that the relative density is not significantly different from 1).

5.1.3 Location and shape decomposition

The difference in the distribution of wage gains between the original cohort and the recent cohort may have various reasons. As indicated above, wage gains have been larger on average in the original cohort than in the recent cohort, which may be because of a general difference in economic growth between the two eras that affected all population members similarly. In such a case, the distribution of wage gains in the recent cohort would differ from the distribution in the original cohort only in its location. However, the structure of wage gains might also have changed, for example, because of rising returns on education, leading to more polarization of wage gains in the recent cohort. In this case, the shape of the two distributions would also be different. To separate location effects from effects of distributional differences net of location, so-called location and shape decompositions can be useful. `reldist` does not perform such decompositions directly, but it offers an option to obtain the relative distribution based on data that have been location- or shape-adjusted.

The following commands produce a graph containing three panels, shown in figure 4.[12] The first panel shows the overall (unadjusted) relative density (same as above). The second panel shows how the relative density looks if we only allow a difference in location but keep the distributional shape fixed. This is achieved by applying option `adjust(:shape scale)`. The option instructs `reldist` to adjust the original cohort distribution such that it has the same shape and scale as the recent cohort distribution but keeps its location. (Technically, this is implemented by applying a location shift to the recent cohort distribution and then replacing the original cohort distribution by this counterfactual distribution; specifying `scale` is necessary because, conceptually, `reldist` treats the scale as a separate element of a distribution that can be adjusted.) The third panel shows the relative density if the location difference between the two distributions is removed but the distributional shapes are allowed to be different. The corresponding option is `adjust(location)`, which shifts the recent cohort distribution such that it has the same location as the original cohort distribution but keeps its shape and scale.[13]

```
. local gropts olabel(-.5(.5)3, grid) histopts(color(%50)) /*
> */ xlabel(0(.2)1) xtitle(Proportion of original cohort) /*
> */ ylabel(0(.5)4, angle(0) grid) ytitle("") noci
. reldist pdf chpermwage [pw = wgt], by(cohort) histogram
> graph(`gropts' title("Overall RD") name(a, replace) nodraw)
  (output omitted)
. reldist pdf chpermwage [pw = wgt], by(cohort) histogram
> adjust(:shape scale)
> graph(`gropts' title("Location shift") name(b, replace) nodraw)
  (output omitted)
```

12. Confidence intervals for the relative density curve have been omitted using graph option `noci` so that the plots are less busy.
13. Handcock and Morris (1999) do the decomposition the other way around, equivalent to specifying `adjust(shape scale)` and `adjust(:location)`.

```
. reldist pdf chpermwage [pw = wgt], by(cohort) histogram
> adjust(location)
> graph(`gropts' title("Shape shift") name(c, replace) nodraw)
  (output omitted)
. graph combine a b c, rows(1) imargin(zero)
```

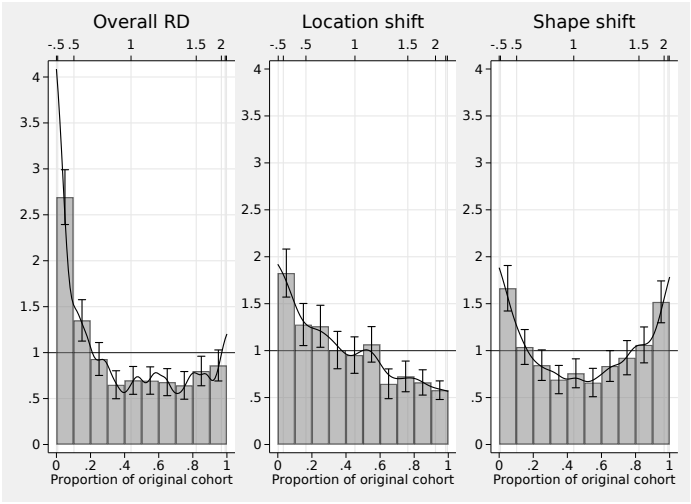

Figure 4

The results indicate that the difference between the recent cohort distribution and the original cohort distribution is not only a matter of location; there is also a substantial difference in distributional shape. In particular, the recent cohort distribution appears more polarized than the original cohort (also see below).

5.1.4 Distributional divergence

To determine the relative contributions of location and shape differences to the overall distributional divergence between the two cohorts, Handcock and Morris (1999) suggest comparing the entropy (Kullback–Leibler divergence) of the unadjusted and adjusted relative distributions. Such an analysis can be obtained by `reldist divergence`:[14]

14. Alternative measures offered by `reldist divergence` are the χ^2 divergence and the dissimilarity index (TVD).

```
. reldist divergence chpermwage [pw = wgt], by(cohort)
> compare(adjust(location))
Relative distribution divergence              Number of obs     =      3,937
  F1: cohort = recent (1979)                  Comparison obs    =      2,103
  F0: cohort = original (1966)                Reference obs     =      1,834
  Adjustment (alternate model)                Histogram bins    =         20
      F1: location                            Statsistic        =    entropy
      F0: (none)
```

chpermwage	Coefficient	Std. err.	[95% conf.	interval]
main	.1726182	.021244	.1309679	.2142686
alternate	.0670518	.0126801	.0421917	.091912
difference	.1055664	.0179497	.0703748	.140758

Three divergence values are reported in the above output: the divergence of the unadjusted relative distribution (labeled as `main`), the divergence of the relative distribution after location-adjusting the recent cohort (labeled as `alternate`), and the difference between these two measures. The first value is the overall divergence, the second value quantifies the divergence because of differences in distributional shape, and the third value quantifies the contribution of the difference in location.[15] We can use `nlcom` (see [R] **nlcom**) to compute the percentage contributions of the location and shape effects to the overall divergence:

```
. nlcom (loc:_b[difference]/_b[main]*100) (shape:_b[alternate]/_b[main]*100)
        loc: _b[difference]/_b[main]*100
      shape: _b[alternate]/_b[main]*100
```

chpermwage	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
loc	61.15599	6.246685	9.79	0.000	48.91271	73.39927
shape	38.84401	6.246685	6.22	0.000	26.60073	51.08729

We see that in this example, the difference in location appears to be more relevant (60%) than the difference in shape (40%). Qualitatively, the results are similar to the ones reported by Handcock and Morris (1999), but note that the precise values are different. Handcock and Morris performed a slightly different decomposition (see footnote 15). More importantly, however, the Kullback–Leibler divergence is quite sensitive to the details of the computation of the underlying relative density. By default, `reldist divergence` obtains the divergence from a 20-bin histogram; changing the number of bins may change the results substantially. Furthermore, the divergence measures could also be obtained from a kernel density estimate of the relative density (see option `pdf`), which would yield yet another set of results (substantially depending on the bandwidth).

15. As discussed above, the last value has a cross-entropy interpretation. Note that `reldist divergence` could also be used to compute alternative decompositions, for example, between the overall relative distribution and a shape-adjusted relative distribution, treating the location effect as a cross-entropy (as in Handcock and Morris [1999]).

5.1.5 Polarization analysis

As stated above, the recent cohort distribution appears more polarized than the original cohort distribution. A measure to quantify the polarization is the MRP computed by `reldist mrp`:

```
. reldist mrp chpermwage [pw = wgt], by(cohort)
Median relative polarization                    Number of obs     =      3,937
  F1: cohort = recent (1979)                    Comparison obs    =      2,103
  F0: cohort = original (1966)                  Reference obs     =      1,834
  Adjustment: location
```

chpermwage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
MRP	.1832597	.0191808	9.55	0.000	.1456544	.220865
LRP	.190353	.0303527	6.27	0.000	.1308445	.2498615
URP	.1761664	.0291428	6.04	0.000	.11903	.2333029

The results indicate that the recent cohort distribution is indeed more polarized because the value of the MRP is positive, of substantial magnitude (the possible range of the MRP is between -1 and 1), and significantly different from 0. Furthermore, the breakup into polarization of the lower half (LRP) and the upper half (URP) of the distribution suggests that the degree of relative polarization is similar in both tails.

5.1.6 Covariate balancing

Education may be one important determinant of the wage distribution as well as the distribution of wage gains over an occupational career. Hence, if the educational distribution changed between the original cohort and the recent cohort, we may be comparing apples with oranges. That is, one reason for the difference in the distribution of wage gains in the two cohorts may be that the cohorts have a different educational composition. This indeed seems to be the case if we look at the relative density of educational levels between the cohorts.[16] The resulting graph is shown in figure 5.

16. Option `categorical` instructs `reldist` to treat `endeduc` as a factor variable and to compute the relative density as the ratio of relative frequencies between the two cohorts at each level. Confidence intervals have been suppressed in the graph using option `noci`.

```
. replace endeduc = 8 if endeduc<8
(34 real changes made)
. reldist hist endeduc [pw = wgt], by(cohort) categorical
Relative histogram                              Number of obs     =       3,937
  F1: cohort = recent (1979)                    Comparison obs    =       2,103
  F0: cohort = original (1966)                  Reference obs     =       1,834
```

endeduc	Coefficient	Std. err.	[95% conf.	interval]
endeduc				
8	.9383436	.220892	.50527	1.371417
9	1.485883	.3551772	.7895346	2.182232
10	1.59819	.3734487	.8660189	2.330361
11	.9276922	.1673159	.5996581	1.255726
12	1.41295	.0657943	1.283956	1.541945
13	1.012919	.1136963	.7900094	1.235828
14	.7208455	.0737943	.5761668	.8655241
15	.6660931	.0984461	.473083	.8591032
16	.8683801	.0644533	.7420152	.994745
17	.5069374	.0751882	.3595259	.6543489
18	.7644302	.0823954	.6028885	.9259719

```
(evaluation grid stored in e(at))
. reldist graph, noci olabel(8(1)18, prune(.05)) color(%50)
> xlabel(0(.1)1) xtitle(Proportion of original cohort)
> ylabel(0(.2)1.6, angle(0) grid) ytitle(Relative density)
```

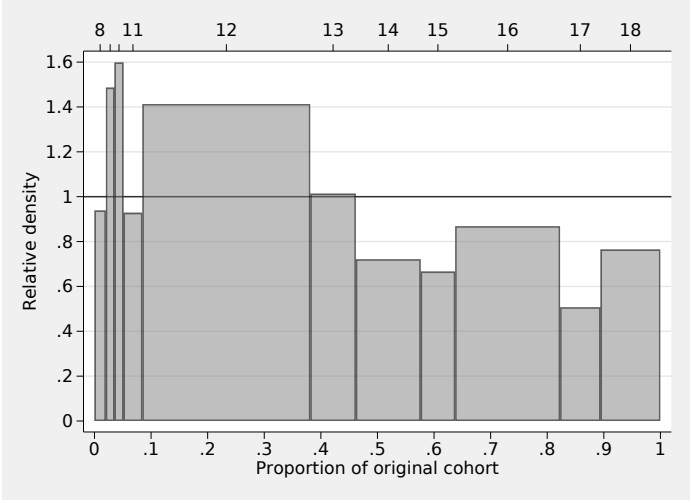

Figure 5

Lower educational levels appear to be more frequent in the recent cohort than in the original cohort (relative density mostly larger than 1), and higher educational levels appear to be less frequent (relative density below 1). Looking at the table, we see that in many cases the confidence interval does not include 1, meaning that these differences between the cohorts are statistically significant.

The question now is whether these differences in educational composition affect the relative distribution of wage gains. Similarly to above in the context of location and shape effects, we can identify the contribution of compositional differences by comparing unadjusted and adjusted relative distributions. The adjustment, however, is now accomplished by reweighting one of the distributions such that its educational composition becomes equal to the educational composition in the other cohort. Option `balance()` can be used in `reldist` to apply such balancing. Here is an example (graph in figure 6) that displays the overall relative distribution (left panel), the relative distribution after the recent cohort has been reweighted (right panel), and the relative distribution between the raw and reweighted recent cohort (middle panel; the purpose of the middle panel is to show how reweighting changes the distribution of the recent cohort):

```
. local gropts olabel(-.5(.5)3, grid) histopts(color(%50)) /*
>  */ xlabel(0(.2)1) xtitle(Proportion of original cohort) /*
>  */ ylabel(0(.5)4, angle(0) grid) ytitle("") noci
. reldist pdf chpermwage [pw = wgt], by(cohort) histogram
> graph(`gropts' title("Overall RD") name(a, replace) nodraw)
  (output omitted)
. reldist pdf chpermwage [pw = wgt], by(cohort) histogram
> balance(i.endeduc, contrast)
> graph(`gropts' title("Education effect") name(b, replace) nodraw)
  (output omitted)
. reldist pdf chpermwage [pw = wgt], by(cohort) histogram
> balance(i.endeduc)
> graph(`gropts' title("Education-adjusted RD") name(c, replace) nodraw)
  (output omitted)
. graph combine a b c, rows(1) imargin(zero)
```

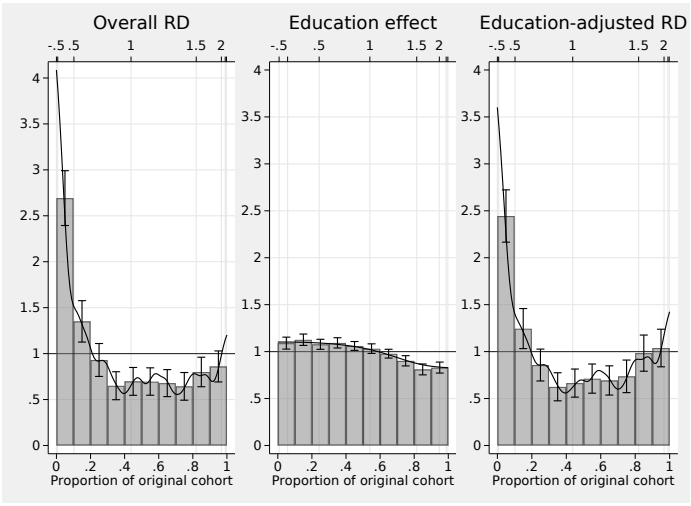

Figure 6

Adjusting the educational composition does seem to make the distribution of wage gains somewhat more equal between the two cohorts. The comparison between the raw recent cohort and the reweighted recent cohort (middle panel) shows that low (high) wage gains are more (less) frequent in the raw data than in the reweighted data. That is, as expected, reweighting the recent cohort generally shifts the distribution of wage gains upward, thus making it more equal to the distribution of wage gains in the original cohort (the effect of the reweighting is statistically significant, as can be inferred from the confidence intervals that have been included for the histogram). Overall, however, the contribution of the difference in educational composition only seems to be of minor importance: there is only a small difference between the overall relative distribution (left panel) and the education-adjusted relative distribution (right panel).

5.1.7 Location adjustment by means of covariate balancing

Note that reweighting can be used as an alternative method for location adjustments. The default method, provided by option `adjust()`, implements the adjustments by transforming the outcome values. The same goal, however, can also be reached by altering the PDF of the data while leaving the outcome values unchanged. This is what reweighting does if we include the outcome variable in the balancing equation. Here is a replication of the location and shape decomposition from above using `balance()` instead of `adjust()`. I use entropy balancing to obtain the weights, which ensures that the means of the two distributions will be exactly the same. The graph is shown in figure 7.

```
. local gropts olabel(-.5(.5)3, grid) histopts(color(%50)) /*
>  */ xlabel(0(.2)1) xtitle(Proportion of original cohort) /*
>  */ ylabel(0(.5)4, angle(0) grid) ytitle("") noci
. reldist pdf chpermwage [pw = wgt], by(cohort) histogram
> graph(`gropts' title("Overall RD") name(a, replace) nodraw)
  (output omitted)
. reldist pdf chpermwage [pw = wgt], by(cohort) histogram
> balance(eb: chpermwage, contrast)
> graph(`gropts' title("Location shift") name(b, replace) nodraw)
  (output omitted)
. reldist pdf chpermwage [pw = wgt], by(cohort) histogram
> balance(eb: chpermwage)
> graph(`gropts' title("Shape shift") name(c, replace) nodraw)
  (output omitted)
. graph combine a b c, rows(1) imargin(zero)
```

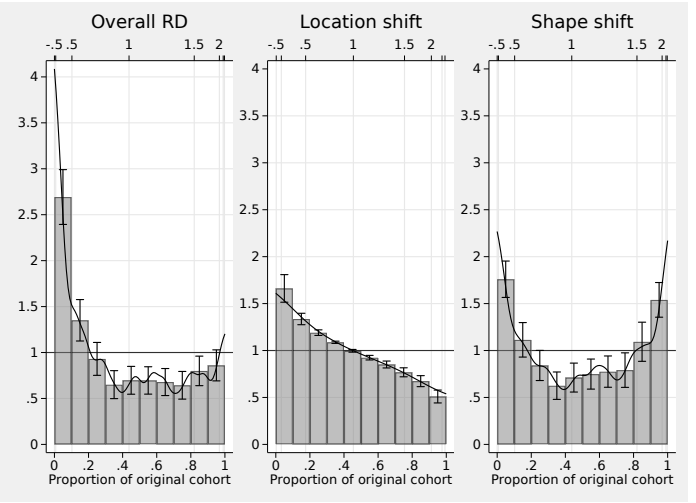

Figure 7

The two approaches lead to qualitatively similar results.[17] One advantage of the reweighting approach, however, is that heaping in the data will have fewer adverse effects on the results.[18]

5.2 Processing results from reldist

5.2.1 Postestimation hypothesis testing

`reldist` stores its results in `e()` just like any other estimation command. Hence, we can use postestimation commands such as `test` (see [R] **test**) to test hypotheses, or we can use `coefplot` (Jann 2014) to draw graphs.

I use the National Longitudinal Study of Young Women 1988 data shipped with Stata to analyze wages of unionized and nonunionized workers. For example, we might be interested in relative wage polarization. An obvious hypothesis is that wages are more polarized among nonunionized workers than among the unionized, but the pattern may be different depending on education. Here are the results for the MRP between nonunionized and unionized workers for different levels of qualification:

17. Although, note that `adjust()`, as used above, adjusts the medians of the distributions, whereas `balance()` adjusts the means. For a more valid comparison, suboption `mean` could be specified within `adjust()`.
18. Note that reweighting could be used for location and scale adjustment by including the square of the outcome variable as an additional covariate in the balancing equation.

```
. sysuse nlsw88, clear
(NLSW, 1988 extract)
. reldist mrp wage, by(union) swap over(collgrad) multiplicative
Median relative polarization                  Number of obs     =      1,878
  F1: union = Nonunion                        Comparison obs    =      1,417
  F0: union = Union                           Reference obs     =        461
  Adjustment: location (mult)
           0: collgrad = Not college grad
           1: collgrad = College grad
```

wage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
0						
MRP	.0654444	.0358179	1.83	0.068	-.0048027	.1356916
LRP	-.0015699	.0572336	-0.03	0.978	-.113818	.1106783
URP	.1324587	.0571956	2.32	0.021	.020285	.2446324
1						
MRP	.1486059	.0591766	2.51	0.012	.032547	.2646647
LRP	.1985118	.0818497	2.43	0.015	.0379858	.3590378
URP	.0987	.0920773	1.07	0.284	-.0818847	.2792846

Option `swap` has been specified to flip the two groups so that the nonunionized are the comparison group and the unionized are the reference group. The option `multiplicative` has been specified because—based on economic theory—a proportional location shift makes more sense for wages than an additive shift. As hypothesized, the results suggest that wage polarization is generally more pronounced among nonunionized workers, although the MRP is only marginally significant for respondents without a college degree. A follow-up question might thus be whether we can conclude from the data that relative polarization between nonunionized and unionized workers is stronger among college graduates than among workers without a college degree. We can use `test` to test the two MRP estimates against each other:

```
. test [0]MRP = [1]MRP
 ( 1)  [0]MRP - [1]MRP = 0
       F(  1,  1877) =    1.45
            Prob > F =    0.2294
```

The test is negative; that is, we cannot reject the null hypothesis that the two MRP estimates are the same (p-value of 0.229). The same result could also be obtained using `lincom` (see [R] **lincom**) instead of `test`.

5.2.2 Creating graphs from multiple results

When comparing wages between unionized and nonunionized workers, it may be relevant to make the two groups more comparable by taking background characteristics into account. Possibly, some of the difference in the wage distributions is because of differential composition with respect to these characteristics and not because of unionization status per se. Here is how you could plot the relative density curves based on

raw data and on balanced data in a single graph (figure 8) using `estimates store` (see [R] **estimates store**) and `coefplot` (Jann 2014):

```
. reldist pdf wage, by(union) notable balance(grade i.race i.south tenure)
  (output omitted)
. estimates store balanced
. reldist pdf wage if e(sample), by(union) notable
  (output omitted)
. estimates store unbalanced
. coefplot balanced unbalanced, at recast(line)
> ciopts(recast(rarea) color(%50) lcolor(%0))
> xtitle("Proportion of non-unionized workers")
> ytitle("Relative density") yline(1)
```

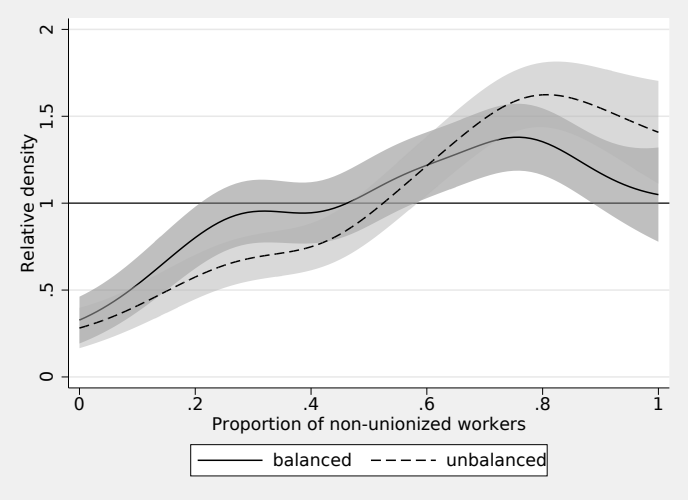

Figure 8

We see that the wage distributions of unionized and nonunionized workers become more similar once we control for background characteristics, especially in the upper part of the distribution.

5.2.3 Working with IFs

The `predict` command can be used to store the IFs that `reldist` uses for standard error estimation. For example, we may want to test whether relative polarization between nonunionized and unionized workers is more pronounced for wages than for working hours. `reldist` does not support analyzing two variables at the same time. However, we can store the IFs and then use them to test the MRP for wages against the MRP for working hours:

```
. reldist mrp wage if hours<., by(union) swap multiplicative
Median relative polarization                    Number of obs     =      1,877
  F1: union = Nonunion                          Comparison obs    =      1,416
  F0: union = Union                             Reference obs     =        461
  Adjustment: location (mult)
```

wage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
MRP	.123268	.0303101	4.07	0.000	.063823	.182713
LRP	.0573649	.0494239	1.16	0.246	-.0395667	.1542964
URP	.1891712	.0482137	3.92	0.000	.094613	.2837294

```
. predict MRPwage
. replace MRPwage = MRPwage + _b[MRP] / e(N)
(1,877 real changes made)
. reldist mrp hours if wage<., by(union) swap
Median relative polarization                    Number of obs     =      1,877
  F1: union = Nonunion                          Comparison obs    =      1,416
  F0: union = Union                             Reference obs     =        461
  Adjustment: location
```

hours	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
MRP	.0712359	.0261141	2.73	0.006	.0200202	.1224516
LRP	.1601944	.0644048	2.49	0.013	.0338818	.286507
URP	-.0177227	.0421322	-0.42	0.674	-.1003535	.0649082

```
. predict MRPhours
. replace MRPhours = MRPhours + _b[MRP] / e(N)
(1,877 real changes made)
. total MRPwage MRPhours
Total estimation                        Number of obs = 1,877
```

	Total	Std. err.	[95% conf.	interval]
MRPwage	.123268	.0303101	.063823	.182713
MRPhours	.0712359	.0261141	.0200202	.1224516

```
. lincom MRPwage - MRPhours
 ( 1)  MRPwage - MRPhours = 0
```

Total	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
(1)	.0520321	.0378415	1.38	0.169	-.0221838	.1262481

```
. drop MRPwage MRPhours
```

The MRP is higher for wages than for working hours, but the difference does not appear to be statistically significant. In the example, I first stored the IFs and then recentered them by adding the point estimates back in (on the use of recentered IFs, see, for example, Firpo, Fortin, and Lemieux [2009] and Rios-Avila [2020]). The IFs returned by `reldist` are scaled such that `total` (see [R] **total**) can be used for estimation of standard errors (note how `total` reproduced the results from `reldist` in the example).

This is why I divided the point estimate by N before adding it back in. Alternatively, multiply the IF by N, add the point estimate as is, and then use `mean` (see [R] **mean**) instead of `total`. Furthermore, note that weights are not incorporated into the IFs. That is, if weights have been applied to `reldist`, the weights will also have to be applied when calling `total` or `mean` (the same is true for clustering).

5.3 Survey estimation

`reldist` fully supports estimation for complex survey data, but the `svy` prefix command (see [SVY] **svy**) cannot be used for technical reasons if the variance estimation method is set to `linearized` (Taylor-linearized variance estimation). You can use option `vce(svy)` instead of the `svy` prefix in this case. Here is an example:

```
. webuse nmihs, clear
. svyset [pweight=finwgt], strata(stratan)
Sampling weights: finwgt
             VCE: linearized
     Single unit: missing
        Strata 1: stratan
 Sampling unit 1: <observations>
           FPC 1: <zero>
. reldist mrp birthwgt, by(childsex) vce(svy)
(running reldist_svyr on estimation sample)
Survey: Median relative polarization
Number of strata =     6                        Number of obs     =      9,946
Number of PSUs   = 9,946                        Population size   =  3,895,562
                                                Design df         =      9,940
  F1: childsex = 2                              Comparison obs    =      4,911
  F0: childsex = 1                              Reference obs     =      5,035
------------------------------------------------------------------------------
             |             Linearized
   birthwgt | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
------------------------------------------------------------------------------
         MRP |  -.0349405   .0155133    -2.25   0.024    -.0653496   -.0045313
         LRP |   .0024726   .0233231     0.11   0.916    -.0432454    .0481907
         URP |  -.0723535   .0252147    -2.87   0.004    -.1217795   -.0229275
------------------------------------------------------------------------------
```

Results indicate that the birthweight distribution is somewhat less polarized for girls (`childsex = 2`) than for boys (`childsex = 1`) and that this is because of a difference in distributional shape in the upper part of the distribution (overall relative polarization is driven by the URP). Option `vce(svy)` also works with variance estimation methods other than `linearized` (for example, see [SVY] **svy brr**), although in these cases one could also apply `svy` as a prefix command.[19]

19. A fine distinction is that with `vce(svy)`, the bandwidth for kernel density estimation (relevant for `reldist pdf` and `reldist divergence` with option `pdf`) will only be estimated once and then held constant across replications. With `svy` as a prefix command, bandwidth estimation will be repeated in each replication.

6 Acknowledgments

I thank Eric Melse for valuable comments on earlier versions of the software that helped improve the command, Blaise Melly for a nudge on how to obtain the IFs for quantiles of relative ranks, and Philippe Van Kerm for pointing me to the work by Joseph Gastwirth. Furthermore, I thank an anonymous reviewer for comments that helped improve the article.

7 Programs and supplemental materials

To install a snapshot of the corresponding software files as they existed at the time of publication of this article, type

```
. net sj 21-4
. net install st0656     (to install program files, if available)
. net get st0656         (to install ancillary files, if available)
```

8 References

Abramson, I. S. 1982. On bandwidth variation in kernel estimates—A square root law. *Annals of Statistics* 10: 1217–1223. https://doi.org/10.1214/aos/1176345986.

Alderson, A. S., J. Beckfield, and F. Nielsen. 2005. Exactly how has income inequality changed?: Patterns of distributional change in core societies. *International Journal of Comparative Sociology* 46: 405–423. https://doi.org/10.1177/0020715205059208.

Bernhardt, A., M. Morris, and M. 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: 302–328. https://doi.org/10.1086/230726.

Bernhardt, A., M. Morris, M. S. Handcock, and M. A. Scott. 2001. *Divergent Paths. Economic Mobility in the New American Labor Market*. New York: Russell Sage Foundation.

Bliege Bird, R., D. W. Bird, B. F. Codding, C. H. Parker, and J. H. Jones. 2008. The "fire stick farming" hypothesis: Australian Aboriginal foraging strategies, biodiversity, and anthropogenic fire mosaics. *Proceedings of the National Academy of Sciences of the United States of America* 105: 14796–14801. https://doi.org/10.1073/pnas.0804757105.

Botev, Z. I., J. F. Grotowski, and D. P. Kroese. 2010. Kernel density estimation via diffusion. *Annals of Statistics* 38: 2916–2957. https://doi.org/10.1214/10-AOS799.

Chernozhukov, V., I. Fernández-Val, and B. Melly. 2013. Inference on counterfactual distributions. *Econometrica* 81: 2205–2268. https://doi.org/10.3982/ECTA10582.

Clementi, F., V. Molini, and F. Schettino. 2018. All that glitters is not gold: Polarization amid poverty reduction in Ghana. *World Development* 102: 275–291. https://doi.org/10.1016/j.worlddev.2017.07.019.

Cox, N. J. 1998. distplot: Stata module to generate distribution function plot. Statistical Software Components S337502, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s337502.html.

———. 2004. ppplot: Stata module for P-P plots. Statistical Software Components S438002, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s438002.html.

Ćwik, J., and J. Mielniczuk. 1989. Estimating density ratio with application to discriminant analysis. *Communications in Statistics—Theory and Methods* 18: 3057–3069. https://doi.org/10.1080/03610928908830077.

———. 1993. Data-dependent bandwidth choice for a grade density kernel estimate. *Statistics & Probability Letters* 16: 397–405. https://doi.org/10.1016/0167-7152(93)90074-S.

Del Giudice, M. 2011. Sex differences in romantic attachment: A meta-analysis. *Personality and Social Psychology Bulletin* 37: 193–214. https://doi.org/10.1177/0146167210392789.

Deville, J.-C. 1999. Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology* 25: 193–203.

DiNardo, J., N. M. Fortin, and T. Lemieux. 1996. Labor market institutions and the distribution of wages, 1973–1992: A semiparametric approach. *Econometrica* 64: 1001–1044. https://doi.org/10.2307/2171954.

Duncan, O. D., and B. Davis. 1953. An alternative to ecological correlation. *American Sociological Review* 18: 665–666. https://doi.org/10.2307/2088122.

Eggers, A. C., and A. Spirling. 2016. Party cohesion in Westminster systems: Inducements, replacement and discipline in the House of Commons, 1836–1910. *British Journal of Political Science* 46: 567–589. https://doi.org/10.1017/S0007123414000362.

Firpo, S., N. M. Fortin, and T. Lemieux. 2009. Unconditional quantile regressions. *Econometrica* 77: 953–973. https://doi.org/10.3982/ECTA6822.

Gastwirth, J. L. 1975. Statistical measures of earnings differentials. *American Statistician* 29: 32–35. https://doi.org/10.2307/2683677.

Hampel, F. R. 1974. The influence curve and its role in robust estimation. *Journal of the American Statistical Association* 69: 383–393. https://doi.org/10.2307/2285666.

Handcock, M. S. 2016. *reldist: Relative Distribution Methods*. R package version 1.6-6. https://CRAN.R-project.org/package=reldist.

Handcock, M. S., and E. M. Aldrich. 2002. Applying relative distribution methods in R. Working Paper No. 27, University of Washington. https://dx.doi.org/10.2139/ssrn.1515775.

Handcock, M. S., and P. L. Janssen. 2002. Statistical inference for the relative density. *Sociological Methods & Research* 30: 394–424. https://doi.org/10.1177/0049124102030003005.

Handcock, M. S., and M. Morris. 1998. Relative distribution methods. *Sociological Methodology* 28: 53–97. https://doi.org/10.1111/0081-1750.00042.

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

Hao, L., and D. Q. Naiman. 2010. *Assessing Inequality*. Thousand Oaks, CA: SAGE.

Jann, B. 2004. duncan: Stata module to calculate dissimilarity index. Statistical Software Components S447202, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s447202.html.

———. 2005. moremata: Stata module (Mata) to provide various functions. Statistical Software Components S455001, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s455001.html.

———. 2007. Univariate kernel density estimation. Technical report, online publication. http://fmwww.bc.edu/repec/bocode/k/kdens.pdf.

———. 2008. The Blinder–Oaxaca decomposition for linear regression models. *Stata Journal* 8: 453–479. https://doi.org/10.1177/1536867X0800800401.

———. 2014. Plotting regression coefficients and other estimates. *Stata Journal* 14: 708–737. https://doi.org/10.1177/1536867X1401400402.

———. 2017. kmatch: Stata module for multivariate-distance and propensity-score matching, including entropy balancing, inverse probability weighting, (coarsened) exact matching, and regression adjustment. Statistical Software Components S458346, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s458346.html.

———. 2020. Influence functions continued. A framework for estimating standard errors in reweighting, matching, and regression adjustment. Working Papers 35, University of Bern Social Sciences. https://ideas.repec.org/p/bss/wpaper/35.html.

Le Breton, M., A. Michelangeli, and E. Peluso. 2012. A stochastic dominance approach to the measurement of discrimination. *Journal of Economic Theory* 147: 1342–1350. https://doi.org/10.1016/j.jet.2011.05.003.

Morris, M., A. D. Bernhardt, and M. S. Handcock. 1994. Economic inequality: New methods for new trends. *American Sociological Review* 59: 205–219. https://doi.org/10.2307/2096227.

Panek, T., and J. Zwierzchowski. 2020. Median relative partial income polarization indices: Investigating economic polarization in Poland during the years 2005–2015. *Social Indicators Research* 149: 1025–1044. https://doi.org/10.1007/s11205-020-02274-2.

Parzen, E. 2004. Quantile probability and statistical data modeling. *Statistical Science* 19: 652–662. https://doi.org/10.1214/088342304000000387.

Reardon, S. F., and J. B. Townsend. 1999. seg: Stata module to compute multiple-group diversity and segregation indices. Statistical Software Components S375001, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s375001.html.

Rios-Avila, F. 2020. Recentered influence functions (RIFs) in Stata: RIF regression and RIF decomposition. *Stata Journal* 20: 51–94. https://doi.org/10.1177/1536867X20909690.

About the author

Ben Jann is a professor of sociology at the University of Bern, Switzerland. His research interests include social science methodology, statistics, social stratification, and labor market sociology. He is the principal investigator of TREE, a large-scale multicohort panel study in Switzerland on transitions from education to employment (http://www.tree.unibe.ch).

A Appendix

A.1 Variance estimation by means of influence functions (IFs)

Influence functions (IFs; Hampel 1974) provide a convenient approach to estimate the sampling variances of the different statistics discussed above. Intuitively, an IF is an approximation of how a function of a distribution changes once some data mass is added at a specific point in the distribution. Random sampling can be seen as a process that modifies the distribution in such a way and hence leads to variation in statistics computed from the distribution. It can be shown that, asymptotically, this variation (that is, the sampling variance) is equal to the expectation of the square of the IF divided by the sample size (for example, Deville [1999]). Therefore, to obtain an estimate of the sampling variance from a given sample, we can evaluate the IF at each observation in the data and then compute the sampling variance of the mean of these values using textbook formulas.[20] More generally, once IFs are available for a set of statistics, the variance matrix of these statistics can be obtained by taking a mean estimate (using `mean`) of the IFs (or a total estimate using `total`, depending on the scaling of the IFs). Sampling weights or other complex survey characteristics do not change the form of the IF and can be accounted for when computing the mean (or total) estimate. This makes the IF approach very general.

A.1.1 One-parameter setting

There is a close connection between IFs and the method of moments (see Jann [2020]). Let h_i^θ be the moment condition for estimating θ in a simple one-parameter setting, such that $\widehat{\theta}$ satisfies

$$0 = \frac{1}{W}\sum_{i=1}^{n} w_i \widehat{h}_i^\theta$$

where $\widehat{h}_i^\theta$ denotes h_i^θ with θ set to $\widehat{\theta}$. Observation i's value of the empirical IF of $\widehat{\theta}$ can then be obtained as

$$\text{IF}_i(\widehat{\theta}) = \frac{1}{-\widehat{\Delta}^\theta}\widehat{h}_i^\theta$$

where

$$\widehat{\Delta}^\theta = \frac{1}{W}\sum_{i=1}^{n} w_i \left.\frac{\partial h_i^\theta}{\partial\theta}\right|_{\theta=\widehat{\theta}}$$

is an estimate of the expectation of the derivative of h^θ at point $\theta = \widehat{\theta}$. Consider the mean estimator

$$\widehat{\overline{y}} = \frac{1}{W}\sum_{i=1}^{n} w_i Y_i$$

for which the moment condition is given as

$$h_i^{\overline{y}} = Y_i - \overline{y}$$

20. Because the mean of an IF is 0 by definition, the expectation of the squared IF is equal to the variance of the IF.

Because

$$\widehat{\Delta}^{\overline{y}} = \frac{1}{W}\sum_{i=1}^{n} w_i \left.\frac{\partial h_i^{\overline{y}}}{\partial \overline{y}}\right|_{\overline{y}=\widehat{\overline{y}}} = \frac{1}{W}\sum_{i=1}^{n} w_i(-1) = -1$$

the IF simplifies to

$$\text{IF}_i(\widehat{\overline{y}}) = Y_i - \widehat{\overline{y}}$$

Assuming a survey design without clustering or stratification, the sampling variance of $\widehat{\overline{y}}$ can then be estimated as

$$\widehat{V}(\widehat{\overline{y}}) = \frac{1}{W(W-W/n)}\sum_{i=1}^{n} w_i^2 \left\{\text{IF}_i(\widehat{\overline{y}})\right\}^2$$

This is equivalent to the textbook formula for the variance of the mean, as can easily be seen if $\text{IF}_i(\widehat{\overline{y}})$ is replaced by its definition. The general point is that we can use the same variance formula also in other situations. That is, the variance of a statistic can be obtained by applying the above formula (or a variant of it depending on survey design) to its IF, whatever the statistic might be.

A.1.2 Multiple-parameter setting

Deriving the IF becomes more involved if a statistic includes auxiliary parameters that are estimated from the data. Think of a system of equations with moment conditions $h^{\theta_1}, h^{\theta_2}, \ldots, h^{\theta_p}$ where θ_1 depends on $\theta_2, \ldots, \theta_p$ (that is, all θ_j appear as arguments in the moment condition for θ_1). The IF for θ_1 can then be written as

$$\text{IF}_i(\widehat{\theta}_1) = \frac{1}{-\widehat{\Delta}^{\theta_1}}\left\{\widehat{h}_i^{\theta_1} + \sum_{j=2}^{p}\widehat{\Delta}_{\theta_j}^{\theta_1}\text{IF}_i(\widehat{\theta}_j)\right\} \tag{16}$$

where $\widehat{h}_i^{\theta_1}$ denotes the value of $h_i^{\theta_1}$ with $\theta = (\theta_1, \ldots, \theta_p)$ set to $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_p)$ and

$$\widehat{\Delta}_{\theta_j}^{\theta_1} = \frac{1}{W}\sum_{i=1}^{n} w_i \left.\frac{\partial h_i^{\theta_1}}{\partial \theta_j}\right|_{\theta=\widehat{\theta}}$$

is an estimate of the expectation of the partial derivative of h^{θ_1} by θ_j at point $\widehat{\theta}$. If parameters θ_j, $j \geq 2$, themselves depend on further parameters, their IFs will have an analogous form. That is, multiple-parameter problems can be solved recursively by applying (16) repeatedly.

One implication of (16) is that, if $\gamma = t(\theta)$, where $t(\theta)$ is a simple transformation function of $\theta = (\theta_1, \ldots, \theta_p)$ that does not involve the data (that is, a linear or nonlinear combination of the elements in θ), the IF for $\widehat{\gamma}$ can be written as

$$\text{IF}_i(\widehat{\gamma}) = \left.\frac{\partial t(\theta)}{\partial \theta_1}\right|_{\theta=\widehat{\theta}} \text{IF}_i(\widehat{\theta}_1) + \cdots + \left.\frac{\partial t(\theta)}{\partial \theta_p}\right|_{\theta=\widehat{\theta}} \text{IF}_i(\widehat{\theta}_p)$$

This means that the IF of a statistic that is defined as an aggregate of other statistics can be obtained as an aggregate of the IFs of these statistics.

A.1.3 Subpopulation estimation

Because the relative size of a subsample is subject to sampling error, IFs should always be evaluated for all observations in the data and also when only a subpopulation is analyzed (even though for many statistics, this may not change the results). Furthermore, the relative distribution is typically computed using data from two subsamples so that the IFs below will inherently contain multiple components based on different observations. Using a full-sample approach is thus inevitable. Subpopulation IFs defined in terms of all observations can be obtained by including appropriate subpopulation indicators in the moment conditions.

Consider the IF for subpopulation mean

$$\widehat{\overline{y}}_S = \frac{1}{W_S} \sum_{i \in \mathcal{S}} w_i S_i Y_i$$

where S_i is an indicator for whether observation i belongs to subsample $\mathcal{S}$ and W_S is the sum of weights in the subsample. The full-sample moment condition for $\widehat{\overline{y}}_S$ can be written as

$$h_i^{\overline{y}_S} = S_i(Y_i - \overline{y}_S)$$

Because

$$\widehat{\Delta}^{\overline{y}_S} = \frac{1}{W} \sum_{i=1}^{n} w_i(-S_i) = -\frac{W_S}{W}$$

the IF for $\widehat{\overline{y}}_S$ becomes

$$\text{IF}_i(\widehat{\overline{y}}_S) = \frac{W}{W_S} S_i(Y_i - \widehat{\overline{y}}_S)$$

The IF will be 0 for observations outside subsample $\mathcal{S}$. However, taking the standard error of the mean of this IF across all observations will provide a consistent standard error for $\widehat{\overline{y}}_S$.

In practice, it may be convenient to omit the global W from the definition of the IF and only divide by the relevant subpopulation size. In this case, the appropriate standard error is provided by the standard error of the total of the IF. An advantage of defining IFs in this way is that they can be computed from the subsample data alone without knowing the total sum of weights.

A.1.4 Overview of IFs for various statistics

Using the methods above, we can obtain IFs for various statistics that are relevant in the context of relative distribution analysis. Table 1 provides an overview. The IFs have been derived for statistics that are conditional on $S = 1$, where S is an indicator for whether an observation belongs to subsample $\mathcal{S}$. For unconditional statistics, set S to 1 for all observations.

Table 1. IFs for various statistics

Statistic (conditional on $S = 1$)	Empirical IF
Mean $\widehat{\bar{y}}_S = \frac{1}{W_S}\sum_{i\in\mathcal{S}} w_i Y_i$	$\text{IF}_i = \frac{W}{W_S} S_i\Big(Y_i - \widehat{\bar{y}}_S\Big)$
Empirical CDF $\widehat{F}_{Y\mid S}(y) = \frac{1}{W_S}\sum_{i\in\mathcal{S}} w_i \mathbb{1}(Y_i \le y)$	$\text{IF}_i = \frac{W}{W_S} S_i\Big\{\mathbb{1}(Y_i \le y) - \widehat{F}_{Y\mid S}(y)\Big\}$
Kernel PDF (assuming h fixed) $\widehat{f}_{Y\mid S}(y) = \frac{1}{W_S}\sum_{i\in\mathcal{S}} w_i \frac{1}{h} K\left(\frac{y - Y_i}{h}\right)$	$\text{IF}_i = \frac{W}{W_S} S_i\left\{\frac{1}{h} K\left(\frac{y - Y_i}{h}\right) - \widehat{f}_{Y\mid S}(y)\right\}$
Histogram PDF $\widehat{f}_{Y\mid S}(a,b) = \frac{1}{W_S}\sum_{i\in\mathcal{S}} w_i \frac{\mathbb{1}(a < Y_i \le b)}{b-a}$	$\text{IF}_i = \frac{W}{W_S} S_i\left\{\frac{\mathbb{1}(a < Y_i \le b)}{b-a} - \widehat{f}_{Y\mid S}(a,b)\right\}$
Quantile $\widehat{q}_{Y\mid S}(p) = \widehat{F}^{-1}_{Y\mid S}(p)$	$\text{IF}_i = \frac{W}{W_S} S_i\left[\frac{p - \mathbb{1}\big(Y_i \le \widehat{q}_{Y\mid S}(p)\big)}{\widehat{f}_{Y\mid S}\{\widehat{q}_{Y\mid S}(p)\}}\right]$
Median $\widehat{\tilde{y}}_S = \widehat{q}_{Y\mid S}(0.5)$	$\text{IF}_i = \frac{W}{W_S} S_i\left\{\frac{0.5 - \mathbb{1}(Y_i \le \widehat{\tilde{y}}_S)}{\widehat{f}_{Y\mid S}(\widehat{\tilde{y}}_S)}\right\}$
Variance $\widehat{\sigma}^2_{Y\mid S} = \frac{1}{W_S - \frac{W_S}{n_S}}\sum_{i\in\mathcal{S}} w_i (Y_i - \widehat{\bar{y}}_S)^2$	$\text{IF}_i = \frac{W}{W_S} S_i\Big\{c \times (Y_i - \widehat{\bar{y}}_S)^2 - \widehat{\sigma}^2_{Y\mid S}\Big\},\ c = \frac{1}{1-\frac{1}{n_S}}$
Standard deviation $\widehat{\sigma}_{Y\mid S} = \sqrt{\widehat{\sigma}^2_{Y\mid S}}$	$\text{IF}_i = \frac{W}{W_S} S_i\left\{\frac{c \times (Y_i - \widehat{\bar{y}}_S)^2 - \widehat{\sigma}^2_{Y\mid S}}{2\widehat{\sigma}_{Y\mid S}}\right\},\ c = \frac{1}{1-\frac{1}{n_S}}$
IQR $\text{IQR}_{Y\mid S} = \widehat{q}_{Y\mid S}(0.75) - \widehat{q}_{Y\mid S}(0.25)$	$\text{IF}_i = \text{IF}_i\{\widehat{q}_{Y\mid S}(0.75)\} - \text{IF}_i\{\widehat{q}_{Y\mid S}(0.25)\}$

The variance is an example of a multiparameter statistic, but it is a special case because the IF for the auxiliary parameter drops out of the equation. Furthermore, note that a quantile can be defined as the value $\widehat{q}_Y(p)$ that solves

$$0 = \frac{1}{W}\sum_{i=1}^{n} w_i[\mathbb{1}(Y_i \leq q_Y(p)) - p]$$

so that $\widehat{f}_Y\{\widehat{q}_Y(p)\}$ provides an estimate of the expectation of the derivative of the relevant moment condition.

A.2 IF for the relative CDF

Using notation as introduced in section 3, the empirical CDF of the relative ranks can be written as

$$\widehat{G}(r) = \widehat{F}_{X|D}(\widehat{q}_r) = \frac{1}{W_D}\sum_{i\in\mathcal{D}} w_i \mathbb{1}(X_i \leq \widehat{q}_r)$$

where $\widehat{q}_r$ is shorthand notation for $\widehat{q}_{Y|R}(r) = \widehat{F}^{-1}_{Y|R}(r)$. The moment conditions for $G(r)$ and q_r are

$$\begin{aligned} h_i^G &= D_i\{\mathbb{1}(X_i \leq q_r) - G(r)\} \\ h_i^q &= R_i\{\mathbb{1}(Y_i \leq q_r) - r\} \end{aligned}$$

Working through (16) yields

$$\text{IF}_i\left\{\widehat{G}(r)\right\} = \frac{W}{W_D}\widehat{h}_i^G + \widehat{f}_{X|D}(\widehat{q}_r)\text{IF}_i(\widehat{q}_r) \tag{17}$$

Because, according to table 1,

$$\text{IF}_i(\widehat{q}_r) = \frac{W}{W_R}R_i\left\{\frac{r - \mathbb{1}(Y_i \leq \widehat{q}_r)}{\widehat{f}_{Y|R}(\widehat{q}_r)}\right\}$$

equation (17) can be written as

$$\text{IF}_i\{\widehat{G}(r)\} = W\frac{D_i}{W_D}\{\mathbb{1}(X_i \leq \widehat{q}_r) - \widehat{G}(r)\} + W\frac{R_i}{W_R}\frac{\widehat{f}_{X|D}(\widehat{q}_r)}{\widehat{f}_{Y|R}(\widehat{q}_r)}\{r - \mathbb{1}(Y_i \leq \widehat{q}_r)\}$$

The density ratio in the second term is equal to the relative density by definition, so we could replace it by $\widehat{g}(r)$. Both variants should yield a consistent estimate of the standard error.

A.2.1 Location and scale adjustment

For the relative CDF based on location-adjusted (and possibly scale-adjusted) data, replace q_r by $\widetilde{q}_r = t(q_r, \theta)$ in the above formulas, where $t(y, \theta)$ is a scalar transformation

function depending on a set of location and scale parameters $\theta = (\theta_1, \ldots, \theta_K)$. More specifically, if $t_D(y, \theta)$ is the transformation function applied to the comparison data and $t_R(y, \theta)$ is the transformation function applied to the reference data, we have

$$\widetilde{q}_r = t(q_r, \theta) = t_D^{-1}\{t_R(q_r, \theta), \theta\} \tag{18}$$

The adjusted relative CDF then becomes

$$\widehat{\widetilde{G}}(r) = \widehat{F}_D(\widehat{\widetilde{q}}_r) = \frac{1}{W_D}\sum_{i\in\mathcal{D}} w_i \mathbb{1}(X_i \leq \widehat{\widetilde{q}}_r)$$

with $\widehat{\widetilde{q}}_r = t(\widehat{q}_r, \widehat{\theta})$ and $\widehat{q}_r = \widehat{F}^{-1}_{Y|R}(r)$, such that the IF can be written as

$$\text{IF}_i\{\widehat{\widetilde{G}}(r)\} = \frac{W}{W_D}\widehat{h}_i^{\widetilde{G}} + \widehat{f}_{X|D}(\widehat{\widetilde{q}}_r)\text{IF}_i(\widehat{\widetilde{q}}_r) \tag{19}$$

with

$$h_i^{\widetilde{G}} = D_i\{\mathbb{1}(X_i \leq \widetilde{q}_r) - \widetilde{G}(r)\} \tag{20}$$

and

$$\text{IF}_i(\widehat{\widetilde{q}}_r) = \tau_{q_r}\text{IF}_i(\widehat{q}_r) + \sum_{k=1}^{K}\tau_{\theta_k}\text{IF}_i(\widehat{\theta}_k) \tag{21}$$

where

$$\tau_{q_r} = \left.\frac{\partial t(q_r, \theta)}{\partial q_r}\right|_{q_r=\widehat{q}_r, \theta=\widehat{\theta}} \qquad \text{and} \qquad \tau_{\theta_k} = \left.\frac{\partial t(q_r, \theta)}{\partial \theta_k}\right|_{q_r=\widehat{q}_r, \theta=\widehat{\theta}}$$

For example, in the case of an additive location adjustment (of either the reference distribution or the comparison distribution), we have $t(q_r, \theta) = q_r - \mu_{Y|R} + \mu_{X|D}$, such that

$$\text{IF}_i(\widehat{\widetilde{q}}_r) = \text{IF}_i(\widehat{q}_r) - \text{IF}_i(\widehat{\mu}_{Y|R}) + \text{IF}_i(\widehat{\mu}_{X|D})$$

where expressions for the three IFs included in $\text{IF}_i(\widehat{\widetilde{q}}_r)$ can be found in table 1 (μ is either the median or the mean). Likewise, in the case of a multiplicative adjustment, we have $t(q_r, \theta) = q_r \times \mu_{X|D}/\mu_{Y|R}$, such that

$$\text{IF}_i(\widehat{\widetilde{q}}_r) = \frac{\widehat{\mu}_{X|D}}{\widehat{\mu}_{Y|R}}\text{IF}_i(\widehat{q}_r) + \frac{\widehat{q}_r}{\widehat{\mu}_{Y|R}}\text{IF}_i(\widehat{\mu}_{X|D}) - \frac{\widehat{q}_r\widehat{\mu}_{X|D}}{(\widehat{\mu}_{Y|R})^2}\text{IF}_i(\widehat{\mu}_{Y|R})$$

In the case of additive location and scale adjustment, we have $t(q_r, \theta) = (q_r - \mu_{Y|R}) \times s_{X|D}/s_{Y|R} + \mu_{X|D}$, such that

$$\begin{aligned}\text{IF}_i(\widehat{\widetilde{q}}_r) = {} & \frac{\widehat{s}_{X|D}}{\widehat{s}_{Y|R}}\text{IF}_i(\widehat{q}_r) - \frac{\widehat{s}_{X|D}}{\widehat{s}_{Y|R}}\text{IF}_i(\widehat{\mu}_{Y|R}) + \frac{\widehat{q}_r - \widehat{\mu}_{Y|R}}{\widehat{s}_{Y|R}}\text{IF}_i(\widehat{s}_{X|D}) \\ & - \frac{(\widehat{q}_r - \widehat{\mu}_{Y|R})\widehat{s}_{X|D}}{(\widehat{s}_{Y|R})^2}\text{IF}_i(\widehat{s}_{Y|R}) + \text{IF}_i(\widehat{\mu}_{X|D})\end{aligned}$$

where s is either the IQR or the standard deviation. Finally, for a logarithmic location and scale adjustment, we have $t(q_r,\theta) = \exp[\{\ln q_r - \mu_{\ln(Y)|R}\} \times s_{\ln(X)|D}/s_{\ln(Y)|R} + \mu_{\ln(X)|D}]$, such that

$$\begin{aligned}\mathrm{IF}_i(\widehat{\widetilde{q}}_r) = \widehat{\widetilde{q}}_r \Bigg[& \frac{\widehat{s}_{\ln(X)|D}}{\widehat{q}_r \widehat{s}_{\ln(Y)|R}} \mathrm{IF}_i(\widehat{q}_r) - \frac{\widehat{s}_{\ln(X)|D}}{\widehat{s}_{\ln(Y)|R}} \mathrm{IF}_i\{\widehat{\mu}_{\ln(Y)|R}\} + \frac{\ln \widehat{q}_r - \widehat{\mu}_{\ln(Y)|R}}{\widehat{s}_{\ln(Y)|R}} \mathrm{IF}_i\{\widehat{s}_{\ln(X)|D}\} \\ & - \frac{\{\ln \widehat{q}_r - \widehat{\mu}_{\ln(Y)|R}\}\widehat{s}_{\ln(X)|D}}{\{\widehat{s}_{\ln(Y)|R}\}^2} \mathrm{IF}_i\{\widehat{s}_{\ln(Y)|R}\} + \mathrm{IF}_i\{\widehat{\mu}_{\ln(X)|D}\} \Bigg]\end{aligned}$$

In the case of a shape adjustment, one of the two distributions is replaced by a location-adjusted (and possibly shape-adjusted) variant of the other distribution. The same formulas as above can be applied after choosing the appropriate transformation function and replacing some of the components. For example, if the comparison distribution is shape- and scale-adjusted (and the reference distribution remains unchanged), the relevant transformation functions are $t_D(y,\theta) = y - \mu_{X|R} + \mu_{X|D}$ and $t_R(y,\theta) = y$, such that $t(q_r,\theta) = q_r - \mu_{X|D} + \mu_{X|R}$. The main moment condition will be conditional on subsample $\mathcal{R}$ instead of $\mathcal{D}$, meaning that D_i and X_i in (20) have to be replaced by R_i and Y_i. This further implies that W_D in the first term of (19) has to be replaced by W_R and that the density in the second term is $\widehat{f}_{Y|R}(\widehat{\widetilde{q}}_r)$ instead of $\widehat{f}_{X|D}(\widehat{\widetilde{q}}_r)$. If the reference distribution is shape- and scale-adjusted (and the comparison distribution remains unchanged), the transformation function again is $t(q_r,\theta) = q_r - \mu_{X|D} + \mu_{X|R}$, but $\widehat{q}_r$ is now based on the comparison distribution, that is, $\widehat{q}_r = \widehat{F}^{-1}_{X|D}(r)$, such that the definition of $\mathrm{IF}_i(\widehat{q}_r)$ in (21) changes.

A.3 IF for the relative histogram

For the IF of a histogram estimate of the relative density, note that for each bin the histogram density is equal to the difference between two points on the relative CDF divided by the bin width. That is,

$$\widehat{g}(a,b) = \frac{\widehat{G}(b) - \widehat{G}(a)}{b-a}$$

The IF for $\widehat{g}(a,b)$ can thus be obtained as

$$\mathrm{IF}_i\{\widehat{g}(a,b)\} = \frac{\mathrm{IF}_i\{\widehat{G}(b)\} - \mathrm{IF}_i\{\widehat{G}(a)\}}{b-a}$$

A.4 IF for the relative PDF

The relative density estimate for continuous data can be written as

$$\widehat{g}(r) = \frac{1}{W_D} \sum_{i \in \mathcal{D}} w_i K_c(r, \widehat{r}_i, h) \qquad \text{with } \widehat{r}_i = \widehat{F}_{Y|R}(X_i)$$

where $K_c(\cdot)$ is a boundary-corrected kernel function as described in section 3.3.1. Note that each individual $\widehat{r}_i$, $i \in \mathcal{D}$, has its own moment condition:

$$h_i^g = D_i\{K_c(r, r_i, h) - g(r)\}$$
$$h_i^{r_j} = R_i\{\mathbb{1}(Y_i \leq X_j) - r_j\} \qquad \text{for each } j \in \mathcal{D}$$

This leads to

$$\text{IF}_i\{\widehat{g}(r)\} = \frac{W}{W_D}\left\{\widehat{h}_i^g + \sum_{j\in\mathcal{D}} \widehat{\Delta}_{r_j}^g \text{IF}_i(\widehat{r}_j)\right\}$$
$$= \frac{W}{W_D}\left[\widehat{h}_i^g + \frac{R_i}{W_R}\sum_{j\in\mathcal{D}} \delta_j\{\mathbb{1}(Y_i \leq X_j) - \widehat{r}_j\}\right]$$

with $\widehat{\Delta}_{r_j}^g = \delta_j/W$ and $\delta_j = w_j K_c'(r, \widehat{r}_j, h)$, where $K_c'(r, \widehat{r}_j, h)$ is the derivative of $K_c(r, r_j, h)$ with respect to r_j at point $\widehat{r}_j$. The sum in the second part of the equation looks computationally burdensome [complexity $O(n_R n_D)$ once we evaluate the IF for all observations], but it can be simplified. Let

$$\lambda_i = \sum_{j\in\mathcal{D}} \delta_j \mathbb{1}(Y_i \leq X_j) \qquad \text{and} \qquad \Lambda = \sum_{j\in\mathcal{D}} \delta_j \widehat{r}_j$$

such that

$$\text{IF}_i\{\widehat{g}(r)\} = \frac{W}{W_D}\left\{\widehat{h}_i^g + \frac{R_i}{W_R}(\lambda_i - \Lambda)\right\}$$

Term λ_i is equivalent to a "reverse" (summation from the top) and nonnormalized CDF of X weighted by δ_j and can be obtained for all observations in a single run across the data.[21]

A.4.1 Location and scale adjustment

For the relative PDF based on location-adjusted (and possibly scale-adjusted) data, define $\widetilde{x}_i = t^{-1}(X_i, \theta)$ and replace r_i by $\widetilde{r}_i = F_{Y|R}(\widetilde{x}_i)$ in the above formulas. Function $t(x, \theta)$ is as defined in (18); if $t_D(x, \theta)$ is the transformation function applied to the comparison data and $t_R(x, \theta)$ is the transformation function applied to the reference data, then

$$t^{-1}(x, \theta) = t_R^{-1}\{t_D(x, \theta), \theta\} \tag{22}$$

The adjusted relative PDF can thus be written as

$$\widehat{\widetilde{g}}(r) = \frac{1}{W_D}\sum_{i\in\mathcal{D}} w_i K_c(r, \widehat{\widetilde{r}}_i, h) \qquad \text{with } \widehat{\widetilde{r}}_i = \widehat{F}_{Y|R}(\widehat{\widetilde{x}}_i) \text{ and } \widehat{\widetilde{x}}_i = t^{-1}(X_i, \widehat{\theta})$$

21. For the adaptive kernel (see footnote 9), a complication arises because the local bandwidth depends on preliminary density estimates. This should only be of secondary importance for the variance estimate so that applying the above equation with h replaced by the relevant local bandwidth (that is, treating the local bandwidth as fixed) should produce acceptable results in practice.

such that the IF becomes

$$\text{IF}_i\{\widehat{\widetilde{g}}(r)\} = \frac{W}{W_D}\left\{\widehat{h}_i^{\widetilde{g}} + \sum_{j\in\mathcal{D}} \widehat{\Delta}_{\widetilde{r}_j}^{\widetilde{g}}\,\text{IF}_i(\widehat{\widetilde{r}}_j)\right\}$$

with

$$\begin{aligned}
\widehat{h}_i^{\widetilde{g}} &= D_i\{K_c(r,\widehat{\widetilde{r}}_i,h) - \widehat{\widetilde{g}}(r)\}\\
\widehat{\Delta}_{\widetilde{r}_j}^{\widetilde{g}} &= \delta_j/W \qquad \text{with } \delta_j = w_j K_c'(r,\widehat{\widetilde{r}}_j,h)\\
\text{IF}_i(\widehat{\widetilde{r}}_j) &= \frac{W}{W_R}\left\{\widehat{h}_i^{\widetilde{r}_j} + \widehat{\Delta}_{\widetilde{x}_j}^{\widetilde{r}_j}\text{IF}_i(\widehat{\widetilde{x}}_j)\right\}\\
\widehat{h}^{\widetilde{r}_j} &= R_i\{\mathbb{1}(Y_i \le \widehat{\widetilde{x}}_j) - \widehat{\widetilde{r}}_j\}\\
\widehat{\Delta}_{\widetilde{x}_j}^{\widetilde{r}_j} &= \frac{W_R}{W}\widehat{f}_{Y|R}(\widehat{\widetilde{x}}_j)\\
\text{IF}_i(\widehat{\widetilde{x}}_j) &= \sum_{k=1}^{K}\tau_{jk}\text{IF}_i(\widehat{\theta}_k) \qquad \text{with } \tau_{jk} = \left.\frac{\partial t^{-1}(X_j,\theta)}{\partial\theta_k}\right|_{\theta=\widehat{\theta}}
\end{aligned}$$

Similar to above, computational complexity can be reduced by rewriting the IF as

$$\text{IF}_i\{\widehat{\widetilde{g}}(r)\} = \frac{W}{W_D}\left\{\widehat{h}_i^{\widetilde{g}} + \frac{R_i}{W_R}(\lambda_i - \Lambda) + \sum_{k=1}^{K}\kappa_k\text{IF}_i(\widehat{\theta}_k)\right\} \tag{23}$$

with

$$\lambda_i = \sum_{j\in\mathcal{D}}\delta_j\mathbb{1}(Y_i \le \widehat{\widetilde{x}}_j), \qquad \Lambda = \sum_{j\in\mathcal{D}}\delta_j\widehat{\widetilde{r}}_j, \qquad \kappa_k = \frac{1}{W}\sum_{j\in\mathcal{D}}\delta_j\widehat{f}_{Y|R}(\widehat{\widetilde{x}}_j)\tau_{jk}$$

In the case of a shape adjustment, the same formulas can be used, but various components have to be replaced to account for the switch in subsamples.

A.5 IF for the discrete relative density

The categorical relative density is defined as

$$\widehat{g}^k = \widehat{p}_k^D/\widehat{p}_k^R$$

with moment conditions

$$h_i^{p_D^k} = D_i\{\mathbb{1}(X_i = k) - p_k^D\} \quad \text{and} \quad h_i^{p_R^k} = R_i\{\mathbb{1}(Y_i = k) - p_k^R\}$$

where $\widehat{p}_k^D = \widehat{P}_D(X = k)$ and $\widehat{p}_k^R = \widehat{P}_R(Y = k)$. The IF can thus be written as

$$\text{IF}_i(\widehat{g}^k) = \frac{1}{\widehat{p}_k^R}\text{IF}_i(\widehat{p}_k^D) - \frac{\widehat{p}_k^D}{(\widehat{p}_k^R)^2}\text{IF}_i(\widehat{p}_k^R)$$

with

$$\text{IF}_i(\widehat{p}_k^D) = \frac{W}{W_D}D_i\{\mathbb{1}(X_i = k) - \widehat{p}_k^D\} \quad \text{and} \quad \text{IF}_i(\widehat{p}_k^R) = \frac{W}{W_R}R_i\{\mathbb{1}(Y_i = k) - \widehat{p}_k^R\} \tag{24}$$

A.6 IFs for divergence measures

Divergence measures are obtained as aggregates of relative density estimates. Hence, their IFs can be written as aggregates of the IFs of the density estimates. Assuming the divergence measures are computed from a kernel density estimate on a regular grid or from a histogram density with K evenly sized bins, as described in section 3.4, we get

$$\text{IF}_i(\widehat{\chi^2}) = \frac{2}{K}\sum_{k=1}^{K}(\widehat{g}_k - 1)\text{IF}_i(\widehat{g}_k)$$

$$\text{IF}_i(\widehat{\text{KL}}) = \frac{1}{K}\sum_{k=1}^{K}\{1 + \ln(\widehat{g}_k)\}\text{IF}_i(\widehat{g}_k)$$

$$\text{IF}_i(\widehat{\text{TVD}}) = \frac{1}{2K}\sum_{k=1}^{K}\text{sign}(\widehat{g}_k - 1)\text{IF}_i(\widehat{g}_k)$$

where $\widehat{g}_k$ is the kernel density estimate at evaluation point r_k or the histogram estimate for bin k. For divergence measures computed from categorical data, the IFs can be written as

$$\text{IF}_i(\widehat{\chi^2}) = \sum_{k=1}^{K} 2\left(\frac{\widehat{p}_k^D}{\widehat{p}_k^R} - 1\right)\text{IF}_i(\widehat{p}_k^D) + \left\{1 - \left(\frac{\widehat{p}_k^D}{\widehat{p}_k^R}\right)^2\right\}\text{IF}_i(\widehat{p}_k^R)$$

$$\text{IF}_i(\widehat{\text{KL}}) = \sum_{k=1}^{K}\left\{\ 1 + \ln\left(\frac{\widehat{p}_k^D}{\widehat{p}_k^R}\right)\right\}\text{IF}_i(\widehat{p}_k^D) - \frac{\widehat{p}_k^D}{\widehat{p}_k^R}\text{IF}_i(\widehat{p}_k^R)$$

$$\text{IF}_i(\widehat{\text{TVD}}) = \frac{1}{2}\sum_{k=1}^{K}\text{sign}(\widehat{p}_k^D - \widehat{p}_k^R)\left\{\text{IF}_i(\widehat{p}_k^D) - \text{IF}_i(\widehat{p}_k^R)\right\}$$

where $\text{IF}_i(\widehat{p}_k^D)$ and $\text{IF}_i(\widehat{p}_k^R)$ are as defined in (24).

A.7 IFs for polarization indices

The MRP can be written as

$$\text{MRP} = \frac{1}{W_D}\sum_{i\in\mathcal{D}} w_i(4|\widehat{\widetilde{r}}_i - 0.5| - 1) \qquad \text{with } \widehat{\widetilde{r}}_i = \widehat{F}_{Y|R}(\widehat{\widetilde{x}}_i) \text{ and } \widehat{\widetilde{x}}_i = t^{-1}(X_i, \widehat{\theta})$$

where $t^{-1}(x,\theta)$ is as defined in (22). We see that the MRP has the same structure as an estimate of the relative PDF based on location-adjusted (and possibly scale-adjusted) data. We can thus obtain the IF using (23) with $h_i^{\widetilde{g}}$ replaced by

$$h_i^{\text{MRP}} = D_i\{(4|\widetilde{r}_i - 0.5| - 1) - \text{MRP}\}$$

and δ_j set to

$$\delta_j = w_j\, 4\,\text{sign}(\widehat{\widetilde{r}}_j - 0.5)$$

Likewise, the IF for the LRP can be obtained by using

$$h_i^{\text{LRP}} = D_i[\{8|\widetilde{r}_i - 0.5|\mathbb{1}(\widetilde{r}_i < 0.5) - 1\} - \text{LRP}]$$

and

$$\delta_j = w_j\, 8\,\text{sign}(\widehat{\widetilde{r}}_j - 0.5)\mathbb{1}(\widehat{\widetilde{r}}_j < 0.5)$$

Note that $\mathbb{1}(\widetilde{r}_i < 0.5)$ always selects half the comparison data because the data have been median-adjusted. Assuming that it is fixed should not introduce significant bias into the variance estimates. The IF for the URP can be derived analogously.

A.8 IFs for descriptive statistics

Like the MRP, summary statistics of the relative ranks such as the mean or the standard deviation have a structure that is very similar to the relative PDF. For the mean $\widehat{\mu}$ of the (possibly adjusted) relative ranks, the IF can be obtained by replacing $h_i^{\widetilde{g}}$ in (23) by

$$h_i^{\mu} = D_i(\widetilde{r}_i - \mu)$$

and setting δ_j to

$$\delta_j = w_j$$

Likewise, for the variance $\widehat{\sigma}^2$ of the relative ranks, we can use

$$h_i^{\sigma^2} = D_i\{(\widetilde{r}_i - \mu)^2 - \sigma^2\} \qquad \text{and} \qquad \delta_j = 2w_j(\widehat{\widetilde{r}}_j - \widehat{\mu})$$

The IF for the standard deviation $\widehat{\sigma}$ is given as

$$\text{IF}_i(\widehat{\sigma}) = \frac{1}{2\widehat{\sigma}}\text{IF}_i(\widehat{\sigma}^2)$$

For quantile $\widehat{q}(p)$ of the relative ranks, it is easier to follow a different approach. Note that the quantile can be written as

$$q(p) = G^{-1}(p) = F_{Y|R}\{F_{X|D}^{-1}(p)\}$$

That is, a quantile of the relative ranks of $F_{X|D}$ with respect to $F_{Y|R}$ is equivalent to a point on the relative CDF of $F_{Y|R}$ with respect to $F_{X|D}$. We can thus obtain the IF in the same way as for the relative CDF (see page 942) but with swapped distributions. Finally, the IF for the IQR is given as

$$\text{IF}_i(\text{IQR}) = \text{IF}_i\{\widehat{q}(0.75)\} - \text{IF}_i\{\widehat{q}(0.25)\}$$

A.9 IFs in case of covariate balancing

If covariates are balanced using the reweighting approach, the IFs need to be adjusted to account for the fact that the balancing weights have been estimated. I will discuss two reweighting methods below: IPW based on logistic regression and IPW based on

entropy balancing. Deriving the IFs is relatively easy in these cases because the weights are obtained from a parametric model. Nonparametric reweighting methods such as matching are more challenging; I leave it to future research to work out the details for such methods.

Given covariates Z, logit IPW and entropy balancing both estimate a vector of coefficients γ from which the balancing weights $\widetilde{w}$ are computed. To obtain the IF for a reweighted relative distribution statistic, we could follow the procedures outlined above but additionally incorporate γ by applying (16) to each of the parameters within the definition of the statistic that depends on the balancing weights. However, working through the math shows that this can be simplified. Let θ be a covariate-adjusted relative distribution statistic, and let $\text{IF}_i^*(\widehat{\theta})$ be a provisional IF that assumes the balancing weights as fixed (the provisional IF is obtained in the same way as without covariate adjustment, only that the base weights w are replaced by the balancing weights $\widetilde{w}$). The final IF that no longer assumes the balancing weights as fixed can then be computed as

$$\text{IF}_i(\widehat{\theta}) = \frac{\widetilde{w}_i}{w_i}\text{IF}_i^*(\widehat{\theta}) + \widehat{\Delta}_\gamma^{\text{IF}^*}\,\text{IF}_i(\widehat{\gamma})$$

For both logit IPW and entropy balancing, the vector of partial derivatives of $\text{IF}_i^*(\widehat{\theta})$ by γ can be written as

$$\frac{\partial \text{IF}_i^*(\widehat{\theta})}{\partial \gamma'} = \frac{\widetilde{w}_i}{w_i}(1-T_i)\text{IF}_i^*(\widehat{\theta})Z_i$$

such that

$$\widehat{\Delta}_\gamma^{\text{IF}^*} = \frac{1}{W}\sum_{i=1}^{N}\widetilde{w}_i(1-T_i)\text{IF}_i^*(\widehat{\theta})Z_i$$

where T is an indicator for the "treatment" group ($T=R$ if the comparison group is reweighted, $T=D$ if the reference group is reweighted; the two groups are assumed distinct and exhaustive). The definition of the balancing weights and the expression for $\text{IF}_i(\widehat{\gamma})$ differ between the two procedures, as is described next.

A.9.1 Logit IPW

In the case of logit IPW, the balancing weights are defined as

$$\widetilde{w}_i = \begin{cases} w_i\frac{\widehat{p}_i}{1-\widehat{p}_i}c & \text{if } T_i = 0 \\ w_i & \text{if } T_i = 1 \end{cases} \qquad \text{with} \qquad \widehat{p}_i = \frac{e^{Z_i\widehat{\gamma}}}{1+e^{Z_i\widehat{\gamma}}}$$

where Z_i is a row vector of covariates (typically including a constant), $\widehat{\gamma}$ is the corresponding coefficient vector estimated by logistic regression, and c is a scaling factor ensuring that the sum of weights remains constant (see section 3.6; c can be treated as fixed). As discussed in Jann (2020), the logistic regression moment conditions for γ can be written as

$$h_i^\gamma = Z_i'(T_i - p_i)$$

such that

$$\text{IF}_i(\widehat{\gamma}) = (-\widehat{\Delta}^{\gamma})^{-1} Z_i'(T_i - \widehat{p}_i) \qquad \text{with} \qquad \widehat{\Delta}^{\gamma} = \frac{1}{W}\sum_{i=1}^{N} w_i Z_i' \widehat{p}_i(1-\widehat{p}_i) Z_i$$

A.9.2 Entropy balancing

For entropy balancing, the weights are given as

$$\widetilde{w}_i = \begin{cases} w_i e^{z_i\widehat{\beta}+\widehat{\alpha}} c & \text{if } T_i = 0 \\ w_i & \text{if } T_i = 1 \end{cases}$$

where z_i is a row vector of k covariates without constant and $\widehat{\beta}$ and $\widehat{\alpha}$ are the entropy-balancing coefficients. The estimation of β involves a vector of auxiliary parameters μ, the means of Z in the treatment group. Furthermore, α involves auxiliary parameter p, the treatment probability. Based on Jann (2020), the entropy-balancing moment conditions can be written as

$$h_i^{\beta} = (1-T_i)\frac{\widetilde{w}_i}{w_i c}(z_i' - \mu), \qquad h_i^{\alpha} = (1-T_i)\left(\frac{\widetilde{w}_i}{w_i c} - \frac{p}{1-p}\right),$$
$$h_i^{\mu} = T_i(z_i' - \mu), \qquad h_i^{p} = T_i - p$$

such that

$$\text{IF}_i(\widehat{\beta}) = (-\widehat{\Delta}^{\beta})^{-1}\{\widehat{h}_i^{\beta} + \widehat{\Delta}_{\mu}^{\beta}(-\widehat{\Delta}^{\mu})^{-1}\widehat{h}_i^{\mu}\} = (-\widehat{\Delta}^{\beta})^{-1}\left(\widehat{h}_i^{\beta} - \widehat{h}_i^{\mu}\right) \tag{25}$$

with

$$\widehat{\Delta}^{\beta} = \frac{1}{W}\sum_{i=1}^{N} w_i \widehat{h}_i^{\beta} z_i = \frac{1}{W}\sum_{i=1}^{N} \frac{\widetilde{w}_i}{c}(1-T_i)(z_i' - \widehat{\mu})z_i$$

The simplification on the right in (25) follows from

$$\widehat{\Delta}^{\mu} = \frac{1}{W}\sum_{i=1}^{N} -w_i T_i \,\text{diag}_k(1) = -\text{diag}_k(W_T)/W$$
$$\widehat{\Delta}_{\mu}^{\beta} = \frac{1}{W}\sum_{i=1}^{N} -\frac{\widetilde{w}_i}{c}(1-T_i)\,\text{diag}_k(1) = -\text{diag}_k(\Omega)/W$$

where $\Omega = \sum_{i:T_i=0} \widetilde{w}_i/c$ because moment condition h^{α} ensures that $\Omega = W_T$. Furthermore, for α we get

$$\begin{aligned}\text{IF}_i(\widehat{\alpha}) &= \frac{1}{-\widehat{\Delta}^{\alpha}}\left\{\widehat{h}_i^{\alpha} + \widehat{\Delta}_p^{\alpha}\frac{1}{-\widehat{\Delta}^{p}}\widehat{h}_i^{p} + \widehat{\Delta}_{\beta}^{\alpha}\text{IF}_i(\widehat{\beta})\right\} \\ &= -\frac{\Omega}{W}\left\{\widehat{h}_i^{\alpha} - \frac{T_i - \widehat{p}}{1-\widehat{p}} + \widehat{\Delta}_{\beta}^{\alpha}\text{IF}_i(\widehat{\beta})\right\} \quad \text{with} \quad \widehat{\Delta}_{\beta}^{\alpha} = \frac{1}{W}\sum_{i=1}^{N}\frac{\widetilde{w}_i}{c}z_i\end{aligned}$$

Finally, let

$$Z_i = (z_i, 1), \qquad \widehat{\gamma} = (\widehat{\beta}', \widehat{\alpha})', \qquad \text{IF}_i(\widehat{\gamma}) = \{\text{IF}_i(\widehat{\beta})', \text{IF}_i(\widehat{\alpha})\}'$$

so that the results have the same format as for the logit IPW.