b
UNIVERSITÄT
BERN

Faculty of Business, Economics and Social Sciences

Department of Social Sciences

University of Bern Social Sciences Working Paper No. 37

Relative distribution analysis in Stata

Ben Jann

Current version: September 26, 2020
First version: September 18, 2020

http://ideas.repec.org/p/bss/wpaper/37.html
http://econpapers.repec.org/paper/bsswpaper/37.htm

University of Bern
Department of Social Sciences
Fabrikstrasse 8
CH-3012 Bern

Tel. +41 (0)31 631 48 11
Fax +41 (0)31 631 48 17
info@sowi.unibe.ch
www.sowi.unibe.ch

Relative distribution analysis in Stata

Ben Jann
Institute of Sociology
University of Bern
ben.jann@soz.unibe.ch

September 26, 2020

Abstract

In this paper 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. Another example would be the analysis of changes in the distribution of earnings over time. Of interest are the relative cumulative distribution (relative CDF), the relative density (relative PDF), as well as summary measures such as the median relative polarization (MRP). The presented software can be used to estimate these quantities and 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: Stata, `reldist`, relative distribution, relative density, median relative polarization, divergence, reweighting, influence functions

Contents

1 Introduction

Although earlier work on relative distributions and related approaches can be found in the statistical literature (e.g., 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 et al., 1994; Bernhardt et al., 1995, 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 widespread use, which might, in part, be due to lack of statistical software supporting such analyses (although an R package does exist; see Handcock and Aldrich, 2002; Handcock, 2016).

In this article I provide an introduction to relative distributions methods, discuss issues that are relevant for estimation, and 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, distributional divergence measures, as well as descriptive statistics of the relative data, and supports the decomposition of the relative distribution by adjusting for location, scale, and shape differences, or for differences in covariate distributions.

2 Theory

In this section I will briefly 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 Distribution function and density

Let Y be a continuous outcome variable of interest. Y is assumed a random variable with distribution function

$$F_Y(y) = P(Y \leq y), \quad y \in \mathbb{R} \tag{1}$$

That is, for any value y, the distribution function provides the probability that Y will take on a value that is smaller than or equal to y. The density function of Y is then defined as the first derivative of the distribution function, that is,

$$f_Y(y) = F'_Y(y) = \frac{\mathrm{d}F_Y(y)}{\mathrm{d}y} \tag{2}$$

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

$$F_Y(y) = \int_{-\infty}^{y} f_Y(t)\,\mathrm{d}t \tag{3}$$

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

$$P(a < Y \leq b) = F_Y(b) - F_Y(a) = \int_a^b f_Y(y)\,\mathrm{d}y \tag{4}$$

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

$$r_Y(y) = F_Y(y) \tag{5}$$

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

Relative ranks can themselves have a distribution which 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 an even distribution.

2.3 The relative distribution function

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 distribution function (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 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

$$G(r) = F_X(F_Y^{-1}(r)), \quad r \in [0, 1] \tag{6}$$

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

$$(F_X(y), F_Y(y)), \quad y \in \mathbb{R} \tag{7}$$

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 equation 6).[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.

2.4 The relative density function

Since $G(r)$ is a distribution function, we can take the first derivative to obtain a density function. That is, using the chain rule, the relative density function (relative PDF) of F_X with respect to F_Y can be derived as

$$g(r) = \frac{\mathrm{d}G(r)}{\mathrm{d}r} = \frac{f_X(F_Y^{-1}(r))}{f_Y(F_Y^{-1}(r))}, \quad r \in [0,1] \tag{8}$$

As can be seen, the relative density is equal to the ratio of the densities of the two distributions at a specific y value (i.e. $g(r)$ 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 density function as 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 (8). 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 distribution, 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 in their mean or median. If a large location difference exist, the relative CDF and PDF will be dominated by this difference. In many application 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

$$\widetilde{Y} = Y - \mu_Y + \mu_X \tag{9}$$

[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).

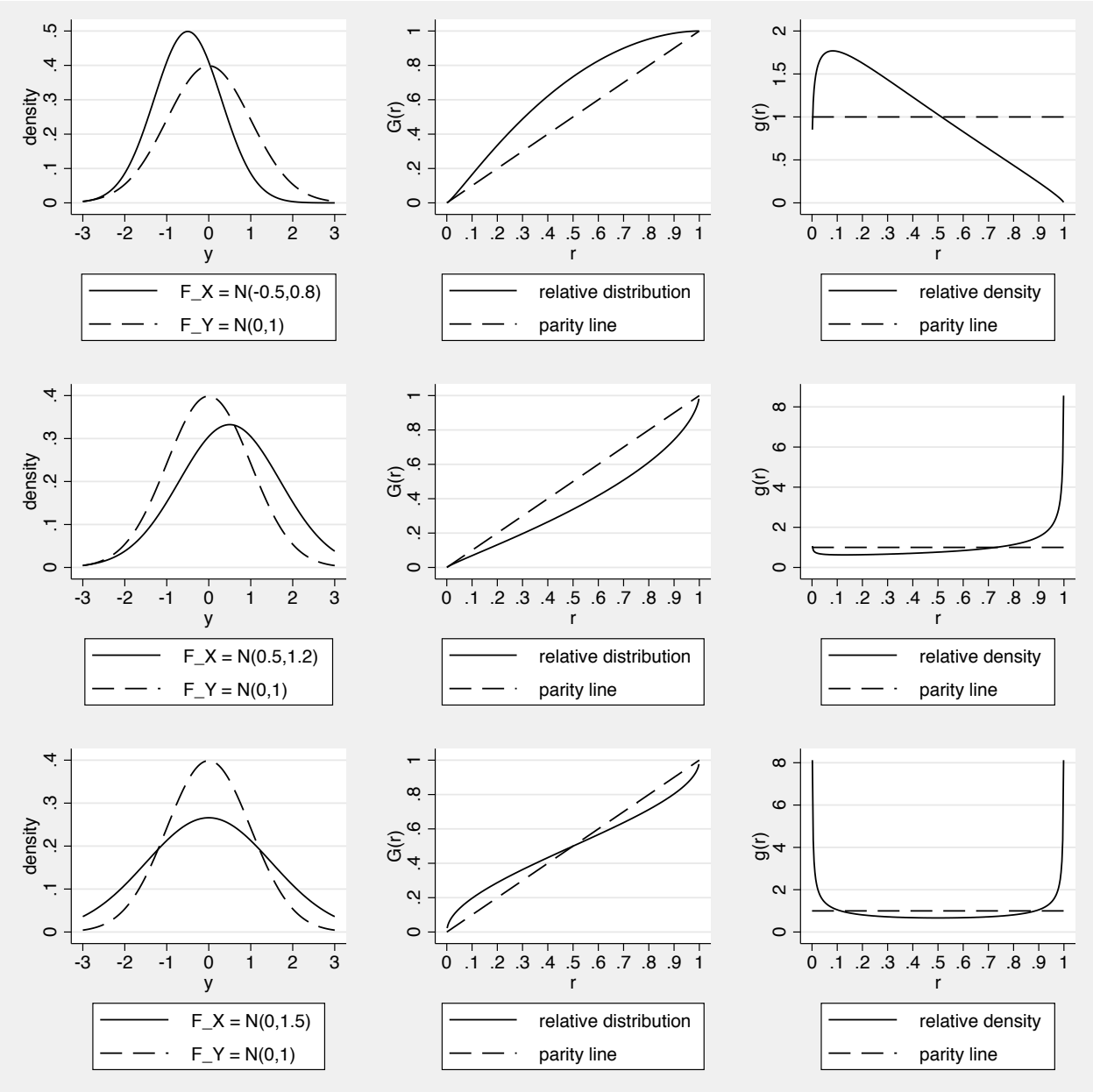


Figure 1: Illustration of the relative distribution

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 $F_Y(t^{-1}(y))$. That is,

$$F_{\widetilde{Y}}(y) = P(Y - \mu_Y + \mu_X \leq y) = P(Y \leq y + \mu_Y - \mu_X) = F_Y(y + \mu_Y - \mu_X) \tag{10}$$

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

$$g(r) = \frac{f_X(F_Y^{-1}(r))}{f_Y(F_Y^{-1}(r))} = \underbrace{\frac{f_{\widetilde{Y}}(F_Y^{-1}(r))}{f_Y(F_Y^{-1}(r))}}_{\text{location effect}} \times \underbrace{\frac{f_X(F_Y^{-1}(r))}{f_{\widetilde{Y}}(F_Y^{-1}(r))}}_{\text{shape effect}} \tag{11}$$

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

$$g_{X\widetilde{Y}}(r) = \frac{f_X(F_{\widetilde{Y}}^{-1}(r))}{f_{\widetilde{Y}}(F_{\widetilde{Y}}^{-1}(r))} \tag{12}$$

or the corresponding adjusted relative CDF

$$G_{X\widetilde{Y}}(r) = F_X(F_{\widetilde{Y}}^{-1}(r)) \tag{13}$$

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

$$\widetilde{X} = X - \mu_X + \mu_Y \qquad \text{with} \qquad F_{\widetilde{X}}(y) = F_X(y + \mu_X - \mu_Y) \tag{14}$$

such that

$$g(r) = \frac{f_X(F_Y^{-1}(r))}{f_Y(F_Y^{-1}(r))} = \underbrace{\frac{f_X(F_Y^{-1}(r))}{f_{\widetilde{X}}(F_Y^{-1}(r))}}_{\text{location effect}} \times \underbrace{\frac{f_{\widetilde{X}}(F_Y^{-1}(r))}{f_Y(F_Y^{-1}(r))}}_{\text{shape effect}} \tag{15}$$

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

$$g_{X\widetilde{X}}(r) = \frac{f_X(F_{\widetilde{X}}^{-1}(r))}{f_{\widetilde{X}}(F_{\widetilde{X}}^{-1}(r))} \quad \text{and} \quad G_{X\widetilde{X}}(r) = F_X(F_{\widetilde{X}}^{-1}(r)) \tag{16}$$

instead of $f_X(F_Y^{-1}(r))/f_{\widetilde{X}}(F_Y^{-1}(r))$. Results from (11) and (15) will generally not be the same, although for some 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 (e.g. 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 $\widetilde{Y} = Y \cdot \mu_X/\mu_Y$ and hence

$$F_{\widetilde{Y}}(y) = F_Y(y \cdot \mu_Y/\mu_X) \tag{17}$$

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 $\widetilde{Y} = (Y - \mu_Y) \cdot s_X/s_Y + \mu_X$ such that

$$F_{\widetilde{Y}}(y) = F_Y((y - \mu_X) \cdot s_Y/s_X + \mu_Y) \tag{18}$$

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

$$F_{\widetilde{Y}}(y) = F_Y(\exp((\ln(y) - \mu_{\ln(X)}) \cdot s_{\ln(Y)}/s_{\ln(X)} + \mu_{\ln(Y)})) \tag{19}$$

2.6 Summary measures

2.6.1 Divergence

Handcock and Morris (1999) suggest the Pearson chi-squared 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 chi-squared divergence between F_X and F_Y is defined as

$$\chi^2 = \int_{-\infty}^{\infty} \frac{(f_X(y) - f_Y(y))^2}{f_Y(y)} \,\mathrm{d}y = \int_0^1 (g(r) - 1)^2 \,\mathrm{d}r \tag{20}$$

The equality between the first and second expression 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

$$\text{KL} = \int_{-\infty}^{\infty} \ln\left(\frac{f_X(y)}{f_Y(y)}\right) f_X(y) \,\mathrm{d}y = \int_0^1 \ln(g(r))\, g(r) \,\mathrm{d}r \tag{21}$$

For both measures, the divergence of F_X with respect to F_Y is not, in general, equal to the divergence of F_Y with respect to F_X. That is, the direction from which we look at the

relative distribution matters. An example for a symmetric divergence measure[2] is the total variation distance

$$\text{TVD} = \int_{-\infty}^{\infty} \frac{1}{2} \left| \frac{f_X(y)}{f_Y(y)} - 1 \right| f_Y(y)\,\mathrm{d}y = \int_0^1 \frac{1}{2} \left| g(r) - 1 \right| \mathrm{d}r \tag{22}$$

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 case of categorical data the total variation distance is equal to the dissimilarity index by Duncan and Davis (1953), which is often used in analyses of segregation (for a Stata implementation see, e.g., Jann, 2004).

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 (9) 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

$$\text{KL} = \text{KL}_{X\widetilde{Y}Y} + \text{KL}_{X\widetilde{Y}} \tag{23}$$

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

$$\text{KL}_{X\widetilde{Y}Y} = \int_{-\infty}^{\infty} \ln\left(\frac{f_{\widetilde{Y}}(y)}{f_Y(y)}\right) f_X(y)\,\mathrm{d}y = \int_0^1 \ln\big(g_{\widetilde{Y}Y}(r)\big)\, g(r)\,\mathrm{d}r \tag{24}$$

This suggest 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

$$\text{MRP} = 4 \cdot E_X(|r_{\widetilde{Y}}(y) - 0.5|) - 1 \qquad \text{MRP} \in [-1, 1] \tag{25}$$

[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 due to the smoothing involved in density estimation.

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 location measure). The justification for the MRP is that the median of the location-adjusted relative ranks is 0.5 and that 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 zero or one for all y with positive density in F_X, such that $E_X(|r_{\widetilde{Y}}(y) - 0.5|) = .5$ and, hence, MRP $= 1$. In the opposite extreme, $r_{\widetilde{Y}}(y)$ will always be .5, leading to an MRP of -1.

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

$$\text{LRP} = 4 \cdot E_X(\text{abs}(r_{\widetilde{Y}}(y) - 0.5) | r_{\widetilde{Y}}(y) \leq 0.5) - 1 \tag{26}$$

$$\text{URP} = 4 \cdot E_X(\text{abs}(r_{\widetilde{Y}}(y) - 0.5) | r_{\widetilde{Y}}(y) > 0.5) - 1 \tag{27}$$

Since the conditional expectations in the definitions of LRP and URP each cover half of 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

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

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

$$F_{Y|D=d}(y) = \int_{-\infty}^{\infty} f_{Z|D=d}(z) F_{Y|D=d,Z}(y|z) \, \mathrm{d}z \tag{28}$$

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,

$$F^C_{Y|D=0}(y) = \int_{-\infty}^{\infty} f_{Z|D=1}(z) F_{Y|D=0,Z}(y|z) \, \mathrm{d}z \tag{29}$$

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 $F^C_{Y|D=0}$.[3]

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. In any case, constructing counterfactual distributions in this 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.

2.7.2 Reweighting

An equivalent but more attactive approach from an applied perspective is to conceptualize covariate-adjustment as reweighting in the spirit of DiNardo et al. (1996). Define

$$P(D = 1|Z = z) = 1 - P(D = 0|Z = z) \tag{30}$$

as the conditional probability of belonging to the comparison group given Z, where Z is a vector of covariates. Furthermore, define

$$\Psi(z) = \frac{P(D = 1|Z = z)/P(D = 1)}{P(D = 0|Z = z)/P(D = 0)} \tag{31}$$

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

$$F^C_{Y|D=0}(y) = \int_{-\infty}^{\infty} f_{Z|D=0}(z) F_{Y|D=0,Z}(y|z)\Psi(z)\,\mathrm{d}z \tag{32}$$

This indicates that the counterfactual distribution can be estimated by simply reweighting the data by an estimate of $\Psi(z)$.[4] Mathematically, (32) is equivalent to (29) because

$$\Psi(z) = \frac{P(D = 1|Z = z)/P(D = 1)}{P(D = 0|Z = z)/P(D = 0)} = \frac{P(D = 1|Z = z) \cdot \frac{f_Z(z)}{P(D=1)}}{P(D = 0|Z = z) \cdot \frac{f_Z(z)}{P(D=0)}} = \frac{f_{Z|D=1}(z)}{f_{Z|D=0}(z)} \tag{33}$$

[3] 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 (i.e., 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, e.g., Jann, 2008).

[4] To reweight the comparison group, we would use factor $1/\Psi(z)$ instead of $\Psi(z)$.

(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 (e.g. logistic regression).[5]

Note that reweighting can also be used as an alternative approach to identify location and shape effects, by modeling Ψ as a function of Y (instead of applying location adjustments as described in section 2.5).

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 weights w as well as a (possibly empty) vector of covariates Z. The complete data is (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 else) 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 else) 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 subsamples (e.g., 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 (e.g., 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 the 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.

[5] 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 $\Psi(z)$.

3.1 The relative distribution function

To obtain an estimate for the relative CDF

$$G(r) = F_{X|D}(F^{-1}_{Y|R}(r)), \quad r \in [0,1] \tag{34}$$

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

$$\hat{r}_i = \frac{1}{W_R} \sum_{j \in \mathcal{R}} w_j \mathbb{1}\{Y_j \leq X_i\} \qquad \text{for all } i \in \mathcal{D} \tag{35}$$

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

$$\widehat{G}(r) = \frac{1}{W_D} \sum_{i \in \mathcal{D}} w_i \mathbb{1}\{\hat{r}_i \leq r\} \tag{36}$$

An issue with this simple computation is that it leads to a step function with jumps at distinct values of $\hat{r}$. Let (i) refer to observations in $\mathcal{D}$ ordered by $\hat{r}$ such that $\hat{r}_{(1)} \leq \hat{r}_{(2)} \leq \cdots \leq \hat{r}_{(n_D)}$. If $\hat{r}_{(i)} < r < \hat{r}_{(i+1)}$, that is, if evaluation point r falls between two values of $\hat{r}$, then $\widehat{G}(r)$ will be equal to the CDF corresponding to the lower value of $\hat{r}$. In the context of the relative distribution it makes more sense to use linear interpolation[6] between the two points and, hence, determine $\widehat{G}(r)$ as

$$\widehat{G}(r) = \widehat{G}_{(i')} + \left(\widehat{G}_{(i'+1)} - \widehat{G}_{(i')}\right) \frac{r - \hat{r}_{(i')}}{\hat{r}_{(i'+1)} - \hat{r}_{(i')}} \tag{37}$$

with

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

where i' is selected such that $\hat{r}_{(i')} < r \leq \hat{r}_{(i'+1)}$ (with $\hat{r}_{(0)} = \widehat{G}_{(0)} = 0$ if $\hat{r}_{(1)} > 0$ and $\hat{r}_{(n_D+1)} = \widehat{G}_{(n_D+1)} = 1$ if $\hat{r}_{(n_D)} < 1$). For values of r that have an exact match in $\hat{r}_i$, $i \in \mathcal{D}$, this leads to the same result as (36). For r values without exact match, (37) is equivalent to picking the result from a linear segmented curve connecting the points given by $(\widehat{G}_{(i)}, \hat{r}_{(i)})$, $i = 1, \ldots, n_D$.

Equation (37) improves on (36) in that it uses interpolation in regions where (36) is flat. It does not, however, take into account that flat regions in (36) 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 due to 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

[6]Interpolation is equivalent to breaking ties proportionally between the comparison distribution and the reference distribution.

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

$$\hat{r}^D_{(j)} = \frac{1}{W_D} \sum_{i \in \mathcal{D}} w_i \mathbb{1}\{X_i \leq y_{(j)}\} \quad \text{and} \quad \hat{r}^R_{(j)} = \frac{1}{W_R} \sum_{i \in \mathcal{R}} w_i \mathbb{1}\{Y_i \leq y_{(j)}\} \tag{38}$$

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

$$\widehat{G}(r) = \begin{cases} \hat{r}^D_{(j^r)} & \text{if } r = 0 \\ \hat{r}^D_{(j_r)} & \text{if } r = 1 \\ 0.5\left(\hat{r}^D_{(j_r)} + \hat{r}^D_{(j^r)}\right) & \text{if } r = \hat{r}^R_{(j)} \text{ for any } j \\ \hat{r}^D_{(j')} + \left(\hat{r}^D_{(j'+1)} - \hat{r}^D_{(j')}\right) \frac{r - \hat{r}^R_{(j')}}{\hat{r}^R_{(j'+1)} - \hat{r}^R_{(j')}} & \text{else} \end{cases} \tag{39}$$

where j_r and j^r denote the smallest and largest value of j, respectively, for which $\hat{r}^R_{(j)} = r$, and where j' is chosen such that $\hat{r}^R_{(j')} < r < \hat{r}^R_{(j'+1)}$. For graphical display we may also directly plot $\hat{r}^D_{(j)}$ against $\hat{r}^R_{(j)}$ and linearly connect the points. All estimates for $\widehat{G}(r)$ obtained using (39) will lie on that curve.

If all values in $\mathcal{Y}$ exist in both distributions, (39) will lead to the same results as (37). Furthermore, for continuous data, at least if the dataset is not very small, results from the two approaches will be very similar. Equation (39), however, leads to more appropriate results than (37) if the data is 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

$$\hat{r}_i = \frac{1}{W_R} \sum_{j \in \mathcal{R}} w_j \mathbb{1}\{Y_j \leq X_i\} \tag{40}$$

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 mid-distribution function instead of the regular CDF (Parzen, 2004) that deducts half a step size from the ranks in regions where the CDF is upright. Let

$$\widehat{P}_R(Y = y) = \frac{1}{W_R} \sum_{j \in \mathcal{R}} w_j \mathbb{1}\{Y_j = y\} \tag{41}$$

be the relative frequency of outcome y in $F_{Y|R}$ (i.e. the step size in the CDF at value y). The relative ranks computed according to the mid-distribution function then are

$$\hat{r}_i = \frac{1}{W_R} \sum_{j \in \mathcal{R}} w_j \mathbb{1}\{Y_j \leq X_i\} - \frac{1}{2} \widehat{P}_R(Y = X_i) \tag{42}$$

Note that (42) differs from (40) only for observations that have ties in $F_{Y|R}$ (i.e. 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 mid-ranks 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. 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, the larger the bias.

Using the mid-rank adjustment removes the bias in the relative ranks. Heaping, however, will still lead to undesirable results such as arbitrary spikes in the relative 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 mid-rank adjustment) can be written as

$$\hat{r}_i = \frac{1}{W_R} \sum_{j \in \mathcal{R}} w_j \mathbb{1}\{Y_j \leq X_i\} - \widehat{P}_R(Y = X_i) \frac{\widehat{P}_D(X = X_i) + 0.5 w_i - \delta_i}{\widehat{P}_D(X = X_i)} \tag{43}$$

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

$$\delta_i = \frac{1}{\sum_{k=1}^{K_i} w_k^{(i)}} \sum_{k=1}^{k_i} w_k^{(i)} \tag{44}$$

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

Due to the random ordering, repeated computation of (43) will lead to slightly different results for the relative density and other estimates unless the weights are constant (or unless there are on ties). One (arbitrary) solution to enforce stable results is to sort the observations within ties in (ascending or descending) order of the weights.

To obtain broken relative ranks *without* mid-rank adjustment, set $0.5 w_i$ in (43) to zero. Whereas the mid-rank adjustment can have a strong effect on results if relative ranks are computed without breaking ties (equation 40 vs. equation 42), the adjustment is only of minor importance in (43), 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

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

instead of Y in the above equations, where $\hat{\mu}_{Y|R}$ is the median or mean of Y in subsample $\mathcal{R}$ and $\hat{\mu}_{X|D}$ is the median or mean of X in subsample $\mathcal{D}$. Location-and-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 (i.e., a comparison distribution that has the same shape as the reference distribution, but a different location), use

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

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

3.3 The relative density function

3.3.1 Kernel density estimation for continuous data

Estimation of the relative density function can be implemented by applying a univariate density estimator to the relative ranks (preferably as defined in equation 43). Compared to a standard density estimation problem, there are two specific complications that should to be taken into account. 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

$$\hat{g}(r) = \frac{1}{W_D} \sum_{i \in \mathcal{D}} w_i K_c(r, \hat{r}_i, h) \tag{45}$$

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

$$K_c(r, \hat{r}_i, h) = \frac{1}{h} K\left(\frac{r - \hat{r}_i}{h}\right) c(r, h) \qquad \text{with} \qquad c(r, h) = \left(\int_{(0-r)/h}^{(1-r)/h} K(x)\,\mathrm{d}x\right)^{-1} \tag{46}$$

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 density function) 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, 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.[7]

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 domain of r. The histogram density estimate for that interval can then be obtained as

$$\hat{g}(a,b) = \frac{\widehat{P}_D(a < r \leq b)}{b-a} = \frac{1}{W_D} \sum_{i \in \mathcal{D}} w_i \frac{\mathbb{1}\{a < \hat{r}_i \leq b\}}{b-a} \tag{47}$$

(with a modification in case of $a = 0$ such that the interval includes the lower bound). A convenient setup is to split the domain 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 over- or underrepresentation multiplier at each level of r), but the binning may make the histogram more easy to interpret.

[7] Estimator 45 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. For the adaptive estimator, replace h by $h_i = h \cdot \sqrt{\widetilde{g}^0 / \hat{g}^0(\hat{r}_i)}$ where $\hat{g}^0(\hat{r}_i)$ is an initial (constant-bandwidth) density estimate and $\widetilde{g}^0$ is the geometric mean of $\hat{g}^0(\hat{r}_i)$ 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.

3.3.3 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

$$\hat{g}_k = \frac{\widehat{P}_D(X = k)}{\widehat{P}_R(Y = k)} \tag{48}$$

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

When plotting the relative density for categorical data, $\hat{g}_k$ can be plotted against $\widehat{P}_R(Y \leq k)$ using a step function, including an additional point at coordinate $(\hat{g}_1, 0)$ 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 range of r from $1/(2K)$ to $1 - 1/(2K)$. The divergence measures can then be estimated as

$$\widehat{\chi}^2 = \frac{1}{K} \sum_{k=1}^{K} (\hat{g}(r_k) - 1)^2 \qquad \widehat{\text{KL}} = \frac{1}{K} \sum_{k=1}^{K} \hat{g}(r_k) \ln(\hat{g}(r_k)) \qquad \widehat{\text{TVD}} = \frac{1}{2K} \sum_{k=1}^{K} |\hat{g}(r_k) - 1| \tag{49}$$

where $\hat{g}(r_k)$ be 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 (49) with $\hat{g}(r_k)$ 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

$$\widehat{\chi^2} = \sum_{k=1}^{K} \frac{\left(\hat{p}_k^D - \hat{p}_k^R\right)^2}{\hat{p}_k^R} \qquad \widehat{\text{KL}} = \sum_{k=1}^{K} \hat{p}_k^D \ln\left(\frac{\hat{p}_k^D}{\hat{p}_k^R}\right) \qquad \widehat{\text{TVD}} = \sum_{k=1}^{K} \frac{1}{2} \left|\hat{p}_k^D - \hat{p}_k^R\right| \tag{50}$$

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

3.5 Median relative polarization

For the polarization indices first compute location-adjusted relative ranks using one of the above methods, where the median is used as location measure. Let $\hat{\tilde{r}}_i$ 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

$$\widehat{\text{MRP}} = \left(\frac{4}{W_D} \sum_{i \in \mathcal{D}} w_i \left|\hat{\tilde{r}}_i - 0.5\right|\right) - 1 \tag{51}$$

Furthermore, using

$$\widehat{\text{LRP}} = \left(\frac{8}{W_D} \sum_{i \in \mathcal{D}} w_i \left|\hat{\tilde{r}}_i - 0.5\right| \mathbb{1}\left\{\hat{\tilde{r}}_i < .5\right\}\right) - 1 \tag{52}$$

$$\widehat{\text{URP}} = \left(\frac{8}{W_D} \sum_{i \in \mathcal{D}} w_i \left|\hat{\tilde{r}}_i - 0.5\right| \mathbb{1}\left\{\hat{\tilde{r}}_i > .5\right\}\right) - 1 \tag{53}$$

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 mid-distribution 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 $\hat{p}_i = \widehat{P}(D = 1|Z = Z_i)$ from the model. To reweight the reference group, define adjusted weights

$$\widetilde{w}_i = \begin{cases} w_i \frac{\hat{p}_i}{1-\hat{p}_i} c_R & \text{if } i \in \mathcal{R} \\ w_i & \text{else} \end{cases} \tag{54}$$

where $c_R = W_R / \sum_{i \in \mathcal{R}} w_i \frac{\hat{p}_i}{1-\hat{p}_i}$ 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

$$\widetilde{w}_i = \begin{cases} w_i \frac{1-\hat{p}_i}{\hat{p}_i} c_D & \text{if } i \in \mathcal{D} \\ w_i & \text{else} \end{cases} \tag{55}$$

with $c_D = W_D / \sum_{i \in \mathcal{D}} w_i \frac{1-\hat{p}_i}{\hat{p}_i}$. 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.

3.7 Standard errors

3.7.1 Variance estimation by means of influence functions

Influence functions provide a convenient approach to estimate the sampling variances of the different statistics discussed above. Intuitively, an influence function is an approximation of how a functional 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 (i.e., the sampling variance) is equal to the expectation of the square of the influence function divided by the sample size. Therefore, to obtain an estimate of the sampling variance from a given sample, we can evaluate the influence function at each observation in the data and then compute the sampling variance of the mean of these values using textbook formulas.[8] More generally, once influence functions are available for a set of statistics, the variance matrix of these statistics can be obtained by taking a mean estimate (using [R] **mean**) of the influence functions (or a total estimate using [R] **total**, depending on the scaling of the influence functions). Sampling weights or other complex survey characteristics do not change the form of the influence function and can be taken into account when computing the mean (or total) estimate. This makes the influence function approach very general.[9]

[8] Since the mean of an influence function is zero by definition, the expectation of the squared influence function is equal to the variance of the influence function.

[9] Also see Rios-Avila (2020) for an overview of (recentered) influence functions for a variety of statistics.

One-parameter setting There is a close connection between influence functions 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 $\hat{\theta}$ satisfies

$$0 = \frac{1}{W}\sum_{i=1}^{n} w_i \hat{h}_i^\theta \tag{56}$$

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

$$\textsc{if}_i(\hat{\theta}) = \frac{1}{-\widehat{\Delta}^\theta}\hat{h}_i^\theta \tag{57}$$

where

$$\widehat{\Delta}^\theta = \frac{1}{W}\sum_{i=1}^{n} w_i \left.\frac{\partial h_i^\theta}{\partial \theta}\right|_{\theta=\hat{\theta}} \tag{58}$$

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

$$\hat{\bar{y}} = \frac{1}{W}\sum_{i=1}^{n} w_i Y_i$$

for which the moment condition is given as

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

Since

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

the influence function simplifies to

$$\textsc{if}_i(\hat{\bar{y}}) = Y_i - \hat{\bar{y}} \tag{59}$$

The sampling variance of $\hat{\bar{y}}$ can then be estimated as

$$\widehat{V}(\hat{\bar{y}}) = \frac{1}{W(W - W/n)}\sum_{i=1}^{n} w_i^2(\textsc{if}_i(\hat{\bar{y}}))^2 \tag{60}$$

This is equivalent to the textbook formula for the variance of the mean, as can easily be seen if $\textsc{if}_i(\hat{\bar{y}})$ 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 if it depending on survey design) to its influence function, whatever that influence function might be.

Multiple-parameter setting Deriving the influence function becomes more involved if a statistic involves 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 may depend on $\theta_2, \ldots, \theta_p$ (that is, all θ_j appear as arguments in the moment condition for θ_1). The influence function for θ_1 can then be written as

$$\text{IF}_i(\hat{\theta}_1) = \frac{1}{-\widehat{\Delta}^{\theta_1}} \left(\hat{h}_i^{\theta_1} + \sum_{j=2}^{p} \widehat{\Delta}_{\theta_j}^{\theta_1} \text{IF}_i(\hat{\theta}_j) \right) \tag{61}$$

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

$$\widehat{\Delta}_{\theta_j}^{\theta_1} = \frac{1}{W} \sum_{i=1}^{n} w_i \left. \frac{\partial h_i^{\theta_1}}{\partial \theta_j} \right|_{\theta=\hat{\theta}} \tag{62}$$

is an estimate of the expectation of the partial derivative of h^{θ_1} by θ_j at point $\hat{\theta}$. If parameters θ_j, $j \geq 2$, themselves depend on further parameters, their influence functions will have an analogous form. That is, multiple-parameter problems can be solved recursively by applying equation (61) repeatedly.

One implication of (61) 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 (i.e. a linear or nonlinear combination of the elements in θ), the influence function for $\hat{\gamma}$ can be written as

$$\text{IF}_i(\hat{\gamma}) = \left. \frac{\partial t(\theta)}{\partial \theta_1} \right|_{\theta=\hat{\theta}} \text{IF}_i(\hat{\theta}_1) + \cdots + \left. \frac{\partial t(\theta)}{\partial \theta_p} \right|_{\theta=\hat{\theta}} \text{IF}_i(\hat{\theta}_p) \tag{63}$$

This means that the influence function of a statistic that is defined as an aggregate of other statistics can be obtained as an aggregate of the influence functions of these statistics.

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

Consider the influence function for subpopulation mean

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

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 $\hat{\bar{y}}_S$ can be written as

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

Since

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

the influence function for $\hat{\bar{y}}_S$ becomes

$$\text{IF}_i(\hat{\bar{y}}_S) = \frac{W}{W_S} S_i(Y_i - \hat{\bar{y}}_S) \tag{64}$$

The influence function will be zero for observations outside subsample $\mathcal{S}$. However, taking the standard error of the mean of this influence function across all observations will provide a consistent standard error for $\hat{\bar{y}}_S$.

In practice it may be convenient to omit the global W from the definition of the influence function 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 influence function. An advantage of defining influence functions in this way is that they can be computed from the subsample data alone, without knowing the total sum of weights.

Overview of influence functions for various statistics Using the methods above, we can obtain influence functions for various statistics that are relevant in the context of relative distribution analysis. Table 1 provides an overview. The influence functions 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.

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

$$0 = \frac{1}{W}\sum_{i=1}^{n} w_i(\mathbb{1}\{Y_i \leq q_Y(p)\} - p) \tag{65}$$

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

3.7.2 Influence function for the relative CDF

Using notation as introduced at the start of this chapter, the empirical CDF of the relative ranks can be written as

$$\widehat{G}(r) = \widehat{F}_{X|D}(\hat{q}_r) = \frac{1}{W_D}\sum_{i\in\mathcal{D}} w_i \mathbb{1}\{X_i \leq \hat{q}_r\} \tag{66}$$

Table 1: Influence functions for various statistics

Statistic (conditional on $S=1$)	Empirical influence function
Mean	
$\hat{\bar{y}}_S = \frac{1}{W_S}\sum_{i\in\mathcal{S}} w_i Y_i$	$\text{IF}_i = \frac{W}{W_S} S_i\big(Y_i - \hat{\bar{y}}_S\big)$
Empirical CDF	
$\widehat{F}_{Y\|S}(y) = \frac{1}{W_S}\sum_{i\in\mathcal{S}} w_i \mathbb{1}\{Y_i \le y\}$	$\text{IF}_i = \frac{W}{W_S} S_i\Big(\mathbb{1}\{Y_i \le y\} - \widehat{F}_{Y\|S}(y)\Big)$
Kernel PDF (assuming h fixed)	
$\hat{f}_{Y\|S}(y) = \frac{1}{W_S}\sum_{i\in\mathcal{S}} w_i \frac{1}{h} K\left(\frac{y - Y_i}{h}\right)$	$\text{IF}_i = \frac{W}{W_S} S_i\left(\frac{1}{h} K\left(\frac{y - Y_i}{h}\right) - \hat{f}_{Y\|S}(y)\right)$
Histogram PDF	
$\hat{f}_{Y\|S}(a,b) = \frac{1}{W_S}\sum_{i\in\mathcal{S}} w_i \frac{\mathbb{1}\{a < Y_i \le b\}}{b-a}$	$\text{IF}_i = \frac{W}{W_S} S_i\left(\frac{\mathbb{1}\{a < Y_i \le b\}}{b-a} - \hat{f}_{Y\|S}(a,b)\right)$
Quantile	
$\hat{q}_{Y\|S}(p) = \widehat{F}^{-1}_{Y\|S}(p)$	$\text{IF}_i = \frac{W}{W_S} S_i\left(\frac{p - \mathbb{1}\big\{Y_i \le \hat{q}_{Y\|S}(p)\big\}}{\hat{f}_{Y\|S}(\hat{q}_{Y\|S}(p))}\right)$
Median	
$\hat{\tilde{y}}_S = \hat{q}_{Y\|S}(0.5)$	$\text{IF}_i = \frac{W}{W_S} S_i\left(\frac{.5 - \mathbb{1}\Big\{Y_i \le \hat{\tilde{y}}_S\Big\}}{\hat{f}_{Y\|S}(\hat{\tilde{y}}_S)}\right)$
Variance	
$\hat{\sigma}^2_{Y\|S} = \frac{1}{W_S - \frac{W_S}{n_S}}\sum_{i\in\mathcal{S}} w_i (Y_i - \hat{\bar{y}}_S)^2$	$\text{IF}_i = \frac{W}{W_S} S_i\Big(c\cdot(Y_i - \hat{\bar{y}}_S)^2 - \hat{\sigma}^2_{Y\|S}\Big),\ c = \frac{1}{1-\frac{1}{n_S}}$
Standard deviation	
$\hat{\sigma}_{Y\|S} = \sqrt{\hat{\sigma}^2_{Y\|S}}$	$\text{IF}_i = \frac{W}{W_S} S_i\left(\frac{c\cdot(Y_i - \hat{\bar{y}}_S)^2 - \hat{\sigma}^2_{Y\|S}}{2\hat{\sigma}_{Y\|S}}\right),\ c = \frac{1}{1-\frac{1}{n_S}}$
Interquartile range	
$\text{IQR}_{Y\|S} = \hat{q}_{Y\|S}(0.75) - \hat{q}_{Y\|S}(0.25)$	$\text{IF}_i = \text{IF}_i(\hat{q}_{Y\|S}(0.75)) - \text{IF}_i(\hat{q}_{Y\|S}(0.25))$

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

$$h_i^G = D_i(\mathbb{1}\{X_i \leq q_r\} - G(r)) \tag{67}$$
$$h_i^q = R_i(\mathbb{1}\{Y_i \leq q_r\} - r) \tag{68}$$

Working through (61) yields

$$\text{IF}_i(\widehat{G}(r)) = \frac{W}{W_D}\hat{h}_i^G + \hat{f}_{X|D}(\hat{q}_r)\text{IF}_i(\hat{q}_r) \tag{69}$$

Since, according to table 1,

$$\text{IF}_i(\hat{q}_r) = \frac{W}{W_R}R_i\left(\frac{r - \mathbb{1}\{Y_i \leq \hat{q}_r\}}{\hat{f}_{Y|R}(\hat{q}_r)}\right) \tag{70}$$

equation (69) can be written as

$$\text{IF}_i(\widehat{G}(r)) = W\frac{D_i}{W_D}(\mathbb{1}\{X_i \leq \hat{q}_r\} - \widehat{G}(r)) + W\frac{R_i}{W_R}\frac{\hat{f}_{X|D}(\hat{q}_r)}{\hat{f}_{Y|R}(\hat{q}_r)}(r - \mathbb{1}\{Y_i \leq \hat{q}_r\}) \tag{71}$$

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

Location and scale adjustment For the relative CDF based on location (and, possibly, scale) adjusted data, replace q_r by $\tilde{q}_r = t(q_r, \theta)$ 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

$$\tilde{q}_r = t(q_r, \theta) = t_D^{-1}(t_R(q_r, \theta), \theta) \tag{72}$$

The adjusted relative CDF then becomes

$$\widehat{\widetilde{G}}(r) = \widehat{F}_D(\hat{\tilde{q}}_r) = \frac{1}{W_D}\sum_{i\in\mathcal{D}} w_i \mathbb{1}\Big\{X_i \leq \hat{\tilde{q}}_r\Big\} \tag{73}$$

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

$$\text{IF}_i(\widehat{\widetilde{G}}(r)) = \frac{W}{W_D}\hat{h}_i^{\widetilde{G}} + \hat{f}_{X|D}(\hat{\tilde{q}}_r)\text{IF}_i(\hat{\tilde{q}}_r) \tag{74}$$

with

$$h_i^{\widetilde{G}} = D_i(\mathbb{1}\{X_i \leq \tilde{q}_r\} - \widetilde{G}(r)) \tag{75}$$

and

$$\text{IF}_i(\hat{\tilde{q}}_r) = \tau_{q_r}\text{IF}_i(\hat{q}_r) + \sum_{k=1}^{K}\tau_{\theta_k}\text{IF}_i(\hat{\theta}_k) \tag{76}$$

where

$$\tau_{q_r} = \left.\frac{\partial t(q_r,\theta)}{\partial q_r}\right|_{q_r=\hat{q}_r,\theta=\hat{\theta}} \qquad \text{and} \qquad \tau_{\theta_k} = \left.\frac{\partial t(q_r,\theta)}{\partial \theta_k}\right|_{q_r=\hat{q}_r,\theta=\hat{\theta}} \tag{77}$$

For example, in 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

$$\text{IF}_i(\hat{\tilde{q}}_r) = \text{IF}_i(\hat{q}_r) - \text{IF}_i(\hat{\mu}_{Y|R}) + \text{IF}_i(\hat{\mu}_{X|D}) \tag{78}$$

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

$$\text{IF}_i(\hat{\tilde{q}}_r) = \frac{\hat{\mu}_{X|D}}{\hat{\mu}_{Y|R}}\text{IF}_i(\hat{q}_r) + \frac{\hat{q}_r}{\hat{\mu}_{Y|R}}\text{IF}_i(\hat{\mu}_{X|D}) - \frac{\hat{q}_r\hat{\mu}_{X|D}}{(\hat{\mu}_{Y|R})^2}\text{IF}_i(\hat{\mu}_{Y|R}) \tag{79}$$

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

$$\begin{aligned}\text{IF}_i(\hat{\tilde{q}}_r) = {} & \frac{\hat{s}_{X|D}}{\hat{s}_{Y|R}}\text{IF}_i(\hat{q}_r) - \frac{\hat{s}_{X|D}}{\hat{s}_{Y|R}}\text{IF}_i(\hat{\mu}_{Y|R}) + \frac{\hat{q}_r - \hat{\mu}_{Y|R}}{\hat{s}_{Y|R}}\text{IF}_i(\hat{s}_{X|D}) \\ & - \frac{(\hat{q}_r - \hat{\mu}_{Y|R})\hat{s}_{X|D}}{(\hat{s}_{Y|R})^2}\text{IF}_i(\hat{s}_{Y|R}) + \text{IF}_i(\hat{\mu}_{X|D})\end{aligned} \tag{80}$$

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

$$\begin{aligned}\text{IF}_i(\hat{\tilde{q}}_r) = \hat{\tilde{q}}_r \Bigg(& \frac{\hat{s}_{\ln(X)|D}}{\hat{q}_r\hat{s}_{\ln(Y)|R}}\text{IF}_i(\hat{q}_r) - \frac{\hat{s}_{\ln(X)|D}}{\hat{s}_{\ln(Y)|R}}\text{IF}_i(\hat{\mu}_{\ln(Y)|R}) + \frac{\ln\hat{q}_r - \hat{\mu}_{\ln(Y)|R}}{\hat{s}_{\ln(Y)|R}}\text{IF}_i(\hat{s}_{\ln(X)|D}) \\ & - \frac{(\ln\hat{q}_r - \hat{\mu}_{\ln(Y)|R})\hat{s}_{\ln(X)|D}}{(\hat{s}_{\ln(Y)|R})^2}\text{IF}_i(\hat{s}_{\ln(Y)|R}) + \text{IF}_i(\hat{\mu}_{\ln(X)|D})\Bigg)\end{aligned} \tag{81}$$

In case of a shape adjustment, one of the two distributions is replaced by a location (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 (75) have to be replaced by R_i and Y_i. This further implies that W_D

in the first term of (74) has to be replaced by W_R and that the density in the second term is $\hat{f}_{Y|R}(\hat{\tilde{q}}_r)$ instead of $\hat{f}_{X|D}(\hat{\tilde{q}}_r)$. 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 $\hat{q}_r$ is now based on the comparison distribution, that is, $\hat{q}_r = \widehat{F}^{-1}_{X|D}(r)$, such that the definition of $\text{IF}_i(\hat{q}_r)$ in (76) changes.

3.7.3 Influence function for the relative histogram

For the influence function 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

$$\hat{g}(a, b) = \frac{\widehat{G}(b) - \widehat{G}(a)}{b - a} \tag{82}$$

The influence function for $\hat{g}(a, b)$ can thus be obtained as

$$\text{IF}_i(\hat{g}(a, b)) = \frac{\text{IF}_i(\widehat{G}(b)) - \text{IF}_i(\widehat{G}(a))}{b - a} \tag{83}$$

3.7.4 Influence function for the relative PDF

The relative density estimate for continuous data can be written as

$$\hat{g}(r) = \frac{1}{W_D} \sum_{i \in \mathcal{D}} w_i K_c(r, \hat{r}_i, h) \qquad \text{with } \hat{r}_i = \widehat{F}_{Y|R}(X_i) \tag{84}$$

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

$$h_i^g = D_i(K_c(r, r_i, h) - g(r)) \tag{85}$$

$$h_i^{r_j} = R_i(\mathbb{1}\{Y_i \leq X_j\} - r_j) \qquad \text{for each } j \in \mathcal{D} \tag{86}$$

This leads to

$$\begin{aligned} \text{IF}_i(\hat{g}(r)) &= \frac{W}{W_D} \left(\hat{h}_i^g + \sum_{j \in \mathcal{D}} \widehat{\Delta}^g_{r_j} \text{IF}_i(\hat{r}_j) \right) \\ &= \frac{W}{W_D} \left(\hat{h}_i^g + \frac{R_i}{W_R} \sum_{j \in \mathcal{D}} \delta_j (\mathbb{1}\{Y_i \leq X_j\} - \hat{r}_j) \right) \end{aligned} \tag{87}$$

with $\widehat{\Delta}^g_{r_j} = \delta_j / W$ and $\delta_j = w_j K_c'(r, \hat{r}_j, h)$, where $K_c'(r, \hat{r}_j, h)$ is the derivative of $K_c(r, r_j, h)$ with respect to r_j at point $\hat{r}_j$. 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

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

such that

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

Term λ_i is equivalent to a "reverse" (summation from the top) and non-normalized CDF of X weighted by δ_j and can be obtained for all observations in a single run across the data.[10]

Location and scale adjustment For the relative PDF based on location (and, possibly, scale) adjusted data, define $\tilde{x}_i = t^{-1}(X_i, \theta)$ and replace r_i by $\tilde{r}_i = F_{Y|R}(\tilde{x}_i)$ in the above formulas. Function $t(x, \theta)$ is as defined in section 3.7.2; 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

$$t^{-1}(x, \theta) = t_R^{-1}(t_D(x, \theta), \theta) \tag{88}$$

The adjusted relative PDF can thus be written as

$$\hat{\tilde{g}}(r) = \frac{1}{W_D}\sum_{i \in \mathcal{D}} w_i K_c(r, \hat{\tilde{r}}_i, h) \qquad \text{with } \hat{\tilde{r}}_i = \widehat{F}_{Y|R}(\hat{\tilde{x}}_i) \text{ and } \hat{\tilde{x}}_i = t^{-1}(X_i, \hat{\theta}) \tag{89}$$

such that the influence function becomes

$$\text{IF}_i(\hat{\tilde{g}}(r)) = \frac{W}{W_D}\left(\hat{h}_i^{\tilde{g}} + \sum_{j \in \mathcal{D}} \widehat{\Delta}_{\tilde{r}_j}^{\tilde{g}} \text{IF}_i(\hat{\tilde{r}}_j)\right) \tag{90}$$

with

$$\hat{h}_i^{\tilde{g}} = D_i(K_c(r, \hat{\tilde{r}}_i, h) - \hat{\tilde{g}}(r)) \tag{91}$$

$$\widehat{\Delta}_{\tilde{r}_j}^{\tilde{g}} = \delta_j / W \qquad \text{with } \delta_j = w_j K_c'(r, \hat{\tilde{r}}_j, h) \tag{92}$$

$$\text{IF}_i(\hat{\tilde{r}}_j) = \frac{W}{W_R}\left(\hat{h}_i^{\tilde{r}_j} + \widehat{\Delta}_{\tilde{x}_j}^{\tilde{r}_j} \text{IF}_i(\hat{\tilde{x}}_j)\right) \tag{93}$$

$$\hat{h}^{\tilde{r}_j} = R_i(\mathbb{1}\{Y_i \leq \hat{\tilde{x}}_j\} - \hat{\tilde{r}}_j) \tag{94}$$

$$\widehat{\Delta}_{\tilde{x}_j}^{\tilde{r}_j} = \frac{W_R}{W}\hat{f}_{Y|R}(\hat{\tilde{x}}_j) \tag{95}$$

$$\text{IF}_i(\hat{\tilde{x}}_j) = \sum_{k=1}^{K} \tau_{jk} \text{IF}_i(\hat{\theta}_k) \qquad \text{with } \tau_{jk} = \left.\frac{\partial t^{-1}(X_j, \theta)}{\partial \theta_k}\right|_{\theta = \hat{\theta}} \tag{96}$$

[10] For the adaptive kernel (see footnote 7) 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 (i.e., treating the local bandwidth as fixed) should produce acceptable results in practice.

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

$$\text{IF}_i(\hat{\tilde{g}}(r)) = \frac{W}{W_D}\left(\hat{h}_i^{\tilde{g}} + \frac{R_i}{W_R}(\lambda_i - \Lambda) + \sum_{k=1}^{K}\kappa_k \text{IF}_i(\hat{\theta}_k)\right) \tag{97}$$

with

$$\lambda_i = \sum_{j\in\mathcal{D}}\delta_j \mathbb{1}\{Y_i \leq \hat{\tilde{x}}_j\}, \qquad \Lambda = \sum_{j\in\mathcal{D}}\delta_j \hat{\tilde{r}}_j, \qquad \kappa_k = \frac{1}{W}\sum_{j\in\mathcal{D}}\delta_j \hat{f}_{Y|R}(\hat{\tilde{x}}_j)\tau_{jk} \tag{98}$$

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

3.7.5 Influence function for the discrete relative density

The categorical relative density is defined as

$$\hat{g}^k = \hat{p}_k^D / \hat{p}_k^R \tag{99}$$

with moment conditions

$$h_i^{p_D^k} = D_i(\mathbb{1}\{X_i = k\} - p_k^D) \quad \text{and} \quad h_i^{p_R^k} = R_i(\mathbb{1}\{Y_i = k\} - p_k^R) \tag{100}$$

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

$$\text{IF}_i(\hat{g}^k) = \frac{1}{\hat{p}_k^R}\text{IF}_i(\hat{p}_k^D) - \frac{\hat{p}_k^D}{(\hat{p}_k^R)^2}\text{IF}_i(\hat{p}_k^R) \tag{101}$$

with

$$\text{IF}_i(\hat{p}_k^D) = \frac{W}{W_D}D_i(\mathbb{1}\{X_i = k\} - \hat{p}_k^D) \quad \text{and} \quad \text{IF}_i(\hat{p}_k^R) = \frac{W}{W_R}R_i(\mathbb{1}\{Y_i = k\} - \hat{p}_k^R) \tag{102}$$

3.7.6 Influence functions for divergence measures

Divergence measures are obtained as aggregates of relative density estimates. Hence, their influence functions can be written as aggregates of the influence functions of the density estimates. Assuming the divergence measures are computed form a kernel density estimate on a regular grid or from a histogram density with K evenly sized bins, as described in section 3.4.1, we get

$$\text{IF}_i(\widehat{\chi^2}) = \frac{2}{K}\sum_{k=1}^{K}(\hat{g}_k - 1)\text{IF}_i(\hat{g}_k) \tag{103}$$

$$\text{IF}_i(\widehat{\text{KL}}) = \frac{1}{K}\sum_{k=1}^{K}\left(1 + \ln(\hat{g}_k)\right)\text{IF}_i(\hat{g}_k) \tag{104}$$

$$\text{IF}_i(\widehat{\text{TVD}}) = \frac{1}{2K}\sum_{k=1}^{K}\text{sign}(\hat{g}_k - 1)\text{IF}_i(\hat{g}_k) \tag{105}$$

where $\hat{g}_k$ is the kernel density estimate at evaluation point r_k or the histogram estimate for bin k. For divergence measures computed from categorical data (see section 3.4.2), the influence functions can be written as

$$\text{IF}_i(\widehat{\chi^2}) = \sum_{k=1}^{K} 2\left(\frac{\hat{p}_k^D}{\hat{p}_k^R} - 1\right)\text{IF}_i(\hat{p}_k^D) + \left(1 - \left(\frac{\hat{p}_k^D}{\hat{p}_k^R}\right)^2\right)\text{IF}_i(\hat{p}_k^R) \tag{106}$$

$$\text{IF}_i(\widehat{\text{KL}}) = \sum_{k=1}^{K}\left(1 + \ln\left(\frac{\hat{p}_k^D}{\hat{p}_k^R}\right)\right)\text{IF}_i(\hat{p}_k^D) - \frac{\hat{p}_k^D}{\hat{p}_k^R}\text{IF}_i(\hat{p}_k^R) \tag{107}$$

$$\text{IF}_i(\widehat{\text{TVD}}) = \frac{1}{2}\sum_{k=1}^{K}\text{sign}(\hat{p}_k^D - \hat{p}_k^R)\left(\text{IF}_i(\hat{p}_k^D) - \text{IF}_i(\hat{p}_k^R)\right) \tag{108}$$

where $\text{IF}_i(\hat{p}_k^D)$ and $\text{IF}_i(\hat{p}_k^R)$ are as defined in section 3.7.5.

3.7.7 Influence functions for polarization indices

The MRP can be written as

$$\text{MRP} = \frac{1}{W_D}\sum_{i\in\mathcal{D}} w_i(4|\hat{\tilde{r}}_i - 0.5| - 1) \qquad \text{with } \hat{\tilde{r}}_i = \widehat{F}_{Y|R}(\hat{\tilde{x}}_i) \text{ and } \hat{\tilde{x}}_i = t^{-1}(X_i, \hat{\theta}) \tag{109}$$

where $t^{-1}(x, \theta)$ is as defined in (88). We see that the MRP has the same structure as an estimate of the relative PDF based on location (and, possibly, scale) adjusted data. We can thus obtain the influence function using (97) with $h_i^{\tilde{g}}$ replaced by

$$h_i^{\text{MRP}} = D_i((4|\tilde{r}_i - 0.5| - 1) - \text{MRP}) \tag{110}$$

and δ_j set to

$$\delta_j = w_j\, 4\,\text{sign}(\hat{\tilde{r}}_j - 0.5) \tag{111}$$

Likewise, the influence function for the LRP can be obtained by using

$$h_i^{\text{LRP}} = D_i((8|\tilde{r}_i - 0.5|\mathbb{1}\{\tilde{r}_i < .5\} - 1) - \text{LRP}) \tag{112}$$

and

$$\delta_j = w_j\, 8\,\text{sign}(\hat{\tilde{r}}_j - 0.5)\mathbb{1}\{\hat{\tilde{r}}_j < .5\} \tag{113}$$

Note that $\mathbb{1}\{\tilde{r}_i < .5\}$ always selects half of the comparison data because the data has been median adjusted. Assuming it fixed should not introduce significant bias into the variance estimates. The influence function for the URP can be derived analogously.

3.7.8 Influence functions 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* $\hat{\mu}$ of the (possibly adjusted) relative ranks, the influence function can be obtained by replacing $h_i^{\tilde{g}}$ in equation (97) by

$$h_i^{\mu} = D_i(\tilde{r}_i - \mu) \tag{114}$$

and setting δ_j to

$$\delta_j = w_j \tag{115}$$

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

$$h_i^{\sigma^2} = D_i((\tilde{r}_i - \mu)^2 - \sigma^2) \qquad \text{and} \qquad \delta_j = 2w_j(\hat{\tilde{r}}_j - \hat{\mu}) \tag{116}$$

The influence function for the *standard deviation* $\hat{\sigma}$ is given as:

$$\text{IF}_i(\hat{\sigma}) = \frac{1}{2\hat{\sigma}}\text{IF}_i(\hat{\sigma}^2) \tag{117}$$

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

$$q(p) = G^{-1}(p) = F_{Y|R}(F^{-1}_{X|D}(p)) \tag{118}$$

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 influence function as in section 3.7.2, but with swapped distributions. Finally, the influence function for the *interquartile range* is given as:

$$\text{IF}_i(\text{IQR}) = \text{IF}_i(\hat{q}(0.75)) - \text{IF}_i(\hat{q}(0.25)) \tag{119}$$

3.7.9 Influence functions in case of covariate balancing

If covariates are balanced using the reweighting approach, the influence functions need to be adjusted to take account of the fact that the balancing weights have been estimated. I will discuss two reweighting methods below: IPW based on logistic regression and entropy balancing. Deriving the influence functions is relatively easy in these cases as the weights are obtained from a parametric model. Non-parametric reweighting methods such as matching are more challenging; I leave it to future research to work out the details for such methods.

IPW and entropy balancing both estimate a vector of parameters β from which the balancing weights are computed. To adjust the influence functions we may apply equation (61) to each relative distribution parameter θ that depends on the balancing weights. That is, for each θ, obtain $\hat{\Delta}^{\theta}_{\beta}$ and then include $\hat{\Delta}^{\theta}_{\beta}\text{IF}_i(\hat{\beta})$ as an additional component in the influence function, where $\text{IF}_i(\hat{\beta})$ is the influence function of $\hat{\beta}$.

To keep notation simple, assume that $\mathcal{D}$ and $\mathcal{R}$ are distinct and exhaustive. Furthermore, let T be an indicator for the "treatment" group. If the reference group is reweighted, then $T = D$; if the comparison group is reweighted, then $T = R$. In case of logistic regression IPW the balancing weights are obtained as

$$\widetilde{w}_i = w_i \left(\frac{\hat{p}_i}{1 - \hat{p}_i} c \right)^{(1-T_i)} \qquad \text{with} \qquad \hat{p}_i = \frac{e^{Z_i \hat{\beta}}}{1 + e^{Z_i \hat{\beta}}} \tag{120}$$

where Z_i is a vector of covariates (typically including a constant) and $\hat{\beta}$ is a coefficient vector estimated by logistic regression (c is a scaling factor ensuring that the sum of weights remains constant; it is irrelevant for the derivation of the influence functions). As discussed in Jann (2020), the logistic regression moment conditions for β can be written as

$$h_i^\beta = Z_i'(T_i - p_i) \tag{121}$$

such that

$$\text{IF}(\hat{\beta}) = \frac{1}{-\hat{\Delta}^\beta} Z_i'(T_i - \hat{p}_i) \qquad \text{with} \qquad \hat{\Delta}^\beta = \frac{1}{W} \sum_{i=1}^{N} w_i Z_i' \hat{p}_i (1 - \hat{p}_i) Z_i \tag{122}$$

For entropy balancing, the weights can be written as

$$\widetilde{w}_i = w_i \left(e^{Z_i \hat{\beta} + c} \right)^{(1-T_i)} \tag{123}$$

where Z_i is a vector of covariates without a constant. The estimation of β involves a vector of auxiliary parameters μ, the means of Z in the treatment group. Based on Jann (2020) the entropy balancing moment conditions can be written as

$$\begin{aligned} h_i^\beta &= (1 - T_i) e^{Z_i \beta + c} (Z_i' - \mu) \\ h_i^\mu &= T_i (Z_i' - \mu) \end{aligned}$$

such that

$$\text{IF}(\hat{\beta}) = \frac{1}{-\hat{\Delta}^\beta} \left(\hat{h}_i^\beta + \hat{\Delta}_\mu^\beta \frac{1}{-\hat{\Delta}^\mu} \hat{h}_i^\mu \right) \tag{124}$$

with

$$\hat{\Delta}^\beta = \frac{1}{W} \sum_{i=1}^{N} w_i \hat{h}_i^\beta Z_i = \frac{1}{W} \sum_{i=1}^{N} \widetilde{w}_i (1 - T_i)(Z_i' - \hat{\mu}) Z_i \tag{125}$$

Since

$$\begin{aligned} \hat{\Delta}^\mu &= \frac{1}{W} \sum_{i=1}^{N} -w_i T_i = \frac{-W_T}{W} \\ \hat{\Delta}_\mu^\beta &= \frac{1}{W} \sum_{i=1}^{N} -w_i (1 - T_i) e^{Z_i \hat{\beta} + c} = \frac{1}{W} \sum_{i=1}^{N} -\widetilde{w}_i (1 - T_i) = \frac{-\widetilde{W}_{!T}}{W} \end{aligned}$$

the influence function simplifies to

$$\text{IF}(\hat{\beta}) = \frac{1}{-\hat{\Delta}^{\beta}} \left(\hat{h}_i^{\beta} - \frac{\widetilde{W}_{!T}}{W_T} \hat{h}_i^{\mu} \right) \tag{126}$$

We can now start integrating $\text{IF}(\hat{\beta})$ into the various relative distribution influence functions discussed above. The required adjustments will always have the same form. Let h_i^{θ} be the moment condition for relative distribution parameter θ in the non-reweighted case. If reweighting is applied, the moment condition becomes

$$\widetilde{h}_i^{\theta} = \frac{\widetilde{w}_i}{w_i} h_i^{\theta} \tag{127}$$

For both IPW and entropy balancing the vector of partial derivatives of $\widetilde{h}^{\theta}$ by β can be written as

$$\frac{\partial \widetilde{h}_i^{\theta}}{\partial \beta} = (1 - T_i) \widetilde{h}_i^{\theta} Z_i = \frac{\widetilde{w}_i}{w_i} (1 - T_i) h_i^{\theta} Z_i \tag{128}$$

The respective adjustment to be made to the relative distribution influence function is thus given by addend

$$\hat{\Delta}_{\beta}^{\theta} \text{IF}_i(\hat{\beta}) \qquad \text{with} \qquad \hat{\Delta}_{\beta}^{\theta} = \frac{1}{W} \sum_{i=1}^{N} \widetilde{w}_i (1 - T_i) \hat{h}_i^{\theta} Z_i \tag{129}$$

For parameters θ that do not depend on the reweighting, that is, for parameters that are computed based on observations for which $T_i = 1$, the addend is zero and drops out of the equation.

Consider the influence function for the relative CDF in a situation in which the comparison group is subject to reweighting (that is, $T = R$). In this case

$$\widetilde{h}_i^{G} = \frac{\widetilde{w}_i}{w_i} h_i^{G} = \frac{\widetilde{w}_i}{w_i} D_i (\mathbb{1}\{X_i \leq q_r\} - G(r)) \tag{130}$$

$$h_i^{q} = R_i (\mathbb{1}\{Y_i \leq q_r\} - r) \tag{131}$$

such that

$$\text{IF}_i(\widehat{G}(r)) = \frac{W}{W_D} \left(\frac{\widetilde{w}_i}{w_i} \hat{h}_i^{G} + \hat{\Delta}_{\beta}^{G} \text{IF}_i(\hat{\beta}) \right) + \hat{f}_{X|\widetilde{D}}(\hat{q}_r) \text{IF}_i(\hat{q}_r) \tag{132}$$

where $f_{X|\widetilde{D}}(q_r)$ is the density of q_r in the reweighted comparison group. Likewise, if the reference group is reweighted ($T = D$) we have

$$h_i^{G} = D_i (\mathbb{1}\{X_i \leq q_r\} - G(r)) \tag{133}$$

$$\widetilde{h}_i^{q} = \frac{\widetilde{w}_i}{w_i} h_i^{q} = \frac{\widetilde{w}_i}{w_i} R_i (\mathbb{1}\{Y_i \leq q_r\} - r) \tag{134}$$

such that

$$\text{IF}_i(\widehat{G}(r)) = \frac{W}{W_D}\hat{h}_i^G + \hat{f}_{X|D}(\hat{q}_r)\widetilde{\text{IF}}_i(\hat{q}_r) \tag{135}$$

with

$$\text{IF}_i(\hat{q}_r) = \frac{W}{-W_R\hat{f}_{Y|\widetilde{R}}(\hat{q}_r)}\left(\frac{\widetilde{w}_i}{w_i}\hat{h}_i^q + \hat{\Delta}_\beta^q \text{IF}_i(\hat{\beta})\right) \tag{136}$$

If reweighting is combined with location and scale adjustment, similar addends have to be integrated into the influence functions of the location and scale measures computed from the reweighted group.

The adjustments for the influence function of the relative PDF and related statistics such as the MRP or the mean of the relative ranks are straightforward if the comparison distribution is reweighted (similar to the adjustments for the relative CDF above). Things are somewhat more involved if reweighting is applied to the reference distribution. In general, the moment conditions of such a statistic θ can be written as

$$h_i^\theta = D_i(\nu(r_i) - \theta) \tag{137}$$

$$\widetilde{h}_i^{r_j} = \frac{\widetilde{w}_i}{w_i}h_i^{r_j} = \frac{\widetilde{w}_i}{w_i}R_i(\mathbb{1}\{Y_i \leq X_j\} - r_j) \qquad \text{for each } j \in \mathcal{D} \tag{138}$$

where $\nu(r_i)$ is a function depending on the definition of the statistic. The influence function can then be written as

$$\text{IF}_i(\hat{\theta}) = \frac{W}{W_D}\left(\hat{h}_i^\theta + \sum_{j\in\mathcal{D}}\widehat{\Delta}_{r_j}^\theta \widetilde{\text{IF}}_i(\hat{r}_j)\right) \tag{139}$$

with

$$\widehat{\Delta}_{r_j}^\theta = \frac{\delta_j}{W}, \quad \delta_j = w_j\frac{\partial\nu(r_j)}{\partial r_j} \qquad \text{for each } j \in \mathcal{D} \tag{140}$$

and

$$\widetilde{\text{IF}}_i(\hat{r}_j) = \frac{W}{W_R}\left(\frac{\widetilde{w}_i}{w_i}\hat{h}_i^{r_j} + \hat{\Delta}_\beta^{r_j}\text{IF}_i(\hat{\beta})\right) \tag{141}$$

with

$$\hat{\Delta}_\beta^{r_j} = \frac{1}{W}\sum_{i=1}^N \widetilde{w}_i(1-T_i)\hat{h}_i^{r_j}Z_i = \frac{1}{W}\sum_{i=1}^N \widetilde{w}_i R_i(\mathbb{1}\{Y_i \leq X_j\} - \hat{r}_j)Z_i \tag{142}$$

The problem with these expressions is that they are computationally burdensome: for each $i \in \mathcal{R}$ we have to evaluate a sum over $\mathcal{D}$, which itself contains another sum over $\mathcal{R}$. The first part of the second component in (139) can be written as

$$\sum_{j\in\mathcal{D}}\frac{\delta_j}{W}\frac{W}{W_R}\frac{\widetilde{w}_i}{w_i}\hat{h}_i^{r_j} = \frac{R_i}{W_R}\frac{\widetilde{w}_i}{w_i}\left(\sum_{j\in\mathcal{D}}\delta_j\mathbb{1}\{Y_i \leq X_j\} - \sum_{j\in\mathcal{D}}\delta_j\hat{r}_j\right) = \frac{R_i}{W_R}\frac{\widetilde{w}_i}{w_i}(\lambda_i - \Lambda) \tag{143}$$

where λ_i and Λ are as defined above and allow efficient computation. The second part can be rewritten as

$$\sum_{j\in\mathcal{D}} \frac{\delta_j}{W}\frac{W}{W_R}\hat{\Delta}_\beta^{r_j}\,\mathrm{IF}_i(\hat{\beta}) = \frac{1}{W_R}\mathrm{IF}_i(\hat{\beta})\sum_{j\in\mathcal{D}}\delta_j\hat{\Delta}_\beta^{r_j} \tag{144}$$

with

$$\begin{aligned}\sum_{j\in\mathcal{D}}\delta_j\hat{\Delta}_\beta^{r_j} &= \sum_{j\in\mathcal{D}}\delta_j\frac{1}{W}\sum_{i=1}^{N}\widetilde{w}_i R_i(\mathbb{1}\{Y_i\le X_j\}-\hat{r}_j)Z_i \\ &= \frac{1}{W}\sum_{i=1}^{N}\widetilde{w}_i R_i\left(\sum_{j\in\mathcal{D}}\delta_j(\mathbb{1}\{Y_i\le X_j\}-\hat{r}_j)\right)Z_i \\ &= \frac{1}{W}\sum_{i=1}^{N}\widetilde{w}_i R_i(\lambda_i-\Lambda)Z_i \end{aligned} \tag{145}$$

Expression (145) can be computed upfront using a single run across the data once λ_i is available.

4 The reldist command

Stata 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
```

or

```
. net from https://raw.githubusercontent.com/benjann/reldist/master/
. net install reldist, replace
. net from https://raw.githubusercontent.com/benjann/moremata/master/
. net install moremata, replace
```

The versions at GitHub might be updated more frequently than the versions at the SSC Archive.

4.1 Syntax

Estimation 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, *subcmd* can be

`pdf`	estimate the density function of the relative distribution, possibly including a histogram of the relative density
`histogram`	estimate a histogram of the relative density
`cdf`	estimate the relative distribution function (equivalent to a so-called probability-probability plot)
`divergence`	estimate the Kullback-Leibler divergence (entropy), the Chi-squared divergence, or the dissimilarity index (total variation distance) of the relative distribution
`mrp`	estimate the median relative polarization index (MRP), as well as its decomposition into a lower and and upper polarization index (LRP and URP)
`summarize`	estimate summary statistics such as the mean or the median of the relative ranks and, optionally, store the relative ranks in a new variable

and `fweight`s, `pweight`s, and `iweight`s are allowed; see [U] **11.1.6 weight**.

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

`reldist graph` [, *graph_options*]

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

Storing influence functions after estimation Command `predict` can be applied after `reldist` to generate the influence functions 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 or, 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 influence functions; see page 41 for a description of available *density_options* (option `boundary()` will have no effect, as unbounded domain is assumed for auxiliary densities).

Command `total` (see [R] **total**) can be applied to the stored influence functions to replicate the standard errors reported by `reldist`.

4.2 Options for reldist

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 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

$$[\, \textit{method}\texttt{:}\,]\ \textit{varlist}\ [\,\texttt{,}\ \textit{options}\,]$$

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(, reference)`.

`contrast` compares the balanced distribution to 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 [R] **logit**. *logit_options* are only allowed if *method* is `ipw`.

`btolerance(#)`, with $\# \geq 0$, 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 (interquartile range).

multiplicative uses a multiplicative adjustment instead of an additive adjustment. *adjust* may only contain one keyword in this case, either location or shape. 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. Error will be returned if the data is not strictly positive.

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. Option nobreak has no effect if specified with reldist histogram and reldist cdf.

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. Option `nomid` has no effect if specified with `reldist histogram` and `reldist 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. Furthermore, it has no effect if specified with `reldist histogram` and `reldist cdf`.

replace allows replacing existing variables. This is relevant for `generate()` with `reldist summarize` and `generate()` in `balance()`.

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, either by providing 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 from the reference distribution, specify `atx(comparison)` or `atx(reference)`, respectively. Alternatively, specify a grid of custom values, either by providing 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 like `discrete`, but additionally requests that the data only contains 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 plug-in), dpi[(#)] (Sheather-Jones direct plug-in estimate, where # specifies the number of stages of functional estimation; 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 et al. (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. Default is bwadjust(1).

boundary(*method*) sets the type of boundary correction method. Choices are renorm (renormalization method), 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) (non-adaptive density estimator).

kernel(*kernel*) specifies the kernel function to be used. *kernel* may be 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 command 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 an evaluation grid on the probability scale. If the relative density is evaluated with respect to specific outcome values (e.g. if `atx()` is specified), the outcome labels will be obtained from the information stored in `e(at)`.

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` like `discrete`, but additionally requests that the data only contains positive integers. Factor-variable 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 command `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`.

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, either by providing 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 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, either by providing 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.

`discrete` 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.

`categorical` like `discrete`, but additionally requests that the data only contains positive integers. Factor-variable 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 command `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 an evaluation grid on the probability scale. If the relative CDF is evaluated with respect to specific outcome values (e.g. if `atx()` is specified), the outcome labels will be obtained from the information stored in `e(at)`.

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 Chi-squared divergence of the relative distribution.

tvd or dissimilarity computes the dissimilarity index (total variation distance) 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 the kernel density estimation. This is only relevant if option pdf is specified. See page 41 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 and the width of bars will be proportional to the reference distribution. Option alt has no effect and options n(), pdf, and adjust() are not allowed if discrete is specified.

categorical like discrete, but additionally requests that the data only contains 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.

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 (interquartile range) 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.

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.

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 influence functions. Likewise, `vce(cluster` *clustvar*`)` computes influence-function 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 influence functions (see page 41 for details; option `boundary()` will have no effect, as unbounded support is assumed for auxiliary densities).

`vce(svy)` computes standard errors taking the survey design as set by [SVY] **svyset** into account. The syntax is equivalent 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 influence functions; use

density_options to specify the details of auxiliary density estimation in this case. For *svy_vcetype* other than `linearized`, *density_options* are not allowed.

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

If a replication technique is used for standard error estimation (i.e. `vce(bootstrap)`, `vce(jackknife)`, `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 bandwidth search in each replication, use `bootstrap`, `jackknife`, or `svy` as a prefix command.

Simulation results suggest that the influence-function 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.

Reporting options

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

`noheader` suppress the output header.

`notable` suppresses the output table containing the estimated coefficients. `table` enforces displaying 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

Main graph options

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

`norefline` suppresses the parity line.

Additional options after reldist pdf

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

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

`noci` omits the confidence spikes of the histogram bars.

`nohistogram` omits the histogram bars.

Additional options after reldist histogram

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

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 zero 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 help [G] ***cline_options***.

Confidence intervals

`level(#)` specifies the confidence level, as a percentage, for confidence intervals.

`citransform` plots transformed confidence intervals depending on the type of the reported statistic (log transform for PDF and histogram density, 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 use `ci(ci_percentile)` to plot percentile confidence intervals. `ci()` and `level()` are not both allowed.

`ciopts(`*options*`)` specifies options to affect the rendition of the confidence intervals. See help [G] ***area_options*** or, after, `reldist histogram` help [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.

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) is to 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 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. 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 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 as follows:

`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 be generated, `prune()` determines the pruning (see above), `at` changes the meaning of the main *numlist*, `format()` specifies the display format for the labels, `tick()` specifies values for which ticks be generated, `line()` specifies values for which added lines be generated, and `y` request outcome labels for the Y axis of the relative CDF (only allowed after `reldist cdf`). The command returns the following macros in `r()`:

`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()`

❑

General graph options

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

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

4.4 Saved 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, chapter8). The data covers 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 contains information on the achieved educational level and there is a variable providing sampling weights.[11]

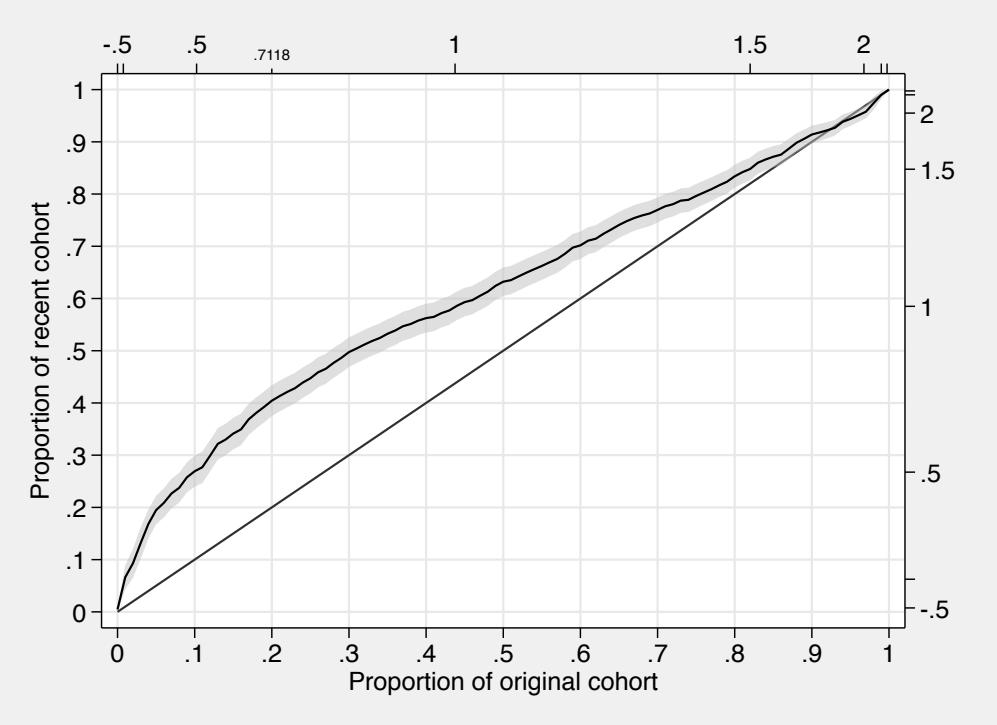
```
. use nls
(NSL data from Handcock and Morris (1999))

. describe

Contains data from nls.dta
  obs:         3,937                          NSL data from Handcock and
                                                Morris (1999)
 vars:             4                          11 Sep 2020 15:46
                                               (_dta has notes)
-------------------------------------------------------------------------------
              storage   display    value
variable 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
     by categories of: 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 $\exp(1.085) - 1) \times 100 = 196$ percent. For the recent cohort the average is only 0.878 (141 percent). We can also see that inequality in wage growth has been more pronounced in the recent cohort than in the original cohort, as the standard deviation of log wage gains is larger. Looking at the median and interquartile range (IQR) instead of the mean and standard deviation leads to qualitatively similar findings.

[11] The data has been obtained from `http://www.stat.ucla.edu/~handcock/RelDist/Data/R/RDBnls.RData`.

The relative CDF The relative CDF of log wage gains between the recent cohort and the original cohort can be obtained as follows:

```
. 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))
```



The horizontal axis of the graph corresponds to cumulative proportions of the original cohort, the vertical axis to cumulative proportions of the recent cohort, both 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 since 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 percent. In the original cohort, 20% experienced a wage growth of at most 104 percent; in the recent cohort, this proportion increased to 40%. That is, relative to the original cohort, wage growth of 104 percent or less is overrepresented by factor 2 in the recent cohort.

❑ Technical note

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 like using a reduced set of evaluation points; option `at(.1(.1).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	Coef.	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.

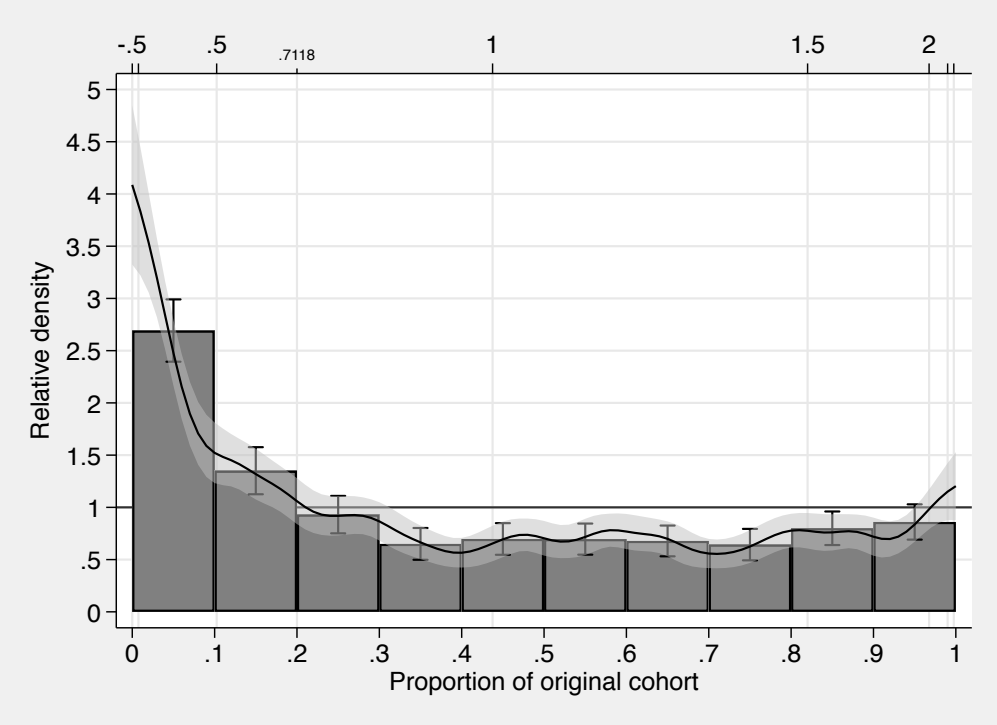
Furthermore, the graph has been produced by first estimating the CDF using `reldist cdf` and then plotting the result using `reldist graph`. We could also have drawn the graph in a single step by including option `graph()` in the call to `reldist cdf` (see examples further down). Options `olabel()` and `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.[12] 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***.

❑

[12]Suboption `add` has been specified in the second `olabel()` option so that the labels from both `olabel()` options are printed, `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 labelling suboptions; see [G] ***axis_label_options***.

The relative PDF Relative over- 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:

```
. 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)
```



A relative density larger than one means that the recent cohort is overrepresented at the corresponding level of wage gains, values lower than one mean that 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).

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 due to a general difference in economic growth between the two areas that affected all population members in a similar way. 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, also the structure of wage gains might have changed, for example due to rising returns on education, leading to more polarization of wage gains in the recent cohort. In this case, also the shape of the two distributions would 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 has been location or shape adjusted.

The following commands produce a graph containing three panels.[13] The first panel shows the overall (unadjusted) relative density (same as above). The second panel shows how the relative density looks like 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 in a way 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.[14]

```
. 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)
```

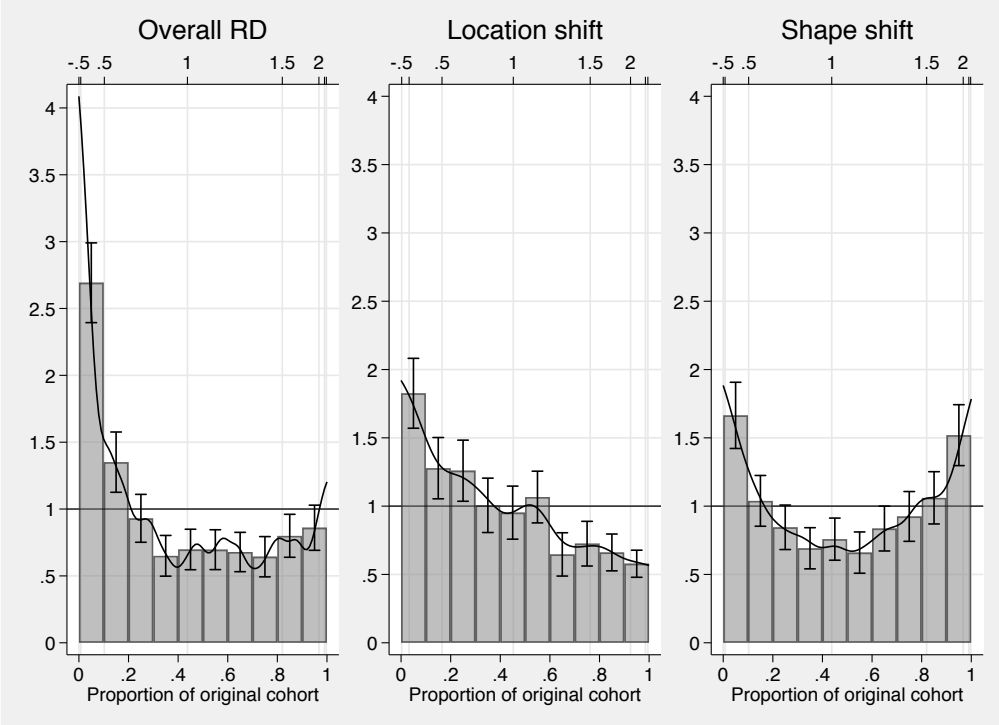
(*output omitted*)

```
. reldist pdf chpermwage [pw = wgt], by(cohort) histogram ///
>     adjust(:shape scale) ///
```

[13] Confidence intervals for the relative density curve have been omitted using graph option `noci`, so that the plots are less busy.

[14] Handcock and Morris (1999) do the decomposition the other way round, equivalent to specifying `adjust(shape scale)` and `adjust(:location)`.

```
>     graph(`gropts´ title("Location shift") name(b, replace) nodraw)
```
(*output omitted*)
```
. 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)
```



The results indicate that the difference between the recent cohort distribution and the original cohort distribution is not only a matter of location, but that 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).

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`:[15]

```
. reldist divergence chpermwage [pw = wgt], by(cohort) ///
```

[15] Alternative measures offered by `reldist divergence` are the Chi-squared divergence and the dissimilarity index (total variation distance).

```
>     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	Coef.	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 (labelled as `main`), the divergence of the relative distribution after location-adjusting the recent cohort (labelled as `alternate`), as well as the difference between these two measures. The first value is the overall divergence, the second value quantifies the divergence due to differences in distributional shape, and the third value quantifies the contribution of the difference in location.[16] We can use [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	Coef.	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. On the one hand, Handcock and Morris (1999) performed a slightly different decomposition (see Footnote 16). 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).

[16] 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).

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	Coef.	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, as 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 zero. Furthermore, the breakup into polarization of the lower half (LRP) and the upper half of the distribution (URP) suggests that the degree of polarization is similar in both tails.

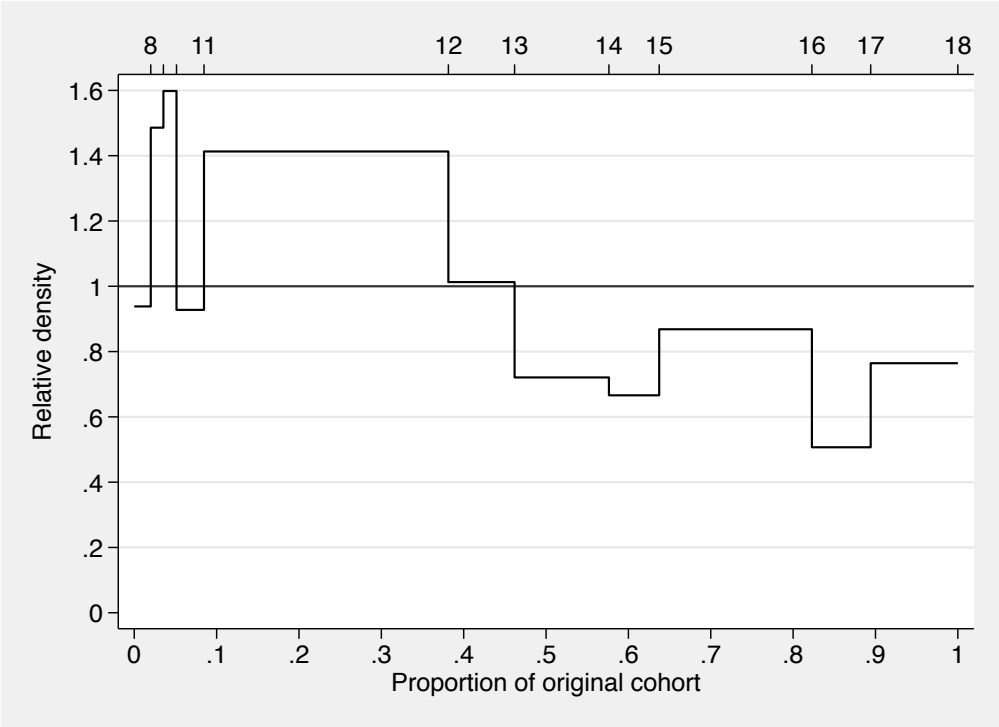
Covariate balancing Education may be on important determinant of the wage distribution as well as the distribution of wage gains over an occupational career. Hence, if the educational distribution changed between the original cohort and the recent cohort, we may be comparing apples with oranges. That 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:[17]

```
. replace endeduc = 8 if endeduc<8
(34 real changes made)

. reldist pdf endeduc [pw = wgt], by(cohort) categorical

Relative density                                Number of obs     =      3,937
  F1: cohort = recent (1979)                    Comparison obs    =      2,103
  F0: cohort = original (1966)                  Reference obs     =      1,834
```

endeduc	Coef.	Std. Err.	[95% Conf.	Interval]
endeduc				
8	.9383436	.220892	.50527	1.371417
9	1.485883	.3551772	.7895346	2.182232

[17] 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`.

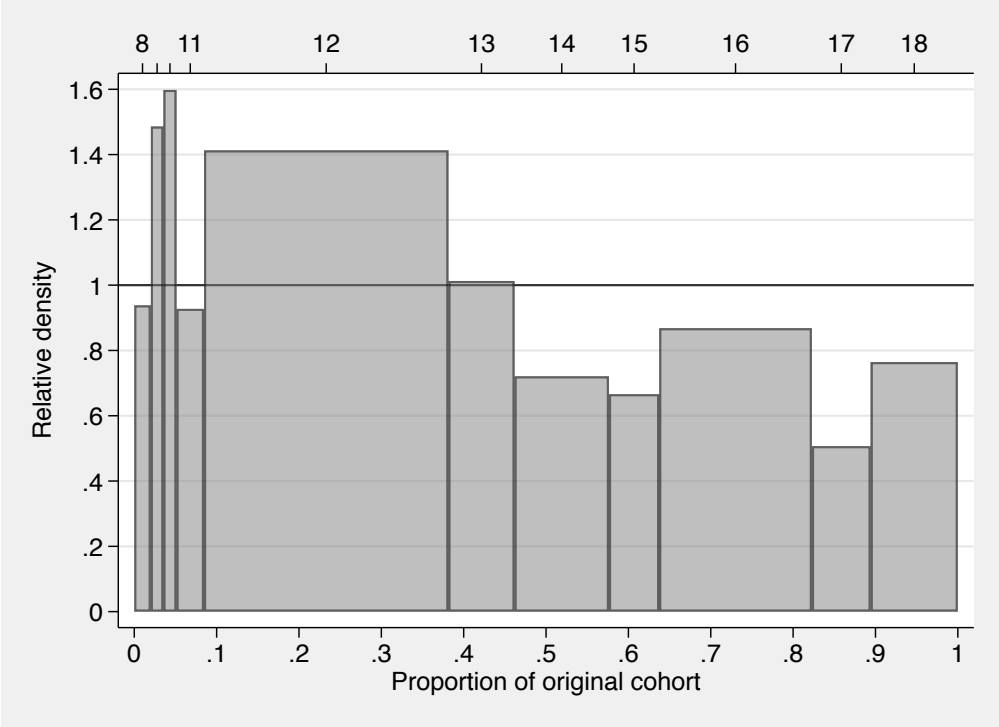
```
        10 |    1.59819   .3734487      .8660189    2.330361
        11 |   .9276922   .1673159      .5996581    1.255726
        12 |    1.41295   .0657943      1.283956    1.541945
        13 |   1.012919   .1136963      .7900094    1.235828
        14 |   .7208455   .0737943      .5761668    .8655241
        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)) ///
>     xlabel(0(.1)1) xtitle(Proportion of original cohort) ///
>     ylabel(0(.2)1.6, angle(0) grid) ytitle(Relative density)
```



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

As suggested by Handcock and Morris (1999), the graph above uses a step function with the steps located at the values of the cumulative distribution in the original cohort. An alternative would be to display the categorical relative density as a histogram in which the width of each bar is proportional to the relative frequency of the corresponding level in the original cohort:

```
. reldist hist endeduc [pw = wgt], by(cohort) categorical ///
>     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))

Relative histogram                              Number of obs     =      3,937
  F1: cohort = recent (1979)                    Comparison obs    =      2,103
  F0: cohort = original (1966)                  Reference obs     =      1,834

(coefficients table suppressed)
```



Numerically, both approaches lead to the exact same results (including standard errors and confidence intervals), but a different style is used for graphical display.

The question now is whether these differences in educational composition affect the relative distribution of wage gains. Similar as above in the context of locaction 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 in a way 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 that displays the overall relative distribution (left panel), the relative distribution after the recent cohort has been reweighted (right panel), as well as 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 original 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*)

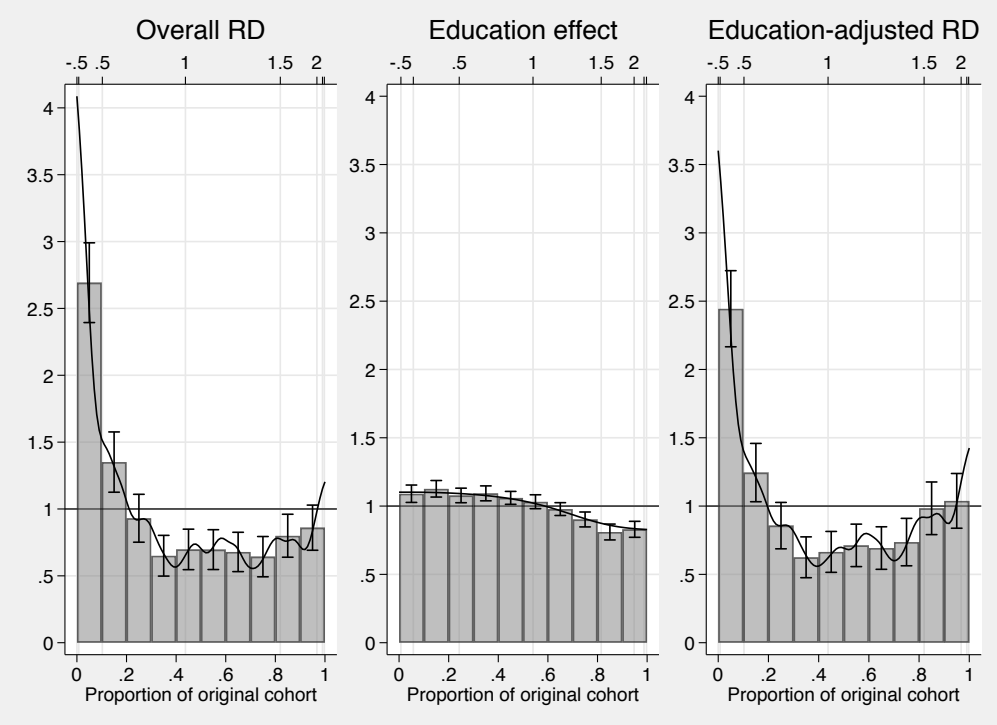
```
. reldist pdf chpermwage [pw = wgt], by(cohort) histogram ///
>     balance(i.endeduc, contrast) ///
>     graph(`gropts´ title("Education effect") name(b, replace) nodraw)
```

(*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)
```



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 upwards, 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).

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 distribution will be exactly the same:

```
. 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*)

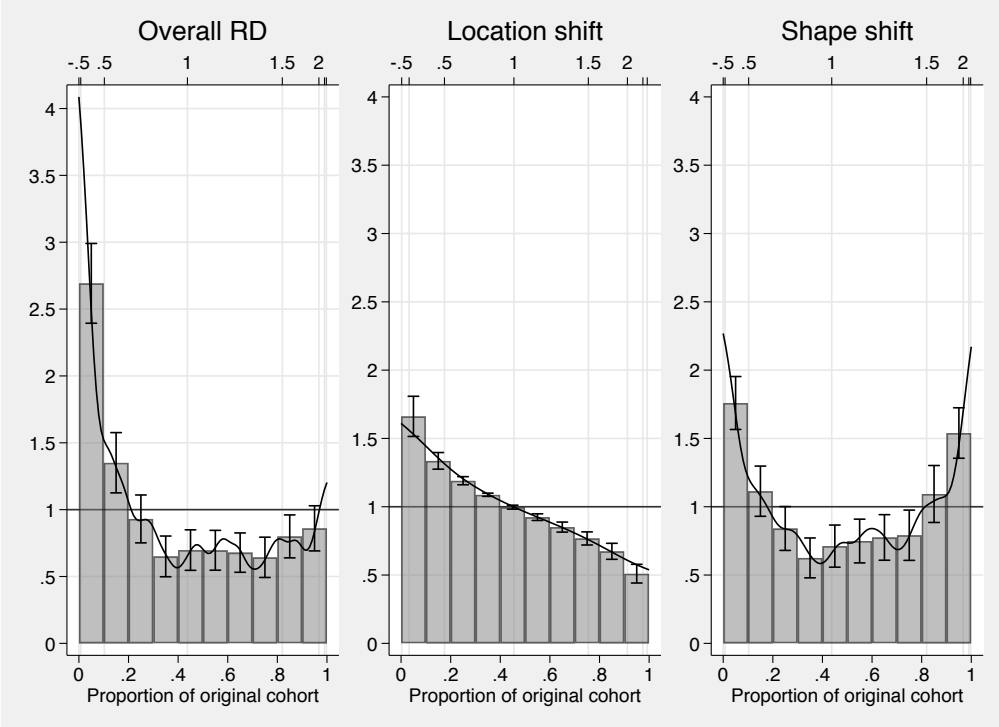
```
. reldist pdf chpermwage [pw = wgt], by(cohort) histogram ///
>     balance(eb: chpermwage, contrast) ///
>     graph(`gropts´ title("Location shift") name(b, replace) nodraw)
```

(*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)
```



The two approaches lead to qualitatively similar results.[18] One advantage of the reweighting approach, however, is that heaping in the data will have less adverse effects on the results.[19]

5.2 Processing results from reldist

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

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

```
. sysuse nlsw88, clear
(NLSW, 1988 extract)
```

[18] 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()`.

[19] 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.

```
. 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	Coef.	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 non-unionized are the comparison group and the unionized are the reference group. 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 non-unionized workers, although the MRP is only marginally significant for respondents without college degree. A follow-up question might thus be whether we can conclude from the data that relative polarization between non-unionized and unionized workers is stronger among college graduates than among workers without 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
```

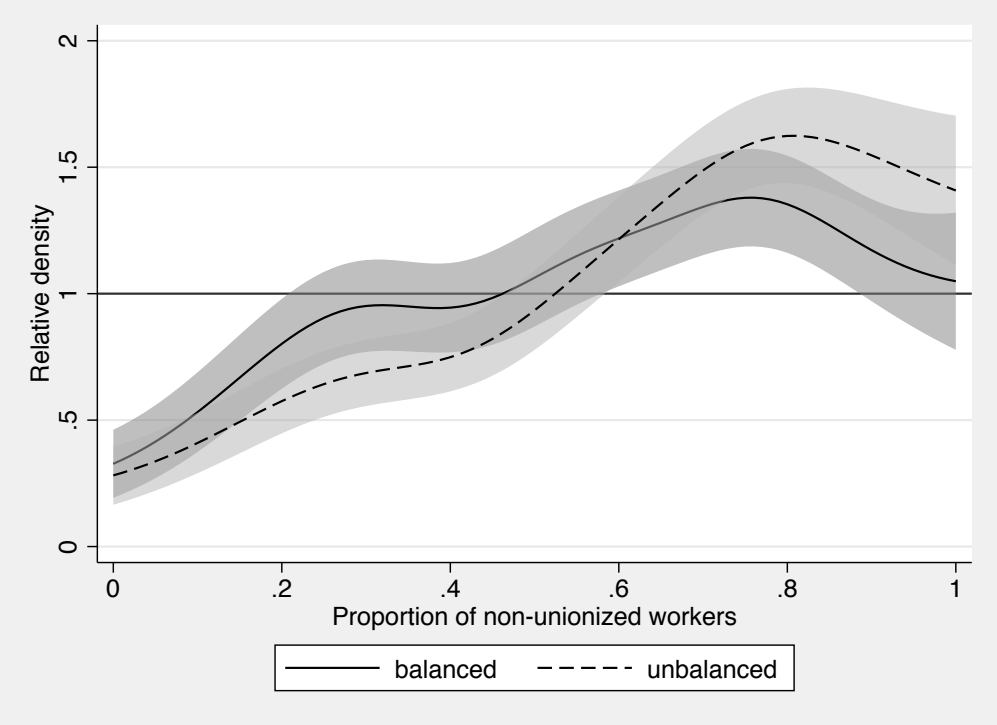
The test is negative; that is, we cannot reject the null hypothesis that the two MRP estimates are the same (p-value of 0.229). The same result can also be obtained using [R] **lincom**:

```
. lincom [1]MRP - [0]MRP
 ( 1)  - [0]MRP + [1]MRP = 0
```

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	.0831615	.0691722	1.20	0.229	-.0525011	.218824

Creating graphs from multiple results When comparing wages between unionized and non-unionized 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 due to differential composition with respect to these characteristics, and not due to 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 using [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)
```



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

Working with influence functions The `predict` command can be used to store the influence functions that `reldist` uses for standard error estimation. For example, we may want to test whether relative polarization between non-unionized 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 influence functions 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	Coef.	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	Coef.	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	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	.0520321	.0380618	1.37	0.172	-.0226159	.1266801

```
. 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 influence functions and then recenterd them by adding the point estimates back in (on the use of recentered influence functions also see Firpo et al., 2009). The influence functions returned by `reldist` are scaled such that [R] **total** can be used for estimation of standard errors (note how [R] **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 influence function by N, add the point estimate as is, and then use [R] **mean** instead of [R] **total**. Furthermore, note that weights are not incorporated into the influence functions. That is, if weights have been applied to `reldist`, the weights will also have to be applied when calling [R] **total** or [R] **mean** (the same is true for clustering).

5.3 Survey estimation

`reldist` fully supports estimation for complex survey data, but the [SVY] **svy** prefix command 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. Example:

```
. webuse nmihs, clear
. svyset [pweight=finwgt], strata(stratan)
     pweight: finwgt
         VCE: linearized
  Single unit: missing
     Strata 1: stratan
         SU 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
```

birthwgt	Linearized Coef.	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 due to a difference in distributional shape in the upper part of the distribution (overall polarization is driven by the URP). Option `vce(svy)` also works with variance estimation methods other than `linearized` (e.g. [SVY] **brr**), although in these cases one could also apply `svy` as a prefix command.[20]

6 Acknowledgements

I thank Eric Melse for valuable comments on earlier versions of the software that helped improving the command. I thank Blaise Melly for a nudge on how to obtain the influence functions for quantiles of relative ranks.

References

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(2): 302–328.

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

Botev, Z. I., J. F. Grotowski, and D. P. Kroese. 2010. Kernel density estimation via diffusion. *Annals of Statistics* 38(5): 2916–2957.

Cox, N. J. 2004. PPPLOT: Stata module for P-P plots. Statistical Software Components S438002, Boston College Department of Economics. Available from `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(8): 3057–3069.

———. 1993. Data-dependent bandwidth choice for a grade density kernel estimate. *Statistics & Probability Letters* 16: 397–405.

[20] 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 prefix command, bandwidth estimation will be repeated in each replication.

DiNardo, J. E., N. Fortin, and T. Lemieux. 1996. Labour Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach. *Econometrica* 64(5): 1001–1046.

Duncan, O. D., and B. Davis. 1953. An Alternative to Ecological Correlation. *American Sociological Review* 18(6): 665–666.

Firpo, S., N. M. Fortin, and T. Lemieux. 2009. Unconditional Quantile Regressions. *Econometrica* 77: 953–973.

Handcock, M. S. 2016. Relative Distribution Methods. Version 1.6-6. Available from `https://cran.r-project.org/web/packages/reldist/index.html`.

Handcock, M. S., and E. M. Aldrich. 2002. Applying Relative Distribution Methods in R. University of Washington Working Paper No. 27. Available from `http://dx.doi.org/10.2139/ssrn.1515775`.

Handcock, M. S., and P. L. Janssen. 2002. Statistical Inference for the Relative Density. *Sociological Methods and Research* 30(3): 394–424.

Handcock, M. S., and M. Morris. 1998. Relative Distribution Methods. *Sociological Methodology* 28: 53–97.

———. 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. Available from `https://ideas.repec.org/c/boc/bocode/s447202.html`.

———. 2005. moremata: Stata module (Mata) to provide various functions. Statistical Software Components S455001. Available from `http://ideas.repec.org/c/boc/bocode/s455001.html`.

———. 2007. Univariate kernel density estimation. Available from `https://doi.org/10.7892/boris.69421`.

———. 2008. The Blinder-Oaxaca decomposition for linear regression models. *The Stata Journal* 8(4): 453–479.

———. 2014. Plotting regression coefficients and other estimates. *The Stata Journal* 14(4): 708–737.

———. 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. Available from `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. University of Bern Social Sciences Working Papers 35. Available from `https://ideas.repec.org/p/bss/wpaper/35.html`.

Morris, M., A. D. Bernhardt, and M. S. Handcock. 1994. Economic Inequality: New Methods for New Trends. *American Sociological Review* 59(2): 205–219.

Parzen, E. 2004. Quantile Probability and Statistical Data Modeling. *Statistical Science* 19(4): 652–662.

Rios-Avila, F. 2020. Recentered influence functions (RIFs) in Stata: RIF regression and RIF decomposition. *The Stata Journal* 20(1): 51–94.