A shorter version of this paper has been published as:


See: http://www.stata-journal.com/article.html?article=st0432

University of Bern Social Sciences Working Paper No. 13

Assessing inequality using percentile shares

Ben Jann

Current version: October 27, 2016
First version: August 12, 2015

http://ideas.repec.org/p/bss/wpaper/13.html
http://econpapers.repec.org/paper/bsswpaper/13.htm
Assessing inequality using percentile shares

Ben Jann
Institute of Sociology
University of Bern

ben.jann@soz.unibe.ch

October 27, 2016

Abstract

At least since Thomas Piketty’s best-selling “Capital in the Twenty-First Century” (2014, Cambridge, MA: The Belknap Press), percentile shares have become a popular approach for analyzing distributional inequalities. In their work on the development of top incomes, Piketty and collaborators typically report top-percentage shares, using varying percentages as thresholds (top 10%, top 1%, top 0.1%, etc.). However, analysis of percentile shares at other positions in the distribution may also be of interest. In this paper I present a new Stata command called \texttt{pshare} that estimates percentile shares from individual-level data and displays the results using histograms or stacked bar charts.

\textit{Keywords:} Stata, \texttt{pshare}, percentile shares, Lorenz curve, concentration curve, inequality, income distribution, wealth distribution, graphics
## Contents

1 Introduction

2 Methods and formulas
   2.1 Lorenz ordinates ........................................ 4
   2.2 Percentiles shares ........................................ 4
   2.3 Point estimation ........................................ 6
   2.4 Variance estimation ...................................... 6
   2.5 Extensions .............................................. 9
      2.5.1 Contrasts ........................................ 9
      2.5.2 Renormalization .................................. 9
      2.5.3 Concentration shares ............................... 10

3 The pshare command
   3.1 Syntax of pshare estimate ................................. 12
   3.2 Syntax of pshare contrast ................................ 15
   3.3 Syntax of pshare stack .................................. 16
   3.4 Syntax of pshare histogram .............................. 17

4 Examples
   4.1 Basic application ........................................ 20
   4.2 Stacked bar charts ...................................... 21
   4.3 Histograms ............................................... 23
   4.4 Contrasts ............................................... 29
      4.4.1 Differences between subpopulations ................ 29
      4.4.2 Differences between outcome variables ............ 32
   4.5 Concentration shares .................................. 33
   4.6 Processing results from pshare ......................... 34

5 Small-sample bias ........................................... 37

6 Discussion .................................................. 39

7 Acknowledgments ............................................. 41
1 Introduction

Empirical inequality literature heavily relies on the Gini coefficient for the analysis of the development of inequality over time or the analysis of differences in inequality between countries. Various distributional changes, however, can give rise to an increase or a decrease of the Gini coefficient and it might be important to obtain more detailed knowledge about these processes. Moreover, even if the Gini coefficient remains stable, significant changes in the shape of a distribution may occur. In addition, the specific values of the Gini coefficient, apart from the minimum and the maximum, are difficult to interpret in an absolute sense.

For these reasons, percentile shares have become increasingly popular for the analysis of distributional inequality. Percentile shares quantify the proportions of total outcome (e.g. of total income) that go to different groups defined in terms of their relative ranks in the distribution. They thus have an intuitive and appealing interpretation and can be used for detailed analysis of distributional changes. The most prominent applications of percentiles shares can be found in the works of Thomas Piketty and collaborators (e.g., Atkinson et al., 2011, Piketty and Saez, 2014, Piketty, 2014) and their “World Top Incomes Database” (http://topincomes.parisschoolofeconomics.eu/). Piketty and collaborators typically study top-income shares, such as the proportion of income that goes to the top 1% or the top 10%, but the income shares of other percentile groups may be interesting too.

In this article I present a new Stata command called \texttt{pshare} that estimates percentile shares of an outcome variable from individual level data. \texttt{pshare} provides standard errors and confidence intervals for the estimated percentile shares and supports estimation from complex samples. Furthermore, \texttt{pshare} provides subcommands for computing differences in percentile shares across variables or subpopulations and for graphing results as stacked bar charts or histograms.\footnote{Some of the functionality of \texttt{pshare} is also covered by the user commands \texttt{sumdist} (Jenkins, 1999) and \texttt{svylorenz} (Jenkins, 2006). However, \texttt{pshare} specifically focusses on percentile shares and provides a more comprehensive implementation. Furthermore, \texttt{sumdist} and \texttt{svylorenz} use somewhat different methods to compute the percentile shares (ties are not broken and flat regions in the distribution function are not interpolated; see below).}
2 Methods and formulas

2.1 Lorenz ordinates

Let $Y$ be the outcome variable of interest (e.g. income). The distribution function of $Y$ is given as $F(y) = \Pr\{Y \leq y\}$ and the quantile function (the inverse of the distribution function) is given as $Q(p) = F^{-1}(p) = \inf\{y | F(y) \geq p\}$ with $p \in [0, 1]$. Based on these definitions the ordinates of the Lorenz curve are given as

$$L(p) = \frac{\int_{-\infty}^{Q_p} y \, dF(y)}{\int_{-\infty}^{\infty} y \, dF(y)}$$

Intuitively, a point on the Lorenz curve quantifies the proportion of total outcome of the poorest $p \cdot 100$ percent of the population. This can easily be seen in the finite population form of $L(p)$, which is given as

$$L(p) = \frac{\sum_{i=1}^{N} Y_i I_{Y_i \leq Q_p}}{\sum_{i=1}^{N} Y_i}$$

with $I_A$ as an indicator function being equal to 1 if $A$ is true and 0 else.

2.2 Percentiles shares

Percentile share $S(p_1, p_2)$, with $p_1 \leq p_2$, is equal to the proportion of total outcome that falls into the quantile interval $(Q_{p_1}, Q_{p_2}]$ or, stated differently, the proportion of total outcome pertaining to the population segment from relative rank $p_1$ to relative rank $p_2$ in the list of ordered outcomes. This is equal to the difference between the Lorenz ordinates for $p_1$ and $p_2$, that is

$$S(p_1, p_2) = L(p_2) - L(p_1)$$

or, in the finite population,

$$S(p_1, p_2) = \frac{\sum_{i=1}^{N} Y_i I_{Y_i \leq Q_{p_2}}}{\sum_{i=1}^{N} Y_i} - \frac{\sum_{i=1}^{N} Y_i I_{Y_i \leq Q_{p_1}}}{\sum_{i=1}^{N} Y_i} = \frac{\sum_{i=1}^{N} Y_i (I_{Y_i \leq Q_{p_2}} - I_{Y_i \leq Q_{p_1}})}{\sum_{i=1}^{N} Y_i}$$

To simplify notation, let $S_\ell = S(p_{\ell-1}, p_\ell)$. Furthermore, let

$$\mathbf{s}(p) = [S_1 \quad S_2 \quad \cdots \quad S_k]$$

be the $1 \times k$ vector of a disjunctive and exhaustive set of percentile shares across the domain of $p$ using cutoffs $p = [p_0 \quad p_1 \quad \cdots \quad p_k]$ with $p_{\ell-1} < p_\ell$ for all $\ell = 0, \ldots, k$ and $p_0 = 0$ and $p_k = 1$. 
Depending on context it may be sensible to normalize percentile shares by the size of the respective population segment (i.e. the proportion of the population covered by the segment, which is equal to \( p_\ell - p_{\ell-1} \)), yielding percentile share density

\[
D_\ell = \frac{S_\ell}{p_\ell - p_{\ell-1}}
\]

\( D_\ell \) is a density in the sense that \( \mathbf{d}(p) \) — a disjunctive and exhaustive set of percentile share densities across the domain of \( p \) — integrates to 1. Note, however, that \( D_\ell \) may be negative if the outcome variable can take on negative values (e.g. debt). A value of \( D_\ell = 1 \) means that each member in the respective population segment has (on average) an outcome value equal to the average outcome in the population. A value of \( D_\ell = 2 \) means that each member in the segment has (on average) twice the population average; a value of \( D_\ell = -0.5 \) means that each member in the segment has (on average) minus half the population average.

Furthermore, percentile shares can be expressed as totals or averages in absolute terms. The finite population form of percentile share totals and averages are given as

\[
T_\ell = \sum_{i=1}^{N} Y_i I_{Y_i \leq Q_{p_\ell}} - \sum_{i=1}^{N} Y_i I_{Y_i \leq Q_{p_{\ell-1}}} = \sum_{i=1}^{N} Y_i \left( I_{Y_i \leq Q_{p_\ell}} - I_{Y_i \leq Q_{p_{\ell-1}}} \right) = S_\ell \sum_{i=1}^{N} Y_i
\]

and

\[
A_\ell = \frac{T_\ell}{(p_\ell - p_{\ell-1}) \cdot N}
\]

respectively. \( T_\ell \) is simply the sum of all outcomes in the respective population segment; \( A_\ell \) is the average outcome among the members of the segment.

Finally, with reference to the generalized Lorenz curve, generalized percentile shares can be defined as

\[
G_\ell = GL(p_\ell) - GL(p_{\ell-1})
\]

where the finite-population form of the generalized Lorenz ordinate \( GL(p) \) is

\[
GL(p) = \frac{1}{N} \sum_{i=1}^{N} Y_i I_{Y_i \leq Q_p}
\]

so that

\[
G_\ell = \frac{1}{N} \sum_{i=1}^{N} Y_i I_{Y_i \leq Q_{p_\ell}} - \frac{1}{N} \sum_{i=1}^{N} Y_i I_{Y_i \leq Q_{p_{\ell-1}}} = \frac{1}{N} \sum_{i=1}^{N} Y_i \left( I_{Y_i \leq Q_{p_\ell}} - I_{Y_i \leq Q_{p_{\ell-1}}} \right)
\]

Note that there is an interesting relation between percentile share averages and generalized percentile shares: percentile share average \( A_\ell \) is equal to \( G_\ell / (p_\ell - p_{\ell-1}) \), that is, \( A_\ell \) is equal to the difference in the generalized Lorenz ordinates for \( p_\ell \) and \( p_{\ell-1} \) divided by the population share \( p_\ell - p_{\ell-1} \).
2.3 Point estimation

The above exposition assumes $Y$ to be continuous. Since empirical data is always discrete, the empirical distribution function is non-smooth and there may be ties or empty sets at the quantiles of interest. For estimation of percentile shares using empirical data it makes sense to break ties proportionally and apply linear interpolation in regions where the empirical distribution function is flat.

Let $w_i$ be sampling weights (equal to 1 in unweighted data) and let subscripts in parentheses refer to sorted observations in ascending order of $Y$. $S_\ell$ can then be estimated from a sample of size $n$ as

$$\hat{S}_\ell = \hat{L}(p_\ell) - \hat{L}(p_{\ell-1})$$

with

$$\hat{L}(p) = (1 - \gamma)\tilde{Y}_{jp-1} + \gamma\tilde{Y}_{jp},$$

where

$$\gamma = \frac{p - \hat{p}_{jp-1}}{\hat{p}_{jp} - \hat{p}_{jp-1}}, \quad \tilde{Y}_{jp} = \frac{\sum_{i=1}^{j_p} w_i Y_i}{\sum_{i=1}^{n} w_i Y_i}, \quad \text{and} \quad \hat{p}_{jp} = \frac{\sum_{i=1}^{j_p} w_i}{\sum_{i=1}^{n} w_i}$$

and where $j_p$ is set such that $\hat{p}_{jp-1} < p \leq \hat{p}_{jp}$. This corresponds to estimating Lorenz ordinates by taking quantiles from the running sum of the ordered $Y$ values (divided by the total of $Y$) according to quantile definition 4 in Hyndman and Fan (1996).

Alternatively, ignoring linear interpolation in flat regions, $L(p)$ can be estimated as

$$\hat{L}(p) = \tilde{Y}_{jp} = \frac{\sum_{i=1}^{j_p} w_i Y_i}{\sum_{i=1}^{n} w_i Y_i}$$

corresponding to quantile definition 1 in Hyndman and Fan (1996).\footnote{The first approach is the default method in the \texttt{pshare} command presented below. The second approach ignoring linear interpolation can be requested by specifying the \texttt{step} option. Note that results from the second approach depend on the sort order within ties of $Y$ if there are sampling weights. To enforce stable results in this case, the \texttt{pshare} command sorts observations in ascending order of the sampling weights among ties of $Y$, but this is an arbitrary decision.}

An estimate for $D_\ell$ is given as $\hat{S}_\ell/(p_\ell - p_{\ell-1})$. For an estimate of $T_\ell$ omit the denominator, $\sum_{i=1}^{n} w_i Y_i$, in the formula for $\tilde{Y}_j$. An estimate for $A_\ell$ can be obtained as $\hat{T}_\ell/((p_\ell - p_{\ell-1})\sum_{i=1}^{n} w_i)$. For an estimate of $G_\ell$ replace the denominator in the formula for $\tilde{Y}_j$ by $\sum_{i=1}^{n} w_i$.

2.4 Variance estimation

An approximate variance matrix for $\mathbf{s}(p)$ can be obtained by employing an estimating equations approach as outlined by Binder and Kovacevic (1995; also see Kovačević and Binder, 1997). Let $\theta$ be the parameter of interest (a percentile share in our case) and let $\mathbf{X}$ be a
vector of nuisance parameters on which $\theta$ depends (the two quantiles determining the Lorenz ordinates in our case). According to Kovačević and Binder (1997), the sampling variance of $\hat{\theta}$ can be approximated by the sampling variance of the total estimator

$$\sum_{i=1}^{n} w_i u_i^*$$

where $w_i$ are sampling weights and $u_i^*$ is the solution of

$$\left(-u_i^\theta + \frac{\partial U^\theta}{\partial \lambda} \left[ \frac{\partial U^\lambda}{\partial \lambda} \right]^{-1} u_i^\lambda \right) \left[ \frac{\partial U^\theta}{\partial \theta} \right]^{-1}$$

with all unknowns in the final solution replaced by their sample counterparts. $u_i^\theta$ and $u_i^\lambda$ are estimating functions such that, in the (finite) population, $\theta$ and $\lambda$ are the solutions to

$$U^\theta = \sum_{i=1}^{N} u_i^\theta = 0 \quad \text{and} \quad U^\lambda = \sum_{i=1}^{N} u_i^\lambda = 0$$

In our case, $\theta = S_j^\ell$ and $\lambda = [Q_{j p_{\ell}}^i, Q_{j p_{\ell-1}}^i]$, where $j$ refers to the analyzed subpopulation. Let $J_i = 1$ if observation $i$ belongs to subpopulation $j$ and $J_i = 0$ else (with $J_i = 1$ for all observations if the entire sample is analyzed). Since

$$S_j^\ell = \frac{\sum_{i=1}^{N} Y_i \left( I_{Y_i \leq Q_{j p_{\ell}}^i} - I_{Y_i \leq Q_{j p_{\ell-1}}^i} \right) J_i}{\sum_{i=1}^{N} Y_i J_i}$$

and

$$\sum_{i=1}^{N} \left( I_{Y_i \leq Q_{p_{\ell}}^j} - p_{\ell} \right) J_i = 0$$

the estimating functions are

$$u_i^\theta = Y_i \left( I_{Y_i \leq Q_{j p_{\ell}}^i} - I_{Y_i \leq Q_{j p_{\ell-1}}^i} \right) J_i - Y_i J_i S_j^\ell \quad \text{and} \quad u_i^\lambda = \left[ \left( I_{Y_i \leq Q_{j p_{\ell}}^i} - p_{\ell} \right) J_i \bigg/ \left( I_{Y_i \leq Q_{j p_{\ell-1}}^i} - p_{\ell-1} \right) J_i \right]$$

Furthermore, given these definitions,

$$\frac{\partial U^\theta}{\partial \theta} = -\sum_{i=1}^{N} Y_i J_i \quad \text{and} \quad \frac{\partial U^\theta}{\partial \lambda} \left[ \frac{\partial U^\lambda}{\partial \lambda} \right]^{-1} = \left[ \frac{E(Y|Y = Q_{j p_{\ell}}^i)}{-E(Y|Y = Q_{j p_{\ell-1}}^i)} \right]'$$

Finally, since $E(Y|Y = Q_{p_{\ell}}) = Q_{p_{\ell}}$, we get

$$u_i^* = \left( Y_i - \hat{Q}_{j p_{\ell}}^i \right) I_{Y_i \leq Q_{j p_{\ell}}^i} - \left( Y_i - \hat{Q}_{j p_{\ell-1}}^i \right) I_{Y_i \leq Q_{j p_{\ell-1}}^i} + p_{\ell} \hat{Q}_{j p_{\ell}}^i - p_{\ell-1} \hat{Q}_{j p_{\ell-1}}^i \right) J_i - Y_i J_i S_j^\ell$$

$$= \frac{\left( Y_i - \hat{Q}_{j p_{\ell}}^i \right) I_{Y_i \leq Q_{j p_{\ell}}^i} - \left( Y_i - \hat{Q}_{j p_{\ell-1}}^i \right) I_{Y_i \leq Q_{j p_{\ell-1}}^i} + p_{\ell} \hat{Q}_{j p_{\ell}}^i - p_{\ell-1} \hat{Q}_{j p_{\ell-1}}^i \right) J_i - Y_i J_i S_j^\ell}{\sum_{k=1}^{n} w_k Y_k J_k}$$
The sampling variance of the total of \( u_i^* \), which serves as an approximation of the sampling variance of \( \hat{S}_i^j \), can then be estimated using standard techniques as implemented in \texttt{total} (see \cite{r} \texttt{total}), possibly accounting for complex survey design. The joint variance matrix for all elements of \( \hat{S}(p) \) can be obtained by applying \texttt{total} to a series of appropriate \( u^* \) variables. Likewise, for joint estimation across several outcome variables or multiple subpopulations, include multiple series of \( u^* \) variables, one series for each outcome variable or subpopulation.\(^3\)

Variance estimators for percentile densities, totals, averages, or generalized shares can be derived analogously. The appropriate \( u^* \) variables are obtained by replacing \( a_i \) and \( b \) in

\[
u_i^* = \left( \frac{(Y_i - \hat{Q}_p^{ij}) I_{Y_i \leq \hat{Q}_p^{ij}} - (Y_i - \hat{Q}_p^{ij+1}) I_{Y_i \leq \hat{Q}_p^{ij+1}} + p_i \hat{Q}_p^{ij} - p_{\ell-1} \hat{Q}_p^{ij}}{b} \right) J_i - a_i
\]

according to the overview in table 1, where \( n_c \) is the number of clusters and \( \omega_i \) is the sum of weights in the cluster to which observation \( i \) belongs.\(^4\)

### Table 1: Definitions of \( a_i \) and \( b \) for different types of percentile shares

<table>
<thead>
<tr>
<th>( a_i )</th>
<th>( b )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S )</td>
<td>( Y_i J_i \hat{\Sigma}_\ell^j )</td>
</tr>
<tr>
<td>( D )</td>
<td>( Y_i J_i (p_\ell - p_{\ell-1}) \hat{\Gamma}_\ell^j )</td>
</tr>
<tr>
<td>( T )</td>
<td>( \frac{1}{n_c \omega_i} \hat{T}_\ell^j )</td>
</tr>
<tr>
<td>( A )</td>
<td>( J_i (p_\ell - p_{\ell-1}) \hat{\Lambda}_\ell^j )</td>
</tr>
<tr>
<td>( G )</td>
<td>( J_i \hat{\Gamma}_\ell^j )</td>
</tr>
</tbody>
</table>

An alternative to the approach outlined above is to estimate the variances using the bootstrap or jackknife method (see \cite{r} \texttt{bootstrap} and \cite{r} \texttt{jackknife}).

\(^3\)When computing the \( u^* \) variables, the \texttt{pshare} command presented below uses definition 4 in Hyndman and Fan (1996) to determine \( \hat{Q}_p^i \) (or definition 1, depending on the method used for estimating the Lorenz ordinates). Furthermore, in analogy to the approach employed for point estimation, ties in \( Y \) are broken when determining \( I\{Y_i \leq \hat{Q}_p^i\} \) (based on observations sorted by \( w_i \) within ties, which is an arbitrary decision to enforce stable results).

\(^4\)Depending on sample design, the denominator in \( a_i \) for \( T \) may require modification, for example, to take account of stratification. A workaround, followed by the \texttt{pshare} command presented below, is to simply set \( a_i \) to zero for \( T \). This is a slight deviation from the approach outlined above (as \( u^* \) will sum to \( \hat{T} \) instead of zero), but the resulting variance estimates are the same in this case. On a related matter, note that \texttt{total} with clusters or weights yields different results than \texttt{svy:total} because the former assumes the number of observations or the sum of weights (and not the number of clusters) to be fixed. Likewise, \texttt{total} with the \texttt{over()} option produces different results than \texttt{svy:total}, even in the absence of clusters or weights, because the subgroup sizes are assumed fixed. Despite this disagreement, the \texttt{pshare} command presented below, which relies on the \texttt{total} command for purpose of variance estimation, always yields results that are consistent with \texttt{svy:total}, irrespective of whether weights and clusters are specified directly or via the \texttt{svy} option.
2.5 Extensions

2.5.1 Contrasts

To analyze distributional differences among subpopulations or across time, it is helpful to compute contrasts between percentile shares. The most intuitive approach is to compute contrasts as arithmetic differences. For example, given percentile share estimates from two subpopulations (or two variables), \( A \) and \( B \), the vector of arithmetic contrasts is

\[
\tilde{s}^A(p) - \tilde{s}^B(p)
\]

with variance matrix

\[
\begin{bmatrix}
I_k & -I_k
\end{bmatrix}
\hat{V}\left\{\begin{bmatrix}
\tilde{s}^A(p) \\
\tilde{s}^B(p)
\end{bmatrix}\right\}
\begin{bmatrix}
I_k & -I_k
\end{bmatrix}^T
\]

where \( I_k \) is the identity matrix of size \( k \) and \( \hat{V}\{\ldots\} \) is the joint variance matrix of the percentile shares across both subpopulations (or variables).

Alternatively, contrasts could be computed as ratios or logarithms of ratios. Generally, let

\[
\begin{bmatrix}
c\left(\hat{S}^A_1, \hat{S}^B_1\right) \\
c\left(\hat{S}^A_2, \hat{S}^B_2\right) \\
\vdots \\
c\left(\hat{S}^A_k, \hat{S}^B_k\right)
\end{bmatrix}
\]

be the vector of percentile share contrasts between subpopulations (or variables) \( A \) and \( B \), with \( c(a, b) \) as a function of \( a \) and \( b \), such as \( c(a, b) = a/b \) (ratio) or \( c(a, b) = \ln(a/b) \) (logarithm of ratio). The variance matrix of the vector can then be approximated by the delta method as

\[
\Delta \hat{V}\left\{\begin{bmatrix}
\tilde{s}^A(p) \\
\tilde{s}^B(p)
\end{bmatrix}\right\} \Delta^T
\]

where \( \Delta \) is a \( k \times 2k \) matrix

\[
\begin{bmatrix}
\frac{\partial c(\hat{S}^A_1, \hat{S}^B_1)}{\partial \hat{S}^A_1} & 0 & \cdots & 0 & \frac{\partial c(\hat{S}^A_1, \hat{S}^B_1)}{\partial \hat{S}^B_1} & 0 & \cdots & 0 \\
0 & \frac{\partial c(\hat{S}^A_2, \hat{S}^B_2)}{\partial \hat{S}^A_2} & \cdots & 0 & 0 & \frac{\partial c(\hat{S}^A_2, \hat{S}^B_2)}{\partial \hat{S}^B_2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \frac{\partial c(\hat{S}^A_k, \hat{S}^B_k)}{\partial \hat{S}^A_k} & 0 & 0 & \cdots & \frac{\partial c(\hat{S}^A_k, \hat{S}^B_k)}{\partial \hat{S}^B_k}
\end{bmatrix}
\]

In Stata, the \texttt{nlcom} command can be used to perform the necessary computations. The derivatives in \( \Delta \) are determined numerically by \texttt{nlcom} (see \texttt{[R] nlcom}).

2.5.2 Renormalization

Percentile shares expressed as proportions or densities are normalized with respect to the total of the analyzed outcome variable in the given (sub-)population. Depending on context, it may be sensible to use a different total for normalization. For example, when analyzing different subpopulations we may want to express results in terms of proportions of the grand
To normalize results to a different total, simply replace denominator $\sum_{i=1}^{n} w_i Y_i$ in the above percentile share estimators by the appropriate total. For example, to normalize to the total of variable $Z$ instead of the total of variable $Y$ (where $Z$ may be the sum of several variables, possibly including $Y$), use $\sum_{i=1}^{n} w_i Z_i$ as denominator. Similarly, if normalizing percentile shares to the total of a reference (sub-)population $r$ instead of subpopulation $j$, replace the standard denominator $\sum_{i=1}^{n} w_i Y_i J_i$ by $\sum_{i=1}^{n} w_i Y_i R_i$, where $J_i$ and $R_i$ are indicators for whether observation $i$ belongs to subpopulation $j$ or $r$, respectively. When normalizing percentile densities to the total of a reference (sub-)population, the relative group sizes have to be taken into account so that the densities reflect multiples of the average outcome in the reference (sub-)population. That is, use

$$\hat{D}^{jr}_\ell = \frac{\hat{S}^{jr}_\ell}{(p_\ell - p_{\ell-1}) P^{jr}_\ell}$$

with

$$\hat{P}^{jr}_\ell = \frac{\sum_{i=1}^{n} w_i J_i}{\sum_{i=1}^{n} w_i R_i}$$

to compute the percentile density in subpopulation $j$ with respect to the total of subpopulation $r$.

For variance estimation several cases have to be distinguished: (1) normalizing to the total of $Z$, (2) normalizing to a fixed total $\tau$, (3) normalizing to the total of $Y$ in reference population $r$, (4) normalizing to the total of $Z$ in reference population $r$, (5) normalizing to a fixed total $\tau$ in reference population $r$. In general, when normalizing densities with respect to a reference population (cases 3 to 5), the relative group size is a further nuisance parameter that has to be taken into account. Solving the equations for the different cases leads to the expressions for $a_i$ and $b$ as shown in table 2 (see the section on variance estimation above for background).

2.5.3 Concentration shares

A further interesting possibility is to determine the relative ranks of the population members using an alternative outcome variable. By default, observations will be ordered by their $Y$ values. We may, however, also order observations by some alternative variable $Z$. The (finite-population) Lorenz ordinates, are then defined as

$$L^Z(p) = \frac{\sum_{i=1}^{N} Y_i I_{Z_i \leq Q^Z_p}}{\sum_{i=1}^{N} Y_i}$$

and the percentile shares reflect the proportion of total $Y$ that is received by different percentile groups of $Z$ (the Lorenz curve is called a concentration curve in this case; see Kakwani, 1977, Lambert, 2001). For example, this could be used to analyze how taxes ($Y$) are distributed across income groups ($Z$).

---

5Depending on sample design, expression $\tau/(n_c \omega_i)$ in $a_i$ for cases (2) and (5) may require modification. An alternative, however, is to simply set $\tau/(n_c \omega_i)$ to zero. See footnote 4 above.
Table 2: Definitions of $a_i$ and $b$ for renormalized percentile shares

<table>
<thead>
<tr>
<th></th>
<th>$a_i$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$S$ [ Z_i J_i \hat{s}_{ij} ]</td>
<td>$\sum_i w_i Z_i J_i$</td>
</tr>
<tr>
<td></td>
<td>$D$ [ Z_i J_i ( p_\ell - p_{\ell-1} ) \hat{D}_{ij} ]</td>
<td>$\sum_i w_i Z_i J_i ( p_\ell - p_{\ell-1} )$</td>
</tr>
<tr>
<td>2</td>
<td>$S$ [ \frac{\tau}{n_{i-i}} \hat{s}_{ij} ]</td>
<td>$\tau$</td>
</tr>
<tr>
<td></td>
<td>$D$ [ \frac{\tau}{n_{i-i}} ( p_\ell - p_{\ell-1} ) \hat{D}_{ij} ]</td>
<td>$\tau ( p_\ell - p_{\ell-1} )$</td>
</tr>
<tr>
<td>3</td>
<td>$S$ [ Y_i R_i \hat{s}<em>{ij}^{jr} ] [ Y_i R_i \hat{s}</em>{ij}^{jr} ]</td>
<td>$\sum_i w_i Y_i R_i$</td>
</tr>
<tr>
<td></td>
<td>$D$ [ ( Y_i R_i - \sum_k w_k Y_k R_k ) \frac{\tau}{\sum_k w_k R_k} R_i + \sum_k w_k Y_k R_k J_i ]</td>
<td>[ \sum_i w_i Y_i R_i ( p_\ell - p_{\ell-1} ) \hat{P}<em>{ijr} \hat{D}</em>{ijr} ]</td>
</tr>
<tr>
<td></td>
<td>[ \times ( p_\ell - p_{\ell-1} ) \hat{P}<em>{ijr} \hat{D}</em>{ijr} ]</td>
<td>[ \times ( p_\ell - p_{\ell-1} ) \hat{P}<em>{ijr} \hat{D}</em>{ijr} ]</td>
</tr>
<tr>
<td>4</td>
<td>$S$ [ \text{Like (3), but with all instances of } Y \text{ replaced by } Z. ]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$D$ [                                                   ]</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>$S$ [ \frac{\tau}{n_{i-i}} \hat{s}_{ij}^{jr} ] [ ( Y_i R_i - \sum_k w_k Y_k R_k ) \frac{\tau}{\sum_k w_k R_k} R_i + \sum_k w_k Y_k R_k J_i ]</td>
<td>$\tau ( p_\ell - p_{\ell-1} ) \hat{P}<em>{ijr} \hat{D}</em>{ijr} ]</td>
</tr>
<tr>
<td></td>
<td>$D$ [ \times ( p_\ell - p_{\ell-1} ) \hat{P}<em>{ijr} \hat{D}</em>{ijr} ]</td>
<td>[ \times ( p_\ell - p_{\ell-1} ) \hat{P}<em>{ijr} \hat{D}</em>{ijr} ]</td>
</tr>
</tbody>
</table>

(all sums are across the entire sample)

For purpose of estimation it appears sensible to average $Y$ within ties of $Z$ when computing the concentration curve ordinates, so that results are independent of the sort order of the observations. Furthermore, for variance estimation, we need to replace $\hat{Q}_p$ in the formulas for the $u^*$ variables by $\hat{E}(Y|Z = Q_p^Z)$, the expected value of $Y$ at the $p$-quantile of $Z$.\(^6\)

---

\(^6\)In the **pshare** command presented below, $E(Y|Z = Q_p^Z)$ is estimated by local linear regression using the Epanechnikov kernel and the default rule-of-thumb bandwidth as described in [R] **lpoly**.
3 The pshare command

Four subcommands are provided. `pshare estimate` computes the percentile shares and their variance matrix; `pshare contrast` computes differences in percentile shares between outcome variables or subpopulations based on the results by `pshare estimate`; a stacked bar chart of the results from `pshare estimate` is drawn by `pshare stack`; and a histogram of the results from `pshare estimate` or `pshare contrast` is drawn by `pshare histogram`.

To install `pshare`, type

```
    . ssc install pshare
```

3.1 Syntax of pshare estimate

The syntax of `pshare estimate` is

```
pshare [estimate] varlist [if] [in] [weight] [, options]
```

where `pweight`, `iweight`, and `fweight` are allowed; see `[U] 11.1.6 weight`. For each specified variable, percentile shares (quintile shares by default) are tabulated along with their standard errors and confidence intervals.\(^7\) Only one variable is allowed in `varlist`, if the `over()` option is specified (see below). `pshare` assumes subcommand `estimate` as the default; typing the word “estimate” is only required in case of a name conflict between the first element of `varlist` and the other subcommands of `pshare` (see below). Options are as follows.

Main

- `proportion`, `percent`, `density`, `sum`, `average`, or `generalized` to determine whether percentile shares are expressed as proportions, percentages, densities, totals, averages, or generalized shares. `proportion` is the default.
- `normalize(spec)` to normalize results with respect to the specified total (not allowed in combination with `sum`, `average`, or `generalized`). `spec` is

  \[\overline{}\text{[/}[\text{over:}][\text{total}]\]

  where `over` may be

  - `.` the subpopulation at hand (the default)
  - `#` the subpopulation identified by value `#`
  - `#` the `#`th subpopulation
  - `total` the total across all subpopulations

---

\(^7\)Variance estimation is not supported for `iweights` and `fweights`. To compute standard errors and confidence intervals in case of `fweights`, apply `pshare` to the expanded data (see `[R] expand`).
and `total` may be

- the total of the variable at hand (the default)
- `*` the total of the sum across all analyzed outcome variables
- `varlist` the total of the sum across the variables in `varlist`
- `#` a total equal to `#`

`total` specifies the variable(s) from which the total is to be computed, or sets the total to a fixed value. If multiple variables are specified, the total across all specified variables is used (`varlist` may contain external variables that are not among the list of analyzed outcome variables). `over` selects the reference population from which the total is to be computed; `over` is only allowed if the `over()` option has been specified (see below). Subpopulation sizes (sum of weights) are taken into account for the computation of densities (option `density`) if `over` is provided, so that the densities reflect multiples of the average outcome in the reference population.

`gini` to report the Gini coefficient(s) of the distribution(s) (a.k.a. concentration indices in case of `pvar()`; see below).\(^8\)

**Percentiles**

`nquantiles(#{})` to specify the number of (equally sized) percentile groups to be used or `percentiles(numlist)` to specify a list of percentile cutoffs. The default is `nquantiles(5)`, which corresponds to `percentiles(20 40 60 80)` or, using shorthand as described in [u] 11.1.8 `numlist`, `percentiles(20(20)80)`.

`pvar(pvar)` to construct percentile groups based on variable `pvar` instead of the outcome variable (i.e. to compute percentile shares based on the concentration curve of the outcome variable with respect to `pvar`).

`step` to determine the Lorenz ordinates from the step function of cumulative outcomes. The default is to employ linear interpolation in regions where the step function is flat.

**Over**

`over(varname)` to repeat results for each subpopulation defined by the values of `varname`. `total` is only allowed if `-over()` is specified.

---

\(^8\)Following Lerman and Yitzhaki (1989), the concentration index of \(Y\) with respect to \(Z\) is computed as:

\[
C = 2 \sum_{i=1}^{n} \tilde{w}_i (Y_i - \bar{Y}) (F_i - F) / \bar{Y},
\]

where \(\tilde{w}_i = w_i / \sum_{i=1}^{n} w_i\) are normalized weights, \(\bar{Y} = \sum_{i=1}^{n} \tilde{w}_i Y_i\) is the mean of \(Y\), \(F = \sum_{i=1}^{n} \tilde{w}_i F_i\) is the mean of \(F\), and \(F_i = \sum_{j=1}^{n} \tilde{w}_j I_{Z_j \leq Z_i} - \sum_{j=1}^{n} \tilde{w}_j I_{Z_j = Z_i} / 2\) is the mid-interval relative rank of \(Z_i\) in the empirical distribution of \(Z\). For the Gini coefficient of \(Y\), set \(Z = Y\).
Contrast/Graph

\texttt{contrast[\{spec\}]} to compute differences in percentile shares between outcome variables or between subpopulations. \texttt{spec} is

\begin{verbatim}
[base][, \texttt{ratio lnratio}]
\end{verbatim}

where \texttt{base} is the name of the outcome variable or the value of the subpopulation to be used as base for the contrasts. If \texttt{base} is omitted, adjacent contrasts across outcome variables or subpopulations are computed (or contrasts with respect to the total if total results across subpopulations have been requested).

Use suboption \texttt{ratio} to compute contrasts as ratios or suboption \texttt{lnratio} to compute contrasts as logarithms of ratios. The default is to compute contrasts as differences.

\texttt{stack[\{options\}]} to draw a stacked bar chart of the results. \texttt{options} are as described for \texttt{pshare stack} below.

\texttt{histogram[\{options\}]} to draw a histogram of the results. \texttt{options} are as described for \texttt{pshare histogram} below.

SE/SVY

\texttt{vce(vcetype)} to determine how standard errors and confidence intervals are computed where \texttt{vcetype} may be:

\begin{verbatim}
  analytic
  cluster clustvar
  bootstrap [, \texttt{bootstrap_options}]
  jackknife [, \texttt{jackknife_options}]
\end{verbatim}

\texttt{analytic} is the default. See \texttt{[r] bootstrap} and \texttt{[r] jackknife} for \texttt{bootstrap_options} and \texttt{jackknife_options}, respectively.

\texttt{svy[\{subpop\}]} for taking the survey design as set by \texttt{svyset} into account; see \texttt{[svy] svyset}. Specify \texttt{subpop} to restrict survey estimation to a subpopulation, where \texttt{subpop} is

\begin{verbatim}
[\texttt{varname}][\texttt{if}]
\end{verbatim}

The subpopulation is defined by observations for which \texttt{varname} $\neq 0$ and for which the \texttt{if} condition is met. See \texttt{[svy] subpopulation estimation} for more information on subpopulation estimation.

The \texttt{svy} option is only allowed if the the variance estimation method set by \texttt{svyset} is Taylor linearization (the default). For other variance estimation methods the usual \texttt{svy} prefix command may be used; see \texttt{[svy] svy}. For example, type type “\texttt{svy brr: pshare} . . .” to use BRR variance estimation. \texttt{pshare} does not allow the \texttt{svy} prefix for Taylor linearization due to technical reasons. This is why the \texttt{svy} option is provided.
nose to suppress the computation of standard errors and confidence intervals. Use the nose option to speed-up computations when analyzing census data. The nose option may also be useful to speed-up computations when using a prefix command that employs replication techniques for variance estimation, such as, e.g., [svy] svy jackknife. Options vce(bootstrap) and vce(jackknife) imply nose.

Reporting

level(#) to set the level of confidence intervals; see [R] level.
noheader to suppress the output header, notable to suppress the coefficient table, and nogtable to suppress the table containing Gini coefficients.
display_options such as cformat() or coeflegend to format the coefficient table. See [R] estimation options.

3.2 Syntax of pshare contrast

pshare contrast computes differences in percentile shares between outcome variables or subpopulations. It requires results from pshare estimate to be in memory, which will be replaced by the results from pshare contrast. The syntax is

pshare contrast [base] [, options ]

where base is the name of the outcome variable or the value of the subpopulation to be used as base for the contrasts. If base is omitted, pshare contrast computes adjacent contrasts across outcome variables or subpopulations (or contrasts with respect to the total if total results across subpopulations have been requested). Options are:

cratio to compute contrasts as ratios instead of differences.
lnratio to compute contrasts as logarithms of ratios instead of differences.
stack[options] to draw a stacked bar chart of the results. options are as described for pshare stack below.
histogram[options] to draw a histogram of the results. options are as described for pshare histogram below.
display_options such as cformat() or coeflegend to format the coefficient table. See [R] estimation options.

9 Alternatively, to compute the contrasts directly, you may apply the contrast() option to pshare estimate (see above).
3.3 Syntax of pshare stack

pshare stack draws a stacked bar chart of percentile shares. It requires results from pshare estimate to be in memory.\textsuperscript{10} The syntax is

\texttt{pshare stack \[, \textit{options}\]}

where the options are as follows.

Main

\texttt{vertical} or \texttt{horizontal} to specify whether a vertical or a horizontal bar plot is drawn; the default is \texttt{horizontal}.

\texttt{proportion} to scale the population axis as proportion (0 to 1). The default is to scale the axis as percentage (0 to 100).

\texttt{reverse} to order percentile groups from top to bottom (the richest are leftmost, the poorest are rightmost). The default is to order percentile groups from bottom to top (the poorest are leftmost, the richest are rightmost).

\texttt{keep(list)} to select and order the results to be included as separate bars, where \texttt{list} is a space-separated list of the names of the outcome variables or the values of the subpopulations to be included. \texttt{list} may also contain \texttt{total} for the overall results if overall results were requested. Furthermore, you may use elements such as \#1, \#2, \#3, \ldots to refer to the 1st, 2nd, 3rd, \ldots outcome variable or subpopulation.

\texttt{sort([options])} to order the bars for the different outcome variables or subpopulations by the level of inequality, where \texttt{options} are \texttt{gini} to sort by Gini coefficients (if Gini coefficients have been computed), \texttt{descending} to sort in descending order, and \texttt{tfirst} or \texttt{tlast} to place the total across subpopulations first or last, respectively. The default is to sort in ascending order of the shares of the top percentile group.

\texttt{gini([fmt])} to set the format for the Gini coefficients included in the graph as secondary axis labels or \texttt{nogini} to suppresses the Gini coefficients. These options are only relevant if the \texttt{gini} option has been specified when calling \texttt{pshare estimate}. The default format is \%9.3g; see \texttt{R format}.

Labels/rendering

\texttt{labels("label 1", "label 2", \ldots)} to specify custom axis labels for the outcome variables or subpopulations.

\textsuperscript{10}You may also draw the chart directly by applying the \texttt{stack()} option to \texttt{pshare estimate} or \texttt{pshare contrast} (see above).
\texttt{plabels("label 1" "label 2" \ldots )} to specify custom legend labels for the bar segments (i.e. the percentile groups).

\texttt{barwidth(\#)} to set the width of the bars as proportion of the spacing between bar positions; the default is \texttt{barwidth(0.75)}.

\texttt{barlook\_options} and \texttt{p\#(barlook\_options)} to affect the rendition of the plotted bars, where \texttt{p\#()} applies to the \#th segment (the \#th percentile group) of the stacked bars; see \texttt{[g]} \texttt{barlook\_options}.

\texttt{values[\%fmt]} to print the values of the percentile shares as marker labels at the center of the bar segments. The default format is \texttt{\%9.3g}; see \texttt{[p]} \texttt{format}.

\texttt{marker\_label\_options} to affect the rendition of the marker labels if \texttt{values()} is specified; see \texttt{[g]} \texttt{marker\_label\_options}. \texttt{marker\_label\_options} may also be included in \texttt{p\#()} to affect the rendition of the marker labels for selected percentile groups.

**Standard twoway options**

\texttt{addplot()} to add other plots to the generated graph; see \texttt{[c]} \texttt{addplot\_option}.

\texttt{twoway\_options} to affect the overall look of the graph, manipulate the legend, set titles, add lines, etc.; see \texttt{[c]} \texttt{twoway\_options}.

**3.4 Syntax of pshare histogram**

\texttt{pshare histogram} draws a histogram of percentile shares or percentile share contrasts. It requires results from \texttt{pshare estimate} or \texttt{pshare contrast} to be in memory.\textsuperscript{11} The syntax is

\texttt{pshare histogram [, options]}

where the options are as follows.

**Main**

\texttt{vertical} or \texttt{horizontal} to specify whether a vertical or a horizontal plot is drawn; the default is \texttt{vertical}.

\texttt{proportion} to scale the population axis as proportion (0 to 1). The default is to scale the axis as percentage (0 to 100).

\texttt{keep(list)} to select and order the results to be included as separate subgraphs, where \texttt{list} is a list of the names of the outcome variables or the values of the subpopulations to

\textsuperscript{11}You may also draw the histogram directly by applying the \texttt{histogram()} option to \texttt{pshare estimate} or \texttt{pshare contrast} (see above).
be included. *list* may also contain *total* for the overall results if overall results were requested. Furthermore, you may use elements such as \#1, \#2, \#3, … to refer to the 1st, 2nd, 3rd, … outcome variable or subpopulation.

\textit{max}([\#, \textit{options}]) to top-code results at \# and \textit{min}([\#, \textit{options}]) to bottom-code results at \#. This is useful if there are large differences in the plotted values and you want to restrict the axis range. The truncated values will be included in the graph as marker labels. \textit{options} are \texttt{format}(%\textit{fmt}) to set the format for the marker labels (default is \%9.3g; see \[R\] \texttt{format}), \textit{marker_label_options} to affect the rendition of the marker labels (see \[G\] \texttt{marker_label_options}), and \texttt{nolabels} to omit the marker labels.

\textit{prange}(\textit{min max}) to restrict range of percentile groups to be included in the graph. \textit{min} and \textit{max} are numbers within [0, 100]. Only results for percentile groups whose lower and upper cumulative population bounds (in percent) are within \textit{min} and \textit{max} will be plotted.

\textit{gini}(%\textit{fmt}) to set the format for the Gini coefficients included in the subgraph labels or \texttt{nogini} to suppresses the Gini coefficients. These options are only relevant if the \texttt{gini} option has been specified when calling \texttt{pshare estimate}. The default format is \%9.3g; see \[R\] \texttt{format}.

\textbf{Labels/rendering}

\textit{barlook_options} to affect the rendition of the plotted bars; see \[G\] \texttt{barlook_options}.

\texttt{step} to use a step function (line plot) instead of a histogram to draw the results. Use \textit{line_options} instead of \texttt{barlook_options} to affect the rendition of the plotted line; see \[G\] \texttt{line_options}. \texttt{step} may be included in \texttt{o\#()}, if \texttt{overlay} has been specified, to apply \texttt{step} to selected outcome variables or subpopulations (see below).

\texttt{spikes}[\#] to use (equally spaced) spikes instead of histogram bars to draw the results. \# specifies the number of spikes; the default is \# = 100. Use \texttt{line_options} instead of \texttt{barlook_options} to affect the rendition of the plotted spikes; see \[G\] \texttt{line_options}. Confidence intervals will be omitted.

\texttt{labels}("label 1" "label 2" ...) to specify custom labels for the subgraphs of the outcome variables or subpopulations.

\texttt{byopts}(%\textit{byopts}) to determine how subgraphs are combined; see \[G\] \texttt{by_option}.

\texttt{overlay} to include results from multiple outcome variables or subpopulations in the same plot instead of creating subgraphs. Use \texttt{o\#(barlook_options)} to affect the rendition of the bars of the \#th outcome variable or subpopulation. Confidence intervals will be omitted. \texttt{overlay} cannot be combined with \texttt{psep}.

\texttt{psep}(["label 1" "label 2" ...]) to use a different rendering for each percentile group and include a corresponding legend in the graph (custom labels for the legend keys can
be specified in parentheses). The default is to draw all bars in the same style. Use 
\texttt{p\#(barlook\_options)} to affect the rendition of the bars of the \#th percentile group.

**Confidence intervals**

\texttt{level(\#)} to specify the confidence level, as a percentage, for confidence intervals. The 
default is the level that has been used for computing the \texttt{pshare} results. \texttt{level()} cannot 
be used together with \texttt{ci(bc)}, \texttt{ci(bca)}, or \texttt{ci(percentile)}.

\texttt{ci(citype)} to choose the type of confidence intervals to be plotted for results that have been 
computed using the bootstrap technique. \texttt{citype} may be \texttt{normal} (normal-based CIs; the 
default), \texttt{bc} (bias-corrected CIs) \texttt{bca} (bias-corrected and accelerated CIs) \texttt{percentile} 
(percentile CIs). \texttt{bca} is only available if BC\textsubscript{a} confidence intervals have been requested 
when running \texttt{pshare estimate} (see \texttt{R bootstrap}).

\texttt{ciopts(options)} to affect the rendition of the plotted confidence spikes. \texttt{options} depend 
on the plot type used for the confidence spikes. The default plot type is capped 
spikes; see \texttt{[c] graph twoway rcap}. To use uncapped spikes, for example, type 
\texttt{ciopts(recast(rspike))}; see \texttt{[c] graph twoway rspike}. \texttt{ciopts()} may be included 
in \texttt{p\#()}, if \texttt{psep} has been specified, to affect the rendition of the confidence spikes for 
selected percentile groups.

\texttt{cibelow} to place confidence interval spikes behind the plotted bars. The default is to draw 
the spikes in front of the bars.

\texttt{noci} to omit confidence interval spikes from the plot.

**Standard twoway options**

\texttt{addplot()} to add other plots to the generated graph; see \texttt{[c] addplot\_option}.

\texttt{twoway\_options} to affect the overall look of the graph, manipulate the legend, set titles, add 
lines, etc.; see \texttt{[c] twoway\_options}. 19
4 Examples

4.1 Basic application

By default, `pshare` computes outcome shares of quintile groups. The following example shows the results for wages in the 1988 extract of the NLSW data shipped with Stata:

```
. sysuse nlsw88
(NLSW, 1988 extract)
. pshare estimate wage, percent
```

Percentile shares (percent)  Number of obs  = 2,246

<table>
<thead>
<tr>
<th>wage</th>
<th>Coef.</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-20</td>
<td>8.018458</td>
<td>.1403194</td>
<td>7.743288 – 8.293627</td>
</tr>
<tr>
<td>20-40</td>
<td>12.03655</td>
<td>.1723244</td>
<td>11.69862 – 12.37448</td>
</tr>
<tr>
<td>40-60</td>
<td>16.2757</td>
<td>.2068139</td>
<td>15.87013 – 16.68127</td>
</tr>
<tr>
<td>60-80</td>
<td>22.47824</td>
<td>.2485367</td>
<td>21.99085 – 22.96562</td>
</tr>
<tr>
<td>80-100</td>
<td>41.19106</td>
<td>.6246426</td>
<td>39.96612 – 42.41599</td>
</tr>
</tbody>
</table>

Option `percent` was specified to express results as percentages. We can see, for example, that the 20 percent best earning women in the data receive 41% of the total of wages, whereas the 20 percent poorest earning women only receive 8%. If wages were distributed evenly, then all quintile groups would receive 20%.

To compute decile shares we could type:

```
. pshare estimate wage, percent nquantiles(10)
```

Percentile shares (percent)  Number of obs  = 2,246

<table>
<thead>
<tr>
<th>wage</th>
<th>Coef.</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-10</td>
<td>3.426509</td>
<td>.0702149</td>
<td>3.288816 – 3.564202</td>
</tr>
<tr>
<td>10-20</td>
<td>4.591949</td>
<td>.0813845</td>
<td>4.432352 – 4.751546</td>
</tr>
<tr>
<td>20-30</td>
<td>5.544608</td>
<td>.0842676</td>
<td>5.379357 – 5.709858</td>
</tr>
<tr>
<td>30-40</td>
<td>6.491941</td>
<td>.0934605</td>
<td>6.308663 – 6.675219</td>
</tr>
<tr>
<td>40-50</td>
<td>7.542334</td>
<td>.1023013</td>
<td>7.341719 – 7.742948</td>
</tr>
<tr>
<td>50-60</td>
<td>8.733366</td>
<td>.1131891</td>
<td>8.5114 – 8.955333</td>
</tr>
<tr>
<td>60-70</td>
<td>10.24571</td>
<td>.1284118</td>
<td>9.993888 – 10.49752</td>
</tr>
<tr>
<td>70-80</td>
<td>12.23253</td>
<td>.1367424</td>
<td>11.96438 – 12.50069</td>
</tr>
<tr>
<td>90-100</td>
<td>26.53588</td>
<td>.682887</td>
<td>25.19672 – 27.87503</td>
</tr>
</tbody>
</table>

The results indicate that the 10 percent best earning women get 26.5% of the wages, whereas the lowest paid 10 percent only get 3.4%.

`pshare` does not require the percentile groups to be of equal size. To compute the shares of, say, the bottom 50%, the mid 40% and the top 10%, we could type:
The `percentiles()` option specifies the cutoffs defining the percentile groups. That is, `percentiles(50 90)` indicates to use three groups, 0–50, 50–90, and 90–100. We see that the lower-paid half of women gets about the same share of total wages as the best-paid 10 percent.

### 4.2 Stacked bar charts

`pshare` supports two types of graphical displays of percentile shares. The first type is a stacked bar chart. For example, to compare wage distributions by some occupational groups, we could type:

```
pshare estimate wage if occupation<=4, percentiles(50 90) over(occupation) total gini
```

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

```
Percentile shares (percent) Number of obs = 1,409
1: occupation = Professional/technical
2: occupation = Managers/admin
3: occupation = Sales
4: occupation = Clerical/unskilled
```

<table>
<thead>
<tr>
<th>wage</th>
<th>Coef.</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-50</td>
<td>32.08652</td>
<td>.9560224</td>
<td>30.21114 33.9619</td>
</tr>
<tr>
<td>50-90</td>
<td>44.90042</td>
<td>.8232238</td>
<td>42.64146 45.96118</td>
</tr>
<tr>
<td>90-100</td>
<td>26.98812</td>
<td>1.337874</td>
<td>24.36368 29.61256</td>
</tr>
</tbody>
</table>

```
Percentile shares (percent) Number of obs = 1,409
1: occupation = Professional/technical
2: occupation = Managers/admin
3: occupation = Sales
4: occupation = Clerical/unskilled
```

<table>
<thead>
<tr>
<th>wage</th>
<th>Coef.</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-50</td>
<td>21.78931</td>
<td>1.909258</td>
<td>18.04401 25.53461</td>
</tr>
<tr>
<td>50-90</td>
<td>41.83106</td>
<td>2.046101</td>
<td>37.81733 45.84479</td>
</tr>
</tbody>
</table>
```
<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Gini</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0-50</td>
<td>28.06045</td>
</tr>
<tr>
<td></td>
<td>50-90</td>
<td>44.91512</td>
</tr>
<tr>
<td></td>
<td>90-100</td>
<td>27.02443</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>.3279324</td>
</tr>
<tr>
<td>Gini</td>
<td>1</td>
<td>.273825</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>.3373482</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>.2833736</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>.4357447</td>
</tr>
</tbody>
</table>

Option `over(occupation)` causes results to be computed by the subpopulations defined by the values of `occupation`, option `total` requests total result across subpopulations to be included, and option `gini` causes Gini coefficients to be computed. The `plabels()` option of `pshare stack` provides custom labels for the legend keys, the `values` option causes the values of the shares to be included as marker labels in the graph, and option `nogini` suppresses the Gini coefficients that would be included in the graph as secondary axis labels (see next example).

To sort the bars by level of inequality we could type:

```
pshare stack, plabels("bottom 50\%" "mid 40\%" "top 10\%") values nogini
```
The gini argument in sort() causes bars to be sorted by Gini coefficients, tlast specifies to place the overall results last, and descending requests sorting from highest inequality to lowest inequality. The example also illustrates how to print marker labels only for specific percentile groups. The global option mlabsize(zero) sets the size of the marker labels to zero so that they are invisible, but p3(mlabsize(small)) resets the marker label size for the third percentile group to small.

### 4.3 Histograms

The second type of graphical display supported by pshare is a percentile share histogram. The basic idea is to display a bar chart in which the area of each bar is proportional to the outcome share of the respective percentile group. An example with decile shares is as follows:

```
  . pshare estimate wage, percent nquantiles(10)
  (output omitted)
  . pshare histogram, yline(10)
```
Option *yline*(10) was added to print a reference line at 10%. This would be the share each group would receive in an equal distribution.

If percentile groups are of unequal size, then densities instead of percentages or proportions should be used to construct the histogram (otherwise the areas of the bars would no longer be proportional to the shares). Here is an example in which the top 1% is a separate group:

```
.pshare estimate wage, density percentiles(10(10)90 99)
```

Percentile shares (density)  
Number of obs  = 2,246

<table>
<thead>
<tr>
<th>wage</th>
<th>Coef.</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-10</td>
<td>.3426509</td>
<td>.0070215</td>
<td>.3288816 .3564202</td>
</tr>
<tr>
<td>10-20</td>
<td>.4591949</td>
<td>.0081384</td>
<td>.4432352 .4751546</td>
</tr>
<tr>
<td>20-30</td>
<td>.5544608</td>
<td>.0084268</td>
<td>.5379357 .5709858</td>
</tr>
<tr>
<td>30-40</td>
<td>.6491941</td>
<td>.009346</td>
<td>.6308663 .6675219</td>
</tr>
<tr>
<td>40-50</td>
<td>.7542334</td>
<td>.0102301</td>
<td>.7341719 .7742948</td>
</tr>
<tr>
<td>50-60</td>
<td>.8733366</td>
<td>.0113189</td>
<td>.85114 .8955333</td>
</tr>
<tr>
<td>60-70</td>
<td>1.024571</td>
<td>.0128412</td>
<td>.9993888 1.049752</td>
</tr>
<tr>
<td>70-80</td>
<td>1.223253</td>
<td>.0136742</td>
<td>1.196438 1.250069</td>
</tr>
<tr>
<td>80-90</td>
<td>1.46518</td>
<td>.0149372</td>
<td>1.436226 1.49481</td>
</tr>
<tr>
<td>90-99</td>
<td>2.377868</td>
<td>.0794248</td>
<td>2.222114 2.533622</td>
</tr>
<tr>
<td>99-100</td>
<td>5.135065</td>
<td>.0696951</td>
<td>4.998392 5.271739</td>
</tr>
</tbody>
</table>
```

. pshare histogram, yline(1)
Percentile share densities have an intuitive interpretation. They indicate how much each member in a group gets (on average) in relation to the overall average. In the example we see that the average pay of the lowest 10 percent is only about 35% of the overall average. On the other hand, the members in the top percentage group earn wages that are more than five times the average wage. An alternative interpretation is as follows: Think of 100 representative dollars to be distributed among 100 people. In an equal distribution everyone would get 1 dollar. If, however, you divide the 100 dollars according to the observed distribution, then the density of a particular group indicates how many representative dollars a person in that group would get. In the example above we see that the ten women at the bottom would only get 35 cents each, whereas the top women would get more than 5 dollars (about 15 times as much). We also see that about 60% of the women are below the equal distribution line (that is, receive below average wages).

Note that the percentile density histogram is closely related to the so-called quantile plot (see [r] diagnostic plots and Cox, 1999), also known as Pen’s “Parade of Dwarfs (and a few Giants)” (Pen, 1971, 48-59). The difference is that a quantile plot usually displays individual observations using the original scale of the outcome variable. In the percentile density histogram, the values are averaged within bins and normalized by the population average.

By default, pshare histogram uses the same plot style for all bars in the histogram. Specify the psep option if you want each percentile group to have its own style:

```
.pshare estimate wage, density percentiles(50 90)
Percentile shares (density) Number of obs = 2,246

wage | Coef. Std. Err. [95% Conf. Interval]
-----|----------------------
     |                     |
```
<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0-50</td>
<td>.5519468</td>
<td>.0074846</td>
<td>.5372694</td>
<td>.5666242</td>
</tr>
<tr>
<td>50-90</td>
<td>1.14667</td>
<td>.0105444</td>
<td>1.125992</td>
<td>1.167347</td>
</tr>
<tr>
<td>90-100</td>
<td>2.653588</td>
<td>.0682887</td>
<td>2.519672</td>
<td>2.787503</td>
</tr>
</tbody>
</table>

This may be useful, for example, to highlight single groups:

```
pshare estimate wage, density percentiles(10(10)90 99)
(output omitted)
.pshare histogram, yline(1) psep
> fintensity(100) color(*.8) pstyle(p1) ciopts(lstyle(none))
> p10(pstyle(p2) ciopts(lstyle(p2))) legend(off)
```
In the example, the options on the second line of the code specify the defaults to be used for all percentile groups. Option `p10()` then changes some of the settings for the 10th group.

Furthermore, you may use spikes instead of bars:

```
. pshare histogram, yline(1) psep spikes lwidth(*2)
>   pstyle(p1) p10(pstyle(p2)) legend(off)
```

By default, 100 spikes are uses, one for each percentile. Specify, e.g., `spikes(1000)` to use 1000 spikes. Confidence intervals will be omitted if `spikes` is specified.
Finally, instead of drawing histogram bars, you may also draw a step function using a line plot:

```
. pshare histogram, yline(1) step
```

The `step` option may be helpful if you want to overlay results from different groups in the same plot:

```
. pshare estimate wage, density over(union) n(10)
(output omitted)
. pshare histogram, yline(1) overlay o2(step lwidth(*2))
```
The example illustrates that the wage distribution of unionized women is less unequal than the distribution of non-unionized women, especially at the top.

4.4 Contrasts

4.4.1 Differences between subpopulations

A useful feature of \texttt{pshare} is that contrasts between distributions can be computed. As illustrated in the last example, the distribution of wages among unionized workers is somewhat less uneven than among non-unionized workers. To make the differences between the distributions more visible (and evaluate which differences are significant), the \texttt{pshare contrast} command can be applied:

```
.pshare estimate wage, density over(union) n(10)
(output omitted)
.pshare contrast 0
```

Differences in percentile shares (density)  
Number of obs = 1,878

| wage  | Coef.    | Std. Err. | t     | P>|t|   | [95% Conf. Interval] |
|-------|----------|-----------|-------|-------|----------------------|
| 0-10  | 0.0429197| 0.016305  | 2.63  | 0.009 | 0.0109419 - 0.0748975|
| 10-20 | 0.0528084| 0.0177041 | 2.98  | 0.003 | 0.0180866 - 0.0875301|
| 20-30 | 0.0743417| 0.0204516 | 3.64  | 0.000 | 0.0342315 - 0.1144519|
| 30-40 | 0.0765406| 0.018892  | 4.05  | 0.000 | 0.0394891 - 0.1135922|
From the results we see that the bottom 70% are relatively better off if unionized, the top 30% are relatively worse off. The differences are expressed in representative dollars, that is, the bottom 70% gain around five representative cents, the top 10% loose about a quarter of a representative dollar. However, note that these differences only reflect differences in the distributional shape; they are net of a possible overall difference in the wage levels between unionized and non-unionized workers.

To take the different wage levels of unionized and non-unionized workers into account, specify the *average* option so that the results are expressed as average wages. Furthermore, note that instead of using the `pshare contrast` command, contrasts can also be computed directly by applying the `contrast()` option to `pshare estimate`:

```
pshare estimate wage, average over(union) n(10) contrast(0) histogram
```

(output omitted)
From these results we see that unionized workers are better off across the board (by about 1 to 2 dollars per hour). Hence, from a welfare perspective, one could argue that the wage distribution of unionized women is strictly preferable over the wage distribution of non-unionized women (generalized Lorenz dominance; see, e.g., Lambert, 2001). We also see that the (absolute) gains are somewhat larger in the middle of the distribution than at the top and at the bottom.

Furthermore, in relative terms the differences look as follows:

```
pshare estimate wage, average over(union) n(10) contrast(0, ratio)
(output omitted)
pshare histogram
```
We see that in the lower half of the distribution, unionized workers earn about 30\% more than non-unionized workers; in the upper ranks the advantage of unionized workers shrinks to about 10\%.

4.4.2 Differences between outcome variables

Instead of comparing subpopulations, pshare can also be used to compare distributions of different variables. For example, we could be interested in how the distribution changes once we move from hourly wages to weekly earnings:

```
. generate weekly = hours * wage
(4 missing values generated)
. label variable weekly "weekly earnings"
. pshare estimate wage weekly, density n(10) contrast(wage)
(output omitted)
. pshare histogram, yline(0)
```
We see that weekly earnings are considerably more unequal than hourly wages. Apparently, and as expected by economic theory, women with higher wages do supply more labor, so that they get a larger share of weekly earnings than of hourly wages.

### 4.5 Concentration shares

The relation between two continuous variables can be analyzed by the `pshare` command using the `pvar()` option (percentile shares correspond to differences in concentration curve ordinates in this case). In the last example, we saw that weakly earnings are distributed more unequally than hourly wages, which implies that women with higher wages work longer hours. Hence, it might be interesting to see how labor supply is distributed across wage groups:

```
.pshare estimate hours, pvar(wage) average n(10)
Percentile shares (average) Number of obs = 2,242
```

<table>
<thead>
<tr>
<th>hours</th>
<th>Coef.</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-10</td>
<td>33.05259</td>
<td>.889763</td>
<td>31.30775 34.79744</td>
</tr>
<tr>
<td>10-20</td>
<td>33.6382</td>
<td>.8199639</td>
<td>32.03023 35.24616</td>
</tr>
<tr>
<td>20-30</td>
<td>34.78557</td>
<td>.7480189</td>
<td>33.31869 36.25245</td>
</tr>
<tr>
<td>30-40</td>
<td>37.14429</td>
<td>.6222536</td>
<td>35.92404 38.36454</td>
</tr>
<tr>
<td>40-50</td>
<td>37.73974</td>
<td>.6375459</td>
<td>36.4895  38.98998</td>
</tr>
<tr>
<td>50-60</td>
<td>38.6289</td>
<td>.670502</td>
<td>37.31403 39.94377</td>
</tr>
<tr>
<td>60-70</td>
<td>39.17663</td>
<td>.5903086</td>
<td>38.01902 40.33424</td>
</tr>
<tr>
<td>70-80</td>
<td>38.59946</td>
<td>.5712248</td>
<td>37.47928 39.71965</td>
</tr>
<tr>
<td>80-90</td>
<td>40.03568</td>
<td>.5799854</td>
<td>38.88322 41.17305</td>
</tr>
<tr>
<td>90-100</td>
<td>39.38002</td>
<td>.660688</td>
<td>38.08439 40.67564</td>
</tr>
</tbody>
</table>
The results indicate that average labor supply by women in the bottom 30% of the wage distribution is only about 33 to 35 hours per week, whereas in the upper half of the wage distribution it is about 40 hours per week. To obtain results expressed in relation to the overall average, use the **density** option:

\[
\texttt{. pshare estimate hours, pvar(wage) density n(10)}
\]

<table>
<thead>
<tr>
<th>Percentile shares (density)</th>
<th>Number of obs = 2,242</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-10</td>
<td>0.8880782 .0222773 .8443919 .9317646</td>
</tr>
<tr>
<td>10-20</td>
<td>0.9038126 .0205245 .8635637 .9440616</td>
</tr>
<tr>
<td>20-30</td>
<td>0.934641 .0188478 .8976801 .971602</td>
</tr>
<tr>
<td>30-40</td>
<td>0.9980166 .0159431 .9667519 1.029281</td>
</tr>
<tr>
<td>40-50</td>
<td>1.014016 .0162895 .9820715 1.04596</td>
</tr>
<tr>
<td>50-60</td>
<td>1.037906 .0170757 1.00442 1.071392</td>
</tr>
<tr>
<td>60-70</td>
<td>1.052623 .0153487 1.022524 1.082722</td>
</tr>
<tr>
<td>70-80</td>
<td>1.075704 .0149871 1.045945 1.105464</td>
</tr>
<tr>
<td>80-90</td>
<td>1.093115 .0149871 1.066505 1.127922</td>
</tr>
<tr>
<td>90-100</td>
<td>1.1058088 .0169731 1.071392 1.140312</td>
</tr>
</tbody>
</table>

We see, for example, that the weekly labor supply of women in the top 10% of the wage distribution is about 6% higher than average labor supply. The weekly labor supply of women in the bottom 10% of the wage distribution is 11% below the average.

The same technique could also be used, for example, to study the relation between income and wealth or between received bequests and existing income or wealth (e.g., how much of the sum of all bequests in a given year goes to the wealthiest 10% of the population). Furthermore, it could be used to study the composition of income by sources or to study the effects of redistribution (e.g., how much the different income percentiles contribute to overall taxes and how the empirical tax progression looks like).

### 4.6 Processing results from pshare

**pshare estimate** and **pshare contrast** post their result in the `e()` returns (see [r] **ereturn**; also see [r] **13.5 Accessing coefficients and standard errors**) so that they can be processed by post estimation commands such as **test** ([r] **test**), **lincom** ([r] **lincom**), and **nlcom** ([r] **nlcom**) or tabulated and graphed by programs such as **estout** (Jann, 2005, 2007) and **coefplot** (Jann, 2014).

For example, to compute the Palma ratio of wages (top 10% share divided by bottom 40% share; see, e.g., Cobham et al., 2015), we could type:

\[
\texttt{. pshare estimate wage, percentiles(40 90)}
\]
Percentile shares (proportion)  Number of obs = 2,246

<table>
<thead>
<tr>
<th></th>
<th>Coef.</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-40</td>
<td>0.20055</td>
<td>0.002916</td>
<td>0.1948315 0.2062687</td>
</tr>
<tr>
<td>40-90</td>
<td>0.53409</td>
<td>0.004878</td>
<td>0.5245258 0.5436566</td>
</tr>
<tr>
<td>90-100</td>
<td>0.26536</td>
<td>0.006829</td>
<td>0.2519672 0.2787503</td>
</tr>
</tbody>
</table>

```
.nlcom (Palma: _b[90-100] / _b[0-40])
```

Palma: _b[90-100] / _b[0-40]

|        | Coef.   | Std. Err. | z   | P>|z| | [95% Conf. Interval] |
|--------|---------|-----------|-----|-------|----------------------|
| Palma  | 1.32316 | 0.050604  | 26.15| 0.000 | 1.223972 1.422337    |

Furthermore, the Lorenz ordinates used to compute the percentile shares are stored by `pshare` in `e(L_ll)` (lower bounds) and `e(L_ul)` (upper bounds). To tabulate the Lorenz ordinates together with the percentile shares we could type:

```
.pshare estimate wage
(output omitted)
.estout, cell((b(label/share)) L_ll L_ul) mlabels(none)
```

<table>
<thead>
<tr>
<th>share</th>
<th>L_ll</th>
<th>L_ul</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-20</td>
<td>0.0801846</td>
<td>0.0801846</td>
</tr>
<tr>
<td>20-40</td>
<td>0.1203655</td>
<td>0.2005501</td>
</tr>
<tr>
<td>40-60</td>
<td>0.162757</td>
<td>0.3633071</td>
</tr>
<tr>
<td>60-80</td>
<td>0.2247824</td>
<td>0.5880894</td>
</tr>
<tr>
<td>80-100</td>
<td>0.4119106</td>
<td>1</td>
</tr>
</tbody>
</table>

Finally, `estimates store` (see [r] `estimates store`) can be used to make copies of results from different calls to `pshare` for later usage by commands such as `estout` or `coefplot`. In the following example, `coefplot` is used to plot the top decile share and the top centile share of weekly earnings against time:

```
use http://www.stata-press.com/data/r14/nlswork.dta, clear
(National Longitudinal Survey. Young Women 14-26 years of age in 1968)
.gen weekly = exp(ln_wage) * hours
(67 missing values generated)
.pshare estimate weekly, percent percentile(90) over(year) vce(cluster idcode)
(output omitted)
estimates store p90
.pshare estimate weekly, percent percentile(99) over(year) vce(cluster idcode)
```

35
Across the years, as the respondents grew older, the share of the top decile increased from about 18% to 30%. The share of the top centile increased from 2.5% to about 5%.\footnote{Option \texttt{vce(cluster idcode)} has been added because the data are from a panel study where \texttt{idcode} identifies individuals. Adding the option in the example is not strictly necessary as the variances of the yearly estimates are not affected much by the clustering. It will be relevant, however, once differences between years are analyzed.}
5 Small-sample bias

Estimates of percentile shares are affected by small-sample bias, especially at the top of the distribution. The bias can be substantial if the distribution is highly skewed and the number of observations is small. Consequently, to obtain reliable estimates for shares of small top groups such as, say, the top 0.1% share, large samples are required.

The simulation below provides some results for the relative bias in the estimate of the top 1% share for different sample sizes using a log-normal distribution. The scale parameter of the log-normal distribution is varied between $\sigma = 0.5$ (corresponding to a Gini coefficient of 0.276) and $\sigma = 2$ (corresponding to a Gini coefficient of 0.843).

```
. set seed 3230982
. program mysim, rclass
1.   syntax [, n(integer 1000) Sigma(real 1) ]
2.   drop _all
3.   qui set obs `n`
4.   tempvar y
5.   gen `y´ = exp(rnormal(0, `sigma´))
6.   pshare estimate `y´, nose percentile(99)
7.   local b = 1 - normal(invnorm(0.99) - `sigma´)
8.   return scalar bias = (_b[99-100] - `b´) / `b´
9. end
.
. local i 0
. capture matrix drop R
. foreach sigma in 0.5 1 1.5 2 {
2.   local ++i
3.   local gini = 2*normal(`sigma´/sqrt(2)) - 1
4.   foreach n in 100 500 1000 5000 10000 {
5.       quietly simulate r(bias), reps(10000): mysim, n(`n´) sigma(`sigma´)
6.       quietly ci means _sim_1
7.       matrix tmp = r(mean), r(lb), r(ub)
8.       matrix rownames tmp = s`i´:`n´
9.       matrix R = nullmat(R), tmp`
10.   }
11. }
.
. local i 0
. local plots
.
. foreach sigma in 0.5 1 1.5 2 {
2.   local ++i
3.   local lbl `: di %9.3f 2*normal(`sigma´/sqrt(2)) - 1´
4.   local lbl Gini = `lbl´ ({$\sigma$} = `sigma´)
5.   local plots `plots´ (matrix(R), keep(s`i´:) label("$\sigma$"="`sigma´"))
6. }
.
. coefplot `plots´, ci((R[2] R[3])) vertical nooffset rescale(100)
>  msymbol(d) xtitle(Sample size) recast(connected) ciopts(recast(rcap))
>  ytitle(Bias in %) ylabel(#10, angle(horizontal)) yline(0)
>  title(Bias in top centile share) legend(cols(1) position(0) bplace(se))
```
For example, in a sample of 100 observations, the top centile share is underestimated by about 30% for a log-normal distribution with a Gini coefficient of 0.843. For lower levels of inequality, the underestimation is less severe but still substantial. This is not much of a surprise since in a sample of 100 observation the top centile group only contains a single observation. However, also with a sample size of 1000, the top centile share is underestimated by about 5% for the distribution with a Gini coefficient of 0.843.

The simulation results suggest that for moderately skewed distributions (such as the income distribution with a typical Gini coefficient between around 0.3 and 0.6) there should be a minimum of about 10 observations in the top group to keep the error within acceptable bounds of just a few percent. To estimate the top 0.1% share, for example, a sample size of at least 10000 observations would be required. For accurate estimation of top shares in extremely skewed distributions (such as the wealth distribution with Gini coefficients as high as 0.8 or even 0.9), however, minimum sample size requirements may be considerably higher (such as 50 or even 100 observations in the top group).
6 Discussion

Only a selection of the features of the \texttt{pshare} command were be presented in this article. The command has been designed in a way such that it offers a wide variety of possible applications and can be used in many different situations. For example, much effort has been put into the support for complex survey data, a topic that has not been touched in the presented examples. Nonetheless, a number of limitations and remaining issues are to be mentioned.

First, \texttt{pshare} is designed to be applied to individual-level data. Often, however, data on the distribution of income or wealth is available in form of aggregate tables (typically from tax statistics). In such tables, individual-level units are grouped into outcome brackets and for each bracket the number of units and the outcome total is reported. \texttt{pshare} can be applied to such grouped data by computing the average outcome per bracket and weighting the data by the number of units. However, such a procedure assumes perfect equality within brackets and hence only provides a lower bound of the true inequality in the distribution (see, e.g., Cowell, 2011). It would be worthwhile to develop a companion command for grouped data that also offers upper bound estimates and intermediate estimates.

Second, analytic variance estimation implemented in \texttt{pshare} is only approximate and, possibly, more accurate estimation procedures could be developed. For example, variance estimation for percentile shares based on the concentration curve (that is, if the \texttt{pvar()} option is specified) requires the estimation of the expectation of the outcome variable at specific quantiles of the auxiliary variable. In the current implementation of \texttt{pshare} this is accomplished by local linear regression using a constant bandwidth (see footnote 6). Some preliminary simulations indicate that this procedure generates consistent estimates of standard errors. However, possibly, the accuracy and stability of the standard error estimates could be improved by using a variable bandwidth depending on the local density of the data. Furthermore, \texttt{pshare} reports symmetric, normal-based confidence intervals that may not be very accurate in small samples. A topic for future research could thus be to develop refined estimation of confidence intervals.

Third, as discussed above, percentile shares are affected by small-sample bias. Future research will have to show whether a suitable correction procedure can be designed. A main challenge is to ensure that the correction does not increase the mean squared error (MSE) of the estimates. The problem can be illustrated by a simple bootstrap correction procedure. Let \( \hat{S} \) be the uncorrected estimate in the original sample and \( \overline{S} \) be the mean of the estimates from a number of bootstrap samples. The bias in the bootstrap samples with respect to the original sample is then given as \( \overline{S} - \hat{S} \). The idea is to use the bootstrap bias as an approximation of the bias of the sample with respect to the population. Hence, a corrected estimate of \( S \) can be obtained as \( \hat{S}^{\text{corr}} = \hat{S} - (\overline{S} - \hat{S}) = 2\hat{S} - \overline{S} \). Alternatively, the correction could also be based on ratios or on odds ratios between \( \overline{S} \) and \( \hat{S} \). Findings from simulations with such procedures are that the bootstrap correction mostly removes the bias, unless the distribution is extremely skewed. At the same time, however, MSE increases.
for this is quite obvious: the larger the top share in a given sample turns out to be, the larger will be the bootstrap correction. This inflates sampling variance. Possibly, however, parametric extreme-value estimation may be used to design a correction procedure that does not increase the MSE.
7 Acknowledgments

The histogram and stacked bar plots produced by \texttt{pshare} have been inspired by the graphs shown in a video posted by Evan Klassen on Youtube\textsuperscript{13} and a TED talk by Dan Ariely.\textsuperscript{14} The \texttt{max()} and \texttt{min()} options of \texttt{pshare histogram} have been independently suggested by Hans-Jürgen Andreß and Jonas Meier. Furthermore, I would like to thank Stephen P. Jenkins for his helpful advice.

This research has been supported by the Swiss National Science Foundation (Grant No. 143399).

\textsuperscript{13}http://www.youtube.com/watch?v=slTF_XXoKAQ
\textsuperscript{14}https://www.ted.com/talks/dan_ariely_how_equal_do_we_want_the_world_to_be_you_d_be_surprised
References


