Assessing inequality using percentile shares

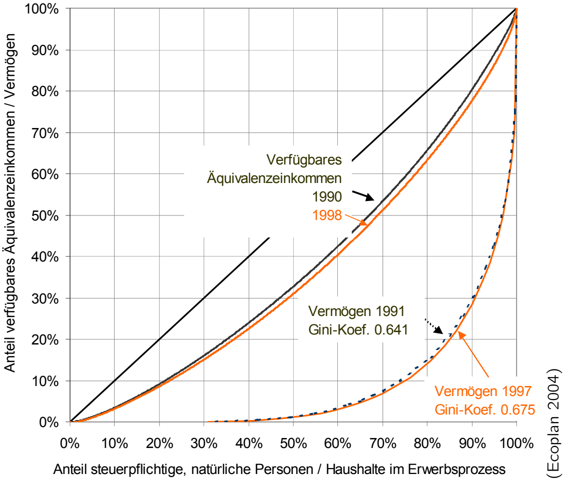
An application to Swiss tax data

Ben Jann

University of Bern, ben.jann@soz.unibe.ch

Seminar at the Institute of Social and Preventive Medicine (ISPM)
University of Bern, April 28, 2016


HAPPY
BIRTHDAY



- ▷ http://www.youtube.com/watch?v=slTF_XXoKAQ
- ▷ https://www.ted.com/talks/dan_ariely_how_equal_do_we_want_the_world_to_be_you_d_be_surprised

Outline

- Motivation
- Estimation of percentile shares
- The `pshare` Stata command
- Examples
- Small sample bias

Estimation of percentile shares

- Outcome variable of interest, e.g. income: Y
- Distribution function: $F(y) = \Pr\{Y \leq y\}$
- Quantile function: $Q(p) = F^{-1}(p) = \inf\{y | F(y) \geq p\}$, $p \in [0, 1]$
- Lorenz ordinates:

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

- Finite population form:

$$L(p) = \sum_{i=1}^{N} y_i \mathcal{I}\{y_i \leq Q_p\} \Big/ \sum_{i=1}^{N} y_i$$

- Percentile share: proportion of total outcome within quantile interval $[Q_{p_{\ell-1}}, Q_{p_\ell}]$, $p_{\ell-1} \leq p_\ell$

$$S_\ell = L(p_\ell) - L(p_{\ell-1})$$

Estimation of percentile shares

- Estimation given sample of size n:

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

$$\widehat{L}(p) = (1-\gamma)\widetilde{Y}_{j-1} + \gamma\widetilde{Y}_j \quad \text{where } \widehat{p}_{j-1} < p \leq \widehat{p}_j \text{ with } \widehat{p}_j = \frac{j}{n}$$

$$\widetilde{Y}_j = \sum_{i=1}^{j} y_{(i)} \Big/ \sum_{i=1}^{n} y_i \quad \text{where } y_{(i)} \text{ refers to ordered values}$$

$$\gamma = \frac{p - \widehat{p}_{j-1}}{\widehat{p}_j - \widehat{p}_{j-1}} \quad \text{(linear interpolation)}$$

- Standard errors
  - approximate standard errors can be obtained by the estimating equations approach as proposed by Binder and Kovacevic (1995)
  - supports complex survey data and joint estimation across subpopulations or repeated measures
  - alternative: bootstrap

Estimation of percentile shares: standard errors

- Let θ be a parameter interest and λ be a vector of nuisance parameters. Furthermore, let $u_\theta(y_i, \theta, \lambda)$ and $u_\lambda(y_i, \lambda)$ be estimating functions such that, in the (finite) population, θ and λ are the solutions to

$$U_\theta(\theta, \lambda) = \sum_{i=1}^{N} u_\theta(y_i, \theta, \lambda) = 0 \quad \text{and} \quad U_\lambda(\lambda) = \sum_{i=1}^{N} u_\lambda(y_i, \lambda) = 0$$

- Following Kovacević and Binder (1997), the sampling variance of $\hat{\theta}$ can be approximated by a variance estimate of

$$\sum_{s} w_i u^*(y_i, \hat{\theta}, \hat{\lambda})$$

where w_i are sampling weights and

$$u^*(y_i, \theta, \lambda) = \left(-u_\theta(y_i, \theta, \lambda) + \frac{\partial U_\theta}{\partial \lambda} \left[\frac{\partial U_\lambda}{\partial \lambda} \right]^{-1} u_\lambda(y_i, \lambda) \right) \left[\frac{\partial U_\theta}{\partial \theta} \right]^{-1}$$

Estimation of percentile shares: standard errors

- For percentile shares, $\theta = S$ and $\lambda = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}$.
- The estimating functions are:

$$u_\theta = y_i \mathcal{I}\{y_i \leq Q_2\} - y_i \mathcal{I}\{y_i \leq Q_1\} - y_i S$$

$$u_\lambda = \begin{bmatrix} \mathcal{I}\{y_i \leq Q_1\} - p_1 \\ \mathcal{I}\{y_i \leq Q_2\} - p_2 \end{bmatrix}$$

- Hence:

$$u^* = \frac{\begin{array}{c} y_i \mathcal{I}\{y_i \leq Q_2\} - y_i \mathcal{I}\{y_i \leq Q_1\} - y_i S \\ -Q_2(\mathcal{I}\{y_i \leq Q_2\} - p_2) + Q_1(\mathcal{I}\{y_i \leq Q_1\} - p_1) \end{array}}{\sum y_i}$$

$$= \frac{\begin{array}{c} (y_i - Q_2)\mathcal{I}\{y_i \leq Q_2\} - (y_i - Q_1)\mathcal{I}\{y_i \leq Q_1\} \\ +Q_2 p_2 - Q_1 p_1 - y_i S \end{array}}{\sum y_i}$$

Estimation of percentile shares: some extensions

- Percentile share "density":
  - particularly useful for graphing

$$D_\ell = \frac{S_\ell}{p_\ell - p_{\ell-1}} = \frac{L(p_\ell) - L(p_{\ell-1})}{p_\ell - p_{\ell-1}}$$

- Totals:

$$T_\ell = \sum_{i=1}^{N} y_i \mathcal{I}\{Q_{p_{\ell-1}} < y_i \leq Q_{p_\ell}\} = S_\ell \cdot \sum_{i=1}^{N} y_i$$

- Averages:
  - again, useful for graphing
  - useful if you are also interested in levels, not just distribution

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

Estimation of percentile shares: some extensions

- Contrasts:
  - useful for comparing distributions, e.g. changes over time
  - standard errors easily computed using delta method

$$S_\ell^A - S_\ell^B \qquad S_\ell^A / S_\ell^B \qquad \ln(S_\ell^A / S_\ell^B) \qquad \ldots$$

- Renormalization (using a different total):
  - useful, e.g., to analyze income components or subpopulation shares

$$L^*(p) = \sum_{i=1}^{N} y_i \, \mathcal{I}\{y_i \leq Q_p\} \, / \, T$$

$$S_\ell^* = L^*(p_\ell) - L^*(p_{\ell-1})$$

with T whatever you like it to be (e.g. the total of variable Z or the total across subpopulations)

Estimation of percentile shares: some extensions

- Concentration shares:
  - compute shares while ordering by a different variable
  - useful for analyzing relations between variables (wealth and income, pre- and post-tax income, etc.)

$$L^Z(p) = \sum_{i=1}^{N} y_i \mathcal{I}\{z_i \leq Q_p^Z\} \Bigg/ \sum_{i=1}^{N} y_i$$

$$S_\ell^Z = L^Z(p_\ell) - L^Z(p_{\ell-1})$$

- Often a combination of renormalization and using a different ordering variable is useful (e.g. to analyze redistribution).

The `pshare` Stata command

- `pshare estimate`
  - estimates the percentile shares and their variance matrix
  - arbitrary cutoffs for the percentile groups
  - joint estimation across multiple outcome variables or subpopulations
  - shares as proportions, densities, totals, or averages
  - etc.
- `pshare contrast`
  - computes contrasts between outcome variables or subpopulations
  - differences, ratios, or log ratios
- `pshare stack`
  - displays percentile shares as stacked bar chart
- `pshare histogram`
  - displays percentile shares as histogram

Example: quintile shares (the default)

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

. pshare estimate wage, percent

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

--------------------------------------------------------------
        wage |      Coef.   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
        0-20 |   8.018458   .1403194      7.743288    8.293627
       20-40 |   12.03655   .1723244      11.69862    12.37448
       40-60 |    16.2757   .2068139      15.87013    16.68127
       60-80 |   22.47824   .2485367      21.99085    22.96562
      80-100 |   41.19106   .6246426      39.96612    42.41599
--------------------------------------------------------------
```

- top 20% percent of the population get 41% of wages
- bottom 20% only get 8% of wages, etc.

Example: bottom 50%, mid 40%, and top 10%

```
. pshare estimate wage, percent percentiles(50 90)
```

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

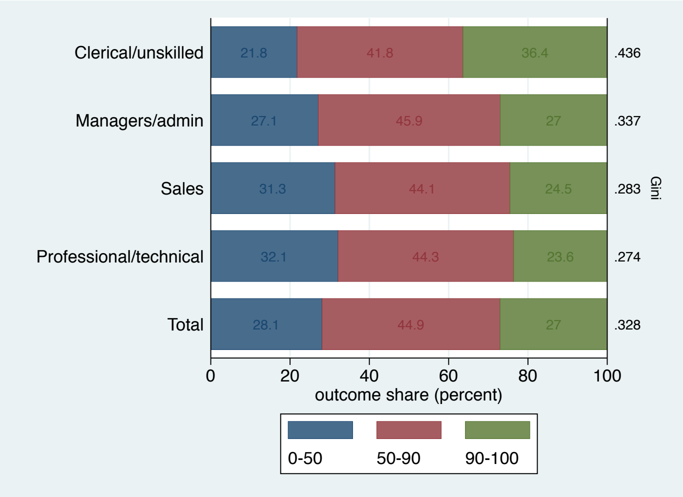
wage	Coef.	Std. Err.	[95% Conf.	Interval]
0-50	27.59734	.3742279	26.86347	28.33121
50-90	45.86678	.4217771	45.03967	46.6939
90-100	26.53588	.682887	25.19672	27.87503

Example: stacked bars plot

```
. pshare estimate wage if occ<=4, percent p(50 90) over(occ) total gini
```

(*output omitted*)

```
. pshare stack, values sort(gini tlast descending)
```

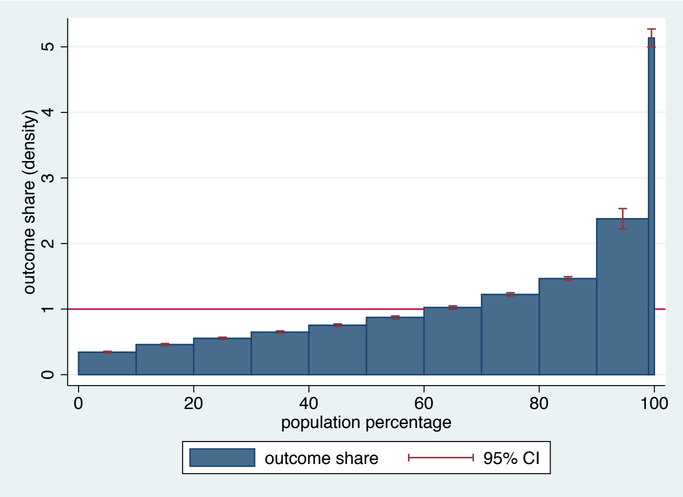


Example: histogram of densities

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

(*output omitted*)

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

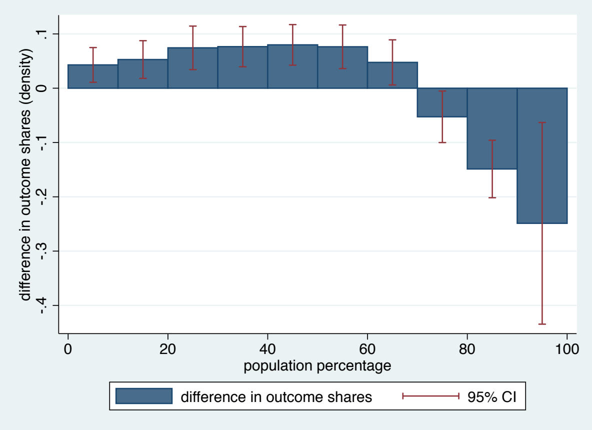


Example: histogram of densities

- Interpretation
  - Take 100 dollars and divide them among 100 people who line up along the x-axis.
  - The heights of the bars shows you how much each one gets.
  - If all get the same, then everyone would get one dollar (red line).
  - However, according to the observed distribution, the rightmost person would get five of the 100 dollars, the next 9 would get about two and a half dollars each, ..., the bottom 10 only get 35 cents each.

Example: contrasts

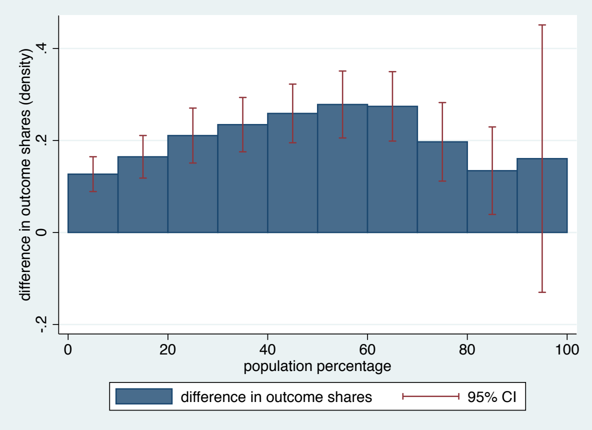
```
. quietly pshare estimate wage, n(10) density over(union)
. quietly pshare contrast 0
. pshare histogram
```



- bottom 70% percent are *relatively* better off if unionized

Example: contrasts

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

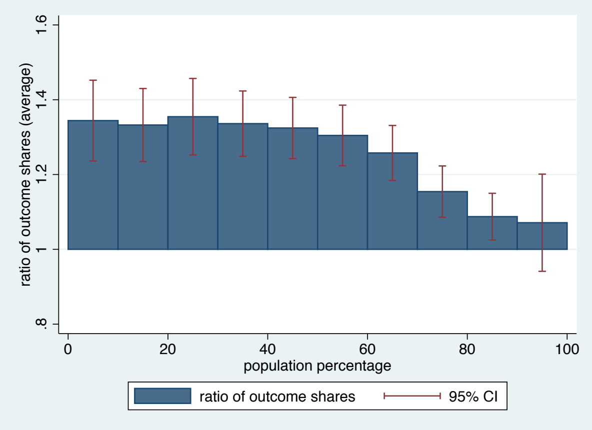


- everyone is *absolutely* better off if unionized (between about 15% and 25% of average nonunion wages)

Example: contrasts

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

(*output omitted*)



- bottom 50% of unionized are about 30% better off than bottom 50% of nonunionized; at the top the advantage shrinks to 10%

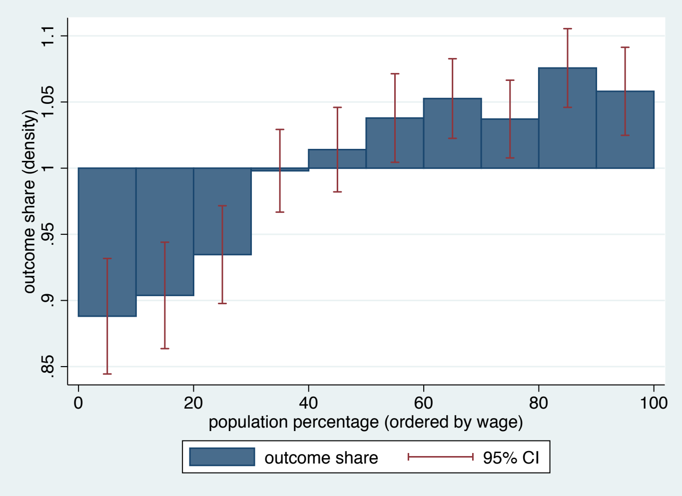
Example: concentration shares

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

hours	Coef.	Std. Err.	[95% Conf.	Interval]
0-10	.8880782	.0222773	.8443919	.9317646
10-20	.9038126	.0205245	.8635637	.9440616
20-30	.934641	.0188478	.8976801	.971602
30-40	.9980166	.0159431	.9667519	1.029281
40-50	1.014016	.0162895	.9820715	1.04596
50-60	1.037906	.0170757	1.00442	1.071392
60-70	1.052623	.0153487	1.022524	1.082722
70-80	1.037115	.0149871	1.007725	1.066505
80-90	1.075704	.0151754	1.045945	1.105464
90-100	1.058088	.0169731	1.024803	1.091372

```
(percentile groups with respect to wage)
. pshare histogram, base(1)
```

Example: concentration shares



- the 10% with the highest wages work 5.8% longer hours
- the 10% with the lowest wages work 11.2% shorter hours

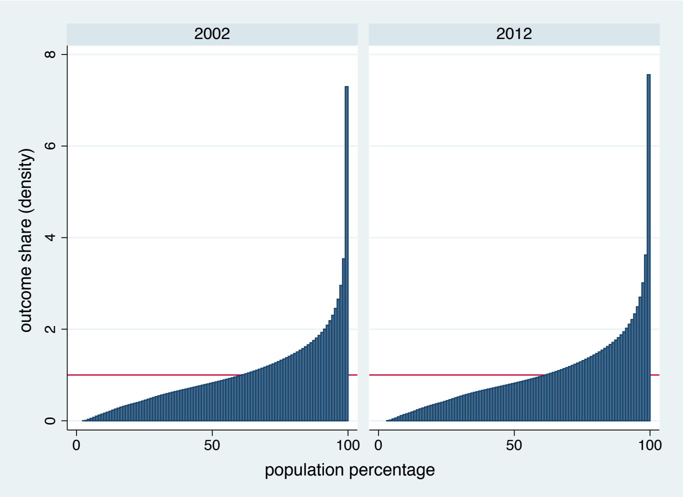
Some examples with “real” data

- Tax data from canton of Bern, Switzerland, 2002 and 2012
- individual level data from personal tax forms
- information on income components, deductions, assets, etc.
- units of analysis in following examples are “tax units”

```
. describe

Contains data from BE-02-12.dta
  obs:     1,153,709
 vars:            10                          28 Apr 2016 15:17
 size:    48,455,778
--------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
--------------------------------------------------------------------------------
year            int     %9.0g                 Year
hhid            double  %10.0g                Household ID
earnings        float   %9.0g                 Labor market income
capincome       float   %9.0g                 Capital income
transfers       float   %9.0g                 Transfer income
tax             float   %9.0g                 Tax
heritage        long    %10.0gc               Received heritage
income          float   %9.0g                 Total income
aftertax        float   %9.0g                 After tax income
wealth          float   %9.0g                 Net wealth
--------------------------------------------------------------------------------
Sorted by:
```

Distribution of total income in 2002 and 2012

```
. pshare estimate income, n(100) nose density over(year)
  (output omitted)
. pshare histogram, yline(1)
```

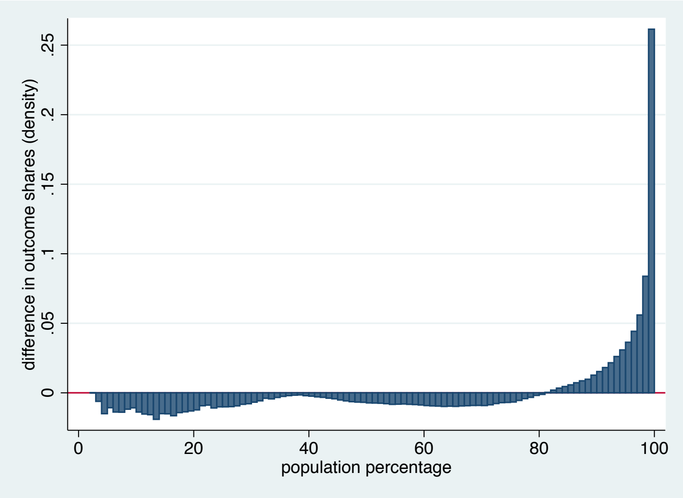


Change in income distribution from 2002 to 2012

```
. pshare contrast
```

(*output omitted*)

```
. pshare histogram, yline(0)
```

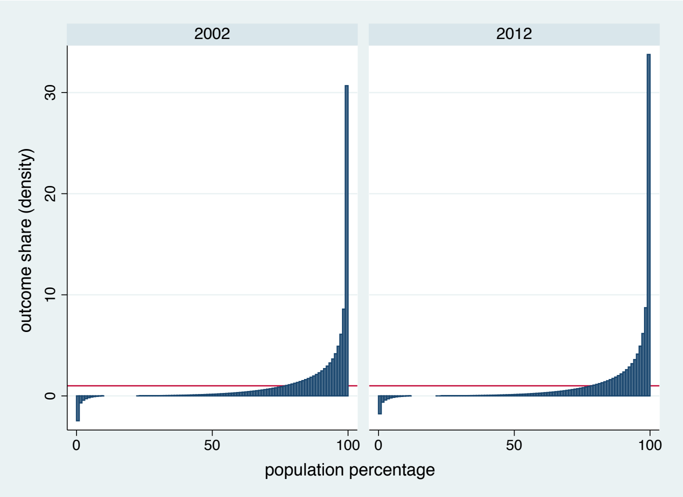


Distribution of net wealth in 2002 and 2012

```
. pshare estimate wealth, n(100) nose density over(year)
```

(*output omitted*)

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

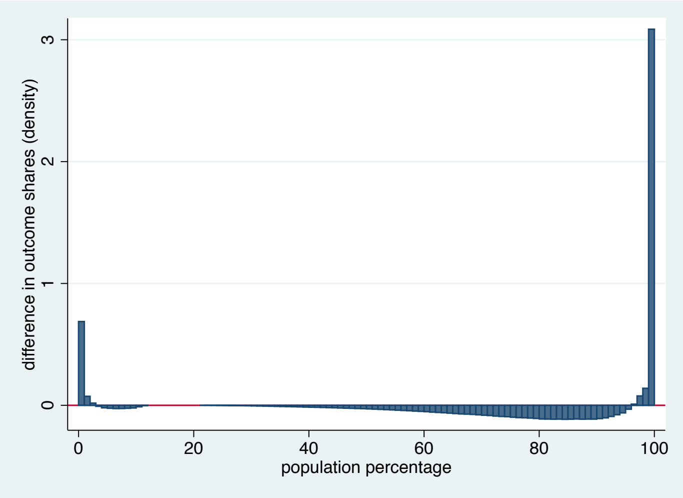


Change in wealth distribution from 2002 to 2012

```
. pshare contrast
```

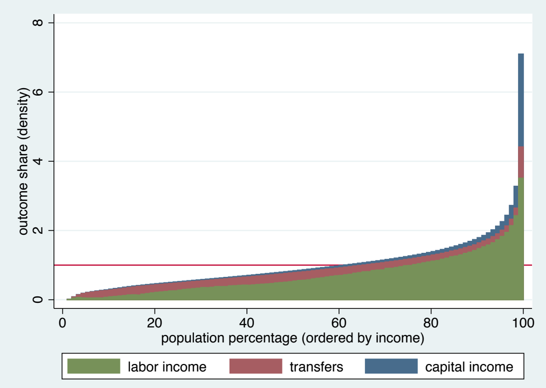
(*output omitted*)

```
. pshare histogram, yline(0)
```



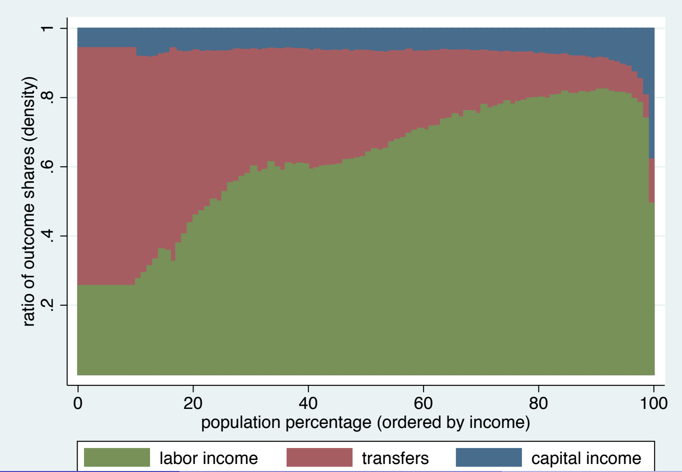
Income composition by income percentiles (2012)

```
. keep if year==2012
(553,976 observations deleted)

. drop year

. drop if hhid>=.
(11,720 observations deleted)

. collapse (sum) earnings-wealth, by(hhid) fast // generate households

. generate earn_trans = earnings + transfers

. quietly pshare estimate income earn_trans earnings, n(100) nose density ///
>         pvar(income) normalize(income)

. pshare histogram, overlay yline(1) fintensity(100) color(*.8) ///
>     legend(order(3 "labor income" 2 "transfers" 1 "capital income") rows(1))
```



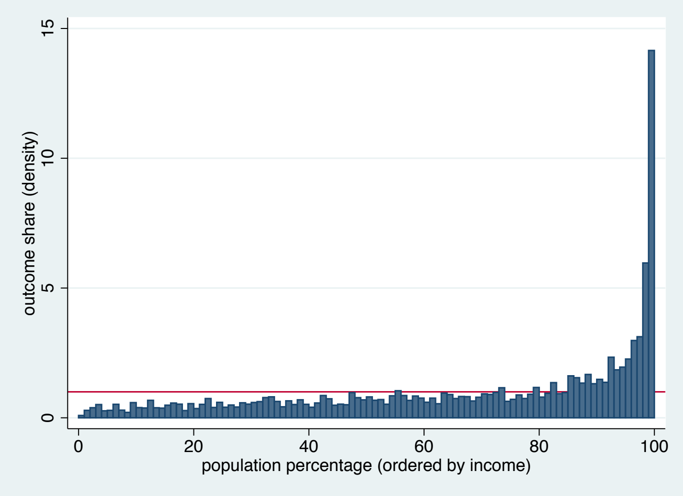
Income composition in relative terms (2012)

```
. generate earn_trans_cap = income

. quietly pshare estimate income earn_trans_cap earn_trans earnings, ///
>     p(10(1)99) nose density pvar(income) normalize(income)

. quietly pshare contrast income, ratio

. pshare histogram, overlay finten(100) color(*.8) base(0) ///
>     legend(order(3 "labor income" 2 "transfers" 1 "capital income") rows(1))
```



Received heritage by income percentiles (2012)

```
. quietly pshare estimate heritage, n(100) nose density pvar(income)
```
(*output omitted*)
```
. pshare histogram, yline(1)
```

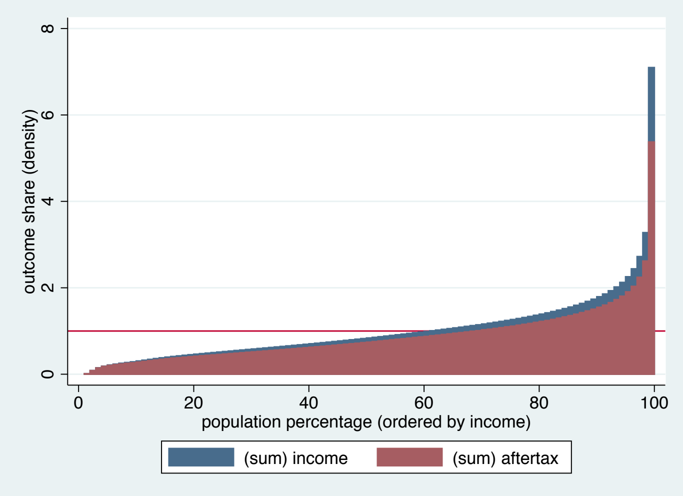


Pre-tax and post-tax income (2012)

```
. quietly pshare estimate income aftertax, n(100) nose density normalize(income)
>  pvar(income)
```

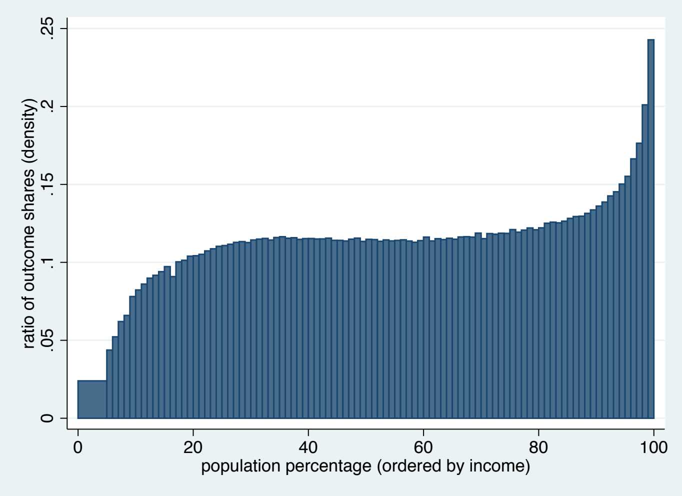
(*output omitted*)

```
. pshare histogram, yline(1) overlay finten(100) color(*.8)
```



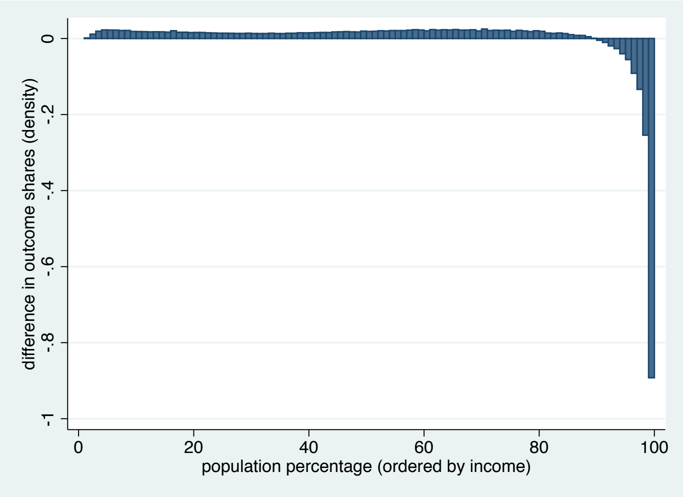
Tax rate by income percentiles (2012)

```
. quietly pshare estimate income tax, p(5(1)99) nose density ///
>     normalize(income) pvar(income)

. quietly pshare contrast income, ratio

. pshare histogram, base(0) ylabel(0(.05).25)
```



"Winners" and "losers" from taxation (2012)

```
. quietly pshare estimate income aftertax, n(100) nose density pvar(income)
. quietly pshare contrast income
. pshare histogram
```

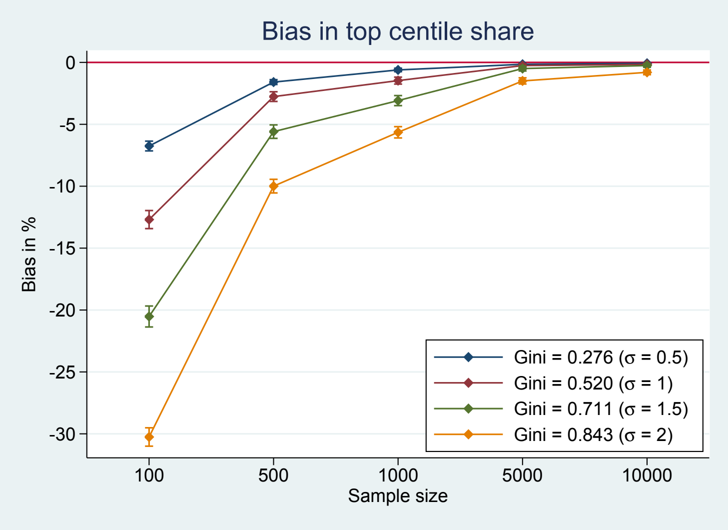


Small sample bias

- Percentile shares are affected by small sample bias.
- The top percentile share is typically underestimated.
- The problem is difficult to fix.
  - Corrections could be derived based on parametric assumptions.
  - Smoothing out the data by adding random noise can be an option, but this also requires parametric assumptions.
  - I evaluated a non-parametric small-sample correction using a bootstrap approach: the bias in bootstrap samples is used to derive correction factors for the main results.
  - This works very well in terms of removing bias (unless the distribution is extremely skewed).
  - **However:** MSE increases compared to uncorrected results!
  - Any ideas? Can Extreme Value Estimation be used to improve the estimates? Or would it be better to leave the point estimates as is and focus on obtaining bias-corrected CIs that have the correct size?

Small sample bias: how bad is the problem?

- Simulation: relative bias in top 1% share using a log-normal distribution



Software and paper

- Software:

```
. ssc install pshare
```

- Paper (forthcoming in the Stata Journal):
  - Jann, Ben. 2015. Assessing inequality using percentile shares. University of Bern Social Sciences Working Papers No. 13. https://ideas.repec.org/p/bss/wpaper/13.html

Lorenz curves

- Should you still be attached to Lorenz curves/concentration curves, I wrote a companion command with similar functionality:

```
. ssc install lorenz
```

- Paper:
  - Jann, Ben. 2016. Estimating Lorenz and concentration curves in Stata. University of Bern Social Sciences Working Papers No. 15. https://ideas.repec.org/p/bss/wpaper/15.html

References

- Ecoplan (2004). Verteilung des Wohlstands in der Schweiz. Bern: Eidgenössische Steuerverwaltung.
- Binder, D. A., M. S. Kovacevic (1995). Estimating Some Measures of Income Inequality from Survey Data: An Application of the Estimating Equations. Survey Methodology 21(2): 137-145.
- Kovacević, Milorad S., David A. Binder (1997). Variance Estimation for Measures of Income Inequality and Polarization – The Estimating Equations Approach. Journal of Offcial Statistics 13(1): 41-58.