Assessing inequality using percentile shares

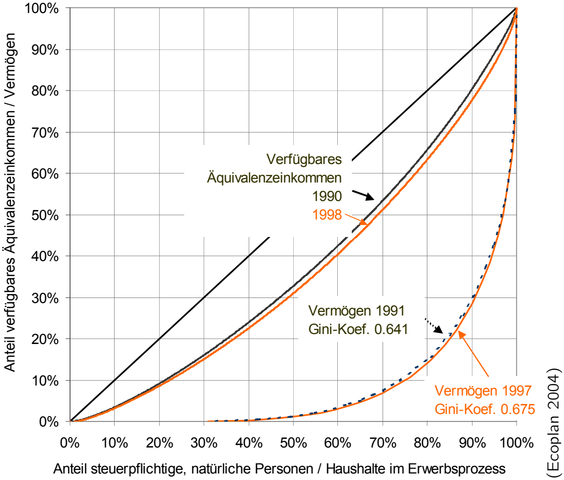
An application to Swiss tax data

Ben Jann

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

BIGSSS Lecture Series
University of Bremen, June 13, 2018


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

Definition 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:

$$L(p) = \frac{1}{\sum_{i=1}^{N} y_i} \sum_{i=1}^{N} y_i \mathbb{1}_{y_i \leq Q_p}$$

Percentile share S_k

$$S_k = L(p_k) - L(p_{k-1})$$

Proportion of total outcome within quantile interval $(Q_{p_{k-1}}, Q_{p_k}]$ with $p_{k-1} \leq p_k$.

Estimation

- Estimation given sample of size n:

$$\widehat{S}_k = \widehat{L}(p_k) - \widehat{L}(p_{k-1})$$

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

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

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

- 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

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

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

where w_i are sampling weights and

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

Standard errors

- For percentile shares, $\theta = S$ and $\lambda = \begin{bmatrix} Q_{p_1} \\ Q_{p_2} \end{bmatrix}$.
- The estimating functions are:

$$u_\theta = y_i \mathbb{1}_{y_i \leq Q_{p_2}} - y_i \mathbb{1}_{y_i \leq Q_{p_1}} - y_i S$$

$$u_\lambda = \begin{bmatrix} \mathbb{1}_{y_i \leq Q_{p_1}} - p_1 \\ \mathbb{1}_{y_i \leq Q_{p_2}} - p_2 \end{bmatrix}$$

- Hence:

$$u^* = \frac{y_i \mathbb{1}_{y_i \leq Q_{p_2}} - y_i \mathbb{1}_{y_i \leq Q_{p_1}} - y_i S - Q_{p_2}(\mathbb{1}_{y_i \leq Q_{p_2}} - p_2) + Q_{p_1}(\mathbb{1}_{y_i \leq Q_{p_1}} - p_1)}{\sum y_i}$$

$$= \frac{(y_i - Q_{p_2})\mathbb{1}_{y_i \leq Q_{p_2}} - (y_i - Q_{p_2})\mathbb{1}_{y_i \leq Q_{p_1}} + Q_{p_2} p_2 - Q_{p_1} p_1 - y_i S}{\sum y_i}$$

Percentile shares as densities, averages, or totals

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

$$D_k = \frac{S_k}{p_k - p_{k-1}} = \frac{L(p_k) - L(p_{k-1})}{p_k - p_{k-1}}$$

- Totals:

$$T_k = \sum_{i=1}^{N} y_i \, \mathbb{1}_{Q_{p_{k-1}} < y_i \leq Q_{p_k}} = S_k \cdot \sum_{i=1}^{N} y_i$$

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

$$A_k = \frac{T_k}{(p_k - p_{k-1}) \cdot N}$$

Contrasts and renormalization

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

$$S_k^A - S_k^B \qquad S_k^A / S_k^B \qquad \ln(S_k^A / S_k^B) \qquad \dots$$

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

$$L^*(p) = \frac{1}{T} \sum_{i=1}^{N} y_i \, \mathbb{1}_{y_i \leq Q_p}$$

$$S_k^* = L^*(p_k) - L^*(p_{k-1})$$

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

Concentration shares

- 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.)

$$L^Z(p) = \frac{1}{\sum_{i=1}^N y_i} \sum_{i=1}^N y_i \, \mathbb{1}_{z_i \leq Q_p^Z}$$

$$S_k^Z = L^Z(p_k) - L^Z(p_{k-1})$$

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

Contents

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

Contents

Quintile shares (the default)

- Distribution of hourly wages in the NLSW 1988 data:

```
. 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.

Bottom 50%, mid 40%, and top 10%

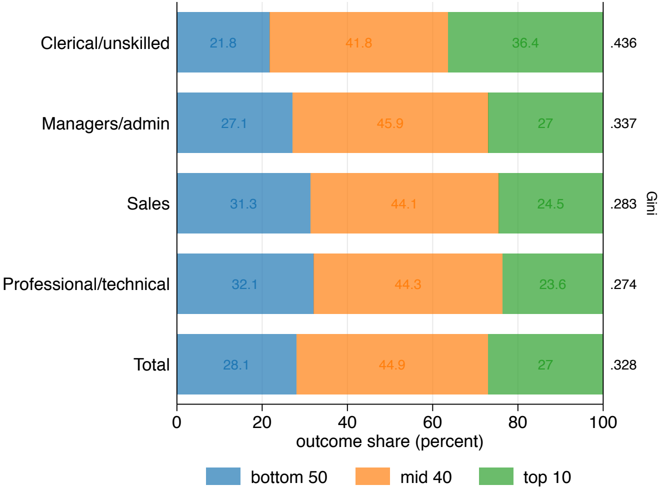
```
. pshare estimate wage, percent percentiles(50 90)
```

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

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) ///
>     legend(order(1 "bottom 50" 2 "mid 40" 3 "top 10") nostack)
```



Palma ratio

- By the way, you can also use `pshare` to compute summary inequality measures that are based on percentile shares, such as, e.g., the Palma ratio (ratio of top 10 to bottom 40).

```
. pshare estimate wage if occ<=4, percent p(40 90) over(occ) total
  (output omitted)
. nlcom (Clerical:     _b[4:90-100]     / _b[4:0-40])     ///
>       (Managers:     _b[2:90-100]     / _b[2:0-40])     ///
>       (Sales:        _b[3:90-100]     / _b[3:0-40])     ///
>       (Professional: _b[1:90-100]     / _b[1:0-40])     ///
>       (Total:        _b[total:90-100] / _b[total:0-40]) ///
>       , noheader
```

wage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Clerical	2.249477	.3525574	6.38	0.000	1.558477	2.940477
Managers	1.396125	.1167446	11.96	0.000	1.16731	1.62494
Sales	1.051786	.0823413	12.77	0.000	.8903995	1.213171
Professional	1.007814	.0911951	11.05	0.000	.8290749	1.186553
Total	1.316197	.0615441	21.39	0.000	1.195573	1.436821

Histogram of densities

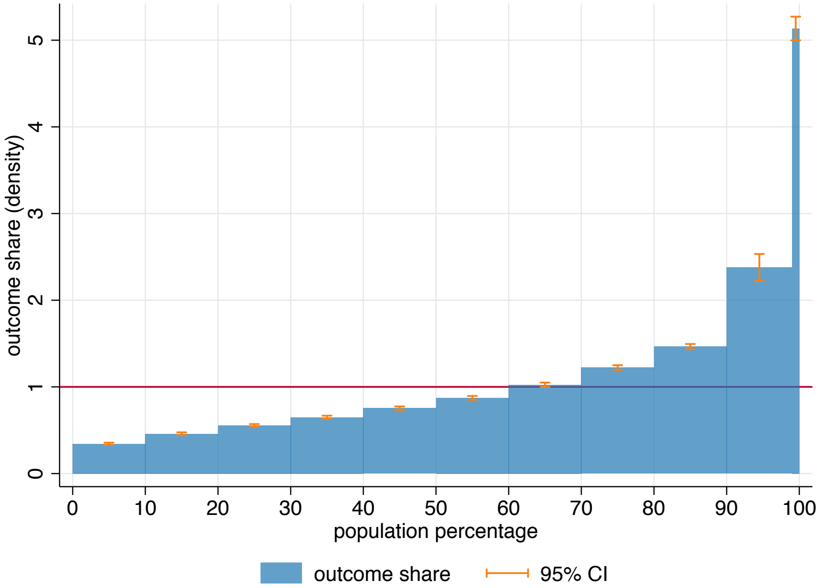
```
. pshare estimate wage, density percentiles(10(10)90 99)
```

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

wage	Coef.	Std. Err.	[95% Conf.	Interval]
0-10	.3426509	.0070215	.3288816	.3564202
10-20	.4591949	.0081384	.4432352	.4751546
20-30	.5544608	.0084268	.5379357	.5709858
30-40	.6491941	.009346	.6308663	.6675219
40-50	.7542334	.0102301	.7341719	.7742948
50-60	.8733366	.0113189	.85114	.8955333
60-70	1.024571	.0128412	.9993888	1.049752
70-80	1.223253	.0136742	1.196438	1.250069
80-90	1.465518	.0149372	1.436226	1.49481
90-99	2.377868	.0794248	2.222114	2.533622
99-100	5.135065	.0696951	4.998392	5.271739

```
. pshare histogram, yline(1) xlabel(0(10)100)
```

Histogram of densities



Histogram of densities: Interpretation

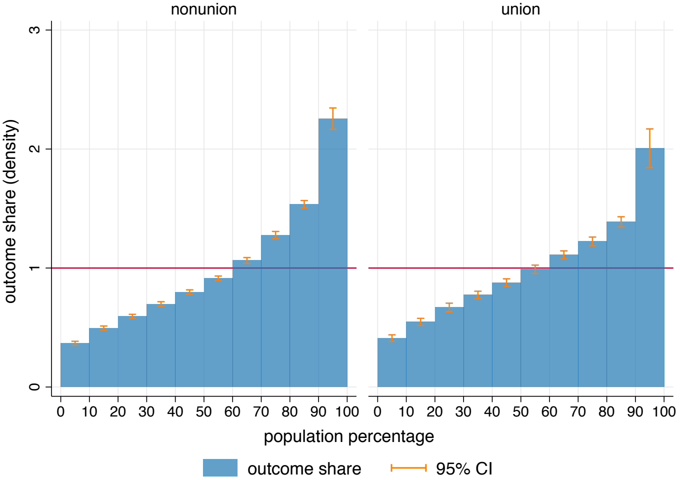
- Take 100 dollars and divide them among 100 people who line up along the X axis.
- The height of the bars shows you how much each one gets.
- If all get the same, then everyone would get one dollar (red line).
- However, according to the observed distribution, the rightmost person would get five of the 100 dollars, the next 9 would get about two and a half dollars each, ..., the bottom 10 only get 35 cents each.
- Stated differently, the top person gets 5 times the average, the bottom 10 only get about a third of the average.

Contrasts

- Distribution of wages among unionized and non-unionized workers:

```
. pshare estimate wage, n(10) density over(union) histogram(yline(1) xlabel(0(10)100))
```

(*output omitted*)



Contrasts

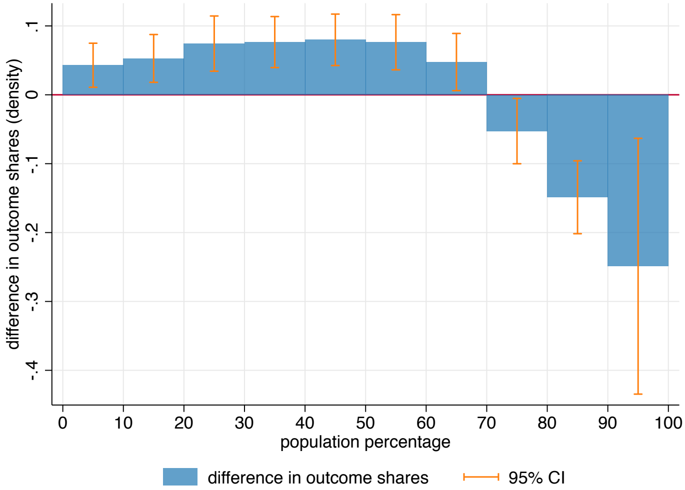
- How does the distribution differ between unionized and non-unionized workers?

```
. pshare contrast 0
Differences in percentile shares (density)      Number of obs     =      1,878
            0: union = nonunion
            1: union = union
```

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
1						
0-10	.0429197	.016305	2.63	0.009	.0109419	.0748975
10-20	.0528084	.0177041	2.98	0.003	.0180866	.0875301
20-30	.0743417	.0204516	3.64	0.000	.0342315	.1144519
30-40	.0765406	.018892	4.05	0.000	.0394891	.1135922
40-50	.0798209	.0190538	4.19	0.000	.0424521	.1171897
50-60	.0763097	.0204552	3.73	0.000	.0361924	.116427
60-70	.0475279	.0211824	2.24	0.025	.0059843	.0890715
70-80	-.0526677	.0242038	-2.18	0.030	-.1001369	-.0051984
80-90	-.1487654	.0269943	-5.51	0.000	-.2017074	-.0958234
90-100	-.2488358	.094742	-2.63	0.009	-.4346464	-.0630251

```
(contrasts with respect to union = 0)
. pshare histogram, yline(0) xlabel(0(10)100)
```

Contrasts



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

Contrasts

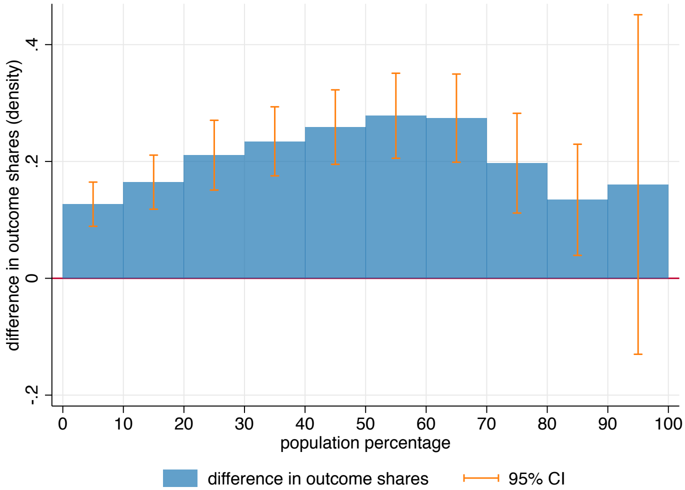
- How do results change if we take into account that unionized workers have higher wages on average than non-unionized workers?

```
. pshare estimate wage, n(10) density over(union) contrast(0) normalize(0:)
Differences in percentile shares (density)      Number of obs     =      1,878
            0: union = nonunion
            1: union = union
```

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
1						
0-10	.126848	.0193006	6.57	0.000	.0889951	.1647009
10-20	.1645575	.0236112	6.97	0.000	.1182506	.2108645
20-30	.2107276	.0304829	6.91	0.000	.1509437	.2705115
30-40	.2344474	.0301437	7.78	0.000	.1753287	.293566
40-50	.258802	.0325174	7.96	0.000	.1950281	.322576
50-60	.2782536	.0371205	7.50	0.000	.2054518	.3510553
60-70	.2741746	.0384939	7.12	0.000	.1986792	.34967
70-80	.1970798	.0435501	4.53	0.000	.111668	.2824915
80-90	.1343646	.0485359	2.77	0.006	.0391746	.2295547
90-100	.1605684	.1482708	1.08	0.279	-.1302246	.4513614

```
(shares normalized with respect to total for union = 0)
(contrasts with respect to union = 0)
. pshare histogram, yline(0) xlabel(0(10)100)
```

Contrasts



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

Contrasts

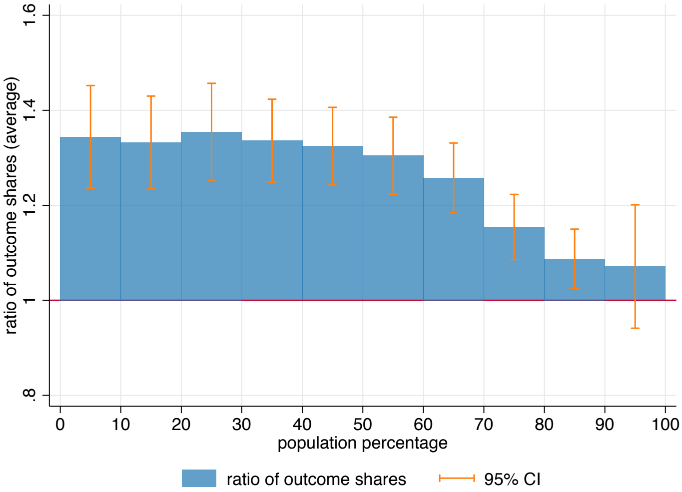
- How high are the benefits of unionization in relative terms at different positions in the distribution?

```
. pshare estimate wage, n(10) average over(union) contrast(0, ratio)
Ratios in percentile shares (average)          Number of obs     =      1,878
            0: union = nonunion
            1: union = union
```

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
1						
0-10	1.344201	.0550809	24.40	0.000	1.236174	1.452227
10-20	1.33242	.0497931	26.76	0.000	1.234764	1.430076
20-30	1.354596	.0521801	25.96	0.000	1.252259	1.456934
30-40	1.336087	.0445377	30.00	0.000	1.248739	1.423436
40-50	1.32447	.0417097	31.75	0.000	1.242668	1.406273
50-60	1.304536	.0412891	31.60	0.000	1.223559	1.385513
60-70	1.257784	.0374506	33.59	0.000	1.184335	1.331233
70-80	1.154327	.0350066	32.97	0.000	1.085671	1.222983
80-90	1.087433	.0318485	34.14	0.000	1.024971	1.149895
90-100	1.071177	.0662822	16.16	0.000	.9411828	1.201172

```
(contrasts with respect to union = 0)
. pshare histogram, yline(1) xlabel(0(10)100)
```

Contrasts



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

Concentration shares

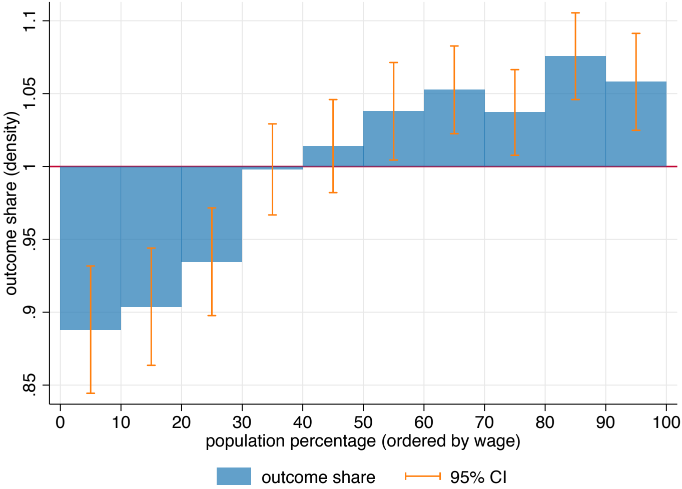
- How are working hours related to wages? Do people with high hourly wages work more, as economic theory would predict?

```
. 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) yline(1) xlabel(0(10)100)
```

Concentration shares



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

Contents

Application to "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 (natural) "tax units"
- see the project website for more information: http://inequalities.ch

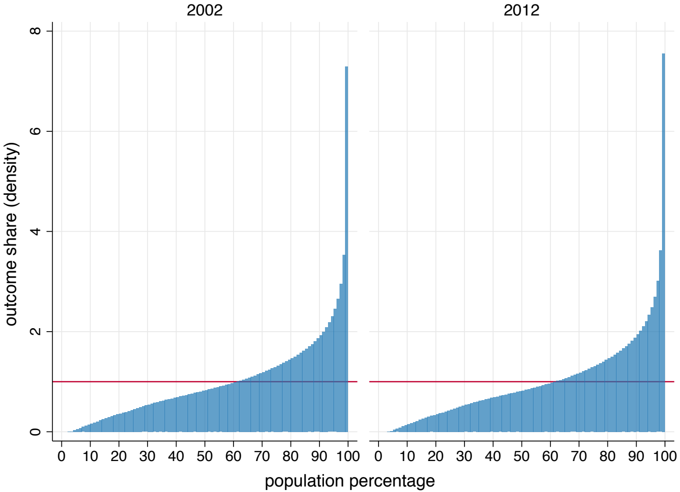
```
. describe

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) xlabels(0(10)100)
```

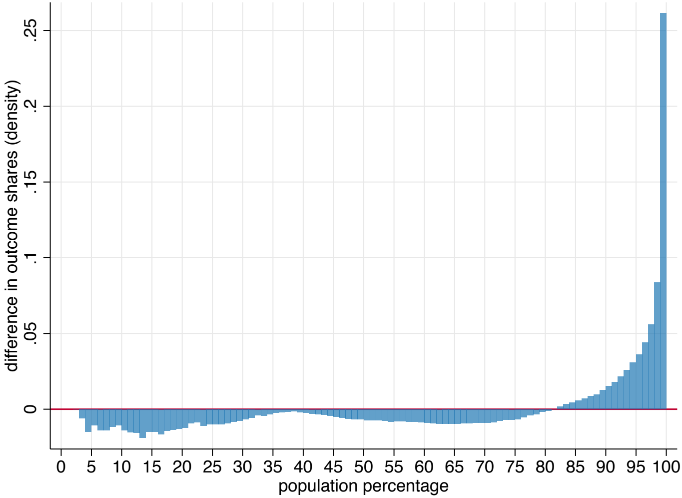


Change in income distribution from 2002 to 2012

```
. pshare contrast
```

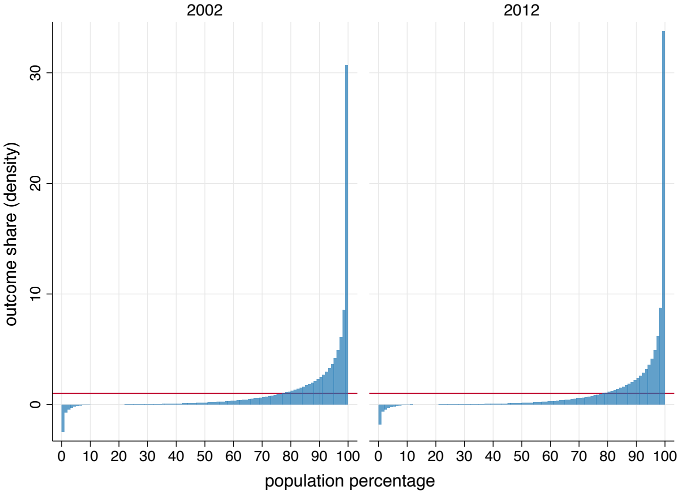
(*output omitted*)

```
. pshare histogram, yline(0) xlabels(0(5)100)
```



Distribution of net wealth in 2002 and 2012

```
. pshare estimate wealth, n(100) nose density over(year)
  (output omitted)
. pshare histogram, yline(1) xlabels(0(10)100)
```

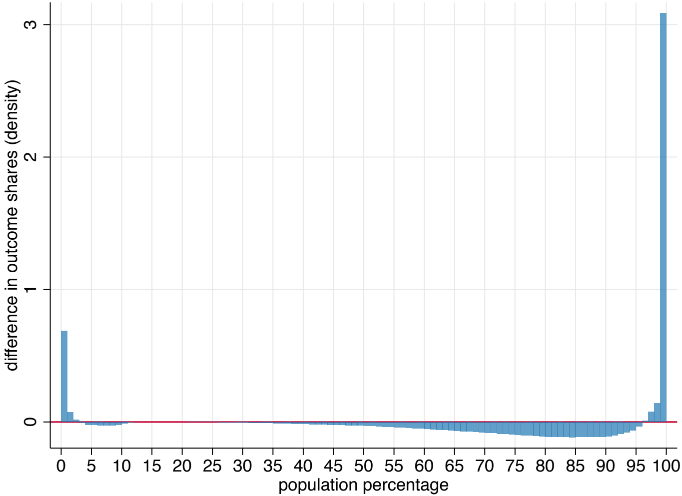


Change in wealth distribution from 2002 to 2012

```
. pshare contrast
```

(*output omitted*)

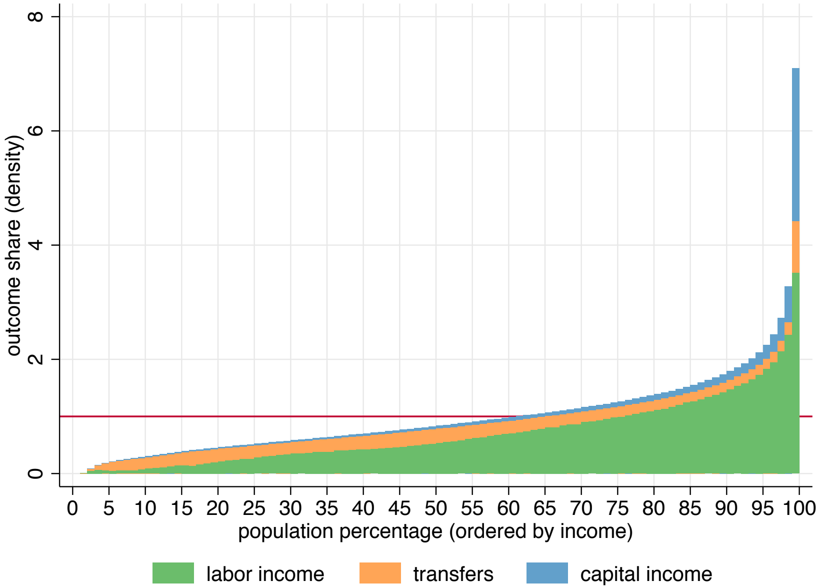
```
. pshare histogram, yline(0) xlabels(0(5)100)
```



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) xlabels(0(5)100) fcolor(%100) fintensity(70) ///
>     legend(order(3 "labor income" 2 "transfers" 1 "capital income") rows(1))
```

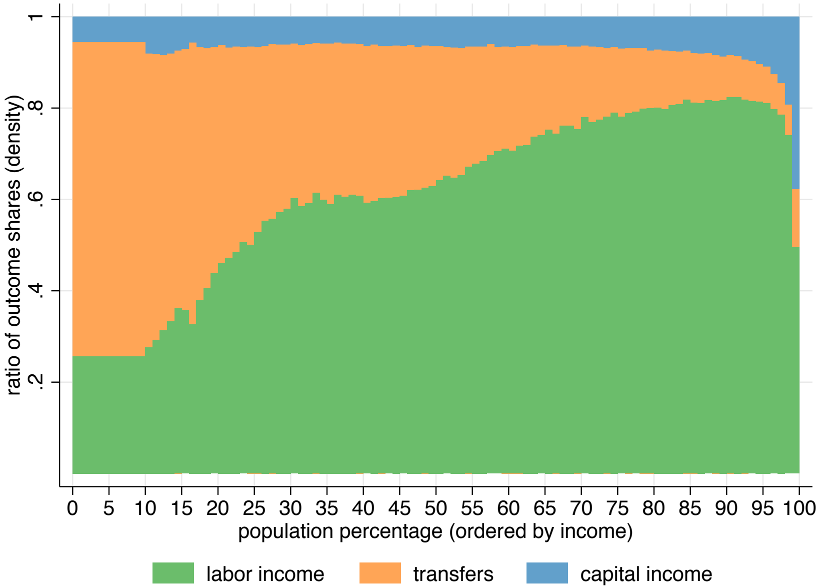
Income composition by income percentiles (2012)



Income composition in relative terms (2012)

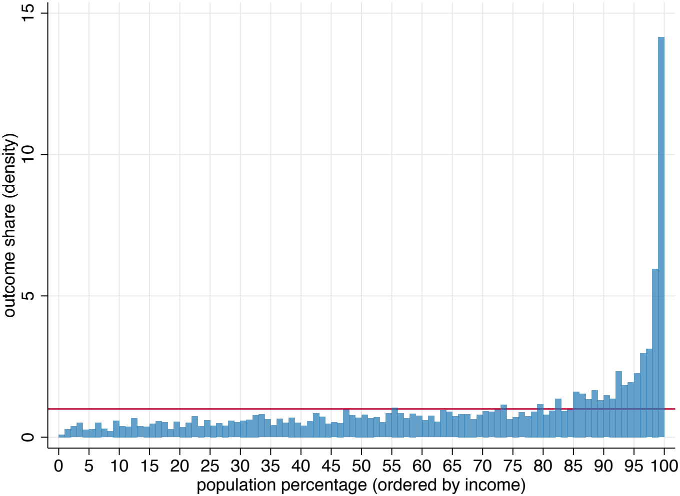
```
. generate earn_trans_cap = income
. quietly pshare estimate income earn_trans_cap earn_trans earnings, ///
>     p(10(1)99) nose density pvar(income) normalize(income)
. quietly pshare contrast income, ratio
. pshare histogram, overlay xlabels(0(5)100) fcolor(%100) fintensity(70) base(0) ///
>     legend(order(3 "labor income" 2 "transfers" 1 "capital income") rows(1))
```

Income composition in relative terms (2012)



Received bequests by income percentiles (2012)

```
. pshare estimate heritage, n(100) nose density pvar(income)
  (output omitted)
. pshare histogram, yline(1) xlabels(0(5)100)
```

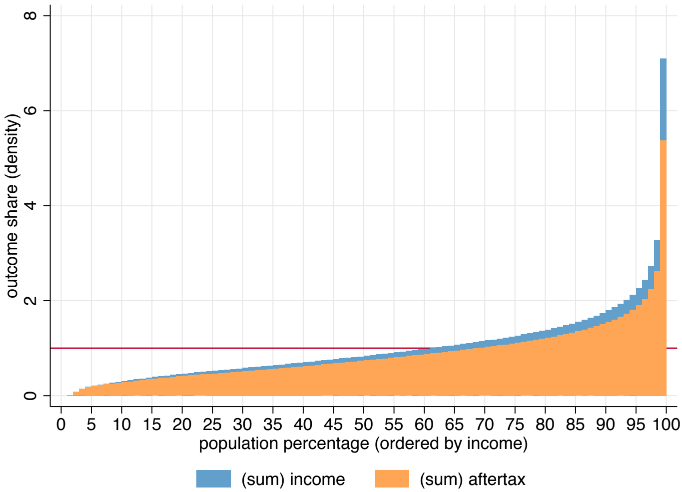


Pre-tax and post-tax income (2012)

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

(*output omitted*)

```
. pshare histogram, yline(1) overlay xlabels(0(5)100) fcolor(%100) fintensity(70)
```



Tax rate by income percentiles (2012)

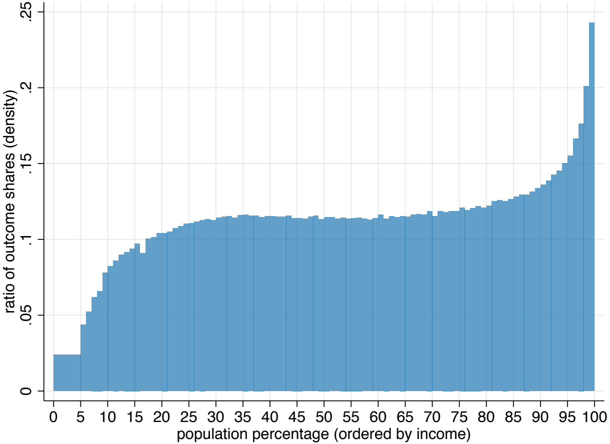
```
. pshare estimate income tax, p(5(1)99) nose density ///
>     normalize(income) pvar(income)
```

(*output omitted*)

```
. pshare contrast income, ratio
```

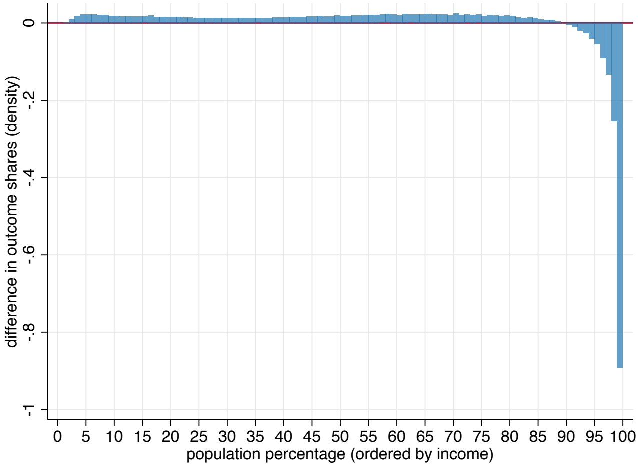
(*output omitted*)

```
. pshare histogram, base(0) ylabel(0(.05).25) xlabels(0(5)100)
```



“Winners” and “losers” from taxation (2012)

```
. pshare estimate income aftertax, n(100) nose density pvar(income)
```
(*output omitted*)
```
. pshare contrast income
```
(*output omitted*)
```
. pshare histogram, yline(0) xlabels(0(5)100)
```



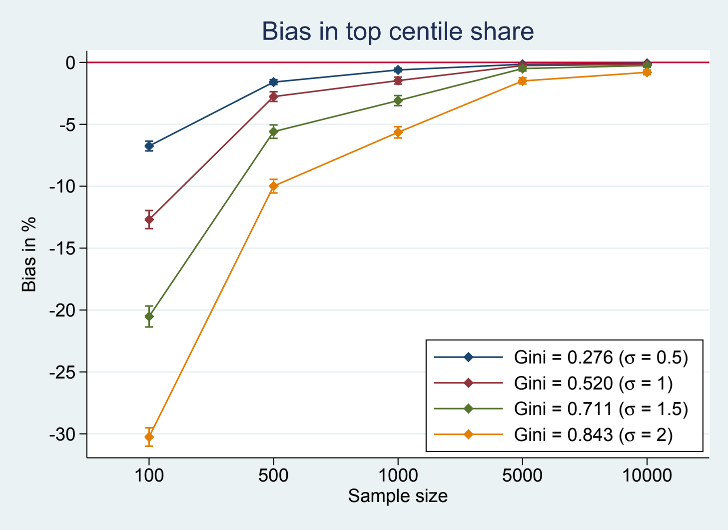
Contents

Small sample bias

- Percentile shares are affected by small sample bias (estimates of Lorenz ordinates have the same problem).
- 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!

Small sample bias: how bad is the problem?

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



Small sample bias: recommendations

- The simulation results suggest that for moderately skewed distributions (such as the income distribution with a typical Gini coefficient between around 0.3 and 0.5) 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.
  - For example, to estimate the top 0.1% share 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 a Gini coefficient as high as 0.8 or event 0.9) minimum sample size requirements may be considerably higher (such as 50 or even 100 observations in the top group).

Contents

Conclusions

- I my opinion, percentile shares are an excellent method to analyze – and visualize – income and wealth distributions.
- The pshare package in Stata (Jann 2016a) provides powerful tools to compute and graph percentile shares in various flavors and also allows comparing distributions between groups or analyzing relations between variables by means of concentration shares.
- To install pshare in Stata, type

```
. ssc install pshare
```

- Should you still be attached to classical Lorenz and concentration curves, there is a companion command with similar functionality called lorenz (Jann 2016b).

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.
- Jann, Ben (2016a). Assessing inequality using percentile shares. The Stata Journal 16(2): 264–300.
- Jann, Ben (2016b). Estimating Lorenz and concentration curves. The Stata Journal 16(4):837–866.
- 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.