

# A new Stata command for computing and graphing percentile shares

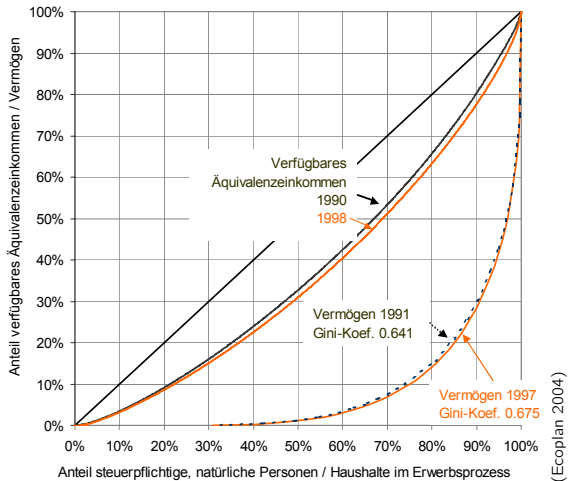
Ben Jann

University of Bern, [ben.jann@soz.unibe.ch](mailto:ben.jann@soz.unibe.ch)

2015 UK Stata Users Group meeting  
London, September 10–11, 2015

HAPPY  
BIRTHDAY





- ▷ [http://www.youtube.com/watch?v=slTF\\_XXoKAQ](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](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` command
- Examples using Bern tax data
- 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:

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

## Estimation of percentile shares

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

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

- Percentile share “density”:

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

- 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:

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

# Estimation of percentile shares

- Estimation given sample of size  $n$ :

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

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

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

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

- Standard errors
  - ▶ using estimating equations approach as proposed by Binder and Kovacevic (1995)
  - ▶ supports complex survey data

# The pshare 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



# Examples

- Tax data from canton of Bern, Switzerland, 2012
- individual level data from personal tax forms, 20% sample
- information on income components, deductions, assets, etc.

# Examples

```
. use taxdata
(Some tax data)

. describe
```

Contains data from taxdata.dta

```
obs:      119,939      Some tax data
vars:      3           27 Jun 2015 23:49
size:     1,079,451
```

---

variable name	storage type	display format	value label	variable label
agecat	byte	%9.0g	agecat	Age group
income	float	%9.0g		Total income
wealth	float	%9.0g		Net wealth

---

Sorted by:

```
. help pshare
```

## Quintile shares (the default)

```
. pshare estimate income
```

```
Percentile shares (proportion)          Number of obs      =      119,939
```

income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
0-20	.029269	.0002685	109.02	0.000	.0287428	.0297952
20-40	.1048592	.0004634	226.30	0.000	.103951	.1057674
40-60	.1645584	.0006001	274.24	0.000	.1633823	.1657345
60-80	.2365146	.0008311	284.59	0.000	.2348856	.2381435
80-100	.4647989	.0018814	247.05	0.000	.4611113	.4684864

Interpretation: The top 20% percent of the population get 46.5% of all income; the bottom 20% only get 2.9% of all income etc.

# Decile shares

```
. pshare estimate income, nquantiles(10)
```

```
Percentile shares (proportion)          Number of obs      =      119,939
```

income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
0-10	.0045258	.0000935	48.42	0.000	.0043426	.004709
10-20	.0247432	.0001932	128.07	0.000	.0243645	.0251219
20-30	.0435825	.0002279	191.23	0.000	.0431359	.0440292
30-40	.0612766	.0002527	242.45	0.000	.0607812	.061772
40-50	.0752199	.0002815	267.19	0.000	.0746681	.0757717
50-60	.0893385	.0003267	273.48	0.000	.0886982	.0899788
60-70	.1065221	.0003835	277.74	0.000	.1057704	.1072739
70-80	.1299924	.0004627	280.95	0.000	.1290855	.1308993
80-90	.1654318	.0005818	284.32	0.000	.1642914	.1665722
90-100	.2993671	.0023598	126.86	0.000	.2947419	.3039922

# Bottom 50%, Mid 40%, and Top 10%

```
. pshare estimate income wealth, percentiles(50 90)
```

```
Percentile shares (proportion)          Number of obs      =      119,939
```

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
<b>income</b>						
0-50	.2093481	.0008937	234.24	0.000	.2075964	.2110998
50-90	.4912848	.0016618	295.64	0.000	.4880278	.4945419
90-100	.2993671	.0023598	126.86	0.000	.2947419	.3039922
<b>wealth</b>						
0-50	-.0237426	.0010954	-21.67	0.000	-.0258896	-.0215956
50-90	.3042619	.0062104	48.99	0.000	.2920896	.3164343
90-100	.7194807	.0057992	124.07	0.000	.7081143	.730847

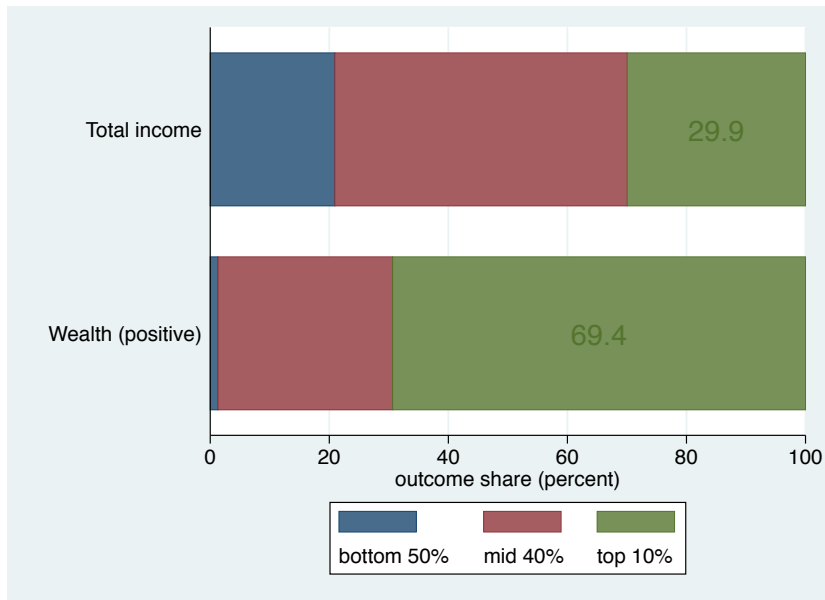
# Stacked bars plot

```
. generate wealth0 = cond(wealth<0, 0, wealth)
. label variable wealth0 "Wealth (positive)"
. pshare estimate income wealth0, percentiles(50 90) percent
Percentile shares (percent)                Number of obs      =      119,939
```

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
income						
0-50	20.93481	.0893717	234.24	0.000	20.75964	21.10998
50-90	49.12848	.1661772	295.64	0.000	48.80278	49.45419
90-100	29.93671	.2359796	126.86	0.000	29.47419	30.39922
wealth0						
0-50	1.314179	.029754	44.17	0.000	1.255862	1.372496
50-90	29.32997	.5773461	50.80	0.000	28.19838	30.46156
90-100	69.35585	.6038125	114.86	0.000	68.17239	70.53931

```
. pshare stack, plabels("bottom 50%" "mid 40%" "top 10%") ///
> values mlabsize(zero) p3(mlabsize(large))
```

# Stacked bars plot



# Histogram of densities

```
. pshare estimate wealth, p(1 5 10(10)90 95 99) density
```

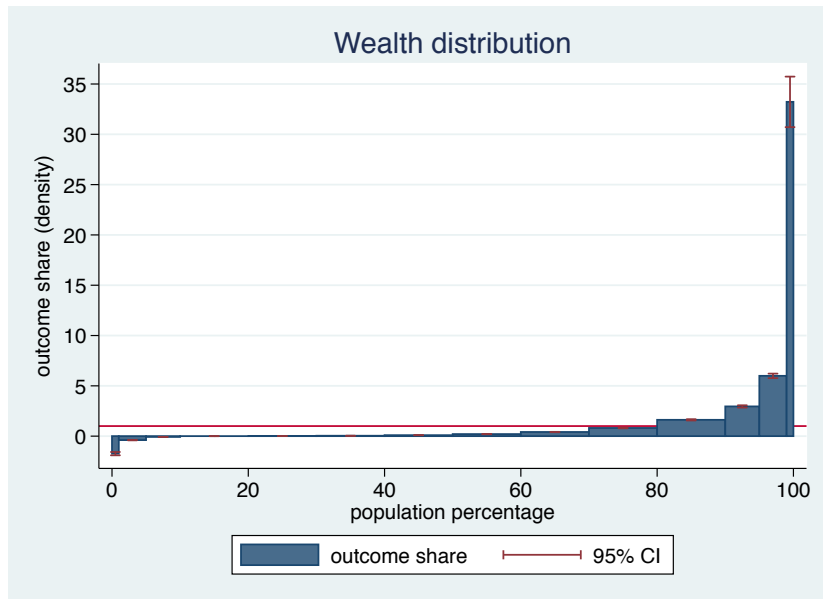
```
Percentile shares (density)          Number of obs      =      119,939
```

wealth	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
0-1	-1.740562	.0843581	-20.63	0.000	-1.905903	-1.575222
1-5	-.3963251	.0106253	-37.30	0.000	-.4171505	-.3754998
5-10	-.0789102	.0027628	-28.56	0.000	-.0843253	-.0734951
10-20	-.0017146	.0001861	-9.21	0.000	-.0020794	-.0013497
20-30	.0062028	.0002404	25.80	0.000	.0057316	.0066741
30-40	.0363781	.0008941	40.69	0.000	.0346257	.0381306
40-50	.0937488	.002118	44.26	0.000	.0895975	.0979
50-60	.2023709	.0044304	45.68	0.000	.1936873	.2110544
60-70	.4100251	.0087438	46.89	0.000	.3928874	.4271628
70-80	.811196	.0168879	48.03	0.000	.778096	.8442961
80-90	1.619027	.0328938	49.22	0.000	1.554556	1.683499
90-95	2.951972	.0589668	50.06	0.000	2.836398	3.067546
95-99	5.990715	.114097	52.51	0.000	5.767087	6.214344
99-100	33.22535	1.281602	25.92	0.000	30.71343	35.73727

```
. pshare histogram, yline(1) ylabel(0(5)35, angle(hor)) ti(Wealth distribution)
```



# Histogram of densities



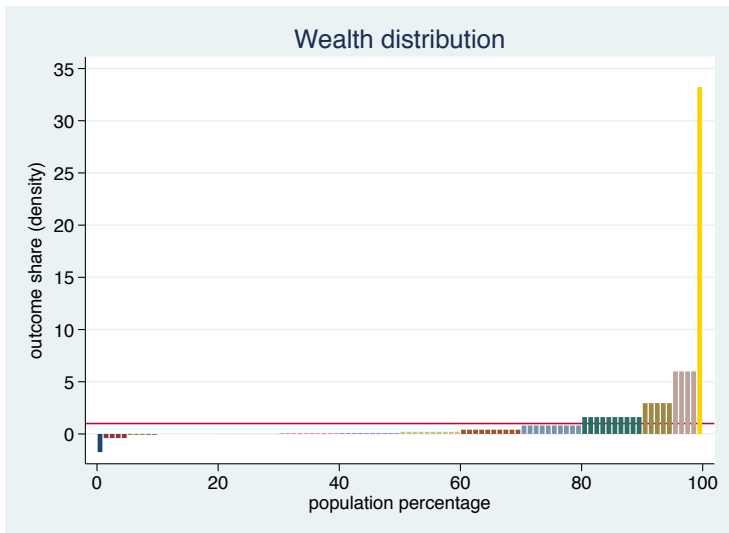
# Histogram of densities

- 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 (i.e. the richest) would get 33 (!) of the 100 dollars, the next four would get 6 dollar each, and so on.
- ▶ At the bottom, there are also some people you would have to take away some money (e.g., you would have to take away 1.74 dollars from the rightmost person).

# Using spikes and group-specific styles

```
. pshare hist, yline(1) ylabel(0(5)35, angle(hor)) ti(Wealth distribution) ///  
> spikes(100) lw(*3) psep legend(off)
```



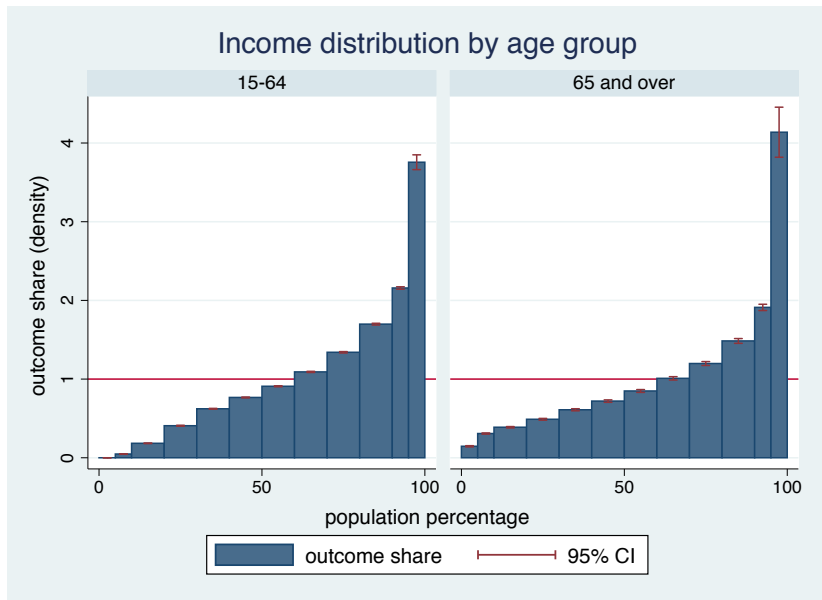
# Analysis of subpopulations

```
. pshare estimate income, p(5 10(10)90 95) over(agecat) density
Percentile shares (density)          Number of obs   =   119,939
    15: agecat = 15-64
    65: agecat = 65 and over
```

income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
15					
0-5	.0002108	.0000446	4.73	0.000	.0001234 .0002982
5-10	.0477019	.001519	31.40	0.000	.0447247 .050679
10-20	.1839598	.0019449	94.59	0.000	.1801479 .1877717
20-30	.4072382	.0030794	132.25	0.000	.4012026 .4132738
30-40	.62253	.0026311	236.60	0.000	.617373 .627687
40-50	.7665669	.0027358	280.20	0.000	.7612048 .7719291
50-60	.9087896	.0031534	288.20	0.000	.9026091 .9149701
60-70	1.091312	.0037406	291.75	0.000	1.08398 1.098643
70-80	1.340543	.0045022	297.75	0.000	1.331719 1.349367
80-90	1.698174	.0056207	302.13	0.000	1.687158 1.70919
90-95	2.157966	.0076957	280.41	0.000	2.142882 2.173049
95-100	3.755896	.0478808	78.44	0.000	3.66205 3.849741
65					
0-5	.1460041	.0039658	36.82	0.000	.1382311 .153777
5-10	.3081252	.0037889	81.32	0.000	.300699 .3155513
10-20	.3879465	.0043662	88.85	0.000	.3793888 .3965043
20-30	.4893443	.0056362	86.82	0.000	.4782975 .5003911
30-40	.6099115	.006742	90.46	0.000	.5966972 .6231257
40-50	.7204667	.0078164	92.17	0.000	.7051468 .7357867
50-60	.8488989	.0091605	92.67	0.000	.8309446 .8668533
60-70	1.009523	.0107486	93.92	0.000	.9884558 1.03059
70-80	1.19784	.0126163	94.94	0.000	1.173113 1.222568
80-90	1.484796	.0155048	95.76	0.000	1.454407 1.515185
90-95	1.911148	.0203231	94.04	0.000	1.871315 1.950981
95-100	4.137269	.1622893	25.49	0.000	3.819184 4.455353

```
. pshare histogram, yline(1) byopts(ti(Income distribution by age group))
```

# Analysis of subpopulations



# Subpopulation contrasts

```
. pshare contrast
```

```
Differences in percentile shares (density)      Number of obs      =      119,939
```

```
15: agecat = 15-64
```

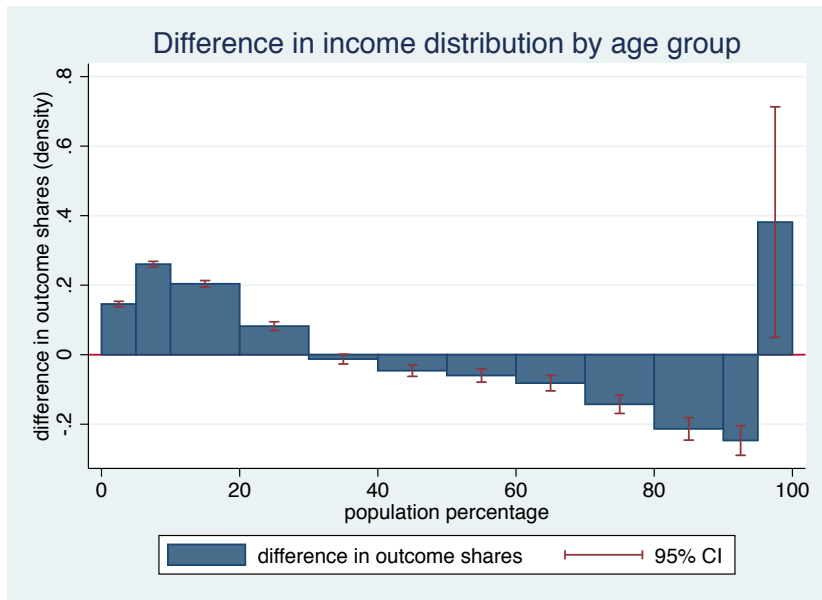
```
65: agecat = 65 and over
```

income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
65						
0-5	.1457933	.0039661	36.76	0.000	.1380198	.1535667
5-10	.2604233	.004082	63.80	0.000	.2524226	.268424
10-20	.2039867	.0047798	42.68	0.000	.1946184	.213355
20-30	.0821061	.0064225	12.78	0.000	.069518	.0946941
30-40	-.0126185	.0072372	-1.74	0.081	-.0268034	.0015664
40-50	-.0461002	.0082813	-5.57	0.000	-.0623314	-.029869
50-60	-.0598907	.009688	-6.18	0.000	-.078879	-.0409023
60-70	-.0817888	.0113809	-7.19	0.000	-.1040952	-.0594824
70-80	-.1427025	.0133956	-10.65	0.000	-.1689575	-.1164474
80-90	-.213378	.0164921	-12.94	0.000	-.2457023	-.1810537
90-95	-.2468179	.0217313	-11.36	0.000	-.2894109	-.2042248
95-100	.3813731	.1692052	2.25	0.024	.0497337	.7130125

```
(contrasts with respect to preceding subpopulation)
```

```
. pshare hist, yline(0) ti(Difference in income distribution by age group)
```

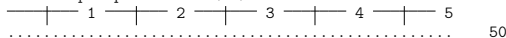
# Subpopulation contrasts



# Bivariate analysis: Wealth by income group

```
. pshare estimate wealth, p(10(10)90 95) pvar(income) density vce(boot)
(running pshare on estimation sample)
```

```
Bootstrap replications (50)
```



```
Percentile shares (density)          Number of obs    =   119,939
                                      Replications      =         50
```

wealth	Observed Coef.	Bootstrap Std. Err.	z	P> z	Normal-based [95% Conf. Interval]	
0-10	.094965	.0082328	11.54	0.000	.0788291	.1111009
10-20	.2175728	.0082347	26.42	0.000	.201433	.2337126
20-30	.506814	.0242173	20.93	0.000	.4593491	.554279
30-40	.5747378	.0162314	35.41	0.000	.5429247	.6065508
40-50	.5637113	.014835	38.00	0.000	.5346353	.5927874
50-60	.6399473	.0116937	54.73	0.000	.6170281	.6628664
60-70	.8189444	.0208025	39.37	0.000	.7781722	.8597166
70-80	1.029533	.0311481	33.05	0.000	.9684839	1.090582
80-90	1.213529	.0293399	41.36	0.000	1.156024	1.271034
90-95	1.75912	.0684403	25.70	0.000	1.624979	1.893261
95-100	6.921371	.2116263	32.71	0.000	6.506591	7.336151

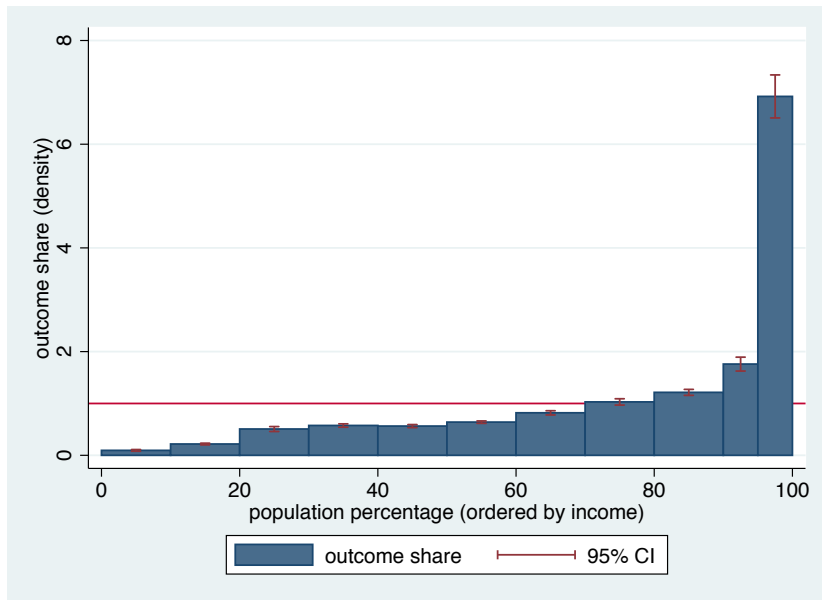
```
(percentile groups with respect to income)
```

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

The results show that the top income households are also the ones among which most of the wealth is accumulated.



## Bivariate analysis: Wealth by income group

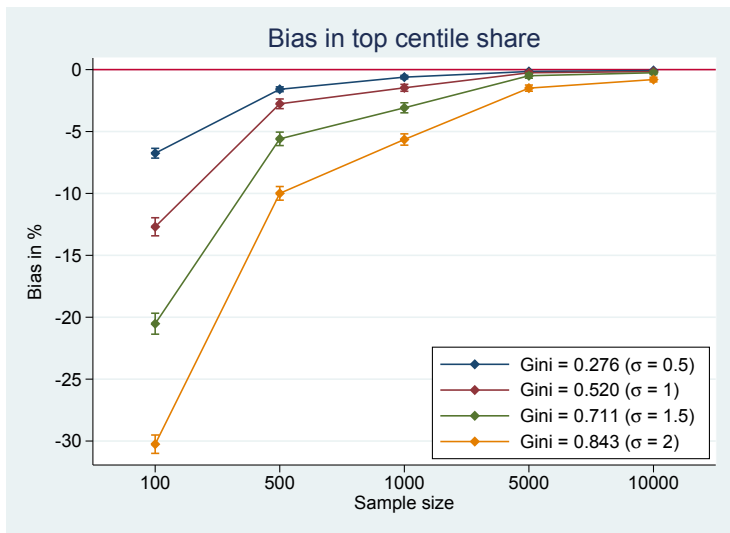


# Small sample bias

- 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!
  - ▶ No idea how to improve on this.

# 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  
. ssc install moremata  
. mata mata mlib index
```

- Paper:

- ▶ 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>

# 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.