Dealing with Extreme Values: Trimming and Bottom- / Top- coding

Goal

Many inequality measures are sensitive to the values at the bottom and/or top of the income distribution, and some are not defined for non-positive values of income (e.g., any measure that calculates a logarithm). Therefore, comparative researchers sometimes 'trim' the distribution (by deleting the top and bottom 1% for example) or impose 'bottom codes' and 'top codes' to provide a common calculation of lower and upper limits, method often referred to as 'winsorising'.

Activity

Use the data for Sweden 2005. Remove all missing and zero values of household disposable income. Using both the trimming and winsorising methods, create the following two new variables:

- variable ***trim***, where the top 1% and bottom 1% of weighted household disposable income (***dpi***) is set to missing (trimming);
- variable ***wins*** where the top 1% and bottom 1% of weighted household disposable income (***dpi***) are set respectively to the value of the 1st and 99th percentile (winsorising).

Compare the mean, median, and the first four and last four observations of the household income before the changes, after trimming, and after winsorising.

Guidelines

- You can find the values of the 1st and 99th percentiles of a variable as well as its first and last four observations by using `summarize`, with the option `detail`.
- In order to recall any of the results calculated by the `summarize` (or other) command(s), use the return codes automatically created by Stata. After you have summarized a variable, you can call `r(mean)` for the mean; `r(p50)` for the median; `r(p10)` for the first decile; `r(p20)` for the second decile, and so on. (In order to find out what statistics are available for each command, you can type `return list` immediately following the command, or look in the Stata manuals.) As a result, you can create a new variable based on the saved results of another variable used in a previous command in the following way:

```
sum <varname1>, de
gen <varname2> = <varname1> if <varname1> >= r(p1) & <varname1> <=r(p99)
```

- If you need information from a command, but do not need to see the results, precede your command by "`quietly`" (abbreviated by `qui`). This can save you from potentially lengthy log files that could be sent to the manual review queue by LISSY.
- You can ask Stata to format variables differently than what it does by default. Use the `format` command for the variables you wish to display differently, and the option `format` on any subsequent command using that same variable. For example, the following commands:

```
format <varlist> %8.0f
sum <varlist>, de format
```

will show the descriptive statistics with all their digits (up to the 8th), and no decimal point.

Program

```
di "** INCOME DISTRIBUTION II – Exercise 12 **"

use hweight dpi if (!mi(dpi) & !(dpi==0)) using $se05h, clear
quietly sum dpi [w=hweight], de
gen trim=dpi if dpi>=r(p1) & dpi<=r(p99)
gen wins=dpi
replace wins=r(p1) if dpi<=r(p1)
replace wins=r(p99) if dpi>=r(p99)
format dpi trim wins %8.0f
sum dpi trim wins [w=hweight], de format
```

Results

		Original values	After trimming	After winsorising
Number of valid observations		*16,268*	*15,918*	*16,268*
Average income		*269,551*	*262,484*	*265,713*
Median income		*223,861*	*223,861*	*223,861*
Income level of the first four observations (smallest incomes)	Smallest:	*-1,053,732*	*43,066*	*43,066*
	2nd smallest:	*-813,940*	*43,487*	*43,066*
	3rd smallest:	*-270,365*	*43,644*	*43,066*
	4th smallest:	*-239,543*	*43,671*	*43,066*
Income level of the last four observations (highest incomes)	4th largest:	*6,542,836*	*803,085*	*806,076*
	3rd largest:	*6,746,146*	*803,952*	*806,076*
	2nd largest:	*7,609,412*	*804,307*	*806,076*
	Largest:	*1,072,029,135*	*806,076*	*806,076*