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

- To find the values of the 1st and 99th percentiles of a variable **AS WELL AS** its first and last four observations you must use the **PROC UNIVARIATE**. The syntax of this procedure is similar to the syntax of the **PROC MEANS.** Few additional features has been added to **PROC UNIVARIATE** to produce an extensive set of statistics. By default, information on distribution such as the extreme value of analysed variables is provided.
- In order to recall any of the results calculated by the **PROC UNIVARIATE**, use the **CALL SYMPUT** routine as done previously. Note that with a **_NULL_** data step, the routine can be called several times:

```
DATA _NULL_;
 SET tmp;
    CALL SYMPUT("y",x);
    CALL SYMPUT("z",w);
RUN;
```

Program

```
OPTIONS NONOTES NOSOURCE NOFMTERR NODATE NONUMBER NOCENTER LABEL LS=max
PS=max ;

%MACRO univ ;
      PROC UNIVARIATE DATA=&pi ;
        WHERE &var not in (. 0);
          VAR &var       ;
        WEIGHT hweight ;
        &out ;
      RUN ;
%MEND univ ;

TITLE "*** AVERAGE & MEDIAN DPI BEFORE ***";
%LET pi  = &se05h ;
%LET var = dpi   ;
%LET out = OUTPUT OUT=tmp P1=fperc P99=lperc ;
%univ

DATA _NULL_;
 SET tmp;
    CALL SYMPUT("a",fperc);
    CALL SYMPUT("b",lperc);
RUN;
DATA tbc ;
 SET &se05h (KEEP=hweight dpi);
    IF (dpi not in (. 0)) ;
    trim = dpi ;
    wins = dpi ;
    botlin =&a;
    IF dpi lt botlin THEN DO;
       trim = . ;
       wins = botlin ;    END;
    toplin = &b;
    IF dpi gt toplin THEN DO;
           trim = . ;
       wins = toplin ;    END;
RUN ;
TITLE "*** AVERAGE & MEDIAN DPI AFTER TRIMMING ***";
%LET pi  = tbc ;
%LET var = trim   ;
%LET out = ;
%univ
TITLE "*** AVERAGE & MEDIAN DPI AFTER WINSORING ***";
%LET pi  = tbc  ;
%LET var = wins ;
%LET out = ;
%univ
```

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*