Running Descriptive Statistics: Sample and Population Values

Goal

This exercise is an introduction to a few of the variables in the household-level and person-level LIS data sets. The exercise concentrates on job syntax, basic descriptive statistics and the use of the weight.

Comparative researchers are typically interested in the characteristics of national populations, not the samples provided. It is very important to understand and use sample weights correctly in order to get representative results for the total underlying population. This exercise shows the differences in statistics between the unweighted sample and the weighted population.

Activity

For Luxembourg 2004 (LU04), create a household-level dataset containing: the household identifier (***casenum***), household weight (***hweight***), number of earners in the household (***d6***), number of children under 18 (***d27***), whether the head of the household is living in a couple (***married***), age of the household head (***d1***), gender of household head (***d3***), gender of the spouse of the household head (***sexsp***), and the household net disposable income (***dpi)***.

Find the unweighted and weighted number of observations, mean, median, minimum, and maximum for the continuous variables (including ***casenum*** and ***hweight***) and the unweighted and weighted frequencies of the categorical variables.

For the same country, use the person-level data to create a dataset containing: the household identifier (***casenum***); the person identifier (***ppnum***); person weight (***pweight***); age (***page***); gender (***psex***); marital status (***pmart***); relationship to the head of the household (***prel***); gross wages and salaries (***pgwage***); and gross wages per unit of time (***pgwtime***).

Find the unweighted and weighted number of observations, mean, median, minimum, and maximum for the continuous variables (including ***casenum***, ***ppnum***, ***pweight*** and ***page***) and the unweighted and weighted frequencies of the categorical variables.

Use the information from your output to answer the following questions:

1. Why can ***sexsp*** have a value of -1, but ***d3*** can never be -1?

   ____

   ____

   ____

   ____

2. Why do the values of ***pgwage*** and ***pgwtime*** differ (check the *Variables Definition List* and the *Lissification Table* for LU04 on line)?

   ____

   ____

____

____

____

Guidelines

- When you open a LIS dataset, use the correct alias for the country/year you wish to use. For example:

```
SET &lu04p;
```

For more information about the syntax of country/year macros, see the job submission instructions on the LIS web site (*Micro-Databases Access* → *Job Submission Instructions*). For a list of available data sets and their 2-digit country codes, go to:

Luxembourg Income Study (LIS) → *List of Datasets*.

- Only keep the variables you will be using:

```
(KEEP= casenum d6 d27 married d1 d3 dpi);
```

This avoids unnecessary burden on the machine so that submitted jobs will run faster.

If you need help determining which variables are categorical, go to the LIS web site and click on *Luxembourg Income Study (LIS)* → *Luxembourg* → *2004* (column: *Lissification Tables*; row: *Wave VI*). The "Value Labels" column of the *Lissification Table* delineates the values of categorical variables.

- For this dataset, the weight inflates to the total population in Luxembourg in 2004. This means you can find the population size by looking at the "Sum of Weight.". Information about sample size and weighted population estimates can be found at *Luxembourg Income Study (LIS)* → *<country>* → *Weighting Procedures.*

- SAS reminder: to run descriptive statistics, use these procedures:

```
PROC MEANS DATA=dataset <statistics>;
  VAR variablelist;
  WEIGHT yourweight;
RUN;
```

```
PROC FREQ DATA=dataset;
  TABLES variablelist / <OPTION>
  WEIGHT yourweight;
RUN;
```

  - By default, SAS generates cumulative frequencies. Add the `NOCUM` option to the "tables" statement. It suppresses display of cumulative frequencies and cumulative percentages in one-way frequency tables and in list format. Add also the `MISSPRINT` option to the "tables" statement in order to display missing value frequencies.

- SAS reminder

Descriptive and frequencies can be run by classification variable(s) adding the statement **BY** ***variable***. Prior to use it, the dataset must be sorted. To sort the dataset, apply the following procedure:

```
PROC SORT DATA=dataset;
  BY variable;
RUN;
```

- **IMPORTANT: Wait to get your results before sending a new job!**

Program

```
OPTIONS NOSOURCE NONOTES NOFMTERR NODATE NOCENTER LABEL NONUMBER LS=200
PS=MAX;

DATA lu4h;
SET &lu04h (KEEP=casenum d1 hweight dpi d6 married d27 d3 sexsp);
RUN;

TITLE "LU04H – Run unweighted descriptives – Continuous household variables";
PROC MEANS DATA=lu4h N MEAN MIN MAX MEDIAN;
 VAR casenum d1 hweight dpi;
RUN;

TITLE "LU04H – Run weighted descriptives – Continuous household variables";
PROC MEANS DATA=lu4h N MEAN MIN MAX MEDIAN SUMWGT;
 VAR casenum d1 hweight dpi;
 WEIGHT hweight ;
RUN;

TITLE "LU04H – Run unweighted descriptives – Categorical household
variables";
PROC FREQ DATA=lu4h;
 TABLES d6 married d27 d3 sexsp / MISSPRINT NOCUM;
RUN;

TITLE "LU04H – Run weighted descriptives – Categorical household variables";
PROC FREQ DATA=lu4h;
 TABLES d6 married d27 d3 sexsp / MISSPRINT NOCUM;
 WEIGHT hweight ;
RUN;

DATA lu4p;
SET &lu04p (KEEP=casenum ppnum pweight page pgwage pgwtime psex pmart prel);
RUN;

TITLE "LU04P – Run unweighted descriptives – Continuous person-level
variables";
PROC MEANS DATA=lu4p N MEAN MIN MAX MEDIAN SUMWGT;
 VAR casenum ppnum pweight page pgwage pgwtime;
RUN;

TITLE "LU04P – Run weighted descriptives – Continuous person-level
variables";
PROC MEANS DATA=lu4p N MEAN MIN MAX MEDIAN;
 VAR casenum ppnum pweight page pgwage pgwtime;
 WEIGHT pweight;
RUN;

TITLE "LU04P – Run unweighted descriptives – Categorical person-level
variables";
PROC FREQ DATA=lu4p;
 TABLES psex pmart prel / MISSPRINT NOCUM;
RUN;
```

```
TITLE "LU04P - Run weighted descriptives - Categorical person-level
variables";
PROC FREQ DATA=lu4p;
 TABLES psex pmart prel / MISSPRINT NOCUM;
 WEIGHT pweight;
RUN;

TITLE;
```

Results

Continuous household-level variables – unweighted results

	# of obs	Mean	Median	Minimum	Maximum
casenum	3,622	1,811.5	1,811.5	1	3,622
hweight	3,622	49.12	28	0.105	466.04
d1	3,622	48.95	48	18	100
dpi	3,622	56,750	48,598	-34,602	686,352

Continuous household-level variables – weighted results

	# of obs	Mean	Median	Minimum	Maximum
casenum	177,910	1,282.1	1,317	1	3,622
hweight	177,910	113.15	96.77	0.105	466.04
d1	177,910	51.04	49	18	100
dpi	177,910	55,371	47,373	-34,602	686,352

Categorical household-level variables

Variable name	Codes	Labels	# of obs in the sample	unweighted percent	weighted percent
d6	0		881	24.32	27.70
	1		1,444	39.87	37.92
	2		1,135	31.34	29.35
	3		130	3.59	3.92
	4		27	0.75	1.04
	5		4	0.11	0.02
	6		1	0.03	0.05
d27	0		2,291	63.25	68.96
	1		614	16.95	13.39
	2		493	13.61	12.52
	3		177	4.89	4.30
	4		31	0.86	0.42
	5		13	0.36	0.35
	6		2	0.06	0.01
	7		1	0.03	0.05
married	0	head not living in couple	1,166	32.19	35.68
	1	married couple	2,064	56.99	57.35
	3	non-married cohabiting couple	382	10.55	6.48
	5	non-married cohabiting couple, both partners same sex	10	0.28	0.49
d3	1	male	2,390	65.99	64.12
	2	female	1,232	34.01	35.88

sexsp	*-1*		*1,166*	*32.19*	*35.68*
	1	*male*	*550*	*15.18*	*14.75*
	2	*female*	*1,906*	*52.62*	*49.57*

Continuous individual-level variables – unweighted results

	# of obs	*Mean*	*Median*	*Minimum*	*Maximum*
casenum	*9,661*	*1,808.2*	*1,796*	*1*	*3,622*
ppnum	*9,661*	*3.11*	*2*	*1*	*63*
pweight	*9,661*	*46.27*	*24.08*	*0.105*	*466.04*
page	*9,661*	*35.47*	*35*	*0*	*100*
pgwage	*9,661*	*16,763*	*0*	*0*	*430,000*
pgwtime	*9,661*	*1,234.6*	*0*	*0*	*25,000*

Continuous individual-level variables – weighted results

	# of obs	*Mean*	*Median*	*Minimum*	*Maximum*
casenum	*447,006*	*1,250.4*	*1,276*	*1*	*3,622*
ppnum	*447,006*	*2.89*	*2*	*1*	*63*
pweight	*447,006*	*113.01*	*95.80*	*0.105*	*466.04*
page	*447,006*	*37.65*	*38*	*0*	*100*
Pgwage	*447,006*	*17,723*	*0*	*0*	*430,000*
Pgwtime	*447,006*	*1,314.2*	*0*	*0*	*25,000*