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.

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 file-naming for the country/year you wish to use. For example in this exercise :

  ```
  get file = lu04h.
  ```

  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*.

  The command get file will load the entire dataset in memory. Keeping only a subset of variables avoids unnecessary burden on the machine so that submitted jobs will run faster. Only keep the variables you will be using. The full command will be :

  ```
  get file = lu04h
       /keep = casenum d6 d27 married d1 d3 dpi.
  ```

- Each SPSS command should be ended by a delimiter, in case of SPSS this is the dot .

- When a command continues over several lines, INDENT each following line with a couple of spaces, and only put the dot at the end of the last line.

- To produce results like the mean, number of cases, minimum , maximum, you can use the command descriptives.

- SPSS reminder:

  - to run descriptives, use `descriptives variables = <variable list>.`

  Unfortunately, unlike the mean, the median will not be produced by the `descriptives` command. Therefore one needs to use the `frequencies` command, while specifying the option `statistics = median`. By default, the frequency command will not only produce the required median, but also the full frequency table. Please note that running frequencies on continuous variables (like wage, weights, identification numbers, etc.) will produce a huge frequency table!! This must be avoided, and can be done by adding the `format = notable` option in order to suppress the table. Your final frequencies command will look like this:

  ```
  frequencies variables = dpi
      / statistics = median
      / format = notable.
  ```

- 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.

- To get weighted results, you need to use the command `weight by` followed by the name of the weighting variable. Each result that follows after this command will give weighted results. The weights stay in effect until another weighting command is encountered. To return to unweighted results, use the command:

weight off.

- 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`** in the weighted summary. Information about sample size and weighted population estimates can be found at *Luxembourg Income Study (LIS)* → *<country>* → *Weighting Procedures*.
- **IMPORTANT: Wait to get your results before sending a new job!**

Program

```
title "** BASICS I - Exercise 2 **" .


get file = lu04h
    / keep = casenum hweight d6 d27 married d1 d3 sexsp dpi .
title " unweighted " .
descriptives variables =  casenum hweight d1 dpi .
* add each time lines for the median only .
frequencies variables =  casenum hweight d1 dpi
           / statistics = median
           / format = notable .
frequencies variables =  d6 d27 married d3 sexsp .
weight by hweight.
title " weighted " .
descriptives variables =  casenum hweight d1 dpi .
frequencies variables =  casenum hweight d1 dpi
           / statistics = median
           / format = notable .
frequencies variables =  d6 d27 married d3 sexsp .

get file = lu04p
    / keep = casenum ppnum pweight page psex pmart prel pgwage pgwtime .
title " unweighted " .
descriptives variables =  casenum ppnum pweight page pgwage pgwtime .
frequencies variables =  casenum ppnum pweight page pgwage pgwtime
           / statistics = median
           / format = notable .
frequencies variables =  psex pmart prel .
weight by pweight.
title " weighted " .
descriptives variables =  casenum ppnum pweight page pgwage pgwtime .
frequencies variables =  casenum ppnum pweight page pgwage pgwtime
           / statistics = median
           / format = notable .
frequencies variables =  psex pmart prel .
```

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*

Categorical individual -level variables

Variable name	*Codes*	*Labels*	*# of obs in the sample*	*unweighted percent*	*weighted percent*
psex	*1*	*male*	*4,808*	*49.77*	*49.57*
	2	*female*	*4,853*	*50.23*	*50.43*
pmart	*1*	*never married*	*4,349*	*45.02*	*41.81*
	2	*married*	*4,266*	*44.16*	*46.68*
	3	*separated*	*91*	*0.94*	*0.84*
	4	*widowed*	*425*	*4.40*	*5.58*
	5	*divorced*	*530*	*5.49*	*5.09*
prel	*1*	*head of household*	*3,622*	*37.49*	*39.80*
	2	*husband/wife*	*2,059*	*21.31*	*22.81*
	3	*partner*	*397*	*4.11*	*2.79*
	4	*own/adopted child*	*3,149*	*32.59*	*31.21*
	5	*step child (child of husband/wife)*	*62*	*0.64*	*0.50*
	6	*step child (child of partner)*	*47*	*0.49*	*0.20*
	7	*child in law*	*21*	*0.22*	*0.18*
	8	*foster child*	*15*	*0.16*	*0.24*
	9	*brother or sister*	*46*	*0.48*	*0.33*
	10	*sister/brother in law by marriage*	*9*	*0.09*	*0.08*
	11	*sister/brother in law by partnership*	*1*	*0.01*	*0*
	12	*mother or father*	*78*	*0.81*	*0.85*
	13	*parent-in-law by marriage*	*37*	*0.38*	*0.22*
	14	*parent-in-law by partnership*	*1*	*0.01*	*0.01*
	15	*grandchild*	*66*	*0.68*	*0.48*
	16	*great grandchild*	*1*	*0.01*	*0*
	17	*grandparent*	*2*	*0.02*	*0.01*
	21	*niece or nephew*	*15*	*0.16*	*0.08*
	23	*aunt or uncle*	*6*	*0.06*	*0.07*
	24	*aunt or uncle of spouse*	*1*	*0.01*	*0*
	25	*cousin*	*1*	*0.01*	*0.01*
	27	*other relative of head*	*1*	*0.01*	*0.01*
	29	*other not related person*	*24*	*0.25*	*0.15*

Answers to question 1-2:

1. ***sexsp** is -1 when the individual does not belong to the universe of those individuals (the subsample) who are asked for that information. In this case, only households with a couple present are asked about spouse's gender, so **sexsp** is always -1 for heads not living in a couple, and never -1 for those households with couples*. Since every household must have a head, ***d3*** must always have a value of 1 or 2.
2. *The* Variable Definition List *explains that, with the exception of **pgwtime** and **pnwtime**, all income variables are recorded in annual amounts (see cell H257). It also states that **pgwtime** contains gross wages for the unit of time (less than a year) that can be most accurately measured in the original data (cell H260). The* Lissification table *tells you that in Luxembourg, this measure is monthly gross income. By looking at the* Contents *(column G) of **pgwage** and **pgwtime**, you can see what constitutes gross income. In the case of LU04, the contents of **pgwage** include all gross income from dependent work, including wages, 13th and 14th month salaries, special or exceptional bonuses, wages from a secondary professional activity, and income from apprenticeships. The variable **pgwtime** includes the same information, but is adjusted by the data provider to account for the number of months worked (e.g., if 2 individuals have the same value for **pgwage**, the individual with the fewest months worked will have a higher value for **pgwtime**.)*

<u>Comments</u>

- File composition

  There are 9,661 observations of the identifier, ***casenum***, which gives us the total sample size (number of persons in this case).

  Without opening the household-level files, we can get the total number of households in the sample by looking at the number of household heads (***prel***=1). In Luxembourg 2004, there are 3,622 household heads. In many cases, you can also find the number of households by looking at the maximum value of ***casenum*** (which here is also 3,622). If these two values differ, then some of the original households have been removed from the main file and have been either included in the shadow file or dropped completely. (Go to *Luxembourg Income*

Study (LIS) → *LIS Policy on the Treatment of Missing Information* and *Luxembourg Income Study (LIS)* → *LIS Policy on the Treatment of Shadow Files* for a discussion about the LIS sample composition and shadow files.)

- Remember that the income variables are the nominal value of the national currency.
- Married variable

  As of Wave V, ***pmart*** is always coded as 2 if married. If more detailed marital information is given by the data provider, never married will be coded as 1 and other marital information (e.g., divorced, separated, widowed) are given codes above 2.

  Please be aware that when information about cohabiting status is not available for each person in the original dataset, a head with a cohabiting partner could be coded in ***pmart*** as single (if never civically married). See ***pparsta*** and ***prel*** for more information about cohabiting status.
- For this dataset, unweighted results on age (as well as other variables) are lower than the weighted ones. This means that younger individuals are over-represented in the sample. The sample (person) weight corrected for this by giving those individuals a lower weight. The unweighted result gives the average for the survey sample, not the Luxembourg average in 2004.
- Size of Population

  For this dataset, the weight also inflates to total population. This means you can find the population size by looking at the “Sum of Wgt.” in the weighted summary.

  - In some datasets, the average weight is equal to 1. In this case, the “Sum of Wgt.” is equal to the number of observations in the sample.
  - In other datasets, the weight “inflates to the population”, i.e. the weight for each unit in the sample is equal to the number of units he/she represents in the population (a “unit” could be a household or an individual). In other words, the average weight multiplied by the sample size gives the total population in the country.
- The person-level LIS weight is ***pweight***. In some cases ***pweight*** is given directly by the data provider and is the inverse of the probability of the individual being included in the sample. In cases where ***pweight*** is not provided (e.g., most household surveys), ***pweight*** for each member in the household is equivalent to the household weight, ***hweight***.