## **Assignment weeks 1 and 2**

## *To answer all the questions below, you must use Stata (and, specifically, DASP, if requested). Be concise and clear in your answers.*

## *The assignment is divided into three exercises (the points assigned to each exercise are indicated next to each exercise). Please answer (A) directly in this file after each question (Q) and please attach the \*.do file (do-file) that you generated. Rename both files as: “Assignment weeks 1-2 - Name, Surname”. Please submit this completed file and the \*.do through the virtual drop box (boîte de dépôt) in the course portal, no later than Tuesday, February 2 11:59 p.m. (*[*Québec time*](https://www.timeanddate.com/worldclock/converter.html?iso=20210203T045900&p1=189)*).*

## **Exercise 1 (4%)**

Assume that the population is of composed 12 households that live in regions *A, B and C*.

|  |  |  |  |
| --- | --- | --- | --- |
| *identifier* | *region* | *income* | *hhsize* |
| 1 | A | 210 | 4 |
| 2 | A | 450 | 6 |
| 3 | A | 300 | 5 |
| 4 | A | 210 | 3 |
| 5 | B | 560 | 2 |
| 6 | B | 400 | 4 |
| 7 | C | 140 | 4 |
| 8 | C | 250 | 2 |
| 9 | C | 340 | 2 |
| 10 | C | 220 | 2 |
| 11 | C | 360 | 3 |
| 12 | C | 338 | 3 |

**Q 1.1:** Using Stata, generate per capita income (*pcinc*).

**A: gen pcinc = income/hhsize**

**Q 1.2:** Using Stata, estimate the average per capita income and the total incomes of our population.

**A: Mean per capita income = 94.45; Total incomes = 3778**

**Q 1.3:** Assume that, the poverty line is equal to 100, generate the variable “per capita poverty gap (*pgap*)”, and then estimate its average (the per capita poverty gap should be normalized by the poverty line).

. sum pgap [aw=hhsize]

Variable | Obs Weight Mean Std. Dev. Min Max

-------------+-----------------------------------------------------------------

pgap | 12 40 .2225 .2368688 0 .65

**A: Average pgap = 0.2225**

**Q 1.4:** Redo question Q 1.3 using DASP.

**A: pcinc estimate = 0.2225**

----------------------------------------------------------------------------------------------

Variable | Estimate STE LB UB Pov. line

-------------+--------------------------------------------------------------------------------

pcinc | 0.222500 0.070494 0.067344 0.377656 100.00

----------------------------------------------------------------------------------------------

**Q 1.5:** Assume that the purchasing power in region B is higher than that of region A by 10% and that of region C is higher than that of region A by 30%. In the case where the region A is the region of reference, generate the variable (deflator) as a price deflator index, and then generate the variable real per capita income (r*pcinc*).

**A: Average rpcinc = 82.76049;**

**Q 1.6:** Redo the question 1.3 and 1.4 using the real per capita income when the poverty line is 120.

**A: Mean pgap with rpcinc = 0.370877**

**rpcinc estimate with GASP = 0.370877**

**Exercise 2 (3%)**

* 1. Using the file data\_1, estimate the average per adult equivalent expenditures without using the sampling weight and by using the DASP command **imean**. What does this statistic refer to?

**The imean command gives the coefficient of variation of the mean of the variable of interest (ae\_exp), and standard error, the lower and the upper bound of the estimate.**

**A:**  Variable | Estimate STE LB UB

------------------+----------------------------------------------------------------

1: mean\_ae\_exp | 42964.714844 1701.506958 39627.800781 46301.628906

**The estimate average ae\_exp is 42,964.71, with a total standard error of 1701.51. The true value of the means relies between 39627.80 and 46301.63, and such range contains the estimate average value.**

* 1. Assume different cases for initialising the sampling design.
* CASE1: Only by using the variable *strata* to initialise the stratification variable of the sampled population.
* -----------------------------------------------------------------------------------
* Variable | Estimate STE LB UB
* ------------------+----------------------------------------------------------------
* 1: mean\_ae\_exp | 42964.714844 1702.926636 39625.007813 46304.421875
* CASE2: Only by using the variable *psu* to initialise the primary sampling unit variable.

-----------------------------------------------------------------------------------

Variable | Estimate STE LB UB

------------------+----------------------------------------------------------------

1: mean\_ae\_exp | 42964.714844 1693.014282 39636.878906 46292.550781

* CASE3: By using the variable *strata* and *psu.*

-----------------------------------------------------------------------------------

Variable | Estimate STE LB UB

------------------+----------------------------------------------------------------

1: mean\_ae\_exp | 42964.714844 1699.352783 39624.207031 46305.222656

* CASE4: By using the variable *strata, psu* and the sampling weight variable*.*

-----------------------------------------------------------------------------------

Variable | Estimate STE LB UB

------------------+----------------------------------------------------------------

1: mean\_ae\_exp | 41993.101563 2213.284668 37642.332031 46343.875000

-----------------------------------------------------------------------------------

For each of these four cases, estimate the average per adult equivalent expenditures and give some explanation on the level of the standard errors compared to that of the question 1.1 and to those of the other cases.

**A: In estimation 1.1, since there is no survey specification (svyset) involved, the estimation assumes that the survey sample was obtained by a simple-random sampling procedure.**

**When using PSU only as survey specification, the results are a little more precise and the STE are lower than when using Strata only.**

**When adding the sampling weight at the survey specification (Case 4), the mean estimate of the aes\_per lowers, but the Standard Error increases**

* 1. Test whether the average per adult equivalent expenditures in region 1 is higher than the double of that of region 3. Briefly discuss the result.

**A: Since Pr(T>t)=0.9222, lower than 0.95 at alpha=0.05, we fail to reject Ho, and therefore, there is no evidence in the data that average ae\_exp in region 1 is higher than the double of that of region 3.**

diff = mean(ae\_expr1) - mean(ae\_expr3d) t = -1.4213

Ho: diff = 0 Satterthwaite's degrees of freedom = 691.723

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0

Pr(T < t) = 0.0778 Pr(|T| > |t|) = 0.1557 Pr(T > t) = 0.9222

* 1. Using the DASP command ***dimean*** test whether the average per adult equivalent expenditures for male household heads is higher than that of female households headed. Briefly discuss the result.

**A: With t = -1.9306 and Prob=0.0268 < alpha=0.05, we can not reject Ho, and therefore the data indicates that average ae\_exp for Males is higher than for females.**

**--------+--------------------------------------------------------------------**

**diff.| -6094.179 3156.551 -1.93064 0.0537 -infinity -899.7074**

**------------------------------------------------------------------------------**

**estimate(diff) = estimate (mean\_D2 - mean\_D1) t = -1.9306**

**Ho: estimate(diff) = 0 degrees of freedom = 1999**

**Ha: est.(diff) < 0 Ha: est.(diff) != 0 Ha: est.(diff) > 0**

**Pr(T < t) = 0.9732 Pr(|T| > |t|) = 0.0537 Pr(T > t) = 0.0268**

**------------------------------------------------------------------------------**

**Exercise 3 (5.5%)**

**Q 3.1** Use the data\_1.dta data file, and then compute the population size of the sampled households.

**A: 14694 by multiplying 2000 households by the average hhsize**

**Q 3.2** Rank the per capita expenditures in ascending order and then generate the variable population share (*ps*) that includes the proportion of the sampled population with corresponding per capita expenditures. Based on this, generate the variable percentiles (*p*) and quantiles (*q*).

**A: gen ps = hhsize/r(sum); gen p = sum(ps); gen q = pcexp**

**Q 3.3** Draw the cumulative distribution curve (X-Axis: the corresponding per capita expenditures and Y-Axis: the percentiles) (range of percentiles: min=0 and max=0.95).

**A: line p pcexp, title(The cumulative distribution curve) xtitle(The per capita expenditure (y)) ytitle(F(y));**

**Q 3.4** Plot the quantile curve (X-axis: percentiles and Y-axis: quantiles) (range of percentiles: min=0 and max=0.95), and briefly discuss the results.

**A: A high percentage of the population have ae\_expenditures below 2,000,000**

**Q 3.5** Using DASP, draw the quantile curve for each of the rural and urban regions (range of percentiles: min=0 and max=0.95), and briefly discuss the results.

**A: c\_quantile ae\_exp, hsize(hhsize) hgroup(zone) min(0.0) max(0.95)**

**The curves have a very similar behaviour between urban and rural areas.**



**Q 3.6** Using DASP, draw the density curves of the per capita expenditures by the sex of the household head (range of per capita expenditures: min=0 and max=1000000) and briefly discuss the results

**A: Both density curves behave very similar to each other. I would lower the max expenditure within the range to 200,000 to ensure normality and get a better shape.**

****