## **Exercices semaines 1 et 2**

## *Pour répondre à toutes les questions ci-dessous, vous devez utiliser Stata (et, spécifiquement, DASP, si demandé). Soyez concis(es) et clair(e)s dans vos réponses.*

## *L’examen est divisé en trois exercices (les points assignés à chaque exercice sont indiqués à côté de chaque exercice). Veuillez répondre (R) directement dans ce fichier après chaque question (Q) et veuillez joindre le fichier \*.do (do-file) que vous avez généré. Renommez ces deux fichiers en : "Exercice semaines 1-2 - Prénom, Nom" et veuillez les* soumettre *par la boîte de dépôt du portail de cours avant mardi le 2 février 23h59 (*[*heure du Québec*](https://www.timeanddate.com/worldclock/converter.html?iso=20210203T045900&p1=189)*).*

## **Exercice 1 (4%)**

Supposons que la population est composée de 12 ménages qui vivent dans les régions *A, B et 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:** À l’aide de Stata, générez le revenu par habitant (*pcinc*).

**R : \*\*Q1.1**

**gen pcinc=income/hhsize**

**Q 1.2:** À l'aide de Stata, estimez le revenu moyen par habitant et le revenu total de notre population.

**R : \* moyenne du revenu par tête**

**sum pcinc [aw=hhsize]**

On trouve**: 94.45**

**\*\* estimation du total de revenu**

**global Rtot r(sum)**

**display $Rtot**

Revenu total**: 3778**

**Q 1.3:** Supposons que le seuil de pauvreté soit égal à 100. Générez la variable « intensité de la pauvreté par habitant (*pgap*) », puis estimez sa moyenne (l’intensité de la pauvreté par habitant devrait être normalisée par le seuil de pauvreté).

**R :**

**\*intensité de la pauvreté par habitant**

**global pline 100**

**gen pgap=0**

**replace pgap=($pline-pcinc)/$pline if pcinc<$pline**

**\*\* estimation de sa moyenne**

**sum pgap [aw=hhsize]**

On trouve : 0.2225

**Q 1.4:** Refaites la question Q 1.3 avec DASP.

**R : ifgt pcinc, pline(100) alpha(1) hsize(hhsize)**

On trouve: 0.222500

**Q 1.5:** Supposons que le pouvoir d'achat dans la région B soit supérieur de 10% à celui de la région A et que celui de la région C soit supérieur de 30% à celui de la région A. Dans le cas où la région A est la région de référence, générez la variable (deflator) en tant qu'indice de déflation des prix, puis générez la variable de revenu réel par habitant (rpcinc).

**R :**

**\*\*déflateur**

**gen deflator=1 if region=="A"**

**replace deflator=0.9 if region=="B"**

**replace deflator=0.7 if region=="C"**

**\*\*revenu réel par habitant**

**gen rpcinc=pcinc/deflator**

**Q 1.6:** Refaites les questions 1.3 et 1.4 en utilisant le revenu réel par habitant lorsque le seuil de pauvreté est de 120.

**R :**

**\*Q1.3bis intensité de la pauvreté par habitant**

**global pline\_r 120**

**gen pgap\_r=0**

**replace pgap\_r=($pline\_r-rpcinc)/$pline\_r if rpcinc<$pline\_r**

**\*\* estimation de sa moyenne**

**sum pgap\_r [aw=hhsize]**

**On trouve : 0.2719907**

**\*\*Q1.4bis**

**ifgt rpcinc, pline(120) alpha(1) hsize(hhsize)**

On trouve**:0.271991**

**Exercice 2 (3%)**

* 1. À l'aide du fichier data\_1, estimez les dépenses moyennes par équivalent adulte sans utiliser le poids de sondage et en utilisant la commande DASP **imean**. À quoi réfère cette statistique?

**R :**

**use data\_1,clear**

**\*Q1.1**

**sum ae\_exp [aw=hhsize]**

repons**e : 42964.72**

**\*\* DASP**

**imean ae\_exp, hs(hhsize)**

On trouve**: 42964.714844**

**Estimation de l’erreur-Type :**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Estimate | STE | LB | UB |
|  |  |  |  |  |
| mean\_ae\_exp | 42964.714844 | 1701.506958 | 39627.800781 | 46301.628906 |

**Commentaire : Il s’agit de la moyenne de la dépense par équivalent adulte sur l'échantillon enquêté. En l’absence du poids de sondage, elle n’est pas extrapolée sur la population.**

* 1. Supposez différents cas d'initialisation du plan d'échantillonnage
* CAS1: Seulement en utilisant la variable *strata* pour initialiser la variable de stratification de la population échantillonnée.
* CAS2 : Seulement en utilisant la variable *psu* pour initialiser la variable d'unité primaire d’échantillonnage (primary sampling unit, PSU).
* CAS3: En utilisant la variable *strata* et *psu.*
* CAS4: En utilisant la variable *strata, psu* et la variable de poids de sondage*.*

Pour chacun de ces quatre cas, estimez les dépenses moyennes par équivalent adulte et donnez quelques explications sur le niveau des erreurs-types par rapport à celui de la question 1.1 et à ceux des autres cas.

**R :**

**\*CAS1**

**svyset \_n, strata(strata) vce(linearized) singleunit(missing)**

**gen numerator=hhsize\*ae\_exp**

**gen denominator=hhsize**

**svy: ratio numerator/denominator**

**Estimation erreurs-types**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Ratio | Std. Err. | [95% Conf. | Interval] |
| \_ratio\_1 | 42964.72 | 1702.927 | 39625.01 | 46304.42 |

**En menant une stratification uniquement, la précision de l’estimation de la dépense moyenne par équivalence adulte au niveau de l’échantillon est restée quasiment au même niveau que que Q1.1. Cela pose la question du bon choix de la variable de stratification (hétérogénéité entre les strates du point de vu de la variable ae\_exp).**

**\*\*CAS 2**

**svyset psu, vce(linearized) singleunit(missing)**

**svy: ratio numerator/denominator**

|  |  |  |  |
| --- | --- | --- | --- |
| Ratio | Std. Err. | [95% Conf. | Interval] |
| 42964.72 | 1693.014 | 39636.88 | 46292.55 |

**En introduisant juste les unités primaires, la précision de l’estimation s’est améliorée par rapport au Q1.1 par ce que l’on réduit les effets de grappes induits par les Unités primaires.**

**\*\*Cas 3**

**svyset psu, strata(strata) vce(linearized) singleunit(missing)**

**svy: ratio numerator/denominator**

|  |  |  |  |
| --- | --- | --- | --- |
| Ratio | Std. Err. | [95% Conf. | Interval] |
| 42964.72 | 1699.353 | 39624.21 | 46305.22 |

**Les erreurs types sont inférieures à celles de Q1.1 mais supérieures à la situation sans stratification ( PSU uniquement). La stratification n’a pas amélioré l’estimation non plus au niveau de cette estimation sur échantillon.**

**Cas4:**

**gen fweight=sweight\*hhsize**

**svyset psu, strata(strata) weight(fweight) vce(linearized) singleunit(missing)**

**svy: ratio numerator/denominator**

|  |  |  |  |
| --- | --- | --- | --- |
| Ratio | Std. Err. | [95% Conf. | Interval] |
|  |  |  |  |
| 41102.29 | 2500.363 | 36187.19 | 46017.38 |

**En prenant en compte les pondérations, l’estimation de la moyenne diffère de celle obtenue sur l’échantillon du fait des pondérations des unités primaires qui change le poids des ménages.**

**Par ailleurs, l’erreur d’échantillonnage est prise en compte dans l’extrapolation et on obtient une erreur-type plus grand.**

* 1. Vérifiez si les dépenses moyennes par équivalent adulte dans la région 1 sont supérieures au double de celles de la région 3. Discutez brièvement ce résultat.

**R : Nous travaillerons avec les pondérations étant donné que l’exercice n’a pas précisé :**

**sum ae\_exp [aw=fweight] if region==1**

**scalar m1=r(mean)**

**sum ae\_exp [aw=fweight] if region==3**

**scalar m3=r(mean)**

**scalar rapport13=m1/m3**

**display rapport13**

2.5979621

|  |  |
| --- | --- |
| region | mean(ae\_exp) |
|  |  |
| 1 | 59713.67 |
| 2 | 39196.69 |
| 3 | 22984.81 |
| 4 | 36195.89 |
|  |  |
| Total | 41993.1 |
|  |  |

Commentaires : La région 1 est celle qui a un niveau de dépense moyenne par équivalent adulte la plus élevé et celle de la région 3, la plus faible.

|  |  |
| --- | --- |
| region | mean(pcexp) |
| 1 | 131 560 |
| 2 | 142 221 |
| 3 | 160 286 |
| 4 | 144 723 |
| Total | 142 794 |

On note cependant qu’en considérant la moyenne de la dépense par tête, la région 3 est celle qui aurait la moyenne la plus élevé.

Cela souligne le rôle que peut jouer le choix des échelles d’équivalence dans les conclusions des études sur les niveaux de vie des populations.

* 1. À l'aide de la commande DASP ***dimean***, évaluez si les dépenses moyennes par équivalent adulte pour les chefs de famille hommes sont plus élevées que celles des femmes chefs de famille. Discutez brièvement ce résultat.

**R :**

**dimean pcexp pcexp, hsize1(hhsize) test(0) cond1(sex==2) hsize2(hhsize) cond2(sex==1) conf(ub)**



H0 : les dépenses moyennes par équivalent des ménages ayant un chef de famille Hommes sont plus élevées que celles des femmes chefs de familles équivaut a diff>0 :

Je ne peux rejeter l’hypothèse H0 par ce que l’erreur statistique commise serait de 97% donc je conclus que les ménages avec un homme à leur tête ont une dépense moyenne par équivalence adulte supérieur à ceux ayant une femme à leur tête.

### Exercice 3 (5.5%)

**Q 3.1** Utilisez le fichier de données data\_1.dta, puis calculez la taille de la population des ménages échantillonnés.

**R :**

**sum hhsize**

**global tpop r(sum)**

**displa $tpop**

on trouve : 14694

**Q 3.2** Ordonnez les dépenses par habitant en ordre croissant et générez ensuite la variable part de population (*ps*) qui comprend la proportion de la population échantillonnée avec les dépenses par habitant correspondantes. Sur cette base, générez les variables centiles (*p*) et quantiles (*q*).

**R :**

**sort pcexp**

**gen ps=hhsize/$tpop**

**gen p=sum(ps)**

**gen q=pcexp**

**Q 3.3** Dessinez la courbe de distribution cumulative (Axe X: les centiles et axe Y: les dépenses par habitant correspondantes) (domaine des centiles: min = 0 et max = 0,95).

**R :**

**Preserve**

**Keep if p<=0.95**

**line p q, title(courbe de distribution cumulative) xtitle(dépenses par habitant) ytitle(centiles (P)) ylab(0(.1) 0.95,ang(hor))**

**restore**

****

**Q 3.4** Tracez la courbe des quantiles (Axe X: centiles et axe Y: quantiles) (domaine des centiles: min = 0 et max = 0,95), et commentez brièvement les résultats.

**R :**

**preserve**

**Keep if p<=0.95**

**line q p , title(courbe des quantiles) xtitle(centiles (p)) ytitle(quantiles )**

**restore**

****

**Commentaire : La courbe indique une répartition assez inégalitaire des dépenses par tête. En effet p90(pcexp)= 255003.75 alors que p10(pcexp)=** **39172.714 soit un rapport 6,5.**

**Q 3.5** En utilisant DASP, dessinez la courbe des quantiles pour chacune des régions rurales et urbaines (domaine des centiles : min = 0 et max = 0,95), et discutez brièvement des résultats.

**R :**

**c\_quantile pcexp, hs(hhsize) hgroup(zone) min(0) max(0.95)**



**Commentaires** : Ces deux courbes indiquent une forte disparité entre les deux entités. On note aussi que le milieu urbain a un niveau de dépenses par tête plus élevés que le milieu rural (dominance complète). En effet, les quantiles du milieu urbain sont systématiquement supérieurs à ceux du milieu rural.

On notera également que le milieu tant à être un peu plus homogène au regard de l’évolution des quantiles en fonction des centiles.

**Q 3.6** À l'aide de DASP, dessinez les courbes de densité des dépenses par habitant en fonction du sexe du chef de ménage (domaine des dépenses par habitant: min = 0 et maximum = 1000000) et discuter brièvement des résultats.

**R :**

**cdensity pcexp, hsize(hhsize) hgroup(sex) popb(1) type(den) min(0) max(1000000) ytitle(Fréquences) xtitle(depenses par tête) title(Courbe de densité des dépenses par habitant en fonction du sexe du CM)**



Commentaires : **Les deux fonctions de densités indiquent :**

* **Une dissemblance des fonctions de densité indiquant que les deux populations ont des dépenses par habitant qui ne suivent pas les mêmes lois (hétérogénéité de la population) ;**
* **Les ménages à chef Homme ont en moyenne moins de dépenses par habitant que les femmes.**
* **Pour les deux sous-groupes il y a des ménages avec des hauts niveaux de revenus mêmes si ils sont peu nombreux et les ménages ayant les des femmes à leur tête sont relativement plus présents dans les catégories à forte dépense par tête.**