## **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 1-2-3 - Prénom, Nom" et veuillez les* soumettre *par la boîte de dépôt du portail de cours avant mardi le 5 février 11h59 a.m. (*[*heure du Québec*](https://www.timeanddate.com/worldclock/converter.html?iso=20190205T165900&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 :** Après avoir ajouté les données dans Stata pour que les variables existent, j’effectue la commande suivante pour générer le revenu par habitant (pcinc) :

*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 :** On estime le revenu moyen par habitant avec la commande suivante :

*sum pcinc [aw = hhsize]*

J’obtiens une estimation du revenu par habitant (pcinc) de la population qui vaut en moyenne 94,45. En d’autres mots, le revenu moyen par habitant de toute la population vaut 94,45 selon l’estimation. Il est important de noter que l’échantillon est pondéré par les poids des ménages, soit *hhsize* afin que les résultats puissent s’appliquer à la population (inférence).

J’estime le revenu total de la population avec la commande suivante :

*egen sum\_inc = sum(income)*

J’obtiens une estimation du revenu total de la population qui vaut 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 :** D’abord,je génèrela variable *pline* qui représente le seuil de pauvreté. Ensuite, je génère la variable *pgap* qui représente l’intensité de la pauvreté par habitant. Cette intensité est normalisée par le seuil de pauvreté puisqu’on divise la différence entre le revenu par habitant et le seuil de pauvreté par le seuil de pauvreté pour les personnes pauvres seulement. Pour les personnes non-pauvres, l’intensité de la pauvreté vaut zéro. Finalement, j’estime la moyenne de l’intensité de la pauvreté par habitant au sein de la population avec la commande *sum pgap [aw = hhsize],* ce qui donne comme résultat 0,2225. (Pour le code complet, voir le do-file).

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

**R :** Avec DASP, j’effectue la commande suivantepour estimer la moyenne de l’intensité de la pauvreté par habitant au sein de la population. J’obtiens une moyenne de 0,2225 comme lors de la question 1.3.

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

**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 :** D’abord, je génère la variable *deflato*r qui vaut 1 pour toutes les régions : *gen deflator = 1*

Ensuite, je remplace pour la région B (exprimée par un 2 dans les données) la valeur de la variable *deflator* par 0.9 puisque le pouvoir d’achat est supérieur de 10% dans cette région. *replace deflator = 0.9 if region == 2*

De la même manière, je remplace *deflator* par 0.7 pour la région C (valant 3) parce que le pouvoir d’achat est supérieur de 30% dans cette région :

*replace deflator = 0.9 if region == 2*

Ainsi, pour la région A, qui est la région de référence, le *deflator* vaut 1. Finalement, je génère le revenu réel par habitant par la commande suivante :

*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 :** D’abord, j’estime le revenu réel moyen par habitant de la population : *sum rpcinc [aw = hhsize]*

Ensuite, je remplace le seuil de pauvreté par 120 (*replace pline = 120)* et je remplace l’intensité de la pauvreté par habitant en utilisant le nouveau seuil de pauvreté et en utilisant le revenu réel par habitant (*rpcinc*) au lien du revenu par habitant (*pcinc*) :

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

Ensuite, j’estime la nouvelle moyenne de l’intensité de la pauvreté par habitant :

*sum pgap [aw = hhsize]*

J’obtiens comme valeur 0,272 environ, soit une valeur légèrement plus élevée qu’en utilisant le revenu par habitant et le seuil de pauvreté d’une valeur de 100.

Finalement, j’estime la moyenne de l’intensité de la pauvreté par habitant (avec le nouveau seuil de pauvreté et le revenu réel par habitant) à l’aide de DASP :

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

J’obtiens la même valeur que précédemment pour la moyenne, soit 0,272 environ.

**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 :** J’effectue la commande *imean ae\_exp, hsize(hhsize)* afin d’estimer les dépenses moyennes par équivalent adulte. À l’aide de la boîte de dialogue de la commande *imean* de DASP, que j’accède en tapant *db imean*, je m’assure qu’il n’y a pas de poids de sondage (weight = none) en vérifiant les données d’échantillonnage par les *survey settings.*

J’obtiens comme valeur estimée 42964,71 approximativement, avec comme erreur-type (standard error) 1701,51.

Cette statistique est un indice d’inégalité, soir le coefficient de variation (coefficient of variation index). Plus la valeur est élevée, plus les valeurs des dépenses par équivalent adulte sont dispersées autour de la moyenne et donc plus l’inégalité des dépenses par équivalent adulte est forte dans 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 : CAS 1** : J’effectue d’abord la commande *svyset \_n, strata(strata)* qui utilise la variable *strata* pour initialiser la variable de stratification de la population échantillonnée. Ensuite, j’estime les dépenses moyennes par équivalent adulte avec la commande *imean ae\_exp, hsize(hhsize)*. J’obtiens comme résultat une moyenne de 42964,71 et un niveau des erreurs-type (STE) de 1702,93. Je remarque que la moyenne est la même qu’à la question 1.1, mais l’erreur-type est légèrement supérieure. Je suppose que le nombre de strates n’est pas vraiment élevé. En effet, le regroupement de strates diminue généralement la précision des estimations, ce qui pourrait expliquer la légère hausse de l’erreur-type. En tapant la commande *svydes*, je peux voir qu’il y a 10 strates desquelles proviennent les observations. Je ne sais pas comment est calculé l’estimation lorsqu’aucune strate n’est spécifiée, mais il semble qu’il n’y a pas vraiment de différence importante pour ce cas.

**CAS 2** : J’effectue d’abord la commande *svyset psu* qui utilise la variable *psu* pour initialiser la variable d’unité primaire d’échantillonnage. Je vérifie les statistiques de l’échantillon en tapant la commande *svydes*. Je peux voir qu’il y a maintenant une seule strate et 412 unités (PSU). Il y a maintenant en moyenne 4.8 observations par PSU, tandis que dans l’exercice 1.1, il n’y avait qu’une observation par PSU. Ensuite, j’estime les dépenses moyennes par équivalent adulte avec la commande *imean ae\_exp, hsize(hhsize)*. J’obtiens comme résultat une moyenne de 42964,71 et un niveau des erreurs-type (STE) de 1693,01. Je remarque que la moyenne est la même qu’à la question 1.1, mais l’erreur-type a diminué. Voici la différence entre les méthodes d’échantillonnage : à la question 1.1, il y a 1 strate et l’on prend 1 observation par PSU dans 2000 PSU, tandis que pour ce cas, il y a toujours 1 strate, mais on prend en moyenne 4.8 observations par PSU dans 421 PSU. Nous avons ainsi diminué le nombre de PSU sondés, mais toujours en obtenant 2000 observations. Ceci a pour effet d’augmenter la précision de notre estimation.

**CAS 3** : J’effectue d’abord la commande *svyset psu, strata(strata)* qui utilise la variable *psu* pour initialiser la variable d’unité primaire d’échantillonnage et la variable *strata* pour initialiser la variable de stratification. Je vérifie les statistiques de l’échantillon en tapant la commande *svydes*. Je peux voir qu’il y a maintenant 10 strates et 412 unités (PSU) et qu’il y a toujours en moyenne 4.8 observations par PSU. Ensuite, j’estime les dépenses moyennes par équivalent adulte avec la commande *imean ae\_exp, hsize(hhsize)*. J’obtiens comme résultat une moyenne de 42964,71 et un niveau des erreurs-type (STE) de 1699,35. Je remarque que la moyenne est la même qu’à la question 1.1, mais l’erreur-type a diminué, toutefois, pas autant que pour le cas 2. Ainsi, les changements dans les statistiques de description de l’échantillonnage occasionnent des changements dans l’erreur-type de l’estimation des dépenses moyennes par équivalent-adulte, mais elles n’ont pas d’influence sur la moyenne, ce qui est normal. Ces statistiques descriptives de l’échantillonnage n’ont d’impact que sur la précision de l’estimation, pas sur la moyenne. Je remarque également que lorsque les strates sont spécifiées, l’erreur-type augmente légèrement. Ainsi, il semble qu’avoir 1 strate est préférable à 10 strates. Ceci doit provenir du fait que l’erreur d’estimation se répète pour chacune des strates sondées. Alors, plus de strates signifie une erreur-type légèrement plus élevée. Également, les PSU augmentent la précision de l‘estimation.

**CAS 4** : J’effectue d’abord la commande *svyset psu, strata(strata) weight(sweight)* qui initialise le plan d’échantillonnage avec les variables *strata* et *psu* et qui utilise le poids de sondage *sweight*. Je vérifie les statistiques de l’échantillon en tapant la commande *svydes*. Il y a toujours 10 states et 421 PSU, avec en moyenne 4.8 observations par PSU. Toutefois, je peux voir que maintenant le poids de sondage *sweight* est utilisé. Ensuite, j’estime les dépenses moyennes par équivalent adulte avec la commande *imean ae\_exp, hsize(hhsize)*. J’obtiens comme résultat une moyenne de 41993,10 et un niveau des erreurs-type (STE) de 2213,28. Je remarque que la moyenne de l’estimation a changé, alors que pour les 3 autres cas elle est restée la même. Ceci vient évidemment du fait que l’on utilise maintenant le poids de sondage, qui accorde un certain poids plus important à certaines observations et un autre poids plus faible à d’autres observations afin de représenter correctement (ou le plus fidèlement possible) la population. Ainsi, la nouvelle moyenne est probablement plus près de la vraie valeur de dépenses moyennes par équivalent-adulte. Également, l’erreur-type (STE) a changé et est maintenant plus élevé, soit 2213,28. Ceci provient du fait que nous utilisons le poids de sondage. Nécessairement, l’erreur-type est affectée par le changement de moyenne et par l’utilisation du poids de sondage.

* 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 :** Je cherche à savoir la différence entre deux moyennes. J’utilise alors la commande *dimean* de DASP. L’hypothèse nulle est que la moyenne des dépenses par équivalent adulte dans la région 1 est supérieure au double de la moyenne des dépenses par équivalent adulte dans la région 2. Cette hypothèse peut être reformulée comme ceci : la moyenne de *ae\_exp* dans la région 1 moins la moyenne de *ae\_exp* dans la région 3 est supérieure à la moyenne de *ae\_exp* dans la région 3. J’estime d’abord les dépenses moyennes par équivalent adulte dans la région 3 à l’aide de la commande suivante : *sum ae\_exp [aw = hhsize] if region==3*. J’utilise alors cette estimation, soit une valeur de 26080.66 comme valeur *test* de différence entre les deux moyennes. J’effectue la commande suivante :

*dimean ae\_exp ae\_exp, hsize1(hhsize) test(26080.66) cond1(region==3 ) hsize2(hhsize) cond2(region==1 ) conf(ub)*

La commande *dimean* fournit un tableau de test où je peux consulter les résultats. Je peux rejeter l’hypothèse nulle que la différence entre les deux moyennes est de 26080.66 en faveur de l’hypothèse alternative que la différence entre les deux moyennes est inférieure à 26080.66 car la p-value est de 0.0498, soit inférieure au seuil de confiance de 5%. Toutefois, je ne peux pas rejeter l’hypothèse nulle en faveur de l’hypothèse que la différence des moyennes est supérieure à 26080.66 car la p-value est égale à 0.95, soit supérieure au seuil de 5%. Je constate toutefois que les moyennes estimées par dimean ne sont pas égales à celles que je peux estimer avec ma première commande, soit *sum ae\_exp [aw = hhsize] if region==3*. Ainsi, il se peut que mes résultats soient erronés. Je remarque que l’erreur-type (standard error) de la moyenne de la région 1 est très élevée. Alors, il se peut que ce soit cette erreur-type qui empêche de pouvoir rejeter l’hypothèse nulle pour une différence supérieure à 26080.66.

* 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 :** J’effectue la commande suivante afin de tester l’hypothèse nulle que la différence entre les dépenses moyennes par équivalent adulte pour les chefs de famille hommes et les dépenses moyennes par équivalent adulte pour les chefs de famille femme est égale à zéro :

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

J’obtiens un tableau de résultat pour ce test. Je ne peux pas dire que la différence est supérieure à zéro et donc que les dépenses moyennes des chefs de famille homme sont supérieures aux dépenses moyennes de chefs de famille femme. En effet, la p-value du test différence > 0 est égale à 0.99, soit supérieur au seuil de confiance de 5%. Toutefois, je peux rejeter l’hypothèse nulle que la différence est égale à zéro et donc que les dépenses moyennes des chefs de famille homme sont égales aux dépenses moyennes des chefs de famille femme en faveur des hypothèses alternatives que la différence est inférieure à zéro (p-value = 0.0006) et qu’elle est différente de zéro (p-value = 0.0012). Alors, selon les résultats du test que j’ai effectué, les dépenses moyennes des chefs de famille homme seraient inférieures aux dépenses moyennes des chefs de famille femme. Pourtant, lorsque je regarde les moyennes estimées par la commande *dimean* pour ces deux groupes, je constate que l’inverse semble être vrai. Peut-être que le nombre d’observations de chefs de famille femme largement inférieur au nombre d’observations de chefs de famille homme entraîne ce résultat.

### 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 :** Je calcule la taille de la population des ménages échantillonnés en effectuant la commande suivante : *egen sum\_hhs = sum(hhsize)* qui utilise la fonction somme (sum(var)). J’effectue la commande *sum sum\_hhs* afin de visualiser le résultat, soit le *summary* de la nouvelle variable *sum\_hhs*, qui m’indique que la taille de la population des ménages échantillonnés vaut 14 694.

**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 :** J’ordonne les dépenses par habitant en ordre croissant avec la commande *sort pcexp*. Ensuite, je génère la variable *ps* avec la commande *gen ps = hhsize/sum\_hhs*. Finalement, je génère les variables centiles (p) et quantiles (q) avec respectivement les commandes suivantes : *gen p = sum(ps)* et *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 :** J’utilise la commande suivante pour tracer la courbe :

*line p pcexp, title(The cumulative distribution curve) ytitle(Centiles (p)) xtitle(Dépenses par habitant(pcexp))* *xscale(range(0 0.95))*

**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 :** J’utilise la commande suivante du module DASP pour tracer la courbe :

*c\_quantile pcexp, hsize(hhsize) min(0.0) max(0.95)*

Je peux voir que le niveau de dépense médian est d’environ 90 000. L’aire sous la courbe représente le revenu moyen de la population. On remarque que les 20 pourcents les plus riches de la population possèdent un revenu très élevé en comparaison à la moitié la moins riche. Ainsi, il semble y avoir de l’inégalité dans les revenus.

**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 :** J’utilise la commande suivante pour tracer les courbes :

*c\_quantile pcexp, hsize(hhsize) hgroup(zone) min(0.0) max(0.95)*

Je remarque que la courbe des quantiles pour la région urbaine est plus élevée que la courbe pour la région rurale. Ainsi, on peut dire que les habitants de la région urbaine sont plus riches que les habitants de la région rurale, ce qui peut représenter la réalité. Les 10% plus riches en zone rurale atteignent le 200 000 de dépenses par habitant, tandis que ce montant est atteint par les 40% plus riches en zone urbaine. On peut également clairement voir sur le graphique que l’aire sous la courbe urbaine est beaucoup plus élevée que l’aire sous la courbe rurale. Cela signifie que les dépenses moyennes en zone rurale sont nettement inférieures aux dépenses moyennes en zone urbaine.

**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 :** J’utilise la commande suivante pour tracer les courbes de densité :

*cdensity pcexp, hsize(hhsize) hgroup(sex) popb(1) min(0) max(1000000) ytitle(Densité f(y)) xtitle(Dépenses par habitant (pcexp))*

Je constate que les dépenses par habitant pour les chefs de famille hommes ont un pic plus élevé et plus mince que celui des chefs de famille femmes. Cela suggère des dépenses par habitant plus uniformes pour les chefs de famille hommes que pour les chefs de famille femmes. La courbe de densité pour les chefs de famille femme est plus étirée vers la droite, ce qui suggère un nombre plus élevé de femmes ayant des revenus élevés. Toutefois, il ne faut pas oublier qu’il y a un plus grand nombre de données pour les chefs de famille homme que pour les chefs de famille femme, ce qui pourrait expliquer que la courbe pour les chefs de famille femme soit plus étirée. Somme toute, on peut affirmer que la vaste majorité des ménages où les chefs de famille sont des hommes au sein de cette population ont des dépenses par habitant allant de 0 à 200 000. Il y a un très faible nombre d’individus ayant des dépenses par habitant au-delà de 200 000. La même conclusion peut être tirée pour les chefs de famille femmes, avec des dépenses par habitant valant en moyenne 100 000.