Corrigé - Série 5
Sources de biais et méthodes d'échantillonnage

Exercice 1

a) Son échantillon est sélectionné dans une sous-population ayant des intérêts différents des autres élèves sur la question. Réponses positives potentiellement surestimées.

b) L'enquêteur étant lui-même une personne handicapée, il incite les gens à répondre favorablement à sa question. Même s'il ne l'était pas, il existe une convention sociale qui dicte "la bonne réponse". De plus, l'échantillon n'est pas aléatoire (à l'aveuglette), et donc aucune marge d'erreur ne pourra être associée à l'estimation de la proportion.

c) Les gens dans les villages n'ont pas la même perception du besoin de conduire que les gens de la ville, n'ayant pas accès au transport en commun, et vivant souvent à plusieurs kilomètres des commerces et services essentiels. Il faut ajouter des citadins à l'échantillon.

d) What the $#"@&(%◇ is a APMSEQ ?

e) Les fréquences dans les choix de réponses devraient être explicitées, car la définition de *parfois* et *souvent* peut varier d'une personne à l'autre.

f) Par la poste... Même avec une enveloppe de retour pré-affranchie, c'est beaucoup trop d'étapes pour le répondant. Seuls les gens motivés ou ayant un intérêt dans la cause répondront, ce qui biaise l'échantillon. On voit d'ailleurs un très faible taux de réponses, donc pour interpréter le résultat comme ils le font, il faut faire l'hypothèse que les non répondants sont comme les répondants, ce qui n'est pas raisonnable ici.

g) 36% moins de caries que qui ? En combien de temps ? Sur combien de personnes a-t-on fait cette étude ? Quel laboratoire ? Pourrait-on citer l'étude ? A-t-elle été publiée ? Etc.

h) L'ADUL, malgré toutes ses bonnes intentions, ne possède probablement pas une liste à jour de tous les numéros de téléphone des diplômés de 1989. De plus, on devrait vérifier si tous les diplômés sont nécessairement "membres" de l'ADUL. Le revenu étant une question délicate, on aimerait peut-être avoir un taux de réponse associé. Quelle était la forme de la question : y avait-il des choix de réponses contenant des classes de revenu ? A-t-on tenu compte de la possibilité que les gens puissent mentir (pour diverses raisons) sur leur revenu annuel réel ? On devrait également afficher la marge d'erreur associée à ce salaire moyen.

i) Énoncé imprécis : Les enseignants sélectionnés étaient des spécialistes de quelle matière ? À quel niveau scolaire ? Quelle était la base de sondage ? La distribution des notes était-elle symétrique ? Serait-il possible que quelques notes très faibles aient tiré la moyenne vers le bas ? Quelle était la taille d'échantillon ?

j) L'intelligence est une variable difficilement mesurable. On devrait plutôt parler de résultat à un test précis mesurant certaines capacités cognitives ou académiques. De plus, on utilise la définition de la médiane ici, et on réfère à la moyenne. Si la distribution des valeurs de la variable appelée ici "intelligence" n'est pas symétrique, la phrase serait fausse même en parlant d'un test particulier.

Exercice 2

a) Échantillonnage à deux phases

b) Échantillonnage par grappes

c) Échantillonnage systématique

d) Échantillonnage à probabilité proportionnelle à la taille

e) Échantillonnage volontaire

f) Échantillonnage par quotas

g) Échantillonnage au jugé

Exercice 3

a) Échantillonnage stratifié avec une taille d'échantillon égale dans chaque strate.

b) Échantillonnage stratifié optimal.

c) Échantillonnage aléatoire simple ou systématique

d) Échantillonnage volontaire.

e) Échantillonnage à 2 degrés ou par grappes si le nombre de jeunes concernés est petit dans chaque école.

Exercice 4

a) Une liste d'unités parmi lesquelles l'enquêteur sélectionnera un échantillon.

b) Être représentatif de la population.

c) Connaître le nombre d'individus dans chaque strate de la population.

d) Connaître la variance de la variable considérée dans chaque strate de la population (ainsi que le nombre d'individus dans chaque strate).

e) Choisir un grand nombre de petites grappes, plutôt qu'un petit nombre de grosses grappes.

f) Les deux méthodes échantillonnent des grandes unités qui contiennent les individus sur lesquels seront prises les mesures ou à qui seront posées les questions. L'échantillonnage en grappes prend tous les individus, et l'échantillonnage à deux degrés sélectionne parmi les individus des unités primaires.

g) Les deux méthodes comportent deux étapes d'échantillonnage. Dans l'échantillonnage à deux degrés, on échantillonne des unités de taille différente aux deux étapes. Dans l'échantillonnage à deux phases, on échantillonne des individus à la première étape, on prend une certaine mesure sur eux, puis on échantillonne à nouveau dans un sous-groupe de ce premier échantillon.

h) On ne peut pas associer d'ereur-type à notre estimation.

Exercice 5

Vous devez sélectionner $n = 150$ personnes dans la ville de Québec.

a) Échantillonnage systématique où $N = 400\,050$. Pas d'échantillonnage : $k = 2667$. Individu 1 : numéro 2590. Individus 2 et 3 : numéros 5257 et 7924.

b) Il y a 281 860 individus dans la population identifiée. Vous devrez donc multiplier le nombre de personnes dans chaque classe par $150/281\,860$ (la fraction de sondage). Si vous préférez, la fréquence relative de chaque classe, $f_i/281\,860$ représente le poids de chaque strate, que vous multipliez par 150. Voici la répartition de l'échantillon :

Classe d'âge	Hommes	Femmes
25 à 29 ans	10	10
30 à 34 ans	8	8
35 à 39 ans	8	8
40 à 44 ans	10	10
45 à 49 ans	11	11
50 à 54 ans	10	11
55 à 59 ans	9	10
60 à 64 ans	8	9
Total	74	77

c) – Seuls les propriétaires peuvent faire partie du jury. L'échantillon sera clairement biaisé.
– Selon les secteurs sélectionnés au premier degré, certains groupes de citoyens pourraient ne pas être représentés dans l'échantillon.
– Une ville n'est pas un territoire si vaste, et les listes de citoyens ne sont pas si difficiles à obtenir. Il n'y a pas lieu de procéder à deux degrés.
– Puisqu'il y a un relevé de taxes par immeuble, on aurait pu sélectionner les immeubles résidentiels comme unité primaire, et un des résidents comme unité secondaire. Cela donne par contre une plus grande probabilité de sélection aux gens habitant seuls, par rapport aux gens vivant dans de grands immeubles à logements.

Exercice 6

a) Avantages :

(a) simple à comprendre et à éxécuter

(b) formules d'estimation simples à calculer

(c) aucune information auxiliaire nécessaire

(d) tous les individus (et tous les échantillons de taille n) ont une chance égale d'être sélectionnés

Inconvénients :

- besoin d'une liste d'individus composant la population
- peut coûter cher (si les individus sont dispersés, par exemple)
- peut être non représentatif en fonction d'un critère important

b) 100^{10} échantillons ordonnés de taille 10. Ils ont tous la même probabilité de sélection $(1/100^{10})$. Si on avait considéré les échantillons non ordonnés, il aurait été plus ardu de les dénombrer, et ils n'auraient pas tous la même probabilité de sélection.

c) $\binom{100}{10}$ échantillons non ordonnés de taille 10. Ils tous la même probabilité de sélection $(1/\binom{100}{10})$. On peut aussi considérer les échantillons ordonnés comme à la question précédente : $\frac{100!}{90!}$ échantillons ordonnés de taille 10. Ils tous la même probabilité de sélection ($\frac{90!}{100!} = \frac{1}{100\times99\times\ldots\times91}$).

d) $100/10 = 10$ échantillons de taille 10. Ils ont tous la même probabilité de sélection $(1/10)$.

Exercice 7

- p_i = probabilité de l'individu i de faire partie de l'échantillon.
- w_i = poids de sondage de l'individu i, le nombre d'unités de la population que l'individu i représente. Lorsqu'un échantillonnage simple ou systématique est effectué, on a que $w_i = \dfrac{1}{p_i}$.

a) Oui. $p_i = 1 - \left[\frac{N-1}{N}\right]^{150} = 1 - \left[\frac{99\,999}{100\,000}\right]^{150} = 0,0014989$

b) Oui. $p_i = \frac{n}{N} = \frac{150}{100\,000}$

c) Oui. $p_i = \frac{n}{N} = \frac{100}{100\,000}$

d) Non. $p_i = \frac{\text{taille de l'individu } i}{\text{somme de toutes les tailles}}$

e) Non. $p_i = \frac{n_h}{N_h} = \begin{cases} \frac{100}{131\,670} = 0.08\% & \text{dans la strate 1} \\ \frac{100}{155\,215} = 0.06\% & \text{dans la strate 2} \\ \frac{100}{88\,525} = 0.11\% & \text{dans la strate 3} \end{cases}$

f) Oui. $p_i = \frac{200}{375\,410} = 0.05\%$

g) Les individus ont la même probabilité de sélection, car il s'agit d'un échantillon aléatoire simple : $p_i = \frac{n}{N} = \frac{100}{375\,410}$.

Par contre, le poids de sondage diffère selon la strate :

$w_i = \frac{1}{20/131\,670} = 6\,584$ dans la strate 1
$w_i = \frac{1}{40/155\,215} = 3\,880$ dans la strate 2
$w_i = \frac{1}{40/88\,525} = 2\,213$ dans la strate 3

Ceux qui ont le poids de sondage le plus élevé sont ceux qui étaient sous-représentés dans l'échantillon.

h) Oui. $p_i = \frac{10}{3\,000}$

i) Non. $p_i = \frac{10}{3\,000} \times \frac{5}{e}$, où e = nombre d'employés dans le restaurant où travaille l'individu i.

j) Premier échantillonnage : Oui. $p_i = \frac{500}{N}$.
Deuxième échantillonnage : Non. Les employés travaillant moins de 35 heures n'ont aucune chance d'être sélectionnés.

Exercice 8

- Revenir à la charge auprès des non répondants : une bonne idée en général (surtout quand le taux de réponse est bas), nécessite du temps (donc des coûts supplémentaires), peut être agaçant pour le répondant (donc orienter ses réponses pour se débarrasser de l'enquêteur), ne garantit pas un taux de non réponses nul.
- Ignorer les non répondants : Ne pas les considérer biaise l'échantillon, car c'est prendre pour acquis que les non répondants ont les mêmes caractéristiques que les répondants et donc qu'ils répondront comme eux, ce qui est faux en général.
- Imputer des valeurs aux non-répondants : on considère encore que les non répondants répondraient comme les répondants, ce qui est risqué.
- Imputer des valeurs aux non-répondants en tenant compte de certaines caractéristiques : C'est l'idéal pour imputer des valeurs. Par contre, il faut connaître les caractéristiques des non-répondants, ce qui n'est pas évident s'ils n'ont pas participé au sondage. S'applique bien lorsqu'il manque certaines réponses dans le questionnaire ou dans un plan stratifié où on connaît au moins la strate du non répondant. Attention de toujours se rappeler qu'il s'agit de données imputées, qui n'ont pas la même valeur que les vraies observations.

Exercice 9

$n = 600$

a) Allocation arbitraire : $n_1 = n_2 = n_3 = 200$

b) Allocation proportionnelle :

$$\begin{aligned} n_1 &= 0,351 \times 600 &= 211 \\ n_2 &= 0,413 \times 600 &= 248 \\ n_3 &= 0,236 \times 600 &= 142 \end{aligned}$$

c) Allocation optimale :

$$\begin{aligned} W_1\,\sigma_1 &= 0,351 \times 3 = 1,053 \\ W_2\,\sigma_2 &= 0,413 \times 1 = 0,413 \\ W_3\,\sigma_3 &= 0,236 \times 2 = 0,472 \\ &\qquad\qquad\quad 1,938 \end{aligned}$$

$$\begin{aligned} n_1 &= \frac{1,053}{1,938} \times 600 = 326 \\ n_2 &= \frac{0,413}{1,938} \times 600 = 128 \\ n_3 &= \frac{0,472}{1,938} \times 600 = 146 \end{aligned}$$

Exercice 10

Classe d'âge	Éch. aléatoire simple	Éch. stratifié (arbitraire)	Éch. stratifié proportionnel	Moyenne échantillonnale	Écart-type échantillonnal
20 à 39 ans	20	33	35	6	1
40 à 59 ans	40	33	41	8	1
60 à 79 ans	40	33	24	7	2
Total	100	99	100		

Classe d'âge	Nombre d'individus	Fréquence relative
20 à 39 ans	131 670	35,1%
40 à 59 ans	155 215	41,3%
60 à 79 ans	88 525	23,6%
Total	375 410	100,0%

a) L'estimation de l'erreur-type de la moyenne $\overline{X}_1$ est $\sqrt{(1-f_1)\frac{s_1^2}{n_1}} = \sqrt{\left(1-\frac{n_1}{131\,670}\right)\frac{1^2}{n_1}}$. Une grande taille d'échantillon donnera une plus grande précision dans la strate 1. Ici, c'est la stratification proportionnelle, mais seulement à cause du $n_1 = 35$, non à cause de la méthode comme telle.

b) Échantillonnage aléatoire simple

$$\begin{aligned} \hat{\mu} &= \frac{20(6)+40(8)+40(7)}{100} = 7,20 \\ \widehat{Var}(\hat{\mu}) &= (1-f)\frac{s^2}{n} \end{aligned}$$

Il faut donc calculer $s^2 = \dfrac{\sum_{i=1}^{100}(x_i - \overline{x})^2}{n-1} = \dfrac{\sum_{i=1}^{100} x_i^2 - n\overline{x}^2}{n-1} = \dfrac{\sum_{i=1}^{100} x_i^2 - 100(7,20)^2}{99}$.

On connaît $s_1,\ s_2,\ s_3$, alors on isole la somme des x_i^2 dans chacune des formules et on obtient :

$$\sum_{i=1}^{n_h} x_i^2 = (n_h - 1)\ s_h^2\ +\ n_h\ \overline{x_h}^2 \quad \text{(dans chaque strate)}$$

$$\sum_{i=1}^{20} x_i^2 = 739, \quad \sum_{i=21}^{60} x_i^2 = 2\,599, \quad \sum_{i=61}^{100} x_i^2 = 2\,116$$

$$\Rightarrow \quad \sum_{i=1}^{100} x_i^2 = 5\,454 \quad \text{(trois strates combinées)}$$

$$s^2 = \frac{\sum_{i=1}^{100} x_i^2 - n\overline{x}^2}{n-1} = \frac{5\,454 - 100(7,20)^2}{99} = 2,73$$

On peut ainsi calculer l'erreur-type associée à $\hat{\mu}$:

$$\widehat{Var}(\hat{\mu}) = (1-f)\frac{s^2}{n} = \left(1 - \frac{100}{375\,410}\right)\frac{2,73}{100} = 0,0273$$

$$\text{Erreur-type}(\hat{\mu}) = \sqrt{0,0273} = 0,1652$$

c) Échantillonnage aléatoire stratifié arbitraire

$$\hat{\mu} = \sum_{h=1}^{3} W_h\,\overline{X}_h = 0,351(6) + 0,413(8) + 0,236(7) = 7,062$$

$$\widehat{Var}(\hat{\mu}) = \sum_{h=1}^{3} W_h^2(1-f_h)\frac{s_h^2}{n_h}$$

$$= \frac{1}{33}\left[0,351^2\left(1 - \frac{33}{131\,670}\right)1^2 + 0,413^2\left(1 - \frac{33}{155\,215}\right)1^2 + 0,236^2\left(1 - \frac{33}{88\,525}\right)2^2\right]$$

$$= 0,0156$$

$$\text{Erreur-type}(\hat{\mu}) = \sqrt{0,0156} = 0,1251$$

d) Échantillonnage aléatoire stratifié proportionnel

$$\begin{aligned}
\hat{\mu} &= \sum_{h=1}^{3} W_h \overline{X}_h = 7,062 \\
\widehat{Var}(\hat{\mu}) &= \sum_{h=1}^{3} W_h^2 (1 - f_h) \frac{s_h^2}{n_h} \\
&= \left(1 - \frac{100}{375\,410}\right) \left[0,351^2 \frac{1^2}{35} + 0,413^2 \frac{1^2}{41} + 0,236^2 \frac{2^2}{24}\right] \\
&= 0,01696 \\
\text{Erreur-type}(\hat{\mu}) &= \sqrt{0,01696} = 0,1302
\end{aligned}$$

e) Dans cet exemple particulier, la meilleure précision est atteinte par la stratification avec allocation arbitraire. Quelques remarques :

- C'est sans surprise que la stratification est meilleure que l'échantillonnage aléatoire simple. C'est le cas dès qu'on a une bonne variable de stratification, i.e. une variable pour laquelle la moyenne diffère d'une strate à l'autre (le nombre moyen d'heures de sommeil diffère d'un groupe d'âge à l'autre).

- La stratification offrant la meilleure précision est l'allocation optimale, qui tient compte de la variance dans chaque strate et non seulement de la taille des strates (comme l'allocation proportionnelle). En plus de piger beaucoup d'unités dans les strates populeuses, on veut en piger beaucoup quand la variance est grande.

  Dans notre exemple, la 2^e strate est la plus populeuse, mais la 3^e strate est la plus variable. Pour une précision optimale, il faudrait considérer l'échantillonnage stratifié proportionnel, mais attribuer un peu plus d'observations dans la strate 3, et un peu moins dans la strate 1. Puisque l'échantillonnage arbitraire répond à ces deux conditions dans notre exemple (par hasard, disons-le), c'est cette méthode qui a maximisé la précision.