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

Exercice 1 - Les sources de biais

À partir des situations ci-dessous, déterminer les sources de biais pouvant fausser l'estimation du paramètre d'intérêt.

a) La directrice d'une école se demande si les élèves aimeraient augmenter les heures de cours d'éducation physique. Elle sollicite l'avis des élèves faisant partie de l'équipe de basketball.

b) À la sortie d'un magasin, une personne en fauteuil roulant effectue un sondage auprès des clients sur les investissements publics pour améliorer l'accès des commerces aux handicapés.

c) Une association qui milite contre la vitesse au volant veut connaître l'opinion des Québécois sur l'âge d'obtention du permis de conduire. Elle interroge un échantillon aléatoire de résidents des trois villages les plus proches.

d) Question de sondage : *Vous impliquez-vous dans une APMSEQ ?* Oui ☐ Non ☐

e) Question de sondage : *À quelle fréquence visitez-vous la bibliothèque municipale ?* Jamais ☐ Parfois ☐ Assez souvent ☐ Très souvent ☐

f) Le Parti Pourlagloire envoie 2000 questionnaires par la poste pour connaître les intentions de vote des citoyens. Parmi les 35 questionnaires retournés, 80% indiquent une intention de vote pour le Parti Pourlagloire.

g) Une publicité affirme qu'un laboratoire indépendant a démontré que les gens qui utilisent le dentifrice Dental ont 36% moins de caries.

h) L'Association des diplômés de l'Université Laval a fait un sondage téléphonique auprès de ses membres qui révèle que le salaire moyen des diplômés de 1989 est de 47 560$.

i) Manchette du jour : On a fait passer un examen d'histoire à un groupe d'enseignants choisis au hasard, et la moyenne des notes fut de 56%. On peut donc conclure que les enseignants sont très faibles en histoire.

j) Titre d'un article : Il y a autant de personnes plus intelligentes que la moyenne que de personnes moins intelligentes que la moyenne.

Certaines questions sont inspirées de *Intersection*, 1^{re} année du 2^{e} cycle, Graficor (Chenelière), 2008, p. 127, et de *Visions*, 1^{re} année du 2^{e} cycle, CEC, 2007, p. 121.

Exercice 2 - Identifier la méthode

À partir des contextes ci-dessous, déterminer le type d'échantillonnage dont il est question.

a) Des chercheurs ont contacté par téléphone 200 anciens sportifs de haut niveau choisis aléatoirement à partir d'une liste fournie par un organisme fédéral. Ceux qui ont admis avoir déjà ressenti des symptômes de dépression ont reçu un questionnaire par la poste contenant les questions les plus importantes pour la recherche.

b) Des récréologues veulent s'informer sur le niveau de participation des jeunes enfants inscrits dans les équipes de soccer au Québec. Ils obtiennent une liste des équipes de niveau Atome en colligeant les données des associations de soccer régionales. Ils sélectionnent au hasard 15 équipes et interrogent tous les membres de chaque équipe.

c) Pour mesurer le temps moyen passé par les clients dans son magasin, un commerçant donne un chronomètre à un client sur dix, dans l'ordre d'entrée au magasin.

d) On veut choisir un syndicat parmi un groupe de 12 syndicats représentant un total de 20 000 membres. On veut s'assurer qu'un syndicat ayant plus de membres ait plus de chances d'être sélectionné. Le 1^{er} syndicat compte 350 membres, le 2^{e} en compte 500, etc. On choisit donc aléatoirement un nombre entre 1 et 20 000. Si le nombre est entre 1 et 350, on choisit le 1^{er} syndicat ; Si le nombre est entre 351 et 850, on choisit le 2^{e} syndicat, etc.

e) Une station de radio invite ses auditeurs à répondre à une question de sondage sur son site web.

f) Un directeur de centre commercial veut connaître le montant moyen dépensé par les clients lors d'une visite. Il demande à un enquêteur de conduire des entrevues auprès 200 clients, en respectant les proportions suivantes : 50% de femmes - 50% d'hommes, 1/3 de jeunes - 1/3 d'adultes - 1/3 de personnes âgées. Il doit rencontrer 100 personnes durant les jours de semaine, et 100 personnes la fin de semaine.

g) Un garde forestier est chargé de récolter des données sur le braconnage en interceptant les véhicules qui lui paraissent suspects.

Exercice 3 - Choisir la méthode

En fonction de la mise en situation proposée, déterminer le meilleur type d'échantillonnage que le chercheur devrait choisir.

a) Une compagnie pharmaceutique veut vérifier les effets secondaires d'un nouveau médicament prêt à être testé sur des femmes ménopausées. Elle doit obtenir la même précision dans tous les sous-groupes de patientes, i.e. dans toutes les combinaisons des facteurs suivants : celles qui prennent des hormones et les autres, celles souffrant d'hypertension et les autres, celles souffrant d'osthéoporose et les autres. On suppose que la variabilité est égale dans tous les sous-groupes.

b) Le conseil municipal veut estimer le niveau d'endettement hypothécaire de ses résidents. Il sait que cette valeur change beaucoup d'un individu à l'autre, mais qu'elle est moins variable à l'intérieur d'un même quartier. Le conseil souhaite avoir une estimation la plus précise possible avec un échantillon le plus petit possible.

c) Le Barreau du Québec veut s'informer sur le revenu et la charge de travail de ses membres. Il a conçu un questionnaire qu'il souhaite envoyer par la poste à 150 avocats membres du Barreau.

d) Pour leur travail de session, une équipe d'étudiants en psychologie doit monter une expérience de 30 minutes pour analyser la perception du temps dans divers environnements. Ils disposent d'un budget de 500$ pour dédommager les éventuels participants. La collecte de données doit être terminée dans trois semaines. Ils n'ont accès à aucune liste d'individus, sauf les listes de courriels de leur université.

e) Le Ministère de la santé s'inquiète du nombre croissant de prescriptions de Ritalin chez les jeunes du primaire. Il veut mener une étude longitudinale (sur plusieurs années) auprès d'un échantillon d'élèves pour comparer plusieurs méthodes d'intervention. Il serait utile que les jeunes sélectionnés appartiennent à quelques écoles seulement.

Exercice 4 - Connaître les méthodes

a) Qu'est-ce qu'une base de sondage ?

b) Quelle est la qualité principale d'un échantillon ?

c) Quelle est la condition nécessaire à l'utilisation de l'échantillonnage stratifié proportionnel ?

d) Quelle est la condition nécessaire à l'utilisation de l'échantillonnage stratifié optimal ?

e) Comment peut-on améliorer la précision de l'échantillonnage par grappes ?

f) Comment l'échantillonnage par grappes peut-il se comparer à l'échantillonnage à deux degrés ?

g) Comment l'échantillonnage à deux degrés se compare-t-il à l'échantillonnage à deux phases ?

h) Quel est l'inconvénient principal des méthodes non probabilistes ?

Questions inspirées de Ouellet, Gilles, *Statistique et Probabilités*, Le Griffon d'Argile, 1998, p. 299.

Exercice 5 - Appliquer les méthodes

Vous êtes consultant pour une firme spécialisée en échantillonnage. La Cour du Québec vous engage pour élaborer une procédure aléatoire de sélection d'un jury pour un procès. Vous devez sélectionner 150 personnes dans la ville de Québec parmi lesquelles 12 seront choisies par le juge et les avocats des deux parties via un processus visant à éliminer les individus trop biaisés qui ne sauraient porter un jugement impartial.

a) Dans un premier temps, vous proposez d'effectuer un échantillonnage systématique à partir d'une liste des résidents adultes fournie par votre client. Cette liste contient 400 050 noms placés en ordre alphabétique. Le numéro du premier individu choisi au hasard est 2590. Donnez les numéros des deux individus suivants qui seront sélectionnés dans votre échantillon.

b) Vous vous entendez avec votre client pour stratifier votre échantillon selon l'âge et le sexe de façon proportionnelle à la répartition dans la population. Vous convenez également de ne sélectionner que des gens âgés entre 25 et 64 ans. En consultant le *Profil des communautés* de 2006 de Statistique Canada pour la ville de Québec, dites combien d'hommes et de femmes de chaque groupe d'âge vous devrez sélectionner dans votre échantillon.

c) Un de vos collègues propose d'effectuer un échantillonnage à deux degrés en se basant sur les relevés de taxes foncières. Il suggère d'utiliser les secteurs de la ville comme unités primaires d'échantillonnage et les propriétés comme unités secondaires. Les propriétaires sélectionnés feraient alors partie du jury potentiel. Critiquez ce plan.

Exercice 6 - Probabilité de sélection de l'échantillon

a) Quels sont les avantages et les inconvénients de l'échantillonnage aléatoire simple ?

b) Combien d'échantillons ordonnés de taille 10 peut-on sélectionner aléatoirement avec remise dans une population de taille 100 ? Ces échantillons ont-ils tous la même probabilité de sélection ?

c) Combien d'échantillons de taille 10 peut-on sélectionner aléatoirement sans remise dans une population de taille 100 ? Ces échantillons ont-ils tous la même probabilité de sélection ?

d) Combien d'échantillons de taille 10 peut-on sélectionner systématiquement dans une population de taille 100 dont la liste ne peut être réordonnée ? Ces échantillons ont-ils tous la même probabilité de sélection ?

Exercice 7 - Probabilité de sélection de l'individu et poids de sondage

- On définit p_i comme la probabilité de l'individu i de faire partie de l'échantillon. C'est sa *probabilité de sélection*.
- On définit w_i comme le *poids de sondage* de l'individu i. Le poids de sondage peut être vu comme le nombre d'unités de la population que l'individu i représente. Lorsqu'un échantillonnage simple ou systématique est effectué, on a que $w_i = \dfrac{1}{p_i}$.

Dans les situations ci-dessous, dites si tous les individus ont une même probabilité de sélection. Lorsque c'est le cas, donnez cette probabilité. Lorsque ce n'est pas le cas, dites quels individus ont une plus grande probabilité d'être sélectionnés.

a) On sélectionne aléatoirement avec remise 150 individus d'une population en contenant 100 000.

b) On sélectionne aléatoirement sans remise 150 individus d'une population en contenant 100 000.

c) On sélectionne systématiquement 100 individus d'une population en contenant 100 000.

d) On sélectionne 100 propriétaires selon un échantillonnage avec probabilité proportionnelle à la taille. On voudrait que la taille soit le nombre d'unités d'habitation dans les propriétés.

e) On sélectionne 100 personnes dans la ville de Québec dans chacune des trois classes d'âge ci-dessous, où la population (et disons la base de sondage) se répartit comme suit :

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%

f) On sélectionne aléatoirement 70 personnes entre 20 et 39 ans dans la ville de Québec, 82 personnes entre 40 et 59 ans, et 48 personnes entre 60 et 79 ans.

g) Dans le cas d'une stratification à posteriori, la pondération se fait après la collecte des données. Supposons qu'on procède à un échantillonnage aléatoire simple sans remise de 100 adultes de Québec et que notre échantillon contient 20 personnes entre 20 et 39 ans, 40 personnes entre 40 et 59 ans, et 40 personnes entre 60 et 79 ans. On veut pondérer les moyennes obtenues dans chaque strate pour obtenir une représentation proportionnelle selon la population. Quels individus auront alors un poids de sondage plus élevé ?

h) On sélectionne aléatoirement 10 restaurants Tim Hortons parmi les 3000 franchises au Canada et aux États-Unis, et on interroge tous les employés de chaque établissement sélectionné sur le nombre d'heures travaillées au cours de la dernière semaine.

i) On sélectionne aléatoirement 10 restaurants Tim Hortons parmi les 3000 franchises. On sélectionne aléatoirement 5 employés par franchise et on les interroge sur le nombre d'heures travaillées au cours de la dernière semaine.

j) On sélectionne aléatoirement 500 employés de Tim Hortons pour connaître le nombre d'heures travaillées au cours de la dernière semaine. On forme ensuite une liste de tous les répondants ayant travaillé plus de 35 heures, et on y pige 50 personnes pour leur soumettre un questionnaire plus détaillé.

Exercice 8 - La non réponse

Il est fréquent que des individus sélectionnés dans un échantillon refusent de participer au sondage ou de répondre à certaines questions. On les appelle les non-répondants. Plusieurs stratégies pourraient être envisagées pour traiter ces manquements.

Critiquez les propositions suivantes.

- On peut revenir à la charge auprès des non répondants en les contactant de nouveau ou en utilisant d'autres méthodes d'enquêtes plus efficaces.
- On peut ignorer les non répondants et calculer la moyenne à partir des réponses obtenues.
- On peut *imputer* des valeurs aux non-répondants, i.e. faire comme s'ils avaient donné comme réponse la valeur moyenne des répondants.
- On peut *imputer* des valeurs aux non-répondants en tenant compte de certaines caractéristiques (par ex. sexe, revenu, âge), i.e. faire comme s'ils avaient donné comme réponse la moyenne des répondants ayant les mêmes caractéristiques qu'eux.

Exercice 9 - Combien d'unités par strate ?

On veut estimer le nombre moyen d'heures de sommeil en tenant compte de la distribution par classe d'âge des adultes de Québec, car on croit que le nombre d'heures de sommeil varie beaucoup d'une classe à l'autre mais varie peu au sein d'une même classe. On veut obtenir un échantillon de taille $n = 600$ dans cette population selon diverses méthodes.

On rappelle la distribution de la population pour chacune des classes d'âge :

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%

Combien d'individus sélectionneriez-vous dans chaque classe d'âge dans les situations suivantes ?

a) On considère que la variance du nombre d'heures de sommeil est identique dans chaque classe d'âge, et on veut la même précision en estimant la moyenne d'heures dans chaque classe.

b) On veut que l'échantillon soit le plus représentatif possible de la population en termes de classes d'âge.

c) On veut une précision maximale pour l'estimation de la moyenne globale, sans égard aux strates prises individuellement. Une étude réalisée dans une autre ville il y a 10 ans nous permet de considérer que les écarts-types dans chaque classe d'âge sont de 3 heures, 1 heure et 2 heures respectivement.

Exercice 10 - Et la précision ?

Considérons à nouveau la situation précédente, où on s'intéresse au nombre moyen d'heures de sommeil chez la population adulte de Québec. On procède à trois échantillonnages, avec des méthodes différentes. Supposons, pour les besoins de l'exercice, que nous avons obtenu les mêmes moyennes échantillonnales et les mêmes écarts-types échantillonnaux dans chaque strate avec les trois méthodes. (Ceci est évidemment très peu probable.) Voici les tailles d'échantillon et les statistiques observées.

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		

On rappelle la distribution de la population pour chacune des classes d'âge :

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) Dans la situation présente, laquelle des trois méthodes donnera la meilleure précision sur l'estimation de la moyenne dans la classe 20-39 ans ?

b) Quelle est l'estimation de la moyenne globale obtenue par échantillonnage aléatoire simple ? Quelle est l'erreur-type qui lui est associée ?

c) Quelle est l'estimation de la moyenne globale obtenue par échantillonnage stratifié arbitraire ? Quelle est l'erreur-type qui lui est associée ?

d) Quelle est l'estimation de la moyenne globale obtenue par échantillonnage stratifié proportionnel ? Quelle est l'erreur-type qui lui est associée ?

e) Laquelle des trois méthodes ci-dessus vous a donné la meilleure précision sur l'estimation de la moyenne ? Auriez-vous pu le déterminer avant de faire les calculs ?