

Corrigé du chapitre 5

Comprendre les données sociales

Mark Paradis, Ph. D.

2 septembre 2026

Voici le corrigé détaillé des exercices de *fin de chapitre* du chapitre 5 de *Comprendre les données sociales*. Le corrigé des exercices situés à l'intérieur de chaque section se trouve dans le livre lui-même, immédiatement après les exercices auxquels il répond.

Faites l'exercice avant de lire le corrigé. Un corrigé lu en premier n'est qu'un texte de plus ; un corrigé lu en second est une vérification de votre raisonnement, et seul le second enseigne quelque chose.

Le corrigé complet, pour tous les chapitres, se trouve à <https://www.markpara-dispolitics.com/books/comprendre-les-donnees-sociales/online/solutions/>.

Chapitre 5 Techniques d'échantillonnage

Corrigé de l'exercice 5.E1. (a) Population cible : les électeurs américains de l'élection présidentielle de 1936. Base de sondage : les détenteurs d'immatriculations automobiles, les abonnés du téléphone et la liste d'abonnés du magazine lui-même.

L'écart est économique. En 1936, dans la septième année de la Dépression, posséder une voiture ou un téléphone désignait un ménage comme relativement prospère, et l'abonnement à un magazine faisait de même. La base surreprésentait donc les électeurs plus fortunés, qui favorisaient Landon, et sous-représentait les électeurs plus pauvres, qui favorisaient Roosevelt.

(b) Principalement une *erreur de couverture* — la base ne correspondait pas à la population. Le biais de non-réponse l'a aggravée : environ un quart seulement des dix millions de bulletins ont été retournés, et les électeurs mécontents de Roosevelt avaient plus de motivation à répondre. Les deux étaient présents, et une réponse complète nomme l'erreur de couverture comme principale tout en signalant la seconde.

Notez ce qui n'était *pas* en cause : la variabilité d'échantillonnage. Avec 2,4 millions de réponses, la variation aléatoire était négligeable.

(c) Parce que la taille de l'échantillon ne réduit que l'erreur d'échantillonnage, et que l'erreur d'échantillonnage n'était pas le problème. Les deux défaillances étaient ici systématiques : elles poussent l'estimation dans une direction constante, et recueillir plus de réponses de la même base déformée reproduit la déformation avec plus de précision. Le *Digest* a produit une estimation extrêmement précise des intentions de vote des Américains prospères qui retournaient des bulletins postaux.

(d) Gallup a employé l'échantillonnage par quotas : on assignait aux intervieweurs des cibles quant au nombre de répondants de chaque âge, sexe et tranche économique à obtenir. Les quotas forçaient l'échantillon à correspondre à la population sur les caractéristiques que la base du *Digest* avait déformées — avant tout la tranche économique.

Cinquante mille répondants tirés par quotas ont battu 2,4 millions de répondants autosélectionnés parce que le plus petit échantillon était apparié à la population sur la variable qui comptait.

La suite compte tout autant : la méthode de Gallup a échoué en 1948, quand la liberté laissée aux intervieweurs dans le choix des personnes à approcher a produit un penchant systématique à elle. C'est cet échec qui a fait de l'échantillonnage probabiliste la norme.

Corrigé de l'exercice 5.E2. (a) Deux raisons. *Représentation garantie* : un échantillon aléatoire simple de Canadiens comprendrait parfois, par hasard, très peu de résidents des petites provinces — l'Île-du-Prince-Édouard représente moins d'un demi pour cent de la population, de sorte qu'un EAS de 10 000 devrait en donner

environ 40, et pourrait en donner nettement moins ou nettement plus par le seul effet du hasard. Stratifier par province fixe le nombre tiré de chacune. *Précision* : bien des variables sociales et économiques diffèrent systématiquement d'une province à l'autre, et retirer de l'estimation cette variation interprovinciale réduit l'erreur d'échantillonnage des chiffres nationaux.

(b) Elle suréchantillonne les petites provinces pour que les estimations *provinciales* soient utilisables. Une répartition proportionnelle donnerait à l'Île-du-Prince-Édouard si peu de répondants que sa propre marge d'erreur serait énorme, et l'un des objets de l'ESG est précisément de publier des résultats par province.

L'ajustement est la *pondération*. Les provinces suréchantillonnées sont surreprésentées dans l'échantillon par rapport à la population : dans les estimations nationales, chacun de leurs répondants doit donc compter pour moins de un, et chaque répondant d'une province sous-échantillonnée pour plus de un. Sans pondération, les chiffres nationaux seraient tirés vers les petites provinces.

(c) Il n'existe aucune liste nationale d'individus qui pourrait servir de base de sondage et, quand bien même il en existerait une, un EAS disperserait les intervieweurs sur des milliers d'adresses distinctes. L'échantillonnage par grappes à plusieurs degrés n'a besoin que d'une liste de territoires au premier degré, puis d'une liste de logements dans les territoires retenus au second.

Le compromis porte sur la précision. Les personnes d'une même grappe se ressemblent, de sorte que chaque répondant supplémentaire du même territoire apporte moins d'information nouvelle qu'un répondant tiré au hasard dans l'ensemble du pays. L'effet de plan qui en résulte fait que la taille effective d'échantillon de l'enquête est inférieure à son nombre d'entrevues, et les marges d'erreur publiées en tiennent compte.

Corrigé de l'exercice 5.E3. (a) L'échantillonnage probabiliste exige une base de sondage — une liste de la population dans laquelle tirer. Aucune liste de ce genre n'existe pour les personnes qui consomment actuellement des drogues illicites, et aucune ne pourrait exister : l'appartenance à cette population n'est consignée nulle part, et beaucoup de ses membres ont de fortes raisons de la dissimuler. Même des bases partielles, comme la clientèle des services de traitement, ne captent que celles et ceux qui ont pris contact avec un système, ce qui est une population différente et non aléatoire.

(b) *L'échantillonnage par boule de neige* est le choix habituel. De premiers participants, recrutés par des services de proximité ou des programmes de réduction des méfaits, en réfèrent d'autres tirés de leurs propres réseaux. Il fonctionne ici parce que la confiance se transmet le long d'une référence comme la démarche d'une chercheuse ne saurait le faire, et parce que la population est définie en partie par son souhait de ne pas être identifiée.

L'échantillonnage dirigé par les répondants — une variante structurée qui suit les chaînes de référence et pondère en conséquence — est la réponse la plus sophistiquée, si vous l'avez rencontrée.

Limite à reconnaître : l'échantillon est borné par les réseaux dans lesquels il commence, de sorte que les consommateurs les plus isolés socialement sont les moins susceptibles d'être joints.

(c) Non. Le 74 % décrit son échantillon, et son échantillon n'a pas été tiré au hasard parmi les consommateurs d'opioïdes de la ville. Les personnes jointes par des réseaux de référence diffèrent vraisemblablement de celles qui ne le sont pas — plus branchées socialement, plus susceptibles d'être en contact avec des services, et plausiblement différentes dans leurs habitudes de consommation.

Ce qu'elle peut valablement conclure : que la consommation quotidienne est répandue chez les personnes qu'elle a jointes, et que c'est une population qui mérite d'être étudiée davantage. Elle peut aussi faire des comparaisons internes — savoir si les consommateurs quotidiens de son échantillon diffèrent des consommateurs occasionnels sur d'autres mesures —, lesquelles résistent mieux aux problèmes de sélection que la proportion principale. Ce qu'elle ne peut pas faire, c'est rapporter 74 % comme une estimation portant sur les consommateurs d'opioïdes de la ville, avec ou sans marge d'erreur.

Corrigé de l'exercice 5.E4. (a)

$$\text{marge} = \frac{1}{\sqrt{1\,847}} \times 100 \approx \frac{1}{42,98} \times 100 \approx \pm 2,3 \%$$

Cela correspond presque exactement au chiffre annoncé, ce qui vous apprend que la firme l'a calculé avec la formule habituelle — comme si l'échantillon avait été tiré au hasard.

(b) Non. L'échantillon a été tiré par quotas dans un panel en ligne, ce qui est de l'échantillonnage non probabiliste. Une marge d'erreur décrit comment une estimation varierait d'un échantillon *aléatoire* à l'autre tirés de la population ; comme aucune sélection aléatoire n'a eu lieu, la quantité qu'elle décrit n'existe pas ici. L'AAPOR déconseille d'en rapporter une pour les échantillons à adhésion volontaire.

Le chiffre n'est pas simplement imprécis — c'est un nombre calculé à partir d'une formule dont les hypothèses n'étaient pas remplies, et il fait paraître le sondage plus digne de confiance que le devis ne l'autorise.

(c) Le parti A devance le parti B de 5 points, et chaque estimation porte environ $\pm 2,3$. Les intervalles — 38,7 à 43,3 contre 33,7 à 38,3 — ne se chevauchent pas tout à fait : 0,35 point les sépare. Sur l'arithmétique, l'avance paraît donc réelle, même si un véritable test de différence exige davantage que la comparaison de deux intervalles (chapitre 13), et une marge aussi près de se toucher ne survivrait pas à un

échantillon un peu plus petit.

Mais la réponse honnête nuance fortement. La marge ne s'applique qu'à l'erreur d'échantillonnage, et cet échantillon a des problèmes de sélection auxquels la marge ne touche pas. Une avance de 5 points dans un panel à adhésion volontaire, deux semaines avant le scrutin, est suggestive, non décisive.

(d) Ce qui manque, par question de la grille. *Q1* : la population cible est « les adultes » d'une province, sans qu'on précise l'admissibilité à voter. *Q2* : la base de sondage n'est pas décrite au-delà de « en ligne » — quel panel, recruté comment. *Q4* : aucun taux de complétion, et rien sur la comparaison entre répondants et non-répondants. *Q5* : aucune taille de sous-groupe, de sorte que les chiffres du parti C et des autres ne peuvent pas être évalués. Absents également : la formulation exacte de la question, le traitement des indécis — souvent une large part deux semaines avant le scrutin — et la pondération appliquée.

Corrigé de l'exercice 5.E5. (a) La population doit être définie assez précisément pour qu'on puisse y échantillonner : non pas « les jeunes du Québec » mais, par exemple, les personnes de 18 à 30 ans résidant au Québec. L'hypothèse doit être directionnelle et testable — « les jeunes adultes qui s'informent par les médias sociaux sont plus susceptibles d'avoir communiqué avec un élu au cours de la dernière année » plutôt que « les médias sociaux influencent l'engagement ».

(b) Des variables classées par type et par échelle, avec des opérationnalisations assez précises pour être mises en œuvre. L'engagement politique est un construit, non une mesure ; un décompte d'activités tirées d'une liste définie est de rapport, l'intérêt autoévalué est ordinal, et avoir voté ou non est nominal.

(c) La méthode doit être justifiée par les trois questions de la §4.1.3, non par la commodité.

(d) C'est la partie pour laquelle le chapitre existe. Cherchez : une base de sondage nommée dont l'écart est repéré ; une méthode précise plutôt qu'« j'interrogerais des gens » ; un *n* visé avec la marge d'erreur qui en découle, calculée ; et des réponses concrètes à la couverture et à la non-réponse — un devis à modes mixtes, des incitatifs, une pondération vers des repères connus.

Concéder un devis non probabiliste et le défendre honnêtement est une réponse plus forte qu'elle n'en a l'air. Proposer un échantillon probabiliste des jeunes adultes du Québec sans budget, c'est un plan qui n'a pas passé l'épreuve de la faisabilité ; proposer un échantillon de convenance sur un campus et dire clairement ce qu'il ne peut pas soutenir, c'est la chose la plus difficile, et la plus utile.

(e) Les enjeux éthiques qui méritent d'être soulevés sont ceux qui tiennent à votre devis plutôt qu'à la recherche en général — par exemple qu'interroger sur l'activité politique peut exposer certains répondants à un risque selon leur statut d'immigration, ou que recruter dans un cours crée une pression à participer.

Corrigé de l'exercice 5.E6. 1. Marges d'erreur approximatives :

$n = 100$	$\pm 10,0 \%$
$n = 400$	$\pm 5,0 \%$
$n = 900$	$\pm 3,3 \%$
$n = 1\,600$	$\pm 2,5 \%$
$n = 2\,500$	$\pm 2,0 \%$
$n = 10\,000$	$\pm 1,0 \%$

2. La courbe descend abruptement au début puis s'aplatit : de grands gains de précision pour de petits échantillons, et des gains progressivement plus faibles à mesure que n grandit. Elle s'approche de zéro sans jamais l'atteindre.

3. Réduire la marge de moitié exige de quadrupler l'échantillon : 1 000 devient 4 000 — environ quatre fois le coût du travail de terrain pour passer d'environ $\pm 3,2$ à environ $\pm 1,6$ points de pourcentage.

Recommander ou non cette dépense dépend de la question de savoir si cette précision change une décision. D'ordinaire elle ne la change pas, et l'argent est mieux investi dans une meilleure base de sondage, un meilleur taux de réponse ou un test plus soigné des questions — parce que ceux-là s'attaquent à l'erreur de couverture et de non-réponse, que la taille de l'échantillon ne peut pas toucher et qui sont d'ordinaire plus grandes que ± 3 points de toute façon. Acheter une précision dont on ne peut rien faire, en laissant le biais intact, est le mauvais échange.

4. L'Île-du-Prince-Édouard représente environ 0,4 % de la population canadienne : une répartition proportionnelle de 1 000 donne donc à peu près 4 répondants. La marge d'erreur serait de $1/\sqrt{4} \times 100 = \pm 50$ points de pourcentage — un chiffre qui ne contient aucune information.

C'est exactement pourquoi les enquêtes nationales suréchantillonnent les petites provinces. L'échantillon devrait être réparti de façon *non proportionnelle*, avec assez de répondants dans chaque province pour soutenir des estimations provinciales, et des pondérations appliquées au moment de calculer les chiffres nationaux.

Corrigé de l'exercice 5.E7. 1. *Approche probabiliste.* Obtenir les listes d'inscription d'un échantillon aléatoire d'établissements postsecondaires, stratifié par province et par type d'établissement, puis tirer un échantillon aléatoire d'étudiants dans chacun — un devis stratifié à plusieurs degrés.

L'obstacle pratique est l'accès : les établissements contrôlent leurs listes d'inscription et sont généralement peu disposés à les communiquer ou à diffuser les enquêtes d'un chercheur, et chacun exigerait une approbation éthique distincte. C'est la raison ordinaire pour laquelle des devis de ce genre ne sont pas tentés par des personnes seules.

2. *Approche non probabiliste.* Un échantillon de convenance ou par quotas :

recruter par des annonces de cours, des associations étudiantes et les médias sociaux, dans l'établissement de la chercheuse et quelques autres, avec des quotas sur l'année d'études et le domaine.

Principale limite : les étudiants qui voient une enquête sur les outils d'IA et y répondent sont vraisemblablement ceux que le sujet intéresse le plus, et l'intérêt est corrélé à l'usage. L'estimation est probablement biaisée vers le haut, et la direction du biais est connaissable même si son ampleur ne l'est pas.

3. *Devis probabiliste* : « On estime que 61 % des étudiants canadiens du postsecondaire ont employé un outil d'intelligence artificielle générative pour un travail de cours, avec une marge d'erreur de $\pm [x]$ points de pourcentage. »

Devis non probabiliste : « Parmi les 800 étudiants qui ont répondu, 61 % ont déclaré avoir employé un outil d'intelligence artificielle générative pour un travail de cours. Comme les répondants se sont autosélectionnés, ce chiffre ne doit pas être lu comme une estimation portant sur les étudiants en général, et surestime probablement l'usage. »

La différence n'est pas une précaution de langage. La seconde phrase affirme quelque chose sur un groupe décrit; la première affirme quelque chose sur une population.

4. Avec un budget modeste, le devis probabiliste n'est pas accessible : le choix honnête est donc le second. Le rapport doit énoncer clairement la méthode de recrutement, refuser de rapporter une marge d'erreur, décrire la direction probable du biais, et présenter le résultat comme établissant que le comportement est répandu plutôt que comme une estimation de sa fréquence.

Corrigé de l'exercice 5.E8. Vos 5 moyennes ne seront celles de personne d'autre. Le raisonnement ci-dessous tient quels que soient les échantillons tirés.

(a)–(b) 5 moyennes d'échantillon, pour des échantillons de 20. Elles différeront les unes des autres — typiquement de plusieurs milliers de dollars — et il faut noter les 5 plutôt que de rapporter celle qui est la plus proche de la vérité.

(c) La vraie moyenne de population se trouve au tableau B.7. La plupart des étudiants constateront que leurs 5 moyennes l'encadrent, certains échantillons s'en écartant d'une marge notable. Quelques-uns les trouveront toutes du même côté, ce qui est instructif en soi : cela arrive, et avec seulement 5 échantillons cela n'indique rien d'anormal.

(d) *L'erreur d'échantillonnage* — la variation entre une estimation d'échantillon et la valeur de population, due au seul fait d'avoir échantillonné plutôt que dénombré. Elle rétrécirait avec des échantillons de 100 parce que l'erreur d'échantillonnage diminue à mesure que n grandit, proportionnellement à $1/\sqrt{n}$. Passer de 20 à 100 multiplie n par 5, de sorte que la dispersion des moyennes d'échantillon devrait diminuer d'environ $\sqrt{5} \approx 2,2$ fois.

L'intérêt de le faire à la main est que le rétrécissement est quelque chose que les étudiants observent au lieu de l'accepter sur parole.

(e) Les chiffres de revenu de *censuspop* ne correspondront pas au revenu médian réel du Canada, et ne doivent pas y correspondre. La préface explique pourquoi : les données du livre préservent la *structure* des relations tout en déformant les niveaux, de sorte que les démonstrations fonctionnent sans qu'aucun chiffre puisse être pris pour un fait sur le Canada. L'étudiante qui signale l'écart et dit correctement qu'il n'a pas d'importance pour ce à quoi le jeu de données sert a compris le devis.