

Corrigé du chapitre 4

Comprendre les données sociales

Mark Paradis, Ph. D.

2 septembre 2026

Voici le corrigé détaillé des exercices de *fin de chapitre* du chapitre 4 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 4 Méthodes de collecte de données

Corrigé de l'exercice 4.E1. *En laboratoire.* Recruter des volontaires pour une séance de recherche, mesurer leur préférence partisane de départ, les assigner au hasard à visionner une véritable publicité de campagne ou une vidéo neutre servant de témoin, puis mesurer de nouveau leur préférence. Données : la préférence avant et après, sur une échelle de cotation (ordinaire, souvent traitée comme un intervalle), plus un choix forcé (nominal). Force : un contrôle complet de l'exposition, de sorte que tout changement est attribuable à la publicité. Limite : les participants regardent attentivement dans une pièce où rien d'autre ne se passe, et ce n'est pas ainsi qu'on croise une publicité de campagne.

Intégrée à une enquête. Dans une enquête nationale en ligne, assigner au hasard la moitié des répondants à voir la publicité insérée dans le questionnaire avant la question de préférence. Données : une mesure de préférence par répondant, plus l'assignation et les données sociodémographiques. Force : un grand échantillon représentatif à peu de frais, et une comparaison entre groupes randomisés. Limite : elle mesure une préférence énoncée quelques instants après l'exposition, laquelle peut ne pas durer.

De terrain. Assigner au hasard des secteurs postaux ou des ménages à recevoir du publipostage de campagne, et mesurer la préférence par une enquête de suivi quelques semaines plus tard — ou, mieux, mesurer la participation et le choix de vote réels là où les registres le permettent. Force : de la vraie publicité, de vrais électeurs, des conditions ordinaires. Limite : aucun moyen de savoir qui a lu le matériel, et une contamination entre secteurs voisins.

Naturelle. Exploiter une frontière où les marchés médiatiques ne coïncident pas avec les circonscriptions, de sorte que certaines reçoivent une forte publicité destinée à une course voisine, pour des raisons étrangères à leur propre politique. Comparer l'évolution des préférences de part et d'autre de la frontière à l'aide de données d'enquête existantes. Force : l'ampleur, le réalisme, et une exposition que personne n'a assignée. Limite : tout l'argument repose sur le fait que les deux côtés de la frontière soient par ailleurs comparables.

Recommandation. Pour un premier résultat, l'expérimentation intégrée à une enquête : elle est abordable, représentative et proprement randomisée. Pour une affirmation sur de vraies campagnes, l'expérimentation de terrain, parce que le résultat de laboratoire peut ne pas survivre au contact de la façon dont les gens croisent réellement la publicité. Décidez à quelle question vous répondez avant de recommander un devis ; la recommandation en découle, et non l'inverse.

Corrigé de l'exercice 4.E2. (a) Des données de traces numériques, et secondaires. Elles ont vraisemblablement été recueillies en fouillant l'interface ou l'archive de Twitter à partir de mots-clés liés au logement, sur une certaine période, en classant

le sentiment de chaque publication à l'aide d'un outil automatique, puis en calculant la part exprimant une inquiétude.

(b) Deux raisons, parmi plusieurs.

La population n'est pas la bonne. Les personnes qui utilisent Twitter ne sont pas les Canadiens — elles sont plus jeunes, plus urbaines, plus scolarisées et plus engagées politiquement. Celles qui publient sur le logement forment un groupe plus étroit encore. Rien dans les données de la plateforme ne peut corriger cela.

Les publications ne sont pas des personnes. Un petit nombre de comptes très actifs produit une part démesurée des publications, et des campagnes organisées amplifient délibérément. Compter les publications pondère les gens par la quantité qu'ils publient ; une estimation de population exige de les pondérer également.

Une troisième raison : les gens publient sur le logement quand ils en sont fâchés, de sorte que la plateforme sélectionne l'inquiétude avant toute mesure.

(c) Une enquête par échantillon probabiliste avec une base de sondage connue — l'Enquête canadienne sur le logement de Statistique Canada, ou un sondage omnibus national dont la méthodologie est publiée. Elle soutient l'inférence à la population parce que chaque membre a une probabilité connue d'être sélectionné, que la formulation des questions est fixe et documentée, et que les résultats peuvent être pondérés d'après des repères de population.

Corrigé de l'exercice 4.E3. (a) VI : le fait que l'école ait reçu ou non le programme de lutte contre l'intimidation. VD : le bien-être des élèves, mesuré à l'aide d'une échelle validée. Variables de contrôle : la taille de l'école, la composition socioéconomique du quartier, et le bien-être mesuré avant le début du programme. La fréquence antérieure de l'intimidation se défend aussi.

(b) Un choix raisonnable est une enquête administrée aux élèves avant et après, dans les écoles participantes et dans des écoles de comparaison.

Que mesure-t-elle ? Un bien-être autodéclaré — ce que les élèves sont disposés à rapporter, et non leur état intérieur.

À quoi peut-elle répondre ? À la question de savoir si le bien-être a davantage changé dans les écoles participantes que dans les écoles de comparaison. Avec une assignation aléatoire des écoles, elle soutient une affirmation causale ; sans elle, une association.

Quels sont ses risques ? La désirabilité sociale, puisque les élèves peuvent ne pas vouloir rapporter leur détresse ; et le contexte scolaire, où les élèves peuvent douter que leurs réponses restent confidentielles.

L'observation structurée des interactions dans la cour d'école est une solution de rechange solide, et combiner les deux serait plus solide encore — chacune échoue là où l'autre tient.

(c) Deux enjeux éthiques. Les *personnes mineures* ne peuvent pas consentir

seules : il faut le consentement parental en plus de l'assentiment de l'élève, et les élèves doivent pouvoir refuser sans conséquence. La *détresse et le dévoilement* : interroger sur l'intimidation peut faire émerger un préjudice en cours, ce qui crée l'obligation d'avoir un protocole de réponse et du soutien disponible, et peut déclencher des obligations de signalement.

Un troisième : priver les écoles de comparaison du programme. Un devis à liste d'attente y répond.

(d) Si les écoles ont été assignées au hasard, il s'agit d'une expérimentation de terrain, qui soutient une affirmation causale sur l'effet du programme sur le bien-être déclaré. Si les écoles ont choisi de l'adopter, le devis est quasi expérimental et ne soutient qu'une association — les écoles qui adoptent un programme de lutte contre l'intimidation diffèrent vraisemblablement de celles qui ne le font pas.

Corrigé de l'exercice 4.E4. (a) Une enquête transversale répétée montre comment les attitudes de la *population* ont changé : la moyenne nationale de 2015 contre celle de 2025. Elle ne peut pas dire si des individus ont changé d'avis, puisque ce ne sont pas les mêmes personnes qui ont répondu chaque fois. Une moyenne nationale stable est compatible avec un renouvellement individuel énorme.

Un panel suit les mêmes répondants, de sorte qu'il peut établir que des individus ont changé, dans quel sens, et dans quel ordre par rapport à d'autres événements de leur vie.

(b) La coupe transversale est moins coûteuse, plus rapide et ne souffre pas d'attrition ; chaque vague peut aussi employer un nouvel échantillon représentatif, de sorte qu'elle demeure représentative de la population du moment. Le panel est coûteux et lent, perd des répondants de façon non aléatoire, et son échantillon vieillit avec lui — après dix ans, il ne représente plus la population adulte, puisque personne de jeune n'y est entré.

(c) Pour la question posée — *l'évolution des attitudes sur dix ans* —, la réponse dépend de ce que « évolution » veut dire. Si ce qui intéresse est de savoir si les vues du pays se sont déplacées, des coupes transversales répétées suffisent et coûtent bien moins cher. Si ce qui intéresse est de savoir si les gens changent d'avis, seul un panel fera l'affaire. Ce qu'il faut faire ici, c'est remarquer l'ambiguïté plutôt que d'en choisir une et d'espérer.

(d) Transversal : l'Étude électorale canadienne, tenue à chaque élection fédérale depuis 1965, ou les vagues de l'Enquête mondiale sur les valeurs. Panel : l'Étude électorale canadienne comporte des volets de panel sur les cycles récents, et l'Étude longitudinale et internationale des adultes de Statistique Canada suit les mêmes individus.

Corrigé de l'exercice 4.E5. (a) Les variables. Le *niveau de revenu du quartier* — de rapport s'il est mesuré par le revenu médian des ménages, ordinal s'il est groupé

en tranches. Le *nombre d'activités parascolaires* — numérique, discret, de rapport. La *réussite scolaire* — de rapport s'il s'agit d'un score à une épreuve ou d'une moyenne, ordinale s'il s'agit d'une tranche de notes. L'unité d'analyse est l'enfant, le quartier étant une variable de contexte, ce qui importe pour ce que les résultats pourront affirmer.

(b) *Primaires* : une enquête auprès des parents ou des élèves dans des écoles échantillonnées parmi les quartiers à revenu élevé et à faible revenu, portant sur la participation, la réussite venant des dossiers scolaires. *Secondaires* : les données de recensement sur le revenu des quartiers, appariées aux dossiers administratifs des commissions scolaires sur l'inscription aux activités offertes par l'école et aux résultats des épreuves standardisées.

(c) *Approche primaire*. Menace à la validité interne : la sélection — les familles qui répondent à une enquête scolaire sont plus susceptibles d'être engagées, ce qui est corrélé à la fois à la participation aux activités et à la réussite. Menace à la validité externe : les écoles échantillonnées peuvent ne pas représenter les écoles québécoises en général.

Approche secondaire. Menace à la validité interne : les dossiers ne saisissent que les activités offertes par l'école, de sorte que les enfants inscrits à des programmes communautaires ou religieux paraissent ne participer à rien — et cette erreur de classement diffère vraisemblablement selon le revenu du quartier. Menace à la validité externe : la couverture se limite au territoire d'une seule commission scolaire et à son offre d'activités particulière.

(d) Pour une première étude exploratoire, l'approche secondaire : elle est rapide, peu coûteuse, couvre beaucoup d'enfants, et établirait si la régularité existe seulement. Pour une évaluation définitive de politique publique, aucune des deux ne suffit — la question est causale et les deux devis sont observationnels. Une évaluation de politique exigerait l'assignation aléatoire d'une intervention donnant accès à des activités, ce qui est faisable d'une manière que l'assignation du revenu d'un quartier ne l'est pas.

Corrigé de l'exercice 4.E6. 1. *À double volet*. La qualité de l'enseignement et l'encadrement pédagogique sont deux choses distinctes. Scindez en deux questions.

2. *À présupposé*, et aussi *orientée*. « Excellents » pose une prémisse que le répondant n'a pas acceptée, et « êtes-vous d'accord » appelle l'accord. Réécriture : « Dans quelle mesure les services de bibliothèque de l'université répondent-ils à vos besoins ? », avec une échelle étiquetée allant de « pas du tout » à « tout à fait ».

3. *Quantificateur vague*. « À quelle fréquence » ne veut pas dire la même chose d'un étudiant à l'autre, et « stress » n'est pas défini. Réécriture : « Au cours des sept derniers jours, combien de jours vous êtes-vous senti assez stressé pour que cela nuise à vos études ? »

4. *Erreur de rappel*, et aussi à *double volet* d'une manière subtile — on demande une moyenne sur trois ans, période pendant laquelle les habitudes d'étude ont presque certainement changé. Réécriture : « Au cours des sept derniers jours, combien d'heures environ avez-vous consacrées à l'étude en dehors des cours ? »

5. *Double négation*. « Ne seriez-vous pas peu disposé » demande de démêler deux négations. Réécriture : « Quelle est la probabilité que vous recommandiez ce programme à un ami ? »

Le biais de réponse : l'acquiescement. Un long questionnaire composé entièrement d'items accord-désaccord invite les répondants à se dire d'accord avec tout, à la fois parce que l'accord est la réponse qui demande le moins d'effort et parce qu'un répondant fatigué fait du satisficing. Les deux se composent : le satisficing fournit le motif, et le format uniforme accord-désaccord fournit l'occasion. La contre-mesure courante consiste à inverser la formulation d'une partie des items, pour qu'un accord uniforme devienne visible comme une contradiction.

Corrigé de l'exercice 4.E7. Votre étude est la vôtre : ce corrigé expose donc ce que chacune des quatre parties doit accomplir, puis reprend le spécimen travaillé de l'exercice 3.E6.

1. *Répondez précisément aux trois questions*. « Une enquête mesure des attitudes » ne répond pas à la question de savoir ce que la méthode mesure. Ce qui a été consigné, auprès de qui, et par quelle procédure, oui.

2. *Cherchez les données secondaires, ne les supposez pas*. Nommez ce que vous avez cherché — des programmes précis, des listes de variables précises — et ce que vous avez trouvé, y compris « rien qui convienne, parce que l'item n'a plus été posé après 2018 ». Un résultat négatif rapporté honnêtement répond pleinement à cette partie, et c'est l'aboutissement réel le plus fréquent.

3. *Trouvez un défaut véritable à chaque question*. Il est facile de rédiger trois questions et de les déclarer bonnes. Les défauts à chercher sont ceux de la §4.2.3 — un quantificateur vague, une prémisse implicite, une exigence de rappel.

4. *Nommez le groupe manquant et sa conséquence*. Non pas « certaines personnes pourraient ne pas répondre », mais lesquelles, et ce que leur absence fait à votre conclusion.

Le spécimen, poursuivi. L'étude du chapitre 3 demandait si le travail rémunéré pendant la session abaisse la moyenne des cégépiennes et des cégépiens.

(1) *Méthode* : une enquête en ligne autoadministrée auprès des étudiants, appariée avec leur consentement aux dossiers de notes de l'établissement. *Que mesure-t-elle ?* Les heures rémunérées hebdomadaires déclarées, et les notes telles que le registraire les a consignées — une autodéclaration et une mesure administrative, ce qui est le bon partage, puisque la note est la variable qu'un étudiant a le plus de raisons de mal déclarer. *À quoi peut-elle répondre ?* À la question de savoir si les étudiants qui

travaillent plus d'heures ont une moyenne plus faible à la même session. Pas à celle de savoir si le travail en est la cause. *Quels sont ses risques ?* L'erreur de rappel sur des heures qui varient d'une semaine à l'autre ; et la sélection, puisque les étudiants sous pression scolaire sont les moins susceptibles de remplir une enquête.

(2) *Recherche de données secondaires* : l'Enquête sur la population active de Statistique Canada porte les heures travaillées par âge et par statut d'étudiant, mais sans lien avec les notes. Le *Canadian Undergraduate Survey Consortium* interroge sur le travail et le temps d'étude, mais couvre les universités plutôt que les cégeps et rapporte les notes par tranche autodéclarée. Résultat : rien de convenable n'existe au niveau voulu, ce qui justifie de recueillir des données primaires — et le dire fait partie de la réponse.

(3) *Trois questions en projet, chacune avec son défaut.*

1. « Combien d'heures par semaine travaillez-vous ? » — *période de référence vague*. Un étudiant à l'horaire irrégulier ne peut pas répondre. Réparation : « Au cours des sept derniers jours, combien d'heures avez-vous travaillé contre rémunération ? »
2. « Le travail nuit-il à vos études ? » — *prémisse implicite et cadrage orienté*. La question présume qu'il y a nuisance et appelle l'accord. Réparation : demander les heures d'étude, et laisser l'association répondre à la question à la place du répondant.
3. « Combien d'heures étudiez-vous chaque semaine la session dernière ? » — *exigence de rappel* sur quatre mois. Réparation : la même fenêtre de sept jours que pour l'item sur le travail, pour que les deux soient comparables.

(4) *Le groupe manquant* : les étudiants qui ont abandonné en cours de session, lesquels sont disproportionnellement ceux qui travaillaient le plus d'heures et réussissaient le moins bien. Ils sont absents de l'enquête comme du fichier de notes. Leur absence retire les cas où l'effet devrait être le plus grand, de sorte que l'étude *sous-estime* l'association — et le sens de ce biais doit être énoncé, parce qu'un lecteur qui suppose que l'attrition ne fait qu'ajouter du bruit lira le résultat comme une preuve plus faible plutôt que comme un plancher.

Corrigé de l'exercice 4.E8. (a) Une expérimentation intégrée à une enquête. Les répondants ont été assignés au hasard à différentes versions de la même question à l'intérieur d'un questionnaire, ce qui en est la caractéristique déterminante.

(b) Le biais de désirabilité sociale, et la non-réponse qu'il entraîne. Le revenu y est sujet pour deux raisons qui se composent : déclarer son revenu appelle le jugement dans les deux sens, et les hauts revenus ont en outre un motif particulier — la discrétion sur la fortune, et la conscience que le chiffre pourrait servir à les situer.

Il en résulte que la non-réponse sur le revenu se concentre au sommet, ce qui biaise vers le bas le revenu moyen estimé. Notez que c'est systématique, de sorte qu'un plus grand échantillon aggrave la chose au lieu de l'améliorer.

(c) La réponse dépend de l'enquête que vous aurez choisie. L'Étude électorale canadienne et l'Enquête mondiale sur les valeurs portent toutes deux des items de confiance. Attendez-vous à ce que la formulation réelle soit plus précise que vous ne l'aviez prévu — nommant une institution particulière, ou précisant une échelle à quatre points plutôt qu'à onze — et attendez-vous à ce que l'objet de la confiance soit plus étroit que « le gouvernement ».

Cet écart est tout le propos. L'item réel mesure quelque chose d'un peu différent de ce que le nom du concept suggère, ce qui est le problème d'opérationnalisation du chapitre 2 apparaissant dans un instrument publié plutôt que dans un exemple de manuel.