

Corrigé du chapitre 10

Comprendre les données sociales

Mark Paradis, Ph. D.

2 septembre 2026

Voici le corrigé détaillé des exercices de *fin de chapitre* du chapitre 10 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 10 Statistiques descriptives bivariées

Corrigé de l'exercice 10.E1. (a) *Catégorielle × continue*. Le domaine d'études a quatre catégories non ordonnées ; l'utilisation des médias sociaux est continue et mesurée au niveau de rapport.

(b) Les outils de la §10.4 : calculer une statistique sommaire — la moyenne, ou la médiane si la distribution est asymétrique — à l'intérieur de chaque domaine, rapporter les tailles de groupe à côté, et présenter la comparaison avec des diagrammes en boîte.

Les diagrammes en boîte valent mieux ici qu'un diagramme à barres des moyennes, parce qu'ils montrent la dispersion à l'intérieur de chaque domaine en plus du centre. Quatre moyennes de groupe qui diffèrent d'environ une heure disent peu de chose si la dispersion interne est de trois heures.

(c) Un tableau sommaire par groupe doit porter, pour chaque domaine, la moyenne, un écart type et la taille du groupe. Par exemple :

Domaine	Moyenne	ÉT	<i>n</i>
Sciences et génie	3,2	1,4	62
Sciences sociales	4,1	1,6	54
Lettres et humanités	3,8	1,5	41
Autre	4,5	1,8	43
Tous	3,8	1,6	200

La configuration. L'utilisation quotidienne moyenne des médias sociaux va de 3,2 heures chez les étudiants en sciences et génie à 4,5 chez ceux des autres domaines — un écart de 1,3 heure. Les sciences sociales et les lettres se situent entre les deux.

La ligne « Tous » n'est pas la moyenne des quatre moyennes de groupe. C'est la moyenne des 200 étudiants, qui pondère chaque domaine par sa taille :

$$\frac{3,2(62) + 4,1(54) + 3,8(41) + 4,5(43)}{200} = 3,8$$

Faire la moyenne des quatre moyennes comme si les groupes étaient de même taille donne 3,9, ce qui est une autre quantité et répond à une autre question.

La mise en garde qui doit l'accompagner. Avec des écarts types allant de 1,4 à 1,8 heures, les différences entre domaines sont plus petites que la variation à l'intérieur de chacun. Les distributions se recouvrent largement, et un étudiant en sciences et génie tiré au hasard peut très bien utiliser davantage les médias sociaux qu'un étudiant en sciences sociales tiré au hasard.

Dites-le explicitement. Les tailles de groupe sont d'ailleurs assez modestes pour que ces moyennes soient elles-mêmes imprécises — ce à quoi répondent les méthodes de la partie III.

Corrigé de l'exercice 10.E2. (a) Diviser chaque cellule par le total de sa ligne.

Âge	N'a pas voté	A voté	Total
18–34	41,3	58,7	100,0
35–54	37,0	63,0	100,0
55+	29,3	70,7	100,0
Tous	36,0	64,0	100,0

Par exemple, pour le groupe le plus jeune : $88/150 = 58,7$ pour cent.

(b) Oui, il y a des indices d'une association. La participation augmente régulièrement avec l'âge : 58,7 pour cent chez les 18–34 ans, 63,0 chez les 35–54 ans et 70,7 chez les 55 ans et plus — un écart de 12 points de pourcentage entre le groupe le plus jeune et le plus âgé.

La configuration est monotone, ce qui renforce l'impression que quelque chose de systématique se produit plutôt qu'une fluctuation fortuite dans un groupe.

(c) Sous l'hypothèse d'indépendance, chaque groupe d'âge voterait au taux d'ensemble de 64,0 pour cent, de sorte que les trois lignes afficheraient toutes 36,0 et 64,0.

De façon équivalente, les effectifs attendus seraient :

$$18-34 : \frac{150 \times 320}{500} = 96 \text{ votants}$$

$$35-54 : \frac{200 \times 320}{500} = 128 \text{ votants}$$

$$55+ : \frac{150 \times 320}{500} = 96 \text{ votants}$$

contre 88, 126 et 106 observés. Les écarts sont modestes dans le groupe du milieu et plus grands aux deux extrémités, ce qui est la configuration monotone revue en effectifs.

Corrigé de l'exercice 10.E3. (a) Une corrélation de $r = 0,61$ indique une association linéaire positive modérément forte : les pays où l'inégalité de revenu est plus élevée tendent à avoir des taux d'homicide plus élevés.

En l'élevant au carré,

$$r^2 = 0,61^2 = 0,372$$

donc environ 37 pour cent de la variation des taux d'homicide entre ces pays est expliqué par la relation linéaire avec le coefficient de Gini — ce qui en laisse 63 pour cent inexpliqué.

(b) Non, la conclusion causale n'est pas justifiée, pour les raisons habituelles et pour une raison propre à ces données.

La causalité inverse est plausible : des taux d'homicide élevés peuvent déprimer l'activité économique dans les quartiers les plus pauvres et creuser l'inégalité.

Les variables cachées abondent. La capacité de l'État, le niveau de pauvreté, la solidité du système de justice, les marchés de la drogue et l'histoire des conflits agissent tous plausiblement à la fois sur l'inégalité et sur l'homicide.

L'unité d'analyse est le pays, de sorte qu'il s'agit d'une relation entre agrégats. En conclure quoi que ce soit à propos des individus — qu'un traitement inégal rend une personne plus susceptible de tuer — serait l'erreur écologique (§5.1.2).

(c) Plusieurs choses aideraient. Un nuage de points, d'abord, pour vérifier que la relation est à peu près linéaire et qu'elle n'est pas portée par quelques pays extrêmes — avec seulement 40 cas, deux ou trois valeurs aberrantes pourraient produire ce coefficient à elles seules.

Au-delà : des mesures dans le temps à l'intérieur des pays, pour que des changements d'inégalité puissent être reliés à des changements ultérieurs de l'homicide ; des données sur les variables cachées candidates, pour que leur contribution puisse être examinée ; et, idéalement, une source de variation de l'inégalité indépendante des autres facteurs — un changement de politique touchant certains pays et pas d'autres.

C'est l'une des relations les plus étudiées et les plus contestées de la criminologie comparée, ce qui est en soi instructif sur la difficulté de la question causale.

Corrigé de l'exercice 10.E4. (a) Chaque heure hebdomadaire d'exercice supplémentaire est associée à un score de santé autodéclaré supérieur de 3,7 points, en moyenne.

Notez la formulation. Les données sont observationnelles, donc « associée à » est défendable et « augmente » ne l'est pas — les gens en meilleure santé font peut-être plus d'exercice, plutôt que l'inverse.

(b)

$$\hat{Y} = 52,4 + 3,7(5) = 52,4 + 18,5 = 70,9$$

Le score de santé prédit est de 70,9.

(c) Le résidu est la valeur observée moins la valeur prédite :

$$e = Y - \hat{Y} = 74 - 70,9 = 3,1$$

Le score de santé de cette personne est supérieur de 3,1 points à ce que la droite de régression prédit pour quelqu'un qui fait 5 heures d'exercice par semaine. Un résidu positif signifie que le cas se situe au-dessus de la droite.

(d) C'est une *extrapolation*, et elle produit ici une valeur impossible :

$$\hat{Y} = 52,4 + 3,7(30) = 52,4 + 111 = 163,4$$

sur une échelle qui va de 0 à 100.

L'impossibilité est utile — elle annonce le problème immédiatement, à la manière décrite à la §9.5.3. Mais le problème de fond existerait même si la prédiction était tombée à l'intérieur de l'échelle. 30 heures d'exercice par semaine est bien au-delà de l'étendue que les données couvraient, et rien dans les données n'appuie l'hypothèse que la relation linéaire se poursuit aussi loin.

Il y a aussi une raison de fond de penser qu'elle ne se poursuit pas. Les scores de santé sont bornés vers le haut, de sorte que la relation doit s'aplatir à mesure que les scores approchent du plafond — une droite ne peut pas décrire un résultat borné sur une large étendue de X .

Corrigé de l'exercice 10.E5. *Le paradoxe de Simpson.* Une relation visible dans des données combinées peut se renverser quand on sépare les données en sous-groupes. Ce qui obtient de moins bons résultats globalement peut en obtenir de meilleurs dans chaque sous-groupe, parce que les groupes comparés contiennent des mélanges de cas différents.

La situation. L'explication la plus probable est que les deux hôpitaux traitent des patients de types différents. Si l'hôpital dont le taux de survie global est le plus bas prend une part plus grande de cas graves — comme le ferait un centre de référence spécialisé ou une grande unité de traumatologie — alors son chiffre global est tiré vers le bas par la composition de sa clientèle plutôt que par la qualité de ses soins.

La gravité des cas est ici la variable cachée : elle est liée à l'hôpital que fréquente un patient, et elle est liée à la survie.

Comment enquêter. Obtenir les taux de survie *stratifiés selon la gravité des cas*, avec la classification que permettent les dossiers cliniques — diagnostic, gravité à l'admission, comorbidités, urgence de l'opération. Comparer les hôpitaux à l'intérieur de chaque strate.

Trois résultats sont possibles, et chacun signifie autre chose. Si l'hôpital au taux global le plus bas fait mieux dans chaque strate, la comparaison globale induisait en erreur et la conclusion du journaliste est fausse. S'il fait moins bien dans les strates aussi, la comparaison globale pointait vers quelque chose de réel. Si la configuration varie d'une strate à l'autre, la réponse honnête est qu'aucun des deux hôpitaux n'est simplement meilleur.

Une mise en garde à ajouter. La stratification n'ajuste que pour ce qui a été mesuré. Si les hôpitaux diffèrent d'une manière non consignée qui agit sur la survie — l'âge des patients, leur situation socioéconomique, le temps d'attente avant l'admission — la comparaison stratifiée reste confondue. Le cas des admissions à Berkeley, à la §10.7.3, est instructif ici : la désagrégation a déplacé la question plutôt que de la trancher.

Corrigé de l'exercice 10.E6. 1. Les trois appariements :

Scolarité × *bénévolat* : catégorielle × catégorielle. Tableau de contingence avec pourcentages en ligne, et diagramme à barres groupées ou empilées (§10.2).

Scolarité × *heures de soin* : catégorielle × continue. Sommaires par groupe avec diagrammes en boîte (§10.4).

Bénévolat × *heures de soin* : catégorielle × continue, avec deux groupes seulement. Comparer les deux sommaires de groupe ; diagrammes en boîte de nouveau.

Notez qu'aucun appariement ici n'est continue × continue, de sorte que rien dans ce fichier n'appelle une corrélation ou une régression.

2. Les pourcentages doivent se calculer à l'intérieur des groupes de *scolarité* — en ligne si la *scolarité* est la variable en ligne — parce que la *scolarité* est la variable indépendante présumée. L'énoncé qui en résulte est « X pour cent des diplômés ont fait du bénévolat contre Y pour cent des non-diplômés », ce qui est ce que demande la question de recherche.

Les pourcentages en colonne rapporteraient la composition scolaire des bénévoles, ce qui reflète surtout combien de personnes de chaque niveau de *scolarité* se trouvent dans l'échantillon.

3. Les heures de soin sont probablement asymétriques à droite — la plupart des répondants en font peu ou pas, une minorité en fait beaucoup — de sorte que la *médiane* est le sommaire le plus sûr, rapportée avec l'écart interquartile.

Ce qu'il faut vérifier d'abord, c'est la distribution elle-même : un histogramme ou un diagramme en boîte à l'intérieur de chaque groupe. Cherchez en particulier un pic important à zéro, qui signifierait que la variable est en réalité un mélange de deux populations (ceux qui font du travail de soin et ceux qui n'en font pas) et qu'un sommaire unique ne décrit ni l'une ni l'autre (§7.4.5).

4. Plusieurs variables cachées conviendraient. L'*âge* est la candidate la plus forte : les cohortes plus âgées ont une *scolarité* moyenne plus faible et sont aussi plus susceptibles d'avoir des responsabilités de soin envers un conjoint ou un parent — l'*âge* pourrait donc produire les deux configurations sans que la *scolarité* agisse sur l'une ou l'autre. Le *revenu* et la *situation d'activité* sont également défendables, puisque tous deux sont liés à la *scolarité* et au temps disponible pour du travail non rémunéré.

Comment enquêter : examiner les relations à l'intérieur de tranches d'*âge*. Si le gradient scolaire du bénévolat persiste à tout *âge*, l'*âge* n'était pas l'explication ; s'il disparaît, il l'était.

Notez la limite : cela ajuste pour l'*âge* parce que l'*âge* a été mesuré. Toute variable cachée absente du fichier reste intacte, et l'enquête contient trois variables.

Corrigé de l'exercice 10.E7. Votre réponse emploiera d'autres variables que celles ci-dessous. Ce qui suit est la même démarche en cinq étapes, menée sur un choix

particulier, pour que vous puissiez comparer la forme de votre réponse à celle d'une réponse complète.

Le spécimen emploie le genre et la participation électorale pour le tableau croisé, le revenu selon le genre pour la comparaison de groupes, et l'âge contre le revenu pour le nuage de points. Les répondants déclarant un autre genre — 11 sur 477 — sont écartés de la comparaison à deux groupes, ce qui en laisse 466. Dites-le quand vous faites cela. Une catégorie écartée en silence est une décision que le lecteur ne peut pas voir.

1. *Le tableau croisé.* Le genre est la variable indépendante présumée, de sorte que les pourcentages se calculent à l'intérieur du genre et que la comparaison se fait entre les lignes. Sur 252 femmes, 160 ont voté — 63,5 pour cent. Sur 214 hommes, 153 ont voté — 71,5 pour cent.

Rapportez les effectifs autant que les pourcentages, et indiquez quelle variable vous avez traitée comme indépendante. Un pourcentage sans son dénominateur ne peut pas être jugé, et un lecteur qui ignore dans quel sens vous avez calculé ne peut pas distinguer votre constat de son image inversée.

2. *Les effectifs attendus.* Sous l'hypothèse d'indépendance, chaque groupe vote au taux d'ensemble, donc pour les femmes :

$$E = \frac{252 \times 313}{466} = 169,3$$

contre 160 observées. Moins de femmes ont voté que l'indépendance ne le prédit, et donc davantage d'hommes.

Rapportez à la fois la taille et le sens de l'écart. Qu'il y ait un écart n'est pas un constat ; de quel côté il va, et de combien par rapport à l'effectif, en est un.

3. *La comparaison de groupes.* Le revenu est asymétrique à droite, donc rapportez la médiane et l'écart interquartile plutôt que la moyenne et l'écart type (§7.5.2). Le revenu médian est de 68,7 milliers de \$ pour les femmes contre 92,7 milliers pour les hommes, avec des écarts interquartiles de 50,8 et 64,1 milliers.

Un centre et une dispersion sont tous deux requis. Les seules moyennes de groupe reproduisent exactement l'erreur que les diagrammes en boîte côte à côte de la §10.4.2 existent pour empêcher : les dispersions sont ici plus grandes que l'écart entre les centres, de sorte que les distributions se recouvrent largement et que beaucoup de femmes gagnent plus que beaucoup d'hommes.

4. *Le nuage de points, avant le coefficient.* Décrivez d'abord ce que vous voyez, dans le vocabulaire de la §10.5 — direction, force, forme — et calculez r seulement ensuite. L'âge contre le revenu donne un nuage légèrement incliné vers le haut : positif, très faible, sans forme nette. Le coefficient vaut $r = 0,15$, ce qui concorde.

Le résultat instructif est une divergence, et il faut presque en espérer une. Une relation qui semblait modérée peut donner un petit r parce qu'elle est curviligne ; un nuage sans forme peut en donner un grand parce qu'une seule valeur aberrante tire la droite. Si votre lecture et votre coefficient divergent, rapportez la divergence et dites laquelle de ces deux choses l'a produite — c'est une meilleure réponse qu'une concordance.

5. *La prudence, énoncée précisément.* « Il pourrait y avoir une variable confondante » ne dit rien. Nommez la variable cachée et le mécanisme : ici, la scolarité est associée à la fois à la composition selon le genre et au revenu dans ce fichier, de sorte qu'une partie de l'écart de revenu peut être un écart de scolarité (§10.7.2).

Deux autres prudences valent la peine d'être cherchées. L'erreur écologique s'applique partout où la comparaison porte sur des groupes et l'affirmation sur des individus. Et socsurvey est simulé : ses relations y ont été construites, de sorte qu'un constat tiré d'ici est une preuve à propos de la simulation plutôt qu'à propos du Canada.

Corrigé de l'exercice 10.E8. 1. Les chiffres les plus récents de Statistique Canada au moment d'écrire ces lignes portent sur 2023 : une espérance de vie à la naissance de **84,0 ans pour les femmes et de 79,6 pour les hommes** (STATISTICS CANADA 2026). Cette publication rapporte les deux séparément plutôt que de donner un chiffre combiné, de sorte que citer les deux — et dire pourquoi — est la bonne réponse et non une échappatoire. Vos chiffres peuvent différer si une année plus récente a été publiée ; indiquez l'année que vous avez employée.

2. worldsim donne au Canada 85,0 ans. C'est au-dessus du chiffre réel pour les femmes et environ cinq ans au-dessus du chiffre pour les hommes. La simulation est optimiste, et pas d'un peu : aucun pays n'a atteint une espérance de vie combinée de 85,0 ans.

3. La vérification la plus utile est celle de l'*étendue*. La droite a été ajustée à cinq pays dont la scolarité moyenne va de 6 à 13 ans, et sa pente décrit la différence entre des pays situés à l'intérieur de cette étendue. Des données nationales réelles permettraient de demander si la relation est à peu près droite sur cette étendue — la forte courbure que le chapitre montre pour le revenu et l'espérance de vie est un avertissement qu'elle pourrait ne pas l'être.

Deux autres vérifications méritent d'être nommées. Cinq pays ne peuvent appuyer aucune estimation de pente un peu précise, et ils ont été *choisis* plutôt que tirés au hasard (§10.5). Et une pente ajustée entre pays à un moment donné n'est pas un énoncé sur ce qui arrive quand la scolarité d'un pays augmente.

4. *Corrélation n'est pas causalité* (§10.7.1) et les *variables cachées* (§10.7.2). Le revenu national est le candidat évident : les pays riches ont à la fois plus de scolarité et de meilleurs systèmes de santé, de sorte que l'association peut se produire sans

que la scolarité agisse le moins du monde sur la longévité.

L'*erreur écologique* est une troisième réponse défendable. Les cas sont ici des pays, et la conclusion du camarade porte sur des personnes.

À dire clairement, parce que c'est tout le propos de l'exercice : cette comparaison vous apprend qu'une conclusion tirée de `worldsim` est une conclusion à propos de la simulation. Elle ne dit rien contre la *méthode* — le nuage de points, la corrélation et la droite se font de la même façon sur des données réelles, et c'est ce qu'enseignent les §10.5 et §10.6.