

Corrigé du chapitre 7

Comprendre les données sociales

Mark Paradis, Ph. D.

2 septembre 2026

Voici le corrigé détaillé des exercices de *fin de chapitre* du chapitre 7 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 7 Mesures de tendance centrale

Corrigé de l'exercice 7.E1. (a) *Moyenne.* Additionnez les 17 valeurs :

$$\begin{aligned}\sum x_i &= 8 + 11 + 12 + 14 + 15 + 15 + 18 + 20 + 22 \\ &\quad + 24 + 25 + 28 + 32 + 40 + 44 + 72 + 95 = 495 \\ \bar{x} &= \frac{495}{17} = 29,12 \text{ minutes}\end{aligned}$$

Médiane. Les valeurs sont déjà ordonnées. Avec $n = 17$,

$$\text{position} = \frac{17 + 1}{2} = 9$$

La neuvième valeur est 22, donc la médiane est de 22 minutes.

Moyenne tronquée. Écartez 8 et 95, ce qui laisse 15 valeurs :

$$\begin{aligned}\sum &= 11 + 12 + 14 + 15 + 15 + 18 + 20 + 22 + 24 \\ &\quad + 25 + 28 + 32 + 40 + 44 + 72 = 392 \\ \bar{x}_{\text{tronquée}} &= \frac{392}{15} = 26,13 \text{ minutes}\end{aligned}$$

Notez la convention d'arrondi : conservez toute la précision jusqu'au bout et n'arrondissez qu'à la fin (voir l'annexe A, §A.9).

(b) L'ordre est moyenne (29,12) > moyenne tronquée (26,13) > médiane (22).

C'est la signature d'une distribution étalée vers la droite. La moyenne se situe le plus haut parce qu'elle est tirée par la longue queue supérieure — les déplacements de 72 et de 95 minutes. La moyenne tronquée tombe entre les deux parce qu'écartez une valeur à chaque extrémité retire le cas le plus extrême mais laisse le reste de la queue. La médiane est la plus basse parce qu'elle ignore entièrement la grandeur.

(c) La *médiane*, 22 minutes.

En appliquant la grille : le temps de déplacement est une variable de rapport, donc les trois mesures sont disponibles. La distribution est nettement étalée vers la droite, avec deux valeurs qui font plus du double de la suivante. Elle n'est pas bimodale. L'étape 4 de la grille pointe donc vers la médiane. Et le but est ici de décrire un déplacement typique plutôt que de calculer un total, de sorte que rien ne vient renverser ce choix.

Une réponse complète rapporte la moyenne à côté et signale l'écart.

(d) Deux phrases défendables : « La moitié des travailleuses et des travailleurs de la ville mettent 22 minutes ou moins pour se rendre au travail. Un petit nombre affrontent des trajets bien plus longs — deux des dix-sept personnes interrogées voyagent plus d'une heure —, ce qui tire la moyenne de la ville au-dessus de ce que

la plupart des gens vivent réellement. »

Ce qui fait fonctionner deux phrases de ce genre, c'est qu'elles rendent l'asymétrie sans vocabulaire technique et ne présentent jamais un chiffre unique comme étant *le* temps de déplacement. Si votre version nomme un seul nombre et s'arrête, relisez-la en vous demandant ce qu'un lecteur imaginerait.

Corrigé de l'exercice 7.E2. (a) Pour les deux années, c'est la *médiane* qui décrit le mieux un ménage typique. La moyenne dépasse la médiane chaque année — de 14 000 \$ en 2015 et de 30 000 \$ en 2025 —, donc le revenu est étalé vers la droite et la moyenne est tirée au-dessus du milieu de la distribution.

(b) Hausses en pourcentage :

$$\text{moyenne : } \frac{104 - 82}{82} \times 100 = 26,8 \%$$

$$\text{médiane : } \frac{74 - 68}{68} \times 100 = 8,8 \%$$

La moyenne a crû trois fois plus vite que la médiane. Comme la médiane suit le ménage du milieu et que la moyenne est tirée par la queue supérieure, cela indique que la croissance du revenu s'est concentrée au sommet. Les ménages du milieu ont connu des gains modestes ; les ménages bien au-dessus du milieu en ont connu de bien plus grands.

Notez aussi que l'écart entre la moyenne et la médiane s'est élargi de 14 000 \$ à 30 000 \$, ce qui est le même constat énoncé une seconde fois.

(c) Les deux affirmations emploient un chiffre exact et les deux sont incomplètes.

L'affirmation de la politicienne, selon laquelle le niveau de vie a sensiblement augmenté, repose sur un chiffre qu'une minorité de ménages a réellement vécu. Une hausse de 26,8 pour cent de la moyenne est compatible avec le fait que la plupart des ménages n'aient presque rien vu.

L'affirmation du critique, selon laquelle le niveau de vie a « à peine bougé », minimise ce qui est arrivé aux ménages qui ont bel et bien gagné, et 8,8 pour cent sur une décennie n'est pas rien — même s'il vaut la peine de demander si ces chiffres sont corrigés de l'inflation, puisqu'une croissance nominale de 8,8 pour cent en dix ans serait un recul réel.

L'énoncé exact emploie les deux : les revenus ont monté dans toute la distribution, et ils ont bien plus monté au sommet qu'au milieu.

(d) Plusieurs choses. Si les chiffres sont corrigés de l'inflation — c'est l'omission la plus importante, puisqu'elle pourrait renverser entièrement la conclusion. Le revenu à plusieurs centiles, et pas seulement au milieu, pour voir où les gains ont atterri. Si la taille des ménages a changé, puisqu'un chiffre par ménage réparti entre plus ou moins de personnes ne veut pas dire la même chose. Et le coût local du

logement, qui peut absorber entièrement un gain de revenu.

Corrigé de l'exercice 7.E3. 1. La distribution est *étalée vers la droite*. La moyenne dépasse largement la médiane, ce qui se produit lorsqu'une queue de valeurs élevées tire la moyenne vers le haut en laissant le cas du milieu là où il est.

2. Le *pays de naissance* seulement. Il est nominal, donc sans ordre ni arithmétique. L'âge et le revenu sont de rapport et admettent les trois mesures ; le niveau de scolarité est ordinal et admet le mode et la médiane.

3. Que la distribution contient des valeurs extrêmes qui influencent la moyenne. Tronquer ne change la réponse que lorsque les valeurs retirées ne ressemblaient pas au reste — une différence sensible est donc un diagnostic de valeurs aberrantes ou d'asymétrie marquée.

4. La moyenne doit être *pondérée*, chaque répondant étant pondéré par le nombre de personnes qu'il représente dans la population. Une moyenne non pondérée décrirait l'échantillon, où les répondants ruraux sont surreprésentés, plutôt que le pays.

5. Parce que la moyenne comme la médiane tendront à tomber dans le creux entre les deux sommets, décrivant un cas qui ne se présente à peu près pas. Les deux groupes devraient être rapportés séparément.

6. Quand chaque valeur ne se présente qu'une seule fois. C'est fréquent avec les variables continues mesurées avec précision — un revenu au dollar près, des temps à la milliseconde — et la réponse consiste habituellement à grouper les données et à rapporter la classe modale à la place.

Corrigé de l'exercice 7.E4. 1. 3, 7, 7, 9, 14. Somme = 40, $n = 5$.

$$\bar{x} = \frac{40}{5} = 8,0$$

Position = $(5 + 1)/2 = 3$, donc la médiane est la troisième valeur : 7.

2. 12, 15, 18, 21. Somme = 66, $n = 4$.

$$\bar{x} = \frac{66}{4} = 16,5$$

n pair, donc la médiane est la moyenne des positions 2 et 3 :

$$\frac{15 + 18}{2} = 16,5$$

3. 2, 4, 4, 4, 9, 11. Somme = 34, $n = 6$.

$$\bar{x} = \frac{34}{6} = 5,67$$

n pair, donc la médiane est la moyenne des positions 3 et 4 — toutes deux égales à 4 :

$$\frac{4 + 4}{2} = 4$$

4. 5, 5, 8, 10, 10, 12, 40. Somme = 90, $n = 7$.

$$\bar{x} = \frac{90}{7} = 12,86$$

Position = $(7 + 1)/2 = 4$, donc la médiane est la quatrième valeur : 10.

(e) Le jeu de données 4, avec un écart de $12,86 - 10 = 2,86$. C'est la valeur isolée de 40 qui le produit : elle se situe bien au-dessus des six autres et contribue démesurément à la somme, tout en n'occupant que le dernier rang et en ne déplaçant donc pas du tout la médiane.

Le jeu de données 3 vient loin derrière, à 1,67, pour la même raison sous une forme atténuée.

(f) Le jeu de données 1 a un mode de 7. Le jeu de données 3 a un mode de 4. Le jeu de données 4 est *bimodal*, 5 et 10 apparaissant chacun deux fois. Le jeu de données 2 n'a aucun mode, puisque chaque valeur ne se présente qu'une fois.

Corrigé de l'exercice 7.E5. (a) Fortement étalée vers la droite, avec un pic énorme à zéro. La plupart des répondants ne déclarent aucun jour d'insécurité alimentaire, et un petit nombre déclarent des valeurs élevées allant jusqu'à 31.

Ce qui rend la moyenne trompeuse n'est pas seulement l'asymétrie mais l'*inflation de zéros* : la distribution est en réalité un mélange de deux populations — la grande majorité qui ne vit aucune insécurité alimentaire, et une minorité qui en vit, parfois gravement. Une moyenne calculée sur les deux ne décrit ni l'une ni l'autre.

(b) Elle devrait *se limiter aux répondants déclarant au moins une journée*. La question porte sur l'expérience de celles et ceux qui l'ont vécue, et inclure les zéros répond à une autre question — cela produirait un chiffre déterminé presque entièrement par le nombre de personnes non touchées.

Pour l'échantillon restreint, la *médiane* est la meilleure mesure, puisque la distribution parmi les répondants touchés restera étalée vers la droite. Rapporter les deux, avec le nombre de répondants touchés, est plus solide encore.

(c) Pour l'échantillon complet, le chiffre approprié est la *moyenne*, et la raison en est que le but a changé. Un chiffre unique rendant compte de l'ampleur de l'insécurité alimentaire dans la population est une question de total : la moyenne multipliée par la population donne le nombre total de personnes-jours d'insécurité alimentaire, ce qui est exactement ce qu'une question de planification des ressources demande.

La médiane de l'échantillon complet serait 0, ce qui est vrai et inutile.

C'est le cas le plus net du chapitre où le but l'emporte sur la forme. La même variable, la même asymétrie, et deux réponses justes différentes parce que les questions diffèrent.

(d) Cet usage est indéfendable, même si le nombre est exact.

Une moyenne de, disons, 0,4 jour sur l'échantillon complet a l'air négligeable, et elle est arithmétiquement exacte. Mais c'est une moyenne sur une population dont la plus grande partie n'est pas touchée du tout, et elle masque qu'une minorité vit une insécurité alimentaire grave et durable. La présenter comme la preuve que le problème est minime revient à employer un chiffre dont la petitesse est produite par la taille de la majorité non touchée plutôt que par la douceur du problème.

Le compte rendu honnête donne la *prévalence* — quelle proportion a vécu une quelconque insécurité alimentaire — à côté d'une mesure de gravité parmi les personnes touchées. Aucun des deux chiffres ne décrit la situation à lui seul, et la moyenne seule la décrit activement mal.

Corrigé de l'exercice 7.E6. (a) Moyenne non pondérée des cinq notes :

$$\frac{3,7 + 3,0 + 2,3 + 4,0 + 3,3}{5} = \frac{16,3}{5} = 3,26$$

(b) Cote pondérée. Multipliez chaque note par ses crédits :

$$\begin{aligned}\sum w_i x_i &= (4)(3,7) + (3)(3,0) + (3)(2,3) \\ &\quad + (1)(4,0) + (5)(3,3) \\ &= 14,8 + 9,0 + 6,9 + 4,0 + 16,5 = 51,2 \\ \sum w_i &= 4 + 3 + 3 + 1 + 5 = 16 \\ \bar{x}_w &= \frac{51,2}{16} = 3,20\end{aligned}$$

(c) La cote pondérée est plus basse, 3,20 contre 3,26.

La moyenne non pondérée donne au séminaire d'un crédit — sa meilleure note, 4,0 — la même influence qu'au cours de méthodes de terrain de cinq crédits. La pondération corrige cela : le 4,0 compte une fois et le 3,3 compte cinq fois, de sorte que les cours les plus lourds tirent la moyenne vers leurs notes. Ses deux plus gros cours sont les méthodes de terrain (3,3) et la statistique (3,7), et la cote se tient près d'eux.

(d) En retirant le séminaire :

$$\bar{x}_w = \frac{51,2 - 4,0}{16 - 1} = \frac{47,2}{15} = 3,15$$

Sa cote baisserait de 0,05, passant de 3,20 à 3,15.

Savoir si cela correspond à l'intuition est tout l'objet de la question. La plupart des gens s'attendent à ce qu'abandonner un 4,0 fasse plus mal que cela, parce que c'est sa meilleure note. Ce n'est pas le cas, parce que le cours porte un crédit sur 16 — environ six pour cent de son poids total. La grandeur de la note compte moins que le poids qui s'y rattache, ce qui est la leçon de la moyenne pondérée sous une forme où les étudiantes et les étudiants ont un intérêt personnel.

Corrigé de l'exercice 7.E7. Votre réponse dépend de l'affirmation que vous trouverez. Quatre choses la rendent complète.

Citez l'affirmation exactement, avec sa source. La paraphrase est dangereuse ici, parce qu'elle vous invite à préciser l'affirmation sans vous en rendre compte — et le flou de l'original est souvent ce que vous êtes censé examiner.

Déterminez la mesure employée, ou établissez qu'elle n'est pas énoncée. Le résultat le plus fréquent, et le plus instructif, est que la source ne le dit pas. Si vous avez vérifié l'article, le rapport auquel il renvoie et la note méthodologique sans le découvrir, vous avez fait l'exercice correctement et appris ce qu'il visait à enseigner. C'est un résultat, non un échec à en trouver un.

Soyez précis quant à l'autre mesure. « La médiane pourrait être différente » n'est pas encore une réponse. « Le revenu est étalé vers la droite ; si ce chiffre de 88 000 \$ est une moyenne, la médiane se situe probablement bien en dessous — peut-être autour de 60 000 \$ » en est une : elle nomme la direction, la raison, et une grandeur sur laquelle vous pourriez vous tromper.

Faites une réécriture à la fois exacte et lisible. Il est facile de produire une phrase techniquement juste et impubliable. La limite d'une seule phrase est délibérée : communiquer une distribution asymétrique en langue courante est une compétence réelle, et la contrainte est l'endroit où vous l'exercez. Si votre phrase a besoin d'une subordonnée expliquant ce qu'est une médiane, elle n'y est pas encore.

Une chose que vous remarquerez peut-être en le faisant : la version trompeuse ressemble d'ordinaire à de la négligence plutôt qu'à un calcul. Cela ne la rend pas moins trompeuse, et c'est une raison d'énoncer quelle mesure vous avez employée dans votre propre travail, même quand personne ne l'a demandé.

Corrigé de l'exercice 7.E8. (a) La moyenne est de 90,4 mille et la médiane de 77,7 mille, donc la moyenne est la plus grande — de 15 pour cent de la médiane, à peu de chose près. C'est la direction qui compte : le revenu est étalé vers la droite dans socsurvey comme il l'est dans la réalité, et la moyenne dépasse la médiane en conséquence.

(b) Il ne fait proprement ni l'un ni l'autre, et c'est là la réponse.

L'écart monte dans les trois premiers groupes de scolarité puis redescend, finissant plus bas qu'il n'avait commencé. En pourcentage de la médiane de chaque

groupe, il vaut 5,9 pour cent dans le groupe le moins scolarisé, culmine à 10,9 pour cent dans le groupe 2, et redescend à 4,3 pour cent chez les plus scolarisés — le plus petit écart relatif des cinq.

La lecture tentante, selon laquelle une queue droite plus longue s'accumulerait à mesure que la scolarité monte, n'est donc pas soutenue. C'est exactement l'affirmation que démonte la §7.5.5 : l'ordre de cinq nombres bruités n'est pas un résultat, et avec des groupes de la taille de ceux-ci il bougerait sensiblement si l'enquête était tirée de nouveau.

Ce qui tient, c'est la *direction*. La moyenne dépasse la médiane dans les cinq groupes, pour une raison qu'on peut énoncer d'avance — le revenu est étalé vers la droite partout — et c'est là une affirmation d'une autre nature qu'un ordre.

(c) Les deux chiffres sont 7,6 en laissant les codes en place et 2,8 en les recodant comme valeurs manquantes, sur une échelle qui va de 0 à 10. Les 24 répondants ayant dit « ne sait pas » ou « refus » sont comptés comme s'ils avaient exprimé une confiance quasi totale envers le gouvernement, et ils déplacent la moyenne des 453 vraies réponses de près de cinq points.

L'observation importante est que *rien ne vous avertit*. Le logiciel renvoie un nombre dans les deux cas, 7,6 se tient confortablement à l'intérieur de l'intervalle 0–10, et un lecteur qui n'aurait pas vérifié y verrait un résultat portant sur une population plutôt confiante. Seul le recodage révèle une population majoritairement méfiante. C'est l'argument en faveur de l'étape d'inspection de la section 6.2.3 : arriver en mettant vous-même les mains dans les données.

Cela ne peut pas arriver à *income* dans ce jeu de données, et il vaut la peine de savoir pourquoi : *income* est plafonnée plutôt que codée par valeurs sentinelles, de sorte que sa valeur la plus élevée est 235 mille et que les refus y sont stockés comme manquants plutôt que comme un nombre. La convention qu'emploie un jeu de données est un fait sur ce jeu de données, consigné dans son dictionnaire des variables, et c'est en supposant l'une ou l'autre qu'on commet cette erreur.

(d) La médiane de *socsurvey* ne correspondra pas à celle du Canada, et elle ne devrait pas y correspondre. La préface explique la conception : les jeux de données préservent la structure des relations tout en déformant les niveaux, pour que les démonstrations fonctionnent sans qu'aucun chiffre puisse être pris pour un fait sur le Canada.

L'écart entre les deux n'est donc pas une erreur à expliquer. Une conclusion tirée de *socsurvey* est une conclusion sur *socsurvey*, et l'exercice vaut la peine d'être fait précisément parce que le remarquer est facile à dire et plus difficile à retenir pendant qu'on analyse.