

Corrigé du chapitre 8

Comprendre les données sociales

Mark Paradis, Ph. D.

2 septembre 2026

Voici le corrigé détaillé des exercices de *fin de chapitre* du chapitre 8 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 8 Mesures de dispersion

Corrigé de l'exercice 8.E1. (a) La somme des valeurs est 800, donc

$$\bar{x} = \frac{800}{18} = 44,44$$

Avec $n = 18$, la médiane est la moyenne de la neuvième et de la dixième valeur :

$$\text{médiane} = \frac{34 + 36}{2} = 35$$

La somme des écarts au carré vaut 14 710, d'où

$$s = \sqrt{\frac{14\,710}{17}} = 29,42$$

(b) Avec n pair, les moitiés sont nettes — 9 valeurs chacune.

inférieure : 18, 21, 23, 25, **26**, 28, 30, 32, 34

supérieure : 36, 39, 42, 46, **51**, 58, 67, 84, 140

Chaque moitié a un effectif impair, donc chaque quartile est sa valeur du milieu : $Q_1 = 26$ et $Q_3 = 51$.

Le résumé en cinq nombres est 18, 26, 35, 51, 140.

(c)

$$\text{ÉIQ} = 51 - 26 = 25$$

$$\text{barrière inférieure} = 26 - 1,5(25) = -11,5$$

$$\text{barrière supérieure} = 51 + 1,5(25) = 88,5$$

Une seule valeur, 140, dépasse la barrière supérieure. Rien ne tombe sous la barrière inférieure, qui est négative et donc inatteignable.

(d)

$$\text{asymétrie} = \frac{3(44,44 - 35)}{29,42} = 0,96$$

Forte asymétrie à droite.

(e) Un portrait : « Chez 18 personnes interrogées, l'âge au premier vote était asymétrique à droite (asymétrie 0,96). La médiane était de 35 ans, avec un écart interquartile de 25 ans ; la moyenne de 44,44 ans était tirée vers le haut par la queue, et une valeur de 140 a été signalée comme aberrante. »

(f) Un premier vote à 140 ans est impossible. Il ne s'agit pas d'une véritable valeur extrême mais d'une erreur de saisie — le plus probablement une transposition ou un chiffre dupliqué — et il faut la vérifier contre le dossier d'origine.

Si elle peut être corrigée, corrigez-la. Sinon, elle doit être recodée comme manquante et le cas exclu de l'analyse de cette variable, l'exclusion étant rapportée.

Ce que cela change : la retirer fait tomber n à 17, la moyenne à 38,8 et l'écart type de 29,42 à 17,8 — une baisse de 40 pour cent, parce qu'une seule valeur impossible fournissait l'essentiel de l'écart au carré. La médiane ne bouge que de 35 à 34, et l'ÉIQ de 25 à 20.

Notez que l'ÉIQ bouge tout de même — de 20 pour cent, ce qui n'est pas rien. La robustesse signifie qu'une seule valeur extrême ne peut pas entraîner loin une statistique, non que retirer un cas laisse les quartiles intacts : supprimer une observation quelconque déplace les positions à partir desquelles les quartiles sont lus. Ce qui compte, c'est la comparaison. L'écart type a perdu 40 pour cent de sa valeur et l'ÉIQ 20, pour une seule et même suppression.

Ce contraste est l'argument du chapitre en miniature : les mesures robustes avaient presque raison depuis le début, et les mesures conventionnelles étaient gravement faussées par une seule erreur qu'aucune statistique n'aurait pu détecter à elle seule. Une représentation graphique l'aurait montrée immédiatement.

Corrigé de l'exercice 8.E2. (a) Coefficient de variation et asymétrie pour chacune :

Province	CV	Asymétrie
A	31,0 %	0,41
B	67,6 %	0,75
C	16,9 %	0,25

Pour la province B, par exemple :

$$CV = \frac{48}{71} \times 100 = 67,6 \% \quad \text{asym.} = \frac{3(71 - 59)}{48} = 0,75$$

(b) Par inégalité, de la plus forte à la plus faible : **B, A, C**.

La province B a de loin la plus forte variabilité relative, avec un écart type valant les deux tiers de sa moyenne, et la plus forte asymétrie à droite — sa moyenne dépasse sa médiane de 12 000 \$, ce qui indique un groupe important de hauts revenus au-dessus d'un ménage typique bien plus modeste.

La province A est intermédiaire sur les deux mesures : un CV de 31,0 pour cent et une moyenne qui dépasse sa médiane de 3 000 \$. La province C est la plus égalitaire : ses revenus se concentrent étroitement, son CV atteint à peine 16,9 pour cent, et sa moyenne et sa médiane sont à moins de 1 000 \$ l'une de l'autre, ce qui indique une quasi-symétrie.

Notez que les trois ont la même moyenne. Chaque différence relevée ici est une différence de dispersion et de forme, ce qui est tout l'objet de l'exercice.

(c) **La province C.** Sa médiane de 70 000 \$ est la plus proche de 71 000 \$, sa distribution est presque symétrique et sa dispersion est la plus faible — un ménage

situé à la moyenne se trouve donc en plein milieu d'une distribution serrée.

Dans la province B, un ménage gagnant 71 000 \$ serait bien au-dessus de la médiane de 59 000 \$, et donc plus à l'aise que la plupart.

(d) **La province B**, et le raisonnement se tire des seules statistiques sommaires.

Un seuil de 45 000 \$ se situe sous chacune des trois médianes. La question est de savoir quelle part de chaque distribution tombe sous ce seuil, ce qui dépend de la dispersion — et le seuil se trouve à un nombre très différent d'écarts types sous la moyenne selon la province.

Les revenus de la province C se concentrent à environ 12 000 \$ de 71 000 \$, si bien que 45 000 \$ se situe à 2,17 écarts types sous la moyenne et que très peu de ménages descendent aussi bas. L'écart type de la province B est de 48 000 \$, si bien que le même seuil n'est qu'à 0,54 écarts types sous sa moyenne — à peine la moitié de la distance, en termes relatifs — et que sa queue inférieure s'étend bien au-delà. La province A se situe entre les deux, à 1,18.

L'asymétrie à droite renforce ce constat : la médiane de la province B, à 59 000 \$, est bien plus proche du seuil que les autres, de sorte qu'une plus grande part de ses ménages se situe dans l'intervalle intermédiaire.

C'est le point de politique publique de la §8.1.3 sous forme numérique : le même seuil appliqué à trois provinces de moyennes identiques produit trois programmes complètement différents.

Corrigé de l'exercice 8.E3. 1. Toutes les valeurs du jeu de données sont identiques. Un écart type nul signifie qu'il n'y a d'écart à la moyenne nulle part, ce qui n'est possible qu'en l'absence totale de variation.

2. Parce que l'écart type décrit les *observations*, et non une estimation de celles-ci. Recueillir plus de revenus ne rend pas les revenus moins variables — cela donne une estimation plus précise de leur variabilité. Ce qui diminue avec la taille de l'échantillon, c'est l'*erreur* type, qui décrit de combien une statistique d'échantillon varierait d'un échantillon à l'autre.

3. Oui, *asymétrique à droite*. La médiane se situe 2 unités au-dessus de Q_1 et 23 unités sous Q_3 , de sorte que la moitié supérieure des 50 pour cent centraux est 11,5 fois plus large que la moitié inférieure. Les valeurs sont tassées sous la médiane et étalées au-dessus.

4. Parce que le coefficient de variation divise par la moyenne, ce qui n'a de sens que sur une échelle dotée d'un vrai zéro. Le Celsius a un zéro arbitraire, donc la moyenne peut être proche de zéro ou négative — et diviser par une moyenne proche de zéro produit un rapport énorme ou dépourvu de sens. Le CV exige une variable de rapport.

5. Que l'un des deux contient des valeurs aberrantes ou des queues lourdes. L'ÉIQ couvre la moitié centrale et ignore les extrêmes ; l'écart type, lui, est gonflé

par elles. Un grand écart entre les deux mesures est un signal de valeurs extrêmes, de la même façon qu'un grand écart entre la moyenne et la médiane signale une asymétrie.

6. La moyenne augmente de 5. L'écart type est *inchangé*. Ajouter une constante déplace toute la distribution sans modifier son étalement, et chaque écart à la moyenne reste exactement le même.

Corrigé de l'exercice 8.E4. Votre réponse dépend du rapport que vous trouverez. Quatre choses la rendent complète.

Faire un inventaire exact. La plupart des diffusions gouvernementales rapportent une moyenne ou une médiane et guère plus. Si vous constatez qu'aucune mesure de dispersion n'a été donnée, vous avez trouvé le cas courant, et cette absence est le résultat — non le signe que vous auriez mal choisi votre rapport.

Repérer le manque par rapport au cadre. Parcourez les sept étapes de la §8.8.3 et dites lesquelles ne sont pas remplies. C'est en général la forme, les valeurs aberrantes et souvent n .

Déduire correctement l'asymétrie. Si une moyenne et une médiane sont toutes deux données, le sens découle immédiatement de laquelle est la plus grande, et l'ampleur approximative de la taille de l'écart relativement aux chiffres en jeu. Si un écart type est également donné, calculez le coefficient de Pearson — l'exercice ne le demande pas, mais il transforme un sens en une grandeur.

Garder un jugement proportionné. Il y a deux façons d'aller trop loin : déclarer le rapport sans valeur parce qu'il omet des statistiques, ou l'accepter sans esprit critique. Dites précisément ce qui peut être conclu — en général le centre approximatif — et ce qui ne le peut pas, en général tout ce qui touche à la dispersion, à la forme, ou à la fidélité avec laquelle ce centre décrit un cas typique.

Une chose que vous remarquerez probablement, et elle vaut d'être remarquée : une moyenne publiée est très souvent présentée comme si elle décrivait un cas typique, et l'information permettant de le vérifier est très rarement donnée.

Corrigé de l'exercice 8.E5. 1. 12, 14, 15, 17, 19, 21, 24, 26. Étendue = $26 - 12 = 14$. Somme = 148, donc $\bar{x} = 18,5$; $s = 4,93$. Avec $n = 8$, médiane = $(17 + 19)/2 = 18$. Les moitiés sont 12, **14**, **15**, 17 et 19, **21**, **24**, 26, d'où $Q_1 = 14,5$ et $Q_3 = 22,5$. ÉIQ = 8. Résumé en cinq nombres : 12, 14,5, 18, 22,5, 26.

2. 5, 8, 8, 9, 11, 12, 14, 15, 17. Étendue = 12. Somme = 99, donc $\bar{x} = 11$; $s = 3,87$. Avec $n = 9$, médiane = 11 — la valeur en position 5. En l'incluant dans les deux moitiés : 5, 8, **8**, 9, 11 et 11, 12, **14**, 15, 17, d'où $Q_1 = 8$ et $Q_3 = 14$. ÉIQ = 6. Résumé en cinq nombres : 5, 8, 11, 14, 17.

3. 30, 32, 33, 35, 36, 38, 40, 41, 43, 45. Étendue = 15. Somme = 373, donc $\bar{x} = 37,3$; $s = 4,95$. Avec $n = 10$, médiane = $(36 + 38)/2 = 37$. Les moitiés sont 30, 32, **33**, 35, 36 et 38, 40, **41**, 43, 45, d'où $Q_1 = 33$ et $Q_3 = 41$. ÉIQ = 8.

Résumé en cinq nombres : 30, 33, 37, 41, 45.

4, 2, 4, 5, 6, 8, 9, 11, 13, 15, 45. Étendue = 43. Somme = 118, donc $\bar{x} = 11,8$; $s = 12,35$. Avec $n = 10$, médiane = $(8 + 9)/2 = 8,5$. Les moitiés sont 2, 4, **5**, 6, 8 et 9, 11, **13**, 15, 45, d'où $Q_1 = 5$ et $Q_3 = 13$. ÉIQ = 8. Résumé en cinq nombres : 2, 5, 8,5, 13, 45.

(e) **Le jeu de données 4**, où $s = 12,35$ contre un ÉIQ de 8 — l'écart type dépasse l'écart interquartile, ce qui est inhabituel.

La cause en est la valeur 45. Elle se situe très au-dessus des neuf autres, et le carré de son écart contribue de façon écrasante à la variance. L'ÉIQ l'ignore entièrement, puisqu'elle se trouve hors de la moitié centrale.

Le jeu de données 3 fait le contraste : il a le même n et aucune valeur extrême, et avec $s = 4,95$ contre un ÉIQ de 8 l'écart type est confortablement plus petit, comme il l'est habituellement pour une distribution symétrique sans valeurs extrêmes.

C'est le signal de la question 5 de E3 mis en pratique — un grand écart entre les deux mesures indique des valeurs aberrantes, et la règle des barrières le confirme ici : $13 + 1,5(8) = 25$, donc 45 est bien au-delà de la barrière supérieure.

Corrigé de l'exercice 8.E6. (a) Moyenne 90,4 milliers de dollars, médiane 77,7 milliers, écart type 48,2 milliers, ÉIQ 60,8 milliers, sur 477 personnes interrogées.

La moyenne se situe bien au-dessus de la médiane et l'écart type dépasse la moitié de la moyenne — la signature d'une distribution de revenus asymétrique à droite, et le coefficient de Pearson met un nombre dessus : 0,79.

(b) Avec $Q_1 = 53,1$ et $Q_3 = 113,9$ milliers, la barrière supérieure tombe à 205,1 milliers de dollars et la règle signale 19 ménages. Aucun n'est signalé du côté bas, parce que la barrière inférieure est négative et que le revenu ne l'est pas.

Ces 19 ménages ne sont pas des erreurs. Des revenus de ce niveau sont parfaitement possibles, et la règle les signale parce que le gros de la distribution est comprimé par rapport à sa queue — ce qui est précisément ce qu'est une asymétrie à droite. Rapportez le nombre et conservez-les; supprimer des cas parce qu'une règle de barrières les a nommés, c'est la façon dont une distribution asymétrique se change discrètement en distribution symétrique (§8.7.2).

(c) Le coefficient de variation ne se déplace *pas* dans un seul sens, et le remarquer est tout l'exercice.

En allant du groupe le moins scolarisé au plus scolarisé, il vaut 36,1, 46,3, 46,9, 39,5 et 33,7 pour cent — il monte pour trois groupes, culmine au groupe 3, puis redescend pour finir *sous* son point de départ.

Ce qui tient, et qui tient d'un bout à l'autre, c'est que la dispersion *absolue* s'élargit : l'écart type passe de 21,4 milliers de dollars dans le groupe le moins scolarisé à 50,5 milliers dans le plus scolarisé. Comme la moyenne monte elle aussi, le rapport des deux n'a pas à suivre l'un ou l'autre, et ici il ne les suit pas.

C'est la distinction pour laquelle le CV existe : dispersion absolue et dispersion relative sont deux questions différentes, et une variable peut s'étaler davantage en dollars tout en devenant plus uniforme en proportion de son propre centre.

Résistez à l'attrait de l'ordre apparent. Cinq statistiques de groupe qui montent trois fois puis redescendent, c'est à quoi le bruit ressemble d'habitude ; les groupes comptent ici entre 62 et 110 personnes, et la §7.5.5 travaille exactement ce piège sur exactement ces cinq groupes.

(d) L'écart type vaut 21,0 en laissant les codes en place et 1,9 une fois qu'ils sont recodés — gonflé d'un facteur d'environ 11, sur une échelle qui ne va que de 0 à 10.

Seules 24 personnes portent un code sentinelle, et c'est là le point. La distorsion est pire qu'elle ne l'était pour la moyenne au chapitre 7 parce que la variance *élève au carré* les écarts : une personne enregistrée à 99 sur une échelle de 0 à 10 se situe à quelque 96 points du centre, et 96 au carré fait plus de neuf mille. Une poignée de cas de ce genre peut peser plus lourd que toutes les réponses authentiques réunies.

Cela ne peut pas arriver à *income* dans ce jeu de données, et il vaut la peine de savoir pourquoi : *income* est plafonné plutôt que codé par sentinelles, de sorte que sa plus grande valeur est de 235 milliers de dollars et que les refus sont stockés comme manquants plutôt que comme un nombre. La convention qu'emploie un jeu de données est un fait au sujet de ce jeu de données, consigné dans son dictionnaire des variables, et présumer l'une ou l'autre est la façon dont cette erreur se commet.

La leçon générale : toute statistique bâtie sur des écarts au carré est plus vulnérable aux mauvaises données qu'une statistique bâtie sur des rangs, et le logiciel n'avertit ni dans un cas ni dans l'autre.

(e) Un portrait : « Chez 477 personnes interrogées, le revenu des ménages était asymétrique à droite (asymétrie 0,79). La médiane était de 77,7 milliers de dollars, avec un écart interquartile de 60,8 milliers ; la moyenne de 90,4 milliers était tirée vers le haut par la queue, et 19 ménages ont été signalés par la règle des barrières puis conservés. Le revenu est borné en bas à zéro et non borné en haut, ce qui explique la forme qu'il a (§8.8.1). »