

Corrigé du chapitre 6

Comprendre les données sociales

Mark Paradis, Ph. D.

2 septembre 2026

Voici le corrigé détaillé des exercices de *fin de chapitre* du chapitre 6 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 6 Organiser et visualiser les données

Corrigé de l'exercice 6.E1. (a) Les quatre motifs.

I correspond à ce que les statistiques sommaires laissent croire : une relation à peu près linéaire, avec une dispersion ordinaire autour d'elle. C'est le seul des quatre pour lequel la corrélation est une description honnête.

II est une courbe nette. Les points suivent un arc régulier, qui monte puis redescend. Un résumé en ligne droite est tout simplement de la mauvaise forme, et la corrélation sous-estime une relation presque parfaite — mais non linéaire.

III est une droite parfaite avec une valeur aberrante loin d'elle. Sans ce seul point, la corrélation serait presque exactement de 1 ; avec lui, la droite ajustée est tirée à l'écart de toutes les autres observations.

IV a tous ses points sauf un à la même valeur de x , et un seul cas très à droite. Ce cas unique crée toute la corrélation. Retirez-le et il ne reste aucune relation, parce que x ne varie presque pas.

(b) Que les statistiques sommaires décrivent des traits particuliers d'une distribution et restent muettes sur tout le reste. La moyenne, la variance et la corrélation sont identiques dans les quatre cas ; ces trois nombres ne peuvent donc pas distinguer une relation linéaire d'une courbe, ni d'une droite affectée d'une valeur aberrante, ni d'un motif produit entièrement par un seul cas.

Notez ce que cela ne montre *pas*. Les statistiques ne sont pas fausses — chacune est calculée correctement et rapporte exactement ce qu'elle rapporte. L'erreur consisterait à les traiter comme une description complète.

(c) Portez les données sur un graphique avant d'en calculer quoi que ce soit, et portez-les de nouveau avant de croire ce que vous avez calculé. Précisément : un coefficient de corrélation ne devrait jamais être rapporté sans que le nuage de points ait été examiné, parce que le même coefficient est produit par des formes aux significations entièrement différentes et aux suites entièrement différentes.

Corrigé de l'exercice 6.E2. Votre réponse dépend du tableau que vous trouvez. Quatre choses la rendent complète.

Construire correctement le tableau de fréquences. Des fréquences absolues dont la somme donne le total annoncé, des fréquences relatives en pourcentages, et une note si l'arrondi les empêche de totaliser exactement 100. N'ajoutez une colonne cumulée que si la variable est ordinale.

Laisser la variable choisir le graphique, non votre préférence. Variable catégorielle, diagramme à barres ; variable numérique, histogramme. Si vous choisissez un diagramme circulaire, défendez-le contre la §6.3.5 — et cette défense n'est possible que pour deux ou trois catégories de tailles nettement différentes.

Inclure les éléments exigés dans l'esquisse. Des axes étiquetés avec leurs unités, un titre descriptif, une note de source, et un axe partant de zéro si le graphique

emploie des barres.

Rendre la version trompeuse précise. « On pourrait faire paraître l'écart plus grand » n'est pas encore une réponse. « Faire partir l'axe vertical de 40 plutôt que de 0 ferait remplir tout le panneau par un écart de six points » en est une : elle nomme une technique de la §6.7 et explique le mécanisme qui produit l'impression.

Corrigé de l'exercice 6.E3. (A) *Revenu annuel du ménage.* Asymétrique à droite. Borné par le bas à zéro, non borné par le haut, avec un petit nombre de revenus très élevés qui étirent la queue supérieure. *Médiane*, parce que la queue tire la moyenne au-dessus du ménage typique. *Vérification* : lisez le dictionnaire des variables avant toute chose, et voyez comment le fichier enregistre un refus. Certains jeux de données emploient un code sentinelle comme 999999, qui entre dans chaque calcul comme s'il s'agissait d'un revenu ; d'autres écrètent par le haut, ce qui plafonne la variable au lieu de cela ; socsurvey écrête. Regardez ensuite les zéros — un ménage déclarant exactement zéro revenu peut être un cas authentique ou une valeur manquante codée 0.

(B) *Santé autodéclarée, de 1 à 5, système de santé public solide.* Asymétrique à gauche. La plupart des répondants déclareront une bonne ou une très bonne santé, ce qui entasse les données vers le haut de l'échelle et laisse la queue s'étirer vers le bas — un effet de plafond. *Médiane*, puisque la variable est ordinale et la distribution asymétrique. *Vérification* : confirmez le sens du codage. Que 1 signifie « excellente » ou « mauvaise » détermine si la distribution est asymétrique à gauche ou à droite, et les dictionnaires de variables diffèrent.

(C) *Nombre d'enfants de moins de 18 ans.* Asymétrique à droite et discrète, avec un fort pic à zéro, puisque la plupart des ménages ne comptent aucun enfant de moins de 18 ans. *Médiane*, même si la moyenne est aussi rapportée pour cette variable parce que les questions de politique publique portent sur des totaux. *Vérification* : confirmez que les ménages sans enfant figurent bel et bien dans les données — si la question n'a été posée qu'aux parents, le pic à zéro disparaît et la distribution veut dire autre chose.

(D) *Âge au premier mariage.* À peu près symétrique, ou légèrement asymétrique à droite. Bornée par le bas par les minimums légaux et concentrée dans la fin de la vingtaine, avec une queue modeste de mariages plus tardifs. La *moyenne* se défend si la distribution est presque symétrique ; la médiane est plus sûre.

Guettez ici la bimodalité. Si les hommes et les femmes se marient à des âges moyens différents et que les deux groupes figurent dans le même histogramme, deux pics peuvent apparaître — et le remède est de séparer les groupes plutôt que de décrire la forme combinée (§6.4.3). *Vérification* : voyez si la variable ne porte que sur les personnes qui se sont mariées, ce qui exclut toutes celles qui ne l'ont pas fait et rend l'échantillon sélectionné.

Corrigé de l'exercice 6.E4. (a) Deux problèmes.

La superficie n'est pas la population. Les divisions de recensement du Nord sont immenses et comptent très peu d'habitants. Teintées sur une carte, elles dominent le champ visuel, de sorte que l'impression du lecteur sur « quelle part du Canada accuse un retard » est portée par le territoire plutôt que par les ménages. La carte exagère la portée géographique du résultat.

Le revenu médian des ménages par division masque la variation à l'intérieur des divisions. Une division au revenu médian faible peut contenir à la fois des collectivités aisées et des collectivités pauvres, et la carte ne peut pas le montrer. Lire la carte comme un énoncé sur les ménages du Nord est le sophisme écologique (§6.6.3).

Une troisième réponse se défend : les bornes de classes. Là où tombent les limites entre les teintes détermine quelles divisions paraissent foncées, et des découpages à intervalles égaux, par quantiles ou par seuils naturels produiraient des cartes visiblement différentes à partir de données identiques.

(b) Une échelle *séquentielle* — du clair au foncé dans une seule teinte, le foncé signifiant un revenu plus élevé. Le revenu médian va du bas vers le haut sans point milieu significatif qui le partagerait en deux sortes de valeurs, ce qu'une échelle divergente sous-entendrait. Une échelle divergente conviendrait à une carte proche mais différente : le revenu rapporté à la médiane nationale, où le chiffre national est un véritable point milieu séparant le dessus du dessous.

(c) Une carte à *symboles proportionnels*. La question révisée demande un effectif — combien de ménages — et les effectifs suivent la population. Sur une choroplèthe, une division urbaine densément peuplée contenant des dizaines de milliers de ménages à faible revenu occuperait une petite tache, tandis qu'une division nordique en contenant quelques centaines en occuperait une grande.

Une carte à symboles proportionnels détache la quantité de la superficie : chaque division porte alors un cercle dimensionné selon son effectif. La règle générale : les choroplèthes pour les taux, les symboles proportionnels pour les effectifs.

Corrigé de l'exercice 6.E5. 1. *Diagramme à barres*, trié par fréquence. Un diagramme circulaire est moins bon parce que neuf catégories ne se comparent pas en secteurs, et que plusieurs seront de tailles voisines.

2. *Histogramme*. Un diagramme à barres est moins bon — il traite chaque revenu distinct comme sa propre catégorie, ce qui, pour une variable continue, produit des centaines de barres à un seul cas et ne montre rien.

3. *Polygones de fréquences*, un par groupe, portés en pourcentages à l'intérieur de chaque groupe. Deux histogrammes superposés sont moins bons parce que les barres se cachent l'une l'autre, et porter des effectifs serait pire encore si les deux groupes diffèrent en taille.

4. *Petits multiples*, douze panneaux partageant une même échelle. Un seul graphique à douze courbes est moins bon parce qu'aucun codage ne garde douze séries distinguables, et que le lecteur doit suivre chaque courbe derrière onze autres.

5. *Diagramme à points*. Un histogramme est moins bon : avec quinze cas, les classes en contiennent une ou deux chacune, de sorte que le graphique montre le découpage en classes plutôt que les données.

6. *Nuage de points*. Tout le reste est moins bon, parce que les deux variables sont numériques et que la question porte sur leur relation — ce que seul un nuage de points affiche.

Le motif autour duquel l'exercice est construit : le graphique découle du type de variable et de la question, dans cet ordre. Deux variables numériques et une question de relation donnent toujours un nuage de points ; une variable catégorielle et une question de distribution donnent toujours un diagramme à barres.

Corrigé de l'exercice 6.E6. Votre réponse dépend de la variable que vous choisissez. Quatre choses la rendent complète.

Classer la variable avant de choisir le graphique. Un histogramme d'une variable nominale enregistrée en nombres entiers — la région codée de 1 à 7, par exemple — est précisément l'échec que la §6.2.3 existe pour prévenir. Demandez-vous d'abord de quelle sorte de variable il s'agit, et laissez la réponse choisir le graphique.

Consigner vos choix honnêtement. La largeur de classe, l'ordre des catégories, l'étendue des axes et le traitement des valeurs manquantes sont tous des choix, y compris quand vous ne vous êtes pas aperçu de les faire. Dites ce que vous avez retenu et pourquoi — et « j'ai accepté la largeur de classe par défaut du logiciel » est une entrée légitime, à condition de l'écrire. L'intérêt de la consigne est qu'un réglage par défaut accepté en connaissance de cause se défend, alors qu'un réglage jamais vu ne se défend pas.

Évaluer votre propre graphique pour de bon. Il est facile de produire un graphique et de le déclarer conforme. Cherchez quelque chose à changer : une note de source absente, un axe sans étiquette, un quadrillage plus lourd qu'il n'a besoin de l'être. Si vous ne trouvez absolument rien, c'est probablement que vous n'avez pas repris la liste de la §6.8 point par point.

Faire la version trompeuse délibérément, et la nommer. C'est la partie la plus instructive de l'exercice. Prenez votre propre graphique honnête, tronquez-en l'axe, nommez la technique que vous venez d'employer, et dites pourquoi vous remettriez la version honnête. Si vous ne trouvez pas le moyen de rendre votre graphique trompeur, relisez la §6.7 — presque n'importe quel graphique peut l'être.

Une chose que vous remarquerez peut-être, et qui mérite de l'être : la version trompeuse ressemble souvent davantage aux graphiques publiés dans les journaux que la version honnête.

Corrigé de l'exercice 6.E7. Sept problèmes.

Effet de relief. Déforme la perception — la perspective fait paraître les barres du fond plus petites et rend ambiguë la hauteur de chaque barre. Interdit par l'APA. Correction : des barres plates, à deux dimensions.

Codage par la couleur seule, et de surcroît en rouge et vert. Échoue pour les lecteurs atteints de la forme la plus courante de déficience de la vision des couleurs et dans toute reproduction en niveaux de gris. Correction : si les barres sont les cinq modalités d'une même variable, un seul remplissage est correct et les couleurs sont entièrement inutiles.

Axe tronqué (de 15 à 40 pour cent). Une barre code la quantité par la longueur ; les longueurs disent donc autre chose que les parts de vote. Correction : partir de zéro, sans exception pour les diagrammes à barres.

Axe vertical sans étiquette. Le lecteur ne peut savoir ce qui est mesuré ni en quelles unités. Correction : l'étiqueter, avec son unité.

Trame de fond grise et quadrillage épais. De l'encre non informative qui dispute l'attention aux données. Correction : fond blanc, quadrillage retiré ou réduit à de fines lignes horizontales.

Ordre alphabétique. Le parti est une variable nominale : tout ordre est donc un choix, et l'ordre alphabétique range les données selon un hasard d'orthographe. Correction : trier par part de vote, en ordre décroissant.

« Figure 1 » comme légende. C'est un numéro, non un titre. Notez que sa *position* n'est pas le problème : depuis l'APA 7, le numéro d'une figure et son titre en italique se placent tous deux au-dessus de la figure, exactement comme pour un tableau (§6.8.4). Ce qui manque, c'est le titre. Correction : laisser le numéro où il est, mettre un titre descriptif en italique sur la ligne en dessous, et ajouter la source dans une *Note*, en italique sous la figure.

Des conseils plus anciens vous diront de faire passer toute la légende en dessous. C'était la règle de l'APA 6, et quantité de guides encore en ligne ne se sont pas mis à jour.

Quel changement compte le plus, à lui seul ? L'axe tronqué. Les autres rendent le graphique plus laid, plus difficile à lire ou non conforme ; l'axe tronqué le rend *faux* — un lecteur qui le regarde attentivement et qui lit l'axe en repartira quand même avec une impression déformée des écarts entre les partis. L'effet de relief et les couleurs sont de vrais problèmes, d'un autre ordre, et il vaut la peine de les nommer comme tels — mais ce n'est pas ce qui rend la figure trompeuse.

Corrigé de l'exercice 6.E8. Votre réponse portera sur d'autres régions métropolitaines que celles ci-dessous ; le raisonnement est le même quelles que soient les trois retenues.

(a) Trois séries sur un même graphique, distinguées par le type de trait et la forme

du marqueur plutôt que par la couleur seule (§6.3.4). Trois couleurs produiraient un graphique qui échoue dans l'impression de ce livre même.

(b) Douze courbes se croisent sans cesse et aucun codage ne les garde distinguables. Le remède est le recours aux petits multiples — douze panneaux partageant une même échelle — ou la mise en évidence de deux ou trois séries en noir sur les neuf autres en gris pâle. L'un et l'autre conviennent ; l'échelle partagée, elle, n'est pas facultative.

(c) Probablement oui, et c'est là tout le propos de la question. Une rupture de définition se voit généralement comme une marche : la série fait un saut à une vague et continue à un autre niveau, d'une manière qu'aucun processus social graduel ne produirait. Regardez les vagues 11 à 13 et vous devriez la voir.

Ce que cela suggère : une discontinuité inexpliquée dans une série chronologique que vous n'avez pas recueillie est plus souvent un changement de mesure qu'un changement dans le monde, et le dictionnaire des variables est l'endroit où vérifier. C'est la même leçon qu'à la §4.5.3 — les définitions dérivent, et une rupture peut ressembler exactement à un résultat.

(d) La série simulée ne correspondra pas aux niveaux réels, et ne le doit pas. Ce que vous devriez constater, c'est que l'allure générale et l'ordre entre les RMR sont globalement vraisemblables alors que les chiffres eux-mêmes sont faux — ce que la préface promet exactement. La ressemblance est dans la tendance et dans l'ordre, non dans le niveau, et c'est le devis qui fonctionne comme prévu plutôt qu'un défaut de celui-ci.