

Corrigé du chapitre 11

Comprendre les données sociales

Mark Paradis, Ph. D.

2 septembre 2026

Voici le corrigé détaillé des exercices de *fin de chapitre* du chapitre 11 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 11 Notions de base en probabilité

Corrigé de l'exercice 11.E1. Assemblons d'abord les quatre cellules en un tableau — cela vaut la peine même si l'exercice les donne en liste, parce que chaque partie se ramène ensuite à lire la bonne ligne ou la bonne colonne.

	Satisfait	Insatisfait	Total
Propriétaire	148	72	220
Locataire	96	84	180
Total	244	156	400

(a)

$$P(\text{satisfait}) = \frac{244}{400} = 0,610$$

(b)

$$P(\text{propriétaire}) = \frac{220}{400} = 0,550$$

(c) Se restreindre aux 220 propriétaires et compter les satisfaits :

$$P(\text{satisfait} \mid \text{propriétaire}) = \frac{148}{220} = 0,673$$

(d) Se restreindre aux 180 locataires :

$$P(\text{satisfait} \mid \text{locataire}) = \frac{96}{180} = 0,533$$

(e) Non, elles ne sont pas indépendantes.

L'indépendance exige $P(A \mid B) = P(A)$:

$$P(\text{satisfait}) = 0,610 \quad P(\text{satisfait} \mid \text{propriétaire}) = 0,673$$

Ces valeurs diffèrent, donc les événements sont dépendants. Savoir qu'un répondant est propriétaire de son logement fait passer la probabilité qu'il soit satisfait de 61,0 à 67,3 pour cent.

Le test par multiplication concorde :

$$P(\text{propriétaire}) \times P(\text{satisfait}) = 0,550 \times 0,610 = 0,336$$

contre une probabilité conjointe réelle de $148/400 = 0,370$. Les deux événements se produisent ensemble plus souvent que l'indépendance ne le prédirait.

Un troisième chemin : comparer directement les deux probabilités conditionnelles. Les propriétaires sont satisfaits 67,3 pour cent du temps et les locataires 53,3 pour cent — un écart de 14 points de pourcentage, ce qui est l'association énoncée dans le langage du chapitre 10.

(f) La règle d'addition générale :

$$P(\text{propriétaire} \cup \text{satisfait}) = 0,550 + 0,610 - 0,370 = 0,790$$

Vérification par complément : les seuls répondants dans aucune des deux catégories sont les 84 qui louent et sont insatisfaits, donc

$$\frac{400 - 84}{400} = \frac{316}{400} = 0,790$$

Corrigé de l'exercice 11.E2. (a) La règle d'addition générale, puisque 280 personnes ont fait les deux :

$$P(V \cup \text{Bén}) = 0,620 + 0,410 - 0,280 = 0,750$$

(b) Des 620 qui ont voté, 280 ont aussi fait du bénévolat, donc 340 ont voté sans faire de bénévolat :

$$P(V \cap \text{Bén}^c) = \frac{340}{1\,000} = 0,340$$

Cela se calcule aussi comme $P(V) - P(V \cap \text{Bén}) = 0,620 - 0,280$, qui est la même soustraction énoncée en probabilités plutôt qu'en effectifs.

(c) Ne faire ni l'un ni l'autre est le complément de faire au moins l'un des deux :

$$P(\text{ni l'un ni l'autre}) = 1 - 0,750 = 0,250$$

(d) Se restreindre aux votants :

$$P(\text{Bén} \mid V) = \frac{280}{620} = 0,452$$

(e) Se restreindre aux bénévoles :

$$P(V \mid \text{Bén}) = \frac{280}{410} = 0,683$$

Les parties (d) et (e) valent la peine d'être comparées. Les deux comptent les mêmes 280 personnes, et elles diffèrent parce que les groupes dont elles sont une fraction n'ont pas la même taille — 620 votants contre 410 bénévoles.

La lecture de fond : la plupart des bénévoles votent (68 pour cent), alors que moins de la moitié des votants font du bénévolat (45 pour cent). Le bénévolat est l'acte le plus exigeant, donc c'est le plus petit groupe et celui qui prédit le plus fortement l'autre.

C'est la distinction $P(A \mid B)$ contre $P(B \mid A)$ dans un cas où les deux nombres ont du sens et où ni l'un ni l'autre n'est un piège — ce qui en fait une bonne

vérification de la distinction comprise plutôt que mémorisée.

Corrigé de l'exercice 11.E3. *Une distribution de données* décrit des valeurs réellement observées. L'histogramme de 500 revenus montre comment ces 500 cas se répartissaient — c'est le résumé de données en main, et il aurait une autre allure si un autre échantillon avait été tiré.

Une distribution de probabilité décrit ce qui arriverait à long terme, ou sur l'ensemble des cas possibles. La courbe normale n'est le résumé d'aucun jeu de données particulier ; c'est un modèle mathématique qui précise la probabilité de chaque intervalle de valeurs.

La distinction oppose la description au modèle. L'une rapporte ce qui est arrivé ; l'autre énonce ce qui a tendance à arriver.

Pourquoi cela importe pour la distribution d'échantillonnage. Une distribution d'échantillonnage est une distribution de probabilité, et ses résultats ne sont pas des cas mais des échantillons entiers. Elle décrit les valeurs que prendrait une moyenne d'échantillon sur tous les échantillons possibles d'une taille donnée — dont la quasi-totalité n'ont jamais été tirés.

C'est pourquoi la distinction doit être solide d'abord. Si une distribution n'est jamais que le résumé de données observées, une distribution portant sur des échantillons jamais tirés n'a nulle part où loger. La distribution d'échantillonnage est un énoncé à propos d'un processus hypothétique, et raisonner sur des processus hypothétiques est exactement ce que rend possible un modèle de probabilité.

Une façon utile de tenir les trois niveaux séparés : la distribution de données décrit les cas de votre échantillon ; la distribution de population décrit les cas que vous n'avez pas échantillonnés ; la distribution d'échantillonnage décrit les échantillons que vous n'avez pas tirés.

Corrigé de l'exercice 11.E4. (a) Prenons 100 000 personnes hypothétiques.

Séparer selon le statut de la condition. 0,5 pour cent l'ont :

$$100\,000 \times 0,005 = 500 \text{ avec la condition}$$

$$99\,500 \text{ sans}$$

Appliquer les taux du test.

$$500 \times 0,92 = 460 \text{ vrais positifs}$$

$$99\,500 \times 0,06 = 5\,970 \text{ faux positifs}$$

	Test positif	Test négatif	Total
A la condition	460	40	500
Sans condition	5 970	93 530	99 500
Total	6 430	93 570	100 000

Conditionner sur un résultat positif.

$$P(\text{condition} \mid \text{positif}) = \frac{460}{6\,430} = 0,072$$

Environ 7 pour cent.

(b) L'affirmation n'est pas justifiée, et l'écart est spectaculaire : un résultat positif porte à peu près 7 pour cent de probabilité d'avoir la condition, et non une quasi-certitude.

Ce qu'il faut dire au responsable. Le chiffre de 92 pour cent est $P(\text{positif} \mid \text{condition})$ — la performance du test sur les gens qui l'ont. La question qui importe à une personne qui reçoit un résultat est $P(\text{condition} \mid \text{positif})$, et ce sont deux quantités différentes.

La raison pour laquelle elles divergent autant est le taux de base. Seulement 500 personnes sur 100 000 ont la condition, de sorte que même un taux de faux positifs de 6 pour cent appliqué aux 99 500 autres produit 5 970 faux positifs — treize fois les 460 vrais. La plupart des résultats positifs viennent de personnes en santé, et aucune amélioration de l'exactitude du test sur les malades ne peut y changer quoi que ce soit, parce que l'arithmétique est gouvernée par la taille du groupe en santé.

L'implication pratique pour le programme de dépistage. Dépister les 500 000 résidents produirait 32 150 résultats positifs, dont 2 300 seraient authentiques. Les 29 850 autres personnes subiraient des examens de suivi, avec l'anxiété et les coûts qui les accompagnent, pour une condition qu'elles n'ont pas.

Ce n'est pas un argument contre le dépistage en soi — c'est un argument pour le cibler sur les groupes à risque plus élevé, où le taux de base est plus haut et où un résultat positif est donc plus informatif, et pour communiquer clairement ce que signifie réellement un résultat positif.

Corrigé de l'exercice 11.E5. (a) Votre choix initial avait une probabilité de $1/4$ d'être le bon, et rien de ce que fait l'animateur ne change cela. L'animateur pouvait de toute façon ouvrir 2 portes à chèvre, où que soit la voiture, de sorte que le fait qu'il le fasse ne vous apprend rien sur votre propre porte.

$$P(\text{gagner en restant}) = \frac{1}{4} = 0,25$$

(b) Les $3/4$ restants de la probabilité étaient répartis sur les 3 autres portes. L'animateur en a éliminé 2, de sorte que tout se concentre sur la seule porte encore

fermée :

$$P(\text{gagner en changeant}) = \frac{3}{4} = 0,75$$

Changer est 3 fois meilleur que rester.

En le travaillant cas par cas, comme l'a fait la §11.6.1 : dans le cas unique sur 4 où la voiture est derrière la porte 1, changer perd. Dans chacun des 3 autres cas, l'animateur est contraint d'ouvrir les 2 portes à chèvre parmi celles qui restent, et changer tombe sur la voiture. 3 gains sur 4.

(c) Avec n portes, votre choix initial a une probabilité de $1/n$. L'animateur ouvre $n - 2$ portes à chèvre, ce qui laisse exactement une solution de rechange, et les $(n - 1)/n$ restants s'y concentrent :

$$P(\text{gagner} \mid \text{changer}) = \frac{n - 1}{n}$$

La formule reproduit les cas familiers. À $n = 3$ elle donne $2/3$; à $n = 4$, $3/4$; à $n = 100$, $99/100$.

Elle rend aussi la structure visible. À mesure que n grandit, l'avantage de changer approche la certitude, parce que votre choix initial devient de pire en pire pendant que l'élimination faite par l'animateur devient de plus en plus informative. La version à cent portes de la §11.6.1 est cette formule à $n = 100$, et elle convertit l'intuition de la plupart des gens précisément parce que l'effet est trop grand pour être écarté.

La condition qui fait fonctionner tout cela, à n'importe quel n : l'animateur sait où est la voiture et ne l'ouvre jamais. Retirez cela et l'avantage disparaît, comme l'a montré l'exercice 11.6.1.

Corrigé de l'exercice 11.E6. 1. La participation électorale. Élections Canada rapporte la participation officielle à la 45^e élection générale, tenue le 28 avril 2025, à **69,0 pour cent** des électeurs inscrits — 68,0 pour cent lorsque le dénominateur est l'ensemble des électeurs admissibles plutôt que des inscrits (ELECTIONS CANADA 2025). Indiquez lequel des deux vous avez employé ; ce sont deux quantités différentes.

La location. Le recensement de 2021 a dénombré 5,0 millions de ménages locataires, soit un taux de **33,1 pour cent** de l'ensemble des ménages (STATISTICS CANADA 2023).

Vos chiffres peuvent différer si une élection ou un recensement plus récent a été publié. Donnez l'année.

2. Ni l'un ni l'autre n'est une probabilité pour le Canada, et la raison est la même dans les deux cas : une probabilité est un énoncé à propos d'un processus de sélection précisé, et le processus sélectionne ici dans un jeu de données.

$P(\text{a voté}) = 0,673$ décrit le tirage au hasard d'un répondant parmi les 477 de socsurvey. Pour que ce soit une probabilité à propos du Canada, socsurvey devrait être un échantillon aléatoire d'adultes canadiens — et c'est une simulation, donc ce n'est l'échantillon de rien.

$P(\text{est locataire}) = 0,327$ décrit le tirage au hasard d'une personne parmi les 50 000 de censuspop. Pour que ce soit une probabilité à propos du Canada, censuspop devrait être la population canadienne.

Notez ce que cela ne dit *pas*. L'arithmétique est juste dans les deux cas et la méthode est exactement celle qu'on emploierait sur de vraies données. Ce qui échoue, c'est la classe de référence, non le calcul.

3. Le chiffre du logement est la comparaison la plus difficile, et il l'est d'une manière facile à manquer, parce que les deux nombres ont l'air si proches.

Les lignes de censuspop sont des *personnes*. 0,327 est donc la part des *personnes* qui vivent en logement loué — 16 366 sur 50 000. Les 33,1 pour cent de Statistique Canada sont la part des *ménages* qui louent. Ce sont deux quantités différentes sur deux dénominateurs différents, et elles diffèrent dès que les ménages locataires et propriétaires n'ont pas la même taille.

Calculé sur les ménages plutôt que sur les personnes, censuspop donne 6 416 ménages locataires sur 19 400, soit 33,1 pour cent. *C'est* le chiffre comparable à celui de Statistique Canada, et comparer 32,7 à 33,1 aurait été comparer un taux de personnes à un taux de ménages et appeler « concordance » une quasi-concordance.

C'est le problème de l'unité d'analyse de la §5.1.2 arrivé à l'intérieur d'une probabilité. Avant de comparer deux proportions, vérifiez que les dénominateurs comptent la même sorte de chose.

La comparaison des taux de participation a sa propre difficulté, qui mérite d'être nommée même si elle est plus petite : Élections Canada compte des bulletins déposés par rapport à un registre, alors qu'une enquête compte les personnes qui *disent* avoir voté, et la participation autodéclarée est systématiquement plus élevée que le chiffre officiel. Deux nombres qui se réclament tous deux de la « participation » ne mesurent pas nécessairement le même événement.

Un dernier point, et c'est celui à retenir. La concordance entre censuspop et le vrai recensement n'est la preuve de rien : la simulation a été construite pour ressembler au Canada, de sorte que ressembler au Canada est une propriété de sa construction et non un constat. Un jeu de données simulé ne peut jamais que confirmer ce qu'on y a mis.