

Corrigé du chapitre 14

Comprendre les données sociales

Mark Paradis, Ph. D.

2 septembre 2026

Voici le corrigé détaillé des exercices de *fin de chapitre* du chapitre 14 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 14 Le test du khi carré

Corrigé de l'exercice 14.E1. (a) Les totaux marginaux sont : lignes 140, 140, 88, 112; colonnes 214, 192, 74; $N = 480$.

En appliquant $E = (\text{ligne} \times \text{colonne})/N$:

<i>Attendus</i>	Appui	Opposition	Sans opinion
Parti A	62,42	56,00	21,58
Parti B	62,42	56,00	21,58
Parti C	39,23	35,20	13,57
Indépendants	49,93	44,80	17,27

Le plus petit effectif attendu est 13,57, confortablement au-dessus de 5. Les douze cellules passent, donc la condition tient.

(b) En additionnant $(O - E)^2/E$ sur les douze cellules :

$$\chi^2 = 19,70 \quad df = (4 - 1)(3 - 1) = 6$$

(c) La valeur critique à $\alpha = 0,05$ avec $df = 6$ est 12,59. Puisque $19,70 > 12,59$, on rejette H_0 . La valeur p vaut environ 0,003.

L'appartenance partisane et l'appui à la réforme électorale sont associés.

(d) Pour un tableau 4×3 , $\min(r - 1, c - 1) = \min(3, 2) = 2$:

$$V = \sqrt{\frac{19,70}{480 \times 2}} = \sqrt{0,0205} = 0,143$$

Avec $\min(r - 1, c - 1) = 2$, les repères sont 0,07 pour petit, 0,21 pour moyen et 0,35 pour grand. Une valeur de 0,14 tombe entre petit et moyen — une association modeste.

(e) *Un compte rendu complet* :

« Un test d'indépendance du khi carré a trouvé une association significative entre l'appartenance partisane et l'appui à la réforme électorale, $\chi^2(6, N = 480) = 19,70$, $p = 0,003$, V de Cramér = 0,14. Les partisans du parti A étaient favorables à la réforme (55,7 pour cent d'appui, 30,0 pour cent d'opposition) tandis que ceux du parti B s'y opposaient (32,9 pour cent d'appui, 52,9 pour cent d'opposition). Les électeurs du parti C et les indépendants se situaient entre les deux, et la proportion sans opinion était semblable dans les quatre groupes. »

(f) Les plus fortes contributions viennent des partis A et B :

<i>Contributions</i>	Appui	Opposition	Sans opinion
Parti A	3,89	3,50	0,12
Parti B	4,32	5,79	0,12
Parti C	0,04	0,29	1,45
Indépendants	0,09	0,01	0,09

Les partis A et B fournissent ensemble 17,73 des 19,70 — près de 90 pour cent. Le parti C et les indépendants ne fournissent presque rien ; leurs distributions d'opinion sont proches de ce que prédit l'indépendance.

Sur le fond : l'association tient presque entièrement au fait que les deux grands partis prennent des positions opposées. Les partisans du parti A appuient la réforme au-dessus du taux d'ensemble, ceux du parti B s'y opposent, et le petit parti et les indépendants n'ont rien de remarquable.

Il s'agit d'un clivage partisan sur une politique précise plutôt que d'une relation générale entre identité partisane et attitudes envers les réformes — ce qui est un constat plus précis que ne le transmet le khi carré significatif à lui seul, et qui ne s'obtient que par la décomposition.

Corrigé de l'exercice 14.E2. (a) Pour un tableau 3×4 , $\min(r - 1, c - 1) = \min(2, 3) = 2$:

$$V = \sqrt{\frac{18,4}{600 \times 2}} = \sqrt{\frac{18,4}{1\,200}} = \sqrt{0,01533} = 0,124$$

(b) Avec $\min(r - 1, c - 1) = 2$, les repères sont 0,07 pour petit, 0,21 pour moyen et 0,35 pour grand.

Une valeur de 0,12 est petite — au-dessus du seuil « petit » et bien en dessous de « moyen ». La situation d'emploi et la santé autoévaluée sont associées, faiblement.

Notez l'importance d'employer la bonne ligne de repères. Jugée contre les repères de $\min = 1$ (0,10, 0,30, 0,50), 0,12 paraîtrait petite de la même façon ; mais pour un tableau plus grand les seuils baissent davantage, et une comparaison mécanique contre la mauvaise ligne peut mal classer un effet (§14.4).

(c) La conclusion n'est pas soutenue, pour trois raisons.

Le test établit une association, non une causalité. Ce sont les trois structures de la §3.5, et le test du khi carré n'en distingue aucune.

La causalité inverse est hautement plausible ici — davantage que dans la plupart des exemples. Une mauvaise santé cause le chômage : les personnes malades perdent leur emploi, peinent à en trouver un, et quittent la population active. La catégorie « hors de la population active » contient en particulier bien des personnes que leur santé empêche de travailler.

Les variables confondantes abondent. L'âge, la scolarité, le revenu, la région et une maladie chronique antérieure agissent tous plausiblement à la fois sur la situation d'emploi et sur la santé.

À noter aussi : l'association est faible ($V = 0,12$), de sorte que même causale elle expliquerait peu de la variation de la santé.

Ce qu'il faudrait. Des données longitudinales établissant l'ordre temporel — la santé mesurée avant la perte d'emploi, puis de nouveau après. Idéalement une

expérience naturelle : les fermetures d'usine et les licenciements massifs sont le devis standard ici, parce qu'ils rendent la perte d'emploi indépendante de la santé antérieure de la personne. Et la mesure des variables confondantes candidates, pour qu'on puisse examiner leur contribution.

Le devis de la fermeture d'usine, ou quelque chose du genre, est ce qu'emploie la littérature — ce qui vaut d'être nommé pour cette raison.

Corrigé de l'exercice 14.E3. (a) Les totaux marginaux sont : lignes 74, 68, 23 ; colonnes 80, 72, 13 ; $N = 165$.

<i>Attendus</i>	Chaque jour	Chaque semaine	Jamais
Urbain	35,88	32,29	5,83
Banlieue	32,97	29,67	5,36
Rural	11,15	10,04	1,81

1 cellule ne respecte pas la règle stricte : rural–jamais, avec un effectif attendu de 1,81.

Notez que l'effectif observé de cette cellule est 2, mais que la condition porte sur les effectifs *attendus* plutôt qu'observés — une distinction qui vaut d'être vérifiée, puisque les étudiants éprouvent souvent le mauvais tableau.

(b) *Oui, de justesse*. La règle souple tolère jusqu'à 20 pour cent des cellules sous 5, aucune n'étant sous 1.

Ici, $1/9 = 11,1$ pour cent des cellules tombent sous 5, ce qui est sous l'allocation de 20 pour cent. L'effectif attendu minimal vaut 1,81, ce qui dépasse 1.

Le tableau passe donc la règle souple et échoue à la règle stricte. Un test du khi carré serait défendable, et le résultat devrait être rapporté avec la réserve notée.

(c) Deux options, avec une recommandation.

Regrouper des catégories. La colonne « jamais » est le problème : elle ne contient que 13 des 165 répondants. La fusionner avec « chaque semaine » pour donner une catégorie « moins que quotidiennement » produirait un tableau 3×2 dont tous les effectifs attendus dépasseraient 10.

C'est défendable sur le fond — le contraste qui compte est peut-être bien quotidien contre non quotidien — et cela doit être rapporté comme une décision prise.

Employer le test exact de Fisher. Des extensions au-delà du 2×2 existent et sont disponibles dans les logiciels courants. Cela n'exige aucune modification des données.

Recommandation : regrouper les colonnes, pourvu que le contraste quotidien contre non quotidien réponde à la question de recherche. La catégorie est clairsemée parce que peu de gens n'emploient jamais le transport collectif, et une distinction à trois niveaux dont un niveau est à peine peuplé ne sert pas l'analyse.

Si le groupe « jamais » est justement l'objet de l'étude — si la question porte précisément sur les non-usagers du transport collectif — alors le regroupement détruit la question de recherche et c'est le test exact de Fisher qui est la bonne réponse.

Notez aussi qu'avec $N = 165$ cette étude a peu de puissance, de sorte qu'un résultat non significatif n'établirait pas grand-chose, quelle que soit la voie choisie.

Corrigé de l'exercice 14.E4. Votre article différera de celui de tout autre lecteur. Ce qui suit est ce que couvre une réponse complète, dans l'ordre où l'exercice le demande.

L'inventaire est-il exact et complet ? La plupart des comptes rendus publiés de khi carré donnent la statistique, les degrés de liberté et la valeur p . Beaucoup omettent N des parenthèses, et une proportion appréciable omet entièrement la taille d'effet. L'exercice porte sur ce qui *manque* ; n'énumérer que ce qui est présent, c'est le manquer.

Les degrés de liberté concordent-ils ? C'est la vérification la plus utile dont dispose une lectrice, et elle attrape de vraies erreurs. Si l'article décrit un tableau 3×4 et rapporte $df = 11$, quelque chose cloche — $(r - 1)(c - 1)$ donne 6, et $rc - 1$ donne 11, qui est l'erreur d'origine de Pearson (§14.2.4).

La taille d'effet a-t-elle été calculée correctement ? À partir de χ^2 et de N , vous pouvez retrouver vous-même le V de Cramér, en employant la ligne de repères qui correspond aux dimensions du tableau.

La conclusion est-elle proportionnée ? Deux défaillances à surveiller : des verbes causaux là où le devis est transversal, et un petit V décrit comme s'il était important. Vérifiez aussi si l'article rapporte les pourcentages ; sans eux, le sens de l'association n'est pas disponible.

Une régularité vaut d'être remarquée sur plusieurs articles : à quelle fréquence un résultat de khi carré est rapporté sans taille d'effet, et si ceux qui l'omettent tendent à avoir de grands échantillons. Si c'est le cas, c'est la §13.6.1 à l'œuvre dans une littérature plutôt que dans un manuel.

Corrigé de l'exercice 14.E5. Votre réponse emploiera d'autres variables que celles de quiconque ; le raisonnement est le même quelle que soit la paire choisie.

(a) Énoncez les hypothèses avant de calculer quoi que ce soit, et sous la forme qui convient au khi carré — sur l'indépendance plutôt que sur la valeur d'un paramètre. Écrire $H_0 : \mu = \dots$ signale qu'on a importé le mauvais gabarit.

(b) Les effectifs et les pourcentages en ligne sont tous deux exigés, avec le sens du pourcentage énoncé (§10.2.2). La vérification des effectifs attendus devrait rapporter la valeur attendue minimale plutôt que se contenter d'affirmer que la condition tient.

(c) Un compte rendu complet comprend la statistique avec df et N , la valeur p

exacte, le V de Cramér avec la bonne ligne de repères, et les pourcentages qui montrent le sens.

(d) Les trois issues instruisent toutes, et laquelle survient dépend des variables choisies.

Persiste : la variable de contrôle n'explique pas l'association. La plupart des tentatives aboutiront ici.

S'évanouit : la variable de contrôle la produisait. C'est l'issue la plus instructive et elle demande de chercher un peu — dans *socsurvey*, la scolarité est la candidate la plus probable pour la produire, puisqu'elle est liée à quantité d'autres variables.

Diffère d'une strate à l'autre : une interaction. À rapporter avec soin, puisqu'une analyse combinée l'aurait dissimulée entièrement.

Une chose à surveiller dans votre propre travail : stratifier à répétition jusqu'à ce qu'un résultat intéressant apparaisse, puis ne rapporter que celui-là. Cet exercice y invite. Remarquer la tentation, et rapporter chaque tentative, est tout le propos de la boîte d'éthique de la §14.7 — et c'est plus facile à voir chez soi ici que dans un article publié.

(e) *Réel contre simulé*. Le supplément sur le vote de l'Enquête sur la population active de Statistique Canada publie la participation par scolarité, et la configuration va dans le même sens que celle de *socsurvey* : la participation monte avec la scolarité, et l'écart entre les moins et les plus scolarisés est grand.

Un test du khi carré trancherait-il la concordance des deux ? Non, et la raison vaut d'être énoncée avec précision. Un test d'indépendance demande si deux variables sont associées à l'intérieur d'un tableau. Il ne peut pas comparer un tableau à un ensemble de pourcentages extérieurs — c'est une question d'ajustement (§14.6), et même alors la réponse serait ici sans intérêt.

Fait comme test d'ajustement contre les chiffres réels, le test rejetterait presque certainement, pour deux raisons qui n'ont rien à voir avec le Canada. *socsurvey* est simulé, de sorte que ses pourcentages ont été choisis plutôt qu'observés. Et la participation autodéclarée dans les enquêtes dépasse la participation officielle partout où on l'a mesurée, de sorte que même une véritable enquête échouerait à cette comparaison.

Ce à quoi la comparaison sert, c'est le sens et l'ordre de grandeur du gradient. Ceux-là concordent, ce qui vous dit que la simulation est un endroit raisonnable pour s'exercer — et rien du tout sur le Canada.