

Corrigé du chapitre 13

Comprendre les données sociales

Mark Paradis, Ph. D.

2 septembre 2026

Voici le corrigé détaillé des exercices de *fin de chapitre* du chapitre 13 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 13 Les tests d'hypothèses

Corrigé de l'exercice 13.E1. (a) *Les hypothèses.*

$$H_0 : \mu = 8,0 \quad H_1 : \mu \neq 8,0$$

Bilatéral, et la justification compte. Le groupe de défense s'attend à des temps plus longs, ce qui suggère un test unilatéral — mais un constat de temps de réponse *plus courts* que ce que la ville affirme serait important sur le fond, et un test unilatéral ne pourrait pas le rapporter.

Un test unilatéral se défend si le groupe déclare à l'avance qu'un résultat inférieur à l'affirmation mènerait à la même conclusion qu'une absence de résultat. C'est difficile à soutenir ici.

(b) *Les six étapes.*

1. *Hypothèses* : comme ci-dessus.

2. *Seuil de signification* : $\alpha = 0,05$.

3. *Conditions* : les 36 appels forment un échantillon aléatoire ; les appels sont indépendants ; $n = 36$ dépasse 30, de sorte que le théorème central limite s'applique ; $\sigma = 3,0$ est traité comme connu d'après les registres de la ville.

4. *Statistique de test.*

$$SE = \frac{3,0}{\sqrt{36}} = \frac{3,0}{6} = 0,50$$

$$z = \frac{9,1 - 8,0}{0,50} = \frac{1,1}{0,50} = 2,20$$

5. *Valeur p.* D'après la table z , l'aire au-delà de $z = 2,20$ vaut 0,0139. En bilatéral :

$$p = 2 \times 0,0139 = 0,028$$

6. *Décider et interpréter.* Puisque $0,028 \leq 0,05$, on rejette H_0 .

Conclusion de fond : le temps de réponse moyen des services d'urgence diffère des 8,0 minutes que la ville affirme, et l'échantillon suggère qu'il est plus long — environ 9,1 minutes, à peu près une minute de plus qu'annoncé.

(c) *Taille d'effet et compte rendu APA.*

$$d = \frac{9,1 - 8,0}{3,0} = 0,37$$

$$M = 9,1, \quad z = 2,20, \quad p = 0,028, \quad d = 0,37, \quad n = 36$$

Un intervalle à 95 % irait de $9,1 \pm 1,96(0,50) = [8,1, 10,1]$, ce qui vaut la peine d'être ajouté : il exclut 8,0, ce qui concorde avec le test, et il montre que l'estimation

n'est pas précise.

(d) *L'erreur.* La ville a énoncé $P(H_0 \mid \text{données}) = 0,028$, alors que la valeur p est $P(\text{données} \mid H_0) = 0,028$. Le conditionnel a été renversé.

Ce que la valeur p dit réellement : si le temps de réponse moyen véritable était bel et bien de 8,0 minutes, une moyenne d'échantillon s'écartant d'au moins 1,1 minutes surviendrait environ 2,8 fois sur cent. Cela ne dit rien directement de la probabilité que les registres soient exacts — la calculer exigerait une probabilité a priori que le cadre ne fournit pas.

Il y a aussi une seconde erreur qui mérite d'être relevée. « Nos registres sont exacts » n'est pas l'hypothèse nulle. La nulle dit que la moyenne véritable vaut 8,0 ; les registres pourraient être exacts sur autre chose, ou avoir raison sur la moyenne et tort sur la variabilité. Confondre une affirmation portant sur un paramètre précis avec la fiabilité générale des données de la ville exagère ce que le test a examiné.

Corrigé de l'exercice 13.E2. *La chercheuse B a raison sur le fond ; le chercheur A a raison sur l'arithmétique ; et ils répondent à des questions différentes.*

Ce que A a établi. Avec $n = 5\,000$ et $p = 0,04$, la différence se distingue de zéro. C'est un constat réel : la valeur de population n'est probablement pas exactement la valeur de référence.

Ce que B a remarqué. Une taille d'effet de $d = 0,04$ vaut un vingt-cinquième d'écart type — un cinquième du seuil « petit » de Cohen. Deux individus séparés par un tel écart sont indiscernables en pratique.

Pourquoi les deux peuvent être vrais. La statistique de test porte $\sqrt{n}$ à son dénominateur, de sorte qu'à $n = 5\,000$ même une différence insignifiante franchit le seuil (§13.6.1). La signification a établi que la différence n'est pas nulle ; la taille d'effet a établi qu'elle ne vaut pas grand-chose. Ce sont deux affirmations distinctes et aucune ne réfute l'autre.

Comment trancher. La réponse de A — « mais $p < 0,05$, donc c'est significatif » — est la position la plus faible, parce qu'elle traite la signification comme si elle répondait à la question de B. Elle n'y répond pas. La bonne issue est de dire que le constat est statistiquement détectable et négligeable sur le fond, et que l'article doit l'écrire.

Quels renseignements supplémentaires aideraient.

L'intervalle de confiance de l'effet, qui montrerait si les données excluent des effets assez grands pour compter. Un intervalle de $[0,01, 0,07]$ trancherait la question de façon décisive en faveur de B.

Une valeur de référence de ce qui compte comme significatif sur le fond dans ce domaine. Les tailles d'effet ne sont petites ou grandes que relativement à ce qu'un domaine obtient d'ordinaire, et la recherche sur la participation civique peut avoir ses propres normes.

Le coût et la portée de ce qui est évalué. Un effet minuscule produit par une politique appliquée à des millions de personnes sans frais peut valoir la peine, et c'est un jugement qu'aucune statistique ne rend.

Corrigé de l'exercice 13.E3. (a) *Étude d'origine* : $0,049 \leq 0,05$, donc on rejette H_0 .
Réplication : $0,21 > 0,05$, donc on ne rejette pas.

(b) *Non, cela n'établit pas que le constat d'origine était faux.* Plusieurs explications sont compatibles avec ces résultats.

L'étude d'origine était un faux positif. À $p = 0,049$, elle franchissait le seuil de justesse, et avec 30 participants par groupe ce n'était pas une grande étude. Une certaine proportion de tels constats sont des erreurs de type I.

L'effet véritable est réel mais plus petit que ce que l'étude d'origine estimait. C'est la lecture la plus probable. Une étude sous-puissante n'atteint la signification que lorsque son estimation d'échantillon est anormalement grande (§13.3.4), de sorte que les effets publiés issus de petites études sont systématiquement surestimés. Le $d = 0,18$ de la réplication est peut-être plus près de la vérité que le $0,52$ de l'étude d'origine.

La réplication différait sur un point pertinent. Population, contexte, façon de livrer le programme, ou façon de mesurer le stress. Les réplifications directes sont plus difficiles à réussir qu'il n'y paraît.

La réplication était elle-même sous-puissante. Avec 100 participants par groupe, détecter $d = 0,18$ n'est pas garanti — de sorte qu'un résultat non significatif reste compatible avec un petit effet réel.

(c) *C'est la réplication qui devrait peser le plus lourd,* pour trois raisons.

Elle est beaucoup plus grande — 200 participants contre 60 — donc son estimation est plus précise.

Elle a été menée en connaissant le résultat d'origine, ce qui veut d'ordinaire dire que le devis et les mesures ont été examinés plus attentivement.

Et elle n'est pas soumise à la sélection qui a produit l'étude d'origine. Celle-ci a été publiée en partie *parce qu'elle* était significative ; la réplication a été menée indépendamment de ce qu'elle allait trouver.

La meilleure lecture d'ensemble : le programme a probablement un petit effet positif, plus petit que ne le suggérait l'étude d'origine et peut-être trop petit pour justifier l'intervention. Cette conclusion emploie les deux études plutôt que d'en choisir une, ce que ferait une lecture méta-analytique.

« La réplication gagne » est le bon réflexe arrêté trop tôt ; « les études se contredisent, donc nous ne savons rien » néglige qu'elles s'accordent sur la direction et ne diffèrent que par l'ampleur.

Corrigé de l'exercice 13.E4. Un plan défendable, avec le raisonnement de chaque élément.

(a) *Hypothèses.*

$$H_0 : \rho = 0 \quad H_1 : \rho \neq 0$$

où ρ est la corrélation de population entre les heures d'usage quotidien des médias sociaux et le score à une échelle de bien-être.

Bilatéral, parce qu'une association positive serait aussi intéressante qu'une association négative — l'attente courante est négative, et trouver l'inverse serait un résultat de poids.

Pourquoi cela réduit le p-hacking : fixer la direction à l'avance retire l'option de changer de queue après avoir vu le sens des données (§13.2.2).

(b) *Taille de l'échantillon.* Déterminée par un calcul de puissance : détecter une corrélation de $r = 0,20$ — un petit effet plausible dans cette littérature — avec 80 pour cent de puissance à $\alpha = 0,05$ en bilatéral exige environ 200 participants.

Pourquoi cela réduit le p-hacking : cela supprime l'arrêt facultatif. La chercheuse recrute 200 personnes et analyse une fois, plutôt que de vérifier à mesure que les données s'accumulent (§13.5).

(c) *Seuil de signification et test.* $\alpha = 0,05$; corrélation de Pearson entre les heures d'usage quotidien et le score total de bien-être.

Pourquoi cela réduit le p-hacking : préciser le test à l'avance empêche d'en essayer plusieurs et de déclarer celui qui a marché.

(d) *Traitement des valeurs aberrantes.* Les cas déclarant un usage supérieur à 16 heures par jour seront exclus comme invraisemblables ; aucune autre exclusion. Les analyses seront présentées avec et sans ces cas.

Pourquoi cela réduit le p-hacking : les règles d'exclusion sont un degré de liberté classique du chercheur, puisqu'il est plus facile de conclure que « ce cas a l'air faux » quand le retirer aide. Fixer la règle à l'avance, sur un critère de fond, supprime cette latitude.

(e) *Variable de résultat.* Le score total à l'échelle de bien-être est l'unique résultat principal. Les sous-échelles peuvent être examinées, mais seront présentées comme exploratoires.

Pourquoi cela réduit le p-hacking : c'est l'élément le plus important. Tester 5 sous-échelles et déclarer celle qui a atteint la signification est le passage B de la §13.8, et nommer un résultat principal à l'avance l'empêche.

Le fil commun : chaque élément supprime un choix qui pourrait autrement être fait après avoir vu les données. Le préenregistrement fonctionne non pas en surveillant les intentions, mais en liant le chercheur futur à un moment où il n'a aucun intérêt dans le résultat.

Tout plan qui précise les cinq éléments avec des justifications sensées fait l'affaire. Les nombres précis comptent moins que de voir à quoi sert chaque contrainte.

Corrigé de l'exercice 13.E5. Votre réponse emploiera une variable différente de celle ci-dessous ; le raisonnement est le même quelle que soit celle que vous choisissiez.

(a) Énoncez les hypothèses *avant* de calculer quoi que ce soit, et justifiez la valeur de référence plutôt que d'en choisir une proche de la moyenne de l'échantillon. Calculer la moyenne d'abord et choisir ensuite une valeur de référence est exactement la pratique contre laquelle la §13.7 met en garde — et remarquer que vous avez été tentée de le faire vaut mieux que de l'éviter par chance.

(b) Le compte rendu doit contenir les six étapes, avec des conditions réellement vérifiées plutôt que simplement affirmées. Pour *socsurvey*, l'échantillon est assez grand pour que le théorème central limite s'applique confortablement ; la condition de tirage aléatoire tient par construction, puisque les données sont simulées.

Un compte rendu complet donne M , la statistique de test, la valeur p exacte et n .

(c) Pour la plupart des variables de *socsurvey* à $n = 477$, une différence modérée par rapport à une valeur de référence sera significative, et le d de Cohen sera souvent petit.

Le résultat instructif est un résultat significatif assorti d'un petit d — le paradoxe de la taille d'échantillon dans vos propres données. La question est de savoir si cela change l'interprétation, et cela devrait : déclarer les deux, puis dire que l'effet est négligeable, est ce que demande la §13.6.

(d) Les deux moitiés s'accorderont d'ordinaire sur la direction et pourront fort bien se contredire sur la signification, puisque chacune a la moitié de l'échantillon et d'autant moins de puissance.

Ce que cela suggère au sujet des études isolées est le point de l'exercice. Si deux moitiés aléatoires du *même* jeu de données peuvent aboutir à des conclusions différentes, alors deux études indépendantes de la même question le peuvent certainement — et le verdict d'une étude isolée est une preuve plus faible que ne le laisse croire l'assurance de sa rédaction.

C'est le problème de la réplication de la §13.7 démontré au moindre coût possible.

(e) *Réel contre simulé.* Élections Canada rapporte un taux de participation officiel de **69,0 pour cent** des électeurs inscrits à la 45^e élection générale, tenue le 28 avril 2025 (ELECTIONS CANADA 2025).

Le taux de participation de *socsurvey* vaut $321/477 = 0,673$ — proche du chiffre officiel, et proche sans bonne raison. Deux choses méritent d'être séparées.

Les dénominateurs diffèrent. Élections Canada compte des bulletins par rapport à un registre d'électeurs. Une enquête compte les personnes qui *disent* avoir voté, parmi celles qui ont accepté de répondre. La participation autodéclarée est fiablement plus élevée que le chiffre officiel, parce que les personnes qui n'ont pas voté sont à

la fois moins susceptibles de répondre et plus susceptibles de mal se souvenir.

Et l'échantillon est simulé. Quelle que soit la concordance, elle est une propriété de la façon dont socsurvey a été construit, non une preuve au sujet du Canada. Un test de $H_0 : \pi = 0,69$ sur ce jeu de données vérifierait si la simulation concorde avec un chiffre qu'elle a été écrite pour imiter — ce qui est une question sur le code, et il n'existe pas de test d'hypothèse pour cela.