

Corrigé du chapitre 17

Comprendre les données sociales

Mark Paradis, Ph. D.

2 septembre 2026

Voici le corrigé détaillé des exercices de *fin de chapitre* du chapitre 17 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 17 L'analyse de la variance

Corrigé de l'exercice 17.E1. (a) $df_{\text{intergroupe}} = k - 1 = 3$ et $df_{\text{intragroupe}} = N - k = 100 - 4 = 96$.

Ils s'additionnent à $3 + 96 = 99$, qui est $N - 1$. La vérification passe.

	Source	df	Somme des carrés	Carré moyen	F
(b)	Intergroupes	3	180	60	6,26
	Intragroupes	96	920	9,583	
	Total	99	1100		

Chaque carré moyen est sa somme des carrés divisée par ses degrés de liberté, et $F = 60/9,583 = 6,26$, avec $p < 0,001$.

(c) $\eta^2 = 180/1100 = 0,164$.

(d) *Les résultats moyens diffèrent d'un groupe à l'autre sur les quatre groupes*, $F(3, 96) = 6,26$, $p < 0,001$, $\eta^2 = 0,164$. Le groupement rend compte d'environ seize pour cent de la variation, et ce test n'établit pas quels groupes diffèrent.

Corrigé de l'exercice 17.E2. (a) *Test t à échantillons indépendants* (§15.3). Deux groupes, un résultat continu, et une assignation au hasard — la différence est donc un effet plutôt qu'une association.

(b) *Analyse de la variance à un facteur* (ce chapitre). 5 groupes excluent le test t , et faire à la place les dix comparaisons deux à deux porterait un taux par famille d'environ quarante pour cent. Notez que le niveau de scolarité est *ordonné*, ce que l'analyse de la variance ignore ; une régression le traitant comme un prédicteur numérique se servirait de cette information, au prix de supposer les échelons égaux.

(c) *Régression linéaire simple* (chapitre 16). Deux variables continues, et la question est de combien, non de savoir si les groupes diffèrent.

(d) *Test du khi carré d'indépendance* (chapitre 14). Deux variables catégorielles et aucun résultat continu du tout.

(e) *Test t à échantillons appariés* (§15.4). Les mêmes vingt personnes apparaissent deux fois, de sorte que les deux colonnes sont liées cas par cas et que les traiter comme des groupes indépendants serait l'erreur de la §15.4.3.

Corrigé de l'exercice 17.E3. (a) $df_1 = 3$ veut dire $k = 4$ groupes. $df_2 = N - k$, donc $N = 1\,996 + 4 = 2\,000$.

(b)

$$\eta^2 = \frac{45,2 \times 3}{45,2 \times 3 + 1\,996} = 0,064$$

(c) Un F de 45,2 est énorme et serait rapporté comme décisif. L'effet auquel il correspond représente 6,4 pour cent de la variation — réel, et petit. Les deux sont compatibles parce que le F monte avec la taille de l'échantillon et que le η^2 ne le fait pas, et que cette étude compte 2 000 cas.

(d) Qu'ils ont rapporté la taille d'effet, qu'ils l'aient voulu ou non. Le F et ses deux nombres de degrés de liberté déterminent exactement le η^2 , de sorte qu'une lectrice peut toujours le récupérer. L'omission n'est donc pas un manque d'information mais un choix sur ce qu'on met sous les yeux de la lectrice — et la quantité laissée de côté est celle qui parle de l'importance du résultat.

Corrigé de l'exercice 17.E4. (a) L'égalité des variances, et gravement. Le plus grand écart type vaut 9,8 contre un plus petit de 2,1 — un rapport de 4,7, donc les variances — qui sont les carrés — diffèrent d'un facteur $4,7^2 = 22$. La règle empirique de la §17.4.1 demande un rapport des écarts types inférieur à environ 2.

(b) Les moyennes de groupe elles-mêmes ne sont pas touchées : chacune est calculée à partir de son propre groupe et rien ne les met en commun. Ce dont il faut se méfier, c'est du $CM_{\text{intragroupe}}$, et donc du F , de la valeur p et de tout intervalle bâti sur l'erreur type mise en commun. L'estimation survit ; la confiance annoncée, non.

(c) Employer la correction de Welch (§17.4.3). Elle coûtera des degrés de liberté — avec un groupe de six et un groupe bien plus variable que les autres, probablement beaucoup — et cette réduction est la comptabilité honnête de la quantité d'information indépendante réellement présente.

(d) Oui, et cela rend le résultat plus intéressant plutôt que moins. Un groupe témoin en liste d'attente qui varie quatre fois plus que les groupes traités laisse penser que les traitements ont fait quelque chose que la seule comparaison des moyennes ne saisit pas : ils ont rendu les gens plus semblables. Une différence de variance est un résultat, non seulement un inconvénient, et elle vaut la peine d'être rapportée comme tel.

Corrigé de l'exercice 17.E5. (a) D'abord, les trois tests t sont sans correction, de sorte que leur taux d'erreur par famille est de 14,3 pour cent plutôt que de 5. Ensuite — et c'est pire — seuls les deux significatifs ont été rapportés. Choisir quelles comparaisons publier d'après leurs résultats fait que les valeurs p rapportées ne décrivent ni les tests qui ont été faits ni ceux qui ne l'ont pas été.

(b) Cela réglerait le second problème et non le premier. Rapporter les trois tests sans correction est honnête quant à ce qui a été fait et sous-estime encore le taux d'erreur.

(c) Le test omnibus, qu'elle a déjà, suivi d'une démarche post hoc rapportant *toutes* les 3 comparaisons avec des valeurs p ajustées et des intervalles — l'analyse de la §17.5.3. Si les comparaisons qui l'intéressaient avaient été précisées avant la collecte des données, elle pourrait plutôt les faire comme des comparaisons planifiées et le dire (§17.5.4).

Corrigé de l'exercice 17.E6. Vos chiffres dépendent du téléchargement que vous employez, mais la forme du raisonnement, non. Ce à quoi s'attendre, et pourquoi :

(a) Le groupe de revenu est fortement lié à l'espérance de vie dans *worldsim*, de sorte que le F est très grand et le η^2 élevé — bien plus élevé que tout ce qui figure ailleurs dans ce chapitre. Rapportez les deux nombres de degrés de liberté, qui seront $k - 1$ et $N - k$ pour les 52 pays de la vague 1.

(b) Plus raisonnable que pour *region*, sans être parfaite : les groupes de revenu sont de tailles inégales et le groupe le plus pauvre a la plus grande dispersion d'espérance de vie. Comparez les écarts types plutôt que de les tester.

(c) Peu de chose. Welch coûtera quelques degrés de liberté et la conclusion ne bougera pas, ce qui est le cas que la §17.4.3 décrit comme l'ordinaire — et la raison d'employer la correction par défaut plutôt que seulement quand elle change quelque chose.

(d) *region* donne un η^2 plus petit que *income_grp* parce que les régions sont intérieurement hétérogènes exactement de la façon qui gonfle la $SC_{\text{intragroupe}}$: une région contient des pays à des niveaux de revenu très différents, et c'est le revenu qui détermine l'espérance de vie ici. Grouper selon quelque chose de plus proche de la cause met davantage de la variation entre les groupes et moins à l'intérieur d'eux. La variable de groupement est un choix, et le η^2 mesure en partie la qualité de ce choix.

Corrigé de l'exercice 17.E7. Vos chiffres dépendent de l'année de recensement et des provinces que vous choisissiez, mais ce que vous devriez pouvoir en dire, non. Ce à quoi s'attendre, et pourquoi :

(a)–(b) Rapportez le F avec ses deux nombres de degrés de liberté, la valeur p et le η^2 . Avec trois provinces et une poignée de régions métropolitaines chacune, le N sera petit — peut-être quinze ou vingt — de sorte que les degrés de liberté seront petits et la valeur critique d'autant plus grande.

(c) Vous ne traiteriez pas le η^2 comme une estimation d'une valeur de population : à cette taille d'échantillon il est gravement biaisé vers le haut (§17.5.2), et une grande valeur est le résultat attendu, que les provinces diffèrent réellement ou non. Vous le rapporteriez comme une description de ces régions et vous le diriez.

(d) Deux raisons, et elles tirent en sens contraires. Le revenu médian d'une région métropolitaine est une moyenne sur des centaines de milliers de personnes, de sorte que le bruit qui rend petit le η^2 à l'échelle individuelle a été moyenné et effacé — ce qui gonfle le η^2 et fait paraître les groupements régionaux bien plus nets. Mais le F dépend aussi de la taille de l'échantillon, et 39 085 individus donnent un F de 420 là où vingt régions ne peuvent pas en approcher. L'analyse à l'échelle des régions montrera donc typiquement un η^2 *plus grand* et un F *plus petit* — ce qui est la démonstration la plus claire de ce livre que les deux répondent à des questions différentes.