

Corrigé du chapitre 12

Comprendre les données sociales

Mark Paradis, Ph. D.

2 septembre 2026

Voici le corrigé détaillé des exercices de *fin de chapitre* du chapitre 12 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 12 Inférence et estimation

Corrigé de l'exercice 12.E1. (a) L'erreur type :

$$s_{\bar{x}} = \frac{0,94}{\sqrt{400}} = \frac{0,94}{20} = 0,047$$

La marge d'erreur :

$$ME = 1,97 \times 0,047 = 0,093$$

L'intervalle :

$$3,62 \pm 0,093 \Rightarrow (3,53, 3,71)$$

En format APA :

$$M = 3,62, \text{ IC à } 95 \% [3,53, 3,71], n = 400$$

(b) Oui — et plus que compatible.

La cible de 3,50 tombe *sous* l'intervalle tout entier, dont la borne inférieure est 3,53. Les données ne sont donc pas seulement compatibles avec une population égale ou supérieure à 3,50 : elles sont incompatibles avec une population *inférieure* à 3,50.

La distinction compte. Une valeur à l'intérieur de l'intervalle voudrait dire que les données ne peuvent pas écarter la cible. Une valeur sous l'intervalle veut dire que les données indiquent que la population la dépasse.

Notez l'étroitesse de la marge : 3,50 rate la borne inférieure de 0,03, de sorte que c'est une conclusion que l'échantillon appuie plutôt qu'une qu'il établit de façon décisive.

(c) En employant la formule de taille d'échantillon :

$$n = \left(\frac{1,96 \times 0,94}{0,05} \right)^2 = \left(\frac{1,842}{0,05} \right)^2 = (36,85)^2 = 1\,357,8$$

En arrondissant vers le haut, $n = 1\,358$.

Il vaut la peine de remarquer à quel point cette précision coûte cher. Passer d'une marge d'erreur de 0,093 à une de 0,05 — soit moins que la diviser par deux — exige de faire passer l'échantillon de 400 à 1 358, soit 3,4 fois plus de répondants.

Corrigé de l'exercice 12.E2. (a) L'intervalle de l'équipe A est le plus large : il couvre 11,8 points contre 5,8 pour l'équipe B et 3,0 pour l'équipe C.

La raison est la taille de l'échantillon. L'erreur type vaut $\sigma/\sqrt{n}$, donc :

$$n = 25 \Rightarrow \sigma_{\bar{x}} = \frac{15}{5} = 3,0$$

$$n = 100 \implies \sigma_{\bar{x}} = \frac{15}{10} = 1,5$$

$$n = 400 \implies \sigma_{\bar{x}} = \frac{15}{20} = 0,75$$

Chaque multiplication de n par quatre divise l'erreur type par deux, et les erreurs types ci-dessus se divisent exactement : 3,0, 1,5, 0,75.

Les largeurs des intervalles héritent de cette division, mais à l'arrondi près. Les largeurs exactes sont 11,76, 5,88 et 2,94, qui se divisent proprement ; soustrayez les bornes *telles qu'imprimées* et vous obtenez plutôt 11,8, 5,8 et 3,0, parce que chaque borne est donnée à une décimale et qu'une largeur porte l'arrondi de deux bornes. Le motif est réel ; c'est l'affichage qui le brouille, ce qui vaut la peine de savoir avant de conclure qu'un tableau publié contient une erreur d'arithmétique.

(b) En vérifiant chaque intervalle contre $\mu = 50$:

Équipe	Intervalle	Contient $\mu = 50$?
A	(42,2, 54,0)	Oui
B	(48,9, 54,7)	Oui
C	(47,9, 50,9)	Oui

Les trois contiennent la vraie valeur. La question invite à vérifier plutôt qu'à supposer, et la vérification est le point.

Si l'un avait raté, cela n'indiquerait pas une erreur. Une procédure à 95 pour cent échoue dans environ un échantillon sur vingt, par construction, et une équipe dont l'intervalle rate n'a rien fait de mal — elle a tiré un échantillon inhabituel, ce qui arrive à une fréquence connue. C'est exactement ce que montre la figure 12.3, où un intervalle sur vingt rate.

Ce qui serait une erreur, c'est de conclure d'un seul intervalle raté que la méthode est brisée, ou que l'équipe a été négligente.

Remarquez aussi que l'intervalle de l'équipe C est le plus étroit et celui qui passe le plus près d'exclure μ — sa borne supérieure est 50,9. Un plus grand échantillon donne une estimation plus précise, ce qui veut aussi dire moins de marge avant que l'intervalle rate. La précision et le risque de passer tout près voyagent ensemble.

(c) Environ 5.

Un intervalle de confiance à 95 pour cent est construit de telle sorte que 95 pour cent de tels intervalles contiennent le paramètre, donc 5 pour cent ne le contiennent pas. Sur 100 équipes :

$$100 \times 0,05 = 5$$

C'est une espérance plutôt qu'une garantie. Le nombre réel varierait — trois, ou sept, ou parfois davantage — parce que le compte est lui-même soumis à la variation d'échantillonnage. Ce que les 95 pour cent garantissent, c'est le taux à long terme, non le résultat d'un ensemble particulier de cent études.

Corrigé de l'exercice 12.E3. L'affirmation contient une intuition défendable et deux erreurs.

Ce que veut dire la marge d'erreur. C'est la demi-largeur d'un intervalle de confiance autour d'une estimation. Une marge de ± 3 points sur une candidate qui recueille 50 pour cent veut dire que l'intervalle pour l'appui à cette candidate va de 47 à 53, et que les intervalles construits ainsi capturent le vrai chiffre dans 95 pour cent des échantillons.

Erreur 1 : appliquer la marge d'erreur à une différence. Le journaliste compare l'écart entre deux candidats à une marge d'erreur calculée pour chacun séparément. Ce n'est pas la bonne quantité. Une différence entre deux estimations a sa propre erreur type, plus grande que celle de chacune, de sorte que le seuil pour distinguer un écart de zéro est plus grand que la marge annoncée — environ 1,5 fois celle-ci dans les situations de sondage usuelles.

La règle du journaliste est donc prudente, et prudente pour la mauvaise raison. Elle se trouve à exiger plus de preuve qu'une comparaison naïve, tout en reposant sur un calcul qui ne s'applique pas.

Erreur 2 : « à égalité statistique ». Un intervalle qui inclut zéro n'établit pas que la différence est nulle. Il dit que les données ne peuvent pas distinguer la différence de zéro — un énoncé sur la preuve, non sur la course.

Une candidate qui mène 51–47 est plus probablement devant que derrière. Le sondage manque simplement de précision pour le confirmer. Appeler cela une égalité affirme quelque chose que les données n'appuient pas, dans la direction opposée à l'erreur que l'expression voulait éviter.

Une réécriture défendable. « A mène par 4 points, ce qui est à l'intérieur de la marge d'erreur du sondage — l'avance apparaît donc dans les données, mais cet échantillon n'est pas assez précis pour la confirmer. »

L'erreur d'échantillonnage n'est en outre qu'une source d'erreur parmi d'autres dans un sondage. La non-réponse, la couverture et la formulation des questions (§5.4) ne sont pas du tout captées par la marge d'erreur, et le chiffre annoncé sous-estime donc l'incertitude totale.

Corrigé de l'exercice 12.E4. (a)

$$n = \left(\frac{1,96 \times 18\,000}{2\,000} \right)^2 = \left(\frac{35\,280}{2\,000} \right)^2 = (17,64)^2 = 311,2$$

En arrondissant vers le haut, $n = 312$ ménages.

(b) Avec $n = 200$:

$$\sigma_{\bar{x}} = \frac{18\,000}{\sqrt{200}} = \frac{18\,000}{14,14} = 1\,273$$

$$ME = 1,96 \times 1\,273 = 2\,495$$

Environ 2 500 \$ — nettement pire que les 2 000 \$ qu'elle voulait. Le manque de taille d'échantillon est de 36 pour cent et la perte de précision d'environ 25 pour cent, ce qui est la relation en racine carrée à l'œuvre.

(c) À 99 pour cent de confiance, la valeur critique monte à 2,576 :

$$n = \left(\frac{2,576 \times 18\,000}{2\,000} \right)^2 = (23,18)^2 = 537,5$$

En arrondissant vers le haut, $n = 538$ ménages.

La comparaison entre les trois parties est la leçon. En maintenant la marge d'erreur à 2 000 \$, porter la confiance de 95 à 99 pour cent fait passer l'échantillon nécessaire de 312 à 538 — une hausse de 72 pour cent du travail de terrain pour 4 points de pourcentage de confiance.

Avec un budget de 200 elle ne peut atteindre ni l'une ni l'autre des cibles. Ses options réalistes sont d'accepter une marge d'erreur voisine de 2 500 \$ et de la rapporter clairement, ou de chercher un financement pour le plus grand échantillon. Ce qu'elle ne doit pas faire, c'est rapporter la précision de 2 000 \$ qu'elle avait planifiée.

Corrigé de l'exercice 12.E5. *Le lien.* Un intervalle de confiance contient exactement les valeurs qui sont compatibles avec les données au niveau annoncé. Un test d'hypothèse demande si une valeur précise est compatible avec les données. L'intervalle répond donc, pour toutes les valeurs possibles à la fois, à la question à laquelle un test répond pour une seule.

Concrètement : si une valeur proposée tombe à l'extérieur d'un intervalle à 95 pour cent, un test bilatéral de cette valeur à $\alpha = 0,05$ la rejeterait. Si elle tombe à l'intérieur, le test ne rejeterait pas.

Étant donné l'intervalle (42,3, 57,8) :

$\mu_0 = 40$. Cette valeur tombe à l'extérieur de l'intervalle, sous sa borne inférieure de 42,3. Les données sont donc incompatibles avec une moyenne de population de 40, et un test de cette hypothèse nulle la rejeterait à $\alpha = 0,05$.

$\mu_0 = 50$. Cette valeur tombe à l'intérieur de l'intervalle, et près de son centre. Les données sont compatibles avec une moyenne de population de 50, et un test ne rejeterait pas cette hypothèse nulle.

Pourquoi aucun test formel n'est nécessaire. Parce que l'intervalle encode déjà la réponse. Il a été construit à partir de la même moyenne d'échantillon et de la même erreur type qu'un test emploierait, et la borne de l'intervalle est précisément le point où un test changerait de verdict.

Deux mises en garde valent la peine d'être ajoutées.

Ne pas rejeter $\mu_0 = 50$ n'établit pas que μ vaut 50. L'intervalle contient aussi 45, 52 et 57, et les données sont également compatibles avec toutes ces valeurs. C'est l'asymétrie que développe le chapitre 13 : la preuve peut écarter des valeurs et ne peut en confirmer aucune.

Et cette correspondance vaut pour un test bilatéral au niveau correspondant. Un test unilatéral, ou un α différent, correspond à un intervalle construit autrement.

L'intervalle offre quelque chose qu'un test n'offre pas : un test rend un verdict sur la seule valeur que la chercheuse a choisi de nommer, alors que l'intervalle permet à une lectrice de vérifier n'importe quelle valeur qui l'intéresse — ce qui est l'argument en faveur d'une diffusion centrée sur l'estimation à la §12.4.5.

Corrigé de l'exercice 12.E6. Votre réponse emploiera une autre variable que celle ci-dessous ; le raisonnement est le même quelle que soit celle que vous choisissiez. La version travaillée emploie la variable *income* de *socsurvey*, en milliers de dollars.

(a) Trois chiffres, avec trois sens différents.

$$\bar{x} = 90,4 \quad s = 48,2 \quad s_{\bar{x}} = \frac{48,2}{\sqrt{477}} = 2,21$$

La *moyenne* estime la moyenne de la population. L'*écart type* décrit de combien les répondants individuels varient. L'*erreur type* décrit de combien la moyenne de l'échantillon varierait sur des échantillonnages répétés de 477.

Remarquez la taille du troisième contre le deuxième : l'erreur type vaut environ $1/\sqrt{477}$ — soit à peu près un 22^e — de l'écart type. Les estimations de la moyenne sont bien plus stables que la variable qu'elles résument.

(b) Avec $t^* = 1,96$:

$$90,4 \pm 1,96 \times 2,21 \Rightarrow (86,1, 94,8)$$

Rapporté dans le format de la §12.4.1, cela donne $M = 90,4$, IC à 95 % $[86,1, 94,8]$, $n = 477$. Les quatre éléments sont exigés : un intervalle sans son n , ou sans son niveau de confiance, ne peut pas être jugé par une lectrice.

(c) Attendez-vous à ce que neuf ou dix des 10 contiennent la moyenne de l'échantillon complet, puisqu'une procédure à 95 pour cent rate environ une fois sur vingt.

Les résultats instructifs sont les résultats inattendus. Dix sur dix est courant — avec seulement 10 intervalles, voir zéro raté est plus probable que le contraire. Huit sur dix arrive aussi, et n'indique pas une erreur.

Le point est que 95 pour cent est un taux à long terme, et que 10 essais sont bien trop peu pour l'observer de façon fiable. C'est le même point qu'à l'exercice 12.E2(c), atteint cette fois dans des données que vous avez tirées vous-même.

(d) Les intervalles deviennent *plus étroits* — d'un facteur d'environ $\sqrt{100/30} = 1,83$ — parce que l'erreur type rétrécit avec la racine carrée de la taille de l'échantillon.

Ce qui ne change *pas*, c'est le taux de couverture. À peu près la même proportion d'intervalles contiendra la moyenne de l'échantillon complet, parce que le niveau de confiance gouverne la fréquence à laquelle la procédure réussit, indépendamment de n . Un plus grand échantillon achète de la précision, non de la fiabilité.

Cette distinction est la chose la plus utile de l'exercice, et c'est celle qu'on attend le plus souvent dans l'autre sens.

(e) Le programme des déclarants de Statistique Canada situe le revenu moyen après impôt des familles canadiennes à 82,6 milliers de dollars en 2023 (STATISTICS CANADA 2025). L'intervalle construit en (b) part de 86,1 milliers vers le haut, de sorte que le chiffre publié tombe à *l'extérieur*, de 3,5 milliers.

Ce que cela ne montre pas. Ce n'est pas une preuve que Statistique Canada se trompe, et ce n'est pas une preuve que l'échantillon est non représentatif d'une façon que vous pourriez corriger. L'intervalle est un énoncé sur la population d'où socsurvey a été tiré — et socsurvey n'a pas été tiré du Canada. Il a été simulé. Un intervalle de confiance ne peut jamais porter que sur le processus qui a engendré les données, et ici ce processus était un générateur de nombres aléatoires.

Ce que cela montre. Exactement à quel point cette erreur est facile à commettre. Chaque étape était correcte : l'arithmétique, le format, l'interprétation. L'intervalle est un intervalle à 95 pour cent valable. Il ne porte simplement pas sur le Canada, et rien à l'intérieur du calcul ne le dit — c'est pourquoi la question de savoir de quelle population un échantillon provient se règle avant que cette arithmétique commence (§5.1.2).

Il y a un second écart qui mérite d'être nommé : le chiffre publié porte sur des *familles*, alors que l'unité de socsurvey est un répondant individuel. Deux chiffres appelés « revenu moyen » n'ont pas à être des moyennes de la même chose.