Chapitre 1
Tirage aléatoire simple

1.1 Introduction

On effectue un échantillonnage lorsqu'on tente de connaître certaines caractéristiques d'une *population*, en n'en observant qu'une partie, un *échantillon*. La population peut être de natures très diverses : μ

- L'ensemble des résidents du Canada. C'est une population dans le sens courant du terme.
- L'ensemble des enfants à l'école primaire au Québec
- L'ensemble des fermes avicoles de l'Estrie.
- L'ensemble des employés d'Hydro-Québec.
- L'ensemble des pièces fabriquées dans un lot.
- L'ensemble des membres d'une Caisse populaire.

L'échantillonnage, qui connaît aujourd'hui un essor considérable, est devenu le moyen privilégié d'obtenir l'information nécessaire à la gestion d'un pays ou d'une entreprise. Dans la plupart des pays, dont le Canada, des agences gouvernementales maintiennent à jour des données sur l'état et les activités de la nation. Statistique Canada prélève régulièrement des données de toutes sortes :

- démographiques (distribution de l'âge, du sexe, de l'état matrimonial, des naissances et des décès, tailles et composition des familles);
- agricoles (production de diverses denrées, surfaces cultivées, taille du cheptel);
- industrielles et commerciales (production, stocks, exportations, effectifs, rendement); et
- des données liées au travail (taux de chômage, salaires, effectifs par branches d'activité).

Certaines de ces données sont issues du Recensement, une opération au cours de laquelle on tente d'atteindre *tous* les ménages de la population. Mais un recensement est très coûteux ; on ne le fait pas plus d'une fois par cinq ou dix ans. Entre les recensements, la collecte d'information se poursuit au moyen d'échantillonnage. (Le Recensement lui-même est en partie basé sur l'échantillonnage, puisqu'un bon nombre des questions ne sont posées qu'à une partie de la population).

Les sondages d'opinion sont probablement les exemples d'échantillonnage les plus médiatisés et les mieux connues du grand public. Les maisons de sondages sont mises en vedette pour leurs sondages politiques. Mais les sondages politiques ne constituent qu'une minuscule partie de leurs activités: leurs plus gros contrats ont plutôt à faire avec les goûts et préférences des consommateurs.

En plus des données prélevées pour les besoins de l'état et des compagnies, l'échantillonnage permet d'accumuler les éléments d'information, de nature scientifique ou médicale, qui éclairent les politiques sociales. Les affirmations au tableau 1.1.1 illustrent le type d'information qui ne peut provenir que d'un échantillonnage. Il est évident que beaucoup d'incertitude entoure ces proposi-

tions. Certaines sont sans doute carrément fausses; d'autres probablement erronées. La plupart d'entre elles sont probablement modérément erronées. Jusqu'à quel point peut-on se fier à des conclusions comme celles-ci? Quelle est la marge d'erreur des quantités énoncées? La fiabilité des conclusions dépend de la méthodologie employée, et de la taille des échantillons. Le rôle de la statistique est de proposer des procédures fiables, et, à l'aide de la théorie des probabilités, d'estimer les risques d'erreur.

Nous n'allons pas pousser à fond l'étude, extrêmement complexe, des méthodes possibles de sélection et d'analyse d'un échantillon à l'échelle du monde ou d'un état. Nous allons, cependant, pouvoir présenter les idées essentielles de la théorie de l'échantillonnage dans le cadre plus restreint des projets d'échantillonnage que l'on rencontre normalement dans une entreprise.

1.2 Population et paramètres

Le but d'un échantillonnage est d'estimer certaines caractéristiques d'une population à partir des données d'un échantillon. Ces caractéristiques sont mesurées par ce qu'on appelle des *paramètres* de la population. Voici certains des paramètres qu'on traitera dans ces notes:

- La *moyenne* $\bar{y}_U$ d'une population: par exemple, le revenu moyen des ménages d'un certain quartier;
- Le *total t* d'une population: par exemple, la production totale de blé dans les fermes d'une certaine région;
- Une *proportion p*: par exemple, la proportion des employés d'une compagnie qui serait favorable à un plan de soins dentaires;
- Un *effectif* N_c: par exemple, le *nombre* d'employés favorables à un plan de soins dentaires;
- Un *quotient R:* par exemple, le nombre de postes de radio par personne dans les ménages d'une population de ménages.
- Un *écart-type S*: par exemple , l'écart-type des longueurs de boulons dans un lot.

Ces paramètres ne peuvent être calculés que si l'on a accès à *toutes* les unités de la population. Lorsque cela est impossible, on recourt à l'échantillonnage: on tire un échantillon, c'est-à-dire, une *partie* de la population, et on s'appuie sur les données de l'échantillon pour *estimer* les paramètres de la population. Cette estimation est forcément approximative et sujette à erreur. La science de l'échantillonnage consiste à mesurer les risques, à fournir des méthodes qui les minimisent, et à déterminer la taille de l'échantillon qui permet de limiter l'erreur probable à une certaine marge raisonnable.

Tableau 1.1.1

Quelques conclusions issues d'un échantillonnage

- Il n'y a plus que 25 rhinocéros blancs au Parc national Garamba au nord-est du Congo.
- 20% des puits d'eau potable au Bangladesh sont contaminés par des quantités excessives d'arsenic.
- Les opérations minières déversent 250 tonnes de mercure chaque année dans l'Amazone.
- 85% des ordinateurs du monde sont munis du système d'opérations Windows 95.
- Le médicament Tamoxifen a un effet bénéfique sur le cancer du sein, mais peut provoquer le cancer de l'utérus.
- On estime à 16 000 le nombre de satanistes pratiquants en Grande-Bretagne.
- Il y a 500 000 mines plantées en Angola.
- En moyenne, les personnes qui cessent de fumer, ont un gain de poids de 11,25 kg.
- La population féline tue 75 millions d'oiseaux chaque année en Grande-Bretagne.
- Les fétus féminins exercent leurs mâchoires 30% de plus que les fétus masculins.
- 39% des vaches laitières aux États-Unis ont des selles contaminées par le microbe *Campylobacter*.
- Il ne reste plus que 800 lynx ibériques en Espagne et au Portugal.
- 750 millions de personnes dans les pays en voie de développement souffrent du goitre.
- L'âge de la mère est inférieur à 17 ans dans 40% des naissances en Afrique.
- 120 millions de filles et de femmes dans le monde ont subi une excision génitale.
- À chaque minute, une femme meurt des complications d'un grossesse ou d'un accouchement.
- Il y a 20 millions d'avortement illicites dans le monde chaque année. 95% de ceux-là sont dans des pays en voie de développement.
- Il y a 1,06 milliards de jeunes dans le monde; et il y aura 1,42 milliards de personnes âgées en 2050.
- Si la fortune de Bill Gates était empilée en billets de 100$, la pile aurait une hauteur de 250 kilomètres.
- Il y a 300 millions de fumeurs en Chine; 73 % des hommes vietnamiens sont des fumeurs.
- Dans le monde, 800 000 personnes subissent un pontage coronarien chaque année.
- Les femmes qui ont pris la pilule contraceptive ont 25% moins de chance de subir une fracture de la hanche.
- 31,2 % des Japonais se servent d'un cellulaire; le chiffre est de 99% pour les jeunes Finlandais de 18 à 24 ans.
- L'usure des pneus en Europe génère 40 000 tonnes d'un cancérigène reconnu.
- En Australie, 80% des jeunes de 14 à 16 ans se mettent régulièrement au régime.
- La famine qui a sévi en Corée du Nord ces trois dernières années a fait 2 500 000 victimes.
- La population mondiale atteindra 7,7 milliards en l'an 2040.
- 79 % des Nord Américains sont branchés à l'Internet.
- Il y a 14 millions de cochons au sud des Pays-Bas.
- 42% des Britanniques admettent avoir commis des infidélités conjugales, contre 38% des Italiens et 22% des Espagnols.
- La Communauté européenne contribue 41% à l'effet de serre dans le monde.
- 447 millions d'Africains n'auront pas accès à de l'eau potable en l'an 2000 si les gouvernements n'investissent pas davantage.
- Il y aura 20 millions de nouveaux cas de cancer par année en l'an 2020.
- L'espérance de vie au Zimbabwe est 39 ans.
- La campagne menée par les Nations Unies a réduit de 90% le nombre de nouveaux cas de polio.
- Il y a 1 milliard de bouteilles de champagnes stockées dans les caves françaises.
- Un Cambodgien sur 250 a perdu des bras ou des jambes à cause des mines.
- 4 millions d'enfants meurent chaque année de maladies qui auraient pu être pu être prévenues par des vaccins.
- Le nombre annuel de décès dus à l'asthme est estimé à 180 000.
- 90% des pensionnaires des orphelinats russes ont au moins un parent vivant.
- Chaque année, 3 millions d'enfants naissent avec des déformations génitales majeures. La plupart en meurent durant les trois premières années de leur vie.
- 20 % des Japonais croient aux prédictions de Nostradamus.

Nous commencerons, dans ce chapitre, par définir ces concepts dans un cadre précis, celui de l'estimation d'une moyenne $\overline{y}_U$. Dans le prochain chapitre, nous utiliserons la théorie développée dans celui-ci pour développer des techniques propres aux autres paramètres mentionnés ci-dessus.

Considérons donc une population $\mathcal{P}$ de N unités, à chacune desquelles est associée la valeur (monétaire, souvent, mais pas toujours) d'une certaine variable. Dénotons par

$$y_1, y_2, y_3, \ldots, y_N$$

les N valeurs. Supposons qu'on prélève un échantillon ω de taille n de cette population, et qu'on

observe les données échantillonnales $\{y_i | i \in \omega\}$.

1.3 Estimation ponctuelle et par intervalle de confiance

Pour concrétiser, supposons qu'on s'intéresse à la population présentée au tableau A.01 en annexe, une population de professeurs dont on ne connaît que la taille $N = 200$. On s'intéresse au salaire Y : on voudrait en estimer la moyenne $\bar{y}_U$ sans recenser la population entière. On prélève donc un échantillon de taille $n = 50$. Les 50 valeurs $y_1, \ldots, y_{50}$ de la variable Y = salaire en 2001 sont présentées au tableau 1.3.1.

Tableau 1.3.1
***Échantillon de 50 professeurs tiré de la population du tableau* A.1**
(*Données extraites du tableau* A.02)

36098	50000	33659	51951	40366	38659	54268	41198	35000	44044
35854	33537	51951	56951	58780	54146	56220	32660	51951	36951
35976	42195	54756	51585	54390	44044	51707	62561	45854	64146
60000	41098	65732	53049	41951	45467	29878	57683	58171	50000
62195	54390	51220	54146	36929	52927	28390	56463	58902	62317

Estimation ponctuelle

On se contentera donc d'une *estimation* de $\bar{y}_U$. C'est-à-dire, on calculera la moyenne $\bar{y}_\omega$ des données de l'échantillon au lieu de calculer la moyenne réelle $\bar{y}_U$. Pour l'échantillon ci-dessus, nous obtenons la valeur suivante de $\bar{y}_\omega$

$$\bar{y} = 48\,447{,}32$$

Nous affirmerons donc que la moyenne de la population est 48 447,32 $— sachant fort bien, pourtant, que ce chiffre n'est qu'une *estimation* de la moyenne réelle $\bar{y}_U$, et que cette estimation est probablement erronée.

Un calcul basé sur les données d'un échantillon, telle la moyenne échantillonnale $\bar{y}_\omega$, est ce qu'on appelle un *estimateur*, ou *estimateur ponctuel* (l'adjectif « ponctuel » est utilisé par opposition à « par intervalle de confiance », une notion qui sera définie plus bas). Ainsi, $\bar{y}$ est un estimateur ponctuel de μ:

$\bar{y}_\omega$: *Estimateur ponctuel de* $\bar{y}_U$.

Intervalle de confiance

L'estimation ponctuelle faite, il importe de déterminer une marge d'erreur, question de savoir à quel point l'estimation peut être erronée. Un indice de la marge d'erreur est donné par un *intervalle de confiance*, un intervalle dont on peut dire avec un certain degré de confiance qu'il con-

tient la valeur réelle du paramètre. Voici une formule pour déterminer un intervalle de confiance (nous en discuterons l'origine plus bas) :

$$\bar{y}_\omega - 1{,}96\hat{\sigma}_{\bar{y}} \leq \bar{y}_U \leq \bar{y}_\omega + 1{,}96\hat{\sigma}_{\bar{y}}$$

où

$$\hat{\sigma}_{\bar{y}} = \sqrt{1-f}\,\frac{s}{\sqrt{n}},$$

$f = n/N$ est appelé *fraction d'échantillonnage* et

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

est l'écart-type de l'échantillon :

Avec l'échantillon du tableau 1.3.1 nous obtenons

$$s = 10\,025{,}95,$$

et par conséquent

$$\hat{\sigma}_{\bar{y}} = \sqrt{1-\frac{50}{200}}\,\frac{10025{,}95}{\sqrt{50}} = 1\,227{,}92.$$

L'intervalle de confiance est donc

$$48\,447{,}32 - 1{,}96(1\,227{,}92) \leq \bar{y}_U \leq 48\,447{,}32 + 1{,}96(1\,227{,}92)$$

$$48\,447{,}32 - 2406{,}73 \leq \bar{y}_U \leq 48\,447{,}32 + 2406{,}76$$

$$46040{,}59 \leq \bar{y}_U \leq 50854{,}05$$

On peut affirmer avec une certaine confiance que

$$\bar{y}_U \text{ se situe entre } 46040{,}59 \text{ \$ et } 50854{,}05\text{\$};$$

ou encore, que

$$\bar{y}_U = 48\,447{,}32 \text{ \$} \pm 2406{,}73 \text{ \$}.$$

Le nombre 2406,73 $ est la *marge d'erreur*.

Remarque *Le facteur* $\sqrt{1-f}$ *, est appelé* facteur de correction pour population finie. *Si la population était infinie (N = ∞), on aurait f = n/N = 0 et le facteur disparaîtrait de la formule. En pratique, N est souvent très grand, la fraction d'échantillonnage f est minuscule, le facteur de correction est négli-*

geable et peut être omis.

***Exemple* 1.3.1** *Estimation d'une moyenne*

D'une population de N = 8427 comptes à recevoir, on prélève un échantillon de taille n = 30 afin d'estimer la valeur moyenne des comptes. Voici les résultats, en dollars:

240,82	232,50	740,81	860,32	224,10	7,15	324,12	240,12	190,08	182,75
160,21	148,22	132,19	119,25	113,85	108,30	107,10	101,19	99,21	93,12
88,13	80,15	78,13	72,15	67,13	65,14	41,10	32,17	10,02	9,15

a) Estimer la moyenne $\bar{y}_U$ de la population

b) Déterminer un intervalle de confiance pour la moyenne $\bar{y}_U$ de la population.

Solution

a) Nous avons

$$\sum_{i\in s} y_i = 4\,968{,}68\,, \quad s = \sqrt{35930{,}69}\,, \quad \hat{\sigma}_{\bar{y}} = \sqrt{1-\frac{30}{8427}}\,\frac{189{,}55}{\sqrt{30}} = 34{,}55.$$

Le facteur de correction n'est pas important ici: s'il avait été omis, on aurait eu

$$\hat{\sigma}_{\bar{y}} = \frac{189{,}55}{\sqrt{30}} = 34{,}61,$$

assez proche de la valeur 34,55 calculée plus haut.

b) $165{,}62 - 1{,}96(34{,}55) \le \bar{y}_U \le 165{,}62 + 1{,}96(34{,}55)$, soit

$$97{,}9 \le \bar{y}_U \le 233{,}3.$$

Ceci signifie qu'on peut affirmer, avec un certain degré de confiance, que la valeur moyenne de la population se situe entre 97,9 $ et 233,3 $. ■

L'intérêt d'une estimation par intervalle de confiance est qu'elle est fort probablement correcte—contrairement à une estimation ponctuelle qui, elle, est presque certainement fausse : Si on dit « $\bar{y}_U$ est *égal* à 48 447,32 $ » on dit, à moins d'une chance inouïe, une fausseté. Tandis que si on dit, modestement, « $\bar{y}_U$ se situe entre 46 040,59 $ et 50 854,05$ », on a probablement raison. On peut le dire avec un certain *niveau de confiance*. Quel est le sens exact de cette notion de *niveau de confiance*?

Niveau de confiance

Quel est, précisément, le degré de confiance associé à la proposition « $\bar{y}_U$ se situe entre 46 040,59 $ et 50 854,05 $ »? Conventionnellement, on attribue à cet intervalle un niveau de confiance de 95 %, c'est-à-dire, on affirme « avec 95% de confiance » que « $\bar{y}_U$ se situe entre 45 991,48 $ et 50 903,16 $ ». L'intervalle est d'ailleurs appelé ***intervalle de confiance à* 95 %**. Voici ce que ce pourcentage signifie.

Nous suivons une certaine procédure. Elle consiste à construire l'intervalle [$\bar{y}_\omega - 1{,}96\,\hat{\sigma}_{\bar{y}}$; $\bar{y}_\omega + 1{,}96\,\hat{\sigma}_{\bar{y}}$] à partir des données de l'échantillon, et puis de dire que $\bar{y}_U$ est dedans. On ne peut pas être sûr que m y sera, puisque cet intervalle est aléatoire : il varie d'un échantillon à l'autre. Dépendant des données de l'échantillon, il peut contenir $\bar{y}_U$ ou pas. S'il contient m (comme dans l'exemple actuel), l'affirmation qu'on a faite est vraie. S'il ne contient pas m, l'affirmation est fausse. Quelle est la probabilité que l'intervalle contienne μ? Cette probabilité est précisément ce qu'on entend par *niveau de confiance* :

Niveau de confiance ≡ *Probabilité que l'intervalle de confiance contienne le paramètre*

Le problème est de savoir quelle est, précisément, cette probabilité de recouvrement. La réponse n'est pas sans équivoque :

La probabilité que l'intervalle
[$\bar{y}_\omega - 1{,}96\,\hat{\sigma}_{\bar{y}}$; $\bar{y}_\omega + 1{,}96\,\hat{\sigma}_{\bar{y}}$]
recouvre $\bar{y}_U$ *est d'environ 95 % lorsque n est grand*

Dans la prochaine section nous devrons apporter quelques précisions et faire quelques mises en garde, car le chiffre conventionnel de 95% est approximatif.

1.4 Justification théorique des formules

La procédure décrite dans la section précédente est fondée sur une théorie dont nous présentons ici ses grandes lignes. Pour rendre concrète la discussion nécessairement théorique qui suit, nous nous appuierons sur l'exemple déjà traité, celui de la population présentée au tableau A.01. Nous supposons que la population est entièrement connue (chose impossible en pratique; n'oubliez pas qu'il s'agit ici d'un traitement purement théorique.) Elle est constituée de $N = 200$ professeurs, et on s'intéresse au salaire en 2001. Le tableau 1.4.1 présente l'ensemble des 200 valeurs $y_1, \ldots, y_{200}$ (en ordre croissant) et la figure 1.4.1 en présente l'histogramme.

Ces données ne sont pas connues du statisticien. Celui-ci prélève donc un échantillon de taille *n*. Supposons que $n = 10$ et qu'il obtienne les résultats suivants:

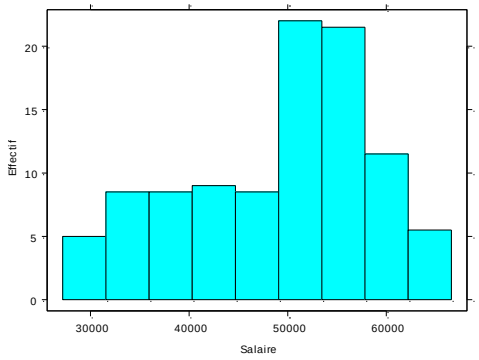
Professeur	4	48	58	63	67	71	86	103	106	197
Salaire (*y*)	62 317	52 561	63 780	57 805	56 220	62 073	48 537	53 049	58 902	44 044

La ligne « *Salaire y* » présente les valeurs échantillonnales $y_1, \ldots, y_{10}$.

Tableau 1.4.1
Salaires d'une population de 200 *professeurs d'université*

27561	34268	39146	44268	49268	51707	53293	54634	57195	60000
28390	34512	39775	44390	49512	51707	53415	54756	57317	60000
28390	35000	39775	44878	49512	51951	53415	55000	57561	60244
29268	35000	40000	45467	49512	51951	53537	55244	57683	61829
29814	35506	40366	45467	49736	51951	53537	55428	57805	62073
29878	35854	40366	45467	49736	51951	53537	55488	57805	62073
30000	35854	41098	45854	50000	52195	53659	55610	57927	62073
30854	35976	41198	45854	50000	52439	53659	55854	57927	62195
31098	36098	41198	45976	50000	52439	53780	55976	58049	62195
31237	36829	41585	46829	50122	52439	53780	55976	58049	62317
31707	36829	41951	47195	50366	52561	54005	56220	58049	62317
32439	36929	42195	47927	50610	52561	54024	56463	58049	62561
32561	36929	42683	48171	50732	52561	54146	56463	58049	62805
32660	36951	42805	48415	50732	52561	54146	56707	58171	62927
32660	36951	42927	48537	51159	52683	54146	56851	58415	63780
33537	38049	42927	48537	51159	52683	54146	56851	58659	64024
33659	38352	43415	48659	51159	52927	54268	56951	58780	64146
33902	38659	44024	48780	51220	52927	54390	56951	58780	65610
34083	38780	44044	48902	51463	53049	54390	57073	58902	65732
34146	38902	44044	49146	51585	53171	54512	57073	59024	66220

Figure 1.4.1
Distribution des salaires en 2001 – Données du tableau A.01



Paramètres et estimateurs

Dénotons par $\bar{y}_U$ la moyenne de la population et par $\bar{y}_\omega$ la moyenne de l'échantillon:

$$\bar{y}_U = \frac{1}{200}\sum_{i=1}^{200} y_i$$

Dans la population du tableau A.01, la moyenne μ est en fait égale à $\bar{y}_U$ = 49 034,90 $. C'est la

quantité que le statisticien doit estimer à partir des données de l'échantillon.

Il est intuitivement raisonnable d'estimer la moyenne $\bar{y}_U$ de la population par la moyenne $\bar{y}_U$ de l'échantillon. La valeur observée $\bar{y}$ de $\bar{y}_\omega$ est:

$$\bar{y} = \frac{1}{10}\sum_{i=1}^{10} y_i$$

$$= \frac{62317 + 52561 + 63780 + 57805 + 56220 + 62073 + 48537 + 53049 + 58902 + 44044}{10}$$

$$= 55\,928{,}80.$$

Puisque, dans ce cas-ci, nous savons que $\bar{y}_U$ = 49 034,90, nous savons que le statisticien commet une erreur dont l'amplitude est

$$|\bar{y} - \bar{y}_U| = |55\,928{,}80 - 49034{,}90| = 6\,893{,}90\ \$.$$

Le statisticien s'est donc trompé de 6 893,90 $ dans l'estimation de μ. C'est une erreur relativement importante, mais elle aurait pu être pire.

Le tableau 1.4.2 présente trois scénarios. Dans le premier, le statisticien a eu de la chance: il est tombé tout près de la moyenne de la population, son erreur n'est que de 2471,20 $, ce qui est peu lorsqu'on considère la moyenne de la population est de plus de 49 000 $. Dans les deux derniers scénarios, des statisticiens malchanceux tombent sur des échantillons très différents de la population et commettent d'importantes erreurs d'estimation: le premier sous-estime la moyenne de 19 385,90 $, le second la surestime de 14 977,30 $.

Tableau 1.4.2
Quelques échantillons possibles tirés de la population de 200 *professeurs*
La moyenne de la population est $\bar{y}_U$ = 49 034,90.

Unités choisies	*Valeurs des y*	$\bar{y}$	*Erreur absolue* $\lvert \bar{y} - \bar{y}_U \rvert$
95 ; 96 ; 97 ; 98 ; 99 ; 100 ; 101 ; 102 ; 103 ; 104	51159 ; 51159 ; 51159 ; 51220; 51463 ; 51585 ; 51707 ; 51707; 51951 ; 51951	51506,10	|51 506,10 − 49 034,90| = 2471,2
1 ; 2 ; 3 ; 4 ; 5 ; 6 ; 7 ; 8 ; 9 ; 10	27561 ; 28390 ; 28390 ; 29268 ; 29814 ; 29878 ; 30000 ; 30854 ; 31098 ; 31237	29649,00	|29 649 − 49 034,90| = 19 385,90
191 ; 192 ; 193 ; 194; 195 ; 196 ; 197; 198 ; 199 ; 200	62317 ; 62561 ; 62805 ; 62927 ; 63780 ; 64024 ; 64146 ; 65610 ; 65732 ; 66220	64012,20	|64012,2-49 034,90| = 14 977,30

Ces trois scénarios sont *possibles*, le plus optimiste comme les plus pessimistes. Aucun, cependant, n'est *probable*: On ne devrait pas compter sur une estimation quasi-parfaite, comme dans le premier cas; ni être paralysé par la crainte d'une erreur grave, comme dans les deux derniers. La confiance qu'on accorde à un échantillonnage est fondée sur l'assurance qu'avec un échantillon

suffisamment grand, la *probabilité* d'une erreur importante est relativement faible.

Une expérience théorique

Cette assurance est-elle justifiée? Puisque nous avons devant nous les données de la population entière, nous pouvons nous faire une bonne idée des probabilités d'erreur : il suffit de refaire l'expérience plusieurs fois et d'examiner les résultats. C'est ce que nous avons fait : nous avons prélevé un échantillon de taille $n = 10$ de cette même population. Nous l'avons fait 10 000 fois, et nous avons calculé la moyenne $\bar{y}$ à chaque fois. La distribution des moyennes obtenues est présentée au tableau 1.4.3. La figure 1.4.1 en donne une représentation graphique.

Une information de ce genre nous permet porter un jugement sur la qualité de l'échantillonnage. On constate, par exemple, que

1 la moyenne $\bar{y}$ est rarement inférieure à 40 000 ou supérieure à 54 000;
2 elle se situe très souvent entre 47 000 et 52 000 et donc en pratique $\bar{y}$ se situera presque sûrement dans cet intervalle.

Tableau 1.4.3
Distribution des moyennes des salaires de 10 000 *échantillons tirés de la population A*.01

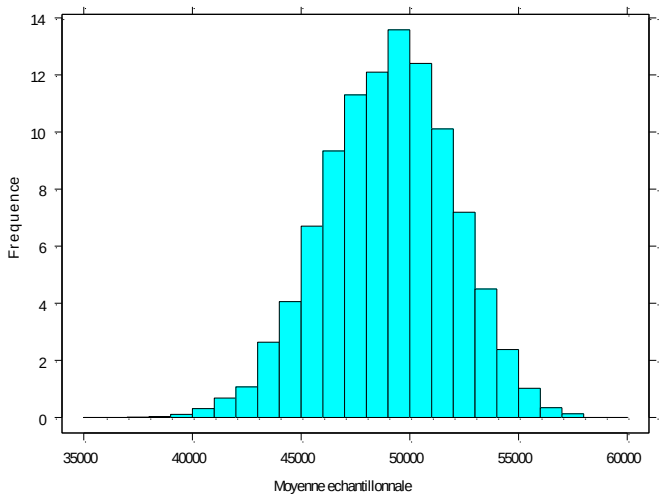
Moyenne $\bar{y}$	*Fréquence*	*Moyenne* $\bar{y}$	*Fréquence*
36000-37000	0	48000-49000	0,1210
37000-38000	0,0001	49000-50000	0,1358
38000-39000	0,0003	50000-51000	0,1240
39000-40000	0,0011	51000-52000	0,1011
40000-41000	0,0031	52000-53000	0,0719
41000-42000	0,0068	53000-54000	0,0450
42000-43000	0,0107	54000-55000	0,0238
43000-44000	0,0264	55000-56000	0,0102
44000-45000	0,0406	56000-57000	0,0034
45000-46000	0,0670	57000-58000	0,0013
46000-47000	0,0934	58000-59000	0
47000-48000	0,1130		1

Supposons que l'erreur maximale qu'on est disposé à tolérer est de 5 000 $. Ceci signifie que $\bar{y}$ doit se situer entre 49 034,90 - 5 000 = 44 034,90 ≈ 44 000 et 49 034,90 + 5 000 = 54 034,90 ≈ 54 000. Dans le tableau 1.4.3, on peut approcher la probabilité que $\bar{y}$ se situe entre 44 000 et 54 000 : 0,0406 + 0,0670 + 0,0934 + 0,1130 + 0,1210 + 0,1358 + 0,1240 + 0,1011 + 0,0719 + 0,0450 = 0,9128 (la fréquence exacte, calculée à partir des 10 000 données brutes est, elle aussi, égale à 0,9128).

Il s'agit de la probabilité d'une estimation correcte (à plus ou moins 5 000 $), et elle se doit d'être assez forte. Une probabilité de 91,28 % de se situer à moins de 5 000 $ de la vraie moyenne, est-ce assez? Il ne nous appartient pas d'en décider : c'est l'expérimentateur qui doit le faire en te-

nant compte des conditions matérielles du contexte.

Figure 1.4.1
Distribution des moyennes de 10 000 *échantillons de taille* 10 *tirés de la population* A.01



Remarque *Si l'on compare les résultats obtenus ici à des sondages réels dans lesquels on estime des moyennes de revenus, on trouve la précision ici excellente :* 91,28 *% de chance de se trouver à moins de* 5 000 *$ de la vraie moyenne, c'est bon. Comment se fait-il qu'un échantillon si petit donne de si bons résultats? Il s'agit ici d'une population particulière : ce sont tous des professeurs d'une même institution, et leurs salaires ne varient pas énormément. La dispersion des salaires est faible. Ce sont des condition idéales pour estimer la moyenne avec précision et peu d'effort (c'est-à-dire, avec n petit).* ■

La moyenne et l'écart-type des moyennes échantillonnales

Revenons sur terre : en pratique on ne connaît pas la population, et il est hors de question de tirer 10 000 échantillons pour voir comment se comporte $\bar{y}_\omega$. Pourtant, l'information que nous avons tirée de notre expérience est précisément celle dont nous avons besoin pour évaluer la précision de notre estimation: Quelles sont les valeurs possibles de $\bar{y}_\omega$? Quelles sont les chances que $\bar{y}_\omega$ s'éloigne gravement de $\bar{y}_U$? Peut-on obtenir quelque réponse en n'utilisant que les données de l'échantillon?

Certaines réponses sont possibles. Il est clair que $\bar{y}$ est une variable aléatoire : sa valeur dépend d'une expérience aléatoire, le tirage d'un échantillon. Elle a donc une certaine distribution, et cette distribution a une moyenne $E(\bar{y}_\omega) = \mu_{\bar{y}}$ et un écart-type $\sqrt{V(\bar{y}_\omega)} = \sigma_{\bar{y}}$. Qu'est-ce qu'on peut dire à propos de la distribution de $\bar{y}_\omega$, de sa moyenne, de son écart-type?

Espérance de $\bar{y}_\omega$ et estimateur sans biais

On peut interpréter ces paramètres concrètement. Considérons l'ensemble de toutes les valeurs possibles de $\bar{y}_\omega$. Alors

$$\mu_{\bar{y}} = \mathrm{E}(\bar{y}_\omega) = \textit{moyenne de tous les } \bar{y} \textit{ possibles.}$$

C'est $\mu_{\bar{y}}$ que nous avons tenté d'approcher dans la section précédente lorsqu'on a calculé la moyenne des 10 000 échantillons que nous avons tirés. Ce concept — la moyenne des valeurs possibles d'un estimateur — joue un rôle important dans la théorie de l'estimation: $\mu_{\bar{y}}$ est la valeur *attendue* de $\bar{y}$, et le théorème suivant énonce une propriété importante de cette moyenne:

$$\mathrm{E}(\bar{y}_\omega) = \bar{y}_\mathrm{U}$$

En mots : la moyenne de tous les $\bar{y}$ possibles est égale à la moyenne de la population. Ceci ne nous aide pas à calculer $\bar{y}_\mathrm{U}$ ou $\mathrm{E}(\bar{y}_\omega)$: $\bar{y}_\mathrm{U}$, étant un paramètre de la population, est inconnu. Mais le théorème est intéressant quand même, car il signifie que, bien que la valeur de $\bar{y}_\omega$ s'écarte inévitablement de $\bar{y}_\mathrm{U}$, sa tendance est de viser juste : $\bar{y}_\omega$ n'a tendance ni à surestimer ni à sous-estimer la moyenne $\bar{y}_\mathrm{U}$ de la population. Un estimateur qui jouit de cette propriété est appelé *estimateur* ***sans biais***. Voici une définition formelle de cette notion :

Définition

Un estimateur $\hat{\theta}$ d'un paramètre θ *est dit* sans biais *si son espérance est égale au paramètre estimé* : $E(\hat{\theta}) = \theta$.

Écart-type de l'estimateur

Mais le fait que $\bar{y}_\omega$ soit sans biais n'empêche pas qu'elle s'écarte de $\bar{y}_\mathrm{U}$, parfois de façon importante. Cette tendance à s'écarter de $\bar{y}_\mathrm{U}$ est mesurée par l'écart-type $\sigma_{\bar{y}}$ de $\bar{y}$:

$$\sigma_{\bar{y}} = \textit{Écart-type de tous les } \bar{y} \textit{ possibles}$$

On souhaite, bien sûr, que $\sigma_{\bar{y}}$ soit *petit*, puisque l'écart entre $\bar{y}_\omega$ et $\bar{y}_\mathrm{U}$ représente une erreur d'estimation. De quoi dépend $\sigma_{\bar{y}}$? On devine que $\sigma_{\bar{y}}$ dépend en partie de la dispersion de la *population* : lorsque la population elle-même est très dispersée, les $\bar{y}$ le sont aussi. Il existe en effet un lien entre $\sigma_{\bar{y}}$ et $S = \sqrt{\dfrac{\sum_{i=1}^{N}(y_i - \bar{y}_U)^2}{N-1}}$. La relation précise entre *ces deux quantités* est

celle-ci :

$$\sigma_{\bar{y}} = \sqrt{1-f}\ \frac{S}{\sqrt{n}}$$

où $f = n/N$ est appelé ***fraction d'échantillonnage*** et $\sqrt{1-f}$ est appelé le ***facteur de correction pour population finie***. Dans le cas présent (mais pas en pratique), on peut calculer S, puisqu'on suppose connues toutes les données de la population. On a

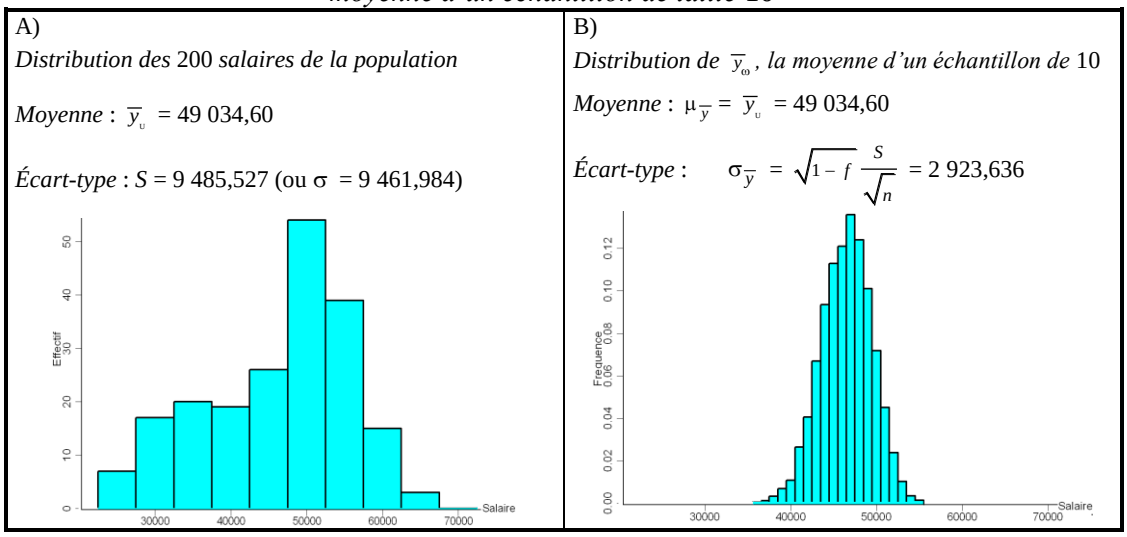
$$S = \sqrt{\frac{\sum_{i=1}^{200}(y_i - \bar{y}_U)^2}{200-1}} = 9\,485{,}527.$$

L'écart-type des $\bar{y}$ est donc

$$\sigma_{\bar{y}} = \sqrt{1-\frac{n}{N}}\ \frac{S}{\sqrt{n}} = \sqrt{1-\frac{10}{200}}\ \frac{9485{,}527}{\sqrt{10}} = 2\,923{,}636$$

La dispersion de la population est bien supérieure à celle des $\bar{y}$ (2 923,636 comparé à 9 485,527) : le revenu d'un individu est plus variable que le revenu moyen de 10 individus.

Figure 1.4.2
Distribution des salaires d'une population et distribution de la moyenne d'un échantillon de taille 10



Remarque *Il est intéressant de noter le rôle de la fraction d'échantillonnage f. Lorsque la population est grande, f est petite et le facteur de correction est négligeable. C'est le cas dans la plupart des sondages: l'échantillon est très petit comparé à la population, le facteur de correction peut être négligé, et la formule de l'écart-type devient simplement*

$$\sigma_{\bar{y}} = \frac{S}{\sqrt{n}} \approx \frac{\sigma}{\sqrt{n}},$$

où

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \bar{y}_U)^2}{N}}\,.$$ ■

La figure 1.4.2 résume les notions introduites jusqu'ici.

Formule d'intervalle de confiance

On commence à entrevoir comment ce développement mènera à la formule d'intervalle de confiance. Comme première approximation, nous pouvons supposer que la variable aléatoire $\bar{y}_\omega$ se situe à 1,96 écarts-types de sa moyenne $\mu_{\bar{y}}$ (et donc de $\bar{y}_U$) avec une probabilité d'environ 95 % :

$$P(|\,\bar{y}_\omega - \bar{y}_U\,| \le 1{,}96\sigma_{\bar{y}}) \approx 0{,}95.$$

Quelques manipulations algébriques mènent à

$$P(\bar{y}_\omega - 1{,}96\sigma_{\bar{y}} \le \bar{y}_U \le \bar{y}_\omega + 1{,}96\sigma_{\bar{y}}) \approx 0{,}95$$

On a donc la formule

$$\bar{y}_\omega - 1{,}96\sigma_{\bar{y}} \le \bar{y}_U \le \bar{y}_\omega + 1{,}96\sigma_{\bar{y}}\text{, où } \sigma_{\bar{y}} = \sqrt{1-f}\ \frac{S}{\sqrt{n}}$$

On y est presque, mais pas tout à fait, car cette formule n'est pas utilisable en pratique : Le S qui y figure (l'écart-type de la population), n'est pas connu. On doit donc estimer S à partir de l'échantillon. L'estimateur naturel est s, l'écart-type de l'échantillon,

$$s = \sqrt{\frac{\sum_{i\in\omega}(y_i - \bar{y})^2}{n-1}}$$

Ce qui donne $\sigma_{\bar{y}}$ comme estimateur de $\sigma_{\bar{y}}$:

$$\sigma_{\bar{y}} = \sqrt{1-f}\ \frac{S}{\sqrt{n}}$$

Nous obtenons enfin la formule utilisée au début du chapitre :

$$\bar{y}_\omega - 1{,}96\,\sigma_{\bar{y}} \le \bar{y}_U \le \bar{y}_\omega + 1{,}96\,\sigma_{\bar{y}}\,.$$

Normalité de $\bar{y}_\omega$

Nous avons affirmé que $\bar{y}_\omega$, comme toutes les variables dont nous avons traité jusqu'ici, se situe à moins de 1,96 écarts-types avec une probabilité voisine de 95 %. Rappelons, cependant, que ce

chiffre (95 %) suppose une loi *normale*. Si $\bar{y}_\omega$ n'est pas normale, la probabilité de 95 % n'est qu'approximative. Jusqu'à quel point la distribution de $\bar{y}_\omega$ peut-elle s'écarter d'une normale? Il existe un théorème, appelé *théorème limite central*, qui a pour remarquable conclusion que $\bar{y}$ est à peu près normale si l'échantillon est assez grand :

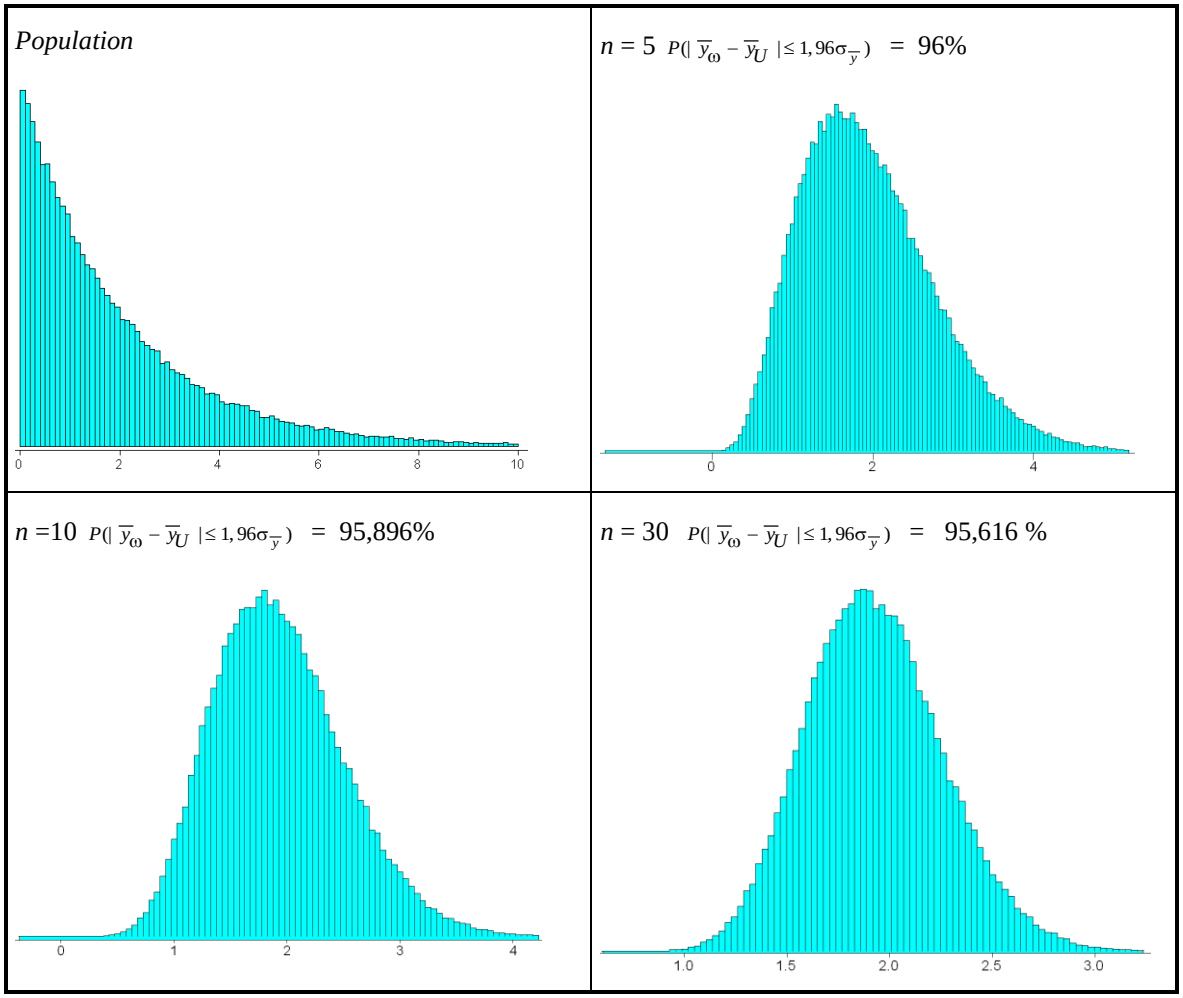
Si n est assez grand, on peut supposer que
$$\bar{y} \sim \mathcal{N}(\mu_{\bar{y}}; \sigma^2_{\bar{y}})$$

Cette conclusion ne serait pas étonnante si on supposait que la *population* était normale. Or, comme le montre la figure 1.4.2, $\bar{y}_\omega$ est presque normale bien que la population, elle, ne l'est pas. Et cela malgré le fait que $n = 10$ n'est pas « grand ».

Quand est-ce qu'on peut dire que n est grand? Il n'y a pas de réponse simple : cela dépend de la population. À la figure 1.4.2, on voit que la population a une distribution qui, sans être tout à fait normale, ne s'en écarte pas trop. C'est ce qui explique pourquoi la distribution de $\bar{y}_\omega$ est si proche d'une normale malgré un échantillon petit. Pour des populations dont l'asymétrie est plus prononcée, un échantillon plus grand est nécessaire. Dans la plupart des applications, cependant, un échantillon de taille 30 suffit à valider l'approximation normale.

La figure 1.4.3 montre ce qui se passe avec une population asymétrique. La distribution de $\bar{y}_\omega$ est asymétrique lorsque $n = 5$; un peu moins asymétrique lorsque $n = 10$; et presque normale lorsque $n = 30$. La figure présente aussi une donnée importante pour l'application actuelle. L'intervalle de confiance est basé sur la supposition que $P(|\bar{y}_\omega - \bar{y}_U| \leq 1{,}96\sigma_{\bar{y}}) \approx 0{,}95$, ce qui est vérifié si $\bar{y}_\omega$ est normale. Mais cette égalité pourrait être vérifiée même si $\bar{y}_\omega$ n'est *pas* normale. On constate que la probabilité est proche de 95 %, même avec $n = 5$.

Figure 1.4.3
Distribution salaires d'une population et distribution de la moyenne d'un échantillon de taille 10



Estimation de l'écart-type

Une dernière mise au point. Partant de l'hypothèse que $\overline{y}_\omega$ est normale, nous avons conclu que

$$P(|\overline{y}_\omega - \overline{y}_U| \le 1{,}96\sigma_{\overline{y}}) \approx 0{,}95,$$

et de là que

$$\overline{y}_\omega - 1{,}96\sigma_{\overline{y}} \le \overline{y}_U \le \overline{y}_\omega + 1{,}96\sigma_{\overline{y}}$$

est un intervalle de confiance à 95 %. Le problème est que ce n'est pas exactement la formule que nous avons employée. La formule que nous avons employée est plutôt

$$\overline{y}_\omega - 1{,}96\hat{\sigma}_{\overline{y}} \le \overline{y}_U \le \overline{y}_\omega + 1{,}96\hat{\sigma}_{\overline{y}}.$$

La différence n'est pas anodine. Pour que cette formule soit valide il faudrait que

$$P(|\,\bar{y}_\omega - \bar{y}_U\,| \le 1{,}96\hat{\sigma}_{\bar{y}}) \approx 0{,}95,$$

ce qui n'est vrai que lorsque l'échantillon est vraiment grand. C'est pourquoi un échantillon de taille 10 est rarement satisfaisant en pratique, même si ce nombre est suffisant pour assurer la normalité de $\bar{y}_\omega$.

Récapitulation *La formule d'un intervalle de confiance est approximative. Elle repose sur un théorème dont la conclusion est bien celle énoncée, mais dont les hypothèses ne sont pas nécessairement vérifiées dans les faits.*

1. *Le théorème affirme ceci : si l'échantillon est assez grand* $\bar{y}_\omega \sim \mathcal{N}(\mu\,;\,\sigma^2_{\bar{y}})$, où $\sigma^2_{\bar{y}} = (1-f)\dfrac{S^2}{n}$.
2. *Par conséquent l'intervalle* $[\,\bar{y}_\omega - 1{,}96\,\sigma_{\bar{y}}\;;\;\bar{y}_\omega + 1{,}96\,\sigma_{\bar{y}}\,]$ *a une probabilité d'environ* 95 *% de contenir* μ
3. *La taille d'échantillon nécessaire pour que cette approximation soit bonne dépend de la forme de la population. À moins que celle-ci soit très asymétrique, un échantillon de taille* 20 *suffirait pour que l'approximation normale soit adéquate.*
4. *Cependant, en pratique on doit remplacer* $\sigma_{\bar{y}}$ *par* $\hat{\sigma}_{\bar{y}} = \sqrt{1-f}\,\dfrac{s}{\sqrt{n}}$ *où s est l'analogue échantillonnal de S.*
5. *En remplaçant* $\hat{\sigma}_{\bar{y}}$ *par* $\sigma_{\bar{y}}$, *on modifie la probabilité que l'intervalle* $[\,\bar{y}_\omega - \bar{y}_\omega\;1{,}96\,\hat{\sigma}_{\bar{y}}\;;\;\bar{y}_\omega + 1{,}96\,\hat{\sigma}_{\bar{y}}\,]$ *recouvre* $\bar{y}_U$, *qui alors n'est plus de 95% mais probablement un peu inférieure à 95 %. Par conséquent, nous devrions majorer la limite de* 20 *proposée au numéro* 3.
6. *Une dernière question: quand est-ce que n est assez grand? Nous n'avons pas de réponse claire à cette question. Nous avons, cependant, plusieurs études empiriques — c'est-à-dire, des simulations comme celle qui a donné la figure* 1.4.2 *— dont les résultats sont généralement encourageants. À moins que l'échantillon ne soit très petit, ou que la population ne soit très éloignée d'une population normale, les niveaux de confiances réels sont assez proches de ceux qu'ils devraient être théoriquement.*
7. *Afin d'avoir une règle commune, nous nous entendons pour dire qu'un échantillon de taille* 30 *est adéquat* ■

1.5 Tests d'hypothèses

Une branche importante de la statistique concerne ce qu'on appelle des ***tests d'hypothèses***. Une hypothèse en statistique est une affirmation à propos d'un paramètre inconnu telle la moyenne μ d'une population. Un *test* d'hypothèse est une procédure qui doit mener à la décision de rejeter ou non l'hypothèse. Dans cette section, nous considérerons une hypothèse particulière, appelée ***hypothèse nulle*** et dénotée par H_o, de la forme

$$H_o:\ \bar{y}_U = \mu_o$$

où μ_o est une valeur particulière de $\bar{y}_U$.

La procédure est simple. On détermine un intervalle de confiance

$$\bar{y}_\omega - 2\,\hat{\sigma}_{\bar{y}} \le \bar{y}_U \le \bar{y}_\omega + 2\,\hat{\sigma}_{\bar{y}}$$

et on *rejette* H_o si et seulement si l'intervalle *ne recouvre pas* μ_o.

***Exemple* 1.5.1** *Test d'hypothèse*

Considérons encore la population présentée au tableau A.01, les professeurs d'une certaine université. Supposons que la moyenne nationale pour les professeurs d'université est de 45 000 $, et supposons qu'un veuille tester l'hypothèse que la moyenne dans cette université est égale à la moyenne

nationale. C'est-à-dire, on veut tester l'hypothèse H_o : $\bar{y}_U$ = 45 000. L'intervalle de confiance étant (38 150 ; 41 440), nous pouvons rejeter l'hypothèse que $\bar{y}_U$ = 45 000$. On peut affirmer que la moyenne dans cette université est *inférieure* à la moyenne nationale. ■

Quelles sont les propriétés de ce test? Il en a une, essentiellement, et c'est celle-ci:

> *Si* H_o *est vraie, la probabilité qu'on la rejette quand même est à peu près égale à* 5 %

Les réserves formulées dans la section précédente concernant une interprétation trop stricte du niveau de confiance s'appliquent également ici: théoriquement 0,05 est la probabilité de rejet lorsque la population est normale et on emploie $\sigma_{\bar{y}}$ plutôt que $\hat{\sigma}_{\bar{y}}$ pour construire l'intervalle de confiance. Mais nous savons que la population n'est pas toujours normale et c'est toujours $\hat{\sigma}_{\bar{y}}$ qu'on emploie, jamais $\sigma_{\bar{y}}$. Néanmoins, le raisonnement qui permet une décision demeure vrai *grosso modo*: on rejette H_0 lorsque la valeur observée de $\bar{y}_\omega$ est trop peu probable sous cette hypothèse. La probabilité des valeurs de $\bar{y}_\omega$ que nous déclarons peu probables n'est peut-être pas 0,05 — notre définition de peu probable — mais elle est presque certainement assez petite pour valider l'argument.

Remarque Habituellement, lorsqu'on détermine un intervalle de confiance dans le seul but de tester une hypothèse à propos de $\bar{y}_U$*, l'écart-type* $\hat{\sigma}_{\bar{y}}$ *est calculé sans le facteur de correction* $\sqrt{1-f}$ *. La raison est qu'un test porte rarement sur la moyenne ou le total de la population concrète et finie qu'on échantillonne; il porte sur le paramètre d'une population infinie conceptuelle de laquelle la population finie est elle-même un échantillon aléatoire. Supposons, par exemple, que la population est l'ensemble que tous les chèques reçus par une compagnie pendant l'année et* μ *est le délai moyen entre la réception d'un chèque et son dépôt en banque. On considère qu'un délai de 2 jours est acceptable, et on se propose donc de tester l'hypothèse que* $\bar{y}_U = 2$*. Mais quel est le sens de* $\bar{y}_U$ *dans ce cas?* $\bar{y}_U$ *est-elle vraiment la moyenne des chèques de l'année? En fait non: on s'intéresse moins aux chèques d'une année donnée qu'au système de traitement des chèques, lequel peut être caractérisé par une certaine moyenne* μ *qui n'est pas la moyenne d'un nombre fini de données mais plutôt la moyenne de l'ensemble de tous les chèques imaginables.* ■

1.7 Développement mathématique

Soit $\mathcal{P} = \{y_1; y_2; \ldots ; y_N\}$ une population. Soit $\bar{y}_U = \dfrac{\sum_{i=1}^{N} y_i}{N}$ la moyenne de la population et $S^2 = \sqrt{\dfrac{\sum_{i=1}^{N}(y_i - \bar{y}_U)^2}{N-1}}$ sa variance. On tire un échantillon aléatoire simple de n unités. L'espace échantillon Ω est constitué de tous les choix possibles de n unités parmi les N unités de la population. Il est donc de cardinalité $\binom{N}{n}$. Un échantillon aléatoire simple est défini comme un

échantillon tiré de telle sorte que pour tout $\omega \in \Omega$, $p(\omega) = \dfrac{1}{\binom{N}{n}}$. Soit $\bar{y}_\omega = \dfrac{1}{n}\sum_{i\in\omega} y_i$ la moyenne de l'échantillon ω.

Quelques résultats préalables

- Chaque unité *i* est contenue dans $\binom{N-1}{n-1}$ échantillons
- Chaque paire d'unités distinctes *i* et *j* est contenue dans $\binom{N-2}{n-2}$ échantillons

Théorème L'estimateur $\bar{y}_\omega$ est sans biais pour $\bar{y}_U$

$$\mathrm{E}(\bar{y}_\omega) = \bar{y}_U$$

Démonstration

$$\mathrm{E}(\bar{y}_\omega) = \sum_{\omega\in\Omega} \bar{y}_\omega p(\omega) = \sum_{\omega\in\Omega} \bar{y}_\omega \frac{1}{\binom{N}{n}} = \frac{1}{\binom{N}{n}} \sum_{\omega\in\Omega}\left(\frac{1}{n}\sum_{i\in\omega} y_i\right) = \frac{1}{\binom{N}{n}n}\sum_{i=1}^{N}\sum_{\omega\ni i} y_i \text{ . Mais}$$

$$\sum_{\omega\ni i} y_i = y_i \times (\text{nombre d'échantillons } \omega \text{ qui contiennent l'unité } i) = y_i\binom{N-1}{n-1}.$$

Donc $\mathrm{E}(\bar{y}_\omega) = \dfrac{1}{\binom{N}{n}n}\sum_{i=1}^{N} y_i\binom{N-1}{n-1} = \dfrac{\binom{N-1}{n-1}1}{\binom{N}{n}n}\sum_{i=1}^{N} y_i = \dfrac{1}{N}\sum_{i=1}^{N} y_i = \bar{y}_U$ ■

Théorème La variance de $\bar{y}_\omega$ est

$$\mathrm{V}(\bar{y}_\omega) = \left(1-\frac{n}{N}\right)\frac{S^2}{n}$$

où $S^2 = \dfrac{\sum_{i=1}^{N}(y_i - \bar{y}_U)^2}{N-1}$

Démonstration

1.8 Résumé

1 Un *estimateur sans biais* pour un paramètre est un estimateur dont l'espérance est égale au paramètre.

2 La moyenne échantillonnale $\bar{y}$ est un estimateur sans biais de la moyenne $\bar{y}_U$ d'une population : $\mathrm{E}(\bar{y}_\omega) = \bar{y}_U$.

3 L'écart-type de $\bar{y}$ est $\sigma_{\bar{y}} = \sqrt{1-f}\ \dfrac{S}{\sqrt{n}}$ où $f = n/N$ et S est l'écart-type corrigé de la population, défini par $S = \sqrt{\dfrac{\sum_{i=1}^{N}(y_i - \bar{y}_U)^2}{N-1}}$.

4 Si *n* est assez grand, le *théorème central limite* permet de conclure que $\bar{y}_\omega \sim \mathcal{N}(\bar{y}_U\ ;\ \sigma^2_{\bar{y}})$.

5 On peut donc conclure que $P(|\bar{y}_\omega - \bar{y}_U| \le 1{,}96\sigma_{\bar{y}}) \approx 0{,}95$ et donc que

$$\bar{y}_\omega - 1{,}96\sigma_{\bar{y}} \le \bar{y}_U \le \bar{y}_\omega + 1{,}96\sigma_{\bar{y}}$$

serait un intervalle de confiance à 95 % si $\sigma_{\bar{y}}$ pouvait être calculé.

6 Pour estimer $\sigma_{\bar{y}}$, on remplace *S* dans la formule par l'équivalent échantillonnal $s = \sqrt{\frac{\sum_{i\in\omega}(y_i - \bar{y}_\omega)^2}{n-1}}$ et on obtient alors l'estimateur $\hat{\sigma}_{\bar{y}} = \sqrt{1-f}\ \frac{s}{\sqrt{n}}$.

7 La formule d'un intervalle de confiance à à peu près 95 % devient alors $\bar{y}_\omega - 1{,}96\,\hat{\sigma}_{\bar{y}} \leq \bar{y}_U \leq \bar{y}_\omega + 1{,}96\,\hat{\sigma}_{\bar{y}}$.

8 Un intervalle de confiance permet de tester une hypothèse. Il suffit d'observer si l'intervalle de confiance recouvre ou non la valeur postulée μ_o : on rejette l'hypothèse que $\bar{y}_U = \mu_o$ si l'intervalle ne recouvre pas μ_o.

1.7 Exercices

1.1 D'une population de 850 transactions, on prélève un échantillon de taille 20. Les montants des transactions obtenus dans l'échantillon sont:

23,90	110,52	79,95	146,65	19,51	26,62	27,67	65,79	12,38	12,44
72,14	57,37	62,93	135,88	15,22	36,35	31,98	7,39	33,05	46,96

a) Estimer la moyenne de la population et l'écart-type de l'estimateur.

b) Déterminer un intervalle de confiance à 95 % pour la moyenne de la population.

1.2 Mêmes données qu'au numéro 1. Supposant que la valeur aux livres totale de toutes les transactions est 32 831,82$. Peut-on avec 95 % de confiance affirmer que ce montant est erroné?

1.3 Le tableau A.02 présente des données sur un échantillon de 50 professeurs tiré de la population de 200 professeurs présentée au tableau A.01. Déterminer un intervalle de confiance pour le salaire moyen à l'entrée.

1.4 Le tableau A.04 présente des données sur un échantillon de 50 paroisses tiré de la population de 210 paroisses québécoises présentée au tableau A.03. Déterminer un intervalle de confiance pour le nombre moyen de mariages par paroisse.

1.5 [*Tableau* A.05] Déterminez un intervalle de confiance à 95 % pour la moyenne μ de la variable « pouls » (nombre de battements à la minute) [Notez que la population est infinie, et donc qu'il n'y a pas de facteur de correction]. Ensuite déterminez deux intervalles de confiance séparés, l'un pour les hommes, l'autre pour les femmes. Pourquoi les deux derniers intervalles sont-ils plus larges que celui pour la moyenne de la population entière? Pourquoi l'intervalle pour les femmes est-il plus large que celui pour les hommes?

1.6 [*Tableau* A.04] Déterminez un intervalle de confiance pour le nombre moyen de mariages par paroisse (il y a 210 paroisses dans la population). La vraie moyenne (tableau A.03) est 2,3524. Votre intervalle la contient-il?

1.7 [*Tableau* A.09] A1 et B1 sont des scores dans des tests de compréhension donnés, respectivement, avant et après une période d'apprentissage. Donc la variable *y* = B1 - A1 est une mesure de l'effet de l'apprentissage. Déterminez un intervalle de confiance pour la moyenne de la variable *y* et dites si on peut affirmer avec confiance que l'apprentissage a eu un effet. Vous considérerez que la population est infinie.

1.8 Un vérificateur du ministère du revenu tire un échantillon de 10 comptes de dépenses de votre compagnie afin d'estimer la proportion des frais éligible à une exemption de la TVQ. Vous avez de graves réserves concernant la manière dont le vérificateur a procédé pour tirer l'échantillon (il vous semble qu'il a délibérément choisi les plus gros comptes). Vous prenez note des valeurs totales des 10 comptes tirés. Les voici

6 067$	4 643$	9 410$	6 621$	6 573$
5 928$	19 478$	4 634$	2 265$	3 109$

Sachant que le montant total des 780 comptes de la population est de 3 197 321 $, vos doutes concernant le mode de tirage sont-ils confirmés?.

1.9 On prélève un échantillon de 40 enfants afin d'estimer le nombre moyen m de caries dentaires parmi les 1 300 élèves d'une certaine école. Les résultats de l'examen dentaire sont présentés dans le tableau suivant :

Nombre de caries	0	1	2	3	4	Total
Effectifs (nombre d'enfants)	28	7	2	2	1	40

a) Déterminez un intervalle de confiance à 95 % pour m

b) Le nombre moyen de caries dentaires dans la même école était de 0,9 il y a cinq ans. Peut-on conclure avec confiance que le nombre de caries a diminué?

1.10 Considérons une population de taille *N*, de moyenne m et d'écart-type corrigé *S*. Soit $\overline{y}_1$ et s_1 la moyenne et l'écart-type corrigé d'un échantillon de taille 50 tiré de cette population; $\overline{y}_2$ et s_2 la moyenne et l'écart-type corrigé d'un échantillon de taille 100 tiré de cette même population. Inscrivez « < », « > », « = », « ≈ » dans les espaces blancs selon que le membre de gauche est inférieur, supérieur, égal, ou à peu près égal au membre de gauche. Marquez « ? » si aucun lien ne peut être établi entre les deux quantités. Vous supposerez que $\overline{y}_1$ et $\overline{y}_2$ sont toutes deux de loi normale.

a) $E(\overline{y}_1)$ ____ $E(\overline{y}_2)$

b) $\sigma_{\overline{y}_1}$ ____ $\sigma_{\overline{y}_2}$

c) $E(\overline{y}_1)$ ____ $\overline{y}_U$

d) $\mu_{\overline{y}_2}$ ____ $\overline{y}_U$

e) s_1 ____ S

f) $P(\overline{y}_1 > \overline{y}_U + 1{,}96)$ ____ $P(\overline{y}_2 > \overline{y}_U + 1{,}96)$

g) $P(\overline{y}_1 > \overline{y}_U + 1{,}96\,\sigma_{\overline{y}_1})$ ____ $P(\overline{y}_2 > \overline{y}_U + 1{,}96S)$

h) $P(\overline{y}_1 > \overline{y}_U + 1{,}96\,\sigma_{\overline{y}_1})$ ____ $P(\overline{y}_2 > \overline{y}_U + 1{,}96\,\sigma_{\overline{y}_1})$

i) $\sigma_{\overline{y}_1}$ ____ S

j) $P(\overline{y}_1 > \overline{y}_U + 1{,}96\,\sigma_{\overline{y}_1})$ ____ $P(\overline{y}_2 > \overline{y}_U + 1{,}96\,\sigma_{\overline{y}_2})$

k) s_1 ____ s_2

l) $P(|\overline{y}_1 - \overline{y}_U| > 1{,}96\,\sigma_{\overline{y}_1})$ ____ $P(|\overline{y}_2 - \overline{y}_U| > 1{,}96\,\sigma_{\overline{y}_2})$

m) $P(\overline{y}_1 > \overline{y}_U + 1{,}96S)$ ____ $P(\overline{y}_2 > \overline{y}_U + 1{,}96S)$

n) $P(|\overline{y}_1 - \overline{y}_U| > 1{,}96\,\sigma_{\overline{y}_1})$ ____ 0,05

o) $P(\overline{y}_1 > \overline{y}_U \mu + 1{,}96\,\sigma_{\overline{y}_1})$ ____ 0,05

p) $P(\overline{y}_1 > \overline{y}_U - 1{,}96\,\sigma_{\overline{y}_1})$ ____ $P(\overline{y}_2 > \overline{y}_U - 1{,}96\,\sigma_{\overline{y}_2})$

q) $P(|\overline{y}_1 - \overline{y}_U| < 1{,}96)$ ____ $P(|\overline{y}_2 - \overline{y}_U| < 1{,}96)$

r) $P(|\overline{y}_1 - \overline{y}_U| > 1{,}96\,\sigma_{\overline{y}_2})$ ____ 0,05

s) $\hat{\sigma}_{\overline{y}_1}$ ____ $\sigma_{\overline{y}_1}$

t) $\hat{\sigma}_{\overline{y}_1}$ ____ s_1

u) $\hat{\sigma}_{\overline{y}_1}$ ____ $\hat{\sigma}_{\overline{y}_2}$

v) $\hat{\sigma}_{\overline{y}_1}$ ____ S

1.11 Nous avons tiré un grand nombre d'échantillons de taille 20 de la population de professeurs décrite au tableau A.01. Pour chaque échantillon, nous avons calculé $\overline{y}$, la moyenne des salaires en 2001.

Voici la distribution de la variable $\overline{y}$ (chaque intervalle contient la limite supérieure)

$\bar{y}$	*Fréquence*	$\bar{y}$	*Fréquence*	$\bar{y}$	*Fréquence*
39000-39500	0,00000	45000-45500	0,01715	51000-51500	0,05666
39500-40000	0,00001	45500-46000	0,02595	51500-52000	0,04237
40000-40500	0,00001	46000-46500	0,03772	52000-52500	0,02845
40500-41000	0,00003	46500-47000	0,05175	52500-53000	0,01745
41000-41500	0,00011	47000-47500	0,06512	53000-53500	0,01075
41500-42000	0,00021	47500-48000	0,07855	53500-54000	0,00543
42000-42500	0,00046	48000-48500	0,08934	54000-54500	0,00236
42500-43000	0,00083	48500-49000	0,09688	54500-55000	0,00106
43000-43500	0,00169	49000-49500	0,09694	55000-55500	0,00044
43500-44000	0,00361	49500-50000	0,09359	55500-56000	0,00019
44000-44500	0,00672	50000-50500	0,08486	56000-56500	0,00008
44500-45000	0,01150	50500-51000	0,07172	56500-57000	0,00001

La moyenne de la population est $\bar{y}_U$ = 49 034,9 et l'écart-type corrigé est S = 9 485,527.

a) Faites un histogramme de la distribution de $\bar{y}$. La distribution de $\bar{y}$ vous semble-t-elle assez proche d'une normale?

b) Utiliser le tableau ci-dessus pour estimer les probabilités suivantes (vous devrez faire quelques approximations, car le tableau n'est pas assez détaillé):

i) $P(48500 < \bar{y} \leq 49500)$; ii) $P(48000 < \bar{y} \leq 50000)$; iii) $P(|\bar{y} - \mu| \leq 3000)$

c) Déterminer $\sigma_{\bar{y}}$

d) Utiliser le tableau ci-dessus pour estimer la probabilité que l'intervalle déterminé par la formule $\bar{y} - 1{,}96\,\sigma_{\bar{y}} \leq \bar{y}_U \leq \bar{y} + 1{,}96\,\sigma_{\bar{y}}$ contienne m (notez que cette formule utilise $\sigma_{\bar{y}}$ alors que, en pratique, c'est $\hat{\sigma}_{\bar{y}}$ qu'on utilise puisque $\sigma_{\bar{y}}$ est inconnue). Théoriquement, à quoi cette probabilité devrait-elle être égale?

e) Déterminer chacune des probabilités demandées en b) en supposant que $\bar{y}$ suit une loi normale. Comparez vos résultats avec ceux obtenus en b). (Si la différence est importante, c'est que l'hypothèse de normalité n'est pas exacte).

f) La formule d'intervalle de confiance est basée sur l'hypothèse que la variable $Z = \dfrac{\bar{y} - \bar{y}_U}{\hat{\sigma}_{\bar{y}}} \sim \mathcal{N}(0\,;\,1)$. Pour mettre cette hypothèse à l'épreuve, nous avons calculé la valeur de Z pour les 10 000 échantillons que nous avons tirés. Voici la distribution des Z :

Z	*Fréquence*	*Z*	*Fréquence*	*Z*	*Fréquence*
- 6,0 - -2,0	0,01730	- 0,7 - -0,6	0,03550	0,7 - 0,8	0,02810
- 2,0 - -1,9	0,00340	- 0,6 - -0,5	0,03290	0,8 - 0,9	0,02780
- 1,9 - -1,8	0,00560	- 0,5 - -0,4	0,04130	0,9 - 1,0	0,02210
- 1,8 - -1,7	0,00710	- 0,4 - -0,3	0,03840	1,0 - 1,1	0,02170
- 1,7 - -1,6	0,00940	- 0,3 - -0,2	0,04280	1,1 - 1,2	0,01960
- 1,6 - -1,5	0,01000	- 0,2 - -0,1	0,04170	1,2 - 1,3	0,01680
- 1,5 - -1,4	0,01000	- 0,1 - 0,0	0,04160	1,3 - 1,4	0,01320
- 1,4 - -1,3	0,01280	0,0 - 0,1	0,04350	1,4 - 1,5	0,01070
- 1,3 - -1,2	0,01740	0,1 - 0,2	0,04250	1,5 - 1,6	0,01140
- 1,2 - -1,1	0,02010	0,2 - 0,3	0,03960	1,6 - 1,7	0,01290
- 1,1 - -1,0	0,02280	0,3 - 0,4	0,03680	1,7 - 1,8	0,00830
- 1,0 - -0,9	0,02460	0,4 - 0,5	0,03640	1,8 - 1,9	0,00760
- 0,9 - -0,8	0,02930	0,5 - 0,6	0,03340	1,9 - 2,0	0,00600
- 0,8 - -0,7	0,03200	0,6 - 0,7	0,02900	2,0 - 6,0	0,03660

L'hypothèse de normalité semble-t-elle raisonnable? Faites-vous en une idée à l'aide d'un histogramme et de quelques comparaisons de probabilités (comparaisons entre les probabilités fournies par le tableau ci-dessus et celles qui découleraient de l'hypothèse de normalité.)

1.12 Considérez la population composée de 8 unités dont les valeurs sont:

3	6	24	27	30	36	51	57

Supposez qu'on tire un échantillon de taille 3 de cette population. L'ensemble des échantillons possibles est présenté dans le tableau 1.8.1. L'intervalle de confiance a été calculé à l'aide de la formule $\bar{y} - 1{,}96\,\hat{\sigma}_{\bar{y}} \leq \mu \leq \bar{y} + 1{,}96\,\hat{\sigma}_{\bar{y}}$; la variable « incl » prend la valeur 1 si l'intervalle de confiance recouvre la moyenne de la population, et la valeur 0 sinon.

a) Calculez μ et *S* pour la population.

b) Vérifiez numériquement que $\mu_{\bar{y}} = \bar{y}_U$ et que $\sigma_{\bar{y}} = \sqrt{1-f}\ \frac{S}{\sqrt{n}}$.

c) Est-ce que *s* est un estimateur sans biais de *S*?

d) Déterminez la probabilité de commettre une erreur (*i*) de plus de 2 unités; (*ii*) de plus de 5 unités; (*iii*) de plus de 25 % dans l'estimation de la moyenne.

e) Quelle est la probabilité de se tromper de plus de 20 % dans l'estimation de l'écart-type *S* de la population?

f) Quel est le niveau de confiance de l'intervalle de confiance donné par la formule $\bar{y} - 2\,\hat{\sigma}_{\bar{y}} \leq \bar{y}_U \leq \bar{y} + 2\,\hat{\sigma}_{\bar{y}}$?

Dans ce qui suit, il faudra utiliser l'ordinateur

g) Quel est le niveau de confiance de l'intervalle de confiance donné par la formule $\bar{y} - 3\,\hat{\sigma}_{\bar{y}} \leq \bar{y}_U \leq \bar{y} + 3\,\hat{\sigma}_{\bar{y}}$?

h) Quel est le niveau de confiance de l'intervalle de confiance donné par la formule $\bar{y} - 2\,\sigma_{\bar{y}} \leq \bar{y}_U \leq \bar{y} + 2\,\sigma_{\bar{y}}$?

1.13 [À faire à l'aide d'Excel] Considérez une population dont la distribution des pointures des chaussures (*y*) est la suivante:

y	*f*	*y*	*f*
5	0,02	10,5	0,04
5,5	0,03	11	0,04
6	0,04	11,5	0,03
6,5	0,06	12	0,03
7	0,1	12,5	0,02
7,5	0,12	13	0,02
8	0,13	13,5	0,01
8,5	0,1	14	0,01
9	0,07	14,5	0,01
9,5	0,05	15	0,01
10	0,05	15,5	0,01

Prélevez plusieurs centaines d'échantillons de taille 5 et tâchez de répondre aux questions suivantes du mieux que possible, par simulation. La population est considérée infinie, ce qui veut dire que s et *S* sont confondus, et que 1-*f* = 1.

a) Dans quelle mesure vos simulations confirment-elles que $\bar{y}$ est un estimateur sans biais de $\bar{y}_U$?

b) Calculez la variance des $\bar{y}$. Quelle devrait, théoriquement être cette variance? Énoncez le théorème qui justifie votre réponse.

c) Lorsque la taille de l'échantillon est grande, la distribution de la variable $\bar{y}$ est à peu près symétrique et d'allure assez proche d'une distribution normale. Est-ce le cas lorsque $n = 5$? (faites un graphique).

d) Si la distribution des $\bar{y}$ était réellement normale, la probabilité que $\bar{y}$ s'éloigne de plus de 1,96 écarts-types de la moyenne est de 5%. Quelle est cette probabilité, en réalité? (Il s'agit d'estimer cette probabilité - plus vous prélèverez d'échantillons plus votre estimation sera juste).

e) L'intervalle de confiance calculé à partir de la formule $\bar{y} \pm 1{,}96s/\sqrt{n}$ est appelé « intervalle de confiance à 95% » à cause d'une probabilité de recouvrement qui est présumée être de 95%. Quelle est, en fait, la probabilité de recouvrement (c'est à dire, la probabilité que l'intervalle de confiance recouvre la vraie moyenne)?

f) Est-ce que vos simulations semblent indiquer que l'estimateur *s* est sans biais pour s?

g) La théorie statistique nous dit que, dans certaines conditions, l'intervalle de confiance calculé par la formule $0{,}35896s^2 < s^2 < 9{,}2576s^2$ est un intervalle de confiance à 95%. Estimez le niveau réel de cet intervalle de confiance.

Tableau 1.8.1

Moyennes, variances, variances estimés des $\bar{y}$, et intervalles de confiance pour les 56 échantillons possibles de la population de l'exercice 1.12

$\bar{y}$	s^2	$\hat{\sigma}^2_{\bar{y}}$	*Intervalle de confiance*	*Incl.*	$\bar{y}$	s^2	$\hat{\sigma}^2_{\bar{y}}$	*Intervalle de confiance*	*Incl.*
11	129	26,88	(0,632- 21,368)	0	28	507	105,625	(7,445 -48,555)	1
12	171	35,63	(0,063- 23,937)	0	30	657	136,875	(6,601 -53,399)	1
13	219	45,63	(-0,509- 26,509)	0	24	252	52,500	(9,509 -38,491)	1
15	333	69,38	(-1,658- 31,658)	1	29	507	105,625	(8,445 -49,555)	1
20	723	150,60	(-4,546- 44,546)	1	31	651	135,625	(7,708 -54,292)	1
22	921	191,9	(-5,704- 49,704)	1	31	525	109,375	(10,083 -51,917)	1
18	171	35,63	(6,063- 29,937)	1	33	657	136,875	(9,601 -56,399)	1
19	201	41,88	(6,058- 31,942)	1	38	777	161,875	(12,554 -63,446)	1
21	279	58,13	(5,752- 36,248)	1	27	9	1,875	(24,261 -29,739)	1
26	579	120,60	(4,034- 47,966)	1	29	39	8,125	(23,299 -34,701)	1
28	741	154,40	(3,150- 52,850)	1	34	219	45,625	(20,491 -47,509)	1
20	219	45,63	(6,491- 33,509)	1	36	333	69,375	(19,342 -52,658)	1
22	291	60,63	(6,428- 37,572)	1	30	36	7,500	(24,523 -35,477)	1
27	576	120,00	(5,091- 48,909)	1	35	201	41,875	(22,058 -47,942)	1
29	732	152,50	(4,302- 53,698)	1	37	309	64,375	(20,953 -53,047)	1
23	309	64,38	(6,953- 39,047)	1	37	183	38,125	(24,651 -49,349)	1
28	579	120,60	(6,034- 49,966)	1	39	279	58,125	(23,752 -54,248)	1
30	729	151,90	(5,352- 54,648)	1	44	309	64,375	(27,953 -60,047)	1
30	603	125,60	(7,583- 52,417)	1	31	21	4,375	(26,817 -35,183)	1
32	741	154,40	(7,150- 56,850)	1	36	171	35,625	(24,063 -47,937)	1
37	876	182,50	(9,981- 64,019)	1	38	273	56,875	(22,917 -53,083)	1
19	129	26,88	(8,632- 29,368)	1	38	147	30,625	(26,932 -49,068)	1
20	156	32,50	(8,598- 31,402)	1	40	237	49,375	(25,947 -54,053)	1
22	228	47,50	(8,216- 35,784)	1	45	252	52,500	(30,509 -59,491)	0
27	513	106,90	(6,324- 47,676)	1	39	117	24,375	(29,126 -48,874)	1
29	669	139,40	(5,389- 52,611)	1	41	201	41,875	(28,058 -53,942)	1
21	171	35,63	(9,063- 32,937)	1	46	201	41,875	(33,058 -58,942)	0
23	237	49,38	(8,947- 37,053)	1	48	117	24,375	(38,126 -57,874)	0

1.14 [*À faire avec l'aide d'Excel*] Imaginez deux populations — deux distributions comme celles du numéro précédent — l'une symétrique, l'autre pas; et deux tailles d'échantillon, l'une petite et l'autre grande. Et tâchez de faire des simulations qui permettent de tirer les conclusions suivantes.

a) La distribution de la moyenne échantillonnale $\bar{y}$ s'approche plus d'une normale lorsque la population est déjà relativement symétrique.

b) La distribution de la moyenne échantillonnale $\bar{y}$ s'approche plus d'une normale lorsque l'échantillon est grand.

Vous allez devoir prendre des cas extrêmes pour que vos résultats soient clairs: le petit échantillon doit être très petit et la population asymétrique doit être très asymétrique. Vous devriez juger de la normalité ou non normalité de la distribution des $\bar{y}$ de deux façons: d'abord, graphiquement, ensuite par des comparaisons avec les probabilités suivantes

La probabilité qu'une normale s'écarte de la moyenne de plus de…	0,5s	s	1,5s	2s
… de la moyenne est égale à …	0,62	0,32	0,13	,05

Développements mathématiques

Échantillon aléatoire simple

Il y a plusieurs façons de tirer un échantillon d'une population. Dans ce chapitre, quand on parlera d'échantillon, il s'agira d'un *tirage aléatoire simple*, c'est-à-dire, un échantillon tiré de telle façon que tout ensemble de n unités a même probabilité de constituer l'échantillon. Soit Ω est l'ensemble des $\binom{N}{n}$ sous-ensembles ω de $\mathcal{P}$ de taille n.

Définissons les paramètres suivants :

Moyenne de la population : $\mu = \frac{1}{N}\sum_{i=1}^{N} v_i = \bar{v}$

Variance de la population : $S^2 = \frac{\sum_{i=1}^{N}(v_i - \bar{v})^2}{N-1}$

Un *échantillon aléatoire simple* (*éas*) est un échantillon diré de telle sorte que, pour tout $\omega \in \Omega$, la probabilité que ω constitue l'échantillon est $\frac{1}{\binom{N}{n}}$.

$$p(\omega) = \frac{1}{\binom{N}{n}} \text{ pour tout } \omega \in \Omega.$$

Proposition 1.1

Chaque unité v de la population se trouve dans l'échantillon avec probabilité $\frac{n}{N}$.

Démonstration

L'événement $\{\omega \in \Omega\}$ est de cardinalité $\binom{N-1}{n-1}$ (le nombre de façons de choisir un échantillon de taille n qui contient v. Donc $P(\{\omega \in \Omega\}) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = n/N$ ◄

Proposition 1.2

Chaque paire d'unités v_1, v_2 se trouve dans l'échantillon avec probabilité $\frac{n(n-1)}{N(N-1)}$.

Démonstration

L'événement $\{v_1 \in \Omega ;\ v_2 \in \Omega\}$ est de cardinalité $\binom{N-2}{n-2}$ (le nombre de façons de choisir un échantillon de taille n qui contient v_1 et v_2. Donc $P(\{\omega \in \Omega\}) = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}$.

Considérons maintenant l'estimateur de μ, $\bar{y} = \frac{1}{n}\sum_{i\in\omega} v_i$ la moyenne et $s = \sqrt{\frac{\sum_{i\in\omega}(y_i-\bar{y})^2}{n-1}}$ l'écart-type des données de l'échantillon◄

Proposition 1.3 $\bar{y}$ est un estimateur sans biais de $\mu = \frac{1}{N}\sum_{i=1}^{N} v_i$

Démonstration $E(\bar{y}) = \sum_{\omega\in\Omega} \bar{y}(\omega)p(\omega) = \frac{1}{\binom{N}{n}}\sum_{\omega\in\Omega}\bar{y}(\omega) = \frac{1}{\binom{N}{n}}\sum_{\omega\in\Omega}\left(\frac{1}{n}\sum_{i\in\omega} v_i\right) = \frac{1}{\binom{N}{n}n}\sum_{i=1}^{N}\sum_{i\in\omega} v_i$

où la somme $\sum_{i\in\omega} v_i$ représente la somme sur tous les échantillons ω qui. Or le nombre d'échantillons qui contiennent l'unité v_i est $\binom{N-1}{n-1}$, comme on le voit dans la démonstration de la proposition 1.1. Donc $\sum_{i\in\omega} v_i = \binom{N-1}{n-1} v_i$. D'où, $E(\bar{y}) = \frac{1}{\binom{N}{n}n}\sum_{i=1}^{N}\sum_{i\in\omega} v_i = \frac{1}{\binom{N}{n}n}\sum_{i=1}^{N}\binom{N-1}{n-1} v_i =$

$\frac{\binom{N-1}{n-1}}{\binom{N}{n}n}\sum_{i=1}^{N} v_i = \frac{\binom{N-1}{n-1}}{\binom{N}{n}n} N\mu = \mu.$ ◄

Proposition 1.4 La variance de $\bar{y}$ est $\mathrm{Var}(\bar{y}) = (1-f)\frac{S^2}{n}$

Démonstration Rappelons que $\sum_{i=1}^{N}(v_i-\bar{v})^2 = \sum_{i=1}^{N} v_i^2 - \frac{\left(\sum_{i=1}^{N} v_i\right)^2}{N}$, et que

$\left(\sum_{i=1}^{N} v_i\right)^2 = \sum_{i=1}^{N} v_i^2 + \sum_{i=1}^{N}\sum_{j\neq i} v_i v_j = \sum_{i=1}^{N} v_i^2 + 2\sum_{i=1}^{N-1}\sum_{j>i} v_i v_j$.

$\mathrm{Var}(\bar{y}) = E(\bar{y}^2) - \mu^2$. Déterminons d'abord $E(\bar{y}^2)$.

$E(\bar{y}^2) = \frac{1}{\binom{N}{n}}\sum_{\omega}\bar{y}^2$

$= \frac{1}{n^2\binom{N}{n}}\sum_{\omega\in\Omega}\left(\sum_{i\in\omega} v_i\right)^2 = \frac{1}{n^2\binom{N}{n}}\sum_{\omega\in\Omega}\left(\sum_{i\in\omega} v_i^2 + 2\sum_{i\in\omega}\sum_{\substack{j\in\omega\\ j>i}} v_i v_j\right)$

$= \frac{1}{n^2\binom{N}{n}}\sum_{\omega\in\Omega}\left(\sum_{i\in\omega} v_i^2\right) + \frac{2}{n^2\binom{N}{n}}\sum_{\omega\in\Omega}\left(\sum_{i\in\omega}\sum_{\substack{j\in\omega\\ j>i}} v_i v_j\right)$

$= \frac{1}{n}\sum_{\omega\in\Omega}\left(\frac{1}{n}\sum_{i\in\omega} v_i^2\right)\Big/\binom{N}{n} + \frac{2}{n^2\binom{N}{n}}\sum_{i=1}^{N-1}\sum_{j>i}\left(\sum_{i\in\omega,\, j\in\omega} v_i v_j\right)$

$$= \frac{1}{nN}\sum_{i=1}^{N} v_i^2 + \frac{2}{n^2\binom{N}{n}}\sum_{i=1}^{N-1}\sum_{j>i}\binom{N-2}{n-2} v_i v_j$$

$$= \frac{1}{nN}\sum_{i=1}^{N} v_i^2 + \frac{\binom{N-2}{n-2}}{n^2\binom{N}{n}}\left[2\sum_{i=1}^{N-1}\sum_{j>i} v_i v_j\right]$$

$$= \frac{1}{nN}\sum_{i=1}^{N} v_i^2 + \frac{n(n-1)}{n^2 N(N-1)}\left[\left(\sum_{i=1}^{N} v_i\right)^2 - \sum_{i=1}^{N} v_i^2\right]$$

$$= \frac{1}{nN}\left[1-\frac{n-1}{N-1}\right]\sum_{i=1}^{N} v_i^2 + \frac{(n-1)}{nN(N-1)}\left(\sum_{i=1}^{N} v_i\right)^2$$

$$= \frac{1}{nN}\left[1-\frac{n-1}{N-1}\right]\sum_{i=1}^{N} v_i^2 + \frac{(n-1)}{nN(N-1)}\left(\sum_{i=1}^{N} v_i\right)^2$$

$$= \frac{(N-n)}{nN(N-1)}\sum_{i=1}^{N} v_i^2 + \frac{(n-1)N}{n(N-1)}\mu^2$$

Donc

$$\mathrm{Var}(\bar{y}) = E(\bar{y}^2) - \mu^2 = \frac{(N-n)}{nN(N-1)}\sum_{i=1}^{N} v_i^2 + \frac{(n-1)N}{n(N-1)}\mu^2 - \mu^2$$

$$= \frac{(N-n)}{nN(N-1)}\sum_{i=1}^{N} v_i^2 + \left[\frac{(n-1)N}{n(N-1)} - 1\right]\mu^2$$

$$= \frac{(N-n)}{nN(N-1)}\sum_{i=1}^{N} v_i^2 + \left[\frac{(n-1)N}{n(N-1)} - 1\right]\mu^2$$

$$= \frac{(N-n)}{nN(N-1)}\sum_{i=1}^{N} v_i^2 - \frac{N-n}{n(N-1)}\mu^2$$

$$= \frac{N-n}{n(N-1)N}\left[\sum_{i=1}^{N} v_i^2 - N\mu^2\right] = \frac{N-n}{nN}\frac{\sum_{i=1}^{N} v_i^2 - N\mu^2}{(N-1)} = \frac{N-n}{nN}S^2 = (1-f)\frac{S^2}{n} = \left(1-\frac{n}{N}\right)S^2 \blacktriangleleft$$

Proposition 1.5 s^2 est un estimateurs sans biais de S^2.

Démonstration Nous montrerons que $E[(n\text{-}1)s^2] = (n\text{-}1)S^2$. $(n\text{-}1)s^2 = \sum_{i\in\omega}(v_i - \bar{y})^2 =$

$\sum_{i\in\omega} v_i^2 - n\bar{y}^2$. Alors $E[(n\text{-}1)s^2] = E\left[\sum_{i\in\omega} v_i^2 - n\bar{y}^2\right] = E\left[\sum_{i\in\omega} v_i^2\right] - nE\left[\bar{y}^2\right] =$

$$nE\left[\frac{1}{n}\sum_{i\in\omega} v_i^2\right] - n\{Var(\bar{y}) + [E(\bar{y})]^2\} = n\left[\frac{1}{N}\sum_{i=1}^{N} v_i^2\right] - n\left\{\left(1-\frac{n}{N}\right)\frac{S^2}{n} + \mu^2\right\} =$$

$$\frac{n}{N}\left[\sum_{i=1}^{N} v_i^2\right] - n\mu^2 - n\left\{\left(1-\frac{n}{N}\right)\frac{S^2}{n}\right\} = \frac{n}{N}\left\{\sum_{i=1}^{N} v_i^2 - N\mu^2\right\} - \left\{\left(1-\frac{n}{N}\right)S^2\right\} = (n-1)S^2 \text{ . C'est}$$

ce qu'il fallait démontrer◄

Une autre approche

Soit $z_1, z_2, \ldots, z_N$ N variables aléatoires définies par

$$z_i = \begin{cases} 1 & \text{si l'unité } i \text{ est dans l'échantillon} \\ 0 & \text{sinon} \end{cases}$$