Chapitre 5

Intervalles de confiance

5.1 Introduction

Une estimation ponctuelle, qui propose un seul chiffre comme estimation d'un paramètre, est presque certainement erronée. Si, par exemple, un échantillon de ménages donne un revenu moyen par ménage de $\bar{x} = 47\,525\$$, on se permet d'affirmer que « la moyenne de la *population* est de 47 525$ », mais on sait bien que cette estimation est entachée d'erreur puisqu'il est peu probable que $\bar{X}$ ait pris exactement la même valeur que μ. Une affirmation plus modeste, comme « la moyenne de la population se trouve entre 45 525$ et 49 525$ » a de meilleures chances d'être vraie.

C'est ce qu'on appelle une estimation par *intervalle de confiance* : on entoure $\bar{X}$ d'un intervalle [*a* ; *b*] et on affirme « μ se situe dans [*a* ; *b*] ». Cette affirmation n'est pas *nécessairement* vraie, mais avec *a* et *b* bien choisis, elle serait *probablement* vraie. On peut, en fait, s'arranger pour que la probabilité de dire vrai soit aussi élevée qu'on veut, par exemple, 95 %. Ce chiffre, appelé *niveau de confiance* et désigné par 1 - α, est fixé d'avance et les valeurs de *a* et de *b* sont alors déterminées en conséquence. On pourra dire alors que μ est se situe dans l'intervalle [*a* ; *b*] « avec 95 % de confiance ». Traditionnellement, le niveau de confiance de 90 % (parfois moins) à 99 % (parfois plus).

Définition *Niveau de confiance*

Soit $\boldsymbol{X} = [X_1; X_2; \ldots ; X_n]$ un échantillon aléatoire simple. Un intervalle de confiance de niveau 1 - α pour un paramètre θ est un intervalle aléatoire $[a(\boldsymbol{X})\ ;\ b(\boldsymbol{X})]$ tel que

$$\mathrm{P}\big(\theta \in [a(\boldsymbol{X})\,;\,b(\boldsymbol{X})]\big) \geq 1-\alpha\,. \qquad (5.1.1)$$

La notion d'intervalle de confiance est applicable à tout paramètre et dans ce chapitre nous montrerons comment construire un intervalle de confiance pour une moyenne, pour une proportion et pour une variance (ou écart-type).

5.2 Intervalle de confiance pour μ — σ connu ou *n* grand

On montre ici comment déterminer un intervalle de confiance pour la moyenne μ d'une population. Pour illustrer les principes sans s'embarrasser de difficultés techniques, nous supposerons d'abord que l'écart-type σ de la population est *connu*. Considérons un exemple dans lequel un intervalle de confiance est choisi arbitrairement. Il illustrera la propriété essentielle qu'un intervalle de confiance doit satisfaire.

Exemple 5.2.1 *Intervalle de confiance de largeur fixe arbitraire*

On tire un échantillon de 80 ménages afin d'estimer μ, le revenu moyen des ménages d'une ville. Supposons pour l'instant que l'écart-type des revenus de la population est connu, σ = 2000 $. De façon arbitraire, on décide que l'intervalle de confiance sera $[\bar{X} - 400; \bar{X} + 400]$. En d'autres termes, si l'échantillon donne une moyenne de 47525$, alors on affirmera que la moyenne de la population est entre 47 125 $ et 47 925 $. Quelle est la probabilité qu'une affirmation basée sur cette règle soit vraie ?

Solution Cette affirmation sera vraie si $\mu \in [\bar{X} - 400; \bar{X} + 400]$, ou encore, si $\bar{X} - 400 \leq \mu \leq \bar{X} + 400$. La probabilité de cet événement est $\mathrm{P}(\bar{X} - 400 \leq \mu \leq \bar{X} + 400) = \mathrm{P}(-400 \leq \bar{X} - \mu \leq 400)$. On suppose que $\bar{X}$ est de loi normale (en partie parce qu'il est plausible que la distribution des revenus des ménages ne soit pas trop asymétrique, et en partie parce que le théorème limite central nous

permet d'affirmer que $\bar{X}$ est à peu près normale de toute façon). Son espérance est μ et son écart-type est $\sigma/\sqrt{n} = 223{,}6068$.

Alors $P(-400 \le \bar{X} - \mu \le 400) = P\left(-\frac{400}{223,6068} \le \frac{\bar{X}-\mu}{223,6068} \le \frac{400}{223,6068}\right) = 0{,}9264.$ ■

Dans cet exemple, nous avons choisi une marge, puis examiné les conséquences de ce choix, dont le fait qu'il donne une probabilité de 92,64 % de dire vrai lorsqu'on affirme que μ est dans l'intervalle. Le *niveau de confiance* est donc de 92,64 %. Ce niveau de confiance peut être considéré comme adéquat ou pas. S'il ne l'est pas, il faudra élargir l'intervalle, et procéder ainsi jusqu'à ce qu'on obtienne un intervalle d'un niveau adéquat.

Il est clair qu'il vaut mieux inverser l'ordre et commencer par fixer le niveau de confiance pour ensuite déterminer la marge en conséquence.

Remarque *Le lecteur pourrait, avant de poursuivre la lecture, expérimenter avec le dernier exemple et tenter d'obtenir un intervalle de confiance à un niveau supérieur : il faudra remplacer la marge* « 400 » *dans* $[\bar{X} - 400 \le \mu \le \bar{X} + 400]$ *par un chiffre…plus grand ou plus petit ? Quel est le niveau de l'intervalle* $[\bar{X} - 500 \le \mu \le \bar{X} + 500]$ *? Et de l'intervalle* $[\bar{X} - 600 \le \mu \le \bar{X} + 600]$ *?* ■

Exemple 5.2.2 *Un autre intervalle de confiance pour* μ *de même niveau*

Dans l'exemple 5.2.1, déterminer un intervalle de confiance de la forme $[\bar{X} - a \ ; \ \bar{X} + a]$ à un niveau de confiance fixé à 95 %.

Solution Il faudra choisir a de telle sorte que

$$P(\bar{X} - a \le \mu \le \bar{X} + a) = 0{,}95 \Leftrightarrow P(-a \le \bar{X} - \mu \le a) = 0{,}95$$

$$\Leftrightarrow P\left(-\frac{a}{223,6068} \le \frac{\bar{X}-\mu}{223,6068} \le \frac{a}{223,6068}\right) = 0{,}95.$$

Étant donné que $\frac{\bar{X}-\mu}{223,6068} \sim \mathcal{N}(0 ; 1)$,

$$\frac{a}{223,6068} = 1{,}96 \text{ et } a = 1{,}96(223{,}6068) = 438{,}2693.$$

L'intervalle de confiance sera donc

$$[\bar{X} - 438{,}2693 ; \bar{X} + 438{,}2693].$$

On peut vérifier qu'il s'agit bien d'un intervalle de confiance à 95 % :

$$P(\bar{X} - 438{,}2693 \le \mu \le \bar{X} + 438{,}2693) = P(-438{,}2693 \le \bar{X} - \mu \le 438{,}2693)$$

$$= P\left(-1{,}96 \le \frac{\bar{X}-\mu}{223,6068} \le 1{,}96\right) = 0{,}95.$$ ■

Présentation générale

Soit $X_1;\ldots; X_n$ un échantillon aléatoire d'une population $\mathcal{N}(\mu ; \sigma^2)$. Alors

$$\bar{X} \sim \mathcal{N}(\mu ; \sigma_{\bar{X}}^2) \text{ où } \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \qquad (5.2.1)$$

et

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \sim N(0\,;\,1). \tag{5.2.2}$$

Du fait que $P\left(-z_{\alpha/2} \le \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \le z_{\alpha/2}\right) = 1 - \alpha$, nous obtenons, en inversant les inégalités,

$$P\left(\bar{X} - z_{\alpha/2}\sigma_{\bar{X}} \le \mu \le \bar{X} + z_{\alpha/2}\sigma_{\bar{X}}\right) = 1\text{-}\alpha,$$

ce qui signifie que la probabilité que l'intervalle

$$\left[\bar{X} - z_{\alpha/2}\sigma_{\bar{X}}\,;\bar{X} + z_{\alpha/2}\sigma_{\bar{X}}\right] \tag{5.2.3}$$

contienne μ est $1 - \alpha$, et de ce fait, c'est un *intervalle de confiance à* $100(1 - \alpha)\%$.

Remarque *En « inversant les inégalités » plus haut, ce que nous avons fait, essentiellement, c'est identifier l'ensemble des valeurs de* μ *qui satisfont les inégalités* $-z_{\alpha/2} \le \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \le z_{\alpha/2}$. ■

Le cas où σ *est inconnu mais n est grand*

Normalement, la formule $[\,\bar{X} - z_{\alpha/2}\sigma_{\bar{X}} \le \mu \le \bar{X} + z_{\alpha/2}\sigma_{\bar{X}}\,]$ ne peut être calculée en pratique étant donné que $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, il faut connaître σ, ce qui n'arrive que dans des situations très exceptionnelles. Nous devons alors estimer $\sigma_{\bar{X}}$, ce qu'on fait naturellement en remplaçant σ par un estimateur S dans la formule $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$. Puisque

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

est un estimateur sans biais de σ^2, nous estimerons $\sigma^2_{\bar{X}}$ par

$$\hat{\sigma}^2_{\bar{X}} = \frac{S^2}{n}$$

et $\sigma_{\bar{X}}$ par

$\hat{\sigma}_{\bar{X}} = \dfrac{S}{\sqrt{n}}$.

Si n est grand, la statistique

$$T = \frac{\bar{X} - \mu}{\hat{\sigma}_{\bar{X}}} \tag{5.2.4}$$

suit à peu près une loi $\mathcal{N}(0 ; 1)$. La procédure décrite dans cette section reste intacte, avec pour seule modification le remplacement de $\sigma_{\bar{X}}$ par $\hat{\sigma}_{\bar{X}}$ dans la formule $Z = \dfrac{\bar{X} - \mu}{\sigma_{\bar{X}}}$.

Exemple 5.2.3 *Estimation d'une moyenne*

D'une grande population de comptes de banque, on prélève un échantillon de taille $n = 30$ afin d'estimer la valeur moyenne d'un compte ainsi que le montant total 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 μ de la population et l'écart-type de l'estimateur;

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

Solution

a) Nous avons $\sum_{i=1}^{30} x_i = 4\,968{,}68$; $\bar{x} = 165{,}62$; $\sum_{i=1}^{30} x_i^2 = 1\,864\,916{,}176$; $S = 189{,}55$; $\hat{\sigma}_{\bar{X}} = 189{,}55/\sqrt{30} = 34{,}61$.

b) Un intervalle de confiance approximatif à 95% est donné par

[165,62-1,96(34,61); 165,62+1,96(34,61)] = [165,62- 67,83 ; 165,62 + 67,83] = [97,79 ; 233,45].

La demi-largeur de l'intervalle, 67,83, est parfois appelée *marge d'erreur*. ■

5.3 Intervalle de confiance pour μ — σ inconnu

En général, la statistique

$$T = \frac{\bar{X} - \mu}{\hat{\sigma}_{\bar{X}}} \tag{5.3.1}$$

ne suit pas une loi $\mathcal{N}(0 ; 1)$: ce n'est que lorsque n est grand que T suit *à peu près* une loi $\mathcal{N}(0 ; 1)$. *Lorsque la population est normale*, cependant, T suit une loi connue, la *loi de Student* dont le seul paramètre, le *nombre de degrés de liberté*, dépend uniquement de n et non des paramètres μ ou σ.

La loi de Student

La fonction de densité de la loi de *Student* est unimodale et symétrique par rapport à l'origine. Sa forme est semblable à celle de la loi $\mathcal{N}(0 ; 1)$, sauf qu'elle décroît moins rapidement à mesure que la variable s'éloigne de l'origine. Son espérance est sa variance sont

$$\mathrm{E}(T) = 0 \text{ et } \mathrm{Var}(T) = \frac{\nu}{\nu - 2}, \quad \nu > 2. \tag{5.3.2}$$

On remarque que la variance de T est supérieure à 1, mais tend vers 1 lorsque ν tend vers l'infini. On désigne par $t_{\nu;\beta}$ le point à droite duquel la surface sous la courbe de *Student* à ν degrés de liberté est égale à β :

$$P(T > t_{\nu;\beta}) = \beta. \tag{5.3.3}$$

Par un raisonnement identique à celui suivi plus haut dans le cas où σ est connu, nous obtenons la formule suivante d'un intervalle de confiance à 100(1- α)% :

$$\bar{X} - t_{n-1;\alpha/2}\hat{\sigma}_{\bar{X}} \leq \mu \leq \bar{X} + t_{n-1;\alpha/2}\hat{\sigma}_{\bar{X}}\,. \tag{5.3.4}$$

Exemple 5.2.4 *Estimation d'une moyenne, utilisant la loi de Student*

Reprenons les données de l'exemple 5.2.3 et déterminons un intervalle de confiance à 95 % en tenant compte cette fois-ci que la variable T est de loi de *Student* à 29 degrés de liberté. Seul le point critique change : de $z_{\alpha/2} = 1{,}96$ il passe à $t_{29;\alpha/2} = 2{,}04$.

L'intervalle de confiance, [165,62-2,04(34,61); 165,62+2,04(34,61)] = [95,02 ; 236,22] est, bien sûr, plus large que celui obtenu à l'exemple 5.2.3, reflétant la plus grande incertitude due à l'estimation de σ. ■

5.4 Estimation de la variance σ²

Jusqu'ici, nous avons estimé σ^2 uniquement parce qu'il nous était nécessaire de le faire pour estimer la précision de $\bar{X}$ comme estimateur de μ. Mais la variance est un paramètre en soi, d'un intérêt indépendant. Par exemple, une des qualités souhaitées d'un procédé de fabrication est une *constance* dans les mesures des articles qu'il produit. S'il est important, par exemple, que les longueurs de certaines pièces usinées soient égales à 2 cm *en moyenne*, il est également important — sinon plus — de s'assurer que les longueurs des pièces soient très proches de cette moyenne, c'est-à-dire, que la variance soit petite.

Théorème 5.4.1 *La distribution de* S^2

Soit $X_1, X_2, \ldots, X_n$ un échantillon d'une population $\mathcal{N}(\mu;\sigma^2)$. Alors

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2 \sim \chi^2_{n-1} \tag{5.4.1}$$

Soit a et b deux nombres positifs tels que

$$P\left[a \leq \frac{(n-1)S^2}{\sigma^2} \leq b\right] = 1 - \alpha \tag{5.4.2}$$

L'intervalle de confiance est l'ensemble de toutes les valeurs de σ^2 qui satisfont les inégalités $a \leq \frac{(n-1)S^2}{\sigma^2} \leq b$. Ces valeurs sont

$$a \leq \frac{(n-1)S^2}{\sigma^2} \leq b \Leftrightarrow \left\{\frac{(n-1)S^2}{b} \leq \sigma^2 \leq \frac{(n-1)S^2}{a}\right\}.$$

L'intervalle $\left[\frac{(n-1)S^2}{b} \le \sigma^2 \le \frac{(n-1)S^2}{a}\right]$ est donc un intervalle de confiance à 100(1-α) %.

Les valeurs de *a* et *b* ne sont pas uniques : il existe une infinité de valeurs de (*a* ; *b*) pour lesquelles $P\left[a \le \frac{(n-1)S^2}{\sigma^2} \le b\right]$ = 1 - α. Lesquelles choisir ? Une possibilité : prendre pour *a* le point dont la surface à gauche est α/2 ; et pour *b* le point dont la surface à droite est α/2. Ce choix n'est pas optimal, mais il est commode.

Notation On désigne par $\chi^2_{\nu;\beta}$ le point tel que, si $X \sim \chi^2_\nu$, $P(X > \chi^2_{\nu;\beta}) = \beta$

On a donc *a* = $\chi^2_{n;1-\alpha/2}$ et *b* = $\chi^2_{n;\alpha/2}$. La formule d'un intervalle de confiance à 100(1- α) % pour σ^2 est donc

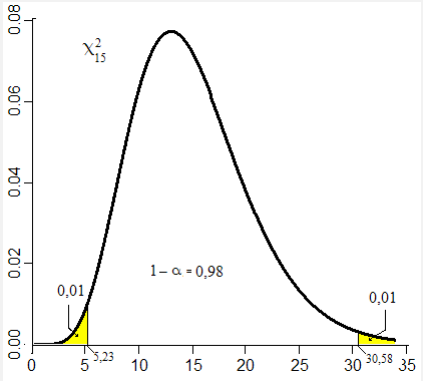
$$\left(\frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}} ; \frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}}\right) \qquad (5.4.3)$$

Il est clair que cet intervalle de confiance pour σ^2 détermine un intervalle de confiance pour σ, construit en prenant les racines carrées des deux bornes.

Exemple 5.4.1 *Intervalle de confiance pour* σ^2

Supposons qu'un échantillon de taille *n* = 16 donne le résultat S^2 = 4,84 et qu'on veuille déterminer un intervalle de confiance de 98 %. Nous posons α = 0,02 et à l'aide des tables, on trouve $\chi^2_{15;0,99}$ = 5,23 et $\chi^2_{15;0,01}$ = 30,58. Donc

$$P\left(5,23 < \frac{15S^2}{\sigma^2} < 30,58\right) = 0,98$$



L'intervalle de confiance est

$$\left(\frac{15S^2}{30,58} ; \frac{15S^2}{5,23}\right),$$

soit

$$2,37 < \sigma^2 < 13,88.$$

Sachant que l'intervalle $\left(\frac{(n-1)S^2}{30,58} ; \frac{(n-1)S^2}{5,23}\right)$ a une probabilité de 0,98 de recouvrir σ^2, on peut affirmer avec 98 % de confiance que l'intervalle [2,37 ; 13,88] contient σ^2.

À ce même niveau de confiance on peut affirmer que

$$\sqrt{2,37} < \sigma < \sqrt{13,88}\,. \qquad \blacksquare$$

5.5 Estimation d'une proportion *p*

Considérons une population dont une proportion p des membres appartient à une certaine classe $\mathcal{A}$. Supposons que dans un échantillon de taille n, on trouve X unités appartenant à la classe $\mathcal{A}$. Si les tirages sont effectués avec remise, ou si la population est grande, alors $X \sim \mathcal{B}(n;p)$. La distribution de X s'approche d'une $\mathcal{N}(np\,;\,npq)$, $q = 1 - p$. On peut déterminer un intervalle de confiance pour p à partir de son estimateur $\hat{p} = \dfrac{X}{n}$. Puisque

$$\hat{p} \sim \mathcal{N}\left(p\,;\,\sigma_{\hat{p}}^2\right) \text{ où } \sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}\,, \tag{5.5.1}$$

$Z = \dfrac{\hat{p} - p}{\sigma_{\hat{p}}}$ est à peu près de loi $\mathcal{N}(0\,;\,1)$, et

$$P\left(-z_{\alpha/2} \le \frac{\hat{p} - p}{\sigma_{\hat{p}}} \le z_{\alpha/2}\right) \approx 1 - \alpha\,. \tag{5.5.2}$$

L'intervalle de confiance est l'ensemble des valeurs de p qui satisfont les inégalités

$$-z_{\alpha/2} \le \frac{\hat{p} - p}{\sigma_{\hat{p}}} \le z_{\alpha/2}$$

La solution de ces inégalités, qu'on peut également écrire sous la forme

$$\hat{p} - z_{\alpha/2}\sigma_{\hat{p}} \le p \le \hat{p} + z_{\alpha/2}\sigma_{\hat{p}}\,, \tag{5.5.3}$$

n'est pas immédiate, car $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$ est fonction de p. Elle peut être résolue: il s'agit de trouver les racines d'une fonction quadratique. Mais une solution plus simple consiste à remplacer $\sigma_{\hat{p}}$ par $\hat{\sigma}_{\hat{p}} = \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$, ce qui donne l'intervalle

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\,. \tag{5.5.4}$$

Cet intervalle est légitime dans la mesure où il est l'ensemble des valeurs de p qui satisfont les inégalités $-z_{\alpha/2} \le \hat{Z} \le z_{\alpha/2}$, où $\hat{Z} = \dfrac{\hat{p} - p}{\hat{\sigma}_{\hat{p}}}$, et la statistique $\hat{Z}$ est, elle aussi à peu près normale si n est assez grand. Cette solution est la plupart du temps tout à fait acceptable, car $-z_{\alpha/2} \le \dfrac{\hat{p} - p}{\hat{\sigma}_{\hat{p}}} \le z_{\alpha/2} \Leftrightarrow -z_{\alpha/2} \le \dfrac{\sqrt{n}(\hat{p} - p)}{\sqrt{\hat{p}(1-\hat{p})}} \le z_{\alpha/2} \Leftrightarrow z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{p(1-p)}} \le \dfrac{\sqrt{n}(\hat{p} - p)}{\sqrt{p(1-p)}} \le z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{p(1-p)}}$. Or à

moins qu'il y ait une grande différence entre p et $\hat{p}$, le facteur $\sqrt{\frac{\hat{p}(1-\hat{p})}{p(1-p)}}$ est généralement proche de 1.

Exemple 5.5.1 *Intervalle de confiance pour p*

Dans un sondage auprès de 2000 citoyens on trouve que 1200 sont en faveur d'une hausse proposée des droits de scolarité. Déterminer un intervalle de confiance à 95 % pour la proportion p de la population en faveur de la hausse.

Solution On a $n = 2000$, et on calcule $\hat{p} = 0{,}6$, $\hat{\sigma}_{\hat{p}} = \sqrt{\frac{(0{,}6)(0{,}4)}{2000}} = 0{,}01095$.

L'intervalle de confiance est $[\hat{p} - 1{,}96\hat{\sigma}_{\hat{p}}^2 ; \hat{p} + 1{,}96\hat{\sigma}_{\hat{p}}^2] = [0{,}579 ; 0{,}622]$.

On peut donc conclure avec 95 % de confiance que le pourcentage de la population en faveur de la hausse se situe entre 57,9% et 62,2 %. Si on avait utilisé la formule plus complexe, nous aurions trouvé [0,578 ; 0,621] pour intervalle de confiance, une différence tout à fait négligeable. ■

La demi-largeur de l'intervalle de confiance, 0,021 ici, est appelée communément *marge d'erreur*. Dans les media, la conclusion tirée d'un intervalle de confiance s'exprime par « une marge d'erreur de 2,1 % 19 fois sur 20 », ce « 19 fois sur 20 » faisant référence au niveau de confiance de 95 %. C'est cette marge d'erreur qui motive le choix de la taille de l'échantillon : un échantillon de 2000 est considéré satisfaisant parce qu'il donne une marge d'erreur jugée acceptable. Théoriquement, donc, on détermine la taille de l'échantillon en fonction de la marge d'erreur voulue. Or cette marge d'erreur, estimée à partir de l'échantillon par $z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, ne peut être calculée qu'une fois l'échantillon prélevé.

Détermination de la taille de l'échantillon

Y a-t-il moyen de déterminer une taille d'échantillon adéquate *avant* de tirer l'échantillon ? Le prochain exemple illustre quelques approches.

Exemple 5.5.2 *Détermination de la taille de l'échantillon*

Déterminer la taille de l'échantillon nécessaire pour estimer une proportion p avec une marge d'erreur de 4 % à un niveau de confiance de 5 %.

Solution La condition qui doit être satisfaite est $1{,}96\sqrt{\frac{p(1-p)}{n}} = 0{,}04 \Leftrightarrow n = \frac{(1{,}96)^2}{(0{,}04)^2}p(1-p)$. Il faut maintenant trouver une façon d'estimer p, ou du moins $p(1\text{-}p)$. Examinons les conséquences sur n des différents choix :

p	0,05 ou 0,95	0,1 ou 0,9	0,15 ou 0,85	0,2 ou 0,8	0,25 ou 0,75	0,3 ou 0,7	0,35 ou 0,65	0,4 ou 0,6	0,45 ou 0,55	0,5
n	114	216	306	384	450	504	546	576	594	600

Première possibilité : Une certaine estimation de p existe grâce à des sondages antérieurs. On peut l'utiliser telle quelle ou, pour plus de sécurité, modifier l'estimation en direction d'une plus grande valeur de $p(1\text{-}p)$. Par exemple, si dans le dernier sondage la probabilité a été estimée à 70 %, on peut prendre 0,7 comme valeur de p, ce qui donnerait un échantillon de taille 504, ou s'approcher un peu plus de ½ et poser p = 65 %, ce qui donne un échantillon de taille 546.

Deuxième possibilité : Quand aucune estimation préalable n'existe, on peut substituer au produit $p(1\text{-}p)$ sa valeur maximale, qui est atteinte lorsque $p = 0{,}5$. C'est une estimation pessimiste qui

donne un échantillon de 600. Quelle que soit la valeur de p, on a l'assurance qu'avec un échantillon de 600, la marge d'erreur ne dépassera pas 4 %, le seuil visé. Cette solution est raisonnable lorsqu'on sait que p ne s'éloigne pas trop de 0,5. Dans un sondage destiné à estimer plusieurs proportions, on peut s'attendre à ce que certaines d'entre elles se situeront dans un voisinage de 0,05.

Troisième possibilité : Une taille d'échantillon calculée sous l'hypothèse que $p = ½$ peut s'avérer inutilement grande (et donc coûteuse) si en fait p est très petit ou trop grand (voir le tableau ci-dessus). Il n'est pas raisonnable de prendre pour p la valeur 1/2 si on doit estimer, par exemple, la proportion p des gens atteints d'une maladie rare, puisqu'on sait que p est petit. En fait, il est toujours possible dans ces situations d'émettre une borne supérieure p_o dont on sait a priori qu'elle est supérieure à p. Si $p_o < 1/2$, une estimation pessimiste de $p(1-p)$ est $p_o(1-p_o)$. Cette estimation est également pessimiste lorsqu'on sait que $p \geq p_o > 1/2$. ■

Dans le dernier exemple, la marge d'erreur est exprimée en termes absolus, l'unité est un *point de pourcentage*. Or en pratique la marge d'erreur tolérable dépend de la proportion p à estimer. Une marge d'erreur de 4 % peut être acceptable si $p = 0,5$ mais pas du tout si $p = 0,05$. C'est pour cela qu'on exprime parfois la marge d'erreur en termes *relatifs* : on dira, par exemple, qu'on souhaite que la marge d'erreur *relative* soit de 20 %, ce qui veut dire que la marge d'erreur absolue doit être de $0,2p$.

La condition à satisfaire est donc $1,96\sqrt{\frac{p(1-p)}{n}} = 0,2p$, ce qui donne

$$n = \frac{(1,96)^2}{(0,2)^2}\frac{(1-p)}{p}. \qquad (5.5.5)$$

Cette quantité ne peut pas être bornée supérieurement à moins de pouvoir borner p *inférieurement*, c'est-à-dire, si on peut affirmer que $p \geq p_o$, alors on peut conclure que $\frac{(1-p)}{p} \leq \frac{(1-p_o)}{p_o}$ et on prendra $n = \frac{(1,96)^2}{(0,2)^2}\frac{(1-p_o)}{p_o}$.

5.6 Une approche générale

L'approche que nous avons suivie est, à quelques détails techniques près, la même pour les trois paramètres étudiés. Elle est basée sur la notion de *pivot* :

Définition *Pivot*

> Un *pivot* est une fonction des observations *et* du paramètre dont la distribution est entièrement connue.

Voici les pivots sur lesquels ont été construits les intervalles étudiés jusqu'ici :

Paramètre	μ (σ connu)	μ (σ inconnu)	p	σ^2
Pivot	$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$	$T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$	$\hat{Z} = \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}}$	$Q = \frac{(n-1)S^2}{\sigma^2}$
Loi	$\mathcal{N}(0\,;\,1)$	t_{n-1}	$\mathcal{N}(0\,;\,1)$ à peu près	χ^2_{n-1}
Équation à résoudre	$-z_{\alpha/2} \leq Z \leq z_{\alpha/2}$	$-t_{\alpha/2;n-1} \leq T \leq t_{\alpha/2;n-1}$	$-z_{\alpha/2} \leq Z \leq z_{\alpha/2}$	$\chi^2_{n-1;1-\alpha/2} \leq Q \leq \chi^2_{n-1;\alpha/2}$

On remarque, dans la première ligne du tableau (« Pivot »), que le pivot fait toujours intervenir le paramètre: μ dans les deux premières cases; p et σ^2 dans les deux dernières, respectivement.

Dans la deuxième ligne du tableau, notez que les lois ne comportent aucun paramètre: elles sont entièrement connues. C'est ce qui permet de trouver deux points, *a* et *b* tels que le pivot se situe dans l'intervalle [*a* ; *b*] avec probabilité 1 - α. Le choix de ces points n'est pas unique. Dans les trois premiers cas (μ et *p*), la symétrie de la distribution fait que le choix $a = -b$ est optimal dans le sens qu'il donne l'intervalle de confiance le plus court. Dans le cas de σ^2, le choix proposé n'est pas optimal, mais facile à déterminer.

Remarque *Comment trouver un pivot ? Jusqu'ici, avec la loi normale et la loi de* Student, *nous avons pu déterminer un pivot sans difficulté. Mais est-ce toujours aussi aisé ? Quel serait un pivot, par exemple, pour un échantillon d'une population de loi exponentielle de paramètre* β *? Ce n'est pas compliqué ici non plus, puisque nous pouvons montrer que si X est de loi exponentielle de paramètre* β, *alors X/*β *est de loi exponentielle de paramètre* 1 *et par conséquent* $\Sigma X_i/\beta$ *est de loi gamma de paramètre n et* β. *Il suffira alors de déterminer deux valeurs a et b telles que* $P(a \le \Sigma x_i/\beta \le b) = 1 - \alpha$. *Ce qui pose quelques problèmes techniques de calcul, mais surmontables avec les moyens informatiques disponibles aujourd'hui. Libérés, donc, de ces difficultés techniques, pouvons-nous énoncer une procédure générale qui pourrait être toujours employée, sous réserve d'une solution au problème de calcul ? Théoriquement, il existe une telle procédure dans le cas continu, basée sur le fait que la fonction de répartition d'une variable est une variable de loi uniforme dans* [0 ; 1]. *C'est-à-dire, si X est un variable aléatoire continue de fonction de répartition* $F(x;\theta)$ *alors* $Y = F(X;\theta)$ *est une variable de loi uniforme sur* [0 ; 1], *et c'est donc un pivot. Si* $\hat{\theta}$ *est une fonction des observations, par exemple, un estimateur de* θ, *alors il suffira de résoudre les inégalités* $a \le F(\hat{\theta};\theta) \le b$, *où F est la fonction de répartition de* $\hat{\theta}$. *Pour a et b on peut prendre les valeurs* $a = \alpha/2$ *et* $b = 1 - \alpha/2$. *Ainsi donc, si X = x est la valeur observée d'une variable de loi gamma de paramètres n et* β, *un intervalle de confiance pour* β *est l'ensemble des valeurs de* β *qui satisfont les inégalités* $\alpha/2 \le \frac{1}{\Gamma(n)\beta^n}\int_0^x t^{n-1}e^{-t/\beta}dt \le 1-\alpha/2$ *—ce qui est théoriquement possible bien que souvent difficile numériquement.* ■

Dans un grand nombre d'applications, pour divers paramètres et populations, l'approximation normale fournit des solutions simples. Considérons, par exemple, l'estimation du paramètre λ d'une population de loi de Poisson.

Exemple 5.6.1 *Intervalle de confiance pour le paramètre* λ *d'une loi de Poisson*

Soit $X_1; X_2; \ldots; X_n$ un échantillon aléatoire simple d'une population de loi de Poisson de paramètre λ. Nous savons qu'alors $X = \Sigma X_i$ est de loi de Poisson de paramètre $n\lambda$. Si *x* est la valeur observée de *X*, la fonction de répartition de *X*, *évaluée à x, est* $\sum_{i=0}^{x}\frac{e^{-n\lambda}(n\lambda)^i}{n!}$ et il faut donc résoudre les inégalités $\frac{\alpha}{2} \le \sum_{i=0}^{x}\frac{e^{-n\lambda}(n\lambda)^i}{n!} \le 1-\frac{\alpha}{2}$, ce qui est loin d'être évident. Mais $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, est approximativement de loi normale de moyenne λ et de variance λ/n. Alors $P\left(-z_{\alpha/2} \le \frac{\bar{X}-\lambda}{\sqrt{\lambda/n}} \le z_{\alpha/2}\right) = 1-\alpha$ et la solution des inégalités $-z_{\alpha/2} \le \frac{\bar{X}-\lambda}{\sqrt{\lambda/n}} \le z_{\alpha/2}$ est $\bar{X} + \frac{z_{\alpha/2}^2}{2n} \pm \frac{z_{\alpha/2}\sqrt{\bar{X}}}{\sqrt{n}}\sqrt{1+\frac{z_{\alpha/2}^2}{4n\bar{X}}}$, ce qui est proche de $\bar{X} - z_{\alpha/2}\frac{\sqrt{\bar{X}}}{\sqrt{n}} \le \lambda \le \bar{X} + z_{\alpha/2}\frac{\sqrt{\bar{X}}}{\sqrt{n}}$. Cette dernière expression viendrait naturellement à l'esprit, puisqu'elle correspond à l'intervalle $\hat{\lambda} - z_{\alpha/2}\hat{\sigma}_{\bar{X}} \le \lambda \le \hat{\lambda} + z_{\alpha/2}\hat{\sigma}_{\bar{X}}$, $\hat{\lambda} = \bar{X}$ étant l'estimateur de λ. Son écart-type $\sigma/\sqrt{n} = \sqrt{\lambda/n}$ est estimé par $\sqrt{\hat{\lambda}/n} = \sqrt{\bar{X}/n}$. Pour avoir une idée de la qualité de cette approximation, considérons un échantillon de taille $n = 30$ dans lequel $\bar{X} =$

2/3. L'approximation normale donne l'intervalle [0,428 ; 0,905]. La méthode exacte donne [0,407 ; 1,03]. Plusieurs calculs par ailleurs révèlent que l'approximation normale est tout à fait adéquate, même avec de petits échantillons. ■

RÉSUMÉ

1 *Intervalle de confiance pour* μ, σ *connu* $[\ \overline{X} - z_{\alpha/2}\sigma_{\overline{X}} \le \mu \le \overline{X} + z_{\alpha/2}\sigma_{\overline{X}}\ ]$ où $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$.

2 *Intervalle de confiance pour* μ, σ *inconnu mais n grand* $\overline{X} - z_{\alpha/2}\hat{\sigma}_{\overline{X}} \le \mu \le \overline{X} + z_{\alpha/2}\hat{\sigma}_{\overline{X}}$ où $\hat{\sigma}_{\overline{X}} = \dfrac{S}{\sqrt{n}}$.

3 *La loi de Student* Lorsque la population est normale $T = \dfrac{\overline{X} - \mu}{\hat{\sigma}_{\overline{X}}} \sim t_{n-1}$.

5 *Intervalle de confiance pour* μ, σ *inconnu, population normale* $\overline{X} - t_{n-1;\alpha/2}\hat{\sigma}_{\overline{X}} \le \mu \le \overline{X} + t_{n-1;\alpha/2}\hat{\sigma}_{\overline{X}}$.

6 *Intervalle de confiance pour* σ^2 : $\left(\dfrac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}} ; \dfrac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}} \right)$.

7 *Intervalle de confiance pour une proportion p* $\hat{p} - z_{\alpha/2}\hat{\sigma}_{\hat{p}} \le p \le \hat{p} + z_{\alpha/2}\hat{\sigma}_{\hat{p}}$, $\hat{\sigma}_{\hat{p}} = \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$.