Chapitre 11

Intervalle de confiance d'une estimation

La méthode d'échantillonnage aléatoire présentée dans le chapitre précédent permet de préciser les marges d'erreur des estimateurs, calculés à partir de l'échantillon lui-même. Cet aspect est crucial car une estimation sans indication du degré de précision est douteuse ; elle ne peut être ni appréciée ni distinguée d'une valeur quelconque qui aurait été avancée sur la base de l'intuition ou d'une simple connaissance du sujet.

Ce qui est remarquable dans la méthode d'échantillonnage aléatoire, c'est que l'échantillon contient non seulement l'information nécessaire pour obtenir une estimation de la quantité voulue, mais aussi celle nécessaire pour calculer le degré de précision de l'estimateur. Dans ce chapitre, nous abordons les méthodes pour déterminer la précision des estimateurs.

11.1 Méthode de construction d'un intervalle de confiance

Soit θ un paramètre à estimer de la population et T son estimateur à partir d'un échantillon aléatoire. On évalue la précision de T comme estimateur de θ en construisant un **intervalle de confiance** autour de l'estimateur, qui souvent s'interprète comme une marge d'erreur.

Pour construire cet intervalle de confiance, on procède, en terme général, de la manière suivante. À partir de la loi de distribution de l'estimateur T, on détermine un intervalle calculé sur la base de l'échantillon tel que la probabilité soit importante qu'il englobe la vraie valeur du paramètre recherché. Soit $(T-e, T+e)$ cet intervalle et $(1-\alpha)$ la probabilité d'appartenance, on peut dire que la marge d'erreur e est liée à α par la probabilité :

$$P(T - e \leq \theta \leq T + e) = 1 - \alpha.$$

Le niveau de probabilité associé à un intervalle d'estimation est appelé **niveau de confiance** ou **degré de confiance**.

L'intervalle, $T - e \leq \theta \leq T + e$, est appelé intervalle de confiance de l'estimateur de θ au niveau de confiance $1 - \alpha$. Prenons comme exemple $\alpha = 5\%$, l'intervalle de confiance du paramètre θ à un seuil de probabilité de 95%. Ceci veut dire qu'en utilisant T comme estimateur de θ, en moyenne, sur 100 échantillonnages, 95 fois l'intervalle construit de la façon indiquée comprendra la vraie valeur de l'estimateur et 5 fois il ne l'incluera pas.

La quantité e de l'intervalle de confiance mesure la moitié de l'étendue de l'intervalle. Elle indique donc, dans un certain sens, la marge d'erreur de l'estimateur. Un estimateur est d'autant plus efficace que, pour un niveau de confiance $1 - \alpha$ donné, il conduit à un intervalle de confiance plus petit.

Dans la suite de ce chapitre, nous étudierons l'intervalle de confiance relatif à l'estimation de θ suivant la nature du paramètre θ à estimer, la forme de la loi de distribution de l'estimateur T, la taille de l'échantillon et la connaissance ou l'ignorance de la variance de la population.

11.2 Intervalle de confiance pour la moyenne d'une distribution normale

Souvent, l'échantillon est utilisé pour estimer une moyenne μ concernant la population, par exemple, la moyenne d'âge de la population, le prix moyen d'un litre d'essence ou la durée moyenne de vie d'une marque de pile électrique. Dans ce cas, le paramètre θ à estimer est μ (donc $\theta = \mu$) et l'estimateur à partir de l'échantillon peut être la moyenne des observations, $\bar{X} = (X_1 + X_2 + \cdots + X_n)/n$, où n dénote la taille de l'échantillon.

Si l'échantillon provient d'une population de distribution normale, nous avons vu dans le chapitre précédent que la variable aléatoire $\bar{X}$, suit elle-même

une distribution normale de moyenne μ et d'écart-type $\sigma_{\bar{X}}$, que nous abrégeons par l'expression :

$$\bar{X} \sim N(\mu, \sigma_{\bar{X}}).$$

Suivant la démarche décrite dans la section précédente, il s'agit de trouver l'intervalle autour de μ tel que :

$$P(\bar{X} - e \leq \mu \leq \bar{X} + e) = 1 - \alpha.$$

La quantité e dépend de la nature de la variance de la population. Il se peut que des expériences préalables nous aient fourni une estimation de la variance de la population. Dans ce cas, la variance σ^2 peut être considérée comme connue. Dans le cas contraire, σ^2 est inconnu et il faudra l'estimer sur la base de l'échantillon. Nous allons traiter séparément ces deux situations.

11.2.1 σ connu

Quand l'écart-type σ de la population est connu, la valeur de e est égale à $z_{\alpha/2} \cdot \sigma_{\bar{X}}$. La valeur de $z_{\alpha/2}$ se lit dans la table de Gauss en fonction de la probabilité attribuée au paramètre α. On en déduit donc l'intervalle de confiance de l'estimateur de μ, au seuil de probabilité $1 - \alpha$:

$$\bar{X} - z_{\alpha/2}\sigma_{\bar{X}} \leq \mu \leq \bar{X} + z_{\alpha/2}\sigma_{\bar{X}}.$$

Le raisonnement permettant d'aboutir à cette formule est le suivant.

Étant donné que la moyenne échantillonnale $\bar{X}$ est distribuée selon une loi normale $N(\mu, \sigma_{\bar{X}})$, la variable aléatoire :

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

est distribuée selon la loi normale centrée réduite $N(0, 1)$ (voir paragraphe 9.4.2). Nous avons :

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$

illustré par la figure 11.1.

$$\begin{aligned}
P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) &= P(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \leq z_{\alpha/2}) \\
&= P(-z_{\alpha/2}\sigma_{\bar{X}} \leq \bar{X} - \mu \leq z_{\alpha/2}\sigma_{\bar{X}}) \\
&= P(-\bar{X} - z_{\alpha/2}\sigma_{\bar{X}} \leq -\mu \leq -\bar{X} + z_{\alpha/2}\sigma_{\bar{X}}) \\
&= P(\bar{X} + z_{\alpha/2}\sigma_{\bar{X}} \geq \mu \geq \bar{X} - z_{\alpha/2}\sigma_{\bar{X}}) \\
&= P(\bar{X} - z_{\alpha/2}\sigma_{\bar{X}} \leq \mu \leq \bar{X} + z_{\alpha/2}\sigma_{\bar{X}}).
\end{aligned}$$

Ce dernier résultat donne :

$$P(\bar{X} - z_{\alpha/2}\sigma_{\bar{X}} \leq \mu \leq \bar{X} + z_{\alpha/2}\sigma_{\bar{X}}) = 1 - \alpha.$$

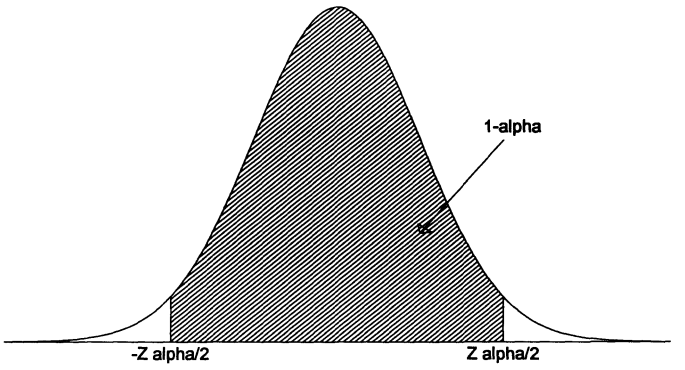

Figure 11.1 : Distribution de Z et intervalle de confiance

Les limites de l'intervalle de confiance pour μ, à un niveau de confiance $1-\alpha$ fixé à l'avance, sont donc :

$$\bar{X} - z_{\alpha/2}\sigma_{\bar{X}} \qquad \text{et} \qquad \bar{X} + z_{\alpha/2}\sigma_{\bar{X}}.$$

Pour un échantillon donné, la variable aléatoire $\bar{X}$ prend la valeur particulière $\bar{x}$ et on a l'intervalle de confiance :

$$\bar{x} - z_{\alpha/2}\sigma_{\bar{X}} \leq \mu \leq \bar{x} + z_{\alpha/2}\sigma_{\bar{X}}$$

où $z_{\alpha/2}$ est la valeur de la variable Z telle que $P(Z \leq z_{\alpha/2}) = 1 - \alpha/2$, et $\sigma_{\bar{X}}$ est l'écart-type de la distribution d'échantillonnage de $\bar{X}$.

Deux situations peuvent se présenter : l'échantillon est tiré soit avec remise soit sans remise. Dans le premier cas, on a :

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \qquad \text{(avec remise)}$$

et dans le deuxième cas :

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \qquad \text{(sans remise)}$$

où N est la taille de la population et n celle de l'échantillon.

Comme nous l'avons déjà vu au paragraphe 10.4, $\bar{X}$ est une variable aléatoire et $\bar{x}$ est une des valeurs de cette variable aléatoire dont les valeurs possibles sont $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$. En pratique, quand on étudie une population quelconque, on ne prend normalement qu'un seul échantillon sur lequel il faut calculer les statistiques nécessaires, à savoir $\bar{X}$ et $\sigma_{\bar{X}}$. C'est à partir de cet échantillon que l'on va tirer des conclusions sur la population. Par conséquent, l'intervalle de confiance de l'estimateur de μ défini sous sa forme générale devient, pour un échantillon donné :

$$\bar{X} - z_{\alpha/2}\sigma_{\bar{X}} \leq \mu \leq \bar{X} + z_{\alpha/2}\sigma_{\bar{X}}.$$

Exemple 11.1 Sur une autoroute contenant 27 postes de ventes d'essence, on a tiré un échantillon aléatoire sans remise de 12 postes différents. Les prix observés en centimes d'un litre d'essence sans plomb sont les suivants :

124	122	125	124
125	124	121	123
125	123	123	123

Supposons que le prix de l'essence sans plomb suive une loi normale d'écart-type $\sigma = 1$ centime, calculons l'intervalle de confiance de l'estimation du prix moyen d'un litre d'essence sans plomb sur l'autoroute.

Dans cet exemple, nous avons :

$$\begin{aligned} N &= 27 \\ n &= 12 \\ \bar{X} &= \bar{x} = 123,5 \\ \sigma &= 1. \end{aligned}$$

L'échantillonnage étant fait sans remise, l'écart-type de l'estimateur $\bar{X}$ est égal à :

$$\begin{aligned} \sigma_{\bar{X}} &= \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \\ &= \frac{1}{\sqrt{12}}\sqrt{\frac{27-12}{27-1}} \\ &= 0,219. \end{aligned}$$

Avec un niveau de confiance $1-\alpha = 95\%$, ceci donne l'intervalle de confiance suivant :

$$\begin{aligned} \bar{X} - z_{\alpha/2} \cdot \sigma_{\bar{X}} &\leq \mu \leq \bar{X} + z_{\alpha/2} \cdot \sigma_{\bar{X}} \\ 123,5 - 1,96 \cdot 0,219 &\leq \mu \leq 123,5 + 1,96 \cdot 0,219 \\ 123,1 &\leq \mu \leq 123,9. \end{aligned}$$

La moyenne du prix de l'essence sans plomb sur l'autoroute en question, y compris celui des stations non-observées, est approximativement entre 123 et 124 centimes par litre.

Si l'échantillon était tiré avec remise, c'est-à-dire si on avait admis la possibilité de retour au même point de vente, le calcul de l'intervalle de confiance de l'estimateur se modifierait comme suit :

$$\begin{aligned} \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} &\leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \\ 123,5 - 1,96 \cdot \frac{1}{\sqrt{12}} &\leq \mu \leq 123,5 + 1,96 \cdot \frac{1}{\sqrt{12}} \\ 122,9 &\leq \mu \leq 124,06. \end{aligned}$$

11.2.2 σ inconnu

Quand l'écart-type σ de la population n'est pas connu, il doit être estimé à partir des informations de l'échantillon. Nous avons vu dans le chapitre précédent que :

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

est un estimateur sans biais de σ^2. Dans l'exemple des prix de l'essence sur l'autoroute, la valeur de S^2 est donc calculée ainsi :

$$\begin{aligned} S^2 &= s^2 = \frac{1}{12-1}[(124^2 + 125^2 + \cdots) - 12 \cdot (123,5)^2] \\ &= 1,5454. \end{aligned}$$

Il s'agit maintenant d'obtenir l'intervalle de confiance de l'estimateur de la moyenne μ de la population en utilisant les valeurs $\bar{X}$ et S^2. Nous cherchons donc la quantité e telle que :

$$\begin{aligned} P(\mid \bar{X} - \mu \mid \leq e) &= 1 - \alpha \\ P\left(\frac{\mid \bar{X} - \mu \mid}{S/\sqrt{n}} \leq e'\right) &= 1 - \alpha \end{aligned}$$

où $e' = e/(S/\sqrt{n})$. Étant donné que le numérateur $\bar{X} - \mu$ et le dénominateur $S/\sqrt{n}$ sont tous deux aléatoires, la loi de probabilité de l'expression $(\bar{X} - \mu)/(S/\sqrt{n})$ est celle d'un ratio de deux variables aléatoires. Cette loi de probabilité n'est plus une distribution normale comme c'était le cas quand la variance σ^2 - le dénominateur du ratio - était fixée (non aléatoire) et connue. Quand X suit une loi normale, le ratio $(\bar{X} - \mu)/(S/\sqrt{n})$ suit une distribution appelée **distribution de Student** ou **distribution t de Student**.

La distribution de Student ressemble à la distribution normale puisque les deux sont symétriques et centrées en zéro. Toutefois, la première est plus plate et dépend de la taille de l'échantillon. Plus la taille de l'échantillon est grande (ou d'une manière équivalente plus le nombre de degrés de liberté augmente), plus la distribution de Student s'approche de la distribution normale (Figure 11.2).

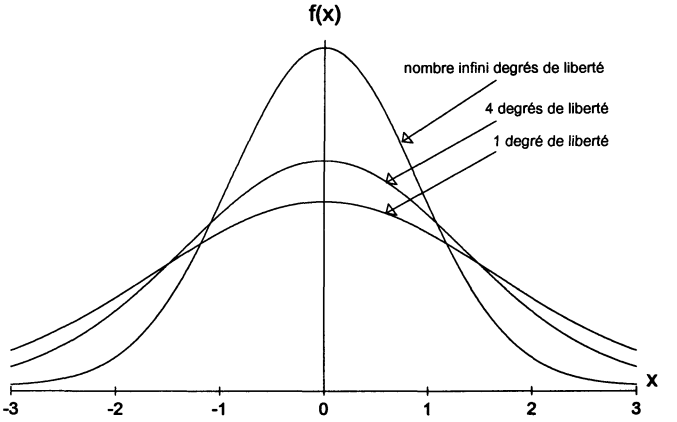

Figure 11.2 : Distribution de Student pour différents degrés de liberté par rapport à la distribution normale

La forme générale de l'intervalle de confiance est la suivante :

$$\bar{X} - t_{(\alpha/2,n-1)} \cdot \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{(\alpha/2,n-1)} \cdot \frac{S}{\sqrt{n}}$$

où $n-1$ représente le degré de liberté, et α représente le seuil de signification de l'intervalle de confiance.

Les calculs pour l'exemple des prix de l'essence sur l'autoroute donnent le résultat suivant :

$$123,5 - t_{(\alpha/2,12-1)} \cdot \frac{\sqrt{1,5454}}{\sqrt{12}} \leq \mu \leq 123,5 + t_{(\alpha/2,12-1)} \cdot \frac{\sqrt{1,5454}}{\sqrt{12}}$$

où $t_{(\alpha/2,12-1)}$ correspond à la valeur t de la distribution de Student pour un niveau de confiance $1-\alpha = 95\%$ et $12-1 = 11$ degrés de liberté. Cette valeur s'obtient à partir de la table des valeurs t de la distribution de Student (voir annexe); elle est égale à $2,201$. Ceci permet de calculer l'intervalle de confiance de l'estimateur de la moyenne μ :

$$\begin{aligned} 123,5 - 2,201 \cdot \frac{\sqrt{1,5454}}{\sqrt{12}} &\leq \mu \leq 123,5 + 2,201 \cdot \frac{\sqrt{1,5454}}{\sqrt{12}} \\ 122,71 &\leq \mu \leq 124,29. \end{aligned}$$

Le prix moyen de l'essence sur l'autoroute est estimé dans l'intervalle allant de 122,71 à 124,29, avec un niveau de confiance de 95%.

Un deuxième exemple de l'utilisation de la distribution de Student est donné ci-dessous.

Exemple 11.2 On dispose de 8 prises de sang recueillies sur une même personne. On obtient pour chaque prise un dosage de cholestérol en grammes de :

246 243 247 248 245 249 242 245

On désire estimer le dosage moyen μ de cholestérol dans le sang de la personne examinée. On construit donc un intervalle de confiance pour l'estimateur de μ avec un niveau de confiance de 95%.

Nous commençons par calculer la moyenne et l'écart-type obtenus sur l'ensemble de l'échantillon :

$$\begin{aligned} \bar{X} &= \frac{\sum x_i}{n} = \frac{1\,965}{8} = 245,625 \\ S^2 &= \frac{\sum (X_i - \bar{X})^2}{n-1} = \frac{39,875}{7} = 5,696 \\ S &= 2,38. \end{aligned}$$

L'erreur-type de la moyenne est égale à :

$$\hat{\sigma}_{\bar{X}} = \frac{S}{\sqrt{n}} = \frac{2,38}{2,83} = 0,84.$$

La valeur du t de Student dans la table pour un seuil de signification de 5% et $7(= 8-1)$ degrés de liberté est 2,365, ce qui nous permet de définir l'intervalle pour μ :

$$\begin{aligned} \bar{X} - t_{(\alpha/2,n-1)} \cdot \hat{\sigma}_{\bar{X}} \leq \mu \leq \quad & \bar{X} + t_{(\alpha/2,n-1)} \cdot \hat{\sigma}_{\bar{X}} \\ 245,625 - 2,365 \cdot 0,84 \leq \mu \leq \quad & 245,625 + 2,365 \cdot 0,84 \\ 243,64 \leq \mu \leq \quad & 247,61. \end{aligned}$$

Le dosage moyen de cholestérol dans le sang de la personne examinée est estimé entre 243,64 et 247,61 grammes avec un niveau de confiance de 95%.

Quand la taille de l'échantillon est assez grande ($n \geq 30$), la distribution de Student s'approche de plus en plus de la distribution normale et les valeurs de $t_{(\alpha/2,n-1)}$ s'approchent des valeurs $z_{\alpha/2}$ correspondantes. Donc, quand n est suffisamment grand, l'intervalle de confiance calculé à partir des valeurs de la distribution normale donne une approximation assez proche de l'intervalle de confiance exact, calculé à partir des valeurs de la distribution de Student.

Exemple 11.3 Sur la base d'un échantillon de 51 objets, on a mesuré une variable X caractérisée par la moyenne :

$$\bar{X} = 12,3$$

et la variance :

$$S^2 = s^2 = 8,9.$$

Supposant que la variable aléatoire X possède une distribution normale de moyenne μ et variance σ^2, le but est d'obtenir l'intervalle de confiance de l'estimation de μ en fonction des résultats de l'échantillon.

La variance étant inconnue, on applique la formule de l'intervalle de confiance selon la distribution de Student :

$$\begin{aligned} 12,3 - t_{(\alpha/2,50)} \cdot \sqrt{\frac{8,9}{51}} &\leq \mu \leq 12,3 + t_{(\alpha/2,50)} \cdot \sqrt{\frac{8,9}{51}} \\ 12,3 - 2,009 \cdot \sqrt{\frac{8,9}{51}} &\leq \mu \leq 12,3 + 2,009 \cdot \sqrt{\frac{8,9}{51}} \\ 11,46 &\leq \mu \leq 13,14. \end{aligned}$$

La taille de l'échantillon étant assez grande ($n = 51$) on aurait pu utiliser la distribution normale au lieu de la distribution de Student et obtenir l'approximation suivante :

$$\begin{aligned} 12,3 - 1,960 \cdot \sqrt{\frac{8,9}{51}} \leq \mu \leq \quad & 12,3 + 1,960 \cdot \sqrt{\frac{8,9}{51}} \\ 11,48 \leq \mu \leq \quad & 13,12. \end{aligned}$$

En comparant cet intervalle et celui obtenu à partir de la distribution de Student, on note que les valeurs sont très proches.

Les choix présentés dans cette section sont résumés ci-dessous :

- **Intervalle de confiance de l'estimation de la moyenne d'une distribution normale**

  1. Variance connue

  $$\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

  2. Variance inconnue

  $$\bar{X} - t_{(\alpha/2,n-1)} \cdot \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{(\alpha/2,n-1)} \cdot \frac{S}{\sqrt{n}}.$$

  Si n est suffisament grand ($n \geq 30$) le résultat ci-dessus peut être approximé par :

  $$\bar{X} - z_{\alpha/2} \cdot \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \cdot \frac{S}{\sqrt{n}}.$$

11.3 Intervalle de confiance pour la moyenne d'une distribution quelconque

Quand la distribution de la variable X n'est pas connue ou lorsqu'elle est connue mais ne suit pas une loi normale, les résultats de la section précédente ne sont pas applicables directement. Toutefois, dans certaines conditions il est quand

même possible de les utiliser pour obtenir un intervalle de confiance approximatif de l'estimation de la moyenne, l'approximation étant d'autant plus rapprochée que le nombre d'observations n (la taille de l'échantillon) est grand et que la distribution est voisine de celle de la loi normale.

• n est grand ($n \geq 30$)

Si l'effectif n de l'échantillon est grand ($n \geq 30$) et si les variables aléatoires $X_1, \cdots, X_n$ sont indépendantes, le ratio :

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

utilisé dans la section précédente pour dériver l'intervalle de confiance, suit approximativement la loi de distribution normale, même si les variables aléatoires $X_1, \cdots, X_n$ elles-mêmes ne suivent pas une distribution normale.

Ceci est le résultat du théorème central limite appliqué à la moyenne échantillonnale $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$. On en déduit que lorsque n est grand, on a approximativement :

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

et

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

On en déduit, quand n est grand, que le ratio :

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim N(0,1)$$

suit approximativement une distribution normale, même si $X_1, \cdots, X_n$ ne suivent pas une distribution normale.

Ce résultat nous permet d'obtenir l'intervalle de confiance approximatif de l'estimateur de μ en utilisant la même procédure de la section précédente quand n est grand.

Exemple 11.4 Dans un test pharmaceutique, on a administré à 64 rats de laboratoire un dosage fixe d'un nouveau produit chimique contre une maladie du sang. Le temps avant que le premier symptôme n'apparaisse au niveau des globules a été mesuré et les résultats obtenus ont été :

$$\begin{aligned} \bar{X} &= 2,13 \text{ minutes} \\ S &= 0,37 \text{ minute.} \end{aligned}$$

Bien que l'analyse des résultats individuels ait montré que la distribution du laps de temps écoulé avant l'apparition d'un symptôme ne suit pas une loi

normale, $n = 64$ étant grand, un intervalle de confiance approximatif de l'estimateur de la moyenne μ de cette durée peut être obtenu à l'aide des résultats de la section précédente, notamment :

$$\begin{aligned} \bar{X} - z_{\alpha/2} \cdot \frac{S}{\sqrt{n}} \leq \mu \leq & \quad \bar{X} + z_{\alpha/2} \cdot \frac{S}{\sqrt{n}} \\ 2,13 - 1,96 \cdot \frac{0,37}{\sqrt{64}} \leq \mu \leq & \quad 2,13 + 1,96 \cdot \frac{0,37}{\sqrt{64}} \\ 2,04 \leq \mu \leq & \quad 2,22. \end{aligned}$$

Cet intervalle de confiance approximatif obtenu, au niveau de confiance de 95%, correspond aux valeurs $\alpha = 5\%$ et $z_{\alpha/2} = 1,96$.

• n n'est pas grand

Si l'effectif n de l'échantillon est restreint, le théorème central limite s'applique mal. Donc l'intervalle de confiance doit s'obtenir directement en fonction de la loi de distribution des $X_1, \cdots, X_n$. Par exemple, si leur distribution est uniforme sur l'intervalle (a, b), on doit chercher la forme de l'intervalle de confiance de l'estimation de la moyenne $\mu = (a+b)/2$ en se basant sur la loi uniforme et non sur la loi normale. Cette démarche est souvent difficile et les formules obtenues compliquées.

En pratique, quand une précision fine n'est pas explicitement demandée, on peut calculer l'intervalle de confiance de l'estimation de la moyenne d'une distribution inconnue (ou connue mais non normale) comme si elle était normale. La fiabilité de cette pratique n'est pas garantie et il se peut que les résultats ainsi obtenus soient très éloignés des résultats théoriques. L'ampleur de cette inexactitude dépend de la taille de l'échantillon et de la forme de la loi théorique de distribution des observation : plus l'échantillon est petit et plus la forme de la loi de distribution est différente de celle de la loi normale, plus l'erreur est considérable.

11.4 Intervalle de confiance pour une proportion

Comme nous l'avons défini au chapitre précédent, nous utiliserons le symbole π pour représenter la proportion d'une population ayant un caractère A défini et le symbole p pour la fraction correspondante dans l'échantillon.

Exemple 11.5 Un sondage effectué sur 300 votants d'une population de 3 000 personnes a montré que 165 personnes avaient l'intention de voter pour l'acceptation du projet soumis au vote. Le pourcentage d'échantillonnage $P = 165/300 = 0,55$ est une estimation de la proportion π de la population.

En général, la valeur p pour un échantillon de taille n peut être considérée comme la moyenne de n variables de Bernoulli, $X_1, X_2, \ldots, X_n$:

$$P = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

où la variable X_i, $i = 1, 2, \ldots, n$, est définie par :

$$X_i = \begin{cases} 1 & \text{si l'observation } i \text{ possède le caractère A} \\ 0 & \text{cas contraire.} \end{cases}$$

La moyenne et la variance de chaque X_i sont exprimées par :

$$\begin{aligned} E(X_i) &= \pi \\ Var(X_i) &= \sigma^2 = \pi(1-\pi). \end{aligned}$$

On en déduit la moyenne et la variance de p pour un échantillon aléatoire simple :

$$\begin{aligned} E(P) &= \frac{E(X_1 + X_2 + \cdots + X_n)}{n} = \pi \\ Var(P) &= \frac{Var(X_1 + X_2 + \cdots + X_n)}{n^2} \\ &= \frac{\sigma^2}{n} = \frac{\pi(1-\pi)}{n}. \end{aligned}$$

Mesurée à partir de l'échantillon, la variance S^2 d'une proportion est égale à $P(1-P)$. Ce qui nous permet de définir l'erreur-type de la distribution d'échantillonnage des pourcentages. Dans le cas de l'exemple 11.5, on obtient :

$$\hat{\sigma}_P^2 = \frac{S^2}{n} = \frac{P(1-P)}{n} = \frac{0,55 \cdot 0,45}{300} = 0,000825$$

et

$$\hat{\sigma}_P = \frac{S}{\sqrt{n}} = \sqrt{0,000825} = 0,0287.$$

La taille de la population étant suffisamment grande, nous n'avons pas tenu compte du facteur correctif.

Le calcul de l'intervalle de confiance de la population π dépend de la taille de l'échantillon. Lorsque la taille de l'échantillon est suffisamment grande, nous pouvons considérer que la distribution d'échantillonnage suit approximativement une loi normale. Nous procédons donc de la même manière que pour l'estimation d'une moyenne :

$$P - z_{\alpha/2}\frac{S}{\sqrt{n}} \leq \pi \leq P + z_{\alpha/2}\frac{S}{\sqrt{n}}.$$

Dans notre exemple et avec un niveau de confiance de 95%, nous obtenons l'intervalle suivant :

$$\begin{aligned} 0,55 - 1,96 \cdot 0,0287 \leq \pi \leq \quad & 0,55 + 1,96 \cdot 0,0287 \\ 0,494 \leq \pi \leq \quad & 0,606. \end{aligned}$$

Lorsque la taille de l'échantillon est petite, l'approximation par la loi normale n'est pas adéquate et l'intervalle de confiance devrait être basé directement sur la distribution théorique des observations. Cette distribution est la loi binômiale et le problème revient à chercher deux valeurs p_1 et p_2 telles que la probabilité d'observer P à l'intérieur de ces deux limites soit égale à $1 - \alpha$:

$$P(p_1 \leq \pi \leq p_2) = 1 - \alpha.$$

La loi binômiale étant une loi discrète, trouver une égalité exacte à $1 - \alpha$ n'est pas possible en général, mais il est toujours possible en revanche d'assurer une probabilité juste un peu plus élevée que le seuil de confiance $1 - \alpha$.

Exprimant P par la fraction X/n où X représente le nombre d'individus dans l'échantillon ayant le caractère A, on obtient :

$$P(np_1 \leq X \leq np_2) = 1 - \alpha.$$

Cette probabilité est assurée si :

(i) $P(X \leq np_1) = \dfrac{\alpha}{2}$ et

(ii) $P(X > np_2) = \dfrac{\alpha}{2}$

où la variable X suit une loi binômiale. On a donc :

$$\text{(i)} \quad P(X \leq np_1) = \sum_{k=0}^{np_1 - 1} \binom{n}{k} \pi^k (1-\pi)^{n-k} = \frac{\alpha}{2}$$

$$\text{(ii)} \quad P(X > np_2) = \sum_{k=np_2+1}^{n} \binom{n}{k} \pi^k (1-\pi)^{n-k} = \frac{\alpha}{2}.$$

Les valeurs de n et α étant fixées d'avance, on considère chacune des expressions (i) et (ii) comme une équation de p_1 ou p_2 en fonction de π. Donc pour chaque valeur de π, on obtient une valeur de p_1 à partir de (i) et une valeur de p_2 à partir de (ii). L'ensemble des valeurs p_1 et p_2 ainsi obtenu peut être répresenté par deux courbes. L'intervalle de confiance de l'estimation de la population π s'obtient en trouvant les valeurs p_1 et p_2, sur l'axe vertical correspondant à la proportion p sur l'axe horizontal du diagramme représentant les deux courbes obtenues dans l'échantillon.

11.5 Historique

Selon A. Desrosières (1988), A. L. Bowley fut l'un des premiers à s'intéresser à la notion d'intervalle de confiance. C'est en 1906 que Bowley présenta à la Royal Statistical Society ses premiers calculs d'intervalle de confiance. Essentiels, dans la théorie des intervalles de confiance, le test de Student et la table de Student ont été développés par W. S. Gosset dit *Student*.

11.6 Exercices

1. Dans un test de fabrication de composantes d'une chaîne Hifi, la baisse de puissance de sortie des circuits éléctriques après 2 000 heures d'utilisation a été mesurée. Un essai sur 80 composantes identiques a donné une baisse de puissance égale à 12 watts. Par ailleurs, il est connu que l'écart-type de la baisse de puissance pour ce type de circuit électrique est $\sigma = 2$ watts.

   (a) Calculer l'intervalle de confiance de l'estimation de la baisse de puissance de la fabrication. Utiliser le niveau de confiance de 95%.

   (b) Recalculer l'intervalle pour un niveau de confiance plus élevé, soit 99%.

   (c) Vérifier que l'intervalle obtenu dans (b) est plus large que celui obtenu dans (a). Expliquer ce fait.

2. Un test similaire à l'exercice 1 a été effectué dans une deuxième usine qui vient d'entrer en fonctionnement. N'ayant pas de données antérieures, il est impossible de fixer une valeur pour l'écart-type σ. Cette valeur doit donc être estimée à partir des résultats du test. Les résultats obtenus sur un échantillon de 70 composantes identiques ont donné :

$$\bar{x} = 14 \text{ watts} \qquad S^2 = 5$$

   (a) Calculer l'intervalle de confiance de l'estimation de la baisse de puissance des composantes de cette nouvelle usine. Utiliser le niveau de confiance de 95%.

   (b) Recalculer (a) avec une valeur de 99%.

3. Le tableau suivant présente un extrait du tableau des valeurs boursières de l'exercice 5 du chapitre 6. Nous avons les valeurs de clôture des 3 et 4 août 1999 de 9 actions parisiennes choisies au hasard parmi les 38 actions qui pourraient constituer un portefeuille :

	3 août	4 août
Accor	216,00	218,70
Alcatel	144,00	144,00
AXA	107,00	107,00
CCF	108,30	108,00
L'Oréal	592,50	579,50
Legrand Ord.	189,90	190,00
Michelin (Action "B")	38,50	39,50
Pinault Printemps Redoute	155,00	151,90
Suez Lyonnaise des Eaux	163,10	162,00

(a) Sur la base du tableau ci-dessus, calculer l'intervalle de confiance avec un degré équivalent à 95% de la valeur moyenne de l'ensemble des actions du portefeuille de 38 actions du 3 août 1999. Exprimer vos hypothèses.

(b) Effectuer le même calcul pour les valeurs boursières en date du 4 août 1999.

(c) Déterminer l'intervalle de confiance avec un niveau de confiance de 95% du changement des valeurs boursières entre le 3 et le 4 août 1999.

4. Douze adultes francophones d'intelligence moyenne ont fait l'objet d'une expérience de mémoire. Le temps pris pour apprendre une liste de 5 verbes allemands a été enregistré pour chaque personne. Ceci a donné les résultats suivants :

5,1	minutes	5,5	minutes	4,5	minutes
4,8	"	5,0	"	5,8	"
6,3	"	5,2	"	5,3	"
5,0	"	4,9	"	5,2	"

(a) Calculer la moyenne et l'écart-type de l'échantillon.

(b) Établir l'intervalle de confiance (α=5%) du temps moyen nécessaire à un francophone pour apprendre la liste des 5 verbes allemands.

(c) On dit qu'un francophone ne peut apprendre qu'un verbe par minute. Est-ce que cette affirmation est justifiée par le résultat obtenu dans (b) ?

5. Un échantillon aléatoire de 100 gravures, pris au hasard dans un grand lot, en contient 15 ayant certaines imperfections.

Calculer l'intervalle de confiance exact de l'estimation de la proportion des gravures défectueuses de ce lot. Utiliser le niveau de confiance de 95%.

JERZY NEYMAN

(1894 - 1981)

Jerzy Neyman est né de parents polonais le 16 avril 1894 à Bendery en Russie. Il entra en 1912 à l'Université de Kharkoo pour y étudier la physique et les mathématiques. Il reçut le titre de Docteur en 1923 à l'Université de Varsovie pour sa thèse portant sur des problèmes probabilistes dans l'expérimentation agricole. En 1937, il fut nommé professeur à l'Université de Berkeley, États-Unis., où il créa un département de statistique.

Neyman, un des plus grands bâtisseurs de la statistique moderne, établit en 1928 avec Egon Pearson (fils de Karl Pearson) les fondements de la théorie des tests d'hypothèses. En 1934, il créa la théorie d'échantillonnage et en 1937 le concept d'intervalle de confiance d'une estimation.