Estimations et intervalles de confiance

Résumé

Cette vignette introduit la notion d'estimateur et ses propriétés : convergence, biais, erreur quadratique, avant d'aborder l'estimation ponctuelle de paramètres de loi : proportion, moyenne, variance. La connaissance des lois de ce estimateurs permet l'estimation par intervalle de confiance et donc de préciser l'incertitude sur ces estimations : intervalle de confiance d'une proportion, d'une moyenne si la variance est connue ou non, d'une variance.
Retour au plan du cours.

1 Introduction

Le cadre est le suivant : on dispose de données observées (en nombre fini) et l'on désire tirer des conclusions de ces données sur l'ensemble de la population. On fait alors une hypothèse raisonnable : il existe une loi de probabilité sous-jacente telle que les "valeurs observables" des différents éléments de la population étudiée puissent être considérées comme des variables aléatoires indépendantes ayant cette loi.

Un aspect important de l'inférence statistique consiste à obtenir des "estimations fiables" des caractéristiques d'une population de grande taille à partir d'un échantillon extrait de cette population. C'est un problème de décision concernant des paramètres qui le plus souvent sont :

- l'espérance mathématique μ ;
- la proportion p ;
- la variance σ^2.

Ces paramètres sont a priori inconnus car la taille réelle de la population étant très grande, il serait trop coûteux de tester tous les éléments de la population. Ainsi, comme un échantillon ne peut donner qu'une information partielle sur la population, les estimations que l'on obtiendra seront inévitablement entachées d'erreurs qu'il s'agit d'évaluer et de minimiser autant que possible.

En résumé, estimer un paramètre inconnu, c'est en donner une valeur approchée à partir des résultats obtenus sur un échantillon aléatoire extrait de la population sous-jacente.

Exemple : Un semencier a récolté 5 tonnes de graines de Tournesol. Il a besoin de connaître le taux de germination de ces graines avant de les mettre en vente. Il extrait un échantillon de 40 graines, les dépose sur un buvard humide et compte le nombre de graines ayant évolué favorablement. On remarque que ce contrôle est de type destructif : l'échantillon ayant servi au contrôle ne peut plus être commercialisé. Il s'agit donc d'évaluer la proportion p des graines de la population à grand effectif, présentant un certain caractère X : succès de la germination. Même avec une population d'effectif restreint, un contrôle destructif impose de faire confiance à un échantillon restreint et la valeur exacte de p ne peut être calculée.

Le modèle s'écrit comme n réalisations x_i de v.a.r. indépendantes de Bernoulli X_i définies par :

$$X_i = \begin{cases} 1 \text{ si l'individu } i \text{ présente le caractère } X \\ 0 \text{ sinon.} \end{cases}$$

Il est naturel d'estimer p par $\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$, qui est la proportion des individus ayant le caractère X dans l'échantillon. En effet, la LGN nous assure de la convergence en probabilité de la v.a.r. $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ vers l'espérance de X_1, c'est-à-dire p ; $\overline{X}$ est l'estimateur de la proportion p et p est estimée par la réalisation $\overline{x}_n$ de $\overline{X}$. Dans l'expérience de germination, 36 graines ont eu une issue favorable avec $x_i = 1$. La proportion estimée est $\overline{x} = 40/36 = 0,9$ C'est une estimation dite *ponctuelle*. D'autre part, dans toute discipline scientifique, il est important d'avoir une indication de la qualité d'un résultat ou encore de l'erreur dont elle peut-être affectée. Ceci se traduit en statistique par la recherche d'un intervalle, dit *intervalle de confiance*, dont on peut assurer, avec un risque d'erreur contrôlé et petit, que cet intervalle contient la "vraie" valeur inconnue du paramètre.

Dans la suite nous nous intéresserons donc à deux types d'estimations :

- soit une estimation donnée par valeur scalaire issue des réalisations des v.a.r. X_i : l'estimation *ponctuelle* ;
- soit une estimation donnée par un ensemble de valeurs appartenant à un intervalle : l'estimation par *intervalle de confiance* contrôlé par un risque d'erreur fixé *a priori*.

2 Estimation ponctuelle

2.1 Estimateur

Convergence

DÉFINITION 1. — *Un n-échantillon aléatoire issu d'une v.a.r. X est un ensemble $(X_1, \ldots, X_n)$ de n v.a.r. indépendantes et de même loi que X.*

Soit θ un paramètre associé à la loi de X, par exemple $\theta = \mathbb{E}(X)$ ou $\theta = \text{Var}(X)$. À partir de l'observation d'un échantillon aléatoire $(X_1, \ldots, X_n)$, on souhaite estimer le paramètre θ.

DÉFINITION 2. — *Un estimateur $\widehat{\theta}_n$ de θ est une fonction qui dépend uniquement du n-échantillon $(X_1, \ldots, X_n)$. Il est dit convergent s'il est "proche" de θ au sens de la convergence en probabilité : pour tout $\epsilon > 0$,*

$$\mathbb{P}\left(|\widehat{\theta}_n - \theta| > \epsilon\right) \underset{n\to+\infty}{\longrightarrow} 0.$$

Dans l'exemple de l'introduction, la quantité $\frac{1}{n}\sum_{i=1}^n X_i$ est un estimateur convergent de p et si, par exemple, on a observé 21 pièces défectueuses sur un lot de 1500 pièces prélevées, l'estimation ponctuelle de p obtenue est $\overline{x}_n = 21/1500 = 1,4\%$.

Pour estimer l'espérance μ des variables aléatoires X_i, on utilise la moyenne empirique

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i,$$

car par la LGN, on sait qu'elle converge en probabilité vers l'espérance $\mu = \mathbb{E}(X_1)$.

Le but de la théorie de l'estimation est de choisir, parmi toutes les statistiques possibles, le "meilleur" estimateur convergent, c'est-à-dire celui qui donnera une estimation ponctuelle la plus proche possible du paramètre et ceci, quel que soit l'échantillon.

Exemple : Considérons une v.a.r. X représentant le nombre de grippes attrapées par une personne en un an. On peut supposer que X suit une loi de Poisson de paramètre $\lambda > 0$. Chercher la loi de X, c'est chercher λ, qui n'est autre que l'espérance mathématique de X. Par conséquent, la LGN nous indique que $\overline{X}_n$ est un estimateur convergent de λ : pour tout $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - \lambda\right| \geq \epsilon\right) \underset{n\to+\infty}{\longrightarrow} 0.$$

Grâce à l'inégalité de Chebychev, on peut démontrer le théorème suivant :

THÉORÈME 3. — *Soit $\widehat{\theta}_n$ un estimateur de θ. Si l'on a :*

$$\lim_{n\to+\infty} \mathbb{E}(\widehat{\theta}_n) = \theta \quad \text{et} \quad \lim_{n\to+\infty} Var(\widehat{\theta}_n) = 0,$$

alors $\widehat{\theta}_n$ est un estimateur convergent de θ.

Biais

DÉFINITION 4. — *Soit $\widehat{\theta}_n$ un estimateur convergent d'un paramètre θ. On appelle biais la quantité $\mathbb{E}(\widehat{\theta}_n) - \theta$. L'estimateur $\widehat{\theta}_n$ est dit sans biais si $\mathbb{E}(\widehat{\theta}_n) = \theta$, et biaisé sinon.*

Exemple : La moyenne empirique $\overline{X}_n$ est un estimateur convergent et sans biais de l'espérance mathématique μ.

Écart quadratique moyen

Notons que l'on a

$$\begin{aligned}
\mathbb{E}\left\{(\widehat{\theta}_n - \theta)^2\right\} &= \mathbb{E}\left\{(\widehat{\theta}_n - \mathbb{E}(\widehat{\theta}_n) + \mathbb{E}(\widehat{\theta}_n) - \theta)^2\right\} \\
&= \mathbb{E}\left\{(\widehat{\theta}_n - \mathbb{E}(\widehat{\theta}_n))^2 + (E(\widehat{\theta}_n) - \theta)^2 + 2(\widehat{\theta}_n - \mathbb{E}(\widehat{\theta}_n))(E(\widehat{\theta}_n) -\right. \\
&= \text{Var}(\widehat{\theta}_n) + (\text{biais})^2,
\end{aligned}$$

car le terme $\mathbb{E}\left\{(\widehat{\theta}_n - \mathbb{E}(\widehat{\theta}_n))(\mathbb{E}(\widehat{\theta}_n) - \theta)\right\}$ est nul. Ainsi, pour rendre l'écart quadratique moyen $\mathbb{E}\left\{(\widehat{\theta}_n - \theta)^2\right\}$ le plus petit possible, il faut que

– $\mathbb{E}(\widehat{\theta}_n) = \theta$, donc choisir un estimateur sans biais,
– la variance $\text{Var}(\widehat{\theta}_n)$ soit faible.

On choisira donc, parmi les estimateurs convergents et sans biais, celui qui a la variance la plus petite. En d'autres termes, si $\widehat{\theta}_n$ est un estimateur convergent et sans biais de θ, on a tout intérêt à ce que $\widehat{\theta}_n$ ne varie pas trop autour de sa moyenne. Cette propriété traduit ce que l'on appelle l'efficacité de l'estimateur.

2.2 Estimateur d'une moyenne ou d'une proportion

On considère un n-échantillon $(X_1, \ldots, X_n)$ issu d'une loi de moyenne μ et de variance σ^2, toutes deux inconnues.

1. d'après la LGN, la moyenne empirique $\overline{X}_n$ est un estimateur convergent de μ.
2. l'estimateur $\overline{X}_n$ est sans biais.
3. par indépendance : $\text{Var}(\overline{X}_n) = \frac{\sigma^2}{n}$.
4. loi de $\overline{X}_n$:
   – si $X \sim \mathcal{N}(\mu, \sigma^2)$, alors $\overline{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$.
   – lorsque n est grand, d'après le TCL, la loi de $\overline{X}_n$ est approchée par une loi normale $\mathcal{N}(\mu, \sigma^2/n)$.

L'estimation d'une proportion p est un cas particulier du précédent, au sens où les v.a.r. X_i considérées sont de Bernoulli de paramètre p.

2.3 Estimateur de la variance

DÉFINITION 5. — *La variance empirique associée à un n-échantillon $(X_1, \ldots, X_n)$ est définie par*

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2.$$

DÉFINITION 6. — *Soit $(Y_1, \ldots, Y_n)$ un n-échantillon de v.a.r. de loi $\mathcal{N}(0,1)$. On appelle loi du chi-deux à n degrés de liberté la loi de la v.a.r. $\sum_{i=1}^{n} Y_i^2$, et on la note $\chi^2_{(n)}$.*

Propriétés de la variance empirique :

1. S_n^2 est un estimateur convergent de la variance σ^2.
2. S_n^2 est sans biais.
3. loi de S_n^2 : pas de résultat général. Cependant, si $X \sim \mathcal{N}(\mu, \sigma^2)$, alors la v.a.r. $\frac{n-1}{\sigma^2} S_n^2$ suit une loi du chi-deux à $n-1$ degrés de libertés $\chi^2(n-1)$.

Remarque : Puisque $\mathbb{E}(Y_i) = 0$, on a $\mathbb{E}(Y_i^2) = \text{Var}(Y_i) = 1$. Si V suit une loi $\chi^2_{(n)}$, alors

$$\mathbb{E}(V) = \mathbb{E}(Y_1^2 + \ldots + Y_n^2) = n.$$

Ainsi on retrouve le fait que S_n^2 est un estimateur convergent et sans biais de σ^2 :

$$\mathbb{E}(S_n^2) = \frac{\sigma^2}{n-1} \mathbb{E}(\chi^2_{n-1}) = \sigma^2.$$

3 Estimation par intervalle de confiance

Pour l'estimation ponctuelle, on considère
– un paramètre inconnu θ,
– un ensemble de valeurs observées $(x_1, \ldots, x_n)$, réalisations d'un n-échantillon aléatoire $(X_1, \ldots, X_n)$, et son estimation ponctuelle $\overline{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$.

Les estimations ponctuelles n'apportent pas d'information sur la précision des résultats, c'est-à-dire qu'elles ne tiennent pas compte des erreurs dues aux fluctuations d'échantillonnage. Pour évaluer la confiance que l'on peut avoir en une valeur, il est nécessaire de déterminer un intervalle contenant, avec une certaine probabilité fixée au préalable, la vraie valeur du paramètre : c'est l'estimation par intervalle de confiance.

3.1 Définition d'un intervalle de confiance

Soit $(X_1, \ldots, X_n)$ un n-échantillon aléatoire et θ un paramètre inconnu de la loi des X_i.

DÉFINITION 7. — *Soit $\alpha \in]0,1[$. S'il existe des v.a.r. $\theta_{min}(X_1, \ldots, X_n)$ et $\theta_{max}(X_1, \ldots, X_n)$ telles que*

$$\mathbb{P}\left(\theta \in [\theta_{min}(X_1, \ldots, X_n), \theta_{max}(X_1, \ldots, X_n)]\right) = 1 - \alpha,$$

on dit alors que $[\theta_{min}(X_1, \ldots, X_n), \theta_{max}(X_1, \ldots, X_n)]$ est un intervalle de confiance pour θ, avec coefficient de sécurité $1-\alpha$. On le note $IC_{1-\alpha}(\theta)$.

Dans la pratique, on peut prendre par exemple $\alpha = 5\%$, ce qui nous donne un IC à 95%. Cela signifie qu'il y a 95% de chance que la valeur inconnue θ soit comprise entre $\theta_{\text{min}}(x_1, \ldots, x_n)$ et $\theta_{\text{max}}(x_1, \ldots, x_n)$.

3.2 Intervalle de confiance pour la moyenne et la variance dans le cas d'un échantillon gaussien

Soit $(X_1, \ldots, X_n)$ un n-échantillon de v.a.r. de loi $\mathcal{N}(\mu, \sigma^2)$.

Estimation de l'espérance μ lorsque la variance σ^2 est connue

Pour estimer μ, on utilise la moyenne empirique $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ qui a pour loi $\mathcal{N}(\mu, \sigma^2/n)$. Il en résulte que

$$\sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sigma}\right) \sim \mathcal{N}(0,1),$$

et que

$$\mathbb{P}\left(-z_{1-\alpha/2} \leq \sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sigma}\right) \leq z_{1-\alpha/2}\right) = 1-\alpha.$$

Ceci équivaut à

$$\mathbb{P}\left(\overline{X}_n - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}_n + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.$$

On obtient donc un IC pour l'espérance μ avec coefficient de sécurité $1-\alpha$ dans le cas où σ est connu : il s'agit de l'intervalle aléatoire

$$\left[\overline{X}_n - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X}_n + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

Ainsi, dans les calculs, l'IC est donné par

$$IC_{1-\alpha}(\mu) = \left[\overline{x}_n - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x}_n + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right],$$

où $\overline{x}_n$ est l'estimation ponctuelle de μ associée à la réalisation du n-échantillon $(X_1, \ldots, X_n)$.

Estimation de l'espérance μ lorsque la variance σ^2 est inconnue

Lorsque la variance σ^2 est inconnue, il est alors nécessaire de remplacer dans les formules précédentes cette quantité par la variance empirique, qui en est un estimateur convergent. Il faut donc considérer non plus la quantité $\sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sigma}\right)$ mais plutôt

$$\sqrt{n}\left(\frac{\overline{X}_n - \mu}{S_n}\right),$$

qui ne suit plus une loi normale mais une loi dite de Student à $n-1$ degrés de liberté, que l'on note $\mathcal{T}_{n-1}$. La densité de la loi de Student est une fonction paire, comme la loi normale $\mathcal{N}(0,1)$. On dispose de tables pour obtenir les quantiles de cette loi. On en déduit donc que

$$\mathbb{P}\left(-t_{1-\alpha/2} \leq \sqrt{n}\left(\frac{\overline{X}_n - \mu}{S_n}\right) \leq t_{1-\alpha/2}\right) = 1-\alpha,$$

ce qui équivaut à

$$\mathbb{P}\left(\overline{X}_n - t_{1-\alpha/2}\frac{S_n}{\sqrt{n}} \leq \mu \leq \overline{X}_n + t_{1-\alpha/2}\frac{S_n}{\sqrt{n}}\right) = 1-\alpha.$$

On obtient donc un IC pour μ avec coefficient de sécurité $1-\alpha$, dans le cas où la variance σ^2 est inconnue : il s'agit de l'intervalle aléatoire

$$\left[\overline{X}_n - t_{1-\alpha/2}\frac{S_n}{\sqrt{n}}, \overline{X}_n + t_{1-\alpha/2}\frac{S_n}{\sqrt{n}}\right].$$

Ainsi, dans les calculs, l'IC est donné par

$$IC_{1-\alpha}(\mu) = \left[\overline{x}_n - t_{1-\alpha/2}\frac{s_n}{\sqrt{n}}, \overline{x}_n + t_{1-\alpha/2}\frac{s_n}{\sqrt{n}}\right],$$

où $\overline{x}_n$ et s_n^2 sont les estimations ponctuelles respectives de la moyenne μ et de la variance σ^2.

Remarque : Si les v.a.r. $X_1, \ldots, X_n$ ne sont pas gaussiennes mais que n est assez grand (en pratique supérieur à 30), alors le TCL nous garantit que la moyenne empirique suit approximativement la loi $\mathcal{N}(\mu, \sigma^2/n)$. Ainsi, dans

le cas où l'on souhaite estimer l'espérance lorsque la variance est connue, l'IC est identique à celui déterminé lorsque les v.a.r. $X_1, \ldots, X_n$ suivent la loi $\mathcal{N}(\mu, \sigma^2)$.

Estimation de la variance σ^2

On estime la variance σ^2, supposée inconnue, par la variance empirique $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. On sait que la v.a.r. S_n^2 a pour loi $\frac{\sigma^2}{n-1}\chi^2_{(n-1)}$ et que

$$\mathbb{E}(S_n^2) = \frac{\sigma^2}{n-1}\mathbb{E}(\chi^2_{(n-1)}) = \sigma^2,$$

c'est-à-dire que S_n^2 est un estimateur sans biais de σ^2. De plus, on lit dans des tables les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ de la loi du $\chi^2_{(n-1)}$, respectivement notés $v_{\alpha/2}$ et $v_{1-\alpha/2}$ (il est normal que les quantiles qui nous intéressent ne soient pas opposés car la densité de cette loi n'est pas paire, à l'inverse de la loi normale centrée réduite). On obtient alors

$$\mathbb{P}\left(v_{\alpha/2} \leq \frac{n-1}{\sigma^2} S_n^2 \leq v_{1-\alpha/2}\right) = 1-\alpha.$$

Ceci équivaut à

$$\mathbb{P}\left(\frac{(n-1)S_n^2}{v_{1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S_n^2}{v_{\alpha/2}}\right) = 1-\alpha.$$

On obtient donc un IC pour la variance σ^2 avec coefficient de sécurité $1-\alpha$: il s'agit de l'intervalle aléatoire

$$\left[\frac{(n-1)S_n^2}{v_{1-\alpha/2}}, \frac{(n-1)S_n^2}{v_{\alpha/2}}\right].$$

Ainsi, dans les calculs, l'IC est donné par

$$IC_{1-\alpha}(\sigma^2) = \left[\frac{(n-1)s_n^2}{v_{1-\alpha/2}}, \frac{(n-1)s_n^2}{v_{\alpha/2}}\right],$$

où s_n^2 est l'estimation ponctuelle de σ^2 associée à la réalisation du n-échantillon $(X_1, \ldots, X_n)$:

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x}_n)^2.$$

3.3 Intervalle de confiance pour la proportion

Revenons à l'exemple introductif : on cherche à estimer la proportion π de graines défectueuses du lot de céréales. On prélève un lot de n graines et on note X_i la v.a.r. qui vaut 1 si la graine i germe, et 0 sinon. On estime π par la moyenne empirique $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Les v.a.r. X_i étant de Bernoulli, on peut alors utiliser l'approximation donnée par le TCL. Soit $Z \sim \mathcal{N}(0,1)$ et $z_{1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$. Par le TCL,

$$\frac{\sum_{i=1}^n X_i - n\pi}{\sqrt{n\pi(1-\pi)}} \xrightarrow{\mathcal{L}} Z \sim \mathcal{N}(0,1).$$

Ceci implique que

$$\mathbb{P}\left(-z_{1-\alpha/2} \leq \frac{\sum_{i=1}^n X_i - n\pi}{\sqrt{n\pi(1-\pi)}} \leq z_{1-\alpha/2}\right) \underset{n\to+\infty}{\longrightarrow} \mathbb{P}(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1-\alpha,$$

c'est-à-dire

$$\mathbb{P}\left(\overline{X}_n - z_{1-\alpha/2}\frac{\sqrt{\pi(1-\pi)}}{\sqrt{n}} \leq \pi \leq \overline{X}_n + z_{1-\alpha/2}\frac{\sqrt{\pi(1-\pi)}}{\sqrt{n}}\right) \underset{n\to+\infty}{\longrightarrow} 1-\alpha.$$

Ceci ne fournit pas un IC pour π car les bornes de l'intervalle dépendent de π. Mais on peut montrer que l'on a le même résultat de convergence, en remplaçant π dans les bornes de l'intervalle par son estimateur convergent $\overline{X}_n$. On obtient alors

$$\mathbb{P}\left(\overline{X}_n - z_{1-\alpha/2}\frac{\sqrt{\overline{X}_n(1-\overline{X}_n)}}{\sqrt{n}} \leq \pi \leq \overline{X}_n + z_{1-\alpha/2}\frac{\sqrt{\overline{X}_n(1-\overline{X}_n)}}{\sqrt{n}}\right) \underset{n\to+\infty}{\longrightarrow} 1-\alpha.$$

On dit que l'intervalle

$$\left[\overline{X}_n - z_{1-\alpha/2}\frac{\sqrt{\overline{X}_n(1-\overline{X}_n)}}{\sqrt{n}}, \overline{X}_n + z_{1-\alpha/2}\frac{\sqrt{\overline{X}_n(1-\overline{X}_n)}}{\sqrt{n}}\right]$$

est un IC asymptotique pour le paramètre π, de coefficient de sécurité $1-\alpha$.

Pour $\alpha = 5\%$, on lit dans les tables $z_{1-\alpha/2} = z_{97,5\%} = 1,96$. Ainsi, le semencier en déduit qu'ayant observé 36 graines germées sur 40, l'intervalle de confiance asymptotique pour π est $[0.807, 0.993]$; il suffit de remplacer dans les calculs la moyenne empirique aléatoire $\overline{X}_n$ par l'estimation ponctuelle $\overline{x}_n = 36/40$).

3.4 Exemple

Une entreprise chimique commercialise un polymère servant à la fabrication de microprocesseurs et stocké dans une cuve dont la caractéristique à contrôler est la viscosité ; celle-ci doit être comprise entre 75 et 95 pour pouvoir commercialiser le polymère. Quatre extractions ont été réalisées dans des zones différentes de la cuve et ont conduit aux valeurs de l'échantillon :$x_1 = 78, x_2 = 85, x_3 = 91, x_4 = 76$, réalisation des variables aléatoires X_1, X_2, X_3, X_4. L'entreprise a besoin d'estimer la viscosité et aussi de connaître la précision de cette estimation. Ayant choisi a priori un seuil de 5%, il s'agit de fournir aux clients des intervalles de confiances à 95% pour μ.

Estimations ponctuelles

- Le *modèle* considère que les variables X_i sont indépendantes selon une loi $\mathcal{N}(\mu, \sigma^2)$; μ représente la moyenne de la viscosité dans la cuve tandis que σ^2 prend en compte la variabilité de la viscosité au sein de la cuve et celle due à l'erreur de mesure.
- Les *paramètres* sont la moyenne μ et la variance σ^2.
- Les *estimateurs* sont $\overline{X}$ de μ et S^2 de σ^2.
- Les estimations ponctuelles sont $\overline{x} = 82.5$ et $s = 6.86$.

Intervalle de confiance de μ avec σ^2 connue

Il est admis que la variabilité du processus de fabrication est constante et connue avec $\sigma = 5$. Dans ce cas, l'estimateur de μ est gaussien, $z_{1-\alpha/2} = 1.96$ et les formules précédentes conduisent à l'estimation de l'intervalle de confiance de μ :

$$[82.5 - 1.96 \times 5/2; 82.5 + 1.96 \times 5/2] \quad = \quad [77.6; 87.4].$$

L'intervalle obtenu est bien à l'intérieur de la spécification ($[75; 95]$).

Intervalle de confiance de μ avec σ^2 estimée

La variance n'est plus supposée constante et connue, elle doit être estimée. L'estimation de l'écart-type est $s = 6.86$. Celui-ci est certes plus important que la valeur théorique précédente mais surtout, l'estimateur de la moyenne μ suit maintenant une loi de Student à $n - 1 = 3$ degrés de liberté. La table de la loi en question fournit le $1 - \alpha/2$-quantile $t_{3;0.975} = 3.182$. L'intervalle de confiance devient alors :

$$[82.5 - 3.182 \times 6.86/2; 82.5 + 3.182 \times 6.86/2] \quad = \quad [71.6; 93.4].$$

L'intervalle n'est pas contenu dans la spécification. Notez l'augmentation sensible de la taille de cet intervalle par le simple fait de devoir estimer la variance plutôt que de la supposer connue ;

Intervalle de confiance de σ^2

L'estimateur de la variance suit une loi du chi-deux à $\nu = (n-1) = 3$ degrés de liberté. Attention, la loi n'est pas symétrique et il faut chercher les deux quantiles à gauche et à droite dans la table ; $\chi^2_{3;0.025} = 0.218$ et $\chi^2_{3;0.975} = 9.35$. Avec $s = 6.86$, l'intervalle de confiance s'écrit :

$$\left[\sqrt{\frac{3 \times 6.86^2}{9.35}}, \sqrt{\frac{3 \times 6.86^2}{0.218}}\right] \quad = \quad [3.9; 25.4].$$

La taille de cet intervalle, souligne le manque de précision de l'estimation de l'écart-type, la taille de l'échantillon y est pour beaucoup.