INSTITUT Mines-Télécom

Statistique pour ingénieur

Thème 2 : Échantillonnage, estimation, estimateurs et intervalles de confiance

F. Delacroix & M. Lecomte, 7 novembre 2016

Introduction

Dans ce second thème nous abordons la statistique inférentielle, et plus particulièrement l'estimation. La section 1 traite des notions de **population**, d'**échantillon** et d'**inférence statistique**, tandis que Les outils autour de la notion d'échantillon sont développés dans la section 2 consacrée à l'**échantillonnage**.

L'**estimation statistique**, qui fait l'objet de la section 3, consistera alors à obtenir des valeurs approchées d'un paramètre inconnu à partir de valeurs observées sur un échantillon. Toutefois, plutôt que de parier sur une valeur d'un paramètre inconnu calculé à partir d'un échantillon, on essaie souvent d'encadrer cette valeur avec une grande probabilité de succès entre des bornes calculées à partir de l'échantillon, en faisant appel à certains modèles probabilistes. Il s'agit alors d'un **intervalle de confiance**, dont la construction fait l'objet de la section 4.

L'échantillonnage et l'estimation ont également un écho très important en entreprise dans le domaine de la maîtrise statistique des procédés. Nous avons choisi d'illustrer ces pratiques mises en œuvre par le biais de normes telles qu'ISO 9000 sous la forme de **cartes de contrôle**, ou **cartes de maîtrise** à la section 5.

Comme on l'a souligné précédemment, la statistique s'appuie sur la théorie des probabilités qui permet de modéliser certains phénomènes aléatoires. Les principales notions de probabilités utiles ont été développées dans le thème 1 de ce MOOC.

Table des matières

Introduction 1

1 Hypothèses, caractéristiques d'un échantillon 2
1.1 Population, échantillons, inférence statistique . . . 2
1.2 Les hypothèses de la statistique classique . . . 4

2 Échantillonnage 5
2.1 Statistiques . . . 5
2.2 Distribution d'échantillonnage des moyennes . . . 7
2.2.1 Cas général . . . 7
2.2.2 Cas où n est suffisamment grand . . . 7
2.2.3 Cas des échantillons gaussiens . . . 8
2.3 Distributions d'échantillonnage des variances . . . 9
2.3.1 Cas général . . . 9
2.3.2 Loi du χ^2 . . . 10
2.3.3 Intervalle de confiance pour la variance . . . 11

3 Estimation **12**
3.1 Estimateurs . . . 12
3.2 Qualités d'un estimateur . . . 14
3.3 Méthode du maximum de vraisemblance . . . 17

4 Intervalles de confiance **20**
4.1 Généralités . . . 20
4.1.1 À partir d'un intervalle de probabilité . . . 20
4.1.2 À l'aide de statistiques . . . 20
4.1.3 Intervalle bilatéral vs intervalle unilatéral . . . 21
4.2 Intervalle de confiance pour l'espérance d'une loi normale . . . 22
4.2.1 Cas où la variance est connue . . . 22
4.2.2 Cas où la variance est inconnue . . . 24
4.3 Intervalle de confiance pour la variance σ^2 d'une loi normale . . . 27
4.4 Intervalle de confiance pour une proportion . . . 28
4.4.1 Utilisation de la loi binomiale . . . 29
4.4.2 Approximation par la loi normale . . . 29

5 Contrôle statistique **32**
5.1 Principe des cartes de contrôle . . . 33
5.2 Carte de contrôle p . . . 33
5.3 Cartes de contrôle aux mesures . . . 34
5.3.1 Limites de contrôle pour la carte de l'écart-type . . . 35
5.3.2 Limites de controle pour la carte de la moyenne . . . 37
5.4 Efficacité des cartes de contrôle . . . 37

Conclusion **39**

Exercices **39**
Exercices sur l'estimation . . . 39
Exercice 1 : Estimateurs . . . 39
Exercice 2 : Estimateur obtenu par la méthode du maximum de vraisemblance 39
Exercice 3 : Paramètre d'une loi de Poisson . . . 40
Exercices sur les intervalles de confiance . . . 41
Exercice 4 : Intervalle de confiance pour une moyenne et une variance . . . 41
Exercice 5 : Intervalle de confiance pour une proportion . . . 41
Exercice 6 : Publicité mensongère ? . . . 41
Exercice 7 : Paramètre d'une loi continue . . . 41

1 Vocabulaire et hypothèses de la statistique, caractéristiques d'un échantillon

1.1 Population, échantillons, inférence statistique

La **population** est l'ensemble des individus sur lesquels porte une étude statistique. On la désigne de façon générale par la lettre Ω (qui, en probabilité, correspondra à l'univers). Pour signifier qu'un individu ω appartient à la population Ω, on écrit : $\omega \in \Omega$.

Dans le cas où la population Ω est finie, nous pouvons écrire

$$\Omega = \{\omega_1,\omega_2,\ldots,\omega_N\}$$

où $N = \operatorname{Card}\Omega$ désigne la taille de la population.

Remarque

En statistique inférentielle, les populations étudiées sont de grande taille, souvent N est de l'ordre de plusieurs milliers, voire plusieurs millions, d'individus.

Nous sommes intéressés par une caractéristique particulière des individus de la population. C'est par exemple, dans l'industrie automobile, le nombre de défauts observés sur les véhicules en sortie de la ligne de production. Le caractère observé est formalisé par une variable X qu'on appelle **variable d'intérêt**. Celle-ci peut être quantitative (discrète ou continue) ou qualitative.

En général, il n'est pas possible de déterminer les valeurs de X pour tous les individus de la population. On réalise alors un **sondage** afin d'obtenir une estimation des paramètres caractérisant la population. Un sondage exige le prélèvement dans la population Ω d'un **échantillon**, défini comme étant un n-uplet d'éléments de Ω. On désigne souvent un tel échantillon par la lettre S (*sample* en anglais). On peut écrire

$$S = (\omega_1,\omega_2,\ldots,\omega_n).$$

Généralement, la taille de l'échantillon est notée n (en minuscule !), afin de la différencier de la taille N (majuscule !) de la population Ω.

La question suivante se pose naturellement :

Question

Comment faire pour sélectionner un «bon» échantillon dans la population ?

Il existe différentes procédures d'échantillonnage, appelées aussi méthodes de sondage. On en distingue deux grandes familles :

- les **méthodes aléatoires** ou **probabilistes**, s'appuyant sur le prélèvement *au hasard* d'individus au sein de la population ;
- les **méthodes empiriques**, ou non aléatoires.

La **méthode des quotas**, largement employée par les instituts de sondage, est la plus connue des méthodes empiriques. Elle consiste à construire un échantillon comme un modèle réduit de la population étudiée selon certains critères : région, sexe, âge, catégorie socio-professionnelle, *etc.*

En statistique, on préfère les méthodes de sondage aléatoires, qui contrairement aux méthodes empiriques permettent de quantifier de manière rigoureuse les estimations faites et de déterminer les erreurs commises.

Parmi les méthodes aléatoires, la plus connue est le **sondage aléatoire simple**, parfois noté PESR (pour «à Probabilités Égales et Sans Remise»),qui consiste à prélever, au hasard et sans remise, n individus au sein de la population de taille N. Chaque individu de la population a la même probabilité que les autres d'être prélevé. Cette probabilité est égale à $\tau = \frac{n}{N}$ et appelée **taux de sondage**[1]. Il en résulte que tous les échantillons de taille n ont la même probabilité d'être sélectionnés.

1. Les lecteurs intéressés par la théorie des sondages pourront consulter l'ouvrage de Pascal ARDILLY : Échantillonnage et méthodes d'enquêtes, Dunod 2004.

Une fois l'échantillon prélevé, on veut étendre les propriétés observées sur celui-ci à l'ensemble de la population. C'est ce qu'on appelle l'**inférence statistique**. Les caractéristiques de l'échantillon telle que sa moyenne, sa variance ou une proportion, peuvent s'étendre à toute la population grâce aux méthodes d'estimation qui seront traitées dans les sections suivantes de ce cours.

Exemple 1

Un industriel commercialise du sucre en paquets, et veut connaître la masse moyenne de sucre contenue dans les paquets. Pour cela, il prélève un échantillon de $n = 20$ paquets et détermine la masse de sucre dans chaque paquet de l'échantillon. Il considère que la moyenne calculée dans l'échantillon est une estimation *de la moyenne relative à toute la population, qui est l'ensemble de tous les paquets produits.*

1.2 Les hypothèses de la statistique classique

Le concept-clé en statistique est la **variabilité**, qui signifie que les caractères peuvent prendre des valeurs différentes : ainsi, un processus industriel ne fournit jamais des caractéristiques parfaitement constantes. En statistique, on modélise les données observées à l'aide de **variables aléatoires**. La théorie des probabilités joue alors un rôle fondamental, d'une part en modélisant certains phénomènes aléatoires, d'autre part en permettant l'étude des caractéristiques observées sur l'échantillon.

Sur une population on définit une variable aléatoire X liée à un caractère observé dans la population, par exemple la masse de sucre dans un paquet dans le cas de l'exemple 1. On supposera que la variable aléatoire X est définie sur un espace probabilisé $(\Omega,\mathcal{T},\mathbb{P})$ où :

— Ω est la population étudiée,
— $\mathcal{T}$ est la tribu des événements,
— $\mathbb{P}$ est une mesure de probabilité sur $(\Omega,\mathcal{T})$.

Dans ces conditions, on peut alors formuler les hypothèses de la statistique «classique».

Définition 1 (*Hypothèses de la statistique classique*)

— *Les valeurs observées $(x_1,\ldots,x_n)$ constituent une réalisation d'un n-uplet, noté $(X_1,\ldots,X_n)$, de variables aléatoires ;*
— *les variables aléatoires X_i sont mutuellement indépendantes et suivent la même loi que X.*

Dans la suite, on admettra que ces hypothèses sont vérifiées quand les échantillons sont prélevés de façon non exhaustive, c'est-à-dire avec remise, ou que la taille de la population est suffisamment importante par rapport à celle de l'échantillon.

Par extension, on appelle aussi échantillon le n-uplet de variables aléatoires $(X_1,\ldots,X_n)$.

Il convient cependant de bien distinguer X_i (la variable aléatoire) de x_i (la valeur prise par la variable aléatoire X_i sur un échantillon S donné). On écrit parfois

$$x_i = X_i(S).$$

Remarque 1

Il est à souligner que, lorsque $n \in \mathbb{N}^*$ est fixé, les variables aléatoires X_i ne sont pas définies sur l'espace probabilisé initial $(\Omega,\mathcal{T},\mathbb{P})$. En effet, X_i désigne la valeur observée

sur le $i^{\text{ème}}$ individu d'un échantillon de taille n. Par conséquent, cette variable aléatoire est définie sur un espace probabilisé correspondant à l'ensemble de tous les échantillons de taille n possibles. Cet ensemble est

$$\Omega^n = \{(\omega_1, \ldots, \omega_n) \text{ tel que } \forall i \in \{1, \ldots, n\}, \quad \omega_i \in \Omega\}.$$

Pour que la construction de ce nouvel espace probabilisé soit complète, il faut munir cet ensemble d'une tribu et d'une mesure de probabilité.

La tribu utilisée est le produit tensoriel $\mathcal{T}^{\otimes n} = \mathcal{T} \otimes \cdots \otimes \mathcal{T}$. *Il s'agit de la plus petite tribu sur Ω^n qui contienne les produits cartésiens d'événements $A_1 \times \cdots \times A_n$. Cette considération, nécessaire pour la cohérence de l'exposé, peut être vue comme théorique et n'aura pas d'incidence sur les questions pratiques posées dans le cadre de ce cours.*

La mesure de probabilité est importante et constitue le reflet de l'hypothèse d'indépendance mentionnée dans la définition 1 : il s'agit là encore d'un produit tensoriel $\mathbb{P}^{\otimes n}$ qui peut se comprendre simplement en disant que la probabilité d'un événement de $(\Omega^n, \mathcal{T}^{\otimes n})$ qui est le produit cartésien d'événements de Ω est le produit des probabilités de ces événements :

$$\mathbb{P}^{\otimes n}(A_1 \times \cdots \times A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2) \cdots \mathbb{P}(A_n).$$

La théorie de l'échantillonnage consiste à étudier les propriétés du n-uplet $(X_1, \ldots, X_n)$ et des caractéristiques le résumant, encore appelées **statistiques**, à partir de la loi supposée connue de la variable parente X.

Un cas particulier important est celui où X suit la loi normale, ou loi de Gauss. On dit alors que l'on est dans le cas d'**échantillons gaussiens**.

2 Échantillonnage

Dans cette section, nous abordons la théorie de l'échantillonnage qui consiste, d'une part, à déterminer un échantillon à partir d'une population donnée, et d'autre part à étudier les caractéristiques de cet échantillon afin d'en déduire des propriétés de la population dont il est issu (inférence statistique).

Exemple 2

On prélève au hasard n ampoules électriques dans une production. On cherche à estimer la durée de vie moyenne des ampoules produites.

Le but de cette partie est d'introduire des modèles permettant de «mathématiser» ces questions et de mettre en place les outils classiques autour de la notion d'échantillon.

2.1 Statistiques

Soit X une variable aléatoire réelle définie sur une population Ω. Si nous prélevons un échantillon $\omega = (\omega_1, \ldots, \omega_n)$ de taille n (où $n \in \mathbb{N}^*$), nous observons n réels $x_1, \ldots, x_n$ qui sont les valeurs que prend X sur chacun des individus de l'échantillon : $X(\omega_i) = x_i$.

D'après les hypothèses de la statistique classique (*cf.* section 1.2), ces nombres sont considérés comme des réalisations de n variables aléatoires $X_1,\ldots,X_n$ i.i.d. : indépendantes et identiquement distribuées, c'est-à-dire suivant la même loi de probabilité. On écrira donc

$$\forall i \in \{1,\ldots,n\}, \quad X_i(\omega) = x_i.$$

Par extension, on appelle ***n*-échantillon** le n-uplet $(X_1,\ldots,X_n)$.

En pratique, on s'intéresse à des caractéristiques simples telles que la moyenne ou la variance de l'échantillon. Celles-ci sont elles-mêmes des réalisations de variables aléatoires réelles issues de $(X_1,\ldots,X_n)$ appelées **statistiques**.

Définition 2

*Une **statistique** T est une variable aléatoire fonction de $X_1,\ldots,X_n$:*

$$T = f(X_1,\ldots,X_n).$$

*La loi de probabilité de la variable aléatoire T s'appelle **distribution d'échantillonnage**.*

En pratique, on s'intéresse souvent à la distribution d'échantillonnage des moyennes et à celles des variances.

Exemple 3

Reprenons l'exemple 2 de la fabrication en série d'ampoules électriques. Soit X la variable aléatoire réelle prenant la valeur 1 si l'ampoule est défectueuse et 0 si l'ampoule est bonne.

On contrôle n ampoules issues de la production. On définit ainsi n variables aléatoires $X_1,X_2,\ldots,X_n$, supposées indépendantes, qui suivent la loi de Bernoulli de paramètre p. On écrit

$$X_i \sim \mathcal{B}(1,p),$$

où p désigne la probabilité qu'une ampoule de la production soit défectueuse. Posons

$$K_n = \sum_{i=1}^{n} X_i.$$

La variable aléatoire K_n désigne donc le nombre d'ampoules défectueuses dans l'échantillon. C'est une statistique, et on sait que sa loi de probabilité est la loi binomiale $\mathcal{B}(n,p)$.

Voici d'autres exemples de statistiques :

- la moyenne empirique : $\overline{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$. Dans le cas de l'exemple 3, on a $\overline{X} = \dfrac{K_n}{n}$,
- la variance empirique : $S^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$,
- le minimum : $\min\limits_{1\leqslant i\leqslant n} X_i$,
- le maximum : $\max\limits_{1\leqslant i\leqslant n} X_i$.

2.2 Distribution d'échantillonnage des moyennes

2.2.1 Cas général

Comme indiqué à la section 2.1, à tout échantillon on associe une suite de variables aléatoires réelles $(X_i)_{i\geqslant 1}$ i.i.d.. On suppose de plus que la variable aléatoire parente X admet une espérance μ et une variance σ^2. On a donc

$$\forall i \geqslant 1, \quad \mathbb{E}(X_i) = \mathbb{E}(X) = \mu \quad \text{et} \quad \mathbb{V}(X_i) = \mathbb{V}(X) = \sigma^2.$$

Définition 3

*La statistique $\overline{X}$, appelée **moyenne empirique de l'échantillon**, est définie par*

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

On a alors le théorème suivant, essentiel en statistique.

Théorème 1

$$\mathbb{E}\left(\overline{X}\right) = \mu \quad et \quad \mathbb{V}\left(\overline{X}\right) = \frac{\sigma^2}{n}.$$

Preuve. La démonstration de ce théorème repose simplement sur les propriétés de l'espérance et la variance : l'espérance est linéaire, la variance est quadratique (c'est-à-dire $\mathbb{V}(\lambda Y) = \lambda^2 \mathbb{V}(Y)$) et additive pour des variables indépendantes. On a donc

$$\mathbb{E}\left(\overline{X}\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$
$$\mathbb{V}\left(\overline{X}\right) = \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}.$$

□

Ce résultat montre que l'écart-type de $\overline{X}$ est égal à $\frac{\sigma}{\sqrt{n}}$, plus petit que l'écart-type de X. On constate, comme le laissait prévoir la loi faible des grands nombres, qu'une observation de $\overline{X}$ est en général plus proche de μ qu'une observation de X, et même d'autant plus proche que n est grand.

2.2.2 Cas où n est suffisamment grand

Par «suffisamment grand» on entend en général $n \geqslant 30$.

La loi de probabilité de $\overline{X}$ dépend *a priori* de la loi de X. Le théorème central-limite étudié dans le thème 1 permet d'affirmer que la suite de variables aléatoires (U_n), où

$$U_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}},$$

converge en loi vers $\mathcal{N}(0,1)$. En pratique, cela signifie que, pour n assez grand, la variable aléatoire $\overline{X}$ suit approximativement la loi normale $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ et ce même si la loi de la variable parente n'est pas une loi normale.

Or on sait que, si $U \sim \mathcal{N}(0,1)$, on a

$$\mathbb{P}\left(-1{,}96 \leqslant U \leqslant 1{,}96\right) = 0{,}95.$$

En appliquant ce résultat à la variable

$$U_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}},$$

compte tenu de l'approximation gaussienne donnée par le théorème central-limite, on obtient

$$\mathbb{P}\left(\overline{X} - 1{,}96\frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \overline{X} + 1{,}96\frac{\sigma}{\sqrt{n}}\right) = 0{,}95.$$

On en déduit, si σ est connu et si n est assez grand, un intervalle de confiance aléatoire pour μ au niveau de confiance 95% :

$$IC_{0{,}95}(\mu) = \left[\overline{X} - 1{,}96\frac{\sigma}{\sqrt{n}}; \overline{X} + 1{,}96\frac{\sigma}{\sqrt{n}}\right].$$

Exemple 4

Soit X la longueur des tiges métalliques dans une production en série. On pose $\mathbb{E}(X) = \mu$ et $\mathbb{V}(X) = \sigma^2$ et on suppose que $\sigma = 0{,}1cm$.

Un échantillon de $n = 50$ tiges a donné une moyenne $\overline{x}$ égale à $15cm$. Un intervalle de confiance réel pour μ au niveau de confiance 95% est donc

$$Ic_{0{,}95}(\mu) = \left[15 - \frac{1{,}96 \times 0{,}1}{\sqrt{50}}; 15 + \frac{1{,}96 \times 0{,}1}{\sqrt{50}}\right] = \left[14{,}97; 15{,}03\right].$$

Nous pouvons en conclure que l'intervalle obtenu contient l'espérance μ avec un niveau de confiance de 95%.

2.2.3 Cas des échantillons gaussiens

Supposons que la variable aléatoire parente X suive une loi normale (ou loi de Gauss) d'espérance μ et de variance σ^2. Comme les X_i suivent la même loi, la variable aléatoire $\overline{X}$ est une combinaison linéaire de variables gaussiennes indépendantes, elle suit donc encore une loi normale, d'espérance μ et de variance $\frac{\sigma^2}{n}$.

Dans ce cas, quelle que soit la taille de l'échantillon, la variable aléatoire

$$U = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

suit la loi normale centrée réduite $\mathcal{N}(0,1)$.

Exemple 5

On prélève 25 pièces dans une production industrielle. Une étude préalable a montré que la longueur X des pièces produites suivant une loi normale d'espérance $10mm$ et d'écart-type $2mm$. Entre quelles valeurs a-t-on 90% de chances de trouver le diamètre moyen de ces 25 pièces ?

Compte tenu des données de cet exemple, on a

$$\overline{X} \sim \mathcal{N}\left(10, \frac{2^2}{25}\right) \quad \text{et} \quad U = \frac{\overline{X} - 10}{2/\sqrt{25}} \sim \mathcal{N}(0,1).$$

D'après les tables de la loi normale, on a

$$0{,}90 = \mathbb{P}\left(-1{,}64 \leqslant U \leqslant 1{,}64\right) = \mathbb{P}\left(-1{,}64 \leqslant \frac{\overline{X} - 10}{2/\sqrt{25}} \leqslant 1{,}64\right)$$

d'où l'on tire

$$\mathbb{P}\left(10 - 1{,}64 \times \frac{2}{5} \leqslant \overline{X} \leqslant 10 + 1{,}64 \times \frac{2}{5}\right) = 0{,}90.$$

Ainsi, on a 90% de chances de trouver le diamètre moyen d'un échantillon de 25 pièces entre 9,34mm et 10,66mm.

2.3 Distributions d'échantillonnage des variances

2.3.1 Cas général

Nous avons défini la moyenne empirique de l'échantillon comme la moyenne arithmétique des variables aléatoires X_i. Introduisons de la même façon la variance empirique de l'échantillon, notée S^2.

Définition 4

La statistique

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

s'appelle ***variance empirique de l'échantillon***.

On utilise souvent la formule suivante, qui se démontre simplement en développant la formule de définition de S^2.

Proposition 2

On a

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2.$$

On retrouve ainsi que la variance de l'échantillon est égale à la «moyenne des carrés moins le carré de la moyenne».

Théorème 3

Si $\mathbb{V}(X) = \sigma^2$, *alors*

$$\mathbb{E}\left(S^2\right) = \frac{n-1}{n}\sigma^2.$$

Preuve. Notons $\mu = \mathbb{E}(X)$. D'après la proposition 2 et par linéarité de l'espérance, on a :

$$\mathbb{E}\left(S^2\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(X_i^2\right) - \mathbb{E}\left(\overline{X}^2\right).$$

Or, d'après la formule usuelle de calcul de la variance, on a

$$\mathbb{E}\left(X_i^2\right) = \mathbb{V}\left(X_i\right) + \mathbb{E}\left(X_i\right)^2 = \sigma^2 + \mu^2.$$

De la même façon, d'après le théorème 1 :

$$\mathbb{E}\left(\overline{X}^2\right) = \mathbb{V}\left(\overline{X}\right) + \mathbb{E}\left(\overline{X}\right)^2 = \frac{\sigma^2}{n} + \mu^2.$$

On obtient donc finalement

$$\begin{aligned}\mathbb{E}\left(S^2\right) &= \frac{1}{n}\sum_{i=1}^{n}\left(\sigma^2+\mu^2\right) - \left(\frac{\sigma^2}{n}+\mu^2\right)\\ &= \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 = \frac{n-1}{n}\sigma^2.\end{aligned}$$

□

Remarque 2

On démontre que

$$\mathbb{V}\left(S^2\right) = \frac{n-1}{n^3}\left[(n-1)\mu_4 - (n-3)\sigma^2\right]$$

où $\mu_4 = \mathbb{E}\left((X-\mu)^4\right)$ *est le moment centré d'ordre 4 de* X *(s'il existe).*

À la section 3.1, on déduira du théorème 3 et de la remarque 2 que S^2 est un estimateur biaisé de σ^2.

Le fait que S^2 ait une espérance qui n'est pas égale à S^2 est générateur d'un **biais**, qui peut être corrigé en multipliant S^2 par un facteur correcteur.

Définition 5

On appelle ***variance corrigée de l'échantillon*** *la variable aléatoire* S^{*2} *définie par*

$$S^{*2} = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}X_i^2 - \frac{n}{n-1}\overline{X}^2.$$

D'après la linéarité de l'espérance, on constate alors que $\mathbb{E}\left(S^{*2}\right) = \sigma^2$.

2.3.2 Loi du χ^2

Dans le cas où la variable aléatoire parente X suit une loi normale, on peut préciser la loi de probabilité de la variable aléatoire S^2. Avant cela, introduisons la loi du χ^2.

Définition 6

On dit qu'une variable aléatoire Z ***suit la loi du*** χ^2 *(«loi du chi-deux») à* ν *degrés de liberté (où* $\nu > 0$*) si elle admet pour densité de probabilité la fonction* f *suivante :*

$$f(t) = \begin{cases}\dfrac{1}{2^{\nu/2}\Gamma(\nu/2)}\, t^{\frac{\nu}{2}-1}\, e^{-\frac{t}{2}} & \text{si } t > 0\\ 0 & \text{si } t \leqslant 0.\end{cases}$$

Ce fait sera noté $Z \sim \chi^2_\nu$*. Dans ce cas la variable aléatoire* Z *admet une espérance et* $\mathbb{E}\left(Z\right) = \nu$.

Cette loi apparaît fréquemment comme celle d'une somme de carrés de variables aléatoires indépendantes suivant toutes la loi normale centrée réduite. C'est l'objet de la proposition suivante, dont la démonstration consiste en un raisonnement par récurrence et produit de convolution.

Proposition 4

Soient $Y_1, \ldots, Y_n$ des variables aléatoires indépendantes et suivant toutes la loi normale centrée réduite $\mathcal{N}(0,1)$. Alors la variable aléatoire

$$Z = Y_1^2 + \ldots + Y_n^2$$

suit la loi du χ^2 à n degrés de liberté.

La loi du χ^2 est tabulée : si ν est fixé et $p \in [0,1]$, on peut lire dans les tables la valeur χ^2_p telle que $\mathbb{P}\left(Z \leqslant \chi^2_p\right) = p$. Ce nombre χ^2_p est le **fractile d'ordre p de la loi χ^2_ν**.

2.3.3 Intervalle de confiance pour la variance

On suppose à nouveau que la variable aléatoire X suit une loi normale $\mathcal{N}(\mu,\sigma^2)$. On a alors le théorème suivant, qui se démontre à l'aide du Théorème de Cochran.

Théorème 5

Si la variable aléatoire X suit une loi normale de variance σ^2, alors la variable aléatoire

$$Z = \frac{n\,S^2}{\sigma^2} = \frac{(n-1)\,S^{*2}}{\sigma^2}$$

suit la loi du χ^2 à $n-1$ degrés de liberté.

Soit α un réel strictement positif, par exemple $\alpha = 0{,}05 = 5\%$. Le nombre $1-\alpha$ sera appelé **niveau de confiance**.

Comme la variable aléatoire Z suit la loi χ^2_{n-1}, on peut déterminer les deux fractiles $\chi^2_{\alpha/2}$ et $\chi^2_{1-\alpha/2}$ tels que

$$\mathbb{P}\left(Z \leqslant \chi^2_{\alpha/2}\right) = \frac{\alpha}{2} \quad \text{et} \quad \mathbb{P}\left(Z > \chi^2_{1-\alpha/2}\right) = \frac{\alpha}{2}.$$

Alors

$$\mathbb{P}\left(\chi^2_{\alpha/2} \leqslant Z \leqslant \chi^2_{1-\alpha/2}\right) = 1-\alpha.$$

Or on a

$$\chi^2_{\alpha/2} \leqslant Z \leqslant \chi^2_{1-\alpha/2} \iff \chi^2_{\alpha/2} \leqslant \frac{n\,S^2}{\sigma^2} \leqslant \chi^2_{1-\alpha/2} \iff \frac{n\,S^2}{\chi^2_{1-\alpha/2}} \leqslant \sigma^2 \leqslant \frac{n\,S^2}{\chi^2_{\alpha/2}}$$

donc

$$\mathbb{P}\left(\frac{n\,S^2}{\chi^2_{1-\alpha/2}} \leqslant \sigma^2 \leqslant \frac{n\,S^2}{\chi^2_{\alpha/2}}\right) = 1-\alpha.$$

On a donc obtenu un **intervalle de confiance aléatoire pour** σ^2 au niveau de confiance $1-\alpha$:

$$IC_{1-\alpha}(\sigma^2) = \left[\frac{n\,S^2}{\chi^2_{1-\alpha/2}}, \frac{n\,S^2}{\chi^2_{\alpha/2}}\right].$$

3 Estimation

Nous abordons maintenant une problématique importante en statistique : l'estimation. À partir de l'observation d'un échantillon provenant d'une loi inconnue, il s'agit de déterminer des caractéristiques de cette loi.

Exemple 6

Le nombre de défauts observés sur des véhicules en sortie d'une ligne de production suit une loi de Poisson de paramètre λ. À partir d'un échantillon de n observations on veut déterminer une valeur approchée «fiable» de λ.

En pratique, l'estimation consiste à déterminer des valeurs approchées de paramètres inconnus relatifs à une population à l'aide d'échantillons. Les paramètres à estimer sont le plus souvent :

— l'espérance μ,
— une proportion p,
— la variance σ^2,
— un autre paramètre relatif à une loi de probabilité.

On distingue deux types d'estimations :

— l'**estimation ponctuelle** qui consiste à calculer, à partir de l'échantillon, une valeur «fiable» représentant le paramètre inconnu ;
— l'**estimation par intervalle de confiance** qui consiste à construire un intervalle (une «fourchette») contenant le paramètre inconnu avec un niveau de confiance élevé (par exemple 95%).

Mettre en place ces estimations nécessite un modèle mathématique : l'**estimateur**.

3.1 Estimateurs

Soit X une variable aléatoire définie sur une population Ω et suivant une certaine loi dont on cherche à estimer un paramètre θ. On note $X_1, \ldots, X_n$ un échantillon de X. On rappelle qu'il s'agit de variables aléatoires i.i.d. (indépendantes et suivant la même loi que X).

Exemple 7

Si le paramètre à estimer est $\mu = \mathbb{E}(X_i)$, il est naturel de s'appuyer sur la moyenne empirique de l'échantillon

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

pour obtenir une estimation de μ.

D'après la loi faible des grands nombres, on sait que $\overline{X}$ converge en probabilité vers μ, c'est-à-dire

$$\forall \varepsilon > 0, \quad \mathbb{P}\left(\left|\overline{X} - \mu\right| \geqslant \varepsilon\right) \xrightarrow[n\to+\infty]{} 0.$$

On peut donc penser que, pour des grandes valeurs de n, la variable aléatoire $\overline{X}$ prendra très probablement des valeurs «proches» de μ. On dit que $\overline{X}$ est un **estimateur de** μ.

Donnons la définition générale d'un estimateur. Soit θ un paramètre de la loi de X, et $X_1, \ldots, X_n$ un échantillon de X.

Définition 7

Une suite de variables aléatoires $(\widehat{\Theta}_n)_{n\geqslant 1}$ est un ***estimateur de*** θ *si les deux conditions suivantes sont remplies.*

(1) Pour tout $n \geqslant 1$, la variable aléatoire $\widehat{\Theta}_n$ est une fonction de $X_1, \ldots, X_n$:

$$\widehat{\Theta}_n = f_n(X_1, \ldots, X_n) \quad ;$$

(2) la suite (Θ_n) converge en probabilité vers θ : $\Theta_n \xrightarrow{\mathbb{P}} \theta$, c'est-à-dire

$$\forall \varepsilon > 0, \quad \mathbb{P}\left(\left|\widehat{\Theta}_n - \theta\right| \geqslant \varepsilon\right) \xrightarrow[n\to+\infty]{} 0.$$

En toute rigueur on devrait parler d'**estimateur faiblement consistant de** θ pour indiquer que la convergence est une convergence en probablité vers θ. De plus, en pratique, on «oublie» qu'un estimateur est une suite de variables aléatoires et on dit simplement que $\widehat{\Theta}_n$ est un estimateur de θ.

Il est souvent difficile de prouver directement la convergence en probabilité, c'est pourquoi on utilise en général la condition suffisante donnée par le théorème 6 suivant.

Théorème 6

Si $\widehat{\Theta}_n$, fonction de $X_1, \ldots, X_n$ est tel que

$$\mathbb{E}\left(\widehat{\Theta}_n\right) \xrightarrow[n\to+\infty]{} \theta \quad \text{et} \quad \mathbb{V}\left(\widehat{\Theta}_n\right) \xrightarrow[n\to+\infty]{} 0$$

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

La démonstration de ce lemme repose sur l'inégalité de Markov.

Lemme 7 (*Inégalité de Markov*)

Soit X une variable aléatoire à valeurs positives admettant une espérance. Alors

$$\forall a > 0, \quad \mathbb{P}(X \geqslant a) \leqslant \frac{\mathbb{E}(X)}{a}.$$

Preuve. Soit $a > 0$. On définit une nouvelle variable aléatoire Y par :

$$Y = \begin{cases} 1 & \text{si } X \geqslant a \\ 0 & \text{si } X < a. \end{cases}$$

Autrement dit, Y est la variable aléatoire indicatrice de l'événement $\{X \geqslant a\}$, et suit donc la loi de Bernoulli de paramètre $p = \mathbb{P}(Y = 1) = \mathbb{P}(X \geqslant a)$.

Observons que par définition de Y et puisque $X \geqslant 0$, on a toujours

$$Y \leqslant \frac{X}{a}$$

donc

$$\mathbb{P}(X \geqslant a) = p = \mathbb{E}(Y) \leqslant \mathbb{E}\left(\frac{X}{a}\right) = \frac{\mathbb{E}(X)}{a}$$

□

Preuve [du théorème 6]. Soit $\varepsilon > 0$. On a, d'après l'inégalité de Markov appliquée à la variable aléatoire $\left[\widehat{\Theta}_n - \theta\right]^2$:

$$\mathbb{P}\left(\left|\widehat{\Theta}_n - \theta\right| \geqslant \varepsilon\right) = \mathbb{P}\left(\left[\widehat{\Theta}_n - \theta\right]^2 \geqslant \varepsilon^2\right) \leqslant \frac{\mathbb{E}\left(\left[\widehat{\Theta}_n - \theta\right]^2\right)}{\varepsilon^2}$$

avec

$$\mathbb{E}\left(\left[\widehat{\Theta}_n - \theta\right]^2\right) = \mathbb{V}\left(\widehat{\Theta}_n\right) + \mathbb{E}\left(\widehat{\Theta}_n - \theta\right)^2 = \mathbb{V}\left(\widehat{\Theta}_n\right) + \left[\mathbb{E}\left(\widehat{\Theta}_n\right) - \theta\right]^2.$$

Par hypothèse, ce majorant tend vers 0 lorsque $n \to +\infty$, donc

$$\mathbb{P}\left(\left|\widehat{\Theta}_n - \theta\right| \geqslant \varepsilon\right) \xrightarrow[n\to+\infty]{} 0,$$

c'est-à-dire que $\widehat{\Theta}_n \xrightarrow{\mathbb{P}} \theta$. □

Reprenons l'exemple de la moyenne empirique de l'échantillon.

Exemple 8

La variable aléatoire $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ *vérifie*

$$\mathbb{E}\left(\overline{X}\right) = \mu \quad \textit{et} \quad \mathbb{V}\left(\overline{X}\right) = \frac{\sigma^2}{n}$$

donc remplit les conditions du théorème 6. C'est donc bien un estimateur de μ.

On peut constater sur cet exemple que le théorème 6 généralise la loi faible des grands nombres, qui traite du seul cas de $\overline{X}$.

Exemple 9

Si, dans une production de pièces en série, on note p *la proportion (inconnue) de pièces défectueuses, cette proportion* p *s'interprète comme la probabilité qu'a une pièce choisie au hasard d'être défectueuse.*

Notons K *le nombre de pièces défectueuses dans un échantillon de taille* n*. Alors la variable aléatoire* $F = \frac{K}{n}$ *est un estimateur de* p.

En effet, on sait que la variable aléatoire K suit la loi binomiale $\mathcal{B}(n,p)$ et donc

$$\mathbb{E}(K) = n\,p \quad \text{et} \quad \mathbb{V}(K) = n\,p(1-p).$$

On en déduit que

$$\mathbb{E}(F) = \frac{\mathbb{E}(K)}{n} = p \quad \text{et} \quad \mathbb{V}(F) = \frac{\mathbb{V}(K)}{n^2} = \frac{p(1-p)}{n} \xrightarrow[n\to+\infty]{} 0$$

et le théorème 6 s'applique.

3.2 Qualités d'un estimateur

Le but de la théorie de l'estimation est de déterminer le «meilleur» estimateur d'un paramètre θ. On s'intéresse donc à la précision d'un estimateur.

À n fixé, l'erreur commise en estimant θ par un estimateur $\widehat{\Theta}_n$ est $\widehat{\Theta}_n - \theta$, qui peut se décomposer de la façon suivante :

$$\widehat{\Theta}_n - \theta = \left[\widehat{\Theta}_n - \mathbb{E}\left(\widehat{\Theta}_n\right)\right] + \left[\mathbb{E}\left(\widehat{\Theta}_n\right) - \theta\right].$$

Le premier terme $\left[\widehat{\Theta}_n - \mathbb{E}\left(\widehat{\Theta}_n\right)\right]$ représente les fluctuations de la variable aléatoire $\widehat{\Theta}_n$ «autour» de son espérance, tandis que l'autre terme $\left[\mathbb{E}\left(\widehat{\Theta}_n\right) - \theta\right]$ représente une erreur systématique qu'on appelle le **biais**.

Définition 8

Soit $\widehat{\Theta}_n$ un estimateur de θ.

*(1) On appelle **biais de l'estimateur** $\widehat{\Theta}_n$ la quantité $\mathbb{E}\left(\widehat{\Theta}_n\right) - \theta$.*

*(2) Si pour tout θ on a $\mathbb{E}\left(\widehat{\Theta}_n\right) = \theta$, on dit que $\widehat{\Theta}_n$ **est un estimateur sans biais de** θ.*

(3) Si on a, pour tout θ,

$$\lim_{n\to+\infty} \mathbb{E}\left(\widehat{\Theta}_n\right) = \theta,$$

*on dit que $\widehat{\Theta}_n$ est **asymptotiquement sans biais**.*

Exemple 10

$\overline{X}$ est un estimateur sans biais de μ. On dit parfois que $\overline{X}$ est l'estimateur «classique» de μ.

Exemple 11

On a introduit à la définition 4 la variance empirique de l'échantillon :

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

On a vu au théorème 3 et à la remarque 2 que

$$\mathbb{E}\left(S^2\right) = \frac{n-1}{n}\sigma^2 \quad \text{et} \quad \mathbb{V}\left(S^2\right) \xrightarrow[n\to+\infty]{} 0.$$

Ceci prouve que S^2 est un estimateur biaisé de σ^2.

Ce biais se corrige comme on l'a vu à la définition 5 en définissant la variance corrigée de l'échantillon.

Exemple 12

$$S^{*2} = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

est un estimateur sans biais de σ^2.

En effet, d'après les propriétés de l'espérance et la variance :

$$\mathbb{E}\left(S^{*2}\right) = \frac{n}{n-1}\mathbb{E}\left(S^2\right) = \frac{n}{n-1}\times\frac{n-1}{n}\sigma^2 = \sigma^2 \quad \text{et} \quad \mathbb{V}\left(S^{*2}\right) = \underbrace{\left(\frac{n}{n-1}\right)^2}_{\to 1}\mathbb{V}\left(S^2\right) \to 0.$$

Il est souvent préférable d'utiliser S^{*2} pour estimer σ^2, notamment pour de petits échantillons, même s'il possède une variance légèrement plus élevée que S^2.

Remarque 3

Il serait faux de dire que S^* est un estimateur sans biais de σ. On sait cependant que S^* possède un «léger» biais. Plus précisément,

$$\mathbb{E}\left(S^*\right) = \sigma\sqrt{\frac{2}{n-1}}\frac{\Gamma(n/2)}{\Gamma\left(\frac{n-1}{2}\right)} \xrightarrow[n\to+\infty]{} \sigma$$

lorsque X suit une loi normale de variance σ^2.

La précision d'un estimateur est généralement donnée à l'aide de l'erreur quadratique moyenne.

Définition 9

Soit $\widehat{\Theta}_n$ un estimateur d'un paramètre θ. On appelle ***erreur quadratique moyenne de*** *$\widehat{\Theta}_n$ la quantité*

$$EQM\left(\widehat{\Theta}_n\right) = \mathbb{E}\left(\left[\widehat{\Theta}_n - \theta\right]^2\right).$$

Théorème 8

L'erreur quadratique moyenne d'un estimateur est égale à la somme de sa variance et du carré du biais :

$$EQM\left(\widehat{\Theta}_n\right) = \mathbb{V}\left(\widehat{\Theta}_n\right) + \left[\mathbb{E}\left(\widehat{\Theta}_n\right) - \theta\right]^2.$$

Preuve. Notons le biais $b(\theta) = \mathbb{E}\left(\widehat{\Theta}_n\right) - \theta$. Par définition de l'erreur quadratique moyenne et par linéarité de l'espérance :

$$\begin{aligned}
EQM\left(\widehat{\Theta}_n\right) &= \mathbb{E}\left(\left[\widehat{\Theta}_n - \theta\right]^2\right) = \mathbb{E}\left(\left[\left(\widehat{\Theta}_n - \mathbb{E}\left(\widehat{\Theta}_n\right)\right) + b(\theta)\right]^2\right) \\
&= \mathbb{E}\left(\left[\widehat{\Theta}_n - \mathbb{E}\left(\widehat{\Theta}_n\right)\right]^2 + b(\theta)^2 + 2\,b(\theta)\left[\widehat{\Theta}_n - \mathbb{E}\left(\widehat{\Theta}_n\right)\right]\right) \\
&= \underbrace{\mathbb{E}\left(\left[\widehat{\Theta}_n - \mathbb{E}\left(\widehat{\Theta}_n\right)\right]^2\right)}_{=\mathbb{V}\left(\widehat{\Theta}_n\right)} + b(\theta)^2 + 2\,b(\theta)\,\underbrace{\mathbb{E}\left(\widehat{\Theta}_n - \mathbb{E}\left(\widehat{\Theta}_n\right)\right)}_{=0} \\
&= \mathbb{V}\left(\widehat{\Theta}_n\right) + b(\theta)^2.
\end{aligned}$$

□

Par conséquent, parmi les estimateurs sans biais de θ, les plus précis sont ceux de variance minimale.

De façon générale, on cherche à minimiser l'erreur quadratique moyenne d'un estimateur. Cependant, sous certaines hypothèses, l'inégalité de Cramér-Rao, que nous n'aborderons pas dans ce cours, fournit une borne inférieure à $EQM\left(\widehat{\Theta}_n\right)$. Cette inégalité est développée dans de nombreux ouvrages de statistique.

En pratique, on se contente souvent de rechercher un estimateur sans biais de variance minimale. Toutefois, dans certains cas particuliers, on peut trouver des estimateurs biaisés plus précis que le meilleur estimateur sans biais.

3.3 Méthode du maximum de vraisemblance

Introduite en 1912 par Fisher, la méthode du maximum de vraisemblance permet en général d'obtenir de «bons» estimateurs. Intuitivement, elle consiste à choisir comme estimateur la valeur qui maximise la probabilité d'avoir obtenu l'échantillon observé.

Pour cela, introduisons la fonction de vraisemblance, généralement notée L [2].

Considérons une variable aléatoire X définie sur Ω, et $x = (x_1, \ldots, x_n)$ des observations issues d'un échantillon $(X_1, \ldots, X_n)$.

Dans un premier temps, supposons que X suive une loi discrète dépendant d'un paramètre θ que l'on souhaite estimer et posons, pour tout $t \in X(\Omega)$,

$$f(t,\theta) = P(X = t).$$

Par indépendance et équidistribution des variables aléatoires X_i constituant l'échantillon, on a :

$$\mathbb{P}\left(\{X_1 = x_1\} \cap \{X_2 = x_2\} \cap \cdots \cap \{X_n = x_n\}\right) = f(x_1,\theta)\, f(x_2,\theta) \cdots f(x_n,\theta).$$

On définit alors la **fonction de vraisemblance**, notée L, par

$$L(x,\theta) = \prod_{i=1}^{n} f(x_i,\theta)$$

Supposons maintenant que X admette une densité de probabilité $t \longmapsto f(t,\theta)$ continue et dépendant d'un paramètre θ que l'on souhaite estimer. On ne peut plus procéder comme dans le cas discret puisqu'ici

$$\mathbb{P}\left(\{X_1 = x_1\} \cap \{X_2 = x_2\} \cap \cdots \cap \{X_n = x_n\}\right) = 0.$$

Pour $\varepsilon > 0$, considérons donc l'événement

$$D_\varepsilon = \{X_1 \in [x_1 - \varepsilon, x_1 + \varepsilon]\} \cap \cdots \cap \{X_n \in [x_n - \varepsilon, x_n + \varepsilon]\}.$$

On a, par indépendance des variables X_i :

$$\mathbb{P}(D_\varepsilon) = \prod_{i=1}^{n} \mathbb{P}\left(X_i \in [x_i - \varepsilon, x_i + \varepsilon]\right) = \prod_{i=1}^{n} \left[\int_{x_i-\varepsilon}^{x_i+\varepsilon} f(t,\theta)\, dt\right]$$

Par continuité de la densité, on a

$$\frac{1}{2\varepsilon} \int_{x_i-\varepsilon}^{x_i+\varepsilon} f(t,\theta)\, dt \xrightarrow[\varepsilon\to 0]{} f(x_i)$$

d'où

$$\mathbb{P}(D_\varepsilon) \underset{\varepsilon\to 0}{\sim} (2\varepsilon)^n f(x_1,\theta) \cdots f(x_n,\theta).$$

2. «Likelihood» en anglais

Cet équivalent, pour ε fixé assez petit, nous conduit donc à définir la fonction de vraisemblance comme

$$L(x,\theta) = \prod_{i=1}^{n} f(x_i,\theta).$$

En réalité, on utilise cette expression de la fonction de vraisemblance même lorsque la densité de probabilité présente des discontinuités (par exemple pour la loi uniforme ou la loi exponentielle).

Définition 10

Soit $x = (x_1,\ldots,x_n)$ des observations issues d'un échantillon $(X_1,\ldots,X_n)$. On suppose que la loi de probabilité des variables i.i.d. X_i est discrète ou continue, connue et dépend d'un paramètre θ à estimer. On définit la ***fonction de vraisemblance*** *L de la façon suivante :*

(1) Si X est discrète,

$$L(x,\theta) = \prod_{i=1}^{n} P(X_i = x_i)$$

(2) Si X est une variable aléatoire continue de densité de probabilité $t \longmapsto f(t,\theta)$:

$$L(x,\theta) = \prod_{i=1}^{n} f(x_i,\theta).$$

Lorsque la fonction $\theta \longmapsto L(x,\theta)$ admet un unique maximum atteint en une valeur

$$\widehat{\theta} = g_n(x_1,\ldots,x_n)$$

on utilise cette valeur $\widehat{\theta}$ pour construire un estimateur de θ, en posant

$$\widehat{\Theta}_n = g_n(X_1,\ldots,X_n).$$

On dit alors que $\widehat{\Theta}_n$ est l'estimateur de θ obtenu par la méthode du maximum de vraisemblance.

En pratique, la recherche d'un tel maximum se fait en dérivant par rapport à θ. Étant donnée l'expression de $L(x,\theta)$ sous forme d'un produit, il est souvent plus commode de passer au logarithme et donc de chercher à maximiser $\ln L(x,\theta)$. Le logarithme étant une fonction strictement croissante, L est maximal si et seulement si $\ln L$ l'est.

Exemple 13

Que donne la méthode du maximum de vraisemblance pour le paramètre λ d'une loi de Poisson ?

La loi de Poisson étant une loi discrète, la fonction de vraisemblance s'écrit

$$L(x_1,\ldots,x_n,\lambda) = \prod_{i=1}^{n} P(X = x_i) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda}\prod_{i=1}^{n}\frac{\lambda^{x_i}}{x_i!}$$

Passons au logarithme :

$$\ln L(x_1,\ldots,x_n,\lambda) = -n\lambda + \sum_{i=1}^{n}\ln\left(\frac{\lambda^{x_i}}{x_i!}\right) = -n\lambda + (\ln\lambda)\sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\ln(x_i!).$$

Dérivons par rapport à λ :

$$\frac{\partial L}{\partial \lambda}(x_1,\ldots,x_n,\lambda) = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i.$$

Le maximum obtenu pour cette fonction est

$$\widehat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

et donne donc pour estimateur de λ la variable aléatoire $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.

Ce résultat n'est guère surprenant puisque, pour une variable aléatoire X suivant la loi de Poisson de paramètre λ, on a $\mathbb{E}(X) = \lambda$: on retrouve donc l'estimateur classique de l'espérance.

Exemple 14

Que donne la méthode du maximum de vraisemblance pour l'écart-type d'une loi de Gauss d'espérance μ supposée connue ?

Ici il s'agit d'une variable aléatoire continue dont la densité de probabilité est la fonction $t \longmapsto f(t,\sigma)$ définie par

$$f(t,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)$$

et donc la fonction de vraisemblance vaut

$$L(x_1,\ldots,x_n,\sigma) = \prod_{i=1}^{n}\left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right] = (2\pi)^{-n/2}\sigma^{-n}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right].$$

Passons au logarithme :

$$\ln\left[L(x_1,\ldots,x_n,\sigma)\right] = -\frac{n}{2}\ln(2\pi) - n\ln\sigma - n\frac{v_n}{2\sigma^2}$$

où on a noté $v_n = \dfrac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$. Dérivons maintenant par rapport à σ :

$$\frac{\partial L}{\partial \sigma}(x_1,\ldots,x_n,\sigma) = -\frac{n}{\sigma} - \frac{n\,v_n}{2}\,\frac{-2}{\sigma^3} = \frac{n}{\sigma}\left(\frac{v_n}{\sigma^2} - 1\right).$$

L'unique maximum est donc atteint pour $\widehat{\sigma} = \sqrt{v_n}$. Ainsi l'estimateur de σ obtenu par la méthode du maximum de vraisemblance est

$$\widehat{\Sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2}.$$

On retrouve une formule similaire à la racine carrée de l'estimateur classique S^2 de la variance. Toutefois à la différence de ce dernier, l'espérance μ avait été supposée connue.

4 Intervalles de confiance

Nous avons déjà étudié sur des exemples de la section 2 la construction d'intervalles de confiance. Plutôt que de calculer une estimation ponctuelle du paramètre inconnu, on essaie de l'encadrer avec une forte probabilité entre des bornes calculées à partir de l'échantillon. On parle alors d'estimation par intervalle de confiance.

Exemple 15

Soit p le pourcentage de voix obtenu par un candidat lors d'une élection. Alors que très peu de bulletins ont été dépouillés, on souhaite obtenir une «fourchette» à 95% pour p, c'est-à-dire un intervalle ayant 95% de chances de contenir p.

4.1 Généralités

Soit θ un paramètre inconnu. Pour construire un intervalle de confiance pour θ, considérons un estimateur de θ, noté $\widehat{\Theta}_n$. Choisissons d'emblée un niveau de confiance, noté $1-\alpha$, par exemple $1-\alpha = 0{,}95$. Ici le nombre $\alpha \in [0,1]$ s'appelle le **risque**.

4.1.1 À partir d'un intervalle de probabilité

Étant donnée une valeur θ_0 de θ, supposons que l'on puisse déterminer un intervalle de probabilité de la forme

$$\mathbb{P}\left(t_1(\theta_0) \leqslant \widehat{\Theta}_n \leqslant t_2(\theta_0)\right) = 1-\alpha.$$

Les bornes de cet intervalle dépendent de θ_0 et peuvent être calculées par exemple si l'on connaît la loi de probabilité de $\widehat{\Theta}_n$. En faisant varier θ_0, on obtient ainsi deux fonctions

$$\theta \longmapsto t_1(\theta) \quad \text{et} \quad \theta \longmapsto t_2(\theta)$$

qui représentent les bornes de l'intervalle de probabilité.

On peut traduire graphiquement cette méthode dans un plan où l'on trace les courbes représentatives de ces fonctions, comme illustré à la figure 1. Sur cette représentation, l'intervalle de probabilité correspondant à une valeur θ_0 du paramètre θ se lit sur la verticale.

Si $\widehat{\theta}$ est une valeur prise par l'estimateur $\widehat{\Theta}_n$, on définit deux réels a et b par les relations

$$a = t_2^{-1}(\widehat{\theta}) \quad \text{et} \quad b = t_1^{-1}(\widehat{\theta}),$$

antécédents de $\widehat{\theta}$ respectivement par les fonctions t_2 et t_1. Alors l'intervalle $[a,b]$ est un intervalle de confiance réel pour θ au niveau de confiance $1-\alpha$ obtenu à partir de l'estimation ponctuelle $\widehat{\theta}$.

4.1.2 À l'aide de statistiques

En pratique, on recherche deux statistiques $T_1 = f_1(X_1,\ldots,X_n)$ et $T_2 = f_2(X_1,\ldots,X_n)$ telles que

$$\mathbb{P}\left(T_1 \leqslant \theta \leqslant T_2\right) = 1-\alpha.$$

Dans ce cas, l'intervalle $[T_1,T_2]$ est un intervalle de confiance aléatoire pour θ au risque α (ou au niveau de confiance $1-\alpha$) Il sera noté $IC_{1-\alpha}(\theta)$.

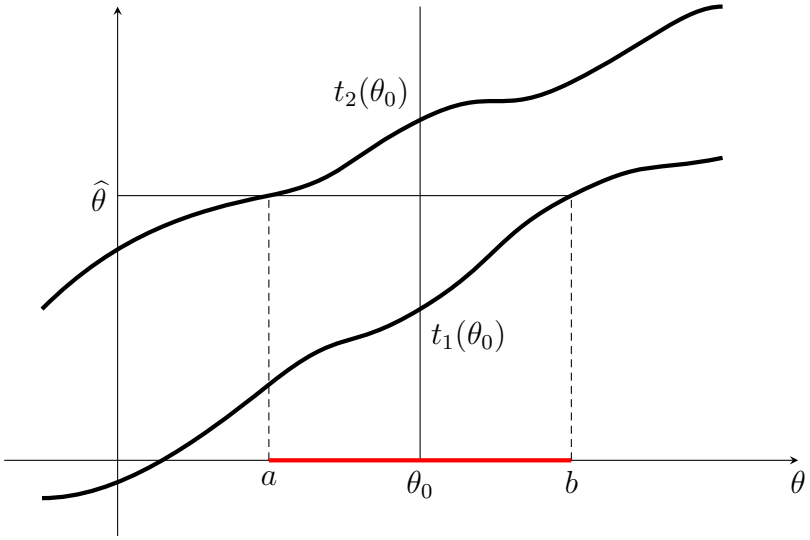

FIGURE 1 – Lecture graphique d'un intervalle de confiance

Supposons qu'à la suite du prélèvement de l'échantillon, la variable aléatoire T_1 prenne la valeur $\widehat{t}_1$ et que T_2 prenne la valeur $\widehat{t}_2$. Il est alors vraisemblable que l'on ait l'encadrement

$$\widehat{t}_1 \leqslant \theta \leqslant \widehat{t}_2.$$

Alors $[\widehat{t}_1,\widehat{t}_2]$ est un intervalle de confiance réel pour θ au risque α (ou au niveau de confiance $1-\alpha$). On le notera $Ic_{1-\alpha}(\theta)$.

4.1.3 Intervalle bilatéral vs intervalle unilatéral

La plupart du temps, les statistiques T_1 et T_2 sont obtenues à partir de la loi de probabilité d'une variable aléatoire Z faisant intervenir un estimateur $\widehat{\Theta}_n$ du paramètre θ. On identifie alors des valeurs de cette variable aléatoire Z comme «très improbables», aux extrémités de l'intervalle des valeurs possibles de Z comme illustré à la figure 2. Les

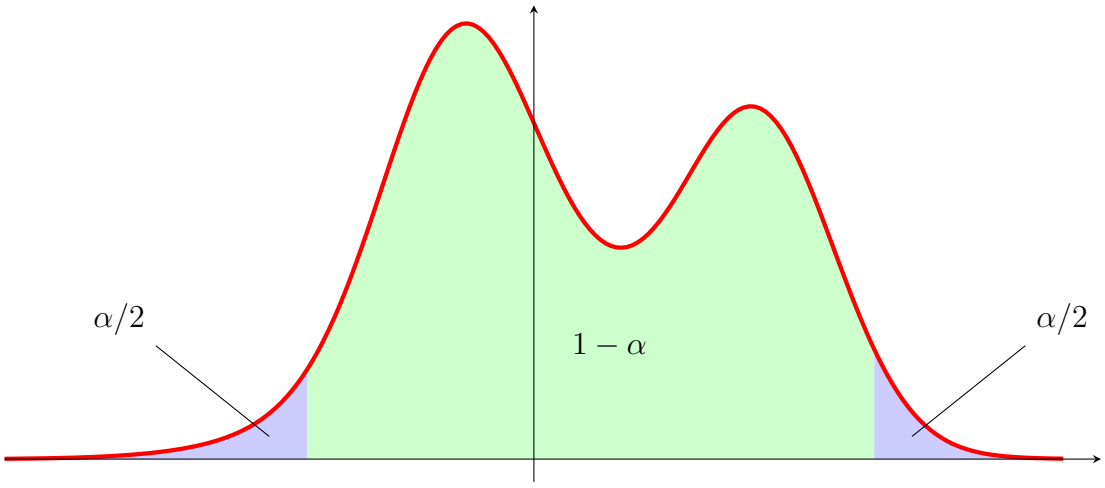

FIGURE 2 – Répartition du risque pour un intervalle de confiance bilatéral

seuils ainsi déterminés permettent alors, en «travaillant» l'encadrement, de déterminer

l'intervalle de confiance. Eu égard à la répartition du risque aux deux extrémités, un tel intervalle de confiance pour θ est qualifié de **bilatéral**

Dans certaines situations bien spécifiques, il peut toutefois être nécessaire de déterminer un intervalle de confiance **unilatéral**, c'est-à-dire un intervalle pour lequel la totalité du risque α est concentrée à l'une des extrémités de l'intervalle des valeurs de Z, comme illustré à la figure 3. La démarche aboutit alors à un intervalle de confiance pour le paramètre θ qui peut être du type $]-\infty,\hat{t}_2]$ ou $[\hat{t}_1,+\infty[$.

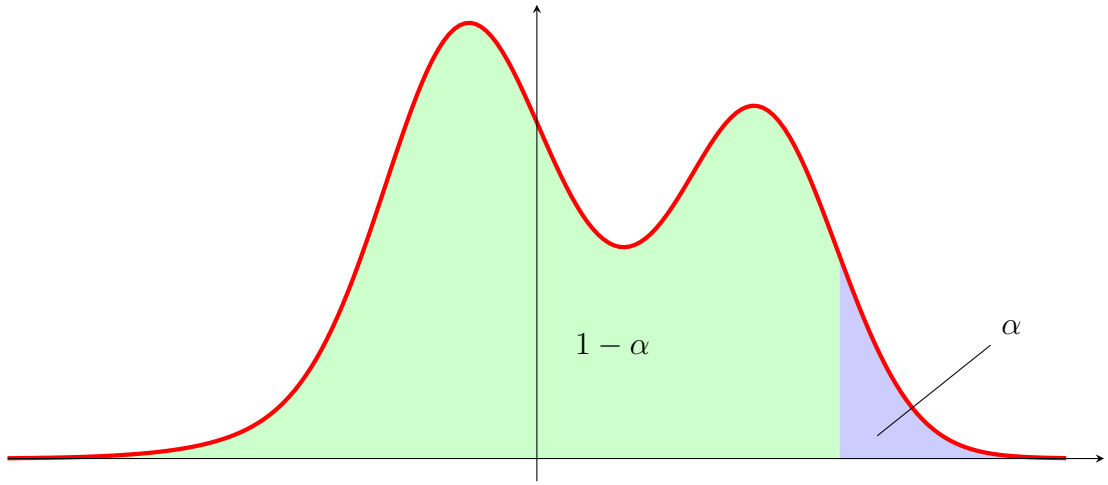

FIGURE 3 – Répartition du risque pour un intervalle de confiance unilatéral

En l'absence de motivation particulière, on choisit la plupart du temps de construire un intervalle bilatéral. Le choix d'un intervalle unilatéral peut se justifier par des considérations telles que la maîtrise d'un risque.

Ainsi, dans le cas de l'exemple 15, si l'idée est de mettre en évidence qu'un candidat à une élection a une forte probabilité d'être élu, on cherchera à construire un intervalle de confiance unilatéral à droite pour la proportion p — inconnue au moment du sondage — de voix qu'il obtiendra.

4.2 Intervalle de confiance pour l'espérance d'une loi normale

Si la variable aléatoire X suit la loi normale d'espérance μ et de variance σ^2, on a vu que la variable aléatoire

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

qui est un estimateur de μ, suit encore une loi normale, cette fois d'espérance μ et de variance $\frac{\sigma^2}{n}$.

4.2.1 Cas où la variance est connue

Supposons que l'écart-type σ soit connu. On sait que la variable aléatoire

$$U = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

suit la loi normale centrée réduite $\mathcal{N}(0,1)$.

À l'aide de tables statistiques, on peut obtenir une valeur $u_{\alpha/2}$ telle que

$$\mathbb{P}\left(U \leqslant u_{\alpha/2}\right) = 1 - \frac{\alpha}{2} \quad \text{ce qui entraîne que} \quad \mathbb{P}\left(-u_{\alpha/2} \leqslant U \leqslant u_{\alpha/2}\right) = 1 - \alpha$$

(*cf.* figure 4). Par exemple, si $1 - \alpha = 0{,}95$ on a $u_{\alpha/2} = 1{,}96$.

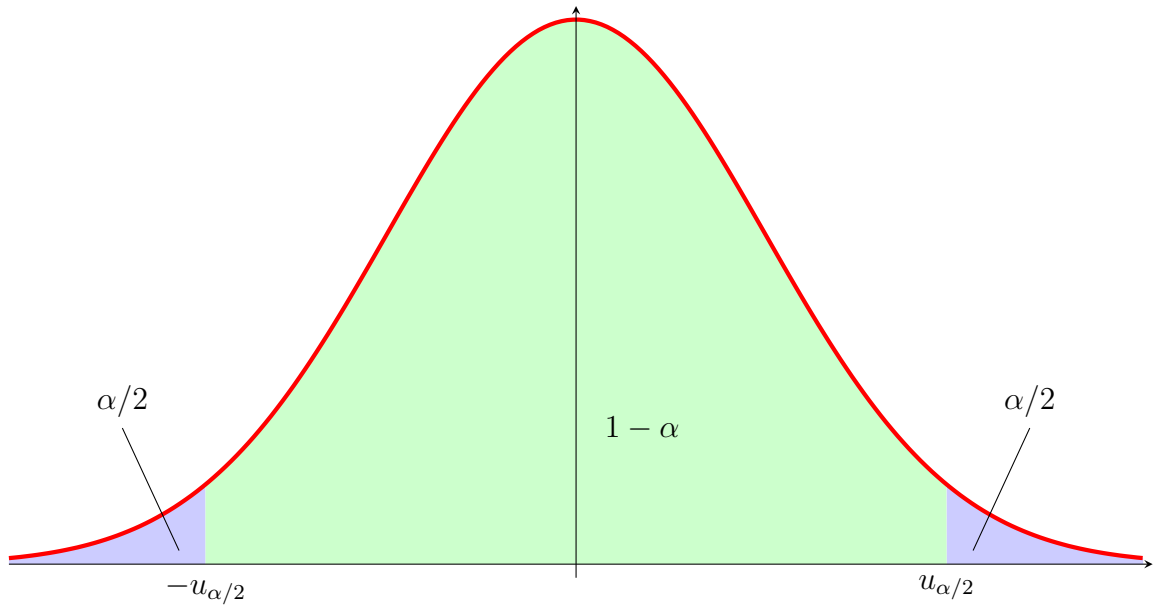

FIGURE 4 – Détermination des seuils pour l'intervalle de confiance avec la loi normale

Or, on a facilement en extrayant μ :

$$-u_{\alpha/2} \leqslant \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leqslant u_{\alpha/2} \iff \overline{X} - u_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \overline{X} + u_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

On en déduit donc un intervalle de confiance aléatoire pour μ au niveau de confiance $1 - \alpha$:

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

Si suite au prélèvement d'un échantillon la variable aléatoire $\overline{X}$ prend la valeur $\overline{x}$, on obtient alors un intervalle de confiance réel pour μ au niveau de confiance $1 - \alpha$:

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

Exemple 16

On a mesuré la capacité, en microfarad, de 25 condensateurs et obtenu une moyenne $\overline{x} = 2{,}09$. On suppose que la capacité d'un condensateur est une variable aléatoire suivant une loi normale d'espérance μ et d'écart-type $\sigma = 0{,}08$. On veut obtenir un intervalle de confiance pour μ au seuil de 1%.

Pour $\alpha = 1\%$, Les tables de la loi normale permettent de déterminer

$$u_{\alpha/2} = u_{0{,}005} = \Phi^{-1}(0{,}995) = 2{,}5758$$

(où Φ désigne la fonction de répartition de $\mathcal{N}(0,1)$).

Les calculs précédents permettent donc de donner un intervalle de confiance réel au niveau de confiance 99% :

$$Ic_{0,99}(\mu) = \left[2{,}09 - \frac{2{,}5758 \times 0{,}08}{\sqrt{25}}; 2{,}09 + \frac{2{,}5758 \times 0{,}08}{\sqrt{25}}\right] = [2.048; 2{,}132].$$

Pour un niveau de confiance de 95%, on aurait eu $u_{\alpha/2} = 1{,}96$ et donc obtenu pour intervalle de confiance

$$Ic_{0,95}(\mu) = \left[2{,}09 - \frac{1{,}96 \times 0{,}08}{\sqrt{25}}; 2{,}09 + \frac{1{,}96 \times 0{,}08}{\sqrt{25}}\right] = [2{,}058; 2{,}122].$$

Il est logique que dans le second cas la longueur de l'intervalle soit plus petite que dans le premier ; en effet la confiance est plus faible donc le risque plus élevé.

4.2.2 Cas où la variance est inconnue

Dans le cas où la variance est inconnue, on utilise la loi de Student, introduite par le statisticien anglais W.S. Gosset (1876-1937). Ce dernier publia en 1908 un article dans lequel il décrivit la fonction de densité de probabilité de la variable aléatoire défini par la différence entre la moyenne d'un échantillon et la moyenne de la population divisée par l'écart-type de l'échantillon.

Fisher proposa en 1912 d'introduire la variable aléatoire $T = \dfrac{U}{\sqrt{Z/\nu}}$ où U suit la loi $\mathcal{N}(0,1)$ et Z la loi du χ^2 à ν degrés de liberté.

La valeur t prise par la variable T est parfois appelée «t de Student».

Définition 11

Une variable aléatoire T ***suit la loi de Student à ν degrés de liberté*** *si elle admet pour densité de probabilité la fonction f définie par*

$$\forall t \in \mathbb{R}, \quad f(t) = \frac{1}{\sqrt{\nu\pi}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}.$$

On notera $\mathcal{T}(\nu)$ la loi de Student à ν degrés de liberté.

Remarques 4

1. *Si $\nu = 1$, il s'agit de la* ***loi de Cauchy****, dont une densité de probabilité est*
$$f(t) = \frac{1}{\pi(1+t^2)}.$$
2. *Pour de grandes valeurs de ν ($\nu > 160$), on peut considérer avec une bonne qualité d'approximation que T suit la loi normale centrée réduite $\mathcal{N}(0,1)$.*

Comme on peut le constater sur la figure 5, les courbes sont symétriques par rapport à l'axe des ordonnées et ressemblent à des courbes Gaussiennes plus «évasées». On admet le théorème suivant.

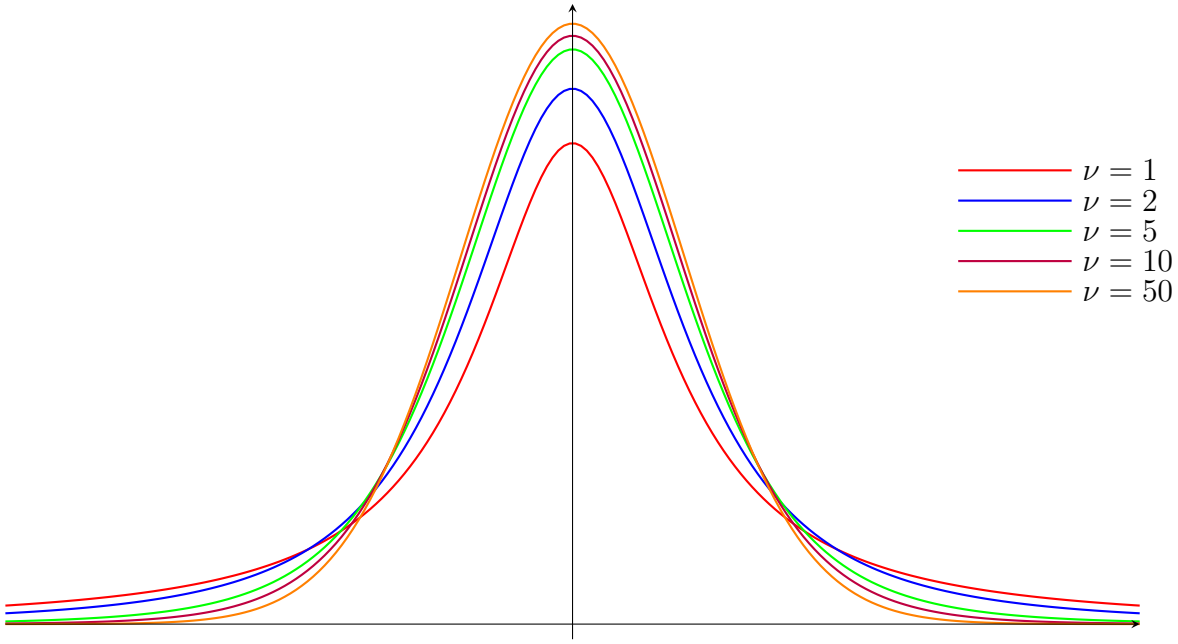

FIGURE 5 – Allure de la densité de la loi de Student

Théorème 9

Soit U une variable aléatoire suivant la loi $\mathcal{N}(0,1)$ et Z une variable aléatoire suivant, indépendamment de U, une loi du χ^2 à ν degrés de liberté (avec $\nu \in \mathbb{N}^*$).

Alors la variable aléatoire $T = \dfrac{U}{\sqrt{Z/\nu}}$ suit la loi de Student à ν degrés de liberté.

Considérons la variable aléatoire $U = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$. On sait que U suit la loi $\mathcal{N}(0,1)$.

Par ailleurs, en notant toujours $S^2 = \dfrac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ la variance empirique de l'échantillon, on a vu au théorème 5 que la variable aléatoire $Z = \dfrac{n\,S^2}{\sigma^2}$ suivait la loi du χ^2 à $\nu = n-1$ degrés de liberté.

Le théorème 9 s'applique donc à la variable aléatoire

$$T = \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{n\,S^2}{\sigma^2(n-1)}}} = \frac{\overline{X} - \mu}{S/\sqrt{n-1}}.$$

On a donc démontré le résultat suivant.

Corollaire 10

Si X est une variable aléatoire qui suit la loi normale $\mathcal{N}(\mu,\sigma^2)$, alors la variable aléatoire

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n-1}}$$

suit la loi de Student $\mathcal{T}(n-1)$.

L'intérêt du corollaire 10 est que la variable aléatoire T ne dépend pas de σ. Cela va

nous permettre de construire un intervalle de confiance pour μ dans le cas où l'écart-type σ est inconnu.

Comme précédemment, on détermine à l'aide d'un logiciel ou de tables statistiques la valeur $t_{\alpha/2}$ telle que

$$\mathbb{P}\left(T \leqslant t_{\alpha/2}\right) = 1 - \frac{\alpha}{2} \quad \text{ce qui entraîne} \quad \mathbb{P}\left(-t_{\alpha/2} \leqslant T \leqslant t_{\alpha/2}\right) = 1 - \alpha.$$

Par exemple, si $\alpha = 5\%$ et $n = 10$, on a $t_{\alpha/2} = 2{,}262$.

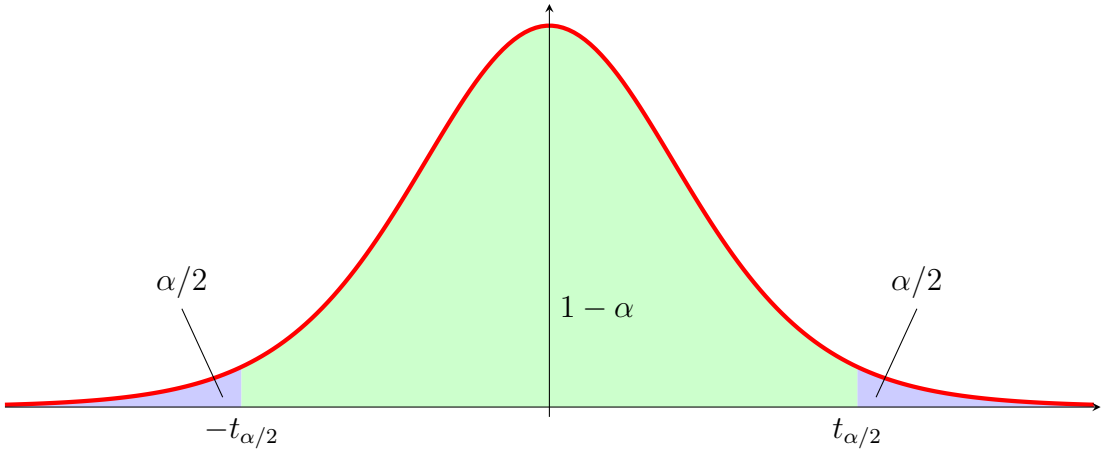

FIGURE 6 – Détermination des seuils pour l'intervalle de confiance avec la loi de Student

De cet encadrement, on peut facilement extraire μ :

$$\begin{aligned} -t_{\alpha/2} \leqslant T \leqslant t_{\alpha/2} &\iff -t_{\alpha/2} \leqslant \frac{\overline{X} - \mu}{S/\sqrt{n-1}} \leqslant t_{\alpha/2} \\ &\iff \overline{X} - t_{\alpha/2}\frac{S}{\sqrt{n-1}} \leqslant \mu \leqslant \overline{X} + t_{\alpha/2}\frac{S}{\sqrt{n-1}}. \end{aligned}$$

On obtient ainsi un intervalle de confiance aléatoire pour μ au niveau de confiance $1 - \alpha$:

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

Suite au prélèvement de l'échantillon, la variable aléatoire $\overline{X}$ prend une valeur $\overline{x}$ et la variable S une valeur s. Alors on obtient l'intervalle de confiance réel pour μ au niveau de confiance $1 - \alpha$:

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

Remarque 5

Si n est assez grand — en pratique $n \geqslant 30$ — le théorème central limite permet d'affirmer que la variable aléatoire $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ suit approximativement la loi normale $\mathcal{N}(0,1)$.

Par conséquent, le corollaire 10 subsiste et les intervalles de confiance demeurent valables même si X ne suit pas une loi normale.

Exemple 17

La masse d'une pièce en cuivre produite en série suit la loi normale d'espérance μ et de variance σ^2 inconnus. Un échantillon de $n = 15$ pièces a donné les résultats suivants (en grammes) :

$$\overline{x} = 10{,}9 \quad et \quad s = 1{,}16.$$

D'après les résultats précédents, la variable aléatoire $T = \frac{\overline{X}-\mu}{S/\sqrt{14}}$ suit la loi de Student $\mathcal{T}(14)$. On détermine à l'aide d'un logiciel ou une table statistique la valeur $t_{\alpha/2}$ telle que

$$\mathbb{P}\left(-t_{\alpha/2} \leqslant T \leqslant t_{\alpha/2}\right) = 1-\alpha \quad \text{avec} \quad \alpha = 10\% \quad t_{\alpha/2} = 1{,}761.$$

On en déduit l'intervalle de confiance réel pour μ au niveau de confiance $1-\alpha = 90\%$:

$$Ic_{0,9}(\mu) = \left[10{,}9 - 1{,}761 \times \frac{1{,}16}{\sqrt{14}}; 10{,}9 + 1{,}761 \times \frac{1{,}16}{\sqrt{14}}\right] = [10{,}35; 11{,}45].$$

4.3 Intervalle de confiance pour la variance σ^2 d'une loi normale

Supposons que la variable aléatoire X suive la loi $\mathcal{N}(\mu,\sigma^2)$, les valeurs de μ et σ étant toutes deux inconnues. On sait que la variance empirique de l'échantillon

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

est un estimateur (biaisé) de σ^2 et on a admis au théorème 5 que la variable aléatoire $Z = \frac{n\,S^2}{\sigma^2}$ suit la loi du χ^2 à $\nu = n-1$ degrés de liberté.

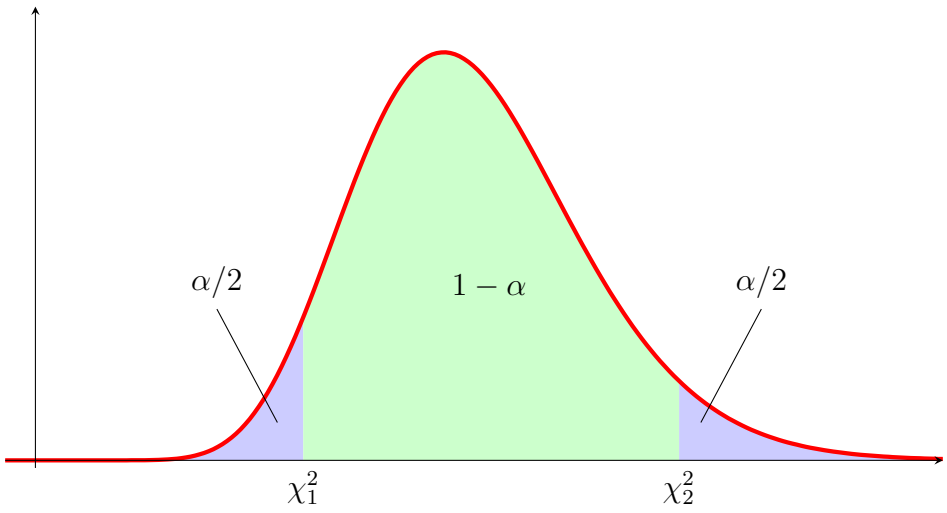

FIGURE 7 – Détermination des seuils pour l'intervalle de confiance avec la loi du χ^2

Pour déterminer un intervalle de confiance au seul α pour σ^2, considérons les réels χ_1^2 et χ_2^2 tels que

$$\mathbb{P}\left(Z \leqslant \chi_1^2\right) = \frac{\alpha}{2} \quad \text{et} \quad \mathbb{P}\left(Z \leqslant \chi_2^2\right) = 1 - \frac{\alpha}{2}.$$

Alors, comme illustré à la figure 7, ce sont les bornes de l'intervalle de probabilité pour Z :

$$\mathbb{P}\left(\chi_1^2 \leqslant \frac{n\,S^2}{\sigma^2} \leqslant \chi_2^2\right) = 1-\alpha.$$

On peut alors extraire un encadrement de σ^2 :

$$\chi_1^2 \leqslant \frac{n\,S^2}{\sigma^2} \leqslant \chi_2^2 \iff \frac{n\,S^2}{\chi_2^2} \leqslant \sigma^2 \leqslant \frac{n\,S^2}{\chi_1^2}.$$

Oa donc trouvé un intervalle de confiance aléatoire pour σ^2 au niveau de confiance $1-\alpha$:

$$IC_{1-\alpha}(\sigma^2) = \left[\frac{n\,S^2}{\chi_2^2}, \frac{n\,S^2}{\chi_1^2}\right].$$

Exemple 18

Dans le cadre d'un concours, un correcteur a corrigé $n = 30$ copies et observé sur cet échantillon une variance $s^2 = 12$ des notes obtenues par les candidats. Comment déterminer un intervalle de confiance pour la variance σ^2 des notes de l'ensemble des copies ?

On admet que les notes suivent une loi normale. Si on se fixe un niveau de confiance $1-\alpha = 90\%$, les tables statistiques permettent de déterminer les valeurs

$$\chi_1^2 = 17{,}71 \quad \text{et} \quad \chi_2^2 = 42{,}56.$$

L'intervalle de confiance réel pour σ^2 obtenu est donc

$$Ic_{0,9}(\sigma^2) = \left[\frac{30\times 12}{42{,}56}; \frac{30\times 12}{17{,}71}\right] = [8{,}45; 20{,}33].$$

On peut aussi en déduire un intervalle de confiance réel pour σ au niveau de confiance 90% :

$$Ic_{0,9}(\sigma) = [2{,}9; 4{,}6].$$

Remarque 6

À cause notamment de l'asymétrie de la densité de χ^2 et contrairement aux intervalles de confiance pour la moyenne obtenus à la section 4.2, l'intervalle de confiance obtenu pour σ^2 n'est pas centré en l'estimation ponctuelle s^2.

4.4 Intervalle de confiance pour une proportion

On considère une population infinie, ou finie à condition que le tirage s'effectue avec remise, dans laquelle une proportion p (inconnue) des individus possède un certain caractère. On souhaite déterminer un intervalle de confiance pour p à partir de la fréquence f observée de ce caractère dans un échantillon de taille n.

Exemple 19

Une entreprise fabrique des cartes électroniques. On s'intéresse à la proportion p de cartes non conformes produites pendant une certaine période.

La proportion p peut s'interpréter comme la probabilité qu'un individu choisi au hasard dans la population ait le caractère étudié.

Étant donné un échantillon de taille n issu de la population étudiée, notons K le nombre d'individus ayant le caractère étudié dans l'échantillon (par exemple une non-conformité). Alors K est une variable aléatoire qui suit la loi binomiale $\mathcal{B}(n,p)$.

Comme on l'a vu à la section 3.1, la variable aléatoire $F = \frac{K}{n}$ est telle que

$$\mathbb{E}(F) = p \quad \text{et} \quad \mathbb{V}(F) = \frac{p(1-p)}{n} \xrightarrow[n\to+\infty]{} 0$$

ce qui montre que F est un estimateur (non biaisé) de p.

La fréquence observée du caractère dans l'échantillon f est ainsi la valeur prise par F, et constitue donc une estimation ponctuelle de p.

Pour obtenir un intervalle de confiance pour p, on procède en général de la façon suivante :

— si n est petit (en pratique $5 \leqslant n \leqslant 100$), on utilise la loi binomiale, et plus précisément les abaques des tables statistiques de cette loi ;

— si n est grand et p (ou f) pas trop petit (en pratique $n \geqslant 100$ et $n\,f(1-f) > 18$), on utilise une approximation par la loi normale.

4.4.1 Utilisation de la loi binomiale

Notons $1-\alpha$ le niveau de confiance souhaité pour l'intervalle de confiance. On va décliner dans le cas de la proportion p la méthode exposée à la section 4.1.1. L'une des différences est qu'ici la loi de probabilité est discrète.

Pour tout réel $p \in]0,1[$, on détermine donc deux nombres entiers $c_1(p)$ et $c_2(p)$ tels que

$$\mathbb{P}(K \leqslant c_1(p)) = \frac{\alpha}{2} \quad \text{et} \quad \mathbb{P}(K \geqslant c_2(p)) = \frac{\alpha}{2},$$

c'est-à-dire

$$\sum_{j=0}^{c_1(p)} \binom{n}{j} p^j (1-p)^{n-j} = \frac{\alpha}{2} \quad \text{et} \quad \sum_{j=c_2(p)}^{n} \binom{n}{j} p^j (1-p)^{n-j} = \frac{\alpha}{2}$$

Il est alors possible de construire le graphe de la figure 8 en portant p en abscisse et $\frac{c_1(p)}{n}$ et $\frac{c_2(p)}{n}$ en ordonnée.

Sur ce graphique, la droite horizontale d'ordonnée $f = \frac{k}{n}$ permet de déterminer l'intervalle de confiance réel $[p_1,p_2]$ au niveau de confiance $1-\alpha$.

Remarque

Il s'agit d'un intervalle de confiance bilatéral, c'est-à-dire que le risque α est symétriquement réparti aux deux extrémités de l'intervalle.

En pratique, on procède souvent par lecture d'abaques de tables statistiques.

4.4.2 Approximation par la loi normale

Si $n > 100$ et $n\,f(1-f) > 18$, on peut affirmer avec une erreur d'approximation acceptable que la variable aléatoire K suit la loi normale d'espérance n,p et de variance

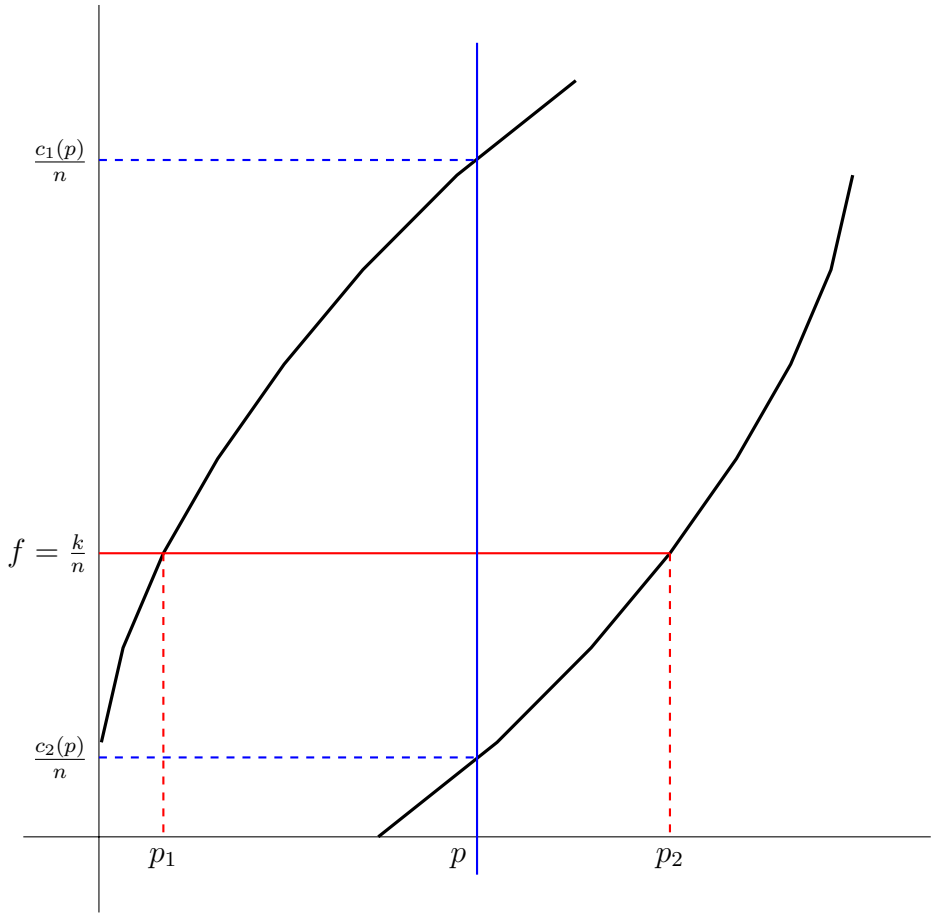

FIGURE 8 – Intervalle de probabilité pour la loi $\mathcal{B}(n,p)$

$n\,p(1-p)$, conséquence du théorème de Moivre-Laplace. Alors, la variable aléatoire $F = \frac{K}{n}$ suit la loi normale d'espérance p et de variance $\frac{p(1-p)}{n}$, et la variable aléatoire

$$U = \frac{F-p}{\sqrt{\frac{p(1-p)}{n}}}$$

suit la loi normale centrée réduite $\mathcal{N}(0,1)$.

Fixons $\alpha \in]0,1[$. On peut alors, à l'aide des tables de la loi normale ou d'un logiciel, déterminer le réel $u_{\alpha/2}$ tel que $\mathbb{P}\left(U \leqslant u_{\alpha/2}\right) = \frac{\alpha}{2}$. On a alors

$$\mathbb{P}\left(-u_{\alpha/2} \leqslant U \leqslant u_{\alpha/2}\right) = 1-\alpha.$$

On peut alors en déduire l'intervalle de probabilité symétrique pour F :

$$p - u_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leqslant F \leqslant p + u_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}.$$

Les bornes de l'intervalle de probabilité ainsi obtenu sont les solutions de l'équation en y suivante :

$$(y-p)^2 = u_{\alpha/2}^2 \frac{p(1-p)}{n}.$$

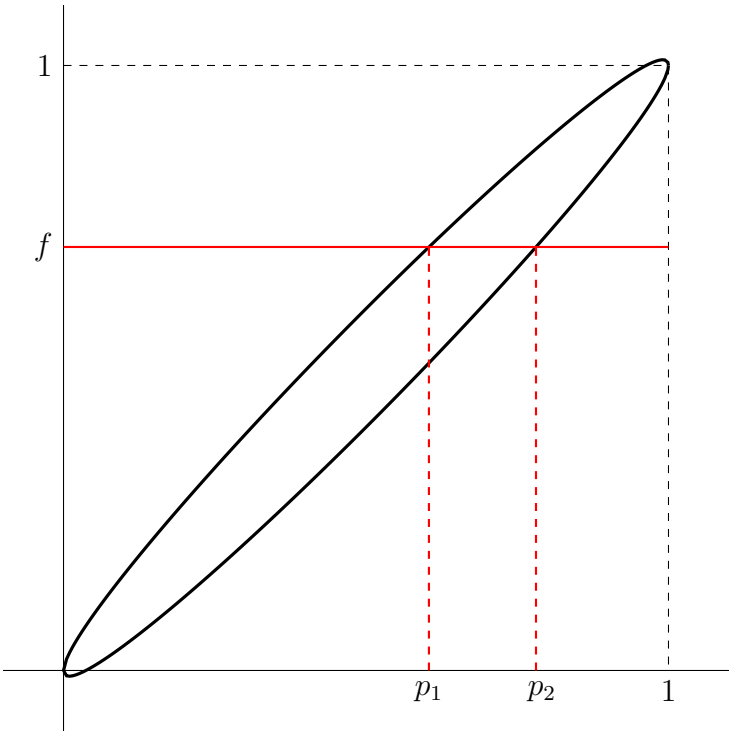

FIGURE 9 – Intervalle de probabilité obtenu par le théorème de Moivre-Laplace

Il s'agit de l'équation d'une ellipse passant par l'origine et par le point de coordonnées (1,1) dans le plan (p,y), comme l'illustre la figure 9. Étant donnée une valeur $y = f$ observée sur un échantillon donné, les bornes de l'intervalle de confiance s'obtiennent en résolvant en p l'équation suivante :

$$(f-p)^2 = u_{\alpha/2}^2 \frac{p(1-p)}{n} \qquad (E).$$

On obtient

$$\begin{aligned}(E) &\iff p^2\left(1+\frac{u_{\alpha/2}^2}{n}\right) - p\left(2f+\frac{u_{\alpha/2}^2}{n}\right) + f^2 = 0 \\ &\iff p = \frac{2f+\frac{u_{\alpha/2}^2}{n} \pm \sqrt{\frac{u_{\alpha/2}^4}{n^2} + 4\,f\frac{u_{\alpha/2}^2}{n} - 4\,f^2\frac{u_{\alpha/2}^2}{n}}}{2\left(1+\frac{u_{\alpha/2}^2}{n}\right)}.\end{aligned}$$

Cette formule étant «un peu» encombrante, on considère en général une approximation de ces solutions à l'aide d'un développement limité au premier ordre en $\frac{1}{n}$. Ainsi, le premier terme donne :

$$\frac{2f+\frac{u_{\alpha/2}^2}{n}}{2\left(1+\frac{u_{\alpha/2}^2}{n}\right)} = f + o\left(\frac{1}{n}\right).$$

Le second terme se réduit alors à

$$\sqrt{\frac{u_{\alpha/2}^4 + 4\,f\,n\,u_{\alpha/2}^2 - 4\,f^2 n\,u_{\alpha/2}^2}{4n^2 + 8\,u_{\alpha/2}^2 n + 4\,u_{\alpha/2}^4}} = \sqrt{\frac{f\,n\,u_{\alpha/2}^2 - f^2 n\,u_{\alpha/2}^2}{n^2}} + o\left(\frac{1}{n}\right)$$
$$= u_{\alpha/2}\sqrt{\frac{f(1-f)}{n}} + o\left(\frac{1}{n}\right).$$

On obtient ainsi le théorème suivant.

Théorème 11

Si $n > 100$ et $n\,f(1-f) > 18$, l'intervalle de confiance réel pour p au niveau de confiance $1-\alpha$ est

$$Ic_{1-\alpha}(p) = \left[f - u_{\alpha/2}\sqrt{\frac{f(1-f)}{n}}, f + u_{\alpha/2}\sqrt{\frac{f(1-f)}{n}}\right].$$

Exemple 20

On souhaite estimer la proportion p de cyclistes parisiens portant un casque. Sur un échantillon de 400 cyclistes, on a observé une proportion $f = 36\%$. On souhaite un intervalle de confiance pour p au niveau de confiance 95%.

On a $n\,f(1-f) = 92{,}16 > 18$, l'approximation par la loi normale est donc légitime. Le théorème 11 donne donc l'intervalle de confiance

$$Ic_{0{,}95}(p) = \left[0{,}36 - 1{,}96\sqrt{\frac{0{,}36 \times 0{,}64}{400}}; 0{,}36 + 1{,}96\sqrt{\frac{0{,}36 \times 0{,}64}{400}}\right] = [0{,}31, 0{,}41].$$

Il est donc fort probable que la proportion cherchée soit comprise entre 31% et 41%.

5 Contrôle statistique

La qualité est au cœur des préoccupations de l'entreprise. Elle est devenue un point-clé de leur compétitivité. Pour un constructeur automobile, par exemple, il est vital de s'assurer que les véhicules livrés sont conformes aux attentes des clients.

Pour maîtriser un processus de production, les entreprises mettent en place des méthodes statistiques permettant de créer et de fabriquer des produits de qualité. Dans le cas de fabrication en moyennes et grandes séries, l'utilisation de techniques statistiques permet notamment d'éviter le contrôle de toutes les unités produites (contrôle à 100%) et de prévenir les effets de déréglages au lieu de les subir. L'ensemble de ces méthodes et actions permettant d'évaluer de façon statistique les paramètres d'un processus de production s'appelle la «maîtrise statistique des processus» (MSP) ou, en Anglais, «statistical process control» (SPC).

Dans cette section, nous allons présenter une méthode statistique de contrôle largement utilisée dans l'Industrie. Il s'agit des **cartes de contrôle**, ou **cartes de maîtrise** (en Anglais : control charts).

5.1 Principe des cartes de contrôle

Les cartes de contrôle ont été introduites en 1931 par Walter Shewhart [3] pour améliorer la qualité de la production au sein de l'usine Western Electric à Chicago. Elles se sont largement développées en Europe depuis cette date grâce à la mise en place de normes qualité (ISO 9000 notamment). Elles s'appuient sur la théorie de l'échantillonnage et de l'estimation.

Une carte de contrôle est un graphique permettant de suivre l'évolution d'un processus de production et de savoir si le processus a dérivé, auquel cas on dit qu'il est «hors contrôle».

5.2 Carte de contrôle p

Exemple 21

Dans une ligne de production de semi-conducteurs, on considère qu'une fabrication de N unités contient une proportion p de pièces non conformes. Comme p est en général inconnu, la première étape consiste à l'estimer à l'aide de plusieurs échantillons de taille n suffisante. On obtient ainsi une valeur p_0 que l'on considère comme une estimation ponctuelle de p et que l'on désigne parfois comme la ***valeur cible****.*

Si K est le nombre de pièces non conformes observés dans un échantillon de taille n, nous avons vu que K est variable aléatoire qui suit la loi binomiale $\mathcal{B}(n,p)$, d'espérance np et de variance $n\,p(1-p)$, que nous pouvons estimer respectivement par $n\,p_0$ et $n\,p_0(1-p_0)$.

Si on est dans les conditions d'application du théorème de Moivre-Laplace (en pratique $n\,p_0(1-p_0) > 18$), nous pouvons considérer que K suit approximativement la loi normale $\mathcal{N}\,(n\,p_0,n\,p_0(1-p_0))$. Alors, la variable aléatoire

$$U = \frac{K - n\,p_0}{\sqrt{n\,p_0(1-p_0)}}$$

suit approximativement la loi normale centrée réduite $\mathcal{N}(0,1)$.

Les tables statistiques permettent d'écrire $\mathbb{P}\,(-3 \leqslant U \leqslant 3) = 0{,}997$, c'est-à-dire

$$\mathbb{P}\left(p_0 - 3\sqrt{\frac{p_0(1-p_0)}{n}} \leqslant \frac{K}{n} \leqslant p_0 + 3\sqrt{\frac{p_0(1-p_0)}{n}}\right) = 0{,}997.$$

On obtient ainsi un intervalle de probabilité à 0,997 pour la variable aléatoire $\frac{K}{n}$, qui représente la proportion d'unités non conformes observée dans l'échantillon.

On peut donc considérer qu'il est très peu probable que la proportion d'unités non conformes observée dans un échantillon de taille n n'appartienne pas à l'intervalle

$$\left[p_0 - 3\sqrt{\frac{p_0(1-p_0)}{n}}, p_0 + 3\sqrt{\frac{p_0(1-p_0)}{n}}\right].$$

Si c'est néanmoins le cas, on considérera que le processus est **hors contrôle**.

3. physicien et statisticien américain, 1891–1967

La **carte de contrôle** p est un diagramme centré sur p_0 (proportion estimée d'unités non conformes) et borné par les limites de contrôle

$$Lc_i(p) = p_0 - 3\sqrt{\frac{p_0(1-p_0)}{n}} \qquad \text{(limite inférieure)}$$
$$Lc_s(p) = p_0 + 3\sqrt{\frac{p_0(1-p_0)}{n}} \qquad \text{(limite supérieure).}$$

Ceci nécessite bien sûr de supposer que la taille n des échantillons prélevés reste constante. On reporte en ordonnée les proportions $\frac{k_i}{n}$ d'unités non conformes trouvées dans les échantillons successifs. On obtient alors le graphique de la figure 10.

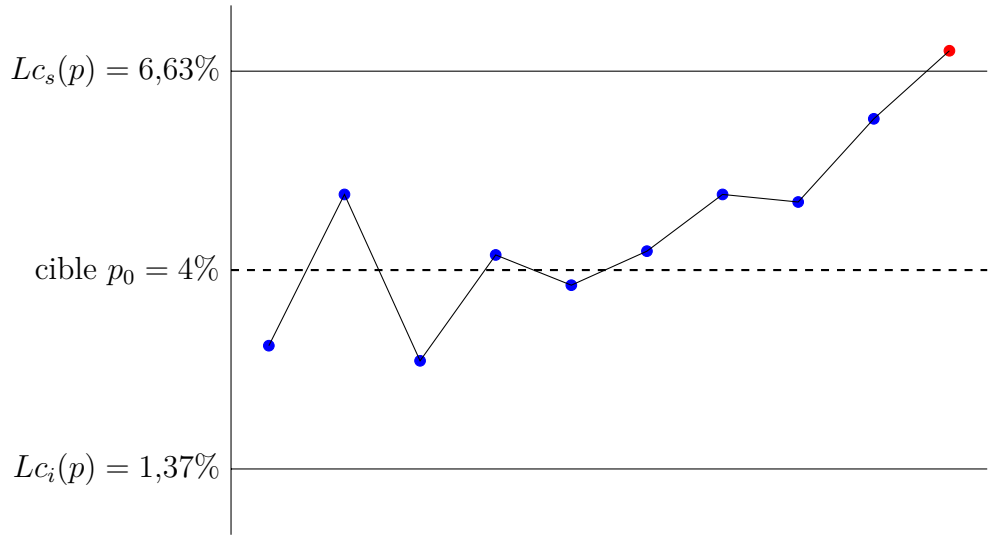

FIGURE 10 – Exemple de carte de contrôle p avec une cible $p_0 = 4\%$ et un échantillon de taille $n = 500$

Sur cette figure 10, on observe notamment que le dernier échantillon est au-delà de la limite de contrôle supérieure, donc que le processus de production est hors contrôle.

L'établissement de ce type de carte de contrôle peut parfois être délicat car l'utilisation de la loi normale pour de petites valeurs de p nécessite des échantillons de taille importante, donc généralement des contrôles automatisés. Pour des petites valeurs de p, il est parfois plus judicieux d'approcher la loi binomiale par la loi de Poisson de paramètre np.

Parfois il n'est pas possible d'assurer une taille d'échantillon constante. Dans ce cas, on peut être amené à faire ajuster les limites de contrôle dynamiquement.

5.3 Cartes de contrôle aux mesures

Le principe du contrôle statistique aux mesure consiste, après avoir établi une référence à partir d'un nombre suffisant de pièces (supérieur à 100) pendant une période stable de fabrication, à prélever régulièrement des échantillons de taille n constante, et à comparer leurs moyennes et écarts-types à la moyenne et à l'écart-type de référence.

Prenons l'exemple d'une fabrication de médicaments.

Exemple 22

Une caractéristique importante lors de la fabrication de médicaments est la masse, qui est une variable aléatoire notée X, d'un comprimé. Il n'est évidemment pas possible de vérifier la masse de chaque comprimé produit.

Si le processus de production est bien maîtrisé, nous admettons que X suit la loi normale $\mathcal{N}(\mu,\sigma^2)$ avec $\mu = 63\,mg$ et $\sigma = 0{,}1\,mg$.

Afin de vérifier que le processus est sous contrôle, on met en place, en s'appuyant sur les moyennes $\overline{x}$ et écarts-types s des échantillons prélevés, deux cartes de contrôle :

— la **carte de la moyenne** (*cf.* figure 11) surveille le réglage du processus de production. Si les points sont tous situés à l'intérieur des limites de contrôle, on ne peut pas conclure à un déréglage. Par contre, si un point sort des limites on a une forte probabilité d'un «décentrage», qu'il faut corriger par un réglage.

— La **carte de l'écart-type** (*cf.* figure 12) surveille la dispersion du processus de production. Si un point se situe au-delà de la limite supérieure de contrôle, cela signifie que la dispersion du processus de production augmente. On arrête alors la ligne de production et on recherche l'origine de la détérioration de la qualité de la production.

Lorsqu'on analyse des cartes de contrôle, on commence généralement par la carte de surveillance de l'écart-type.

5.3.1 Limites de contrôle pour la carte de l'écart-type

On fait l'hypothèse que le processus de production est bien maîtrisé et donc que la variable aléatoire X observée (par exemple la cote ou la masse d'une pièce fabriquée en série) suit la loi normale d'espérance μ et de variance σ^2.

On rappelle qu'au terme du théorème 5, la variable aléatoire $Z = \frac{n\,S^2}{\sigma^2}$ suit la loi du χ^2 à $\nu = n-1$ degrés de liberté, S^2 désignant la variance empirique de l'échantillon. Soit $\alpha > 0$ fixé. En exploitant la table des fractiles de la loi du χ^2, on peut donc déterminer deux nombres $\chi^2_{\alpha/2}$ et $\chi^2_{1-\alpha/2}$ tels qu'on ait l'intervalle de probabilité pour Z à $1-\alpha$:

$$\mathbb{P}\left(\chi^2_{\alpha/2} \leqslant Z \leqslant \chi^2_{1-\alpha/2}\right) = 1-\alpha \quad \text{soit} \quad \sigma\sqrt{\frac{\chi^2_{\alpha/2}}{n}} \leqslant S \leqslant \sigma\sqrt{\frac{\chi^2_{1-\alpha/2}}{n}}.$$

On en déduit la proposition suivante.

Proposition 12

Au risque α de se tromper, l'écart-type de l'échantillon doit être compris entre les limites de contrôle inférieure $Lc_i(\sigma)$ et supérieure $Lc_s(\sigma)$ suivantes :

$$Lc_i(\sigma) = \sigma\sqrt{\frac{\chi^2_{\alpha/2}}{n}} \qquad Lc_s(\sigma) = \sigma\sqrt{\frac{\chi^2_{1-\alpha/2}}{n}}$$

Reprenons la situation l'exemple 22. Si on choisit $\alpha = 0{,}1\%$, par lecture de tables, on obtient

$$\chi^2_{\alpha/2} = 0{,}972 \quad \text{et} \quad \chi^2_{1-\alpha/2} = 29{,}666.$$

Alors, avec des échantillons de taille $n = 10$, on obtient $Lc_i(\sigma) = 0{,}031\,mg$ et $Lc_s(\sigma) = 0{,}172\,mg$. Si tous les échantillons prélevés ont un écart-type compris entre ces deux valeurs,

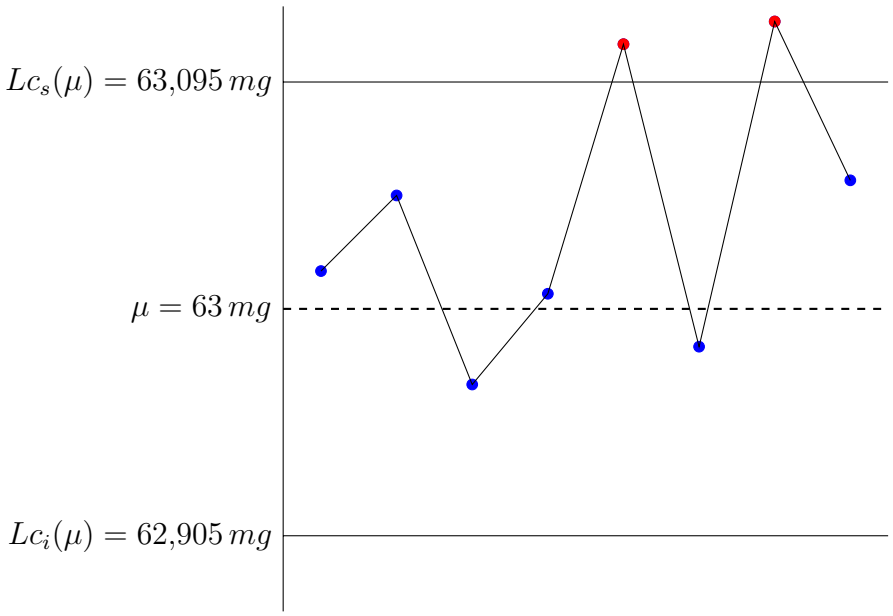

FIGURE 11 – Exemple de carte de contrôle de la moyenne dans le cadre de l'exemple 22

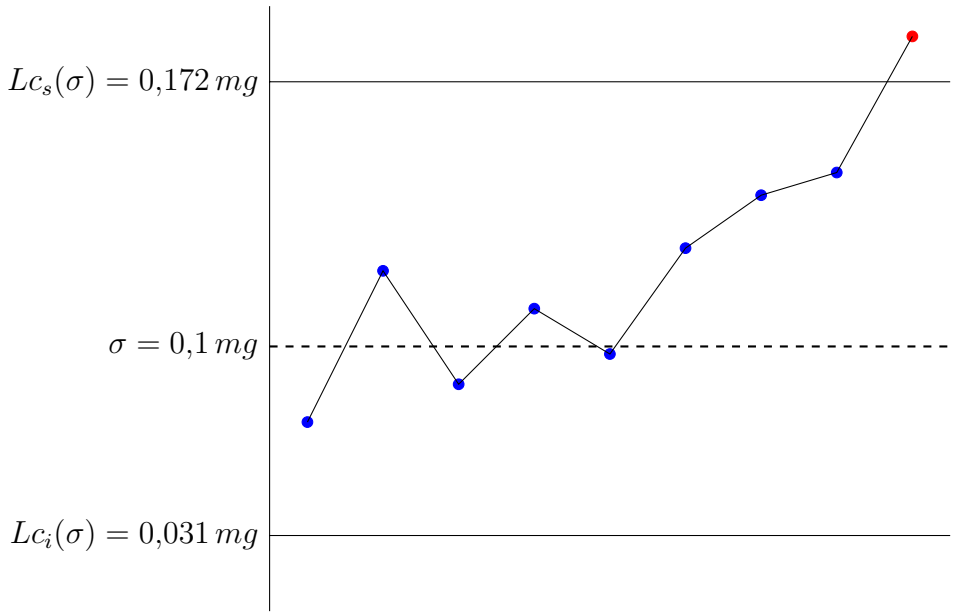

FIGURE 12 – Exemple de carte de contrôle de l'écart-type dans le cadre de l'exemple 22

on accepte l'hypothèse de stabilité de la dispersion. Dans le cas contraire, on estime que le processus est hors contrôle (cas du dernier échantillon de la figure 12).

5.3.2 Limites de controle pour la carte de la moyenne

À nouveau, on fait l'hypothèse que le processus de production est bien maîtrisé et donc que la variable aléatoire X observée (par exemple la cote ou la masse d'une pièce fabriquée en série) suit la loi normale d'espérance μ et de variance σ^2. En pratique, la valeur μ est celle qui a été désignée comme cible, et a été obtenue après étude d'une période stable du processus.

À chaque échantillon de n individus prélevés dans la production, on associe l'estimation ponctuelle $\overline{x}$ de μ. Il s'agit d'une réalisation de la moyenne empirique $\overline{X}$, qui est une variable aléatoire suivant la loi normale $\mathcal{N}\left(\mu,\frac{\sigma^2}{n}\right)$.

Pour $\alpha > 0$, on peut alors déterminer un intervalle de probabilité à $1-\alpha$ pour la variable aléatoire $U = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$: à l'aide du fractile $u_{\alpha/2} = \Phi^{-1}\left(1-\frac{\alpha}{2}\right)$ de la loi $\mathcal{N}(0,1)$. On choisit souvent $\alpha = 0{,}27\%$, ce qui donne $u_{\alpha/2} = 3$. On obtient alors l'intervalle de probabilité pour $\overline{X}$ à $1-\alpha = 99{,}73\%$:

$$\mathbb{P}\left(\mu - 3\frac{\sigma}{\sqrt{n}} \leqslant \overline{X} \leqslant \mu + 3\frac{\sigma}{\sqrt{n}}\right) = 99{,}73\%.$$

On en déduit la proposition suivante.

Proposition 13

Au risque $\alpha = 0{,}27\%$ de se tromper, la moyenne $\overline{x}$ de l'échantillon doit être comprise entre les limites inférieure $Lc_i(\mu)$ et $Lc_s(\mu)$ suivantes :

$$Lc_i(\mu) = \mu - 3\frac{\sigma}{\sqrt{n}} \qquad Lc_s(\mu) = \mu + 3\frac{\sigma}{\sqrt{n}}.$$

Dans la situation de l'exemple 22, avec des échantillons de taille $n = 10$, on obtient

$$Lc_i(\mu) = 62{,}905 \quad \text{et} \quad Lc_s(\mu) = 63{,}095$$

ce qui conduit une carte de contrôle du type de celle de la figure 11, sur laquelle on peut observer deux échantillons différents révélant un processus hors contrôle.

5.4 Efficacité des cartes de contrôle

Supposons la moyenne μ et l'écart-type σ connus et considérons une carte de contrôle de la moyenne. Si un point se trouve hors de la plage de contrôle $[Lc_i(\mu),Lc_s(\mu)]$ explicitée à la proposition 13, on considère que le processus de fabrication est déréglé, et sinon on considère que le processus est bien réglé. Dans ces conditions, il y a deux risques d'erreur :

- le risque α de conclure à tort à un déréglage,
- le risque β de conclure à tort à l'absence de déréglage, c'est-à-dire de ne pas déceler un déréglage existant.

Lors du calcul des limites de contrôle, nous avons fixé la valeur de α à 0,27%. Le risque β est de nature différente. En effet, $1-\beta$ est la probabilité de détecter un déréglage alors que celui-ci existe. En pratique on a donc intérêt à maximiser la valeur de $1-\beta$ puisque cette probabilité traduit la «performance» du dispositif de contrôle.

Lorsque le processus de fabrication est bien maîtrisé, on a $X \sim \mathcal{N}(\mu,\sigma^2)$. Pour calculer β, il faut supposer qu'il est décentré d'une quantité $k\sigma$, et plus précisément supposer que X suit la loi normale $\mathcal{N}(\mu+k\sigma,\sigma^2)$.

Les limites de contrôle pour la moyenne empirique étant toujours celles établies à la proposition 13, on a donc

$$\beta = \mathbb{P}\left(Lc_i(\mu) \leqslant \overline{X} \leqslant Lc_s(\mu)\right).$$

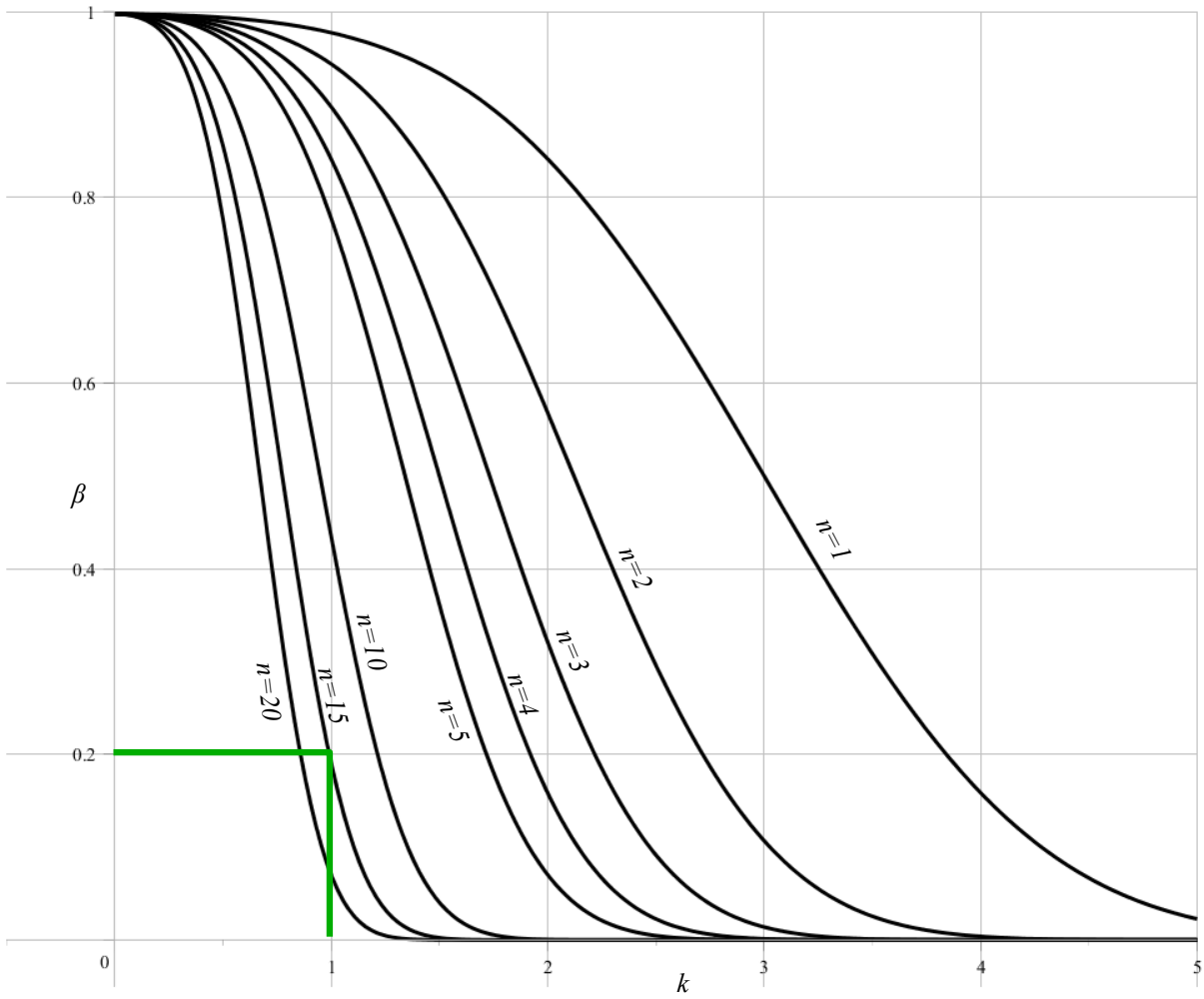

FIGURE 13 – Quelques courbes d'efficacité pour la carte de contrôle de la moyenne

Au vu de la loi de probabilité de $\overline{X}$, la variable aléatoire $U = \dfrac{\overline{X}-\mu-k\sigma}{\sigma/\sqrt{n}}$ suit la loi normale centrée réduite. On a donc

$$\begin{aligned}\beta &= \mathbb{P}\left(\frac{Lc_i(\mu)-\mu-k\sigma}{\sigma/\sqrt{n}} \leqslant U \leqslant \frac{Lc_s(\mu)-\mu-k\sigma}{\sigma/\sqrt{n}}\right)\\ &= \mathbb{P}\left(-3-k\sqrt{n} \leqslant U \leqslant 3-k\sqrt{n}\right) = \Phi(3-k\sqrt{n}) - \Phi(-3-k\sqrt{n})\end{aligned}$$

où Φ désigne la fonction de répartition de $\mathcal{N}(0,1)$.

Définition 12

À taille de l'échantillon n fixée, la courbe représentative de la fonction $k \longmapsto \beta(k)$ est la ***courbe d'efficacité de la carte de contrôle de la moyenne****.*

Faisant varier n, on obtient, comme sur la figure 13, plusieurs courbes d'efficacité de la carte de contrôle de la moyenne, permettant, par exemple, de calibrer la taille des échantillons à prélever en fonction du risque β souhaité pour une valeur du décentrage k.

En pratique, les valeurs admissibles pour β sont comprises entre 5% et 20%.

Exemple 23

Si on souhaite détecter un décentrage de 1σ avec une probabilité $1-\beta = 80\%$, on peut prendre $n = 15$.

Conclusion

Dans ce thème, nous avons exposé les notions de base relatives à la théorie de l'estimation. Ces notions, ainsi que leurs applications vues dans le cadre du contrôle statistique (cartes de contrôle et efficacité) trouveront un prolongement naturel dans les autres thèmes de ce MOOC (tests statistiques et régression linéaire).

Certains concepts tels qu'exhaustivité, efficacité, information de Fisher, liés à la recherche du «meilleur» estimateur d'un paramètre inconnu, n'ont pas été abordés dans ce thème. De même, l'estimation bayésienne, approche qui se révèle très utile dans certains domaines (essais de fiabilité par exemple), n'a pas été traitée. Nous renvoyons les lecteurs intéressés par des compléments à l'ouvrage de Gilbert SAPORTA : Probabilités, Analyse des données et statistique, Éditions Technip, 2006.

Exercices

Exercices sur l'estimation

Exercice 1 : Estimateurs

Soient $X_1, X_2, \ldots, X_n, \ldots$ une suite de variables aléatoires indépendantes suivant la loi uniforme sur $[0,a]$ où a est un paramètre réel strictement positif à estimer. On pose

$$A_n = \sup_{1 \leqslant i \leqslant n} X_i \quad \text{et} \quad B_n = 2\,\overline{X}.$$

1. Déterminer la loi de probabilité de A_n (on pourra utiliser la fonction de répartition).

2. Calculer l'espérance et la variance de A_n et en déduire que A_n est un estimateur de a.

3. Montrer que B_n est un estimateur sans biais de a.

4. Comparer les variances de A_n et de B_n.

Exercice 2 : Estimateur obtenu par la méthode du maximum de vraisemblance

Déterminer un estimateur du paramètre λ d'une loi exponentielle. Celle-ci est définie par la densité de probabilité f suivante :

$$\forall x \in \mathbb{R}, \quad f(x) = \begin{cases} \lambda\, e^{-\lambda x} & \text{si } x \geqslant 0 \\ 0 & \text{sinon.} \end{cases}$$

Exercice 3 : Paramètre d'une loi de Poisson

Dans une ville, on a étudié le nombre d'accidents de la circulation sur une période de 70 jours. En regroupant les jours selon le nombre d'accidents, on a obtenu le tableau suivant

nombre d'accidents	0	1	2	3	4	5
nombre de jours	34	22	11	2	0	1

Soit X la variable aléatoire désignant le nombre d'accidents quotidien. On admet que X suit la loi de Poisson de paramètre λ, c'est-à-dire que

$$X(\Omega) = \mathbb{N} \quad \text{et} \quad \forall n \in \mathbb{N}, \quad \mathbb{P}\left(X = n\right) = e^{-\lambda}\frac{\lambda^n}{n!}.$$

L'objet de cet exercice est de déterminer une estimation ponctuelle de λ par deux méthodes distinctes. Pour cela, on considère un échantillon statistique $(X_1, \ldots, X_n)$ de X, c'est-à-dire n variables aléatoires réelles independantes suivant toutes la loi de Poisson de paramètre λ.

1ère méthode : maximum de vraisemblance.

1. Construire un estimateur de λ par la méthode du maximum de vraisemblance. Est-il sans biais ? En déduire, en utilisant les données du tableau, une estimation ponctuelle de λ.

2ème méthode : on pose $\Sigma_n = \sum_{i=1}^{n} X_i$.

2. Quelle est la loi de probabilité de la variable aléatoire Σ_n ?

3. Montrer que, pour tout entier naturel j,

$$\mathbb{P}\left(\{X_1 = 0\} | \{\Sigma_n = j\}\right) = \left(\frac{n-1}{n}\right)^j.$$

4. En déduire que la variable aléatoire

$$T_n = \left(\frac{n-1}{n}\right)^{\Sigma_n}$$

est un estimateur sans biais de $e^{-\lambda}$.

5. Déduire des questions précédentes un nouvel estimateur de λ et une nouvelle estimation ponctuelle de λ.

6. Comparer les estimateurs obtenus à la question 1 et à la question 5. Conclure.

Exercices sur les intervalles de confiance

Exercice 4 : Intervalle de confiance pour une moyenne et une variance

Une société fabrique des billes pour roulements à billes. On admet que la masse d'une bille est une variable aléatoire suivant la loi normale $\mathcal{N}(\mu,\sigma)$ où μ et σ sont inconnus.

Un échantillon de 30 billes de masses x_i a donné les résultats suivants :

$$\sum_{i=1}^{30} x_i = 69\,g \quad \text{et} \quad \sum_{i=1}^{30} x_i^2 = 163{,}1862\,g^2.$$

1. Déterminer un intervalle de confiance pour μ au niveau de confiance de 95 %.

2. Déterminer un intervalle de confiance pour σ au niveau de confiance de 95 %.

Exercice 5 : Intervalle de confiance pour une proportion

On appelle p la proportion de billes défectueuses dans une production de billes. Déterminer un intervalle de confiance pour p au seuil 5 % dans les deux cas suivants.

1. Dans un échantillon de 100 billes, on a observé 11 billes défectueuses.

2. Dans un échantillon de 500 billes, on a observé 48 billes défectueuses.

Exercice 6 : Publicité mensongère ?

Un fabricant de piles électriques indique sur ses produits que la durée de vie moyenne de ses piles est de 200 heures. Une association de consommateurs prélève un échantillon de 25 piles et observe une durée de vie moyenne de 185 heures avec un écart-type (calculé à partir de l'estimateur biaisé de la variance) de 30 heures.

1. S'agit-il de publicité mensongère ? On précisera la démarche utilisée (hypothèses, raisonnements, calculs, *etc.*).

2. Que faudrait-il faire pour répondre négativement à la question posée ?

Exercice 7 : Paramètre d'une loi continue

Pour $\theta > 0$, on définit la fonction

$$\begin{array}{rccl} f_\theta : & \mathbb{R} & \longrightarrow & \mathbb{R} \\ & x & \longmapsto & \begin{cases} e^{\theta - x} & \text{si } x \geqslant \theta \\ 0 & \text{sinon.} \end{cases} \end{array}$$

L'objet de ce problème est de construire des estimations de θ.

1. Démontrer que f_θ est une densité de probabilité.

2. Soit X une variable aléatoire admettant f_θ pour densité de probabilité. Démontrer que la variable X admet une espérance et une variance et que

$$E[X] = \theta + 1 \quad \text{et} \quad \mathrm{Var}(X) = 1.$$

Dans toute la suite de l'exercice, on considère des variables aléatoires $X_1, \ldots, X_n$ indépendantes admettant f_θ pour densité de probabilité.

3. On pose $U_n = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - 1)$.

3.1 Calculer l'espérance et la variance de U_n.

3.2 Que peut-on en déduire ?

4. Justifier que, si n est assez grand, la variable aléatoire

$$T_n = \sqrt{n}\,(U_n - \theta)$$

suit approximativement la loi normale centrée réduite.

5. On suppose que $n = 100$ et $\overline{x} = 2{,}0706$. À l'aide de la question précédente, déterminer un intervalle de confiance pour θ au niveau de confiance 95%.

6. On admet le résultat suivant :

Théorème (*Inégalité de Bienaymé-Tchebychev*)

Si Y est une variable aléatoire admettant une espérance μ et une variance σ^2, alors

$$\forall k > 0, \qquad P\left(|Y - \mu| \geqslant k\right) \leqslant \frac{\sigma^2}{k^2}.$$

6.1 En appliquant ce théorème à la variable aléatoire $\overline{X}$, déterminer un intervalle de confiance aléatoire pour θ avec un niveau de confiance supérieur ou égal à $1 - \alpha = 95\%$.

6.2 Donner l'intervalle réel ainsi obtenu lorsque $n = 100$ et $\overline{x} = 2{,}0706$.

7. Comparer les intervalles de confiance obtenus aux questions 6.2 et 5. Quel serait le niveau de confiance permettant d'obtenir l'intervalle de confiance de la question 6.2 avec la méthode de la question 5 ?