République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA

FACULTÉ des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire présenté en vue de l'obtention du Diplôme :

MASTER en Mathématiques

Option : Statistique

Par

Saâd Assia

Titre :

Estimation par Intervalle de Confiance et Application

Membres du Comité d'Examen :

Pr.	Yahia Djabrane	UMKB	Président
Pr.	Brahimi Brahim	UMKB	Encadreur
Dr.	Zouaoui Nour El Houda	UMKB	Examinatrice

Juin 2024

Dédicace

A mes chers parents :

Ma mère, symbole de tendresse, qui s'est sacrifiée pour mon bonheur et ma réussite.

Mon père, école de mon enfance, qui a veillé à me protéger, m'aider et m'encourager.

A mon soutien qui m'a encouragé à terminer mes études et à supporter toutes les difficultés avec moi, mon cher époux, Mahmoud.

Aux plaisirs de mon foie et de mes fleurs, mes chers enfants Loujaine et Maram.

A ma seule soeur Kalthoum.

A mes frères Saber, Abd El Kaher, Issam, Marouane.

A toute ma grande famille.

Et à la mémoire des mes grands-pères et mes amirs Rachida et Manel.

REMERCIEMENTS

Tout d'abord, je remercie "Allah Le Tout Puissant" de m'avoir aidé et donné la santé et volonté pour arriver à ce stade.

Mes vifs remerciements sont adressés à mon encadreur Pr. Brahim Brahimi pour ses précieux

conseils, ses orientations pertinentes et sa patience tout au long de la réalisation de ce mémoire.

Je tiens à remercier Pr. Djabrane Yahia et Dr.Zouaoui Nour El Houda qui m'ont fait l'honneur de

faire partie du jury de soutenance.

Je remercie tous les enseignants qui ont contribué à ma formation, ainsi que tous les employés

du département de Mathématiques.

Je remercie tout particulièrement mes parents, pour leur encouragement et soutien sur tous les

aspects, ainsi que toute ma famille.

Je remercie également l'étudiante distingué et généreux qui a apporté une grande contribution tout au long de l'année universitaire, Zernadji Loubna.

Je n'oublie pas l'ensemble de mes amis proches et aussi mes collègues d'études.

A ceux qui ont contribué, de près ou de loin, à la réalisation de ce modeste travail : un grand

merci à vous tous.

Table des matières

Table des figures

Liste des tableaux

Introduction

L'estimation par intervalle de confiance est une méthode statistique fondamentale utilisée pour estimer des paramètres inconnus d'une population à partir d'un échantillon de données. Cette méthode fournit un intervalle de valeurs plausibles pour le paramètre inconnu, avec un niveau de confiance spécifié.

L'importance de l'estimation par intervalle de confiance réside dans le fait qu'elle permet aux chercheurs et aux décideurs de prendre des décisions éclairées en tenant compte de l'incertitude inhérente aux estimations statistiques. En fournissant un intervalle de valeurs plutôt qu'un simple point estimé, cette méthode permet de mieux évaluer la précision et la fiabilité des estimations.

Au XVIIe siècle, le scientifique britannique John Graunt est considéré comme l'un des premiers à avoir abordé le sujet de l'estimation par intervalle de confiance et des statistiques. En 1662, Graunt a publié un livre intitulé "Natural and Political Observations Made upon the Bills of Mortality", dans lequel il utilisait des données statistiques disponibles pour estimer les taux de mortalité et de natalité à Londres. Il a également fourni des estimations de confiance pour ces données, ce qui en fait l'un des premiers à utiliser ce concept en statistique.

Nous exposons notre travail en trois chapitres :

Chapitre 01 : Dans le premier chapitre, les concepts fondamentaux de statistique mathématique sont examinés. Ce chapitre aborde plusieurs sujets de base tels que

les lois de probabilité usuelles, (les loi déscrètes et continues). Il traite également de concepts tels que la moyenne et la variance empirique, ainsi que de théorème centrale limite et la loi forte et faible des grandes nombres.

Chapitre 02 : Dans ce chapitre de mémoire, nous abordons les concepts fondamentaux de l'estimation statistique, notamment l'intervalle de confiance et les méthodes d'estimation ponctuelle et principale. Nous examinons ensuite les différentes méthodes d'estimation, notamment la méthode des moments et la méthode du maximum de vraisemblance, en soulignant leur utilisation pratique dans la construction d'intervalles de confiance. Enfin, nous discutons l'application sur l'estimation par l'intervalle de confiance.

Chapitre 1

Généralités

1.1 Définitions et concepts de base

1.1.1 Les lois usuelles en statistique

Lois discrètes

Définition 1.1.1 (Loi uniforme) *On dit qu'une v.a discrète X est uniforme sur un ensemble de points réels $x_1, x_2, \ldots, x_n, n \geq 1$ si elle admet pour loi de probabilité*

$$P(X = k) = \frac{1}{n} 1\mathrm{I}_{(x_1, x_2, \ldots, x_n)}(k),$$

avec,

$$E(X) = \frac{n+1}{2}, \; V(X) = \frac{n^2 - 1}{12}.$$

Définition 1.1.2 (Loi Poisson) *Une v.a X est dite de Poisson de paramètre $\lambda > 0$ (on écrit $X \sim \mathcal{P}(\lambda)$) si sa loi de probabilité*

$$P(X = k) = \frac{\lambda^k}{k!} \exp(-\lambda) 1\mathrm{I}_{\mathbb{N}}(k),$$

avec,

$$E(X) = V(X) = \lambda.$$

La loi de Poisson est une distribution de probabilité discrète qui décrit le nombre d'événements rares se produisant dans un intervalle fixe de temps ou d'espace, compte tenu d'un taux moyen d'occurrence. Elle est souvent utilisée dans des domaines tels que la modélisation des files d'attente, l'analyse des données de trafic, et la biologie pour modéliser le nombre d'événements tels que les mutations génétiques. La loi de Poisson est caractérisée par le fait que le nombre moyen d'événements est égal à la variance, et elle devient une approximation de la distribution binomiale dans le cas où le nombre d'essais est grand et la probabilité de succès est petite.

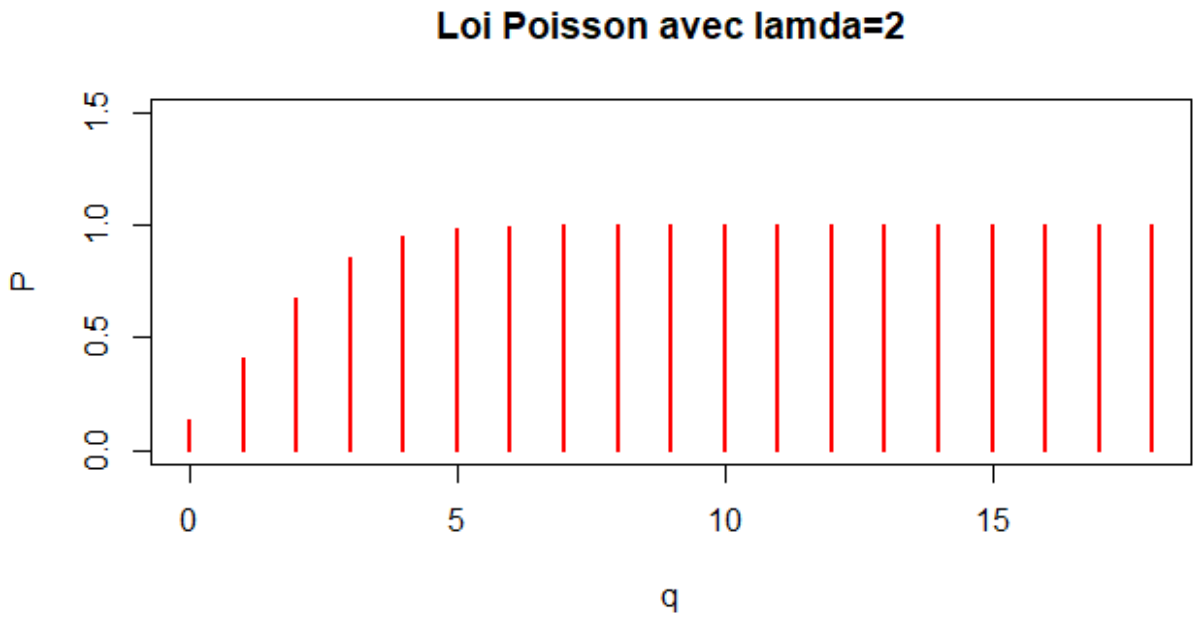

FIG. 1.1 – Loi Poisson.

Définition 1.1.3 (Loi de Bernoulli) *On dit qu'une v.a X suit la loi de Bernouilli de paramètre $0 < p < 1$, et on écrit $X \sim \mathcal{B}(p)$, si elle ne prend que deux valeurs*

possibles 0 *ou* 1 *avec des probabilités respectives* $1-p$ *et* p, *c-à-d*

$$P(X=k)=\begin{cases} 1-p & k=0 \\ p & k=1 \end{cases}$$

tel que, p *le paramètre de succès et*

$$E(X)=p,\ V(X)=p(1-p).$$

La loi de Bernoulli est une distribution de probabilité discrète qui modélise une seule expérience aléatoire avec deux résultats possibles : succès avec une probabilité p *et échec avec une probabilité* $1-p$. *Elle est nommée d'après le mathématicien suisse Jacob Bernoulli. Cette distribution est souvent utilisée pour modéliser des événements binaires tels que la réussite ou l'échec d'un essai, la présence ou l'absence d'un trait, ou encore le résultat d'un lancer de pièce de monnaie. La loi de Bernoulli est la base de nombreuses autres distributions de probabilité, notamment la distribution binomiale qui modélise le nombre de succès dans un nombre fixe d'essais de Bernoulli indépendants.*

Définition 1.1.4 (Loi binomiale) *Une v.a* X *est dite binomiale de paramètres* $n\in\mathbb{N}$ *et* $0<p<1$ *(on écrit* $X\sim\mathcal{B}(n,p)$*) si sa loi de probabilité s'écrit de la forme suivant*

$$P(X=k)=C_n^k p^k(1-p)^{n-k}1\mathrm{I}_{\{0,\dots,n\}}(k)$$

avec,

$$E(X)=np,\ V(X)=np(1-p).$$

Cette loi est utilisée pour calculer la probabilité de différents nombres de succès dans un nombre fixe d'essais, ce qui en fait un outil précieux pour l'analyse statistique des données.

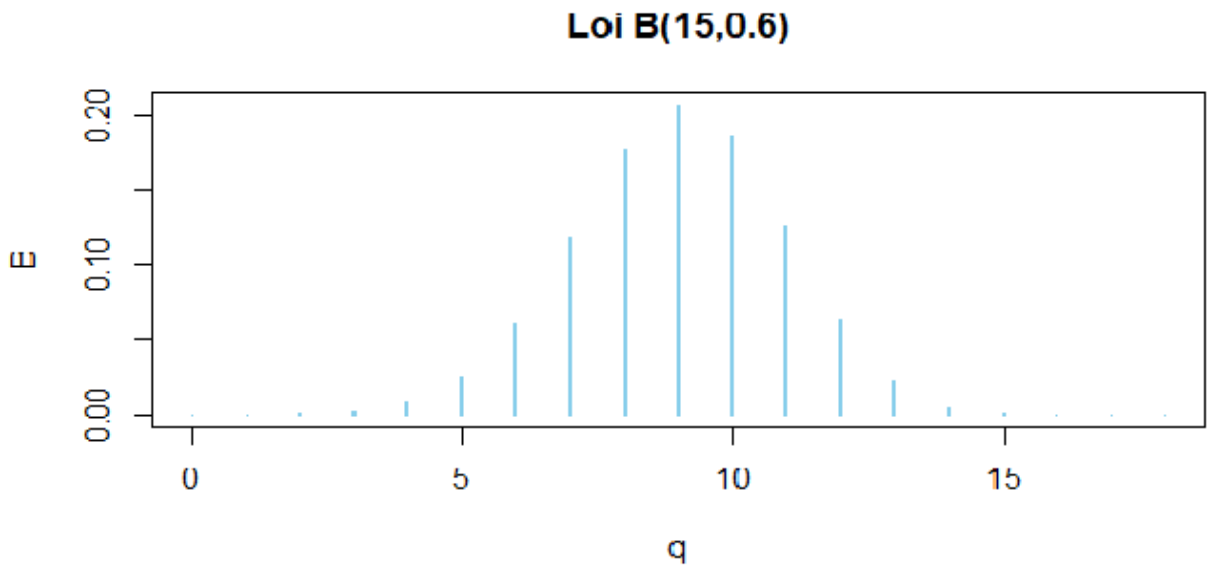

FIG. 1.2 – Loi Binomiale.

Lois continues

Définition 1.1.5 (Loi uniforme) *On dit une v.a continue X suit une loi uniforme sur un intervalle $[a,b](X \sim \mathcal{U}([a,b]))$, où a et b sont deux réels tels que $a < b$, si elle admet pour densité de probabilité*

$$f(x) = \frac{1}{b-a} 1\mathrm{I}_{[a,b]}(x)$$

et on a

$$E(X) = \frac{a+b}{2}, V(X) = \frac{(b-a)^2}{12}.$$

Définition 1.1.6 (Loi exponentielle) *On dit qu'une v.a X suit la loi exponentielle de paramètre $\theta > 0$, et on écrit $X \sim \mathcal{E}(\theta)$, si sa densité de probabilité est définie par*

$$f(x) = \theta \exp(-\theta x) 1\mathrm{I}_{[0,+\infty[}(x),$$

avec,

$$E(X) = \frac{1}{\theta}, V(X) = \frac{1}{\theta^2}.$$

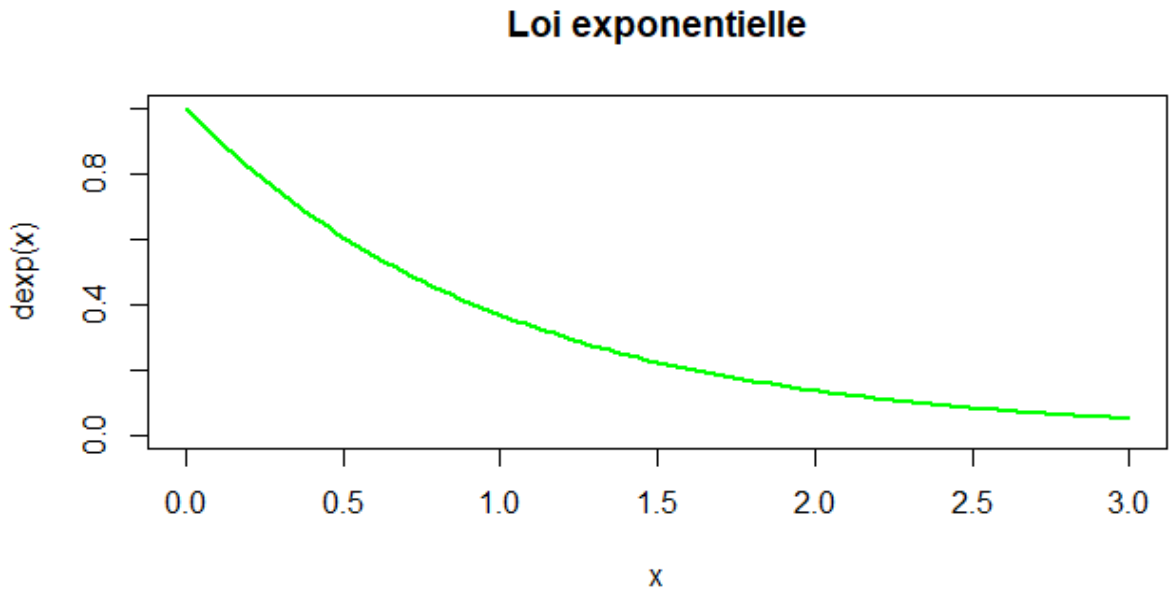

FIG. 1.3 – Loi exponentielle.

Définition 1.1.7 (Loi normale) *Une v.a X suit une loi normale ou gaussienne (loi de Laplace-Gauss) de paramètres $\mu \in \mathbb{R}$ et $\sigma^2 > 0$ ($X \sim \mathcal{N}(\mu, \sigma^2)$) si elle admet pour densité de probabilité la fonction f définie, pour tout nombre réel x, par*

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

tel que,

$$E(X) = \mu, V(X) = \sigma^2,$$

et lorsque X v.a normale centrée.réduite ($X \sim \mathcal{N}(0,1)$) on a

$$E(X) = 0, V(X) = 1.$$

La distribution normale est l'une des distributions de probabilité les plus connues en statistique. Cette distribution se caractérise par une courbe en forme de cloche symétrique autour de sa moyenne, où la plupart des données se situent près de cette valeur et la probabilité diminue avec l'éloignement de celle-ci. La distribution normale est utilisée dans de nombreux domaines tels que la psychologie, l'économie, la physique, etc., en raison de sa compatibilité avec de nombreux phénomènes naturels et sociaux.

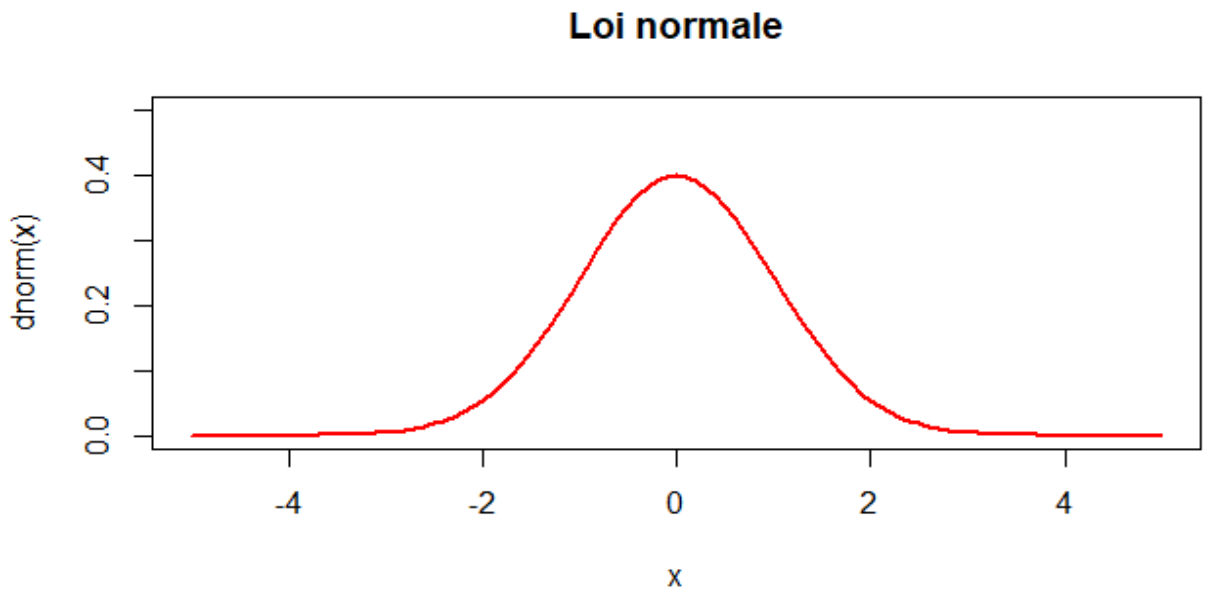

FIG. 1.4 – Loi Normale

Définition 1.1.8 (Loi du khi-deux) *On dit qu'une v.a X suit une loi du khi-deux ou loi de Pearson et on écrit $X \sim \mathcal{X}_n^2$ si elle admet pour densité*

$$f(x) = \frac{2^{-n/2}}{\Gamma(n/2)} \exp(-x/2) x^{n/2-1} 1\mathrm{I}_{[0,+\infty[}(x),$$

où $\Gamma(n/2) = \int_0^{+\infty} u^{n/2-1} \exp(u) du$, désigne la fonction gamma.

La loi du $\mathcal{X}^2$ (chi-deux) est une distribution de probabilité continue qui apparaît

souvent en statistique. Elle est utilisée pour modéliser des variables aléatoires qui représentent des sommes de carrés de variables aléatoires normalement distribuées et indépendantes. La loi du $\mathcal{X}^2$ est couramment utilisée pour tester l'indépendance entre des variables, ainsi que pour tester l'adéquation d'un modèle statistique à des données observées. Elle est également utilisée dans les analyses de régression pour évaluer la significativité des coefficients de régression. La forme de la distribution du $\mathcal{X}^2$ dépend du nombre de degrés de liberté, qui est un paramètre important dans son utilisation.

Définition 1.1.9 (Loi de Student) *Une v.a X est dite variable de Student de degré de liberté $n \geq 1$ (on note $X \sim t_n$) si elle s'écrit sous la forme suivante*

$$X = \frac{Z}{\sqrt{Y/n}},$$

où $Z \sim N(0,1)$ et $Y \sim X_n^2$ sont indépendantes.

La distribution de Student, ou distribution t, est un outil essentiel en statistique pour estimer les moyennes de petites échantillons lorsque l'écart-type de la population est inconnu. Elle a été introduite par William Sealy Gosset sous le pseudonyme de "Student" en raison de son emploi à la Guinness Brewery et de la nécessité de garder confidentiels les détails de ses travaux. Cette distribution diffère de la distribution normale par ses queues plus épaisses, ce qui en fait un outil précieux pour analyser des données où les observations peuvent être plus dispersées. La distribution de Student est largement utilisée dans les études scientifiques, les enquêtes et d'autres domaines où l'analyse des échantillons est nécessaire.

1.1.2 Moyenne et variance empirique

Définition 1.1.10 *Un n-échantillon alèatoire issu d'une v.a X est une suite $(X_1, \ldots, X_n)$ de $n \geq 1$ v.a indépendantes et identiquement distribuées (iid) ayant la même loi que X : Le nombre n est appelé taille de l'échantillon.*

Définition 1.1.11 *On appelle statistique sur un n-échantillon $(X_1, \ldots, X_n)$ toute fonction mesurable des X_i, $i = 1, \ldots, n$.*

Moyenne empirique

Définition 1.1.12 *On appelle moyenne empirique (échantillonnale, expérimentale) la statistique notée $\overline{X_n}$ définie par :*

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

Proposition 1.1.1 *Soit $(X_1, \ldots, X_n)$ une suite de v.a iid de moyenne μ et variance σ^2 on a :*

$$\begin{aligned} E(\overline{X_n}) &= E(\frac{1}{n}\sum_{i=1}^{n} X_i), \\ &= \frac{1}{n}\sum_{i=1}^{n} E(X_i), \\ &= \mu. \end{aligned}$$

et

$$\begin{aligned} V(\overline{X_n}) &= V(\frac{1}{n}\sum_{i=1}^{n} X_i), \\ &= \frac{1}{n^2}\sum_{i=1}^{n} V(X_i), \\ &= \frac{1}{n}\sigma^2. \end{aligned}$$

Variance empirique

Définition 1.1.13 *On appelle variance empirique, la statistique*

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2. \tag{1.1}$$

Proposition 1.1.2 *Soit* $(X_1, \ldots, X_n)$ *une suite de v.a iid de moyenne* μ *et variance* σ^2 *on a :*

$$\begin{aligned} E(S_n^2) &= \frac{1}{n}\sum_{i=1}^{n}E((X_i - \overline{X_n})^2), \\ &= \frac{1}{n}\sum_{i=1}^{n}E({X_i}^2 + \overline{X_n}^2 - 2\mu X_i), \\ &= V(X_i) + E(X_i)^2 + V(\overline{X_n}) + E(\overline{X_n})^2 - 2\mu^2, \\ &= \sigma^2 + \frac{1}{n}\sigma^2. \end{aligned}$$

Remarque 1.1.1 *Le covariance entre* $\overline{X_n}$ *et* S_n^2 *donner par :*

$$cov(\overline{X_n}, S_n^2) = \frac{n-1}{n^2}\mu_3, \text{ avec } \mu_3 \text{ le moment d'ordre 3.}$$

1.2 Mode de convergence

Vu l'utilité de la notion de convergence dans l'estimation statistique, on rappelle, dans cette section, les défnitions et résultats essentiels relatifs aux différents modes de convergence. Soit $(X_n)_{n\geq 1}$ une suite de v.a.

1.2.1 Convergence en loi

Définition 1.2.1 *On dit que $(X_n)_{n\geq 1}$ converge en loi vers une v.a X, et on écrit $X_n \xrightarrow{\mathcal{L}} X$, si*

$$\lim_{n\to+\infty} F_n(x) = F(x),$$

ou F_n et F désignent les fonctions de répartition de X_n et X respectivement.

1.2.2 Convergence en probabilité

Définition 1.2.2 (Convergence en probabilité) *On dit que $(X_n)_{n\geq 1}$ converge en probabilité vers une v.a. X; et on écrit $X_n \xrightarrow{\mathcal{P}} X$, si*

$$\forall \epsilon > 0 : \lim_{n\to+\infty} P(|X_n - X| > \epsilon) = 0.$$

1.2.3 Convergence en moyenne quadratique

Définition 1.2.3 *On dit que $(X_n)_{n\geq 1}$ converge en moyenne quadratique vers une v.a. X, et on écrit $X_n \xrightarrow{mq} X$, si*

$$\lim_{n\to+\infty} E(X_n - X)^2 = 0.$$

1.2.4 Convergence presque sûre

Définition 1.2.4 *On dit que $(X_n)_{n\geq 1}$ converge presque sûrement vers une v.a X, et on écrit $X_n \xrightarrow{ps} X$ si*

$$P(\lim_{n\to+\infty} X_n \neq X) = 0.$$

1.2.5 Liens entre les types de convergence

Les implications suivantes permettent le passage entre certains types de convergence :

$$X_n \xrightarrow{\mathcal{P}} X \Longrightarrow X_n \xrightarrow{\mathcal{L}} X,$$
$$X_n \xrightarrow{ps} X \Longrightarrow X_n \xrightarrow{\mathcal{P}} X,$$
$$X_n \xrightarrow{mq} X \Longrightarrow X_n \xrightarrow{\mathcal{P}} X$$

1.3 Loi forte des grands nombres

La loi forte des grands nombres est un principe fondamental en probabilité et en statistique qui stipule que la moyenne empirique d'une séquence de variables aléatoires indépendantes et identiquement distribuées converge presque sûrement vers l'espérance de cette distribution lorsque le nombre d'observations augmente. En d'autres termes, à mesure que la taille de l'échantillon augmente, la moyenne des échantillons converge vers la moyenne théorique de la population. C'est un concept crucial pour comprendre la fiabilité des estimations basées sur des échantillons.

Théorème 1.3.1 *Soient $(X_1, \ldots, X_n)$ des v.a iid d'espérance μ et de variance σ^2 (finies). Alors, on a*

$$\overline{X_n} \xrightarrow{ps} \mu.$$

1.4 Loi faible des grands nombres

Théorème 1.4.1 *Soient $(X_1, \ldots, X_n)$ des v.a iid d'espérance μ et de variance σ^2 (finies). Alors, on a*

$$\overline{X_n} \xrightarrow{\mathcal{P}} \mu.$$

1.5 Théorème centrale limite

Le théorème central limite est un concept fondamental en statistique qui stipule que la somme (ou la moyenne) d'un grand nombre de variables aléatoires indépendantes et identiquement distribuées suit approximativement une distribution normale, quelle que soit la distribution initiale des variables. C'est un outil essentiel pour comprendre le comportement des échantillons dans divers domaines de la statistique et de la science des données.

Théorème 1.5.1 *Soit $(X_n)_{n\geq 1}$ une suite de v.a iid, d'espérance μ et de variance σ^2 finies. Alors*

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

Chapitre 2

Intervalles de confiance

2.1 Méthode d'estimation ponctuelle

L'estimation ponctuelle et l'intervalle de confiance sont deux concepts clés en statistique qui permettent d'estimer un paramètre inconnu d'une population à partir d'un échantillon de données. L'estimation ponctuelle consiste à estimer ce paramètre par un seul chiffre, généralement la moyenne ou la médiane de l'échantillon. Cependant, cette estimation peut être entachée d'erreur et il est donc important de fournir également un intervalle de confiance, qui donne une fourchette de valeurs plausibles pour ce paramètre, avec un certain niveau de confiance. Ce niveau de confiance est généralement fixé à 95%, ce qui signifie qu'il y a 95% de chances que le véritable paramètre se trouve dans l'intervalle de confiance. L'estimation ponctuelle et l'intervalle de confiance sont des outils essentiels pour interpréter les résultats des études statistiques et pour prendre des décisions éclairées en se basant sur des données probantes.

2.1.1 Définition et propriétés

Soit X une v.a dont la distribution dépend d'un (ou plusieurs) paramètre. La donnée d'un modèle statistique c'est la donnée d'une famille de probabilités $\{P_\theta, \theta \in \Theta\}$, où Θ représente l'espace des valeurs du paramètre inconnu. On dit alors que la loi de X appartient au modèle $\{P_\theta, \theta \in \Theta\}$. On souhaite estimer ce paramètre à partir de l'observation d'un échantillon alèatoire $(X_1, \ldots, X_n)$, de taille $n \geq 1$, extrait de la population X.

Définition 2.1.1 *Un estimateur du paramètre est une statistique (voir la définition 1.1.11), généralement notée par $\widehat{\theta_n}$, contenant le plus d'information possible sur θ. La qualité d'un estimateur s'exprime par sa convergence ou consistance, son biais, son efficacité et/ou sa robustesse.*

Définition 2.1.2 (Consistance) *Un estimateur $\widehat{\theta_n}$ est dit convergent ou consistant s'il est "proche" de θ au sens de la convergence en probabilité, pour tout $\epsilon > 0$*

$$P(|\widehat{\theta_n} - \theta| > \epsilon) \underset{n \to +\infty}{\to} 0.$$

Définition 2.1.3 *On appelle biais d'un estimateur $\widehat{\theta_n}$ la quantité $E(\widehat{\theta_n})$. L'estimateur $\widehat{\theta_n}$ est dit sans biais si $E(\widehat{\theta_n}) = \theta$, sinon il est dit biaisé.*

Dans ce qui suit, on va s'intéresser à l'estimation de trois paramètres très connus, à savoir la moyenne, la proportion et la variance. On considère un n-échantillion $(X_1, \ldots, X_n)$ issu d'une loi de moyenne μ et de variance σ^2, toutes deux inconnues.

2.1.2 Estimation de la moyenne

L'estimateur naturel (non-paramétrique) de l'éspérance μ est la moyenne empirique $\overline{X_n}$ définie par 1.1.12. Dans le cas des lois de probabilités usuelles, il est également obtenu par les méthodes du MV et des moments.

Plus généralement, le $k^{éme}$ moment théorique

$$\mu_k := E(X^k),\ k \geq 1, \tag{2.1}$$

est naturellement estimé par le moment empirique d'ordre k

$$m_k := \frac{1}{n}\sum_{i=1}^{n} X_i^k. \tag{2.2}$$

Propriété 2.1.1 *1. $\overline{X_n}$ est un estimateur convergent (faiblement et fortement) de μ.*

2. $\overline{X_n}$ est un estimateur sans biais.

Proposition 2.1.1 *1. Si $X \sim \mathcal{N}(\mu, \sigma^2)$ alors,*

- *Si la variance σ^2 est connue, on a*

$$\sqrt{n}\frac{\overline{X_n} - \mu}{\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1). \tag{2.3}$$

- *Si la variance σ^2 est inconnue, on a*

$$\sqrt{n-1}\frac{\overline{X_n} - \mu}{\sigma} \xrightarrow{\mathcal{L}} t_{n-1}. \tag{2.4}$$

2. Si X est de distribution quelconque ou même inconnue, la loi de $\overline{X_n}$ est approchée par une loi normale d'éspérance μ et de variance σ^2/n, lorsque n est grand (en pratique $n > 30$).

2.1.3 Estimation de la variance

Lorsque la moyenne de X est inconnue, la variance empirique S_n^2 ; représente l'estimateur naturel de la variance σ^2. Comme $\overline{X_n}$, l'estimateur S_n^2 est obtenu par les méthodes du MV et des moments pour les lois de probabilités usuelles. Dans le cas

où μ est connue, la statistique

$$R_n^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2, \tag{2.5}$$

constitue un estimateur de σ^2 plus précis que S_n^2. En effet, cette dernière est calculée sur la base de $\overline{X_n}$ qui est une valeur approchée de μ, alors que R_n^2 est exprimée en termes de la valeur exacte de μ.

Proposition 2.1.2 *1. S_n^2 est un estimateur convergent (faiblement et fortement) de la variance σ^2.*

2. S_n^2 est un estimateur biaisé pour σ^2. On construit un estimateur sans biais par

$$\widetilde{S_n^2} := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2 = \frac{n}{n-1}S_n^2.$$

3. S_n^2 est un estimateur asymptotiquement sans biais.

4. Si $X \sim \mathcal{N}(\mu, \sigma^2)$, alors

$$\frac{n}{\sigma^2}S_n^2 \sim \mathcal{X}_{n-1}^2. \tag{2.6}$$

Remarque 2.1.1 *R_n^2 est un estimateur convergent (faiblement et fortement), sans biais de σ^2 et vérifiant*

$$\frac{n}{\sigma^2}S_n^2 \sim \mathcal{X}_n^2. \tag{2.7}$$

En effet, nR_n^2/σ^2 est égal à la somme des carrés de n v.a normales centrées réduites indépendantes.

2.1.4 Estimation d'une proportion

Soit $0 < p < 1$ la proportion (pourcentage) d'individus, dans une population, possédant un ceratain caractère. C'est le paramètre de succès dans une expérience de Bernoulli. L'estimateur naturel de p est ce que l'on appelle la fréquence empirique,

notée K_n, calculée sur un échantillon de taille $n \geq 1$. Elle peut être considérée comme une moyenne empirique particulière, où les v.a X_i sont des variables (indépendantes) de Bernoulli de paramètre p. Elle est alors definie par

$$K_n := \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{2.8}$$

Proposition 2.1.3 *1. K_n est un estimateur convergent (faiblement et fortement) de p.*

2. K_n est un estimateur sans biais.

3. La distribution de K_n est donnée par

$$nK_n \sim B(n,p) \text{ et } \frac{K_n - p}{\sqrt{p(1-p)/n}} \sim \mathcal{N}(0,1). \tag{2.9}$$

2.2 Méthodes d'estimation principales

Il existe plusieurs façons de construire un estimateur pour un paramètre donné. Les plus populaires sont la méthode des moments et celle du maximum de vraisemblance (MV).

2.2.1 Méthode des moments

Définition 2.2.1 *On appelle estimateur des moments de θ, toute solution $\widehat{\theta_n}$ des équations des moments*

$$\mu_k = m_k, k \geq 1.$$

où $\mu_k$et m_k désignent les moments théorique et empirique respectivement définis par 2.1 et 2.2. En d'autres termes, si le paramètre s'exprime comme fonction $\varphi(\mu_1, \mu_2, ...)$ des premiers moments, alors son estimateur $\widehat{\theta_n}$ est de la forme $\varphi(m_1, m_2, ...)$.

Exemple 2.2.1 *1.* ***Loi de Bernoulli*** $\mathcal{B}(p)$*: On a* $p = \mu_1$*, donc l'estimateur de* p *par la méthode des moments est* $\widehat{p} = m_1 = K_n$*. Cet estimateur n'est autre que la proportion de succès lors de* n *expériences indépendantes de Bernoulli. On retrouve ainsi le principe d'estimation d'une probabilité par une proportion empirique.*

2. ***Loi exponentielle :*** *Si* X *est une v.a de loi* $\mathcal{E}(\theta)$*, où* $E(X) = 1/\theta$*. Donc l'estimateur de* θ *par la méthode des moments est*

$$\widehat{\theta} = 1/m_1 = 1/\overline{X_n}.$$

2.2.2 Méthode du maximum de vraisemblance (MV)

L'estimateur du maximum de vraisemblance de θ est définie comme suit

$$\widehat{\theta} = \arg\max_{\theta\in\Theta} l(X_1,\ldots,X_n;\theta) = \arg\max_{\theta\in\Theta} \log L(X_1,\ldots,X_n;\theta),$$

avec

$$L(X_1,\ldots,X_n;\theta) = \prod_{i=1}^{n} f(x_i,\theta) \text{ est la fonction de vraisemblance.}$$

avec,

$$f(x,\theta) = \begin{cases} f_\theta(x) & \text{si } X \text{ est continue de densité } f_\theta, \\ P_\theta(x) = P(X = x) & \text{si } X \text{ est discrète.} \end{cases}$$

On obtient l'estimateur de $\hat{\theta}$ en résolvant le système suivant

$$\begin{cases} \dfrac{\partial l(X_1,\ldots,X_n;\theta)}{\partial\theta} = 0. \\ \dfrac{\partial^2 l(X_1,\ldots,X_n;\theta)}{\partial\theta^2} < 0. \end{cases}$$

Remarque 2.2.1 *Les deux méthodes d'estimation ne donnent pas nécessairement le*

même estimateur. Par exemple, pour le paramètre $\theta > 0$ *de la distribution uniforme sur l'intervalle* $[0, \theta]$, *l'estimateur des moments est* $2\overline{X_n}$ *alors que celui du MV est* $\max_{1 \leq i \leq n} X_i$.

2.3 Construction d'un intervalle de confiance

Le problème avec l'estimation ponctuelle est qu'elle n'apporte pas d'information sur la précision des résultats, c'est-à-dire qu'elle ne tient pas compte des erreurs dues aux fluctuations d'échantillonnage. Pour évaluer la confiance que l'on peut avoir en une valeur estimée d'un paramètre, il est nécessaire de déterminer un intervalle contenant, avec une certaine probabilité fixée à priori, la vraie valeur du paramètre. C'est l'estimation par intervalles de confiance.

Définition 2.3.1 *S'il existe des v.a* $\theta_{\min} = \theta_{\min}(X_1, \ldots, X_n) < \theta_{\max} = \theta_{\max}(X_1, \ldots, X_n)$ *telles que*

$$P(\theta \in [\theta_{\min}, \theta_{\max}]) = 1 - \alpha,$$

où $0 < \alpha < 1$, *on dit que l'intervalle* $[\theta_{\min}, \theta_{\max}]$ *est un intervalle de confiance pour* θ, *de niveau* $1 - \alpha$. *On le note par* $IC_{1-\alpha}(\theta)$.

Remarque 2.3.1 *1. Le nombre* α *représente le risque que la vraie valeur de n'appartienne pas à* $IC_{1-\alpha}(\theta)$. *C'est la probabilité de l'erreur qu'on commet en affirmant que* $\theta \in IC_{1-\alpha}(\theta)$. *Par contre, le nombre* $1 - \alpha$ *représente la confiance qu'on a en disant que* $\theta \in IC_{1-\alpha}(\theta)$. *En d'autres termes, il y a* $(1-\alpha)\times 100\%$ *de chances que la valeur inconnue de soit comprise entre* $\theta_{\min}$ *et* $\theta_{\max}$. *Dans les problèmes pratiques (économie, agronomie, sociologie, sciences biomédicales,...), les valeurs usuelles du risque* α *sont* $10\%, 5\%, 1\%$ *correspondant à des niveaux de confiance* $90\%, 95\%$ *et* 99% *respectivement.*

2. *Il est clair que plus le risque est petit, plus l'IC est large, c'est-à-dire moins la précision est bonne.*

3. *Un bon IC est un intervalle dont les bornes $\theta_{\min}$ et $\theta_{\max}$ sont des v.a qui dépendent d'un estimateur perfomant $\widehat{\theta_n}$ de θ.*

Il est à noter que $IC_{1-\alpha}(\theta)$ est un intervalle aléatoire. Dans les calculs, on considère les bornes (non aléatoires) associées à une réalisation $(x_1, \ldots, x_n)$ de l'échantillon $(X_1, \ldots, X_n)$.

En résumé, la construction d'un IC revient à la détermination de $\theta_{\min}$ et $\theta_{\max}$ qui vérifient

$$P(\theta < \theta_{\min}) = \alpha_1 \text{ et } P(\theta > \theta_{\max}) = \alpha_2,$$

avec $\alpha_1 + \alpha_2 = \alpha/2$. En général, on choisit $\alpha_1 = \alpha_2 = \alpha/2$ et on parle alors d'un IC à risques symétriques. D'autre part, un IC de niveau $1-\alpha$ est dit unilatéral s'il est défini par

$$P(\theta > \theta_{\min}) = 1 - \alpha, \text{ ou } P(\theta < \theta_{\max}) = 1 - \alpha.$$

La connaissance de la distribution de probabilité de $\widehat{\theta_n}$ permet d'obtenir les bornes de l'IC comme on va le voir dans les cas usuels qui suivent.

2.3.1 Intervalles de confiance pour les paramètres gaussiens

Soit $(X_1, \ldots, X_n)$ un n-échantillon d'une v.a normale de paramètres μ et σ^2. On souhaite construire des IC pour ces derniers.

Estimation de la moyenne

L'IC de μ est construit autour de son meilleur estimateur qui est la moyenne empirique $\overline{X_n}$ définie par 2.2.

Cas où σ^2 est connue En utilisant la relation 2.3, on peut écrire

$$P(-z_{1-\alpha/2} \leq \sqrt{n}\frac{\overline{X_n} - \mu}{\sigma} \leq z_{1-\alpha/2}) = 1 - \alpha$$

où $z_{1-\alpha/2}$ représente le quantile d'ordre $1-\alpha/2$ de la loi normale standard, alors on a

$$P(\overline{X_n} - \frac{\sigma}{\sqrt{n}}z_{1-\alpha/2} \leq \mu \leq \overline{X_n} + \frac{\sigma}{\sqrt{n}}z_{1-\alpha/2}.$$

Par conséquent, on a

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

Cas où σ^2 est inconnue Cette fois, on utilise la relation 2.4 pour écrire

$$P(-t_{1-\alpha/2} \leq \sqrt{n-1}\frac{\overline{X_n} - \mu}{S_n} \leq t_{1-\alpha/2}) = 1 - \alpha,$$

où $t_{1-\alpha/2}$ représente le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n-1$ ddl, alors on a

$$P(\overline{X_n} - \frac{S_n}{\sqrt{n-1}}t_{1-\alpha/2} \leq \mu \leq \overline{X_n} + \frac{S_n}{\sqrt{n-1}}t_{1-\alpha/2}.$$

D'où

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

On note que lorsque n est grand, les quantiles de Student peuvent être remplacés par ceux de Gauss.

Estimation de la variance

Cas où μ est connue La meilleure estimation de σ^2 est donnée par la variable R_n^2 définie par 2.5. En utilisant la distribution 2.7, on écrit

$$P(v_{\alpha/2} \leq \frac{nR_n^2}{\sigma^2} \leq v_{1-\alpha/2}) = 1 - \alpha,$$

où $v_{\alpha/2}$ et $v_{1-\alpha/2}$ sont les quantiles de la loi du Khi-deux à n ddl, d'ordres resectifs $\alpha/2$ et $1 - \alpha/2$.

On a donc,

$$P(\frac{nR_n^2}{v_{1-\alpha/2}} \leq \sigma^2 \leq \frac{nR_n^2}{v_{\alpha/2}}) = 1 - \alpha,$$

alors,

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

Cas où μ est inconnue Dans ce cas, on estime σ^2 par la variance empirique S_n^2 définie par 1.1. D'après la distribution 2.6, on peut écrire

$$P(\tau_{\alpha/2} \leq \frac{nS_n^2}{\sigma^2} \leq \tau_{1-\alpha/2}) = 1 - \alpha,$$

où $\tau_{\alpha/2}$ et $\tau_{1-\alpha/2}$ sont les quantiles de la loi du Khi-deux à n ddl, d'ordres resectifs $\alpha/2$ et $1 - \alpha/2$.

On a donc,

$$P(\frac{nS_n^2}{\tau_{1-\alpha/2}} \leq \sigma^2 \leq \frac{nS_n^2}{\tau_{\alpha/2}}) = 1 - \alpha,$$

et on obtient

$$IC_{1-\alpha}(\sigma^2) = \left[\frac{nS_n^2}{\tau_{1-\alpha/2}}, \frac{nS_n^2}{\tau_{\alpha/2}}\right].$$

On désire construire un IC pour la proportion dont on a parlé dans le paragraphe 2.1.4. On rappelle que le meilleur estimateur de p est la fréquence empirique K_n

définie par 2.8. D'après 2.9; on peut déterminer un IC exact basé sur la loi binomiale qui est une distribution exacte pour tout n, et/ou un IC approximatif basé sur la loi de Gauss qui est une distribution asymptotique. Les bornes de l'IC exact s'expriment en fonction des quantiles de la loi de Fisher-Snédécor. Les détails sur la construction d'un tel IC se trouvent, par exemple, dans [[3], page 57]. Pour l'IC approximatif, on utilise la loi asymptotique de 2.9 pour avoir

$$P(-z_{1-\alpha/2} \leq \frac{K_n - p}{\sqrt{p(1-p)/n}} \leq z_{1-\alpha/2}) \underset{n\to+\infty}{\to} P(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1-\alpha,$$

où Z est une v.a normale centrée réduite, c'est-à-dire

$$P(K_n - z_{1-\alpha/2}\sqrt{p(1-p)/n} \leq p \leq K_n + z_{1-\alpha/2}\sqrt{p(1-p)/n}) \underset{n\to+\infty}{\to} 1-\alpha.$$

Ceci ne fournit pas un IC car les bornes dépendent de l'inconnue p. Cependant, on a le même résultat de convergence, en remplaçant p (dans les bornes de l'intervalle) par son estimateur convergent K_n (voir détails dans [[6], page 310]). On obtient alors

$$P(K_n - z_{1-\alpha/2}\sqrt{K_n(1-K_n)/n} \leq p \leq K_n + z_{1-\alpha/2}\sqrt{K_n(1-K_n)/n}) \underset{n\to+\infty}{\to} 1-\alpha.$$

On conclusion on a

$$IC_{1-\alpha}(p) \backsimeq \left[K_n - z_{1-\alpha/2}\sqrt{K_n(1-K_n)/n}, K_n + z_{1-\alpha/2}\sqrt{K_n(1-K_n)/n}\right].$$

2.4 Simulation

Dans notre simulation on s'intéresse à estimer les paramètres (moyenne et écart-type) d'une loi normale à partir d'un échantillon, en calculant leurs intervalles de confiance respectifs. Au cours de ce travail on va réaliser les étapes suivantes :

1. Écrire une fonction avec langage R pour calculer l'intervalle de confiance de la moyenne.
2. Écrire une fonction avec langage R pour calculer l'intervalle de confiance de l'écart-type.
3. Donne un exemple d'utilisation.
4. Déscute les résultats.

Par exemple #####

voila un programme en R pour estimer les paramètres (moyenne et écart-type) d'une loi normale à partir d'un échantillon, en calculant leurs intervalles de confiance respectifs :

```
# Fonction pour calculer l'intervalle de confiance de la moyenne
intervalle_confiance_moyenne <- function(x, niveau_confiance = 0.95) {
 n <- length(x)
 moyenne <- mean(x)
 ecart_type <- sd(x)
# Calculer l'erreur standard
 erreur_std <- ecart_type / sqrt(n)
# Calculer les quantiles de la loi normale
 quantile <- qnorm((1 - niveau_confiance) / 2, lower.tail = FALSE)
# Calculer les limites de l'intervalle de confiance
 limite_inf <- moyenne - quantile * erreur_std
 limite_sup <- moyenne + quantile * erreur_std
 return(c(limite_inf, limite_sup))
}
# Fonction pour calculer l'intervalle de confiance de l'écart-type
```

```
intervalle_confiance_ecart_type <- function(x, niveau_confiance = 0.95)
{
 n <- length(x)
 ecart_type <- sd(x)

 # Calculer les quantiles de la loi chi-carrée
 quantile_inf <- qchisq((1 - niveau_confiance) / 2, n - 1)
 quantile_sup <- qchisq(1 - (1 - niveau_confiance) / 2, n - 1)

 # Calculer les limites de l'intervalle de confiance
 limite_inf <- sqrt((n - 1) * ecart_type^2 / quantile_sup)
 limite_sup <- sqrt((n - 1) * ecart_type^2 / quantile_inf)

 return(c(limite_inf, limite_sup))
}
# Exemple d'utilisation
x <- rnorm(50, mean = 10, sd = 2) # Données simulées
intervalle_moyenne <- intervalle_confiance_moyenne(x)
intervalle_ecart_type <- intervalle_confiance_ecart_type(x)
cat("Intervalle de confiance à 95% pour la moyenne :\n")
cat(intervalle_moyenne, "\n")
cat("Intervalle de confiance à 95% pour l'écart-type :\n")
cat(intervalle_ecart_type, "\n")
```

comentaire sur le programme

La fonction `intervalle_confiance_moyenne` calcule l'intervalle de confiance pour la moyenne d'un échantillon x. Elle prend deux arguments : `x` (le vecteur de données) et `niveau_confiance` (le niveau de confiance souhaité, par défaut 0,95 pour un intervalle de confiance à 95%).

La fonction `intervalle_confiance_ecart_type` calcule l'intervalle de confiance pour l'écart-type d'un échantillon x. Elle prend également deux arguments : x (le vecteur de données) et `niveau_confiance` (le niveau de confiance souhaité, par défaut 0,95).

Dans l'exemple fourni, des données sont simulées à partir d'une loi normale avec une moyenne de 10 et un écart-type de 2.

Les intervalles de confiance pour la moyenne et l'écart-type sont calculés à l'aide des fonctions `intervalle_confiance_moyenne` et `intervalle_confiance_ecart_type`, respectivement.

Les résultats (les intervalles de confiance pour la moyenne et l'écart-type) sont affichés à l'aide de la fonction `cat.`

IC de la moyenne		IC de l'écart-type	
`limite_inf`	`limite_sup`	`limite_inf`	`limite_sup`
9.1762	10.3881	1.82614	2.724196

TAB. 2.1 – Intervalles de confiance pour la moyenne et l'écart-type.

Conclusion

L'estimation par intervalles de confiance est un outil puissant et essentiel en analyse statistique, permettant aux chercheurs d'évaluer la précision et la fiabilité des estimations statistiques. En fournissant des limites supérieure et inférieure pour la valeur réelle de la quantité estimée, les intervalles de confiance offrent une information plus complète que de simples estimations ponctuelles.

Dans cette mémoire, nous avons exploré différentes méthodes de construction des intervalles de confiance et discuté de leurs applications aux paramètres courants tels que la moyenne, la variance et la proportion. À travers une étude de simulation utilisant des logiciels d'analyse statistique avancés comme R, nous avons vérifié la validité des résultats théoriques et les avons illustrés clairement à l'aide de graphiques et d'analyses numériques.

Les intervalles de confiance constituent une partie intégrante des études statistiques modernes, contribuant à améliorer la précision des estimations et à fournir un contexte plus large pour comprendre les résultats. Cette mémoire peut être étendue pour inclure des applications supplémentaires, telles que l'estimation de la densité de probabilité et des fonctions de répartition non paramétriques, ce qui renforcerait les capacités d'analyse statistique et approfondirait notre compréhension des données.

Bibliographie

[1] Dusart, P. (2015). Cours de statistiques infÈrentielles. UniversitÈ de Limoges.

[2] DeGroot, M., Shervish, M. (2012). Probability and statistics. Addison-Wesley.

[3] Gaudoin, O. (2017). Principes et Méthodes Statistiques. Notes de cours. INP, Grenoble.

[4] Le Coutre, J.P. (2016). Statistique et ProbabilitÈs. Dunod.

[5] Lejeune, M. (2010). Statistique La ThÈorie et ses Applications. Springer.

[6] Saporta, G. (2006). Probabilités, Analyse des données et Statistique. Technip.

Annexe A : Logiciel R

2.5 Qu'est-ce-que le langage R ?

• Le langage R est un langage de programmation et un environnement mathématique utilisés pour le traitement de données. Il permet de faire des analyses statistiques aussi bien simples que complexes comme des modèles linéaires ou non-linéaires, des tests d'hypothèse, de la modélisation de séries chronologiques, de la classification, etc. Il dispose également de nombreuses fonctions graphiques très utiles et de qualité professionnelle.

• R a été créé par Ross Ihaka et Robert Gentleman en 1993 à l'Université d'Auckland, Nouvelle Zélande, et est maintenant développé par la R Developement Core Team.

L'origine du nom du langage provient, d'une part, des initiales des prénoms des deux auteurs (Ross Ihaka et Robert Gentleman) et, d'autre part, d'un jeu de mots sur le nom du langage S auquel il est apparenté.

Annexe B : Abréviations et Notations

Les différentes abréviations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous :

$E(.)$	:	Espérence mathématique.
$cov(.,.)$	:	Covariance.
$V(.)$	:	Variance.
$v.a$	:	Variable aléatoire.
$i.i.d$	:	Indépendantes et identiquement distribuées.
$1\mathrm{I}_a$	:	Indicatrice de a.
$\xrightarrow{\mathcal{L}}$	:	Convergence en loi.
$\xrightarrow{\mathcal{P}}$	:	Convergence en probabilité.
$\xrightarrow{p.s}$	:	Convergence presque sur.
$N(\mu, \sigma^2)$	:	Loi normal de moyenne μ et de variance σ^2.
$\mathcal{X}^2$	:	Lio de Khi deux.

Résumé

Ce mémoire aborde certains concepts fondamentaux dans les lois usuelles et les modes de convergence. Nous avons également discuté des méthodes d'estimation de la moyenne et de la variance en utilisant les intervalles de confiance. Enfin, nous avons conclu ce travail par une simulation en langage avec R.

ملخص

تتناول هذه المذكرة بعض المفاهيم الأساسية في القوانين الاعتيادية وطرق التقارب. كما تطرقنا أيضًا إلى طرق تقدير المتوسط والتباين باستخدام مجالات الثقة. وأخيرًا، أنهينا هذا العمل بمحاكاة بلغة البرمجة R.

Abstract

This thesis addresses some fundamental concepts in common distributions and modes of convergence. We also discussed methods for estimating the mean and variance using confidence intervals. Finally, we concluded this work with a simulation in the R programming language.