Université Mohamed El Bachir El Ibrahimi de Bordj Bou Arréridj
Faculté des Mathématiques et de l'Informatique
Département des Mathématiques

Mémoire

Présenté par

BOUKHARI Nor El Imane
ZOUAOUI Imane

Pour l'obtention du diplôme de

Master

Filière : Mathématiques
Spécialité : Systéme Dynamique

Thème

LES METHODES STATISTIQUES ET APPLICATIONS

Soutenu publiquement le 06 juillet 2021 devant le jury composé de

SIDHOUM Karima	Président
BELKACEM Naziheddine	Encadrant
DEKKAR Khedra	Examinateur

Promotion 2020/2021

Dédicaces

Je dédie les lettres de mon mémo à :

L'amaour de mon cœur et mon soutien dans la vie, mon père "Said".

Chérie de l'âme et mes souhaits dans la vie, ma mère "Fatiha"

Mon âme sœur et un morceau de mon cœur, ma sœur "Nour El Houda"

Mon jumeau et ma consolation dans la vie, mon frère "Mohammed Seyf Eddine"

Mon deuxième frère et ma bénédiction dans la vie "BAKHOUCHE Hamza"

A toute la famille "BOUKHARI " et " GUEMMOUR ".

Petite amie de ma vie et mon cher "Benzertiha Nour El Houda"

Le frère que ma mère n'a pas mis au monde "HIRECHE Bachir"

Les personnes les plus chéres à moi "ZROUTI Othmane" et sa femme "KHICHANE Djamila", et leur fils "ZROUTI Hamza", "BENRABAH Larbi" et sa femme "KHICHANE Noura"

Mes meilleurs amis "Narimane, khalida, Rania, Hadjer, Chahira, Sylia, Imane, Zahra, Hiba, Hafidha, Youssra, Madiha, Sara, Amira, Karima, Marwa, Asma, Mouna, Ahlam, Nasrin, Meriem,.. "

Mes amies "CHERIET Selman, LOUAIL Amine, Remili Abd Aldjalil. "

BOUKHARI Nor El Imane

Dédicaces

Je dédie les lettres de mon mémo à :

L'amaour de mon cœur et mon soutien dans la vie, mon père "Aziz".

Chérie de l'âme et mes souhaits dans la vie, ma mère "Salma"

Mon âme sœur, mon cher mari "Mounir"

Ma seule et la douceur de mes yeux, ma chére fille "Iline"

Mes chers frères "Oussama, Rabah, Ahmed Amine."

A toute la famille "ZOUAOUI ", " NASRI " et "CHOUTRI "

Mes meilleurs amis "BOUKHARI Nor El Imane, AMROUCHE Meriem, KHALFA Soumia, TOUTI Manel. "

ZOUAOUI Imane

Remerciements

*N*ous tenons à remercier en tout premier dieu **Allah** tout puissant de nous avoir donner la volonté et la puissance pour élaborer ce travail.

*N*ous adressons nos profonds remerciements à notre promoteur **BELKACEM Nazih Eddine** pour ses encouragements ses conseils et pour avoir mis à notre disposition tous les moyens dont nous avions besoin .

*T*enons également à remercier les membres de jury qui ont bien voulu accepter de porter leur jugement sur ce modeste travail que nous souhaitons à la mesure de leur satisfaction .

*N*ous voudrions exprimer nous plus vifs remerciements à tous nos professeurs qui ont contribué à nous transmettre l'inestimable trésor qui est le savoir .

*N*os remerciement s'étendent aussi à monsieur **BENSAID Fares** et madame **BENTERKI Rbiha** pour l'aide qu'il nous a apportée dans notre travail .

TABLE DES MATIÈRES

INTRODUCTION

La statistique est la science dont l'objet est de recueillir, de traiter et d'analyser des données issues de l'observation de phénomènes aléatoires, c'est-à-dire dans lesquels le hasard intervient. L'analyse des données est utilisée pour décrire les phénomènes étudiés, faire des prévisions et prendre des décisions à leur sujet. En cela, la statistique est un outil essentiel pour la compréhension et la gestion des phénomènes complexes. Les méthodes statistiques se répartissent en deux classes :

• La statistique descriptive, statistique exploratoire ou analyse des données, a pour but de résumer l'information contenue dans les données de façon synthétique et efficace. Elle utilise pour cela des représentations de données sous forme de graphiques, de tableaux et d'indicateurs numériques (par exemple des moyennes). Elle permet d'étudier et de suggérer des hypothèses pour une étude ultérieure plus sophistiquée. Les probabilités n'ont ici qu'un rôle mineur.

• La statistique inférentielle va au delà de la simple description des données. Elle a pour but de faire des prévisions et de prendre des décisions au vu des observations. En général, il faut pour cela proposer des modèles probabilistes du phénomène aléatoire étudié et savoir gérer les risques d'erreurs. Les probabilités jouent ici un rôle fondamental. L'informatique et la statistique sont deux éléments du traitement de l'information : l'informatique acquiert et traite l'information tandis que la statistique l'analyse. Les deux disciplines sont donc étroitement liées. Étant donne $x_1, x_2, ..., x_n$ les réalisations de variables aléatoires $X_1, X_2, ..., X_n$ indépendantes et de même lois(généralement dépend d'un paramètre inconnu θ), les techniques de statistiques descriptives permettent de faire les hypothèses sur la nature de la lois de probabilité des $X_1, X_2, ..., X_n$ Les tests d'ajustement permettent de valider ou pas ces hypothèses. Notre travail a pour but de présenter les principes de base d'une analyse statistique de données (description, estimation, tests), ainsi que les méthodes statistiques les plus usuelles. Ces méthodes

seront toujours illustrées par des problèmes concrets. Notre travail est structuré comme suit :

Le premier chapitre présente les notions de base qui est l'estimation et ses différentes variantes, cette notion a une importance capitale en statistiques, en premier lieu on rappelé les définitions et les lois classiques et le principe de l'estimation, puis on décrit en détails deux types fameux d'estimation : l'estimation ponctuelle et l'estimation par l'intervalle de confiance. Des exemples numériques sont introduits pour illustrer les différentes méthodes d'estimation.

Le chapitre suivant introduit deux types de tests statistiques, premièrement on présente le principe, de test d'hypothèse et l'erreur d'un test statistique, ensuite deux types de test statistique sont explorés : les tests paramétriques et les tests non paramétriques.

Le dernier chapitre est consacre aux applications base sur un exemple extrait de l'article de Boitsov et Guzeva, les résultats obtenues dans cet articles sont Validées par une simulation sous Mathematica. Dans l'étude de cet article on applique la méthode d'estimation de maximum de vraisemblance et les tests statiques élaborés dans les chapitres précédents.

CHAPITRE 1

ESTIMATION

1.1 Estimation

1.1.1 Définition

L'estimation est le procédé par lequel on détermine les valeurs inconnues des paramètres de la population à partir des données de l'échantillon. Pour cela, on utilise des distributions théoriques, c'est à dire des variables aléatoires dont on connait les lois de probabilité.

1.1.2 Rappel sur lois classiques utilisées

Loi normale ou loi de Gauss

Une variable aléatoire réelle X suit une loi normale (ou loi gaussienne, loi de Laplace-Gauss) d'espérance μ et d'écart type σ (nombre strictement positif, car il s'agit de la racine carrée de la variance σ^2) si cette variable aléatoire réelle X admet pour densité de probabilité la fonction $p(x)$ définie, pour tout nombre réel x, par :

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}.$$

Une telle variable aléatoire est alors dite variable gaussienne.

Une loi normale sera matée de la manière suivante $\mathcal{N}(\mu, \sigma)$ car elle dépend de deux paramètres μ (la moyenne) et σ (l'écart-type. Ainsi si une variabile aléatoire X suit $\mathcal{N}(\mu, \sigma)$ alors :

$$E(X) = \mu \quad \text{et} \quad V(X) = \sigma^2.$$

Lorsque la moyenne μ vaut 0, et l'écart-type vaut 1, la loi sera notée $\mathcal{N}(0,1)$ et sera appelée loi normale standard. Sa fonction caractéristique vaut $e^{\frac{-t^2}{2}}$. Seule la loi $\mathcal{N}(0,1)$ est tabulée car les autres lois (c'est à-dire avec d'autres paramètres) se déduise de celle-ci à l'aide du théorème suivant : Si Y suit $\mathcal{N}(\mu,\sigma)$ alors $Z = \frac{Y-\mu}{\sigma}$ suit $\mathcal{N}(0,1)$.

On note Φ la fonction de répartition de la loi normale centrée réduite :

$$\Phi(x) = P(Z < x).$$

avec Z une variable aléatoire suivant $\mathcal{N}(0,1)$.

Loi log normale

Définition 1.1

Une variable aléatoire réelle X suit une loi log-normale si elle admet la densité

$$f(x) = \begin{cases} 0 & si \quad t < 0 \\ \frac{1}{\sigma t\sqrt{2\pi}} exp\left(-\frac{(\ln t - m)^2}{2\sigma^2}\right) & si \quad t \geq 0 \end{cases}$$

ou $m \in \mathbb{R}$, $\sigma \in \mathbb{R}^*$.

Cette loi est l'analogue multiplicatif de la loi normale : elle modélise les effets multiplicatifs de phénomènes aléatoires nombreux et indépendants.

Loi du χ^2 (Khi deux)

Définition 1.2

Soit $Z_1, Z_2, ..., Z_\nu$ une suite de variables aléatoires indépendantes de même loi $\mathcal{N}(0,1)$. Alors la variable aléatoire $\sum_{i=1}^{\nu} Z_i^2$ suit une loi appelée loi du Khi-deux à ν degrés de liberté, notée $\chi^2(\nu)$.

Proposition 1.1

❶ Sa fonction caractéristique est $(1 - 2it)^{-\nu/2}$.

❷ La densité de la loi du $\chi^2(\nu)$ est :

$$f(x) = \begin{cases} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2} & \textit{pour} \quad x > 0 \\ 0 & \textit{sinon.} \end{cases}$$

où est la fonction Gamma d'Euler définie par $\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx.$

❸ L'espérance de la loi du $\chi^2(\nu)$ est égale au nombre ν de degrés de liberté et sa variance est 2ν.

❹ La somme de deux variables aléatoires indépendantes suivant respectivement $\chi^2(\nu_1)$ et $\chi^2(\nu_2)$ suit aussi une loi du χ^2 avec $\nu_1 + \nu_2$ degrés de liberté.

Loi de Student

Définition 1.3

Soient Z et Q deux variables aléatoires indépendantes telles que Z suit $\mathcal{N}(0,1)$ et Q suit $\chi^2(\nu)$. Alors la variable aléatoire :

$$T = \frac{Z}{\sqrt{Q/\nu}}.$$

suit une loi appelée loi de Student à ν degrés de liberté, notée $St(\nu)$.

Proposition 1.2

❶ La densité de la loi de Student à ν degrés de liberté est :

$$f(x) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\nu/2)} \frac{1}{(1 + x^2/\nu)^{\frac{\nu+1}{2}}}.$$

❷ L'espérance n'est pas définie pour $\nu = 1$ et vaut 0 si $\nu \geqslant 2$. Sa variance n'existe pas pour $\nu \leqslant 2$ et vaut $\nu/(\nu - 2)$ pour $\nu \geqslant 3$.

❸ La loi de Student converge en loi vers la loi normale centrée réduite.

Remarque 1.1

Pour $\nu = 1$, la loi de Student s'appelle loi de Cauchy, ou loi de Lorentz.

Loi de Fisher-Snedecor

Définition 1.4

Soient Q_1 et Q_2 deux variables aléatoires indépendantes telles que Q_1 suit $\chi^2(\nu_1)$ et Q_2 suit $\chi^2(\nu_2)$ alors la variable aléatoire :

$$F = \frac{Q_1/\nu_1}{Q_2/\nu_2}.$$

suit une loi de Fisher-Snedecor à (ν_1, ν_2) degrés de liberté, noter $F(\nu_1, \nu_2)$.

Proposition 1.3

La densité de la loi $F(\nu_1, \nu_2)$ est :

$$f(x) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \frac{x^{\nu_1/2-1}}{(1+\frac{\nu_1}{\nu_2}x)^{\frac{\nu_1+\nu_2}{2}}} \quad si \quad x > 0 \quad (0\ sinon).$$

Son espérance n'existe que si $\nu_2 \geq 3$ et vaut $\frac{\nu_2}{\nu_2-2}$. Sa variance n'existe que si $\nu_2 \geq 5$ et vaut $2\nu_2^2(\nu_1+\nu_2-2)/\nu_1(\nu_2-2)^2(\nu_2-4)$.

Proposition 1.4

- ➢ Si F suit une loi de Fisher $F(\nu_1, \nu_2)$ alors $\frac{1}{F}$ suit une loi de Fisher $F(\nu_1, \nu_2)$.
- ➢ Si T suit une loi de Student à ν degrés de liberté alors T^2 suit une loi de Fisher $F(1, \nu)$.

♦ Théorème limite central

Théorème 1.1 *(Théorème de la limite centrée)*

Soit $(\Omega, \mathcal{A}, P)$ un espace probabilisé. Soit (X_n) une suite de variables aléatoires indépendantes, de même loi, de classe L^2, de moyenne E et d'écart-type σ. Soit :

$$S_n = \frac{\sum_{j=1}^{n} X_j - nE}{\sqrt{n}\sigma}.$$

Pour tout couple (a, b) avec $a, b \in [-\infty, \infty]$ nous avons :

$$P(a < S_n \leq b) n \to \infty \to \frac{1}{\sqrt{2\pi}} \int_a^b e^{\frac{-x^2}{2}} dx.$$

Remarque 1.2

Ce théorème généralise le théorème de Moivre-Laplace car une variable aléatoire S_n, de loi $B(n, p)$ a même loi qu'une somme $X_1 + ... + X_n$, de n variables aléatoires de Bernouilli indé-

pendantes de paramètre p. Plus généralement, si $X = \sum_{j=1}^{n} X_j$ où les (X_j) sont indépendantes de même loi L^2 alors on peut approximer X par $\mathcal{N}(E(X), Var(X))$.

1.1.3 Principe de l'estimation

On s'intéresse à un phénomène aléatoire et à une VAR X qui lui est associée.

❶ Dans la pratique, la loi de X est inconnue mais on sait souvent qu'elle appartient à une famille de lois connues. Ainsi seul ou plus paramètres $\theta \epsilon \Theta$ de la loi de la variable X est inconnu.

❷ On répète l'expérience plusieurs fois en observant pour chacun de ces répétitions la valeur prise par la variable X. Un lot de tels résultats est appelé un échantillon observé. De manière générale, supposons que l'on possède a réalisations de la variable X (échantillon de taille n), on désigne par X_i, la variable aléatoire donnant la valeur obtenue lors de la i-ème réalisation :

➢ la variable X_i, suit la même loi que la variable X.

➢ les variables $X_1, X_2, ..., X_n$ sont indépendantes.

❸ Le but de l'estimation est de déterminer une valeur approchée du paramètre $\theta \epsilon \Theta$ à partir des résultats observées dans l'échantillon aléatoire. Cette valeur approchée dépendante des valeurs observées est une variable aléatoire appelée estimateur de θ.

❹ Afin obtenir une valeur approchée du paramètre $\theta \epsilon \Theta$, il peut exister plusieurs méthodes d'estimation. Pour les comparer on introduit deux "mesures de qualité" d'une méthode d'estimation appelées : biais d'estimation et le risque quadratique.

❺ Un problème d'importance subsiste dans cette technique par estimation ponctuelle : on ne connait pas la précision des valeurs par lesquelles on approche le paramètre recherché. Cela motive le recours à une autre méthode : l'estimation par intervalle de confiance. Il s'agit alors de fournir un encadrement, auquel le paramètre appartient avec une probabilité maitrisée.

1.2 Estimation ponctuelle

1.2.1 Définition

On cherche à estimer une valeur θ inconnue liée à un certain phénomène aléatoire, en général, la moyenne μ ou la variance σ^2 ou encore l'écart-type σ de la loi du phénomène. Pour ce

faire, on dispose d'observations indépendantes du phénomènes, c-à-d de variables aléatoires $X_1, ..., X_n$, indépendantes et de même loi (celle du phénomène). On parle d'un échantillon. On définit à partir de l'échantillon une nouvelle variable aléatoire notée T dont les valeurs seront proches de celle de la grandeur θ à estimer. Cette nouvelle variable aléatoire T sera appelée estimateur de θ. Il peut y avoir plusieurs estimateurs pour une même grandeur θ , certains meilleurs que d'autres.

1.2.2 Les propriétés des estimation

Le biais

Le biais d'un estimateur $\widehat{\theta}$ du paramètre θ est :

$$Biais(\widehat{\theta}) = E(\widehat{\theta}) - \theta.$$

On dit que $\widehat{\theta}$ est sans biais ou non-biaisé si Biais($\widehat{\theta}$) = 0. Le biais est une mesure de l'erreur systématique faite en approximant θ par $\widehat{\theta}$.

Erreur quadratique moyenne

Définition 1.5

L'erreur quadratique moyenne (EQM) d'un estimateur Erreur $\widehat{\theta}$ du paramètre θ est

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

.

L'EQM est une mesure de la précision d'un estimateur.

Théorème 1.2

Si $\widehat{\theta}$ est un estimateur du paramètre θ alors :

$$EQM(\widehat{\theta}) = V(\widehat{\theta}) + \left[Biais(\widehat{\theta})\right]^2.$$

Le meilleur de deux estimateur $\widehat{\theta_1}$ et $\widehat{\theta_2}$, c'est-à-dire le plus efficace, est celui qui a la plus petite EQM : $\widehat{\theta_1}$ est plus efficace que $\widehat{\theta_2}$ si :

$$EQM(\widehat{\theta_1}) < EQM(\widehat{\theta_2}) \Longleftrightarrow \frac{EQM(\widehat{\theta_1})}{EQM(\widehat{\theta_2})} < 1.$$

Lorsque deux estimateurs son non biaisés, ceci revient à dire que le plus efficace est celui dont la variance est la plus petite.

Convergence

Dénotons par $\widehat{\theta}_n$, un estimateur du paramètre θ calculé à partir d'un échantillon de taille n.

Définition 1.6

Un estimateur $\widehat{\theta}_n$, est convergent si pour tout $\varepsilon > 0$.

$$\lim_{n\to\infty} P\left(|\widehat{\theta}_n - \theta| < \varepsilon\right) = 1.$$

Ceci signifie : si la taille de l'échantillon est assez grande alors on est (presque) certain que l'estimateur $\widehat{\theta}_n$, est très proche de θ.

1.2.3 La méthode du maximum de vraisemblance

Soit X une variable aléatoire réelle de loi paramétrique (discrète ou continue), dont on veut estimer le paramètre θ. Alors on définit une fonction f telle que :

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

Définition 1.7

On appelle fonction de vraisemblance de θ pour une réalisation $(x_1, x_2, ..., x_n)$ d'un échantillon, la fonction de θ :

$$L(x_1, ..., x_n; \theta) = f(x_1, ..., x_n; \theta) = \Pi_{i=1}^{n} f(x_i; \theta).$$

Définition 1.8

La méthode consistant à estimer θ par la valeur qui maximise L (vraisemblance) s'appelle méthode du maximum de vraisemblance

$$\widehat{\theta} = \{\theta / L(\widehat{\theta}) = \sup_{\theta} L(\theta)\}.$$

Ceci est un problème d'optimisation. On utilise généralement le fait que si L est dérivable et si L admet un maximum global en une valeur, alors la dérivée première s'annule en et que la dérivée seconde est négative.

Réciproquement, si la dérivée première s'annule en $\theta = \widehat{\theta}$ et que la dérivée seconde est négative en $\theta = \widehat{\theta}$, alors $\widehat{\theta}$ est un maximum local (et non global) de $L(x_1,...,x_i,...,x_n;\theta)$. Il est alors nécessaire de vérifier qu'il s'agit bien d'un maximum global. La vraisemblance étant positive et le logarithme népérien une fonction croissante, il est équivalent et souvent plus simple de maximiser le logarithme népérien de la vraisemblance (le produit se transforme en somme, ce qui est plus simple à dériver).

Ainsi en pratique

❶ La condition nécessaire

$$\frac{\partial L(x_1,...,x_i,...,x_n;\theta)}{\partial\theta} = 0 \qquad ou \qquad \frac{\partial \ln L(x_1,...,x_i,...,x_n;\theta)}{\partial\theta} = 0$$

permet de trouver la valeur $\widehat{\theta}$.

❷ $\theta = \widehat{\theta}$ est un maximum local si la condition suffisante est remplie au point critique :

$$\frac{\partial L(x_1,...,x_i,...,x_n;\theta)}{\partial\theta^2}(\widehat{\theta}) \leq 0 \qquad ou \qquad \frac{\partial \ln L(x_1,...,x_i,...,x_n;\theta)}{\partial\theta^2}(\widehat{\theta}) \leq 0.$$

Exemple 1 *(Avec une loi discrète)*

On souhaite estimer le paramètre λ d'une loi de Poisson à partir d'un n-échantillon. On a $f(x;\lambda) = P_\lambda(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}$. La fonction de vraisemblance s'écrit :

$$L(x_1,...,x_n;\lambda) = \Pi_{i=1}^n e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} = e^{-\lambda n}\Pi_{i=1}^n\frac{\lambda^{x_i}}{x_i!}.$$

Il est plus simple d'utiliser le logarithme, la vraisemblance étant positive :

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

La dérivée première :

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

s'annule pour $\widehat{\lambda} = \frac{\sum_{i=1}^n x_i}{n}$.

La dérivée seconde :

$$\frac{\partial^2 \ln L(x_1,...,x_n;\theta)}{\partial \theta^2} = -\frac{\sum_{i=1}^{n} x_i}{\lambda^2},$$

est toujours négative ou nulle. Ainsi l'estimation donnée par $\Lambda = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}$ conduit à un estimateur du maximum de vraisemblance égal à $\widehat{\lambda} = \overline{x}$. Il est normal de retrouver la moyenne empirique qui est le meilleur estimateur possible pour le paramètre λ (qui représente aussi l'espérance d'une loi de Poisson).

Exemple 2 *(Avec une loi continue)*

On souhaite estimer les paramètres μ et σ d'une loi normale à partir d'un n-échantillon. La loi normale $\mathcal{N}(\mu,\sigma)$ a pour fonction densité :

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

Écrivons la fonction de vraisemblance pour une réalisation d'un échantillon de n variables indépendantes :

$$f(x_1,...,x_n;\mu,\sigma) = \Pi_{i=1}^{n} f(x_i;\mu,\sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right).$$

Or $\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n}(x_i-\overline{x}+\overline{x}-\mu)^2 = \sum_{i=1}^{n}(x_i-\overline{x})^2 + n(\overline{x}-\mu)^2$, où $\overline{x}$ représente la moyenne de l'échantillon. Ainsi la fonction de vraisemblance peut être écrite sous la forme

$$f(x_1,...,x_n;\mu,\sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2 + n(\overline{x}-\mu)^2}{2\sigma^2}\right).$$

$$\frac{\partial}{\partial\mu}\ln L = \frac{\partial}{\partial\mu}\left(\ln\left(\frac{1}{2\pi\sigma^2}\right)^{n/2} - \left(\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2 + n(\overline{x}-\mu)^2}{2\sigma^2}\right)\right) = 0 - \frac{-2n(\overline{x}-\mu)}{2\sigma^2}$$

On obtient donc l'estimateur par le maximum de vraisemblance de l'espérance :

$$\widehat{\mu} = \overline{x} = \sum_{i=1}^{n} x_i/n.$$

Pour le second paramètre, on calcule :

$$\frac{\partial}{\partial\sigma}\ln L=\frac{\partial}{\partial\sigma}\left(\frac{n}{2}\ln\left(\frac{1}{2\pi\sigma^2}\right)-\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2+n(\overline{x}-\mu)^2}{2\sigma^2}\right)=-\frac{n}{\sigma}+\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2+n(\overline{x}-\mu)^2}{\sigma^3}$$

Donc :

$$\widehat{\sigma}^2=\sum_{i=1}^{n}(x_i-\widehat{\mu})^2/n.$$

que l'on peut traduire par :

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

On vérifie que c'est bien des maxima locaux :

$$\frac{\partial^2\ln L}{\partial\mu^2}=-n/\sigma^2\leq 0,$$

$$\frac{\partial^2\ln L}{\partial\sigma^2}=n/\sigma^2-\frac{3}{\sigma^4}\left(\sum(x_i-\overline{x})^2+n(\overline{x}-\mu)^2\right).$$

Au point $\widehat{\sigma}$,

$$\frac{\partial^2\ln L}{\partial\sigma^2}(\widehat{\sigma})=n/\widehat{\sigma}^2-\frac{3}{\sigma^4}(n\widehat{\sigma}^2+n(\overline{x}-\mu)^2)\leq 0.$$

La méthode fournit un estimateur non biaisé de la moyenne $(E(\widehat{\mu})=\mu)$ mais par contre, l'estimateur de la variance est biaisé $(E(\widehat{\sigma}^2)=\frac{n}{n-1}\sigma^2)$. Néanmoins l'estimateur est asymptotiquement sans biais.

1.2.4 La méthode des moments

Soit X une variable aléatoire continue ayant la fonction de densité $f(x;\theta_1,\theta_2,...,\theta_k)$ ou une variable aléatoire discrète ayant la fonction de masse $p(x;\theta_1,\theta_2,...,\theta_k)$, cette variable se caractérisant par k paramètres inconnus. Si $X_1,X_2,...,X_n$, forment un échantillon aléatoire de taille n des valeurs prises par X, on peut définir comme suit les k premiers moments de cet échantillon par rapport à l'origine :

$$m_t^{'}=\frac{1}{n}\sum_{i=1}^{n}X_i^t,\qquad t=1,2,...,k. \tag{1.1}$$

Les k premiers moments de la population par rapport à l'origine se traduisent, quant à eux, par :

$$\mu_t^{'} = E(X^t) = \begin{cases} \int_{-\infty}^{\infty} x^t f(x;\theta_1,\theta_2,...,\theta_k)dx, & t = 1,2,...,k, \quad \text{si } X \text{ est continue,} \\ \\ \sum_{x \in R_X} x^t p(x;\theta_1,\theta_2,...,\theta_k), & t = 1,2,...,k, \quad \text{si } X \text{ est discrète.} \end{cases} \tag{1.2}$$

Les moments $\{\mu_t^{'}\}$ de la population sont en général des fonctions des k paramètres inconnus $\{\theta_i\}$. En faisant correspondre les moments de l'échantillon à ceux de la population, on equations simultanées à k inconnues (les paramètres θ_i). En d'autres termes,

$$\mu_t^{'} = m_t^{'}, \qquad t = 1,2,...,k. \tag{1.3}$$

La solution de l'équation (1.3), notée $\widehat{\theta}_1, \widehat{\theta}_2, ..., \widehat{\theta}_p$, fournit les estimateurs de moment de $\theta_1, \theta_2, ..., \theta_k$.

Exemple 3

Soit X une variable uniforme sur l'intervalle $[0, a]$. On veut trouver un estimateur de a par la méthode des moments. Or, le premier moment de la population par rapport à zéro est :

$$\mu_1^{'} = \int_0^a x \frac{1}{a} dx = \frac{a}{2}.$$

Quant au premier moment de l'échantillon, il s'agit tout simplement de $\overline{X}$. Par conséquent,

$$\widehat{a} = 2\overline{X}$$

l'estimateur de moment pour le paramètre a n'étant autre que le double de la moyenne de l'échantillon. Dans cet exemple, l'estimateur obtenu ne génère pas une valeur estimée conforme à ce que l'on connaît de la situation. Ainsi, dans le cas d'un échantillon formé des observations $x_1 = 60$, $x_2 = 10$ et $x_3 = 5$, $\widehat{a}$ égalerait 50, une valeur inacceptable puisqu'on sait qu'une des valeurs observées vaut 60, ce qui implique que $a \geqslant 60$.

La méthode des moments fournit souvent d'assez bons estimateurs. En règle générale, les estimateurs de moment présentent une distribution asymptotique normale et se révèlent convergents. Leur variance peut toutefois être plus grande que celle des estimateurs déterminés d'autres méthodes comme celle du maximum de vraisemblance. Les estimateurs fournis par la méthode des moments peuvent aussi, à l'occasion, laisser à désirer, comme c'est le cas à l'exemple 3.

1.3 Estimation par intervalle de confiance

1.3.1 Définition

On ne cherche plus à donner une valeur estimée la meilleure possible du paramètre x (moyenne, proportion, écart-type...) mais un intervalle de valeurs dans lequel la vraie valeur se trouve avec une probabilité donnée (le coefficient de confiance ; dans la pratique, **95%, 99%**...). Si on écrit le coefficient de confiance sous la forme $1-\alpha$, α est appelé le "risque" **(5%, 1%, ...)**. On cherche donc $[a,b]$ tel que :

$$p(x \in [a,b]) = 1-\alpha.$$

1.3.2 Estimation d'une moyenne

La variance σ^2 est supposée connue

La variable aléatoire parente X suit une loi de probabilité de paramètre $m = E(X)$ inconnu et de variance σ^2 connue. Soit $X_1, X_2, ..., X_n$ un $n-$échantillon aléatoire simple de X. On sait alors qu'un bon estimateur ponctuel de m est $\overline{X}$(estimateur sans biais, convergent et efficace) et que :

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n} \rightsquigarrow \mathcal{N}(m, \frac{\sigma}{\sqrt{n}}),$$

et

$$Z = \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \rightsquigarrow \mathcal{N}(0,1).$$

Les tables fournissent la valeur Z_α, pour α donné, telle que $p(\{-Z_\alpha < Z < Z_\alpha\}) = 1-\alpha$. Or

$$-Z_\alpha < Z < Z_\alpha \Longleftrightarrow -Z_\alpha < \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} < Z_\alpha$$

$$\Longleftrightarrow -Z_\alpha \frac{\sigma}{\sqrt{n}} < \overline{X} - m < Z_\sigma \frac{\sigma}{\sqrt{n}} \Longleftrightarrow -Z_\alpha \frac{\sigma}{\sqrt{n}} < m - \overline{X} < Z_\sigma \frac{\sigma}{\sqrt{n}}$$

On obtient ainsi $p(\overline{X} - Z_\alpha \frac{\sigma}{\sqrt{n}} < m < \overline{X} + Z_\alpha \frac{\sigma}{\sqrt{n}}\}) = 1-\alpha$ c'est-à-dire un intervalle de confiance de m au niveau de confiance $1-\alpha$ soit $\left[\overline{X} - Z_\alpha \frac{\sigma}{\sqrt{n}}; \overline{X} + Z_\alpha \frac{\sigma}{\sqrt{n}}\right]$. Dans la pratique, on dispose d'un échantillon non exhaustif tiré au hasard de la population. Cet échantillon fournit une réalisation de par le calcul de la moyenne $\overline{X}$. Ainsi l'échantillon donne une réalisation de l'intervalle de confiance au risque α qui est $\left[\overline{X} - Z_\alpha \frac{\sigma}{\sqrt{n}}; \overline{X} + Z_\alpha \frac{\sigma}{\sqrt{n}}\right]$.

Exemple 4

Un article paru dans le $\boldsymbol{Journal\ of\ Heat\ Transfer}$ décrit une façon de mesurer la conductivité thermique du fer Armco. Les **10** mesures de cette conductivité (en $\boldsymbol{BTU/h-pi-F}$), énumérées ci-après, ont été obtenues à une température de $\boldsymbol{100F}$ avec une puis sance fournie de $\boldsymbol{550W}$.

$$\mathbf{41.60, 41.48, 42.34, 41.95, 41.86,}$$

$$\mathbf{42.18, 41.72, 42.26, 41.81, 42.04.}$$

On veut établir un intervalle de confiance à **95%** pour la conductivité thermique moyenne du fer Armco dans ces conditions, sachant que la moyenne de l'échantillon est de $\boldsymbol{41,924\ BTU/h-pi-F}$ et que l'écart-type σ est de $\boldsymbol{0,10\quad BTU/h-pi-F}$. Si l'on suppose que la conductivité thermique obéit à une loi normale. Dans le cas d'un intervalle à **95%**, $\boldsymbol{1-\alpha=0.95}$, de sorte que $\boldsymbol{\alpha=0.05}$. Or, $\boldsymbol{z_{\alpha/2}=z_{0.05/2}=z_{0,025}=1.96}$ selon la table II de l'annexe. On a ainsi l'intervalle de confiance :

$$\bar{x}\pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}=\mathbf{41.924\pm 1.96(0.10)/\sqrt{10}=41.924\pm 0.062.}$$

Par conséquent, l'intervalle de confiance bilatéral de niveau **95%** est :

$$\mathbf{41.862\leq \mu \leq 41.986.}$$

Il s'agit là d'une plage de valeurs raisonnables pour la moyenne à un degré de confiance de **95%**.

La variance est inconnue

Dans la pratique, si l'espérance $\boldsymbol{m=E(X)}$ est inconnue, a fortiori, la variance $\boldsymbol{\sigma^2=E[(X-m)^2]}$ est également inconnue. Or nous venons de voir que l'intervalle de confiance de $\boldsymbol{m}$ tel qu'il vient d'être défini dépend de σ. Il est alors tentant de remplacer σ par son estimation ponctuelle s fournie par l'estimateur $\boldsymbol{S^2}$. Ce nombre n'est autre que l'écart-type calculé sur l'échantillon de taille $\boldsymbol{n}$ avec $\boldsymbol{n-1}$ degrés de liberté (ddl). Dans ces conditions, on utilise le procédé dit de Studentisation qui consiste à remplacer la variable centrée réduite $\boldsymbol{Z=\frac{\overline{X}-E(\overline{X})}{\sigma(\overline{X})}}$ par la variable $\boldsymbol{T=\frac{\overline{X}-E(\overline{X})}{\frac{S}{\sqrt{n}}}}$ qui suit une loi de Student à $\boldsymbol{n-1}$ ddl. La table de Student nous permet de déterminer le-ta tel que pour $\boldsymbol{n-1}$ ddl. La table de Student nous permet de déter-

miner $t_{n-1,\alpha}$ tel que $n-1$ ddl on ait :

$$p(-t_{n-1,\alpha} \leq T \leq t_{n-1,\alpha}) = 1-\alpha$$

On obtiendra alors l'intervalle de confiance au risque α :

$$\left[\overline{X} - t_{n-1,\alpha}\frac{S}{\sqrt{n}}; \overline{X} + t_{n-1,\alpha}\frac{S}{\sqrt{n}}\right].$$

dont une réalisation sur l'échantillon est $\left[\overline{X} - t_{n-1,\alpha}\frac{S}{\sqrt{n}}; \overline{X} + t_{n-1,\alpha}\frac{S}{\sqrt{n}}\right]$.

Exemple 5

Selon un article paru dans le ***Journal of Testing and Evaluation***, des essais effectués sur des échantillons traités de vêtements de nuit pour enfants ont permis d'obtenir les **20** mesures ci-après du temps de propagation des flammes (en secondes) :

$$\mathbf{9.85, 9.93, 9.75, 9.77, 9.67,}$$

$$\mathbf{9.87, 9.67, 9.94, 9.85, 9.75,}$$

$$\mathbf{9.83, 9.92, 9.74, 9.99, 9.88,}$$

$$\mathbf{9.95, 9.95, 9.93, 9.92, 9.89.}$$

On veut construire un intervalle de confiance à **95%** pour le temps de propagation moyen. Voici la moyenne et l'écart-type de l'échantillon :

$$\overline{x} = \mathbf{9.8475,}$$

$$s = \mathbf{0.0954.}$$

La table IV de l'annexe indique que $t_{0.025;19} = \mathbf{2.093}$. L'intervalle de confiance se traduit ainsi par :

$$\overline{x} \pm t_{\alpha/2;n-1}s/\sqrt{n}$$

$$\mathbf{9.8475 \pm 2.093(0.0954)/\sqrt{20}}$$

$$\mathbf{9.8475 \pm 0.0446.}$$

L'intervalle de confiance à **95%** pour μ est, par conséquent, $[\mathbf{9.8029; 9.8921}]$.

On s'attend, avec un niveau de confiance de **95%**, que le temps de propagation moyen

se situe entre **9.8029** et **9.8921** secondes. Si on répétait cette expérience un grand nombre de fois et qu'on établissait chaque fois l'intervalle de confiance pour μ à **95%** en utilisant cette formule, environ **95%** des intervalles contiendraient la vraie valeur de μ, et **5%** d'entre eux ne la contiendraient pas. Bien sûr, dans les faits, on ne réalise l'expérience qu'une seule fois. On ignore donc si notre unique intervalle contient la vraie moyenne de la population, mais on aura confiance à **95%** que c'est le cas.

Remarque 1.3

Si la taille de l'échantillon est "grande" $(n > 30)$, on peut utiliser la loi normale à la place de la loi de Student. C'est pour cette raison qu'on trouve dans la littérature l'expression :"la loi de Student est la loi des petits échantillons".

1.3.3 Estimation d'une proportion

Dans une population donnée de grande taille, la proportion d'individus p ayant une caractéristique donnée $\mathcal{C}$ est inconnue. On désire déterminer, à partir d'un tirage d'un échantillon non exhaustif de taille n de la population, un intervalle de confiance au risque α de p.

Le tirage de cet échantillon peut être modélisé par un $n-$échantillon au hasard tiré d'une variable aléatoire F qui suit une loi de Bernoulli de paramètre p. Soient donc $X \rightsquigarrow \mathcal{B}(p)$ une loi de Bernoulli de paramètre p et $X_1, X_2, ..., X_n$ un $n-$échantillon aléatoire simple. La fréquence $F = \frac{X_1+X_2+...+X_n}{n}$ est un bon estimateur (estimateur sans biais, convergent et efficace) du paramètre p, où chacune des variables aléatoires X_i, suit une loi de Bernoulli. La fréquence est un estimateur asymptotiquement normal et on utilise l'approximation $F \rightsquigarrow \mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\right)$ pour $n \geq 30, np \geq 5$ et $nq \geq 5$. Ces conditions seront appelées les conditions de normalité.

Les tables statistiques fournissent les valeurs Z_α telles que $p(\{-Z_\alpha < Z < Z_\alpha\}) = 1 - \alpha$ avec $Z \rightsquigarrow \mathcal{N}(0,1)$. On applique cette relation à la variable $Z = \frac{F-p}{\sqrt{\frac{pq}{n}}}$ qui suit une loi normale $\mathcal{N}(0,1)$. On obtient :

$$p\left(\left\{-Z_\alpha < \frac{F-p}{\sqrt{\frac{pq}{n}}} < Z_\alpha\right\}\right) = 1 - \alpha.$$

Remarquons que $-Z_\alpha < \frac{F-p}{\sqrt{\frac{pq}{n}}} < Z_\alpha \Longleftrightarrow -Z_\alpha < \frac{p-F}{\sqrt{\frac{pq}{n}}} < Z_\alpha \Longleftrightarrow F - Z_\alpha\sqrt{\frac{pq}{n}} < p < F + Z_\alpha\sqrt{\frac{pq}{n}}$. On obtient un intervalle de confiance de p au niveau de confiance $1 - \alpha$ soit :

$$\left[F - Z_\alpha\sqrt{\frac{pq}{n}}; F + Z_\alpha\sqrt{\frac{pq}{n}}\right].$$

Pour un risque $\alpha = 5\%$, on trouve $Z_\alpha = 1.96$ et l'intervalle de confiance est :

$$\left[F - 1.96\sqrt{\frac{pq}{n}}; F + 1.96\sqrt{\frac{pq}{n}}\right],$$

pour un risque $\alpha = 1\%$, on trouve $Z_\alpha = 2.58$ et l'intervalle de confiance est :

$$\left[F - 2.58\sqrt{\frac{pq}{n}}; F + 2.58\sqrt{\frac{pq}{n}}\right].$$

Cet intervalle pose un problème pratique important, on peut affirmer que la proportion p appartient à cet intervalle avec une probabilité de $1 - \alpha$ mais les bornes de cet intervalle dépendent de p, la proportion inconnue.

• On remplace p et q par leurs estimations ponctuelles f et $1 - f$. La réalisation de l'intervalle de confiance est alors :

$$\left[f - Z_\alpha\sqrt{\frac{f(1-f)}{n}}; f + Z_\alpha\sqrt{\frac{f(1-f)}{n}}\right].$$

Exemple 6

On a besoin d'estimer rapidement la proportion p d'accidents du travail dans entreprise de construction. On a constaté sur un échantillon de **200** jours ouvrables qu'il y a eu **18** accidents. Déterminer, au risque de **5%**, un intervalle de confiance de la proportion d'accidents. Les données sont $n = 200$ et $f = \frac{18}{200} = 0.09$. Pour un calcul rapide, l'intervalle de confiance numérique est donc, an risque de **5%** :

$$\left[0.09 - 1.96\frac{0.0819}{\sqrt{200}}; 0.09 + 1.96\frac{0.0819}{\sqrt{200}}\right] = [0.07; 0.10].$$

La proportion d'accidents est au risque de **5%** telle que $2\% \leq 15.9\%$. On se rappellera que cette méthode augmente l'amplitude de l'intervalle de confiance. Le calcul fait avec la première méthode donnerait une proportion d'accident p telle que $5\% \leq p \leq 12.99\%$.

Remarque 1.4

❶ On a utilisé l'approximation normale déduite du théorème central limite pour établir l'intervalle de confiance. Il est donc nécessaire que $n > 30$, $np \geq 5$ et $n(1-p) \geq 5$. Dans la pratique, p est inconnue, on vérifie ces conditions sur f donc $n \geq 30, nf \geq 5$ et $n(1-f) \geq 25$.

❷ La longueur de l'intervalle de confiance est $L(\alpha, n) = 2Z_\alpha\sqrt{\frac{f(1-f)}{n}}$.

❸ La précision de l'estimation obtenue est $\frac{1}{2}L(\alpha, n) = Z_\alpha\sqrt{\frac{f(1-f)}{n}}$.

❹ Z_α étant une fonction décroissante de α (risque pris par le statisticien), lorsque $1-\alpha$ augmente, α diminue, Z_α augmente, la longueur de l'intervalle augmente.

❺ Lorsqu'on a choisi la valeur de β, on peut imaginer de déterminer la taille de l'échantillon nécessaire pour atteindre une précision donnée l soit $Z_\alpha\sqrt{\frac{f(1-f)}{n}} < l$. On obtient $n > Z_\alpha^2 \frac{f(1-f)}{l^2}$.

1.3.4 Estimation d'une variance

Théorème 1.3

Soit X une variable aléatoire telle que $X \rightsquigarrow \mathcal{N}(m,\sigma)$ et $X_1, X_2, ..., X_n$ un n-échantillon aléatoire de X. On utilise $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ comme estimateur sans biais et convergent de σ^2. Alors, la variable aléatoire $\frac{n-1}{\sigma^2}S^2$ suit une loi de χ^2 à $n-1$ ddl. On note

$$\frac{n-1}{\sigma^2}S^2 \rightsquigarrow \chi^2$$

.

Un intervalle de confiance au risque α est de la forme $[a;b]$ où a et b sont deux variables aléatoires construites à partir d'un n-échantillon au hasard de X telles que $p(a \leq \sigma^2 \leq b) = 1-\alpha$ or :

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

Posons $U = \frac{(n-1)S^2}{\sigma^2}$, $t_b = \frac{(n-1)S^2}{b}$ et $t_a = \frac{(n-1)S^2}{a}$. Alors,

$$p(a \leq \sigma^2 \leq b) = 1-\alpha \Longleftrightarrow p(t_b \leq U \leq t_a) = 1-\alpha \Longleftrightarrow \left\{p(U < t_b) = \frac{\alpha}{2} \; et \; p(U > t_a) = \frac{\alpha}{2}\right\}$$

.

Comme U suit une loi de χ^2, on détermine les valeurs de t_a et t_b à l'aide d'une table du χ^2 à $n-1$ degrés de liberté. Comme $b = \frac{(n-1)S^2}{t_b}$ et $a = \frac{(n-1)S^2}{t_a}$, l'intervalle de confiance cherché est alors :

$$\left[\frac{(n-1)S^2}{t_a}; \frac{(n-1)S^2}{t_b}\right].$$

Une réalisation de cet intervalle de confiance sur l'échantillon est :

$$\left[\frac{(n-1)s^2}{t_a}; \frac{(n-1)s^2}{t_b}\right]$$

ou s est l'estimation ponctuelle de la variance de la population avec $n-1$ ddl.

Exemple 7

Une variable aléatoire X est distribuée selon une loi normale de paramètres m et σ inconnus.

On dispose d'un n-échantillon associé à X de taille $n = 16$. Sur cet échantillon on observe $\sum_{i=1}^{16}(x_i - \bar{x})^2 = 1500$.

Déterminer un intervalle de confiance de la variance au seuil de confiance de 95%.

La moyenne et la variance de la population sont inconnues. L'intervalle de confiance au seuil 95% de σ^2 est donné par $\left[\frac{(n-1)S^2}{t_a}; \frac{(n-1)S^2}{t_b}\right]$ et sa réalisation sur l'échantillon est donnée numériquement par $\left[\frac{(16-1)s^2}{t_a}; \frac{(16-1)s^2}{t_b}\right]$. On a $s^2 = \frac{1}{15} \times 1500 = 100$. Pour calculer, t_a et t_b on utilise la variable aléatoire U qui suit une loi du χ^2 à $16 - 1 = 15$ ddl. D'après les résultats précédents on a $p(U > t_b) = 0.975$ et $(U < t_a) = 0.025$. La table donne $t_b = 6.26$ et $t_a = 27.5$. On obtient alors l'intervalle de confiance cherché $\left[\frac{1500}{27.5}; \frac{1500}{6.26}\right] = [54.54; 239.60]$.

CHAPITRE 2

TEST D'HYPOTHÈSES

2.1 Principe d'un test statistique

Le principe général de tous les tests statistiques est un raisonnement par l'absurde. En effet, pour déterminer s'il existe une différence d'efficacité entre les deux traitements, on suppose qu'il n'y a pas de différence (cette hypothèse est appelée hypothèse nulle).

2.1.1 L'hypothèse nulle et l'hypothèse alternative

Il existe des centaines de tests statistiques. Chacun de ces tests est associé à une hypothèse nulle que l'on notera $\boldsymbol{H_0}$. Le but d'un test statistique est de démontrer qu'une hypothèse nulle est fausse en la confrontant aux données de notre échantillon. Si les données sont incompatibles avec l'hypothèse nulle, on rejette l'hypothèse nulle. On ne pourra pas par contre démontrer qu'une hypothèse nulle est vraie. L'hypothèse nulle n'est donc pas l'hypothèse d'intérêt ni l'hypothèse scientifique d'une étude. L'hypothèse scientifique d'une étude est l'hypothèse alternative, que l'on notera parfois $\boldsymbol{H_1}$, et qui sera en quelque sorte le contraire de l'hypothèse nulle. Il s'agira donc de formuler l'hypothèse nulle de telle sorte que son rejet implique l'hypothèse alternative.

On démontre une hypothèse alternative en rejetant une hypothèse nulle.

Par exemple, afin de démontrer l'hypothèse alternative (scientifique) suivante :

H_1 : le niveau nutritif a un effet sur la croissance des Onobrychis.

on essaiera de rejeter l'hypothèse nulle :

H_0 : le niveau nutritif n'a aucun effet sur la croissance des Onobrychis.

En d'autres termes, afin de démontrer que le niveau nutritif a un effet sur la croissance des Onobrychis, on essaiera de démontrer qu'il n'est pas possible qu'il n'en ait pas. On recherche en quelque sorte une preuve par l'absurde. On part du contraire de ce que l'on veut démontrer (l'hypothèse nulle) et on essaie d'aboutir à une contradiction entre l'hypothèse nulle et les données, afin de pouvoir conclure ce que l'on veut démontrer (l'hypothèse alternative). Telle est la stratégie d'un test statistique.

2.1.2 Erreurs de première et de seconde espèce

Le résultat d'un test statistique est donc le rejet ou le non-rejet d'une hypothèse nulle H_0. Ceci peut nous mener à deux types d'erreur : rejeter H_0 alors qu'elle est vraie ou ne pas rejeter H_0 alors qu'elle est fausse. On appelle ces erreurs respectivement l'erreur de première espèce et l'erreur de seconde espèce. Notons que lorsque l'on rejette H_0, la seule erreur que l'on peut commettre est une erreur de première espèce, alors que lorsque l'on ne rejette pas H_0, la seule erreur que l'on peut commettre est une erreur de seconde espèce. On peut résumer la situation dans le tableau suivant :

	rejeter H_0	ne pas rejeter H_0
H_0 vraie	erreur de première espèce	bonne décision
H_0 fausse	bonne décision	erreur de seconde espèce

Évidemment, on ne pourra pas commettre une erreur de première espèce si H_0 est fausse, de même que l'on ne pourra pas commettre une erreur de seconde espèce si H_0 est vraie. Si H_0 est vraie, on note par α la probabilité de commettre une erreur de première espèce. Si H_0 est fausse, on note par β la probabilité de commettre une erreur de seconde espèce. On peut résumer la situation dans le tableau suivant :

	probabilité de rejeter H_0	probabilité ne pas rejeter H_0
H_0 vraie	α	$1-\alpha$
H_0 fausse	$1-\beta$	β

Lorsque l'on effectue un test statistique, on pourra choisir la valeur de α, que l'on appellera aussi le seuil du test. On dira que l'on rejette ou que l'on ne rejette pas une hypothèse nulle au seuil α. Évidemment, on aimerait que α soit le plus petit possible. Il y a cependant un conflit

entre α et β. En choisissant α trop petit, on risque d'augmenter considérablement β. Le cas extrême consisterait à choisir $\alpha = 0$, ce qui reviendrait à ne jamais rejeter H_0 (quelles que soient les données observées) et impliquerait alors $\beta = 1$. Il s'agit dès lors d'adopter un compromis, et ce compromis est en général fixé à $\alpha = 5\%$ (selon une convention arbitraire mais souvent raisonnable et largement établie).

La règle de rejet ou de non-rejet d'une hypothèse nulle doit être ainsi définie de façon à ce que si l'hypothèse nulle était vraie (on dira parfois sous l'hypothèse nulle), il y aurait une probabilité de 5% de la rejeter à tort. Cela veut dire que si H_0 était vraie et si on répétait l'expérience (l'échantillonnage) 100 fois, on ne rejetterait H_0 que 5 fois (en moyenne). On aura ainsi :

α = seuil du test = probabilité de rejeter H_0 alors que H_0 est vraie = 5%.

Lorsque l'on effectue un test statistique, on ne choisit pas par contre la valeur de β. On verra au comment on peut calculer β, qui dépendra à la fois de la taille de l'échantillon et du "degré de fausseté" de H_0. La quantité $1 - \beta$ est par ailleurs appelée la puissance du test, qui est donc une mesure de la capacité du test à rejeter à raison une hypothèse nulle qui est fausse.

En résumé, dans un test statistique, on contrôle α mais on ne contrôle pas β , qui pourra être dans certains cas considérablement plus grand que α. En choisissant α petit, on s'assure contre une erreur de première espèce, mais on n'a aucune garantie contre une erreur de seconde espèce. En fait, on considère implicitement qu'il est plus grave de commettre une erreur de première espèce que de commettre une erreur de seconde espèce.

2.2 Tests Paramétriques

Les tests paramétriques se basent sur des distributions statistiques supposées dans les données. Par conséquent, certaines conditions de validité doivent être vérifiées pour que le résultat d'un test paramétrique soit fiable. Par exemple, le test t de Student pour échantillons indépendants n'est fiable que si les données associées à chaque échantillon suivent une distribution normale et si les variances des échantillons sont homogènes.

2.2.1 Tests d'hypothèses sur une moyenne

Le premier test que nous étudions consiste à décider entre les hypothèses suivantes :

$$H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0$$

La question posée est la suivante : les observations remettent-elles en cause l'égalité de la moyenne théorique μ à la valeur spécifiée μ_0 ?

Nous avons déjà répondu à cette question dans le chapitre précédent : l'estimation par intervalle de confiance donne l'ensemble des valeurs possibles de la moyenne théorique μ compte tenu des observations effectuées et du niveau de confiance choisi.

Règle de décision : pour un risque de première espèce α,

- ➢ On accepte l'hypothèse nulle si la valeur μ_0 appartient à l'intervalle de confiance de niveau de confiance $1 - \alpha$;
- ➢ On rejette l'hypothèse nulle sinon.

Définition 2.1

La région critique de la statistique U du test de comparaison de moyennes est de la forme :

$$RC =]-\infty, -\mu_\alpha] \cup [\mu_\alpha, +\infty[,$$

μ_α étant calculé de façon que :

$$P(|U| > \mu_\alpha) = \alpha.$$

la v.a. U suivant la loi normale centrée réduite.

Exemple 8

D'après une société conseil, le salaire annuel moyen d'administrateur de banques de données serait de **49738** Euros. Une enquête effectuée auprès d'un organisme sur un échantillon aléatoire de **36** entreprises de ce secteur donne les résultats suivants concernant la rémunération des administrateurs de banque de données :

Salaire moyen : **50200** Euros Ecart type : **1560** Euros Est-ce que les résultats de cette enquête permettraient de supporter l'affirmation de cette société conseil ? Utiliser un seuil de signification de $\alpha = 0.05$

❶ **Hypothèse Statistique :**

$H_0 : \mu = 49738$

$H_1 : \mu \neq 49738$

❷ **Seuil de signification :**

$\alpha = 0.05$

❸ **Condition d'application du test :**

Grand échantillon $(n \geq 30)$ provenant d'une population de variance connue.

❹ **La statistique :**

La statistique qui convient pour le test est $\overline{X}_n$, l'écart réduit est $Z = \frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ où $\mu_0 = 49738$ et Z suit la loi $\mathcal{N}(0,1)$.

❺ **Règle de décision :**

D'après H_1 et au seuil $\alpha = 0.05$ les valeurs critiques de l'écart réduit sont $z_{\alpha/2} = 1.96$ et $-z_{\alpha/2} = -1.96$ (test bilatéral).

On adoptera la règle de décision suivante :

Rejeter H_0 si $Z > 1.96$ ou $Z < -1.96$, sinon ne pas rejeter H_0.

❻ **Calcul de l'écart réduit :**

Puisque $\overline{X}_n = 50200$, $\sigma = 1560$ et $n = 36$, donc :

$Z = \sqrt{36}\frac{50200-49738}{1560} = 462/260 = 1.77$

❼ **Décision et conclusion :**

La valeur de $Z = 1.77$ se situe dans la région de non-rejet de H_0 donc on ne peut rejeter l'affirmation de la société conseil. L'écart observé entre $\overline{X}_n$, et μ_0 soit $(50200 - 49738 = 462)$, n'est pas statistiquement significatif au seuil $\alpha = 0.05$.

2.2.2 Tests d'hypothèses sur une variance

Le test d'égalité de la variance d'une population à une valeur spécifiée est lui aussi équivalent à l'estimation par intervalle de confiance.

Il consiste à décider entre les hypothèses suivantes :

$$H_0 : \sigma^2 = \sigma_0^2 \qquad H_1 : \sigma^2 \neq \sigma_0^2$$

La question posée est la suivante : les observations remettent-elles en cause l'égalité de la variance théorique σ^2 à la valeur spécifiée σ_0^2 ?

Règle de décision : pour un risque de première espèce α,

➢ On accepte l'hypothèse nulle si la valeur σ_0^2 appartient à l'intervalle de confiance de niveau de confiance $1 - \alpha$;

➢ On rejette l'hypothèse nulle sinon.

Établissons la région critique du test.On rejette l'hypothèse nulle lorsque la valeur observée de la variance empirique S^2 est très différente de la valeur σ_0^2, donc lorsque la v.a. $X^2 = n\frac{S^2}{\sigma_0^2}$ prend une valeur anormalement petite (inférieure à χ_{α^2}) ou anormalement grande (supérieure à $\chi_{1-\alpha^2}$). Les bornes de la région critique χ_{α^2} et $\chi_{1-\alpha^2}$ sont définies par la relation :

$$P(X^2 < \chi^2_{1-\alpha}) = \frac{\alpha}{2}, \qquad P(X^2 > \chi^2_{1-\alpha}) = \frac{\alpha}{2}.$$

la statistique X^2 suivant la loi du χ^2 de degré de liberté $n-1$.

La région critique est donc de la forme :

$$\boxed{RC = [0, \chi_{\alpha^2}] \cup [\chi_{1-\alpha^2}, +\infty]}$$

Théorème 2.1

La loi de la v.a. F ci-dessous est la loi de Fisher de degrés de liberté $n_1 - 1$ et $n_2 - 1$.

$$F = \frac{n_1 S_1^2}{\sigma_1^2} / \frac{n_2 S_2^2}{\sigma_2^2} \times \frac{n_2 - 1}{n_1 - 1}.$$

Considérons maintenant les hypothèses suivantes :

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2 \qquad H_1 : \sigma_1^2 \neq \sigma_2^2$$

Si l'hypothèse nulle est vraie, on a :

$$F = \frac{n_1 S_1^2}{n_2 S_2^2} \times \frac{n_2 - 1}{n_1 - 1} = \frac{n_2 S_1^2}{n_1 - 1} / \frac{n_2 S_2^2}{n_2 - 1}$$

et F devrait être proche de 1 puisque le numérateur et le dénominateur du rapport ci-dessus sont des estimateurs sans biais de la même variance σ^2. Si la v. a. F prend une très grande valeur ou est très proche de 0, on rejette l'hypothèse nulle. La région critique est de la forme :

$$\boxed{RC =]0, f_\alpha[\cup]f_{1-\alpha}, +\infty[}$$

les bornes fa et fia étant choisies dans la table de Ficher Snedecor de façon que :

$$P(F < f_\alpha) = \frac{\alpha}{2}, \qquad P(F > f_{\alpha-1}) = \frac{\alpha}{2}.$$

Exemple 9

Selon la responsable de l'application de tests d'évaluation de l'entreprise PMX, la variabilité des résultats aux tests de dextérité n'excède pas **144** $(\sigma = 12)$.

Les résultats à ce test obtenus par un échantillon aléatoire de **20** employés donnent une somme de carrés des écarts par rapport à la moyenne de **2952**. On suppose que l'échantillon aléatoire provient d'une population normale.

L'hypothèse selon laquelle σ^2 n'excède pas **144** est-elle acceptable au seuil de signification $\alpha = 0.05$?

Effectuer le test selon la démarche usuelle (**7** étapes).

❶ **Hypothèse Statistique :**

$H_0 : \sigma^2 = 144$

$H_1 : \sigma^2 > 144$

❷ **2- Seuil de signification :**

$\alpha = 0.05$

❸ **3- Condition d'application du test :**

Échantillon aléatoire provenant d'une population normale.

❹ **La statistique :**

La statistique qui convient pour le test en supposant H_0 vraie est $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$ où $\sigma_0^2 = 144$. La quantité χ^2 est distribuée selon la loi de khi-deux à **19** $(20-1)$ degrés de liberté.

❺ **Règle de décision :**

D'après H_1 et au seuil $\alpha = 0.05$ la valeur critique de χ^2 est $\chi^2_{0.05;19} = 30.1435$ (test unilatéral à droite).

On adoptera la règle de décision suivante :

Rejeter H_0 si $\chi^2 > 30.1435$, sinon ne pas rejeter H_0.

❻ **Calcul de l'écart réduit :**

On a $\sum(X_i - \overline{X}_n)^2 = (n-1)s^2 = 2952, n = 20$ et $\sigma_0^2 = 144$.

Le calcul de χ^2 donne : $\chi^2 = \frac{2952}{144} = 20.5$.

❼ **Décision et conclusion :**

Puisque $\chi^2 = 20.5 < 30.1435$, on ne peut rejeter H_0. La dispersion des résultats au test de dextérité semble correspondre à la norme requise. Au risque de se tromper **5** fois sur **100**, il n'est pas invraisemblable d'observer une variance de **2952/19 = 155.37** dans un échantillon de taille $n = 20$ lorsqu'on admet que la variance de la population est **144**.

Cette valeur expérimentale ne permet pas d'écarter l'hypothèse selon laquelle $\sigma^2 = 144$.

2.2.3 Tests d'hypothèses sur une proportion

On se propose de tester si la proportion p d'éléments de la population présentant un certain caractère qualitatif peut être considéré ou non comme égale a une valeur hypothétique p_0.
Hypothèse nulle : $H_0 : p = p_0$,
Hypothèse alternative : $H_1 : p \neq p_0 (p > p_0$ ou $p < p_0)$.

Soit un échantillon de grande taille $(n > 30)$ prélevé au hasard d'une population binomiale de sorte que $np \geqslant 5$ et $n(1-p) \geqslant 5$ dans ce cas l'écart réduit

$$Z = \widehat{p} - p_0 / \sqrt{p_0(1-p_0)/n}.$$

est distribué selon la loi normale $\mathcal{N}(0,1)$.

Le tableau suivant résume dans le cas d'un test sur une proportion les règles de décision selon les hypothèses H_0 pour le seuil de signification α.

Hypothèses nulle : $H_0 : p = p_0$

Contre-hypothèses	Régles de décision du test
$H_1 : p \neq p_0$	Rejeter H_0 si $Z > z_{\alpha/2}$ ou $Z < -z_{\alpha/2}$
$H_1 : p > p_0$	Rejeter H_0 si $Z > z_\alpha$
$H_1 : p < p_0$	Rejeter H_0 si $Z < -z_\alpha$

Exemple 10

Un cadre d'une société conseil affirme que **36%** des entreprises ont reçu des commandes par internet. Une enquête sur le commerce électronique par les entreprises indique que sur un échantillon de **256** entreprises **96** font du commerce électronique. Peut-on considérer, au seuil de signification de **5%** que l'affirmation du conseiller est vraisemblable ?

❶ **Hypothèse Statistique :**

$H_0 : p = 0.36$

$H_1 : p \neq 0.36$

❷ **Seuil de signification :**

$\alpha = 0.05$

❸ **Condition d'application du test :**

$np \geq 5, n(1-p) \geq 5$ et $n > 30$.

❹ **La statistique :**

La statistique qui convient pour le test est $\widehat{p}$ l'écart réduit est $Z = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ où $p_0 = 0.36$ et Z suit la loi $\mathcal{N}(0,1)$.

❺ **Règle de décision :**

D'après H_1 et au seuil $\alpha = 0.05$ les valeurs critiques de l'écart réduit sont $z_{\alpha/2} = 1.96$ et $-z_{\alpha/2} = -1.96$ (test bilatéral).

On adoptera la règle de décision suivante :

Rejeter H_0 si $Z > 1.96$ ou $Z < -1.96$ sinon ne pas rejeter H_0.

❻ **Calcul de l'écart réduit :**

$\widehat{p} = 96/256 = 0.375$ et $Z = \frac{0.375-0.36}{\sqrt{\frac{0.36(0.64)}{256}}} = 0.5$

❼ **Décision et conclusion :**

Puisque la valeur prise par Z est 0.5 et que $-1.96 < 0.5 < 1.96$ on ne peut rejeter l'hypothèse nulle H_0, donc l'affirmation du conseiller est vraisemblable au seuil de signification de $\alpha = 5\%$.

2.3 Tests Non Paramétrique

Les tests non-paramétriques ne se basent pas sur des distributions statistiques. Ils peuvent donc être utilisés même si les conditions de validité des tests paramétriques ne sont pas vérifiées.

2.3.1 Tests du khi-deux

Test du χ^2 d'adéquation. Ce test permet de vérifier si un échantillon d'une variable aléatoire Y donne des observations comparables à celles d'une loi de probabilité P définie a priori dont on pense, pour des raisons théoriques ou pratiques, qu'elle devrait être la loi de Y.

Tests du khi-deux d'ajustement

1- Cas d'une variable discrète.

On considère le cas d'un dé à **6** faces. Les hypothèses concernent la loi de probabilité de la face obtenue en lançant le dé.

• Hypothèse nulle H_0 : cette loi est la loi uniforme discrète sur $\{1,...,6\}$ (le dé est parfaitement équilibré).

• Hypothèse alternative : ce n'est pas la loi uniforme discrète sur $\{1,...,6\}$ (les faces n'ont pas toutes la même probabilité, et le dé est mal équilibré).

L'expérience consiste à lancer le dé n fois ($100, 200$ ou 1000 fois par exemple). On compare ensuite les proportions $f_i (i = 1,...,6)$ observées aux probabilités théoriques $p_i (i = 1,...,6)$ de

chaque face du dé. Elle est généralisable à toutes les v.a. discrètes ou qualitatives :

• L'hypothèse nulle est définie par la loi de probabilité supposée vraie, dont la densité est définie par la suite (p_i), $i = 1,...,k$

• L'hypothèse alternative est que la loi de probabilité n'est pas égale à la précédente, sans plus de précision. Pour contrôler l'hypothèse nulle, on compare les proportions n_i/n observées sur un échantillon de taille n aux probabilités théoriques p_i.

Définition 2.2

On appelle X^2 la statistique choisie pour comparer les proportions observées aux probabilités théoriques :

$$X^2 = n\sum_{i=1}^{k}(n_i/n - p_i)^2/p_i = \sum_{i=1}^{k}(n_i - np_i)^2/np_i.$$

Reprenons l'exemple du dé : s'il est bien équilibré, les proportions n_i/n convergent vers les probabilités $p_i = 1/6$ et la valeur prise par la v.a.X^2est faible. Inversement, si la valeur prise par X^2 est élevée, on peut penser que le dé n'est pas bien équilibré puisque les proportions sont différentes de $1/6$.

Le raisonnement est exactement identique dans le cas général, et nous allons déterminer une valeur x_α^2 indiquant la limite à partir de laquelle nous considérons que la valeur de X^2 est trop élevée pour que l'hypothèse nulle soit vraie.

Pour déterminer cette valeur x_α^2, on utilise le risque de première espèce α, qui est la probabilité de rejeter l'hypothèse nulle alors qu'elle est vraie (considérer le dé pipé alors qu'il est bien équilibré).

La valeur x_α^2 est donc telle que :

$$P(X^2 > x_\alpha^2) = \alpha$$

Cette valeur dépend bien entendu de la loi de probabilité de la v.a. X^2 : cette loi est approximativement la loi du χ^2.

2- Cas d'une variable continue

Considérons maintenant le cas des v.a. continues :

• L'hypothèse nulle est définie par une densité théorique de X ;

• L'hypothèse alternative est une loi non précisée de X. La procédure est la suivante :

• On définit des intervalles $I_i, i = 1,...,k$, dont on calcule les probabilités théoriques $p_i = P(X \in I_i)$.

• On répartit les n observations de la v.a. X dans ces intervalles.

● On en déduit la densité observée égale à la suite des proportions n_i/n, où n_i est le nombre d'observations classées dans l'intervalle I_i.

Remarque 2.1

- ➢ Les classes seront choisies toujours a priori, avant le calcul de X^2, et de préférence de probabilité égale.
- ➢ Le calcul des probabilités théoriques peut exiger préalablement l'estimation de paramètres de la densité théorique.
- ➢ Deux densités différentes peuvent donner la même densité par intervalle. L'hypothèse nulle ne les distingue pas l'une de l'autre et le test donne la même valeur de X^2 et par suite , pour un même degré de liberté, la même décision.

Exemple 11

Dans le but de vérifier si un dé est bien équilibré une machine "lance" le de **1000** fois et on observe le nombre de points sur la face visible du dé. Les résultats sont donnés dans le tableau suivant :

Face	1	2	3	4	5	6
Observations	180	167	158	210	135	150

Faire un test au niveau **5%** pour vérifier si le dé est équilibré.

Considérons la va qui donne le nombre de points sur la face visible du dé, on veut confronter les hypothèses :

$$H_0 \ : \ \pi_i \ = \ \frac{1}{6} \ pour \ chaque \ i = 1,2,...,6,$$
$$H_1 \ : \ \pi_i \ \neq \ \frac{1}{6} \ pour \ au \ moins \ un \ i.$$

Le test d'ajustement du khi-deux est de rejeter H_0 si :

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - T_i)^2}{T_i} \geq \chi^2_{k-1;\alpha}$$

où $k = 6$ et $\alpha = 0.05$. On obtient :

x_i	1	2	3	4	5	6
T_i	166.67	166.67	166.67	166.67	166.67	166.67

et ainsi les conditions d'application du test du khi-deux sont respectées.

On observe

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{k} \frac{(n_i - T_i)^2}{T_i} \\ &= \frac{(180 - 166.67)^2}{166.67} + \frac{(167 - 166.67)^2}{166.67} + ... \\ &= 20.468\end{aligned}$$

Or $\chi^2_{5;0.05} = 11.07$ donc on rejette H_0 et on doit conclure avec un niveau de 5% que le de n'est pas équilibré.

Tests du khi-deux d'indépendance

Pour comparer les effectifs théoriques et les effectifs observés, on utilise la même statistique que dans le cas du test d'ajustement.

Définition 2.3

La statistique X^2 utilisée pour comparer les répartitions théoriques et observées est définie par :

$$X^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} (n_{i,j} - n'_{i,j})^2 / n'_{i,j}$$

L'hypothèse d'indépendance est contestable lorsque les effectifs observés $n_{i,j}$, sont très différents des effectifs théoriques $n'_{i,j}$, donc lorsque X^2 prend de grandes valeurs. Il reste à décider à partir de quelle valeur X^2 peut être considéré comme grand. Pour cela, on utilise la loi de X^2 sous l'hypothèse d'indépendance qui est la loi du χ^2 de degré de liberté v :

$$v = (p-1)(q-1).$$

Ce degré de liberté est calculé comme le précédent, par la formule $v = k - l - 1$;

• le nombre de valeurs possibles est $k = p \times q$,

• les paramètres estimés sont les lois de probabilités marginales : $p - 1$ termes pour la loi de $X, q - 1$ pour la loi de Y puisque la somme des probabilités marginales est égale à 1. On a donc $l = (p-1) + (q-1)$,

• Le degré de liberté est égal à : $p \times q - (p-1) - (q-1) - 1 = (p-1)(q-1)$.

Exemple 12

Un échantillon de **1000** personnes ont été interrogées sur leur opinion à propos d'une question qui sera posée à un référendum. On a demandé à ces personnes de préciser leur appartenance politique. Les résultats sont donnés par le tableau suivant :

Appartenance	Réponse		
	Favorable	Défavorable	Indécis
Gauche	210	194	91
Droite	292	151	62

On veut savoir la réponse au référendum est indépendante de l'opinion politique. Pour cela associons les indices de ligne $i = 1$ et 2 à gauche et droite respectivement les indices de colonne $j = 1,2,3$ aux réponses favorable, défavorable et indécis respectivement. On calcule alors les valeurs $\frac{n_{ij}}{n}$ (ici $n = 1000$.) qu'on dispose dans un tableau ainsi que les valeurs $\frac{n_i n_j}{n^2}$ (dans le même tableau entre parenthèses), ce qui donne :

i	j		
	1	2	3
1	0.21(0.248)	0.194(0.170)	0.091(0.076)
2	0.292(0.254)	0.151(0.174)	0.062(0.077)

La quantité $\chi^2(\widehat{p}_n, \overline{p}_n) = n\sum_{i,j} \frac{(n_i n_j/n^2 - n_{ij}/n)^2}{n_i n_j/n^2}$ est alors égale à :
$\mathbf{1000}\left(\frac{(0.248-0.210)^2}{0.248} + \frac{(0.170-0.194)^2}{0.170} + \frac{(0.076-0.091)^2}{0.076} + \frac{(0.254-0.292)^2}{0.254} + \frac{(0.174-0.151)^2}{0.174} + \frac{(0.077-0.062)^2}{0.077}\right)$
$= \mathbf{23.82}$

Dans ce cas on a $k - 1 = 1$ et $l - 1 = 2$ i.e. on utilise la loi $\chi^2(2)$ pour laquelle le seuil de rejet au risque $\alpha = 0,05$ est égal à **5.99**. On doit donc rejeter l'hypothèse H_0 d'indépendance de la réponse et de l'opinion politique. On constate qu'on rejette aussi H_0 au risque **0.01**.

2.3.2 Test de Kolmogornv-Smirnov

Le test de Kolmogorov-Smirnov est un test d'ajustement. La différence avec le test du χ^2 est qu'il est fondé sur les fonctions de répartition plutôt que sur les densités.

L'hypothèse nulle est :

H_0 : la loi P a la même fonction de répartition F qu'une loi continue donnée.

L'idée est que, si l'hypothèse nulle H_0 est vraie, la fonction de répartition empirique $\widehat{F}$ de l'échantillon doit être "proche" (en un sens qui va être précisé) de F.

Fonction de répartition empirique

On cherche à obtenir une estimation de la fonction de répartition à partir de l'échantillon observé afin de la comparer ensuite à la fonction de répartition de la loi théorique.

Pour cela, on commence par trier par ordre croissant les valeurs X_i de l'échantillon.On les appelle traditionnellement des statistiques d'ordre.

La fonction de répartition empirique est définie par :

$$\widehat{F}(x) = \begin{cases} 0 & pour \quad x < X_1 \\ \frac{i}{n} & pour \quad X_i \leq x < X_{i+1} \\ 1 & pour \quad x \geq X_n \end{cases}$$

On estime donc $F(x) = P(X \leq x)$ au moyen de la proportion $\widehat{F}(x)$ d'éléments de l'échantillon qui sont inférieurs ou égaux à x.

Procédure

Voici une description détaillée de la procédure d'exécution du test de Kolmogorov-Smirnov

1. classer les valeurs observées par ordre croissant :

2. calculer les nombres $\frac{i}{n}$; c'est-à-dire les valeurs supérieures de la distribution empirique ;

3. calculer les valeurs absolues des écarts $\left|F(X_i) - \frac{i}{n}\right|$ entre F et les valeurs précédentes ;

4. calculer les nombres $\frac{i-1}{n}$,c'est-à-dire les valeurs inférieures de la distribution empirique ;

5. calculer les valeurs absolues des écarts $\left|F(X_i) - \frac{i-1}{n}\right|$ entre F et les valeurs précédentes ;

6. la distance de Kolmogorov-Smirnov est le plus grand de tous ces écarts ;

7. on conclut le test en acceptant l'hypothèse H_0 si la distance calculée est inférieure à la valeur critique donnée dans la table et en la rejetant sinon.

Le théorème fondamental de la Statistique

On continue l'étude précédente. Soit F_n, la fonction de répartition empirique d'un n-échantillon $(X_1, ..., X_n)$ de variables aléatoires réelles de fonction de répartition F. On veut des précisions sur l'écart uniforme :

$$D_n(\omega) = \sup_{x \in \mathbb{R}} |F_n^{\omega}(x) - F(x)|, \tag{2.1}$$

entre F_n et F. Bien entendu, D_n, est une variable aléatoire puisqu'elle dépend de chaque réalisation du n-échantillon. A priori la loi de D_n semble dépendre de F mais, (et c'est particulièrement remarquable), ce n'est pas le cas si on se restreint à des lois F continues.

Théorème 2.2

Si F est continue, la loi de probabilité de la variable aléatoire D_n, définie en (2.1) est une loi intrinsèque i.e. elle ne dépend pas de F.

Démonstration 1

Posons $Y_i = F(X_i)(i = 1,...,n)$. $(Y_1,...,Y_n)$ est un n-échantillon de la loi uniforme sur $[0,1]$ dont nous noterons U_n^ω la fonction de répartition empirique :

$$U_n^\omega(t) = \frac{1}{n}\sum_{i=1}^{n} \delta_{F(X_i(\omega))}\,(]-\infty,t]) \tag{2.2}$$

Mais $\delta_{F(X_i(\omega))}\,(]-\infty,t]) = \delta_{F(X_i(\omega))}\left(]-\infty,F^{-1}t]\right)$ où F^{-1} est l'inverse généralisée de F. Il résulte alors de (2.2) que $U_n^\omega(t) = F_n^\omega(F^{-1}(t))$. Ainsi, en notant $U(t)$ la fonction de répartition de la loi uniforme, on a :

$$\begin{aligned}
\sup_{t\in\mathbb{R}} |U_n^\omega(t) - U(t)| &= \sup_{t\in]0,1[} |U_n^\omega(t) - t| \\
&= \sup_{t\in]0,1[} |F_n^\omega(F^{-1}(t)) - F(F^{-1}(t))| \\
&= \sup_{t\in\mathbb{R}} |F_n^\omega(t) - F(t)| = D_n(\omega),
\end{aligned}$$

ce qui prouve que $D_n(\omega)$ a la même valeur que pour la loi uniforme.Q.E.D.

Remarque 2.2

La loi de probabilité de D_n, a été tabulée pour différentes valeurs de l'en tier n.

CHAPITRE 3

APPLICATIONS ET SIMULATION SOUS MATHEMATICA

3.1 Exemple extrait du l'article de Boitsov et Guzeva

L'idée de base de cet exemple est la loi gaussienne, c'est-à-dire on suppose que la valeur la plus probable (l'espérance mathématique) $\tilde{m}_x$ de la valeur aléatoire X est la moyenne arithmétique de ses réalisations x_i, définie comme suit :

$$\tilde{m}_x = \overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{3.1}$$

où N est le nombre d'expériences répétées. On suppose que la meilleure caractéristique de la dispersion des résultats, qui est cohérente avec le paramètre (3.1), est la variance S_x^2, définie par la formule suivante :

$$\tilde{\sigma}_x^2 = S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x}_i)^2. \tag{3.2}$$

En tant qu'estimation maximale de vraisemblance du paramètre inconnu θ (par exemple, m_x ou l'estimation de la variance S_x^2), une valeur de $\tilde{\theta}$ est prise à laquelle la densité de probabilité $f(x)$ acquiert la valeur la plus élevée pour les réalisations obtenues $f(x_1, x_2, .., x_N|\tilde{\theta}) = maxf(x_1, x_2, .., x_N|\theta$.

Cette condition conduit à l'équation du maximum de vraisemblance introduite par Fisher.

Considérons le degré de corrélation entre la loi normale et le nombre de charges programmées cycliques nécessaires pour rompre le joint soudé par adhésif (avec un point de soudure),

j	Intervalle au milieu	$f(x) \times 10^6$	$P_j = f(x) \times \Delta x$	$F(x)$
1	457000	0.743	0.111	0.111
2	607000	1.57	0.236	0.347
3	757000	1.94	0.291	0.638
4	907000	1.41	0.212	0.850
5	1057000	0.6	0.090	0.940

TABLE 3.1 – Calculs effectués pour tracer les fonctions $F(x)$ et $f(x)$

qui est formé à l'aide de l'adhésif $UP - 5 - 207$, sur a, les cycles formant la série des mesures expérimentales suivante :
385000, 420000, 585000, 638000, 675000, 700000, 725000, 732000, 740000, 745000, 810000, 850000, 910000, 1090000 et **1130000.**

3.1.1 Estimations MV des Paramètres de la loi Normale

Dans l'exemple 2 section 1.2.3 on étudier l'estimation du MV du paramètres du loi normal. Alors pour notre exemple, nous avons calculé les valeurs suivantes $\overline{x}$ et S : $\overline{x} = 742000$ et $S_x^2 = 42100 \times 10^6$ respectivement par les formules (3.1) et (3.2).Ces valeurs ont été remplacées dans la loi de distribution normale (équations décrivant la densité de probabilité) comme suit :

$$f(x) = \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{(x_i - m_x)^2}{2\sigma_x^2}} \approx \frac{1}{S_x\sqrt{2\pi}} e^{-\frac{(x_i - \overline{x})^2}{2S_x^2}} = \frac{1}{2.05 \times 10^5\sqrt{2\pi}} e^{-\frac{(x_i - 7.24\times 10^5)}{2\times 4.21\times 10^{10}}}, \qquad (3.3)$$

où $-\infty < x_i < +\infty$ et $\sigma_x < 0$. Pour l'expression générale de la loi de Gauss, on obtient $-\infty < m_x < +\infty$.

3.1.2 Représentation graphique des résultats

Traçons la fonction de distribution théorique $F(x)$ et la densité de probabilité théorique $f(x)$ (Fig. 1, 3.2) pour l'échantillonnage susmentionné du nombre de cycles avant la rupture du joint. La série d'intervalles de ces données est présentée. Les valeurs de $F(x)$ sont obtenues par la sommation des chiffres dans la colonne P_i, de haut en bas (Somme cumulée).

Les fonctions théoriques $F(x)$ et $f(x)$ sont représentées en traits rouge sur les Fig. 3.1 et Fig 3.2, respectivement. Le degré de corrélation entre les résultats expérimental et théorique des Fig. 3.1 et Fig 3.2 sont généralement vérifiés par les critères de Kolmogorov ou Pearson (du Khi-deux).

3.1.3 Test de Kolmogorov-Smirnov

Kolmogorov a proposé d'utiliser le maximum valeur absolue de la différence entre les fonctions de distribution statistique $\tilde{F}(x)$ et de distribution théorique correspondante $F(x)$ comme mesure de l'écart entre les distributions théorique et statistique :

$$D_n = max|F(x) - \tilde{F}(x)|. \tag{3.4}$$

Pour notre problème, à partir du [1, Tableau 3.1] et du Tableau 1 de cet exemple, nous établissons que $D_n = \mathbf{0.095}$. Ensuite, nous calculons la valeur

$$\lambda = D_n\sqrt{N} = \mathbf{0.095}\sqrt{\mathbf{15}} = \mathbf{0.368} \tag{3.5}$$

et, à partir des tableaux de référence, nous trouvons la probabilité $P(\lambda) = \mathbf{0.99}$. Ainsi, l'hypothèse de la normalité de la distribution du nombre de cycles avant la rupture de l'articulation pourrait être reconnue comme cohérente avec les données expérimentales. Cependant, le critère de Kolmogorov est directement lié au nombre N d'expériences via la condition (3.5) et, par conséquent, ne peut être utilisé qu'à de grandes valeurs N, il est plus commode de se pencher sur le test de Pearson (parfois appelé distribution χ^2).

3.1.4 Tests du khi-deux

Selon la proposition de Pearson, si les résultats expérimentaux $(x_1, x_2, ..., x_N)$ sont résumés en k intervalles et, pour chacun d'eux, la fréquence correspondante $\tilde{P}_j$, et la probabilité théorique de faire tomber le nombre aléatoire dans ces P_j, les rangs sont calculés, la mesure de l'écart entre les lois statistiques et théoriques est désignée par χ^2 :

$$\chi^2 = N\sum_{j=1}^{k}\frac{(\tilde{P}_j) - P_j}{P_j}. \tag{3.6}$$

L'étape suivante est la détermination du nombre r de degrés de liberté, qui est égal au nombre k de rangs moins le nombre de conditions indépendantes imposées aux fréquences $\tilde{P}_j$. Ces conditions sont composées du nombre l de paramètres estimés dans une loi de distribution (par exemple, à la loi de distribution normale, l'espérance mathématique et la variance sont estimées ; c'est-à-dire $l = \mathbf{2}$) et encore une autre exigence,$\sum_{j=1}^{k} = \tilde{P}_j = \mathbf{1}$. Par conséquent, dans notre exemple,

$$r = k - l - \mathbf{1} = \mathbf{5} - \mathbf{2} - \mathbf{1} = \mathbf{2}. \tag{3.7}$$

En utilisant les tableaux de χ^2, nous trouvons la probabilité que la valeur examinée avec la

j	$\tilde{P}_j$	P_j	$[(\tilde{P}_j - P_j)^2/P_j] \times 10^3$
1	0.133	0.111	4.3603
2	0.200	0.236	5.4915
3	0.400	0.291	40.8281
4	0.133	0.212	29.4386
5	0.134	0.090	21.5111

TABLE 3.2 – Calcul du critère de Pearson

distribution χ^2 et r degrés de liberté dépasse la valeur χ^2 donnée. Les résultats des calculs sont énumérés dans le Tableau 3.2, dans lequel les valeurs $\tilde{P}_j$ et P_j sont tirées du [1, tableau 3.1].

La somme des valeurs de la troisième colonne est égale à **0.102**; par conséquent, selon l'équation (3.6), à $N = 15$, paramètre $\chi^2 = 1.5$ pour la distribution selon la loi χ^2, nous établissons que l'hypothèse de la distribution normale des résultats de test peut être, dans ce cas, acceptée avec une probabilité de **0.45**. Par conséquent, il est raisonnable de vérifier l'hypothèse de la corrélation des résultats expérimentaux avec la loi de distribution log-normale en substituant $Y = \log x$ à la valeur aléatoire x. L'équation (3.3) n'est opérée qu'à $x > 0$, ce qui est cohérent avec la signification physique du problème considéré. En effectuant maintenant toutes les opérations décrites ci-dessus avec le paramètre Y, nous trouvons la valeur $\chi^2 = 0.767$, ce qui correspond à une probabilité de **0.7**.

3.1.5 Application sous Mathématica

```
results = {385000, 420000, 585000, 638000, 675000, 700000, 725000,
    732000, 740000, 745000, 810000, 850000, 910000, 1090000,
    1130000}*10^-5;
% // N
pas = (113/10 - 77/20)/5;
f[x_] = Piecewise[{{0, x <= 4.57 - 0.75}, {.743,
    x <= 4.57 + 0.75}, {1.57, x <= 6.07 + 0.75}, {1.94,
    x <= 7.57 + 0.75}, {1.41, x <= 9.07 + 0.75}, {.6,
    x <= 10.57 + 0.75}}]
p1 = Plot[f[x], {x, 2, 12}, PlotStyle -> Green];
p2 = ListPlot[{{4.57 - 0.75,
     0}, {4.57 - 0.75, .743}, {4.57 + 0.75, .743}, {4.57 + 0.75,
     1.57}, {6.07 + 0.75, 1.57}, {6.07 + 0.75, 1.94}, {7.57 + 0.75,
     1.94}, {9.07 + 0.75, 1.41}, {10.57 + 0.75, .6}}, Filling -> Axis,
```

```
    FillingStyle -> Blue];
p3 = Plot[
   10^6/(2.05 *10^5*
      Sqrt[2 Pi]) Exp[-(x*10^5 - 7.42*10^5)^2/(2 4.21*10^10)], {x, 3,
     12}, PlotStyle -> Red];
Show[p1, p2, p3]
{3.85, 4.2, 5.85, 6.38, 6.75, 7., 7.25, 7.32, 7.4, 7.45, 8.1, 8.5, \
9.1, 10.9, 11.3}
0          x<=3.82
0.743      x<=5.32
1.57       x<=6.82
1.94       x<=8.32
1.41       x<=9.82
0.6        x<=11.32
0          True
```

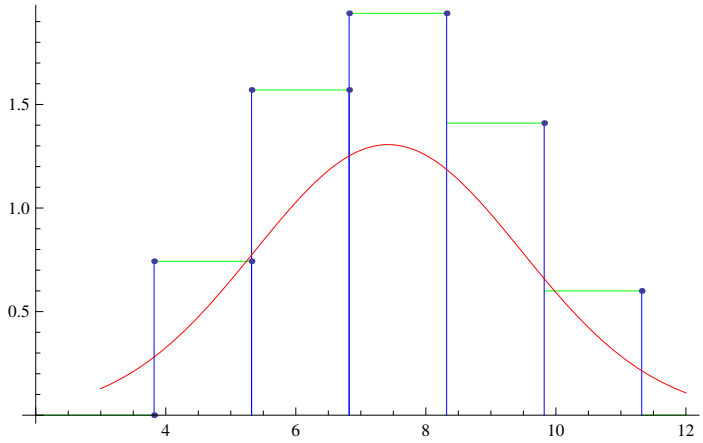

FIGURE 3.1 – Fonction de densité théorique et l'histogramme des résultats

```
Accumulate[{.111, .236, .291, .212, .090}]
{0.111, 0.347, 0.638, 0.85, 0.94}
f[x_] = Piecewise[{{0, x <= 4.57 - 0.75}, {0.111‘,
    x <= 4.57 + 0.75}, {0.347, x <= 6.07 + 0.75}, {0.638,
    x <= 7.57 + 0.75}, {0.85, x <= 9.07 + 0.75}, {0.94,
    x <= 10.57 + 0.75}}]
p1 = Plot[f[x], {x, 2, 12}, PlotStyle -> Green];
p2 = ListPlot[{{4.57 - 0.75, 0}, {4.57 - 0.75, 0.111‘}, {4.57 + 0.75,
     0.111‘}, {4.57 + 0.75, 0.347}, {6.07 + 0.75,
     0.347}, {6.07 + 0.75, 0.638‘}, {7.57 + 0.75,
```

```
     0.638‘}, {7.57 + 0.75, 0.85‘}, {9.07 + 0.75,
     0.85‘}, {9.07 + 0.75, 0.94‘}, {10.57 + 0.75, 0.94‘}},
   Filling -> Axis, FillingStyle -> Blue];
p3 = Plot[CDF[NormalDistribution[7.42, 2.05 ], x], {x, 3, 12},
   PlotStyle -> Red];
Show[p1, p2, p3]
0          x<=3.82
0.111      x<=5.32
1.347      x<=6.82
1.638      x<=8.32
1.85       x<=9.82
0.94       x<=11.32
0          True
```

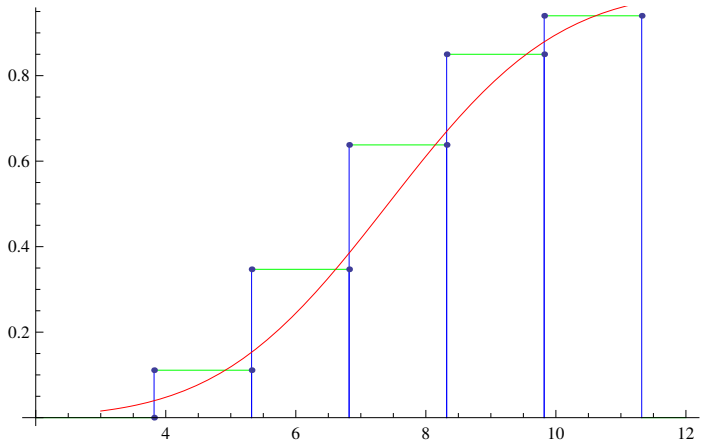

FIGURE 3.2 – Fonction de répartition théorique et empirique

```
{g, {{binCounts}}} =
  Reap[Histogram[results, {77/20, 113/10, 149/100},
    Function[{bins, counts}, Sow[counts]]]];
g
binCounts[[5]] = binCounts[[5]] + 1;
binCounts


{2, 3, 6, 2, 2}
```

Kolmogorov Smirnov Test :

```
H = KolmogorovSmirnovTest[{385000, 420000, 585000, 638000, 675000,
```

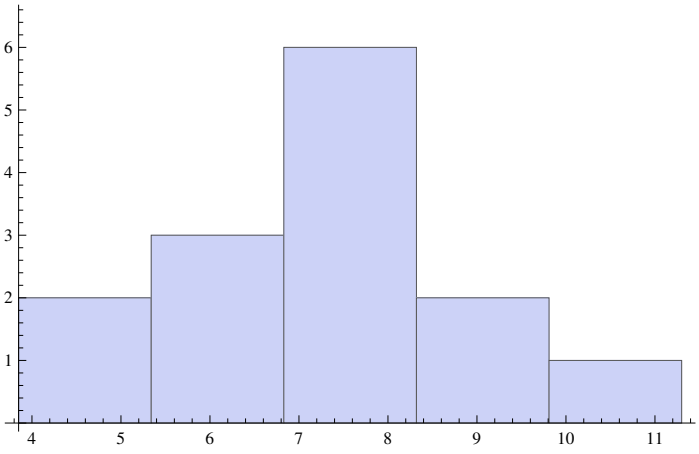

FIGURE 3.3 – L'histogramme du résultats

```
    700000, 725000, 732000, 740000, 745000, 810000, 850000, 910000,
    1090000, 1130000}, NormalDistribution[742000, 205000],
   "HypothesisTestData"];
H["TestDataTable"]
{{"", "Statistic", "P\[Hyphen]Value"},
 {"Kolmogorov\[Hyphen]Smirnov", 0.160829, 0.776549}}
```

ChiSquare Test :

```
tho = {0.111, 0.236, 0.291, 0.212, 0.090};
exp = {0.133, 0.200, 0.400, 0.133, 0.134};
stat = (tho - exp)^2;
stat = stat/tho;
stat = 15*Total[stat]
1.52445
CDF[ChiSquareDistribution[2], stat]
0.533372
1 - %
0.466628
```

Loi Log-Normal

```
list$log = {385000, 420000, 585000, 638000, 675000, 700000, 725000,
   732000, 740000, 745000, 810000, 850000, 910000, 1090000, 1130000} //
   Log
pas = (Log[1130000] - Log[385000])/5 // Simplify
```

```
Mean[list$log] // N
StandardDeviation[list$log] // N
{Log[385000], Log[420000], Log[585000], Log[638000], Log[675000],
 Log[700000], Log[725000], Log[732000], Log[740000], Log[745000],
 Log[810000], Log[850000], Log[910000], Log[1090000], 13.9377}
 1/5 Log[226/77]
 13.4791
 0.294932
 {Log[385000], Log[420000], Log[585000], Log[638000], Log[675000],
 Log[700000], Log[725000], Log[732000], Log[740000], Log[745000],
 Log[810000], Log[850000], Log[910000], Log[1090000],
 Log[1130000 - .7]}
 {Log[385000], Log[420000], Log[585000], Log[638000], Log[675000],
 Log[700000], Log[725000], Log[732000], Log[740000], Log[745000],
 Log[810000], Log[850000], Log[910000], Log[1090000], 13.9377}
 {g, {{binCounts}}} =
  Reap[Histogram[%, {Log[385000], Log[1130000], 1/5 Log[226/77]},
    Function[{bins, counts}, Sow[counts]]]];
g
binCounts
```

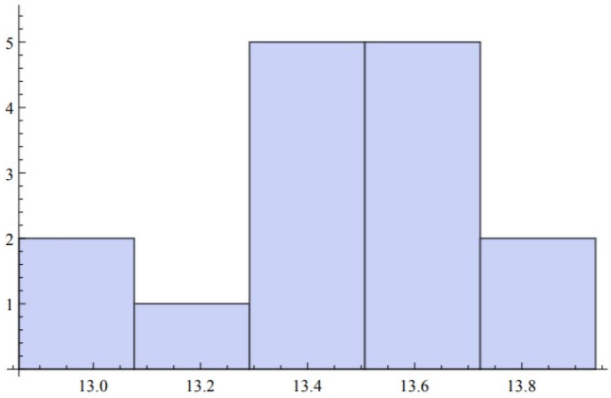

FIGURE 3.4 – L'histogramme logarithme népérienne des résultats

```
{2, 1, 5, 5, 2}
```

Test KHi 2

```
{2, 1, 5, 5, 2}/15 // N
```

```
f[x_] = PDF[
  NormalDistribution[13.479145202918327`, 0.29493186838469765`], x]
Table[12.86099861326992` + pas/2 + pas (i - 1), {i, 1, 5}]
% // f
%*pas
{0.133333, 0.0666667, 0.333333, 0.333333, 0.133333}
1.35266 E^(-5.74813 (-13.4791 + x)^2)
{12.9687, 13.184, 13.3994, 13.6147, 13.8301}
{0.302464, 0.819884, 1.30406, 1.21706, 0.666483}
{0.0651343, 0.176559, 0.280825, 0.262088, 0.143524}
{0.13333333333333333`, 0.06666666666666667`, 0.3333333333333333`,
   0.3333333333333333`, 0.13333333333333333`} - {0.06513428753974712`,
    0.17655876191989417`, 0.28082463107205785`, 0.2620881792619834`,
   0.14352448099426135`};
%^2;
%/{0.30246353822611094`, 0.8198844241986171`, 1.3040629558320436`,
  1.217056653585782`, 0.6664834142401527`}
  {0.0153774, 0.0147292, 0.00211429, 0.00417061, 0.000155832}
  15*Total[{0.0153774, 0.0147292, 0.00211429, 0.00417061, 0.000155832}]
  0.548211
  1 - CDF[ChiSquareDistribution[2], %]
  0.760252
```

CONCLUSION

LES ingénieurs et scientifiques sont souvent sollicités pour réaliser des analyses de risque dans le cadre d'expertises. Ce type d'analyse comporte généralement trois étapes. L'évaluation du risque sa gestion et la communication des résultats qui nécessite la mise en œuvre de méthodes statistiques et de modèles mathématiques. Le traitement et l'analyse de l'information sont au cœur de tous les processus de gestion et de décision. Les méthodes de description, de prévision et de décision se sont considérablement enrichies et développés, ce qui place la statistique appliquée au carrefour de l'observation et de la modélisation. Dans notre mémoire on a simule et détaille un exemple tiré de l'article de Boitsov et Guzeva. Cette étude permet de calculer la fatigue des joints adhésifs à l'aide de procédures développées pour les métaux. Par la méthode du maximum de vraisemblance on estime les paramètres de la loi qui suit la fatigue des joints. En suite par les tests khi-deux et Kolmogorov-Smirnov on test les hypothèses que les résultats expérimentaux suit une loi normal puits une distribution log-normale. En résumé on peut conclure que la deuxième hypothèse est la plus probable. Ce qui est cohérent avec la signification physique du problème considéré.

BIBLIOGRAPHIE

[1] Christian JUTTEN. Détection, Estimation, Information. Univ.Grenoble Alpes-Polytech' Grenoble. Juillet 2018.

[2] Pierre DUSART. Cours de Statistiques inférentielles. Licence 2-S4-MASS. Année 2018.

[3] Jean-Christophe BRETON. STATISTIQUE IUT Biotechnologie 2éme année. veesion du 04 octobre 2008.

[4] Estimation. S.Le Digabel, École Polytechnique de Montréal. MTH2302D. A 2017.

[5] Williaùm W. Hines, Douglas C. Montgomery, David M. Goldsman, Connie M. Borror. Probabilités et statistique pour ingénieurs 2éme édition. Imprimé au Canada.

[6] Probabilité. Echantillonnage, estimation. I.U.T, Département d'informatique. Année 2008-2009.

[7] Gilbert SAPORTA. Probabilités analyse des données et statistique. Editions TECHNIP 25 rue Ginoux, 75015 PARIS, FRANCE.

[8] Valentin Rousson. Statistique appliquée aux sciences de la vie. ISBN 978-2-8178-0393-7 Springer Paris Berlin Heidelberg New York. 2013.

[9] Tests d'hypothèses. S. Le Digabel, Ecole Polytechnique de Montréal. MTH2302D. A 2017.

[10] Marie-Luce Taupin. Rappels sur les tests. Laboratoire LaMME, Université d'Evry val d'Essonne. 2017-2018.

[11] Statistique appliquée a la gestion et au marking. http ://foucart.thierry.free.fr/StatPC.

[12] D.TOUIJAR. Les teste d'hypothèse. 2010-11.

[13] Louis Houde. Module 10 Tests d'hypotheses. Université du Québec a Trois-Rivieres.

[14] Yu.I.Boitsov and T.A.Guzeva. Maximum-Likelihood Estimates of the Normal Distribution Law Parameters of Test Results. February 8, 2007.

Abstract

Our work is structured as follows : the first chapter is made up of three main sections : the first of which presents some generalities on the notion of estimation and the second and third describe in detail two estimation methods, then a second chapter which consists of three sections where we have spoken of two types of tests, the last chapter illustrates some applications of the methods and concepts described in the previous chapters.

Key Words : Maximum likelihood, Estimation, Method of moments, Confidence interval, Hypotheses tests, test chi-2, test Smirnov Kolmogorov, Mathematica

Résumé

Notre travail est structuré comme suit : le premier chapitre est composé de trois section principales : dont la première présente quelques généralités sur la notion d'estimation et la deuxième et la troisième décrivent en détails deux méthodes d'estimation, ensuite un deuxième chapitre qui consiste de trois sections ou on a parle de deux types de tests, le dernier chapitre illustre quelque applications des méthodes et concepts décrits dans les précédents chapitres.

Mots Clés : Estimation, Maximum vraisemblance, Méthode des moments, Intervalle de confiance, Tests des hypothèses, Test khi-deux, Smirnov Kolmogorov, Mathematica.

ملخص

عملنا منظم على النحو التالي: يتكون الفصل الأول من ثلاثة أقسام رئيسية: يقدم الأول منها بعض العموميات حول مفهوم التقدير ويصف الثاني والثالث بالتفصيل طريقتين للتقدير ، ثم الفصل الثاني الذي يتكون من ثلاثة أقسام تحدثنا فيها عن نوعين من الاختبارات ، ويوضح الفصل الأخير بعض تطبيقات الأساليب والمفاهيم الموضحة في الفصول السابقة.

الكلمات المفتاحية: تقدير، الحد الأقصى، طريقة العزوم، مجال الثقة، اختبار الفرضيات، اختبار خي-مربع، ماتيماتيكا.