الجـمهورية الجـزائرية الديمقراطـية الشـعبية
RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
وزارة التعليم العـالي و البحـث العـلمي
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE
جـامعة عمار ثـليجي بالأغـواط
UNIVERSITE AMAR TELIDJI LAGHOUAT
كـلية العلـوم
FACULTÉ DES SCIENCES
قسم الرياضيات
DÉPARTEMENT DE MATHÉMATIQUES

Mémoire de Master

Domaine : Mathématiques et Informatique
Option : Mathematiques appliquées
Spécialisation : Analyse fonctionnelle et applications

Présenté par : Zohra Bentemra

Thème

Introduction à la Statistique Inférentielle

Membres du jury :

Mr.	AbdelAziz RAHMOUNE	MCA	(Université de Laghouat)	Président
Mr.	Boulerbah MERRAD	MAA	(Université de Laghouat)	Examinateur
Mrs.	Nacima MOUSSOUNI	MCA	(Université de Laghouat)	Encadreur

Année universitaire 2025/2026

Remerciements

Je tiens à exprimer ma profonde gratitude à **Madame Nacima Moussouni** pour son encadrement bienveillant, sa disponibilité constante et la richesse de ses conseils qui ont grandement contribué à la réalisation de ce travail. Ses orientations précises et sa rigueur scientifique ont été pour moi une source d'inspiration et de motivation tout au long de cette étude.

Je remercie **Monsieur Abdelaziz Rahmoune**, de nous avoir honnoré de sa présence autant que président jury de soutenance du présent mémoire de Master.

Je remercie également **Monsieur Boulerbah Merrad** pour avoir accepté d'examiner ce travail.

J'adresse également mes sincères remerciements à **Monsieur Fares Yazid**, pour son soutien, son aide précieuse et ses encouragements permanents. Ses remarques pertinentes et sa pédagogie ont largement enrichi cette recherche et m'ont permis d'en améliorer la qualité.

À tous, j'exprime ma reconnaissance la plus sincère pour leur accompagnement et leur confiance.

Acknowledgements

I would like to express my deepest gratitude to **Professor Nacima Moussouni** for her kind supervision, constant availability, and valuable guidance throughout this work. Her scientific rigor and insightful advice have been a great source of inspiration and motivation during the completion of this study.

I thank **Mr. Abdelaziz Rahmoune** for having honored us with his presence as the chairman of the defense jury of this Master's thesis.

I also thank **Mr. Boulerbah Merrad** for having agreed to examine this work.

My sincere thanks also go to **Professor Farès Yazid** for his continuous support, encouragement, and constructive feedback. His valuable comments and pedagogical approach have greatly contributed to improving the quality of this research.

To all of them, I express my heartfelt appreciation for their trust, guidance, and constant encouragement.

Dédicace

À celle qui a été la lumière de ma vie, celle qui m'a appris la patience, la force et la foi, à **ma chère mère**, pour son amour infini, ses sacrifices et son soutien sans faille. Aucune parole ne saurait exprimer toute ma reconnaissance et tout mon amour. Ce modeste travail est dédié à toi, source d'inspiration et de tendresse.

Je remercie également ma famille, mes enseignants, mes amis et tous ceux qui m'ont soutenue de près ou de loin, par leurs encouragements, leurs conseils et leur confiance. Que ce travail soit le fruit de leur appui et de leur présence bienveillante.

À ma mère, avec tout mon amour.

Dedication

To the light of my life, to the one who taught me patience, strength, and faith, to **my beloved mother**, whose unconditional love, endless sacrifices, and unwavering support have been my greatest source of courage and inspiration.
No words can truly express how grateful I am for everything she has done for me. This modest work is dedicated to you, the source of my tenderness and strength.

I also extend my gratitude to my family, teachers, and friends — to all those who supported and encouraged me throughout my academic journey.
May this work be a reflection of their kindness, love, and faith in me.

To my mother, with all my love.

Résumé

Ce travail présente les fondements essentiels de la statistique appliquée et ses principales méthodes d'analyse des données. Dans un premier chapitre, nous avons rappelé les lois de probabilité les plus utilisées, qu'elles soient discrètes ou continues, telles que les lois de Bernoulli, Binomiale, de Poisson, Normale, Khi-deux, Student et Fisher. Le deuxième chapitre a été consacré à la notion d'échantillonnage et à l'estimation ponctuelle, permettant de déduire les caractéristiques d'une population à partir d'un échantillon aléatoire. Le troisième chapitre a abordé la construction des intervalles de confiance, outil indispensable pour estimer les paramètres d'une population avec un certain niveau de précision. Enfin, le quatrième chapitre a traité la théorie des tests statistiques, qui permet de valider ou de rejeter des hypothèses à partir des données observées.

Ainsi, cette étude offre une vue d'ensemble claire et structurée sur les outils fondamentaux de la statistique, indispensables à toute démarche scientifique rigoureuse.

Abstract

This work presents the essential foundations of applied statistics and its main methods of data analysis. The first chapter recalls the most commonly used probability distributions, both discrete and continuous, such as the Bernoulli, Binomial, Poisson, Normal, Chi-squared, Student, and Fisher laws. The second chapter focuses on sampling and point estimation, which make it possible to infer population characteristics from random samples. The third chapter deals with the construction of confidence intervals, a key tool for estimating population parameters with a certain level of precision. Finally, the fourth chapter discusses the theory of statistical tests, which provides a framework for confirming or rejecting hypotheses based on observed data.

Overall, this study offers a clear and structured overview of the fundamental statistical tools essential to any rigorous scientific approach.

Table des matières

Liste des figures

0.1 Introduction

Les statistiques sont une branche des mathématiques où on: collecte, organise, analyse, interpréte et on présente des données. On distingue deux branches principales : la statistique descriptive, qui résume les données observées (ex : moyenne, médiane), et la statistique inférentielle, qui utilise des échantillons pour faire des inférences sur une population plus large.
Dans ce travail, on s'est intéressé à la statistique inférentielle qui nous permet de tirer des conclusions sur une population entière (très large et difficile à étudier) à partir d'un échantillon bien choisit de cette population. Cette stattistique nous permet donc d'estimer des caractéristiques inconnus d'une population (paramètres) ou de tester des hypothèses, à partir des données récoltées de l'echantillon qui représente cette population, à l'inverse des statistiques descriptives qui se contentent de résumer et représenter des données récoltées. La statistique inférentielle est utilisée Dans divers domaines , surtout en médecine, biologie, et aussi pour des sondages d'opinions, et contrôle de qualité, ...etc. Dans ce travail, nous nous sommes intéressés à la statistique inférentielle, où toutes les définitions données sont illustrées par des exemples souvent concrets.

Le manuscrit est organisé de la manière suivante: Le chapitre un est consacré aux notions de base utiles pour la comprehension de la suite du travail. dans les chapitres 2 et 3, on a défini la notion d'échantillonnage, l'estimation ponctuelle. un accent particulier est mis sur la fonction de maximum de vraissemblance qui nous permet de déterminer les estimations ponctuelles des paramètres inconnus de la population. L'estimation ponctuelle fournit une seule valeur comme meilleure estimation d'un paramètre inconnu, pour donner une plage de valeurs qui a une certaine probabilité de contenir ce paramètre, on a utilisé dans le le chapitre 3, la notion d'intervalles de confiance. Ce moyen de calcul est utilisé pour fournir une mesure de l'incertitude autour de l'estimation ponctuelle ceci en considérant un seuil ou un risque d'erreur. Dans le chapitre 4, on a étudié les tests statistiques, à savoir les tests paramétriques, dont les tests de conformité, le test d'homogéneité des populations, ainsi que les tests de comparaison de deux proportions, puis les tests non paramètriques, à savoir les tests d'indépendance et d'adéquation de Khi deux.

Chapitre 1

Rappels sur Lois de probabilités usuelles

1.1 Lois de probabilités discrètes

1.1.1 Loi de Bernouilli

Définition 1.1

Une épreuve de Bernoulli est une expérience aléatoire dont le résultat peut être une réussite , ou un échec, mais pas les deux simultanément.

Définition 1.2

Lors d'une épreuve de Bernoulli, soit $p \in [0,1]$ la probabilité d'un succès et $q = 1 - p$ la probabilité d'un échec Soit X_i la variable aléatoire définie par:

$$X_i = \begin{cases} 1, & \text{il y a réussite à la ième expérience} \\ 0 & \text{s'il y a échec.} \end{cases}$$

Dans ce cas, l'ensemble de définition de la variable aléatoire X_i est: $X_i(\Omega) = \{1, 0\}$.
On dit alors que la variable aléatoire X_i suit une loi de Brenouilli de paramètre p et on note : $X_i \rightsquigarrow \mathscr{B}(p)$.
On a les résultats suivants:

La loi de cette variable est donnée par:

$$P(X_i = x_i) = p^{x_i}(1-p)^{1-x_i}, \qquad x_i = 1, 0.$$

Son espérence et sa variance sont données par:

$$E(X_i) = p, \qquad \sigma^2(X_i) = p(1-p).$$

1.1.2 Loi Binômiale

Définition 1.3

Considérons la variable aléatoire de Bernouilli X_i et soit la variable $X = \sum_{i-1}^{n} X_i$ qui représente le nombre de succès dans le cas de Brenouilli. La variable aléatoire X suit une loi Binômiale de paramètres p et n, où p est le paramètre de Bernouilli et n le nombre de répétitions de l'expérience de Bernouilli.
On note $X \rightsquigarrow \mathscr{B}(n,p)$.

Si une variable aléatoire $X \rightsquigarrow \mathscr{B}(n,p)$, sa loi est donnée par:

$$P(X = x) = C_x^n p^x (1-p)^{n-x}, \qquad x = 0, ..., n.$$

Si une variable aléatoire $X \rightsquigarrow \mathscr{B}(n,p)$, alors :

$$E(X) = np, \quad \sigma^2(X) = np(1-p).$$

Remarque 1.1

-Si dans la loi précédente de la variable X qui suit une loi Binômiale, $n = 1$, on retrouve la loi de Bernouilli.
-Dans la pratiques, les probabilités sont calculées sur les tables de la loi Binômiale, avec différentes valeurs de n et de p. Mais quand n est assez grand, p est assez petit, la lecture de la table de la loi Binômiale, devient difficile voir impossible, pour ceci, on peut démontrer mathématiquement, que quand n est assez grand, p est assez petit, la loi $\mathscr{B}(n,p)$ est approximée à une autre loi discrète appelée loi de Poisson.

1.1.3 Loi de Poisson

Définition 1.4

Soit X une variable aléatoire discrète, On dit que X suit une loi de Poisson de paramètre λ, et on note $X \rightsquigarrow \mathscr{P}(\lambda)$,si sa loi de probabilité est donnée par:

$$P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0, ...$$

Remarque 1.2

Dans la démonstration, $\lambda = np$.
A noter que si: $X \rightsquigarrow \mathscr{P}(\lambda)$,

$$E(X) = \sigma^2(X) = \lambda.$$

1.2 Lois de probabilités continues

1.2.1 Loi exponentielle

Définition 1.5

Soit X une variable aléatoire continue, de densité de probabilité f. On dit que X suit une loi exponentielle de paramètre $\theta > 0$, et on note $X \rightsquigarrow \mathscr{E}(\theta)$, si sa densité de probabilité est donnée par :

$$f(x) = \begin{cases} \theta e^{-\theta x}, & six \geq 0 \\ 0 & sinon. \end{cases}$$

Dans ce cas,

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

En effet:

$$E(X) = \int_{-\infty}^{+\infty} \theta e^{-\theta x} dx = \int_{0}^{+\infty} \theta e^{-\theta x} dx = \frac{1}{\theta}.$$

1.2.2 Loi normale

Cette loi, appelée aussi loi de Gauss est très utile dans tous les domaines. Sous certaines conditions toutes les lois sont approximées à la loi normale de moyenne 0 et de variance 1.

Définition 1.6

Soit X une variable aléatoire continue, de densité de probabilité f. On dit que X suit une loi normale de paramètres m et σ^2 et on note $X \rightsquigarrow \mathscr{N}(m, \sigma^2)$, si sa densité de probabilité est donnée par :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(\frac{x-m}{\sigma}\right)^2}, \quad x \in \mathbb{R}.$$

Remarque 1.3

Si $X \rightsquigarrow \mathscr{N}(m, \sigma^2)$, alors $E(X) = m$, $\sigma^2(X) = \sigma^2$.
-Dans le cas où, $m = 0$, $\sigma^2 = 1$, on parlera de la loi normale centrée et réduite $\mathscr{N}(0, 1)$

Lecture de la table $\mathscr{N}(0, 1)$
Exemple 1 On suppose que les valeurs d'un produit sont distribuées selon une loi de Gauss(Normale) de moyenne m et de variance σ^2; dans une population, 70% des sujets ont une valeur de produit supérieure à 120 et 10% ont une valeur supérieure à 180. Quelles sont les valeurs de m et de σ?
Solution
Soit X la variable aléatoire qui représente les données du produit. On a par hypothèse :

$$P(X > 120) = 0.70, \quad P(X > 180) = 0.10.$$

En utilisant la seconde hypothèse, on aura :

$$1 - P(X \leq 180) = 0.10 \Rightarrow P(\frac{X - m}{\sigma} \leq \frac{180 - m}{\sigma}) = 0.90.$$

La variable aléatoire

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

donc la probabilité précédente est lue sur la table de la loi normale centrée et réduite de la façon suivante :
La valeur la plus proche à 0.90 sur la table de la loi normale centrée et réduite est: 0.8897, on projecte cette valeur sur la première colonne de la table, on trouve la valeur de 1.2, puis sur la première ligne, on trouve la valeur de 0.08, donc de la dernière probabilité, on aura :

$$\Phi(\frac{X - m}{\sigma}) = 0.90 \Rightarrow \frac{180 - m}{\sigma} = 1.28 = 1.2 + 0.08....(1).$$

(Voir Figure 1.1 ci dessous.) D'où la première équation.

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du$$

X	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936

Figure 1.1: Lecture de la table $\mathcal{N}(0, 1)$

En utilisant la première hypothèse, on aura :

$$1 - P(X \leq 120) = 0.70 \Rightarrow P(\frac{X - m}{\sigma} \leq \frac{120 - m}{\sigma}) = 0.3.$$

De là, on aura :

$$\Phi(\frac{120 - m}{\sigma}) = 0.3 \Rightarrow \Phi(-\frac{120 - m}{\sigma}) = 0.7.$$

Car la loi normale est symétrique et la valeur de 0.3 n'est pas sur la table de la loi normale centrée et réduite.
Par la même table $\mathcal{N}(0, 1)$, on trouve:

$$-\frac{120 - m}{\sigma} = 0.52...(2).$$

La valeur la plus proche(en valeur absolue) à 0.7 est 0.6987.
Il suffit de résoudre un système à deux équations (1) et (2) avec deux inconnues, m et σ, on aura : $m = 137.33$ et $\sigma = 33.33$.

1.2.3 Loi de Khi deux

Définition 1.7

Soit X une variable aléatoire qui suit la loi normale centrée réduite , la variable aléatoire définie par $Y = X^2$ suit la loi du khi-deux à un degré de liberté. On note $Y \rightsquigarrow \chi_1^2$

Remarque 1.4

Soit X une variable aléatoire telle que $X \rightsquigarrow \chi_1^2$, alors sa fonction de densité est donnée par:

$$f(X) = x^{\frac{-1}{2}} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}}, \quad x > 0.$$

Définition 1.8

Soit Y une variable aléatoire suivant une loi normale centrée et réduite. Soit $X_1, ..., X_n$ un échantillon aléatoire issu de la variable aléatoire Y.

Considérons la statistique $X = \sum_{i=1}^{n} X_i^2$ cette statistique suit une loi de Khi deux à n *ddl*. On note

$$X = \sum_{i=1}^{n} X_i^2 \rightsquigarrow \chi_n^2.$$

Si $X \rightsquigarrow \chi_n^2$, alors $E(X) = n$, $V(X) = 2n$.

Remarque 1.5

-La fonction de densité d'une variable aléatoire qui suit une loi de Khi deux à n ddl est donnée par:

$$\frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} e^{\frac{-x}{2}} x^{\frac{n}{2}-1}, \quad x > 0.$$

Γ *est la fonction Gamma.*

-Pour $n \geq 30$, $\sqrt{2\chi^2} - \sqrt{2n-1} \rightsquigarrow \mathcal{N}(0.1)$.

Lecture de la table de khi deux.
Exemple 2

$$P(\chi_6^2 \geq c) = 0.025 \Rightarrow c = 14.45.$$

(Voir Figure 1.2 ci dessous).

TABLE DU χ^2

La table donne la probabilité p pour que χ^2 égale ou dépasse une valeur donnée, en fonction du nombre de degrés de liberté dll

dll \ p	0,99	0,975	0,95	0,90	0,10	0,05	0,025	0,01
1	0,0002	0,001	0,004	0,016	2,71	**3,84**	5,02	**6,63**
2	0,02	0,05	0,10	0,21	4,61	**5,99**	7,38	**9,21**
3	0,11	0,22	0,35	0,58	6,25	**7,81**	9,35	**11,34**
4	0,30	0,48	0,71	1,06	7,78	**9,49**	11,14	**13,28**
5	0,55	0,83	1,15	1,61	9,24	**11,07**	12,83	**15,09**
6	0,87	1,24	1,64	2,20	10,64	**12,59**	14,45	**16,81**
7	1,24	1,69	2,17	2,83	12,02	**14,07**	16,01	**18,48**

Figure 1.2: Lecture de la table Khi deux

1.2.4 Loi de Student

Soient X et Y deux variables aléatoires telles que : $X \rightsquigarrow \mathcal{N}(0.1)$ et $Y \rightsquigarrow \chi^2_n$. Alors la statistique T définie ci dessous suit une loi de Student à n ddl. On écrit:

$$T = \sqrt{n}\frac{X}{\sqrt{Y}} \rightsquigarrow T_n.$$

Si $T = \sqrt{n}\frac{X}{\sqrt{Y}} \rightsquigarrow T_n$, alors $E(T) = 0$, $V(T) = \frac{n(n-1)}{2}$.
Si $n \geq 30$, la loi de Student est approchée par la loi normale centrée et réduite.

Lecture de la table de Student

Exemple 3 La loi de Studen test symétrique parrapportà zéro.Calculons $P(|T_{10}| > t) = 0.05$.

$$P(|T_{10}| > t) = 0.05 P(-t \leq T_{10} \leq +t) = 1 - 0.05 = 0.95.$$

Ceci nous donne,

$$P(T_{10} \leq +t) = 1 - \frac{0.05}{2} = 0.975 \Rightarrow t = 2.228.$$

1.2.5 Loi de Fisher snedecor

Définition 1.9

Soient X_1 et X_2 deux variables aléatoires suivant respectivement les lois n et m. Alors la variable aléatoire $F = \frac{X_1/n}{X_2/m}$ suit une loi de Fisher snedecor à (n, m) ddl. On note:

$$F = \frac{X_1/n}{X_2/m} \rightsquigarrow \mathscr{F}_{n,m}.$$

Lecture de la table de Fisher snedecor

Il est impossible de donner toutes les tables de Fisher pour les deux degrés de liberté.

A noter que $F_{n,m}$; à $95\% = \dfrac{1}{F_{m,n}}$ à 5%.

Quel est le $95e$ centile de la loi de Fisher avec 10 degrés de liberté au numérateur et 15 degrés de liberté au dénominateur?

On calcule donc $f_{10,15}$ à 95%. On trouve 2.544 (voir Figure 5.3). On peut aussi écrire: $F_{10,15,0.95} = 2.544$, donc $F_{15,10,0.05} = \frac{1}{2.544} = 0.393$.

-Quel est le 5e centile de la loi de Fisher avec 10 degrés de liberté au numérateur et 15 degrés de liberté au dénominateur?

On calcule donc $f_{10,15}$ à 5%. On trouve 0.3515.

QUANTILES D'ORDRE 0.95 DE LA LOI DE FISHER

Degrés de liberté du numérateur sur la première ligne
Degrés de liberté du dénominateur sur la colonne de gauche

	1	2	3	4	5	6	7	8	9	10
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.

Figure 1.3: Lecture de la table Fisher

Chapitre 2

Échantillonnage, Estimation ponctuelle

2.1 Introduction

En statistiques, les méthodes d'échantillonnage correspondent aux différentes manières de constituer un échantillon de la population étudiée. Si l'échantillon n'est pas constitué de manière aléatoire, il peut ne pas être représentatif de la population c'est-à-dire ne pas posséder les mêmes caractéristiques que la population que l'on souhaite étudier. Les résultats obtenus sur l'échantillon ne peuvent alors être extrapolés à la population. Il existe différentes méthodes d'échantillonnage, aléatoires ou non: Les méthodes non aléatoires et les méthodes aléatoires, dans ce mauscrit, on a utilisé les méthodes aléatoires où tous les éléments de la population ont la même probabilité de faire partie de l'échantillon

2.2 Definitions

Définition 2.1

On appelle échantillon aléatoire de taille n ou n-échantillon, une suite de n variables aléatoires $X_1, X_2, ..., X_n$ indépendantes et identiquement distribuées(iid), c'est à dire: ces variables aléatoires sont indépendantes et suivent toutes la même loi (que la loi d'une variable aléatoire X appelée la variable aléatoire mère).

Définition 2.2

On appelle statistique T liée à l'échantillon $X_1, ..., X_n$, la variable aléatoire $T(X_1, X_2, ..., X_n)$, fonction des variables aléatoires qui constituent l'échantillon aléatoire.

Examples 1 La moyenne empérique ou la moyenne de l'échantillon considéré $X_1, X_2, ..., X_n$, notée

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

est une statistique.
-La variance empérique (appelée aussi variance observée ou variance de l'échantillon) est une statistique, elle est donnée par :

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

Proposition 2.1

Soit X une variable aléatoire de loi normale de moyenne m et de variance σ^2, on note $X \rightsquigarrow \mathcal{N}(m, \sigma^2)$, soit $X_1, X_2, ..., X_n$ un échantillon aléatoire issu de la variable aléatoire X, alors l'espérance et la variance empirique de la moyenne empirique $\bar{X}_n$ sont données respectivement par:

$$\mathbb{E}(\bar{X}_n) = m, \quad \mathbb{V}(\bar{X}_n) = \frac{\sigma^2}{n}$$

2.2.1 Théorème central limite

.

Théorème 2.1 *Soit $X_1, X_2, ..., X_n$ issu d'une variable aléatoire X d'espérance $m < \infty$ et devariance $\sigma^2 < \infty$. Soit $\bar{S}_n = \sum_{i=1}^{n} X_i$, alors pour n assezgrand,lavariable aléatoire*

$$\frac{\bar{S}_n - \mathbb{E}(\bar{S}_n)}{\sqrt{\mathbb{V}(\bar{S}_n)}} \rightsquigarrow \mathcal{N}(0,1).$$

$\mathcal{N}(0,1)$: désigne la loi normale centrée et réduite (centrée: demoyenne0, réduite: de variance 1).

Exemple 4 appliquant ce théorème à la moyenne empirique $\bar{X}_n$.

$$Z = \frac{\bar{X}_n - \mathbb{E}(\bar{X}_n)}{\sqrt{\mathbb{V}(\bar{X}_n)}} = \frac{\bar{X}_n - m}{\sigma/\sqrt{n}} = \sqrt{n}\frac{\bar{X}_n - m}{\sigma} \rightsquigarrow \mathcal{N}(0,1)$$

Application:
Après la correction du concours de doctorat en mathématiques(comportant un grand nombre de candidats), on constate que les notes ont pour moyenne 12 et pour écart-type 3.
On se propose de prélever des échantillons aléatoires non exhaustifs (méthode d'échantillonnage où les individus sont sélectionnés avec remise dans la population) de 100 participants.

1. Quelle est la probabilité d'avoir la moyenne des notes de ces 100 candidats supérieure à 12.5?

2. Quelle est la probabilité d'avoir la moyenne des notes de ces 100 candidats comprise entre 12.5 et 12.9?

Solution

1. On calculera dans cette question $P(\bar{X}_n > 12.5)$.

$$P(\bar{X}_n > 12.5) = P\left(\frac{\bar{X}_n - \mathbb{E}(\bar{X}_n)}{\sqrt{\mathbb{V}(\bar{X}_n)}} > \frac{12.5 - \mathbb{E}(\bar{X}_n)}{\sqrt{\mathbb{V}(\bar{X}_n)}}\right)$$

On a $\mathbb{E}(\bar{X}_n) = m = 12$ et $\mathbb{V}(\bar{X}_n) = \frac{\sigma^2}{n} = \frac{9}{100} = 0.09$.
D'ou l'écart type $\sigma(\bar{X}_n) = \sqrt{0.09} = 0.3$.
Donc la probabilité demandée est:

$$P(\bar{X}_n > 12.5) = P\left(Z > \frac{12.5 - 12}{0.3}\right) = 1 - P\big(Z \leq 1.66\big)$$
$$= 1 - 0.9521 = 0.0479.$$

De la même manière, on calcule la seconde probabilité.

2.

$P(12.5 \leq \bar{X}_n < 12.9) = P(1.66 < Z < 3) = \Phi(3) - \Phi(1.66) = 0.9987 - 0.9521 = 0.0466.$

Φ est définie par :

$$\Phi(z) = P(Z \leq z).$$

C'est la fonction de répartition d'une variable aléatoire qui suit une loi $\mathscr{N}(0,1)$. (Voir Annexe)

Exemple 2: Fréquence
Soit X une variable aléatoire suivant une loi de Bernouilli (voir Annexe) de para mètre p. Considérons l'échantillon aléatoire $X_1, X_2, ..., X_n$ issu de la variable aléatoire X. Définissons la variable aléatoire

$$F = \frac{X_1 + X_2 + ... + X_n}{n}.$$

F est appelée fréquence, la variable aléatoire nF est la sommes des variables de Bernouilli, donc elle suit une loi Binômiale de paramètres n et p . On note $nF \rightsquigarrow \mathscr{B}(n,p)$. On a:

$$\mathbb{E}(nF) = np, \quad \mathbb{V}(nF) = np(1-p).$$

De là, on aura :

$$\mathbb{E}(F) = p, \quad \mathbb{V}(F) = \frac{p(1-p)}{n}.$$

En utilisant le thoérème central limite, on obtient :

$$Z_F = \frac{F - p}{\sqrt{\frac{p(1-p)}{n}}} \rightsquigarrow \mathscr{N}(0,1).$$

La différence entre F et $\bar{X}$ est que dans F e les variables aléatoires de l'échantillon sont des variables de Bernouilli.
Application

On sait par expérience qu'une certaine opération chirurgicale a 90% de chances de réussite. Cette opération est réalisée dans une clinique 400 fois chaque année. Soit la variable aléatoire X, le nombre de réussites dans une année.
Calculer la probabilité que la clinique réussisse au moins 345 opérations dans l'année.
Solution
Soit la variable aléatoire X_i définie par :

$$X_i = \begin{cases} 1 & \text{si la ième opération est réussie} \\ \\ 0 & \text{sinon} \end{cases}$$

Définissons la variable aléatoire $X = \sum_{i=1}^{n} X_i$, qui est le nombre d'opérations réussites

$$X \rightsquigarrow \mathscr{B}(400, 0.9), \mathbb{E}(X) = np = 360, \quad \mathbb{V}(X) = np(1-p) = 36.$$

$n = 400$ est assez grand, la probabilité demandée est :

$$\begin{aligned} P(X \geq 345) &= 1 - P(X < 345) = 1 - P(Z < \frac{346 - 360}{\sqrt{36}}) \\ &= 1 - \Phi(-2.5) = 1 - (1 - \Phi(2.5)) = 0.9938. \end{aligned}$$

La clinique réussira au moins 346 opérations dans l'année, à 99.38%.
À partir des caractéristiques(paramètres) d'un échantillon, que peut-on déduire des caractéristiques de la population dont il est issu?
La réponse à cette question est dans la partie de l'estimation ponctuelle.

2.3 Estimation ponctuelle

Une estimation ponctuelle est une valeur unique utilisée pour estimer un paramètre inconnu d'une population, en se basant sur des données d'un échantillon. C'est une estimation qui fournit un seul nombre, comme la moyenne de l'échantillon. Ceci consiste donc à donner des valeurs approchées aux paramètres d'une population P, m, σ^2 ou (proportion, moyenne, variance) à partir des données récoltées de l'échantillon $(f, \bar{x}_n, s_n'^2)$.

2.3.1 Estimateur

L'objectif de l'estimation est de donner une valeur approchée pour un paramètre θ de la population, ce paramètre peut être la moyenne m, la variance σ^2 (ou l'écart type σ), ou la fréquence p.

Définition 2.3

Considérons un échantillon aléatoire de taille n, $X_1, ..., X_n$, et soit $T(X_1, ..., X_n)$ une statistique, On dit que T est un estimateur d'un paramètre θ de la population, si les réalisations de T appelées estimations se rapprochent (dans un sens donné) vers la valeur de θ.

Exemple 5 $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur de m dans le cas d'une loi normale de moyenne m.

Biais

Définition 2.4

Un estimateur T d'un paramètre θ est dit ***sans biais*** *si*

$$\mathbb{E}(T) = \theta$$

Exemple 6 $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur sans biais de m.

En effet : $\mathbb{E}(\bar{X}_n) = m$.

Mais S_n^2 est estimateur **biaisé** de σ^2, car $\mathbb{E}(S_n^2) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$.

Définition 2.5

Un estimateur T est dit convergent si $\mathbb{E}(T)$ tend vers θ lorsque n tend vers l'infini.

Exemple 7 L'estimateur S_n^2 est convergent puisque :

$$E(S_n^2) = \frac{n-1}{n}\sigma^2 \xrightarrow{n\to\infty} \sigma^2.$$

2.4 Méthode de maximum de vraisemblance

Soit X une variable aléatoire réelle de loi paramétrique(c'est à dire dépendante d'un paramètre, discrète ou continue), dont on veut estimer le paramètre inconnu θ. Soit $X_1, ..., X_n$ un échantillon aléatoire issu de la variable X, de réalisations $x_1, ..., x_n$ Alors on définit une fonction g telle que:

$$g(x_i, \theta) = \begin{cases} P(x_i, \theta) & si \;\; X \;\; discrete \\ \\ f(x_i, \theta) & si \;\; X \;\; continue \end{cases}$$

A noter que P est la loi de probabilité de X, si celle ci est discrète et f est la loi de probabilité (densité de probabilité) de X, si celle ci est continue(voir Annexe).

Définition 2.6

Soit $X_1, ..., X_n$ unéchantillon aléatoire issu de la variable X, de réalisations $x_1, ..., x_n$. On appelle fonction de vraisemblance de θ, pour les réalisations $x_1, ..., x_n$, la fonction de θ définie par :

$$L(x_1, ..., x_n, \theta) = \prod_{i=1}^{n} g(x_i, \theta).$$

La méthode s'appelle méthode du maximum de vraisemblance consistant à estimer θ par la valeur qui maximise L. (vraisemblance). Ainsi, enpratique, la condition suivante va nous permettre de trouver la valeur de l'estimateur de θ noté $\hat{\theta}$.

$$\frac{\partial L(x_1, ..., x_n, \theta)}{\partial \theta} = 0.$$

On a aussi cette condition,

$$\frac{\partial^2 L(x_1,...,x_n,\theta)}{\partial^2\theta}\Big]_{\theta=\hat{\theta}} \leq 0.$$

Examples 2 1-Considérons un échantillon aléatoire $X_1,...,X_n$ issu d'une variable aléatoire $X \rightsquigarrow \mathscr{B}(n,p)$. On dit que la variable aléatoire X suit une loi Binômiale de paramètre p. L'inconnu est donc le paramètre p, en utilisant la méthode du maximumnde vraisemblance, on peut montrer que le meilleur estimateur de ce paramètre est :

$$\hat{p} = \bar{X}_n.$$

2.5 Variables de Student, Khi deux, Fisher

Comme noté précédemment, si la moyenne m de la population est connue(ceci n'est pas vrai dans la réalité), le meilleur estimateur de la variance est $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$. Mais dans la réalité m n'est pas connue, elle est remplacée dans ce dernier estimateur par $\bar{X}_n$, on aura donc un estimateur de la variance donné par: $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$. Or $\mathbb{E}(S_n^2) = \frac{n-1}{n}\sigma^2$, donc il est biaisé, il n'est donc pas le meilleur, l'estimateur corrigé est donc donné par :

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

Remarque 2.1

$$S_n'^2 = \frac{n}{n-1}S_n^2$$

La variable aléatoire

$$T = \frac{\bar{X}_n - m}{S'/\sqrt{n}} \rightsquigarrow T_{n-1}$$

On dit que la variable aléatoire T suit une loi de Student à $(n-1)$ degré de liberté $(n-1$ ddl$)$.Soit la variable aléatoire

$$\chi = \frac{(n-1)S'^2}{\sigma^2} \rightsquigarrow \chi^2_{n-1}$$

On dit que la variable aléatoire suit une loi de khi-deux à $(n-1)$ddl.

Remarque 2.2

Dans le cas des loi normale et Student, l'ensemble des données est symétrique et concentré près de la moyenne.

Chapitre 3

Estimation par Intervalles de confiance

3.1 Introduction

Soit $X_1, ..., X_n$ un échantillon aléatoire. A partir des réalisations $x_1, ..., x_n$, de l'échantillon, on construit un intervalle I tel qu'on peut affirmer avec la probabilité $1-\alpha$ que la valeur du paramètre θ se situe dans l'intervalle I: La probabilité $0 < \alpha < 1$ fixée à l'avance, s'appelle le **risque** ou le seuil de signification, et la probabilité $1 - \alpha$ s'appelle le **niveau de confiance**. On dit que I est l'intervalle de confiance au risque α (ou de niveau $1 - \alpha$). On a donc la définition suivante:

Définition 3.1

Soit X une variable aléatoire dont la loi dépend d'un paramètre θ inconnu. Soit $X_1, ..., X_n$ un échantillon aléatoire issu de la variable aléatoire X et soit $\alpha \in]0, 1[$.

Un intervalle de confiance pour le paramètre θ, de niveau de confiance $1 - \alpha \in]0, 1[$, est l'intervalle qui a la probabilité $1 - \alpha$ de contenir la vraie valeur du paramètre θ.

3.2 Construction

Le principe de la méthode d'estimation par intervalle de confiance est le suivant: Soit T un estimateur de θ (meilleur estimateur possible) dont on connait la loi en fonction de θ. On détermine par la suite un intervalle de probabilité $1 - \alpha$ pour T :

$$P(t_1(\theta) \leq T \leq t_2(\theta)) = 1 - \alpha.$$

Il faut ensuite inverser cet intervalle pour T, dont les bornes dépendent de θ pour obtenir un intervalle de la forme:

$$P(t_2^{-1}(\theta) \leq \theta \leq t_1^{-1}(\theta)) = 1 - \alpha.$$

Par exemple une réduction de mortalité de 20% avec Intervalle de Confiance à 95% $[-35\%; -5\%]$ signifie qu'au mieux la réduction de mortalité est de 35% (borne inférieure), au pire de 5% (borne supérieure), ces 2 valeurs ne sont pas significativement différentes du résultat rapporté (20%). On verra par la suite que la taille de l'intervalle de Confiance est inversement proportionnelle à la taille de l'échantillon: plus ce dernier sera grand, plus l'intervalle de Confiance sera étroit, donc précis.

3.3 Intervalles de confiance classiques

3.3.1 Intervalle de confiance pour la moyenne m

Soit $X_1, ..., X_n$ un échantillon aléatoire issu de la variable aléatoire X suivant une loi $\mathcal{N}(m, \sigma^2)$.
$\bar{X}_n$ est le meilleur estimateur de m, On distinguera deux cas:
$\underline{\sigma \text{ connu}}$ (cas peu réaliste)

$$P(-z_{\frac{\alpha}{2}} \leq \frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}) = 1 - \alpha$$

Donc:

$$P(\bar{X}_n - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq m \leq \bar{X}_n + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

Ainsi, l'intervalle de confiance de m est donné par:

$$IC_m = [\bar{x}_n - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$$

Où $z_{\frac{\alpha}{2}}$ est telle que:

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

Autrement dit:

$$z_{\frac{\alpha}{2}} = \Phi^{-1}(1 - \frac{\alpha}{2})$$

Voir figure ci dessous(Figure3.1).

Remarque 3.1

La table de la loi normale $\mathcal{N}(0, 1)$ est lue de deux manière différentes:

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

Ou bien

$$P(Z \geq z_{\frac{\alpha}{2}}) = 1 - \Phi(z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$$

Exemple8 Les notes à un examen national de sciences sont distribuées normalement avec un écart type de 13 points. Dans un échantillon de 30 élèves, la moyenne observée est de 111 points.

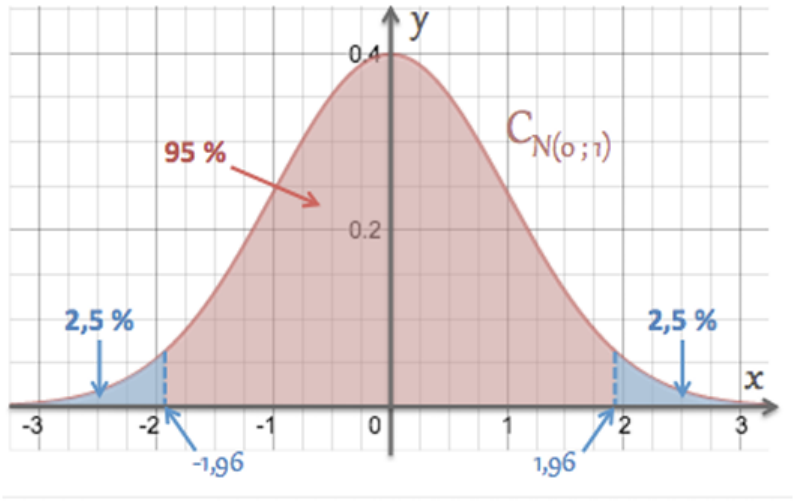

Figure 3.1: Loi normale

1-Estimer la moyenne m des résultats avec un intervalle de confiance à 95%, à 90%, puis à 99% Conclure.

$$IC_m = [\bar{x}_n - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}]$$

σ dans ce cas est connu, la moyenne de l'échantillon est donnée. Il reste à déterminer, $z_{\frac{\alpha}{2}}$ pour les différents niveaux.
$z_{\frac{\alpha}{2}} = \Phi^{-1}(1 - \frac{\alpha}{n})$ pour

$$\alpha = 0.05(1 - \alpha = 0.95), \quad z_{\frac{\alpha}{2}} = 1.96$$

La valeur trouvée est lue sur la table de la loi normale $\mathcal{N}(0, 1)$. (Voir Annexe).
Pour $\alpha = 0.05$,

$$I_{0.95}C_m = [111 - 1.96\frac{1.3}{\sqrt{30}}, 111 + 1.96\frac{1.3}{\sqrt{30}}] = [106.35, 115.65].$$

Pour $\alpha = 0.1$, $z_{\frac{\alpha}{2}} = 2.575$

$$I_{0.90}C_m = [111 - 2.575\frac{1.3}{\sqrt{30}}, 111 + 2.575\frac{1.3}{\sqrt{30}}] = [104.9, 117.1].$$

Pour $\alpha = 0.01$, $z_{\frac{\alpha}{2}} = 1.64$

$$I_{0.99}C_m = [111 - 1.64\frac{1.3}{\sqrt{30}}, 111 + 1.64\frac{1.3}{\sqrt{30}}] = [107.1, 114.9].$$

On peut remarquer que :

$$I_{0.90}C_m \subset I_{0.95}C_m \subset I_{0.99}C_m$$

La longueur d'un intervalle de confiance est égale à deux fois l'erreur , dans notre cas, par exemple pour $\alpha = 0.05$,

$$L(I_{0.95}C_m) = 2 \times z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} = 2 \times 1.96\frac{1.3}{\sqrt{30}} = 9.304.$$

Si la confiance $(1-\alpha)$ diminue, la longueur de l'intervalle diminue et le risque α augmente.

σ inconnu

$$P(-t_{\frac{\alpha}{2}} \leq \frac{\bar{X}_n - m}{\frac{S'_n}{\sqrt{n}}} \leq t_{\frac{\alpha}{2}}) = 1-\alpha$$

Donc:

$$P(\bar{X}_n - t_{\frac{\alpha}{2}} \frac{S'_n}{\sqrt{n}} \leq m \leq \bar{X}_n + t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

Ainsi, l'intervalle de confiance de m est donné par:

$$ICm = [\bar{x}_n - t_{\frac{\alpha}{2}} \frac{s'_n}{\sqrt{n}}, \bar{z}_n + t_{\frac{\alpha}{2}} \frac{t'_n}{\sqrt{n}}]$$

Où $t_{\frac{\alpha}{2}}$ est telle que:

$$P(T \leq t_{\frac{\alpha}{2}}) = \mathscr{T}(t_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$$

Autrement dit:

$$t_{\frac{\alpha}{2}} = (\mathscr{T})^{-1}(1 - \frac{\alpha}{2})$$

A noter que $\mathscr{T}$ est la fonction de répartition d'une variable aléatoire qui suit une loi de Student. Voir figure ci dessous (Figure3.2).

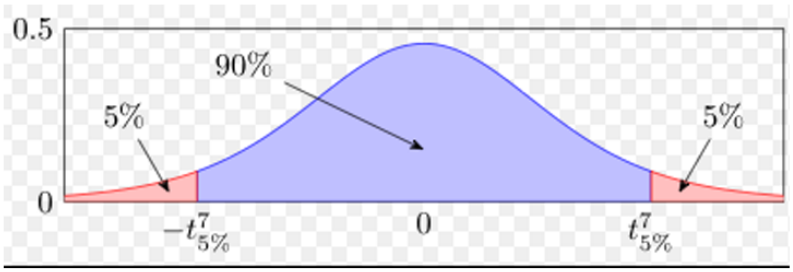

Figure 3.2: Loi de student

Exemple 9 Lors d'une en quête sur la durée de sommeil des enfants de 2 à 3 ans effectuée sur un échantillon de 540 enfants d'une région on a trouvé une moyenne du temps de somme il par nuit de 11, 7 heures. L'écart-type est $1,3$ heures. On veut connaitre à 95%, la moyenne générale du temps de somme il chez tous les enfants de cette région.

Les données de l'échantillon de taille $n = 540$ sont $\bar{x}_n = 11.7h$, $s_n = 1.3$,calculons s'_n.

$$s'_n = \sqrt{\frac{n}{n-1}} s_n = \sqrt{\frac{540}{539}}(1.3) = 1.301.$$

$t_{\frac{\alpha}{2}} = (\mathscr{T})^{-1}(1 - \frac{0.05}{2}) = 0.975 \Rightarrow t_{\frac{\alpha}{2}} = 1.96$. Ainsi l'intervalle de confiance de la moyenne du somme il chez le senfants de la région est:

$$\begin{aligned} IC_m &= [\bar{x}_n - t_{\frac{\alpha}{2}} \frac{s_n^{'}}{\sqrt{n}}, \bar{x}_n + t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}] \\ &= [11.7 - 1.96 \times \frac{1.301}{\sqrt{540}}, 11.7 + 1.96 \times \frac{1.301}{\sqrt{540}}] = [11.60, 11.79]h. \end{aligned}$$

3.3.2 Intervalle de confiance pour la variance

Le meilleur estimateur de la variance est $S_n^{'2}$, la statistique qui regroupe l'inconnue et son estimateur est:

$$\frac{(n-1)S_n^{'2}}{\sigma^2} \rightsquigarrow \chi^2_{n-1}$$

A la différence des loi normale et de Student, la loi de Khi deux n'est pas symétrique. On doit trouver deux constantes c_1 et c_2 de la table de la loi de Khi-deux, telles que:

$$P(c_1 \leq \frac{(n-1)S_n^{'2}}{\sigma^2} \leq c_2) = 1 - \alpha$$

En inversant, on aura:

$$P(\frac{(n-1)S_n^{'2}}{c_2} \leq \sigma^2 \leq \frac{(n-1)S_n^{'2}}{c_2 1}) = 1 - \alpha$$

Ainsi, l'intervalle de confiance de la variance est donné par:

$$IC(\sigma^2) = [\frac{(n-1)s_n^{'2}}{c_2}, \frac{(n-1)s_n^{'2}}{c_1}]$$

Où, c_1 et c_2 sont telles que:

$$P(\chi^2_{n-1} \leq c_1) = \frac{\alpha}{2}, P(\chi^2_{n-1} \geq c_2) = \frac{\alpha}{2}.$$

Voir figure ci dessous(Figure3.3).

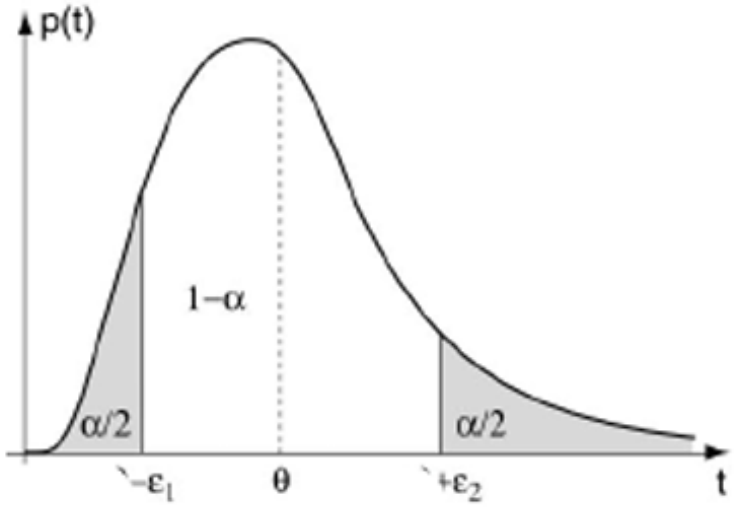

Figure 3.3: Loi de Khi-deux

Chapitre 4

Tests statistiques

4.1 Introduction

La théorie des tests statistiques est une procédure qui consiste à prendre une décision entre deux hypothèses (une hypothèse nulle, H_0, et une hypothèse alternative, H_1) en se basant sur un échantillon de données. Le processus implique de formuler des hypothèses, de définir les risques d'erreur (α et β), de calculer une statistique de test à partir des données, puis de comparer le résultat à un seuil prédéfini pour décider de rejeter ou non l'hypothèse nulle.
Les deux hypothèses font l'objet d'un test d'hypothèses qui permet de contrôler leur validité (accepter ou rejeter) à partir de l'étude d'un ou plusieurs échantillons aléatoires.
Il existe plusieurs types de tests, ils diffèrent en fonction de l'hypothèse de travail, dans ce manuscrit on distingue:

1. Les tests destinés à vérifier si un échantillon peut être considéré comme extrait d'une population donnée, vis-à-vis d'un paramètre comme la moyenne ou la fréquence observée, sont nommés **tests de conformité** .

2. Les tests destinés à comparer plusieurs populations à l'aide d'un nombre équivalent d'échantillons; ce sont les tests d'égalité nommés **test d'homogénéité**.

3. On peut ajouter à cette catégorie le **test d'indépendance** qui cherche à tester l'indépendance entre deux caractères qualitatifs et le **test d'adéquation** qui cherche à vérifier si une variable aléatoire a pour une fonction de répartition, la fonction F ou une autre fonction de répartition. On peut classer les tests selon leurs objectifs (ajustement, indépendance, de moyenne, de variance ...etc).

4.2 Définitions

Définition 4.1

Une hypothèse statistique est une assertion (affirmation) sur la distribution d'une ou plusieurs variables ou sur les paramètres des distributions.

Il existe deux types d'hypothèses: hypothèse simple et multiple.
Si l'hypothèse est réduite à un seul paramètre de l'espace des paramètres, on parlera d'hypothèse simple(appelée hypothèse nulle) sinon on parlera d'hypothèse multiple(qu'on appellera hypothèse alternative).
La formulation d'un test statistique est la suivante :

$$\begin{cases} H_0 : \theta = \theta_0 \\ \\ H_1 : \theta \neq \theta_0 (\theta > \theta_0, \theta < \theta_0) \end{cases}$$

Définition 4.2

Si l'hypothèse alternative est telle que $H_1 : \theta > \theta_0, \theta < \theta_0$, on parlera d'un test unilatéral(unilatéral à gauche ou unilatéral à droite). Voir figures ci dessous Figure 4.1 et Figure 4.2.

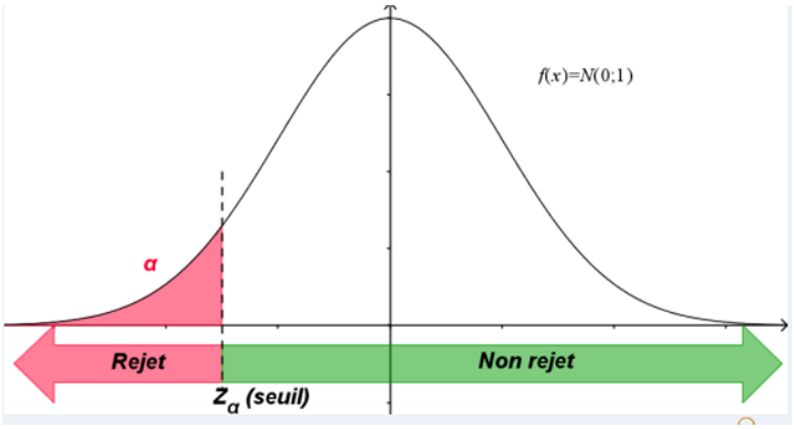

Figure 4.1: test unilatéral gauche

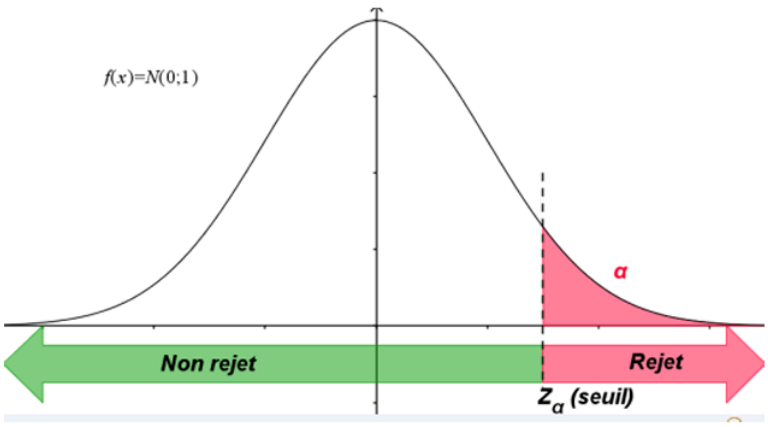

Figure 4.2: test unilatéral droit

Définition 4.3

Si l'hypothèse alternative est telle que $H_1 : \theta \neq \theta_0$, on parlera d'un test bilatéral. Voir figure ci dessous Figure 4.3.

On peut remarquer des figures, qu'il y a deux zones, la zone du rejet et la zone d'acceptation.

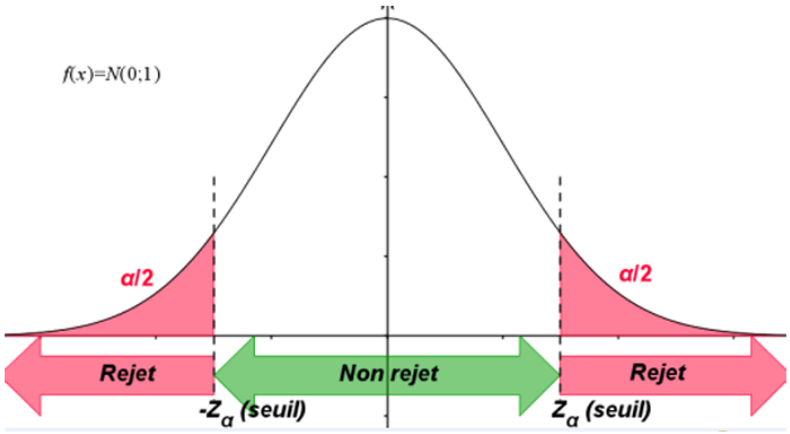

Figure 4.3: test bilatéral

A noter que T est la statistique du test.
Les étapes à suivre pour tester une hypothèse sont les suivantes:

1. définir l'hypothèse nulle (notée H_0) à contrôler,
2. choisir un test statistique ou une statistique pour contrôler l'hypothèse nulle H_0,
3. définir la distribution de la statistique sous l'hypothèse (H_0 est réalisée),
4. définir le niveau de signification du test noté α ,
5. calculer, à partir des données fournies par l'échantillon, la valeur de la statistique,
6. prendre une décision concernant l'hypothèse posée et conclure.

Lorsqu'on effectue un test statistique, on peut accepter à tort l'une des hypothèses, comme on peut les rejeter à tort. On peut donc commettre des erreurs de choix.

4.2.1 Types d'erreurs

Les types d'erreurs qu'on peut commettre dans le choix des hypothèses sont résumés dans le tableau suivant :

Définition 4.4

Décision / Vérité	H_0	H_1
H_0	1-α	β
H_1	α	1-β

Le risrque de première espèce α est définit par :

$$\alpha = P(H_1/H_0 \ \ vraie).$$

Le risque de second espèce est :

$$\beta = P(H_0/H_1 \ \ vraie)$$

.

Remarque 4.1

α est appelé aussi risque d'erreur (du test) ou seuil de signification, $1-\beta$ est appelé puissance du test.

4.2.2 P-valeur : P-value

Définition 4.5

Dan sun test statistique, la valeur-p (en anglais p-value pour probability value), est la probabilité pour un modèle statistique donné sous l'hypothèse nulle d'obtenir une valeur au moins aussi extrême que celle observée.

Dans un test bilatéral, la P-value est définie comme suit

$$P = P(T < -|x_0| \ \ ou \ \ T > +|x_0|).$$

Si la loi de la variable X est symétrique, alors: $P = 2P(T > +|x_0|)$.
Pour un test unilatéral droit,

$$P = P(T \geq x_0).$$

Pour un test unilatéral gauche,

$$P = P(T \leq x_0).$$

4.3 Tests paramétriques

4.3.1 Tests de conformité

Le test de conformité permet de vérifier si un échantillon est représentatif ou non d'une population vis-à-vis d'un paramètre donné (la moyenne, la variance ou la fréquence ...). Pour effectuer ce test, la loi théorique doit être connue au niveau de la population.
a-Comparaison d'une moyenne observée et une moyenne théorique

Soit X une variable aléatoire observée dans une population, suivant une loi normale de paramètres m et σ. Soit un échantillon tiré aléatoirement de cette population de même loi (loi normale).
On veut tester :

$$\begin{cases} H_0 : m = m_0 \\ \\ H_1 : m \neq m_0 \quad (1). \end{cases}$$

La statistique du test varie selon si la variance σ^2 de la population de référence est connue ou non.

a- σ connue

Soit une variable aléatoire X suit la loi normale de moyenne m inconnue et de variance connue σ^2; On considère un échantillon aléatoire issu de cette variable aléatoire X. On sait que : $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ vérifie :

$$\frac{\bar{X}_n - m}{\sigma/\sqrt{n}} \rightsquigarrow \mathcal{N}(0,1),$$

ceci d'après le théorème centrale limite ($n > 30$).
La statistique du test (1) est donc :

$$Z = \frac{\bar{X}_n - m}{\sigma/\sqrt{n}}$$

Notons par:

$$z_{obs} = \frac{\bar{x}_n - m}{\sigma/\sqrt{n}}$$

z_{obs} est la valeur observée de Z (réalisation de Z sur l'échantillon).
Le test est bilatéral (Dans H_1, on a $m \neq m_0$, par contre dans le test unilatéral, on a :$H_1 : m > m_0$ ou $H_1 : m < m_0$).
z_{obs} est comparée, en valeur absolue, à la valeur z_{seuil} lue sur la table de la loi normale centrée et réduite pour un risque d'erreur fixé. z_{seuil} est le fractile de la loi normale centrée et réduite d'ordre $1 - \frac{\alpha}{2}$. Par exemple, pour un risque de première espèce $\alpha = 5\%$, la valeur de z_{seuil} est le fractile de la loi normale centrée et réduite d'ordre $0,975$ est égale à $1,96$ (Voir Annexe).

Décision du test:

-si $|z_{obs}| > |z_{seuil}|$, l'hypothèse H_0 est rejetée au risque d'erreur α .
-si $|zobs| < |zseuil|$,l'hypothèse H_0 est acceptée au risque d'erreur α: l'échantillon est représentatif de la population de référence d'espérance m_0 au risque d'erreur α.

b-σ^2 inconnue

Soit une variable aléatoire X suit la loi normale de moyenne m inconnue et de variance inconnue; On considère un échantillon aléatoire issu de cette variable Xi vérifie : aléatoire X. On sait que : $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ vérifie:

$$\frac{\bar{X}_n - m}{s'/\sqrt{n}} \rightsquigarrow T_{n-1}$$

La statistique du test (1) est donc

$$T = \frac{\bar{X}_n - m}{s'/\sqrt{n}}$$

Notons par :

$$t_{obs} = \frac{\bar{x}_n - m}{s'/\sqrt{n}}$$

Décision du test:

- si $|t_{obs}| > |t_{seuil}|$, l'hypothèse H_0 est rejetée au risque d'erreur α: l'échantillon appartient à une population d'espérance m.
- si $|t_{obs}| < |t_{seuil}|$, l'hypothèse H_0 est acceptée au risque d'erreur α: l'échantillon est représentatif de la population de référence d'espérance m_0 au risque d'erreur α. la valeur t_{seuil} lue sur la table de la loi de Student pour un risque d'erreur α fixé. t_{seuil} est le fractile de la loi Student à $(n-1)$ ddl d'ordre $1 - \frac{\alpha}{2}$ car le test est bilatéral. Par exemple, pour un risque de première espèce $\alpha = 5\%$, et pour une taille de l'échantillon égale à 33, la valeur de t_{seuil} est le fractile de la loi Student à 32 ddl d'ordre de $0,975$ est égale à $2,037$.

b-Comparaison d'une fréquence observée et une fréquence théorique

Soit X une variable qualitative prenant deux modalités (succès $X = 1$, échec $X = 0$); observée sur une population, et un échantillon tiré au hasard de cette population. La probabilité p est la probabilité du succès $P(X = 1) = p$ et la probabilité d'échec est $P(X = 0) = 1 - p$. Le but est de savoir si un échantillon (de grande taille) de fréquence observée $F_{obs} = F/n$, estimation de p, appartient à une population de référence connue de fréquence p_0 (H_0 vraie) ou à une autre population inconnue de fréquence p (H_1 vraie).
On teste donc :

$$\begin{cases} H_0 : p = p_0, \\ \\ H_1 : p \neq p_0, . \end{cases}$$

Soit F_{obs} la réalisation de la variable aléatoire F qui est la fréquence empirique (l'équivalence de $\bar{X}_n$ lorsque la variable aléatoire X est quantitative) et qui suit approximativement une loi normale, c'est à dire :

$$F \rightsquigarrow N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

La statistique du test est :

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

Grâce au théorème central-limite, cette variable aléatoire Z suit approximative ment la loi normale centrée réduite $Z \rightsquigarrow N(0,1)$ si seulement si $n > 30, np_0 \geq 5$.
A partir de l'échantillon, on calcule :

$$z_{obs} = \frac{F_{obs} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

-si $|z_{obs}| > |z_{seuil}|$, l'hypothèse H_0 est rejetée au risque d'erreur α.
-si $|z_{obs}| < |z_{seuil}|$,l'hypothèse H_0 est acceptée au risque d'erreur α.

Example 10: Test d'hypothèse sur la moyenne d'une résistance de matériau.

La résistance à la traction d'un nouveau type d'acier est supposée suivre une loi normale de moyenne inconnue m et d'écart type connu $\sigma = 13$ MPa.
D'après les normes industrielles, la résistance moyenne doit être de $m_0 = 120$ MPa.
Un échantillon de $n = 30$ pièces testées donne une moyenne observée $\bar{x} = 115.5$ MPa.
Peut-on conclure, au niveau de signification $\alpha = 0.05$, que le matériau ne respecte pas la norme ? Les hypothèses sont les suivantes:

$$H_0 : m = 120 = m_0 \qquad \text{contre} \qquad H_1 : m \neq 120$$

(test bilatéral).

2. Statistique de test Comme σ est connue, la statistique de test est

$$Z = \frac{\bar{X} - m_0}{\sigma/\sqrt{n}}.$$

Calculons :

$$\frac{\sigma}{\sqrt{n}} = \frac{13}{\sqrt{30}} \approx 2.3735,$$

$$Z_{\text{obs}} = \frac{115.5 - 120}{2.3735} \approx -1.895.$$

3. Zone de rejet Pour un test bilatéral à $\alpha = 0.05$:

$$z_{1-\alpha/2} = z_{0.975} \approx 1.96.$$

On rejette H_0 si $|Z_{\text{obs}}| > 1.96$.
Or $|Z_{\text{obs}}| = 1.895 < 1.96$, donc on **ne rejette pas** H_0.

4. p-valeur

$$p\text{-value} = 2\big(1 - \Phi(|Z_{\text{obs}}|)\big) = 2(1 - \Phi(1.895)) \approx 2(1 - 0.9712) = 0.0576.$$

Puisque $p = 0.0576 > 0.05$, on ne rejette pas H_0.

5. Intervalle de confiance (95%)

$$\bar{x} \pm z_{0.975}\frac{\sigma}{\sqrt{n}} = 115.5 \pm 1.96 \times 2.3735 = 115.5 \pm 4.65.$$

$$\text{IC}_{95\%}(\mu) = [110.85,\ 120.15].$$

La valeur 120 appartient à cet intervalle, ce qui confirme la Au niveau $\alpha = 0.05$, les données ne permettent pas de conclure que la moyenne de résistance diffère de 120 MPa. L'échantillon observé est compatible avec la norme exigée, bien que la moyenne soit légèrement inférieure.

4.3.2 Tests d'égalité ou d'homogénéité des populations

Soient X_1 et X_2 deux variables aléatoires quantitatives continues observées sur deux populations indépendantes suivant une loi normale de paramètres (m_1, σ_1^2) et (m_2, σ_2^2) respectivement.
On tire deux grands échantillons indépendants l'un de l'autre de ces deux populations de tailles n_1 et n_2 respectivement et n_1 et n_2 sont supposées supérieures à 30. On vérifie que dans les deux populations, les espérances sont égales.
Les hypothèses à tester sont: $H_0 : m_1 = m_2$, contre $H_1 : m_1 \neq m_2$. Pour ceci on envisagera deux cas:
Les variances des deux populations sont connues
Notons par $\bar{X}_1$ la moyenne du premier échantillon (de taille n_1) et $\bar{X}_2$ la moyenne du second échantillon(de taille n_2).
On a $\sigma^2(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.
La statistique du test est :

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \rightsquigarrow \mathcal{N}(0, 1).$$

On calcule :

$$z_{obs} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \rightsquigarrow \mathcal{N}(0, 1).$$

A noter que z_{obs} est calculée sous l'hypothèse H_0, c'est à dire la différence $m_1 - m_2 = 0$.
-si $|z_{obs}| > |z_{seuil}|$, l'hypothèse H_0 est rejetée au risque d'erreur α.
-si $|z_{obs}| < |z_{seuil}|$, l'hypothèse H_0 est acceptée au risque d'erreur α: Les deux échantillons sont issues des deux populations ayant des espérances statistiquement égales à m.
Les deux variances sont inconnues
Avant de tester les hypothèses dans ce cas, on doit vérifier l'égalité des variances(sous section suivante, comparaison de deux variances); On envisagera deux cas:
Les deux variances sont égales: On considérera alors la statistique

$$S'^2 = \frac{(n_1 - 1)S_1'^2 + (n_2 - 1)S_2'^2}{n_1 + n_2 - 2}$$

Où $S_i'^2$ est le meilleur estimateur de la variance σ_i^2 respectivement $(i = 1, 2)$.
La statistique du test est :

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{S'\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \rightsquigarrow T_{n_1+n_1-2}$$

On calcule l'estimation de T, comme suit :

$$T_{obs} = \frac{\bar{x}_1 - \bar{x}_2}{s'\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

on la compare par la suite à t_{seuil} lue sur la table de la loi de Student à $(n_1 + n_2 - 2)$ddl(Voir Annexe).

-Si $|T_{obs}| > |t_{seuil}|$, l'hypothèse H_0 est rejetée au risque d'erreur α: Les deux échantillons sont issues des deux populations ayant des espérances statistiquement et significativement différentes.
-Si $|T_{obs}| < |t_{seuil}|$, l'hypothèse H_0 est acceptée au risque d'erreur α.
Les deux variances sont différentes A noter que les hypothèses sont les mêmes que dans le cas précédent, la statistique du test est :

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{\sqrt{\frac{S_1'^2}{n_1} + \frac{S_2'^2}{n_2}}} \rightsquigarrow T_{n_1+n_1-2}.$$

On calcule l'estimation de T, comme suit :

$$T_{obs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1'^2}{n_1} + \frac{S_2'^2}{n_2}}}$$

on la compare par la suite à t_{seuil} lue sur la table de la loi de Student à (n_1+n_2-2)ddl.

- si $|T_{obs}| > |t_{seuil}|$, l'hypothèse H_0 est rejetée au risque d'erreur α: Les deux échantillons sont issues des deux populations ayant des espérances statistiquement et significativement différentes.
- si $|T_{obs}| < |t_{seuil}|$, l'hypothèse H_0 est acceptée au risque d'erreur α.

Comparaison des variances
Soit X une variable aléatoire observée sur deux populations indépendantes suivant une loi normale. On tire deux échantillons de ces deux différentes populations, indépendants l'un de l'autre, de tailles respectives n_1 et n_2. On veut savoir si les variances, dans ces deux populations, sont égales (ou les différences potentielles sont statistiquement non significatives au seuil α fixé) ou bien ces différences sont statistiquement significatives au seuil α . Le test à réaliser est :

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

La statistique du test est :

$$F_S = \frac{S_1'^2}{S_2'^2}$$

La statistique F_S suit une loi de Fisher snedecor à $(n_1 - 1, n_2 - 1)$ddl (voir Annexe). L'estimation ou l'observée de cette dernière statistique est donnée par :

$$F_{Sobs} = \frac{s_1'^2}{s_2'^2}$$

Remarque 4.2

On pratique, on met toujours au numérateur la plus grande des deux variances ($s_{n1}'^2 > s_{n2}'^2$)*, pour que le rapport des variances soit supérieur à* 1*.*

Pour prendre une décision, la valeur de F_{Sobs} est comparée à la valeur théorique F_{seuil}, lue dans la table de Fisher-Snedecor de degré de liberté $(n_1 - 1; n_2 - 1)$ et pour un risque d'erreur α fixé.

- si $F_{Sobs} > F_{seuil}$ l'hypothèse H_0 est rejetée au risque d'erreur α: Les deux échantillons sont issues des deux populations ayant des variances significativement différentes au risque d'erreur α .
- Si $F_{Sobs} < F_{seuil}$ l'hypothèse H_0 est acceptée au risque d'erreur α: Les deux échantillons sont issues des deux populations ayant des variances significativement égales au risque d'erreur α.

4.3.3 Comparaison de deux propositions (grands échantillons)

Soient p_1 et p_2 deux proportions d'individus d'une certaine modalité A dans les populations M_1 et M_2. n_1 et n_2 sont supposés suffisamment grands (supérieurs à 30). Notons par f_1 et f_2 les estimations de p_1 et p_2 respectivement.
On cherche à tester :

$$\begin{cases} H_0 : p_1 = p_2 = p \\ \\ H_1 : p_1 \neq p_2 \end{cases}$$

D'après ce qu'on a vu précédemment : f_1 et f_2 sont des réalisations des variables

$$\frac{F_i = \sum_{i=1}^{n_i} X_i}{n_i} \quad i = 1, 2$$

On sait que

$$F_1 \rightsquigarrow N\left(p_1, \sqrt{\frac{p_1(1-p_1)}{n_1}}\right)$$

et

$$F_2 \rightsquigarrow N\left(p_2, \sqrt{\frac{p_2(1-p_2)}{n_2}}\right)$$

la loi de la variable $F_1 - F_2$ sera donc donnée par :

$$F_1 - F_2 \rightsquigarrow N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

On définit la variable aléatoire :

$$Z = \frac{F_1 - F_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

Grâce au théorème central limite, la variable aléatoire Z suit une loi normale centrée et réduite $N(0, 1)$.
La valeur de p probabilité de succès commune auxdeuxpopulations n'est en réalité

pas connue, on l'estime par les résultats observés dans les deux échantillons par son estimation :

$$\hat{p} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}, \quad \hat{q} = 1 - \hat{p}$$

où $f_i = \frac{k_i}{n_i}$, $i = 1, 2$, et k_i représente le nombre de succès observés dans l'échantillon 1 et 2 respectivement.
Il s'agit d'un test bilatéral, la valeur de Z observée dans les deux échantillons est :

$$z_{obs} = \frac{f_1 - f_2}{\sqrt{\hat{p}\hat{q}(\frac{1}{n_1} + \frac{1}{n_2})}}$$

La valeur de z_{obs} est comparée, en valeur absolue, à la valeur théorique z_{seuil} lue dans la table de la loi normale centrée et réduite pour un risque d'erreur α fixé. z_{seuil} est le fractile de la loi normale centrée réduite d'ordre $1 - \frac{\alpha}{2}$.
-Si $|Z_{obs}| > z_{seuil}$, l'hypothèse H_0 est rejetée au risque d'erreur α: Les deux échantillons sont issues des deux populations ayant des probabilités de succès statistiquement et significativement différentes p_1 et p_2, au risque d'erreur α.
-Si $|Z_{obs}| < z_{seuil}$, l'hypothèse H_0 est acceptée au risque d'erreur α: Les deux échantillons sont issues des deux populations ayant des probabilités de succès statistiquement et significativement égales p_1 et p_2, au risque d'erreur α.

Exemple 11 Une entreprise souhaite comparer l'efficacité de deux procédés de fabrication d'un même alliage métallique. On mesure la résistance à la traction (en MPa) pour deux échantillons indépendants :

- Procédé A : $n_1 = 16$, $\bar{x}_1 = 115.2$, $s_1^2 = 25$
- Procédé B : $n_2 = 14$, $\bar{x}_2 = 110.4$, $s_2^2 = 20$

On suppose que les résistances suivent des lois normales et que les variances sont égales ($\sigma_1^2 = \sigma_2^2$). Peut-on affirmer, au seuil de signification $\alpha = 0.05$, que les deux procédés diffèrent en moyenne ? On teste alors ces deux hypothèses:

$$H_0 : m_1 = m_2 \qquad \text{contre} \qquad H_1 : m_1 \neq m_2$$

(test bilatéral).

Lorsque les variances sont supposées égales, on utilise la variance combinée :

$$s'^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Calculons :

$$s_p^2 = \frac{(16 - 1) \times 25 + (14 - 1) \times 20}{16 + 14 - 2} = \frac{15 \times 25 + 13 \times 20}{28} = \frac{375 + 260}{28} = 22.68.$$

Ainsi, $s_p = \sqrt{22.68} \approx 4.76$.

La statistique de test est alors :

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

$$t_{\text{obs}} = \frac{115.2 - 110.4}{4.76 \times \sqrt{\frac{1}{16} + \frac{1}{14}}} = \frac{4.8}{4.76 \times 0.370} = \frac{4.8}{1.762} \approx 2.725.$$

Sous H_0, la statistique t suit une loi de Student à $n_1 + n_2 - 2 = 28$ degrés de liberté.

Pour un test bilatéral au niveau $\alpha = 0.05$:

$$t_{0.975,\,28} \approx 2.048.$$

On rejette H_0 si $|t_{\text{obs}}| > 2.048$.

Or $|t_{\text{obs}}| = 2.725 > 2.048$, donc on **rejette** H_0.

La p-valeur (approchée) Avec $t = 2.725$ et 28 degrés de liberté :

$$p\text{-value} \approx 2(1 - F_{t_{28}}(2.725)) \approx 2(1 - 0.994) = 0.012.$$

Donc $p = 0.012 < 0.05$, ce qui confirme le rejet de H_0.

5. Intervalle de confiance (95%) pour $\mu_1 - \mu_2$

$$(\bar{x}_1 - \bar{x}_2) \pm t_{0.975,\,28}\, s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

$$4.8 \pm 2.048 \times 4.76 \times 0.370 = 4.8 \pm 3.61.$$

$$\text{IC}_{95\%}(\mu_1 - \mu_2) = [1.19,\ 8.41].$$

L'intervalle de confiance ne contient pas 0, et la statistique $t_{\text{obs}} = 2.725$ conduit au rejet de H_0. **Au niveau de 5%, on conclut qu'il existe une différence significative entre les moyennes des deux procédés.** Le procédé A donne une résistance moyenne plus élevée que le procédé B.

4.4 Tests non paramétriques

4.4.1 Test d'indépendance de Khi deux

Définition 4.6

Le test d'indépendance est utilisé pour tester l'hypothèse nulle d'absence de relation entre deux variables qualitatives. On peut également dire que ce test vérifie l'hypothèse d'indépendance de ces variables.

L'hypothèse nulle est l'hypothèse de l'absence de relation entre deux variables qualitatives. C'est l'hypothèse d'indépendance. On suppose que la valeur d'une des deux variables ne nous donne aucune information sur la valeur possible de l'autre variable. Lorsqu'il n'existe aucune relation entre deux variables on dit que les variables sont indépendantes l'une de l'autre. Il ne faut pas confondre cette expression avec l'appellation variable indépendante . L'hypothèse alternative est donc qu'il existe une relation entre les variables ou que les deux variables sont dépendantes.
Lorsque l'on a voulu tester l'hypothèse nulle de l'égalité des moyennes de deux échantillons indépendants, nous avons calculé la statistique Z (ou T si les variances sont inconnues). Puis, à l'aide de la distribution de la statistique du test, nous avons déterminé dans quelle mesure la valeur Z obtenue était inhabituelle si l'hypothèse nulle était vraie. Dans le cas de tableau de contingence où l'on travaille avec des effectifs (ou occurrences), nous allons calculer la statistique Khi-deux et comparer sa valeur à l'aide de la distribution Khi-deux dans le but de déterminer dans quelle mesure cette valeur est inhabituelle si l'hypothèse nulle est vraie.
La procédure statistique que nous allons employer pour tester l'hypothèse nulle : On compare les effectifs observés (celles déjà inscrites dans le tableau de données de contingence) avec les effectifs attendus ou théoriques. L'effectif attendu est simplement l'effectif que l'on devrait trouver dans une cellule si l'hypothèse nulle était vraie. Les étapes du test seront présentées en traitant l'exemple illustratif suivant:
Au niveau national, un examen est ouvert à des étudiants de spécialités différentes (Pharmacie et Médecine). Les responsables de l'examen désirent savoir si la formation initiale d'un étudiant influence sa réussite à cet examen. Les résultats sont enregistrés dans le tableau suivant :

	Pharmacie	Médecine	Total($n_{i\bullet}$)
Réussite	268	195	463
Echec	232	240	472
Total($n_{\bullet j}$)	500	435	935

On suit alors les étapes suivantes pour déterminer la statistique du test.

1. Calculer les effectifs théoriques, ces effectifs sont notés n_{ijtheo}, ils sont donnés par la formule :

$$n_{ijtheo} = \frac{n_{i\bullet} \times n_{\bullet j}}{N}$$

Dans note cas par exemple

$$n_{11theo} = \frac{n_{1\bullet} \times n_{\bullet 1}}{N} = \frac{463 \times 500}{935} = 247.6$$

Pour les étudiants en pharmacie ayant réussi l'examen national, l'effectif attendu est 247.6.
Pour les étudiants en pharmacie ayant échoué à l'examen national, l'effectif attendu est :

$$n_{21theo} = \frac{n_{2\bullet} \times n_{\bullet 1}}{N} = \frac{472 \times 500}{935} = 252.4$$

On aura alors le tableau suivant :

	Pharmacie	Médecine	Total
Réussite	268 (**247.6**)	195(**215.4**)	463
Echec	232(**252.4**)	240(**219.6**)	472
Total	500	435	935

Les chiffres notés en gras sont les effectifs théoriques, les autres sont appelés les effectifs observés.

2. Déterminer la statistique du test.
La statistique du test d'indépendance Khi deux est

$$\chi^2_{obs} = \sum_{i=1}^{k}\sum_{j=1}^{p} \frac{(n_{ij} - n_{ijtheo})^2}{n_{ijtheo}}$$

Dans notre cas

$$\chi^2_{obs} = \frac{(268 - 247.6)^2}{247.6} + \frac{(195 - 215.4)^2}{215.4} + ... = 7.16$$

3. La décision :
Le degré de liberté de la distribution de khi-deux ne dépend pas du nombre d'individus statistiques, mais plutôt du nombre de lignes et de colonnes du tableau de contingence.
Degré de liberté = (nombre de rangées-1) × (nombre de colonnes -1).
Dans notre cas : notons χ^2_{theo} la valeur de khi deux lue sur la table statistique, pour un risque $\alpha = 0.05$, sur la table Khi deux à 1 ddl, on trouve $\chi^2_{theo} = 3.84$.

$$\chi^2_{obs} = 7.16 > \chi^2_{theo} = 3.84$$

On rejette l'hypothèse nulle. Cela signifie que la réussite des étudiants à cet examen dépend de leur spécialité au risque 5%.

4.4.2 Test d'adéquation de Khi deux

Soient $X_1, ..., X_n$ un échantillon aléatoire de taille n d'une variable aléatoire mère X, soit F une fonction de répartition donnée. On souhaite tester les hypothèses suivantes :

$$\begin{cases} H_0 : X \text{ a pour fonction de répartition } F \\ \\ H_1 : X \text{ n'a pas pour fonction de répartition } F. \end{cases}$$

Pour cela, on répartit les n observations en k classes $[a_{i-1}, a_i[$, d'effectifs observés n_i, $n = \sum_{i=1}^{n} n_i$, $(1 \leq k \leq n)$ et on calcule, $p_i = F(a_i) - F(a_{i-1})$. On calcule ensuite les effectifs observés np_i.
On peut obtenir ainsi la valeur observée de la variable aléatoire D qui suit la loi de khi-deux à $(k-1)$ ddl.

$$D_{obs} = \sum_{i=1}^{n} \frac{(n_i - np_i)^2}{np_i}$$

La région critique(de rejet) du test au seuil α est : $D \geq D_{seuil}$ où la valeur de D_{seuil} étant lue dans la table de Khi-deux avec $\alpha = P(D \geq D_{seuil}/H_0)$ et D de loi approchée de khi-deux à $(k-1)$ ddl sous H_0.
L'effectif np_i de la classe $[a_{i-1}, a_i[$ doit être supérieur ou égal à 5; sinon on regroupe deux (ou plus) classes consécutives.

Conclusion

En conclusion, ce travail a permis de mettre en évidence les notions fondamentales de la statistique appliquée, en abordant successivement les lois de probabilité usuelles, l'estimation ponctuelle et par intervalles de confiance, ainsi que les tests d'hypothèses statistiques.

Il ressort de cette étude que la statistique ne se limite pas à un ensemble de formules ou de calculs, mais constitue avant tout un outil scientifique rigoureux permettant d'analyser les phénomènes, d'interpréter les données et de tirer des conclusions objectives et fiables.

Nous espérons que cette modeste contribution aura facilité la compréhension des principes essentiels de la statistique et montré le lien étroit entre la théorie et la pratique.

En définitive, la connaissance statistique demeure une composante indispensable de toute démarche scientifique, car elle fait des nombres une véritable langue de la réflexion analytique et du raisonnement logique.

Conclusion

In conclusion, this work has highlighted the fundamental concepts of applied statistics, covering successively the most common probability laws, point and interval estimation, as well as statistical hypothesis testing.

This study shows that statistics is not merely a collection of formulas and calculations, but above all a rigorous scientific tool that allows us to analyze phenomena, interpret data, and draw objective and reliable conclusions.

We hope that this modest contribution has helped to simplify and clarify the essential principles of statistics, while emphasizing the close link between theory and practice.

Ultimately, statistical knowledge remains an indispensable component of any scientific approach, as it transforms numbers into a true language of analytical thinking and logical reasoning.

Bibliography

[1] Jean-Pierre Lecoutre et Marie-Hélène Larmarange, *Statistique - Tome 1 et 2*, Éditions Dunod, Paris, 2020.

[2] Gérard Drouilhet et Jean-Claude Massé, *Probabilités et Statistique*, Éditions De Boeck Supérieur, Bruxelles, 2018.

[3] Michel Bierlaire, *Statistique Inférentielle*, Presses Polytechniques et Universitaires Romandes, Lausanne, 2015.

[4] Philippe Flajolet, *Introduction à la Statistique Appliquée*, Université de Paris-Sud, 2010.

[5] Roger L. Berger et George Casella, *Méthodes statistiques appliquées*, Duxbury Press, New York, 2002.