الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التعليم العالي والبحث العلمي

UNIVERSITE BADJI MOKHTAR - ANNABA
BADJI MOKHTAR – ANNABA UNIVERSITY

جامعة باجي مختار – عنابـــة

Faculté : Des Sciences

Département : De Mathématiques.

Domaine : Maths-Informatique

Filière : Mathématiques

Spécialité : Probabilités-Statistiques

Mémoire

Présenté en vue de l'obtention du Diplôme de Master

Thème:

La statistique avec R

Présenté par : – *Bader Lyna Ikram*

Jury de Soutenance :

Pr Benmostefa F.Z	Pr	U. BADJI Mokhtar Annaba	Président
Pr Chadli A	Pr	U. BADJI Mokhtar Annaba	Examinateur
Dr Chouia S	MCA	U. BADJI Mokhtar Annaba	Encadrant

Année Universitaire : 2022/2023

Table des matières

Remerciements

Louange à Dieu, le tout-puissant est miséricordieux qui m'a prodigué le courage et la force à fin de mener à terme mon travail.

Je remercie particulièrement, ma directrice de mémoire Dr. Chouia S. de m'avoir honoré avec son encadrement, et pour le temps qu'elle a consacré à m'apporter les outils méthodologiques indispensables à la conduite de cette recherche.

Mes sincères remerciements vont aux membres de jury pour l'intérêt qu'ils ont porté à mon travail en acceptant de l'examiner et de l'enrichir par leurs propositions et remarques.

Je remercie mes très chers parents, qui ont toujours été là pour moi, mon père qui m'a fait aimer et choisir les mathématiques, et ma mère dont le soutien, l'optimisme et la confiance qu'elle me témoigne m'ont portée tout au long des mes études .Je remercie ma sœur (Khaoula) et mes frères (Taki Eddine et Riad), pour leurs encouragements.

Je remercie ma deuxième famille mes proches amis (Rayene et Rania sont les meilleures choses qui me soient arrivées à l'université et Yousra mon amie pour la vie, qui a passé avec moi tous les moments importants de ma vie) qui ont toujours été là pour moi. Leur soutien inconditionnel et leurs encouragements ont été d'une grande aide.

À tous ces intervenants, je présente mes remerciements, mon respect et ma gratitude.

La statistique avec R

Résumé

Les logiciels fournissent des outils puissants pour l'analyse mathématique, la visualisation et la modélisation. Ils sont largement utilisés dans les universités, l'industrie pour une gamme d'applications dans les sciences, l'ingénierie, les finances et d'autres domaines.

Le langage de programmation R est devenu de plus en plus populaire ces dernières années en raison de sa polyvalence, de sa vaste gamme d'outils statistiques et de sa communauté d'utilisateurs actifs.

Ce travail traite les différents aspects dela statistique descriptive et de l'inférence statistique avec le logiciel R d'une façon simple. Il est utile aussi à toute personne souhaitant connaitre et surtout utiliser les principales méthodes statistiques avec R.

Introduction générale

1. Introduction

Le logiciel de programmation est un outil essentiel pour l'application des mathématiques dans divers domaines. Voici quelques raisons pour lesquelles le logiciel de programmation est important dans les applications mathématiques:

Calcul efficace : De nombreux problèmes mathématiques impliquent des calculs complexes qui prennent du temps et sont susceptibles d'erreurs lorsqu'ils sont effectués à la main. Un logiciel de programmation peut effectuer ces calculs beaucoup plus rapidement et avec plus de précision, ce qui permet aux chercheurs de se concentrer sur les aspects créatifs du processus de résolution de problèmes.

Visualisation : Un logiciel de programmation peut être utilisé pour créer des représentations visuelles de données et de concepts mathématiques. Cela peut aider les chercheurs à mieux comprendre les tendances et les relations dans leurs données et à communiquer leurs résultats aux autres plus efficacement.

Reproductibilité : Le logiciel de programmation permet aux chercheurs de documenter leur travail d'une manière facile à reproduire. C'est important pour la recherche scientifique, car elle permet à d'autres chercheurs de vérifier les résultats et de s'en inspirer.

Flexibilité : Le logiciel de programmation est très flexible et peut être adapté à un large éventail d'applications. Cela signifie que les chercheurs peuvent utiliser le même logiciel pour résoudre différents problèmes, ou modifier le logiciel en fonction de leurs besoins spécifiques.

Accessibilité : Les logiciels de programmation sont souvent de source ouverte et disponibles gratuitement, ce qui les rend accessibles aux chercheurs disposant de ressources

limitées. Cela contribue à démocratiser l'accès aux outils mathématiques et permet aux chercheurs de différents horizons de contribuer au domaine.

Collaboration : Les logiciels de programmation facilitent la collaboration des chercheurs sur des projets, même s'ils sont situés dans différentes parties du monde. Cela peut aider à accélérer le rythme de la recherche et mener à de nouvelles découvertes.

Globalement, les logiciels de programmation jouent un rôle important dans l'application des mathématiques dans de nombreux domaines différents. Il permet aux chercheurs de résoudre des problèmes complexes plus efficacement, de visualiser leurs données plus efficacement et de collaborer plus facilement avec d'autres. Voici quelques logiciels couramment utilisés en mathématiques:

MATLAB : MATLAB est un environnement de calcul numérique et un langage de programmation utilisés pour l'analyse des données, la visualisation et les mathématiques computationnelles. Il est largement utilisé dans l'ingénierie, la science et les finances.

Mathematica : Mathematica est un logiciel de calcul utilisé pour les mathématiques symboliques, l'analyse numérique et la visualisation. Il est largement utilisé dans les mathématiques, la physique et l'ingénierie.

R : R est un langage de programmation libre et un environnement logiciel utilisé pour le calcul statistique et les graphiques. Il est largement utilisé dans la science des données, les statistiques et l'apprentissage automatique.

Maple : Maple est un logiciel de calcul utilisé pour les mathématiques symboliques, l'analyse numérique et la visualisation. Il est largement utilisé dans les mathématiques, la physique et l'ingénierie.

Python : Python est un langage de programmation à usage général largement utilisé dans le calcul scientifique, l'analyse de données et la visualisation. Il est populaire dans des domaines tels que l'apprentissage automatique, la science des données et la bioinformatique.

SAS : SAS est une suite logicielle utilisée pour l'analyse de données, l'intelligence d'affaires et la modélisation prédictive. Il est couramment utilisé dans la finance, la santé et le gouvernement.

Octave : Octave est un logiciel de calcul numérique libre et open-source utilisé pour l'analyse numérique, l'algèbre linéaire et le traitement des signaux. Il est similaire au MATLAB et est largement utilisé en ingénierie et en science.

Dans l'ensemble, ces logiciels fournissent des outils puissants pour l'analyse mathématique, la visualisation et la modélisation. Ils sont largement utilisés dans les universités, l'industrie et le gouvernement pour une gamme d'applications dans les sciences, l'ingénierie, les finances et d'autres domaines.

R est un langage de programmation libre et open source qui est largement utilisé pour l'analyse statistique, la visualisation de données et l'exploration de données. Il est devenu de plus en plus populaire ces dernières années en raison de sa polyvalence, de sa vaste gamme d'outils statistiques et de sa communauté d'utilisateurs actifs.

Dans ce travail, nous nous intéressons d'appliquées toutes les commandes et paquets utilisés aux statistiques descriptives et à l'inférence statistique avec le logiciel R, pour simplifier l'utilisation et montrer la puissance et la flexibilité pour l'analyse et la visualisation des données.

Le manuscrit est composé de trois parties.

On débute par la partie théorique de la statistique descriptive et l'inférence statistique. Dans la deuxième partie, nous nous sommes intéressé à utiliser toutes les commandes d'analyse statistique, visualisation des données, des paquets utilisés pour l'estimation de maximum de vraisemblance avec des exemples des données réelles.

Enfin, une dernière partie est consacrée à l'analyse de fiabilité. Nous nous sommes intéressé à étudier la distribution q-exponentielle qui est un outil puissant pour modéliser des systèmes complexes et des données en intelligence artificielle et en apprentissage automatique. Sa capacité à capturer les distributions non gaussiennes le rend bien adapté à de nombreuses applications.

Le but de ce travail est de fournir pour différentes tailles d'échantillon, des tableaux de valeurs critiques des tests l'ajustement de Kolmogorov-Smirnov modifiées, Cramer-Von Mises pour la distribution q-exponentielle qui est utilisée pour décrire la fiabilité des systèmes réels. La puissance de ces statistiques est étudié en utilisant certaines alternatives

telles que l'exponentielle et l'exponentielle exponentiée. Tous les calculs sont effectués en utilisant logiciel R et la méthode de Monte-Carlo.

2. La famille de distribution de type q

La famille de distribution de type q est une famille de distributions de probabilité caractérisée par un paramètre q qui est un nombre réel. Ce paramètre contrôle le degré de non-extensivité de la distribution. Dans la limite de q tend vers 1, la distribution de type q se réduit à la distribution de Boltzmann-Gibbs, qui est étendue et obéit au principe d'additivité. Cependant, pour q différent de 1, la distribution devient non-extensive. Une propriété importante des distributions de type q est qu'elles présentent des queues de loi de puissance. Cela signifie que la probabilité d'observer une très grande valeur d'une variable aléatoire suit une distribution de loi de puissance. Cette propriété rend les distributions de type q particulièrement utiles pour les systèmes de modélisation présentant des événements extrêmes ou des phénomènes rares.

La famille de distribution de type Q a trouvé des applications dans une grande variété de domaines, y compris la finance, la physique, la biologie et les sciences sociales. En finance, des distributions de type q ont été utilisées pour modéliser la distribution des rendements financiers, qui sont connus pour présenter des queues de graisse et d'autres caractéristiques non gaussiennes. En physique, les distributions de type q ont été utilisées pour modéliser des systèmes qui présentent des interactions à longue portée, comme les plasmas et les fluides. En biologie, des distributions de type q ont été utilisées pour modéliser la distribution des niveaux d'expression des gènes, qui sont connus pour présenter des caractéristiques non gaussiennes. Dans les sciences sociales, les distributions de type q ont été utilisées pour modéliser les distributions de revenu et de richesse, qui présentent également des queues de puissance-loi. En turbulence, les distributions du type Q ont été utilisées pour modéliser la distribution d'énergie des particules dans les systèmes turbulents, qui présentent des interactions à longue portée. La distribution gaussienne a été trouvée pour fournir un bon ajustement à la distribution d'énergie des particules dans l'écoulement turbulent. En réseaux complexes, les distributions du type Q ont été utilisées pour modéliser la distribution des degrés de nœud dans les réseaux complexes,

tels que les réseaux sociaux et le World Wide Web. La distribution q-exponentielle a été trouvée pour fournir un bon ajustement à la distribution des degrés de nœud dans ces réseaux. En systèmes climatiques, les distributions du type Q ont été utilisées pour modéliser la répartition des phénomènes météorologiques extrêmes, comme les vagues de chaleur et les fortes précipitations.

Il existe de nombreux types de distributions de type q, y compris la distribution q-Gaussienne est peut-être la distribution de type q le plus connu et le plus étudié. Il s'agit d'une généralisation de la distribution gaussienne (normale), et il a été utilisé pour modéliser un large éventail de phénomènes, y compris la distribution d'énergie des particules dans un système avec des interactions à longue portée. Lorsque q=1, la distribution de type q se réduit à la distribution gaussienne. Cependant, pour q différent de 1, la distribution de type q a une forme différente de la distribution gaussienne, et elle a des queues plus lourdes ou plus légères que celles de la distribution gaussienne, selon la valeur de q. La distribution q-Weibull c'est une généralisation q-type de la distribution Weibull, qui est couramment utilisée pour modéliser le temps de défaillance des systèmes mécaniques. Elle a été utilisée pour modéliser le temps de défaillance des systèmes avec des interactions à longue portée, comme les réseaux de communication. Il a une queue plus lourde que la distribution standard de Weibull pour des valeurs de q inférieures à 1. La distribution q-Laplace est une généralisation q-type de la distribution Laplace, qui est couramment utilisée pour modéliser la différence entre deux variables aléatoires exponentielles indépendantes. La distribution q-Laplace a été utilisée pour modéliser une variété de phénomènes, y compris la distribution du revenu et de la richesse. Il a une queue plus lourde que la distribution standard de Laplace pour des valeurs de q inférieures à 1. La distribution q-Pareto est une généralisation q-type de la distribution Pareto, qui est couramment utilisée pour modéliser la distribution du revenu et de la richesse. Elle a été utilisée pour modéliser la distribution de la taille des entreprises et la distribution de la taille des villes. Sa queue est plus lourde que la distribution standard de Pareto pour des valeurs de q inférieures à 1. La distribution q-exponentielle est l'une des nombreuses distributions de type q qui ont été développées pour modéliser une variété de phénomènes,

y compris la distribution des temps inter-événement dans des systèmes complexes avec un comportement non gaussien, la distribution de la richesse et du revenu. Chacune de ces distributions à ses propres propriétés et applications spécifiques, mais elles partagent toutes la caractéristique commune d'être caractérisées par un paramètre q qui est un nombre réel, et qui régit leur forme et leur comportement.

Partie I

Statistique Paramétrique

CHAPITRE 1

Statistique Descriptive

Les statistiques descriptives visent étudier les caractéristiques d'un ensemble d'observations comme les mesures obtenues lors d'une expérience. L'expérience est l'étape préliminaire toute étude statistique.

Définissons maintenant les notions de base:

Qu'est ce que la statistique: Les statistiques, dans le sens populaire du terme, traitent des populations. Leurs objectifs consistent à caractériser une population à partir d'une image plus ou moins constituée à l'aide d'un échantillon issu de cette population. On trouve des applications de la statistique dans tous les domaines : industrie, environnement, médecine, maintenance, marketing, sport, ...

Qu'est ce qu'un test statistique: Un test, qu'il soit statistique ou pas, consiste à vérifier une information hypothétique.

Epreuve statistique: L'épreuve statistique est une expérience que l'on provoque.

Population: On appelle population l'ensemble sur lequel porte notre étude statistique. Cet ensemble est noté Ω.

Unité statistique (individu): On appelle individu ou unité statistique tout élément de la population Ω, il est noté $\omega(\omega$ dans Ω).

Variable statistique (Caractère): Une variable aléatoire est une fonction valeurs numériques (réelles) dénie sur un espace échantillon.

Modalités: Les modalités d'une variable statistique sont les différentes valeurs que peut prendre celle-ci.

Types de Variables: Nous distinguons deux catégories de caractères : les caractères qualitatifs et les caractères quantitatifs.

Variable qualitative: La variable est dite qualitative quand les modalités sont des catégories.

Variable quantitative: Une variable est dite quantitative si toute ses valeurs possibles sont numériques.

On distingue habituellement les variables aléatoires discrètes et les variables aléatoires continues.

Variable quantitative discrète: Si une variable aléatoire X prend un nombre de valeurs ni ou dénombrable (son ensemble de définition est inclus dans $\mathbb{N}$), on parle de variable discrète.

Définition 1.0.1 *Une variable aléatoire est dite de type discret si le nombre de valeurs différentes quelle peut prendre est fini ou infini dénombrable.*

On s'intéresse définir l'ensemble des valeurs possibles et leurs probabilités associées.

Variable quantitative continue: Une variable aléatoire est dite continue si elle peut prendre toutes les valeurs d'un intervalle. En particulier, dans le cas où la variable aléatoire peut prendre toute valeur réelle (son ensemble de définition contient un intervalle de $\mathbb{R}$), on parle de variable aléatoire réelle.

Définition 1.0.2 *Une variable aléatoire qui peut prendre un nombre infini non dénombrable de valeurs est dite variable aléatoire de type continu.*

Dans ce cas, il ne s'agira plus de calculer une probabilité d'apparition d'une valeur donnée mais d'un intervalle.

Série statistique: On appelle série statistique la suite des valeurs prises par une variable X sur les unités d'observation. Le nombre d'unités d'observation est noté n.

Les valeurs de la variable X sont notées

$$x_1, x_2, ..., x_n$$

1.1 Tableau statistique

1.1.1 Effectif partiel (fréquence absolue)

Définition 1.1.1 *Le nombre d'individus qui ont le même x_i, a s'appelle effectif partiel n_i de x_i.*

1.1.2 Effectif cumulé

Définition 1.1.2 *L'effectif cumulé N_i d'une valeur est la somme de l'effectif de cette valeur et de tous les effectifs des valeurs qui précèdent.*

1.1.3 Fréquence partielle (fréquence relative)

Définition 1.1.3 *Pour chaque valeur x_i, on pose par définition*

$$f_i = \frac{n_i}{N}$$

f_i S'appelle la fréquence partielle de x_i.

1.1.4 Fréquence cumulée

Définition 1.1.4 *Pour chaque valeur x_i, on pose par définition*

$$\Delta f_i = f_1 + f_2 + ... + f_i$$

La quantité Δf_i s'appelle la fréquence cumulée de x_i.

1.1.5 Représentation graphique des séries statistiques

Variable qualitative

Le tableau statistique d'une variable qualitative peut être représenté par deux types de graphique. Les effectifs sont représentés par un diagramme en barres et les fréquences par un diagramme en secteurs.

- **Diagramme circulaire (diagramme en secteurs):** Les diagrammes circulaires, consistent partager un disque ou un demi-disque, en tranches, ou secteurs, correspondant aux modalités observes et dont la surface est proportionnelle l'effectif, ou la fréquence, de la modalité.

- **Diagramme en barres:** à chaque marque correspond un bâton. Les hauteurs des bâtons sont proportionnelles aux effectifs.

Variable quantitative discrète

- **Diagramme en bâtonnets :** Quand la variable est discrète, les effectifs sont représentés par des bâtonnets.

Variable quantitative continue

- **Histogramme:** L'histogramme consiste représenter les effectifs (resp. les fréquences) des classes par des rectangles contigus dont la surface représente l'effectif (resp. la fréquence). Pour un histogramme des effectifs, la hauteur du rectangle correspondant la classe j.

1.1.6 Paramètres de position

Les indicateurs statistiques de tendance centrale (dits aussi de position) considérés fréquemment sont la moyenne, la médiane et le mode.

La moyenne

On appelle moyenne de X, la quantité

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} n_i x_i \tag{1.1}$$

La médiane

On appelle médiane la valeur Me de la V.S X:

- 1^{ier} cas: Si $N = P \times 2$, on a :

$$Me = \frac{X_P + X_{P+1}}{2} \tag{1.2}$$

- $2^{éme}$ cas: Si $N = P \times 2 + 1$, on a :

$$Me = X_P \tag{1.3}$$

Le mode

Le mode d'une V.S est la valeur qui a le plus grand effectif partiel (ou la plus grande fréquence partielle) et il est dénoté par Mo.

1.1.7 Paramètres de dispersion

Les indicateurs statistiques de dispersion usuels sont l'étendue, la variance et l'écart-type.

L'étendue

La différence entre la plus grande valeur et la plus petite valeur du caractère, donnée par la quantité

$$e = x_{\max} + x_{\min} \tag{1.4}$$

S'appelle l'étendue de la V.S X.

La variance

On appelle variance d'une série statistique X,

$$Var(X) = \frac{1}{N}\sum_{i=1}^{n} n_i(x_i - \overline{x})^2 = \frac{1}{N}\sum_{i=1}^{n} n_i x_i^2 - \overline{x}^2 \tag{1.5}$$

L'écart-type

La quantité

$$\delta_X = \sqrt{Var(X)} \tag{1.6}$$

S'appelle l'écart-type de la V.S X.

CHAPITRE 2

Estimation des paramètres

2.1 Généralités

Considérons un échantillon de taille n extrait d'une population de taille N et pour laquelle on s'intéresse un caractère X mesuré pour chaque individu de la population.

Le caractère X est considéré comme une variable aléatoire et l'échantillon de valeurs est constitué de n réalisations de cette variable.

On représente cette situation au moyen d'un modèle statistique qui comporte une famille de lois de probabilités parmi lesquelles se trouve la loi suivie par la variable X.

Ces lois de probabilité dépendent en général d'un ou de plusieurs paramètres notés θ. Dans ce cas, on dit qu'on a un modèle statistique paramétrique.

Un des problèmes les plus courants en statistique consiste à trouver la valeur du ou des paramètres pour la population. Mais comme on ne peut pas en général avoir l'information nécessaire, on doit se contenter des valeurs fournies par l'échantillon.

On considère que chaque $X_i$est une variable aléatoire et on suppose qu'elles sont indépendantes entre elles. D'autre part, elles sont identiquement distribuées.

2.2 Estimateurs

Définition 2.2.1 *On appelle estimateur d'un paramètre θ d'une population, toute fonction*

$$T = f(X_1, X_2, ..., X_n) \tag{1.7}$$

et sa réalisation sera noté

$$t = f(x_1, x_2, ..., x_n)$$

Pour un même paramètre, il peut y avoir plusieurs estimateurs possibles; par exemple, Le paramètre λ d'une loi de Poisson admet comme estimateurs possibles la moyenne empirique et la variance empirique. Pour pouvoir choisir, il faut définir les qualités qui font qu'un estimateur sera meilleur. On va voir en particulier les quantités empiriques les plus couramment utilisés :

1. La moyenne empirique.
2. La variance empirique.
3. La fréquence empirique.

La valeur empirique d'un paramètre est également une variable aléatoire car, non seulement, il est calculé à partir d'une variable aléatoire mais aussi d'un échantillon qui lui-même est aléatoirement choisi.

2.2.1 Qualité d'un estimateur

Définition 2.2.2 *On appelle **biais** d'un estimateur, la quantité*

$$B(T) = E(T) - \theta$$

Qui représente l'erreur systématique.

Définition 2.2.3 *Un estimateur T de θ est dit **sans biais** si*

$$E(T) = \theta$$

Définition 2.2.4 *Un estimateur* ***sans biais*** *est dit* ***convergent*** *si*

$$V(T) \underset{n\to\infty}{\to} 0$$

Définition 2.2.5 *Soient T_1 et T_2 deux* ***estimateurs sans biais*** *de θ. T_1 est dit* ***plus efficace*** *que T_2 si*

$$V(T_1) \leq V(T_2)$$

2.2.2 Estimation ponctuelle

Estimer un paramètre, par exemple: une moyenne, une variance, une proportion, etc..., c'est chercher une valeur approché en se basant sur les résultats d'un échantillon. Lorsqu'un paramètre est estimé par un seul nombre déduit des résultats de l'échantillon, ce nombre est appelé une estimation ponctuelle du paramètre.

Estimation de la moyenne

Soit X une variable aléatoire dont on veut estimer la moyenne (ou espérance) $\mu = E(X)$ partir d'un échantillon $(X_1, X_2, ..., X_n)$ de X (on ne suppose rien sur la loi de X).

Définition 2.2.6 *On appelle moyenne empirique de l'échantillon $(X_1, X_2, ..., X_n)$ de X, la statistique*

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

Sa réalisation est

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Qui est la moyenne de l'échantillon aussi appelée moyenne observée.

Propriétés:

$$E(\overline{X}) = \mu$$

$$V(\overline{X}) = \frac{1}{n}\sigma^2$$

On a

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

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

Donc $\overline{X}$ est un estimateur sans biais de $E(X) = \mu$ et de plus il est convergent $V(\overline{X}) = \frac{V(X)}{n} \underset{n\to\infty}{\to} 0$, et $\forall T$, un autre estimateur de , $V(T) > V(\overline{T})$.

Théorème central limite 1 Lorsque la variance σ^2 de la population est connue et que l'échantillon prélevé est grand ($n \geq 30$), alors la moyenne échantillonnée vérifie:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \rightsquigarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \tag{1.8}$$

Alors

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightsquigarrow N(0,1) \tag{1.9}$$

Remarque:

- Le théorème précédent est vrai aussi lorsque la variance est connue, l'échantillon est petit et que la variable aléatoire X suit une loi normale $N(\mu, \sigma^2)$.
- Lorsque la variance σ^2 de la population est inconnue et que l'échantillon prélevé est grand ($n \geq 30$), alors

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \rightsquigarrow N\left(\mu, \frac{s}{\sqrt{n}}\right) \text{ c'est à dire } Z = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \rightsquigarrow N(0,1)$$

Théorème central limite 2 Si la variance de la population est inconnue, si la variable X suit une distribution normale $N(\mu, \sigma^2)$, et si la taille de l'échantillon est petite ($n < 30$), alors

$$T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \rightsquigarrow t_{n-1} \text{Loi de Student } n-1 \text{ degrés de liberté ddl.} \tag{1.10}$$

Estimation de la variance

Soit X une variable aléatoire qui suit une loi normale $N(\mu, \sigma)$. On veut estimer la variance σ^2 de X.

Définition 2.2.7 *On appelle variance empirique de l'échantillon* $(X_1, X_2, ..., X_n)$ *de* X*, la statistique*

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2\right) - \overline{X}^2 \tag{1.11}$$

qui est la variance de l'échantillon, aussi appelée variance observée.

La variance empirique correspond donc à la moyenne des écarts à la moyenne empirique.

Propriétés

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

On a

$$\begin{aligned}
E\left(S^2\right) &= E\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right) = E\left(\frac{1}{n}\left(\sum_{i=1}^{n} X_i^2\right) - \overline{X}^2\right) \\
&= \frac{1}{n}E\left(\sum_{i=1}^{n} X_i^2\right) - E\left(\overline{X}^2\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i^2\right) - E\left(\overline{X}^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[V\left(X_i\right) + \left(E\left(X_i\right)\right)^2\right] - \frac{1}{n}\sigma^2 - \mu^2 \\
&= V\left(X_i\right) + \left(E\left(X_i\right)\right)^2 - \frac{1}{n}\sigma^2 - \mu^2 = \sigma^2 + \mu^2 - \frac{1}{n}\sigma^2 - \mu^2 \\
&= \left(1 - \frac{1}{n}\right)\sigma^2 = \frac{n-1}{n}\sigma^2
\end{aligned}$$

On voit donc qu'à un coefficient prés, l'espérance de la variance empirique est différente de la variance de la population. Cet estimateur est donc biaisé. D'où la nécessité de

trouver un estimateur non biaisé. C'est là qu'intervient la notion de variance empirique modifiée.

Variance empirique modifiée Soit S^{*2} la variance empirique modifiée. Elle se calcule comme suit

$$S^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2\right) - \frac{1}{n-1}\overline{X}^2 \tag{1.12}$$

On peut aisément montrer que :

$$S^{*2} = \frac{n}{n-1}S^2$$

et que

$$E\left(S^{*2}\right) = \sigma^2$$

Variance de la variance empirique S^2 L'expression de la variance de S^2 se présente comme suit :

$$Var(S^2) = \frac{n-1}{n^3}\sigma^2 \tag{1.13}$$

Propriétés :

Si $(X_1, ..., X_n)$ est un échantillon de variables gaussiennes (loi normale), alors les variables $\sqrt{n}\left(\frac{\overline{X}-\mu}{\sigma}\right)$ et $(n-1)\frac{S^{*2}}{\sigma^2}$ sont indépendantes et suivent respectivement la loi normale $N(0,1)$ et la loi de khi-deux à $n-1$ degrés de liberté.

Comme on a $S^{*2} = \frac{n}{n-1}S^2$ la propriété est aussi vraie pour $n\frac{S^2}{\sigma^2}$ qui suit une loi de khi-deux à n degrés de liberté.

Estimation d'une proportion

Considérons une variable aléatoire qui suit une loi de Bernoulli, c'est-à-dire d'une variable aléatoire X qui ne peut prendre que deux valeurs 0 (Échec) ou 1 (succès). Dans cette expérience, il s'agit d'étudier la probabilité de succès. On écrit $X \rightsquigarrow B(f)$.

Dans une expérience de Bernoulli, la moyenne empirique est appelée fréquence empirique. Elle représente l'estimateur du paramétré f. On la note F.

Définition 2.2.8 *La variable aléatoire*

$$f = \frac{K_n}{n}, \text{ Où } K_n \text{représente le nombre de succès} \tag{1.14}$$

S'appelle fréquence empirique.

Loi de probabilité de F L'espérance d'une loi de Bernoulli $B(f)$ est égale à f et la variance à $f(1-f)$. On déduit donc des calculs sur la moyenne empirique que :

$$F \rightsquigarrow N\left(f, \sqrt{\frac{f(1-f)}{n}}\right) \tag{1.15}$$

en effet

$$E(F) = \frac{E(X_1) + ... + E(X_n)}{n} = f$$

et

$$V(F) = \frac{V(X_1) + ... + V(X_n)}{n} = \frac{nf(1-f)}{n^2} = \frac{f(1-f)}{n}$$

donc F est un estimateur sans biais de f.

2.2.3 Estimation par intervalle de confiance

Les estimations ponctuelles, bien qu'utiles, ne fournissent aucune information concernant la précision des estimations, c'est-à-dire quelles ne tiennent pas compte de l'erreur possible dans l'estimation due aux fluctuations d'échantillonnage. La théorie des intervalles de confiance (IC) consiste à construire, autour de l'estimation ponctuelle, un intervalle qui aura une grande probabilité $(1-\alpha)$ de contenir la vraie valeur du paramètre.

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

Définition 2.2.9 *On appelle* ***intervalle de confiance*** *pour θ de niveau $1-\alpha$ (ou de seuil α), un intervalle $[t_1, t_2]$ qui a la probabilité $1-\alpha$ de contenir la vraie valeur de θ*

$$P(t_1 < \theta < t_2) = 1 - \alpha$$

plus le niveau de confiance est élevé, plus la certitude est grande et que la méthode d'estimation produira une estimation contenant la vraie valeur de θ.

- Si on augmente le niveau de confiance $1-\alpha$, on augmente la longueur de l'intervalle.

Intervalle de confiance pour la moyenne

Cas où n, la taille de l'échantillon, est petite n < 30

On suppose que $X \rightsquigarrow N(\mu, \sigma)$

Première cas: Lorsque la taille de l'échantillon est petite ($n < 30$) et $X \rightsquigarrow N(\mu, \sigma)$ de variance **inconnue**:

On a:

$$\frac{\overline{X} - \mu}{S\sqrt{n-1}} \rightsquigarrow t_{n-1} \text{la loi de student à } n-1 \text{ ddl}$$

On cherche dans la table de la loi de Student,α étant fixé, la valeur $t_{n-1;\left(1-\frac{\alpha}{2}\right)}$ telle que:

$$P\left(-t_{n-1;\left(1-\frac{\alpha}{2}\right)} < \frac{\overline{X} - \mu}{S\sqrt{n-1}} < t_{n-1;\left(1-\frac{\alpha}{2}\right)}\right) = 1-\alpha$$

On a

$$P\left(\overline{X} - t_{n-1;\left(1-\frac{\alpha}{2}\right)} \cdot \frac{s^*}{\sqrt{n}} < \mu < \overline{X} + t_{n-1;\left(1-\frac{\alpha}{2}\right)} \cdot \frac{s^*}{\sqrt{n}}\right) = 1-\alpha$$

si $\overline{x}$ est une réalisation de $\overline{X}$ et s une réalisation de S, l'intervalle de confiance de μ de seuil α est

$$IC_\mu = \left[\overline{x} - t_{n-1;\left(1-\frac{\alpha}{2}\right)} \cdot \frac{s^*}{\sqrt{n}}; \overline{x} + t_{n-1;\left(1-\frac{\alpha}{2}\right)} \cdot \frac{s^*}{\sqrt{n}}\right]$$

Deuxième cas: Lorsque la taille de l'échantillon est petite ($n < 30$) et la variance de la population de X est **connue**:

On a:

$$\overline{X} \rightsquigarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \text{ ou bien } \frac{\overline{X} - \mu}{\sigma\sqrt{n}} \rightsquigarrow N(\mu, \sigma)$$

On se fixe le risque α et on cherche dans la table de la loi normale la valeur $u_{1-\frac{\alpha}{2}}$ (est le fractale d'ordre $1-\frac{\alpha}{2}$ de la loi normale centrée réduite) .telle que

$$P\left(-u_{1-\frac{\alpha}{2}} < \frac{\overline{X} - \mu}{\sigma\sqrt{n}} < u_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$P\left(\overline{X} - u_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + u_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Si $\overline{x}$ est une réalisation de $\overline{X}$, l'intervalle de confiance de μ de seuil α est

$$IC_\mu = \left[\overline{X} - u_{1-\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}}; \overline{X} + u_{1-\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}}\right] \tag{1.16}$$

Cas où n, la taille de l'échantillon, est grande n $\geq$ 30

Il n'est plus nécessaire de supposer que X est Gaussienne.

Premier cas: Lorsque la taille de l'échantillon est grande ($n \geq 30$) et $X \rightsquigarrow N(\mu, \sigma)$ de variance **connue**:

On a:

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightsquigarrow N(0,1)$$

Si $\overline{x}$ est une réalisation de $\overline{X}$ et si s une réalisation de S, l'intervalle de confiance de μ de seuil α est

$$IC_\mu = \left[\overline{x} - u_{1-\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}}; \overline{x} + u_{1-\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}}\right] \tag{1.17}$$

Deuxième cas: Lorsque la taille de l'échantillon est grande ($n \geq 30$) et la variance **inconnue**:

On a:

$$\frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \rightsquigarrow N(0,1)$$

On se fixe l'erreur α et on cherche dans la table de la loi normale la valeur $u_{1-\frac{\alpha}{2}}$ telle que

$$P\left(-u_{1-\frac{\alpha}{2}} < \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} < u_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$P\left(\overline{X} - u_{1-\frac{\alpha}{2}}.\frac{s^*}{\sqrt{n}} < \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} < \overline{X} + u_{1-\frac{\alpha}{2}}.\frac{s^*}{\sqrt{n}}\right) = 1 - \alpha$$

si $\overline{x}$ est une réalisation de $\overline{X}$ et s une réalisation de S, l'intervalle de confiance de μ de seuil α est

$$IC_\mu = \left[\overline{x} - u_{1-\frac{\alpha}{2}}.\frac{s^*}{\sqrt{n}}; \overline{x} + u_{1-\frac{\alpha}{2}}.\frac{s^*}{\sqrt{n}}\right] \tag{1.18}$$

Intervalle de confiance pour la variance

On suppose que X est gaussienne.

D'après la section 2.2.2 , on a

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

Remarque

Si la moyenne est μ inconnu, donc

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

est une somme de n variable aléatoire indépendantes qui suivent la loi normale$N(0,1)$ et donc

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

On cherche les deux nombres a et b tels que

$$P\left(\frac{(n-1)\,S^2}{\sigma^2} \geq a \right) = 1 - \frac{\alpha}{2}; P\left(\frac{(n-1)\,S^2}{\sigma^2} \geq b \right) = \frac{\alpha}{2}$$

L'intervalle de confiance au seuil $1-\alpha$ pour la variance inconnue σ^2 de la population est de la forme:

$$IC_{\sigma^2} = \left[\frac{(n-1)\,S^2}{b}; \frac{(n-1)\,S^2}{a} \right] \tag{1.19}$$

Intervalle de confiance pour la proportion

On a

$$f = \frac{K}{n}$$

est le meilleur estimateur de F où F est la proportion de la population possédant le caractère considéré.

On cherche dans la table de $N(0,1)$ la valeur $u_{1-\frac{\alpha}{2}}$ telle que:

$$P\left(-u_{1-\frac{\alpha}{2}} < \frac{F-f}{\sqrt{\frac{f(1-f)}{n}}} < u_{1-\frac{\alpha}{2}} \right) = 1-\alpha$$

$$P\left(f - u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{f\,(1-f)}{n}} < F < f + u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{f\,(1-f)}{n}} \right) = 1-\alpha$$

L'intervalle de confiance au seuil $1-\alpha$ pour la proportion f de la population est de la forme:

$$IC_f = \left[f - u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{f\,(1-f)}{n}}; f + u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{f\,(1-f)}{n}} \right] \tag{1.20}$$

CHAPITRE 3

Tests d'hypothèses

3.1 Généralités

Un test statistique est une procédure de décision qui permet de choisir entre deux hypothèses (contraires) faites sur une population ou sur un ou plusieurs paramètres au vu d'un échantillon aléatoire. On formule une hypothèse de départ, appelée hypothèse nulle et souvent notée H_0 et il s'agit de décider si on rejette ou non cette hypothèse par opposition une contre hypothèse appelée hypothèse alternative et souvent notée H_1. Pour effectuer le test statistique, il faudra choisir un certain risque d'erreur qui est la probabilité de se tromper en prenant la décision retenue. Il existe deux types d'erreur :

- On appelle erreur de première espèce ou erreur de type I, notée α, la probabilité de rejeter H_0 alors qu'elle est vraie. Est aussi appelé niveau ou seuil de signification.
- On appelle erreur de deuxième espèce ou erreur de type II, notée β, la probabilité d'accepter H_0 alors qu'elle est fausse.
- On appelle puissance du test pour H_1 la probabilité de retenir H_1 alors qu'elle est vraie $(1 - \beta)$

Les étapes de construction d'un test statistique

1. Il s'agit d'abord de formuler les hypothèses H_0 et H_1. On choisit en général le risque de type I , α . (souvent donné dans l'énoncé).

2. Détermination de la variable de décision: on détermine la variable de décision Z (qui est une statistique) dont on connaît la loi si H_0 est vraie.

3. On calcule la région critique ou région de rejet W qui est l'ensemble des valeurs de Z qui conduiront à rejeter H_0. Ainsi, si α est fixé, W est déterminé par $\alpha = P$ ($Z \in W$ avec H_0 vraie). Le complémentaire de W est appelé "région d'acceptation". Les points de jonction entre les deux régions sont les points critiques.

4. On calcule la valeur de Z à partir de l'observation de l'échantillon.

5. Conclusion du test : acceptation ou rejet de H_0 selon que la valeur de Z est ou non dans la région d'acceptation.

En fonction de l'hypothèse testée, plusieurs types de tests peuvent être réalisés :

- **Les tests de conformité:** sont destinés à vérifier si un échantillon peut être considéré comme extrait d'une population donnée ou représentatif de cette population, vis-à-vis d'un paramètre comme la moyenne, la variance ou la fréquence observée. Ceci implique que la loi théorique du paramètre est connue au niveau de la population.

- **Les tests d'homogénéité:** sont destinés à comparer plusieurs populations l'aide d'un nombre équivalent d'échantillons. Dans ce cas la loi théorique du paramètre est inconnue au niveau des populations.

3.2 Tests de conformité

3.2.1 Construction générale

Soit X une variable aléatoire dont la loi dépend d'un paramètre inconnu .

- Tester l'hypothèse $H_0 : \theta = \theta_0$,θ_0 étant une valeur numérique; contre $H_1 : \theta \neq \theta_0$.

- Choix de la variable de décision Z qui est l'estimateur de θ ou une fonction simple de l'estimateur de θ.

- Calcul de la région critique :

$$\alpha = P(\text{décider } H_1 \text{ alors que } H_0 \text{ est vraie}) \iff \alpha = P(Z \in W \text{ alors que} \theta \neq \theta_0)$$

3.2.2 Comparaison d'une moyenne observée à une moyenne théorique

Soit X, une variable aléatoire observé sur une population, suivant une loi normale et un échantillon extrait de cette population.

Le but est de savoir si un échantillon de moyenne $\overline{x}$, estimateur de μ, appartient à une population de référence connue d'espérance μ_0 (H_0 vraie) et ne diffère de μ_0 que par des fluctuations d'échantillonnage ou bien appartient à une autre population inconnue d'espérance μ (H_1 vraie).

L'hypothèse testée est la suivante:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases} \tag{1.21}$$

Pour tester cette hypothèse, il existe deux statistiques : la variance σ^2 de la population de référence est connue ou cette variance est inconnue et il faut l'estimer.

Premier cas: la variance σ^2 de la population de référence est connue

On calcule la valeur

$$Z_{obs} = \frac{|\overline{x} - \mu_0|}{\sqrt{\frac{\sigma^2}{n}}} \tag{1.22}$$

On détermine Z_{tab} lue sur la table de la loi normale centré réduite pour un risque d'erreur α fixé, et on décide que:

- si $Z_{obs} > Z_{tab}$ l'hypothèse H_0 est rejetée au risque d'erreur α: l'échantillon appartient à une population d'espérance μ et n'est pas représentatif de la population de référence d'espérance μ_0.

- si $Z_{obs} \leq Z_{tab}$ l'hypothèse H_0 est acceptée : l'échantillon est représentatif de la population de référence d'espérance μ_0.

Deuxième cas: la variance σ_0^2 de la population de référence est inconnue

On calcule la valeur

$$T_{obs} = \frac{|\overline{x} - \mu_0|}{\sqrt{\frac{s^2}{n}}} \tag{1.23}$$

On détermine T_{tab} lue dans la table de Student pour un risque d'erreur α fixé et$(n-1)$ degrés de liberté, et on décide que:

- si $T_{obs} > T_{tab}$ l'hypothèse H_0 est rejetée au risque d'erreur α: l'échantillon appartient une population d'espérance μ et n'est pas représentatif de la population de référence d'espérance μ_0.
- si $T_{obs} \leq T_{tab}$ l'hypothèse H_0 est acceptée: l'échantillon est représentatif de la population de référence d'espérance μ_0.

Remarque Si $n < 30$, la variable aléatoire X étudiée doit impérativement suivre une loi normale $N(\mu, \sigma)$. Pour $n \geq 30$, la variable de student T converge vers une loi normale centrée réduite Z.

3.2.3 Comparaison d'une variance observée à une variance théorique

Soit un échantillon $(X_1, ..., X_n)$ issu d'une population de loi normale, de moyenne μ et de variance σ^2.

Le but est de savoir si un échantillon de moyenne s^2 , estimateur de σ^2, appartient à une population de référence connue de variance σ_0^2 (H_0 vraie) et ne diffère de σ_0^2 que par des fluctuations d'échantillonnage ou bien appartient à une autre population inconnue de variance σ^2 (H_1 vraie).

L'hypothèse testée est la suivante:

$$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2 \end{cases} \tag{1.24}$$

Pour tester cette hypothèse, il existe deux statistiques : la moyenne μ de la population de référence est connue ou cette variance est inconnue et il faut l'estimer.

Premier cas: la moyenne μ de la population de référence est connue

Lorsque la moyenne μ est connue, la statistique T^2 est la meilleure estimation de la variance

$$T^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

Sous l'hypothèse H_0, comme l'échantillon est gaussien, on a la statistique

$$y^2 = \frac{nT^2}{\sigma_0^2} = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma_0}\right)^2$$

Suit une loi du χ^2 à n degrés de libertés. (en tant que somme de carrés de $N(0,1)$) On cherche les deux nombres tabulés a et b tels que:

$$P\left(Y^2 \geq a\right) = \frac{\alpha}{2}; P\left(Y^2 \geq b\right) = 1 - \frac{\alpha}{2}$$

Pour un risque d'erreur fixé α, et on décide que:

- si $y^2 \in]a; b[$ l'hypothèse H_0 est acceptée au risque d'erreur α: l'échantillon est représentatif de la population de référence de variance σ_0 .
- sinon l'hypothèse H_0 est rejetée: l'échantillon n'est pas représentatif de la population de référence de variance σ_0 .

Deuxième cas: la moyenne μ de la population de référence est inconnue

Lorsque la moyenne μ est inconnue, la statistique S^2 est la meilleure estimation de la variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)^2$$

Sous l'hypothèse H_0, comme l'échantillon est gaussien, on a la statistique

$$y^2 = \frac{(n-1)S^2}{\sigma_0^2} = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma_0}\right)^2$$

Suit une loi du χ^2 à $n-1$ degrés de liberté, (en tant que somme de carrés de $N(0,1)$). On cherche les deux nombres tabulés a et b tels que:

$$P\left(Y^2 \geq a\right) = \frac{\alpha}{2}; P\left(Y^2 \geq b\right) = 1 - \frac{\alpha}{2}$$

Pour un risque d'erreur α fixé, et on décide que:

1. si $y^2 \in]a;b[$ l'hypothèse H_0 est acceptée au risque d'erreur α: l'échantillon est représentatif de la population de référence de variance σ_0 .

2. sinon l'hypothèse H_0 est rejetée: l'échantillon n'est pas représentatif de la population de référence de variance σ_0.

3.2.4 Comparaison d'une fréquence observée une fréquence théorique

Soit X une variable qualitative prenant deux modalités (succès $X = 1$, échec $X = 0$) observée sur une population et un échantillon extrait de cette population.

Le but est de savoir si un échantillon de fréquence observée $\frac{K}{n}$, estimateur de f, appartient à une population de référence connue de fréquence f_0 (H_0 vraie) ou une autre population inconnue de fréquence f (H_1 vraie).

L'hypothèse testée est la suivante:

$$\begin{cases} H_0 : f = f_0 \\ H_1 : f \neq f_0 \end{cases} \tag{1.25}$$

On calcule la valeur:

$$Z_{obs} = \frac{\left|\frac{K}{n} - f_0\right|}{\sqrt{\frac{f_0(1-f_0)}{n}}} \tag{1.26}$$

Suit la loi $N(0,1)$.

On détermine Z_{tab} lue sur la table de la loi normale centrée réduite pour un risque d'erreur α fixé, et on décide que:

1. si $Z_{obs} > Z_{tab}$ l'hypothèse H_0 est rejetée au risque d'erreur α : l'échantillon appartient une population de fréquence f et n'est pas représentatif de la population de référence de fréquence f_0.

2. si $Z_{obs} \leq Z_{tab}$ l'hypothèse H_0 est acceptée : l'échantillon est représentatif de la population de référence de fréquence f_0.

3.3 Tests d'homogénéité

3.3.1 Construction générale

On considère deux variables aléatoires X_1 et X_2 définies sur deux populations P_1 et P_2 respectivement. Ces variables aléatoires dépendent d'un paramètre inconnu θ_1 et θ_2 respectivement.

- Tester l'hypothèse $H_0 : \theta_1 = \theta_2$, contre $H_1 : \theta_1 \neq \theta_2$
- On choisit le risque α.
- On dispose d'un échantillon de X_1 et d'un échantillon de X_2 qui fournissent respectivement T_1 un estimateur de θ_1 et T_2 un estimateur de θ_2.

On détermine la variable de décision Z qui est une fonction de T_1 et T_2, et dont on connaît la loi de probabilité si H_0 est vraie.

- α étant connu, on calcule la région critique ou la région d'acceptation.
- On calcule la valeur z de Z partir des résultats des chantillons.

3.3.2 Tests de comparaison de deux moyennes

Soit X un caractère quantitatif continu observé sur deux populations suivant une loi normale et deux échantillons indépendants extraits de ces deux populations.

On fait l'hypothèse que les deux échantillons proviennent de deux populations dont les espérances μ_1 et μ_2 sont égales.

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases} \tag{1.27}$$

Pour tester cette hypothèse, il existe deux statistiques : la variance σ^2 de la population de référence est connue ou cette variance est inconnue et il faut l'estimer.

Premier cas: les variances des populations σ_1^2 et σ_2^2 sont connues

Sous l'hypothèse H_0 avec σ_1^2 et σ_2^2 sont connues, on a

$$\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx N(0,1) \text{ si } \sigma_1^2 \textbf{ et } \sigma_2^2 \textbf{ sont connues} \tag{1.28}$$

On calcule

$$Z_{obs} = \frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{1.29}$$

On détermine Z_{tab} lue sur la table de la loi normale centrée réduite pour un risque d'erreur α fixé , et on décide que:

1. si $Z_{obs} > Z_{tab}$ l'hypothèse H_0 est rejetée au risque d'erreur α: les deux échantillons sont extraits de deux populations ayant des espérances respectivement μ_1 et μ_2.

2. si $Z_{obs} \leq Z_{tab}$ l'hypothèse H_0 est acceptée: les deux échantillons sont extraits de deux populations ayant même espérance μ.

Deuxième cas: les variances des population σ_1^2 et σ_2^2 sont inconnues et égales

Sous l'hypothèse H_0 avec $\sigma_1^2 = \sigma_2^2 = \sigma^2$, on a

$$\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \text{Suit une loi de Student } (n_1 + n_2 - 2) \text{ degrés de liberté} \tag{1.30}$$

La variance commune σ^2 peut être estimée par:

$$s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

On calcule

$$t_{obs} = \frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{1.31}$$

On détermine t_{tab} lue sur la table de Student pour un risque d'erreur α fixé et $(n_1 + n_2 - 2)$ degrés de liberté, et on décide que:

1. si $t_{obs} > t_{tab}$ l'hypothèse H_0 est rejetée au risque d'erreur α: les deux échantillons sont extraits de deux populations ayant des espérances respectivement μ_1 et μ_2.

2. si $t_{obs} \leq t_{tab}$ l'hypothèse H_0 est acceptée: les deux échantillons sont extraits de deux populations ayant même espérance μ.

Troisième cas: les variances des populations σ_1^2 et σ_2^2 sont inconnues et inégales

Si les variances des populations ne sont pas connues et si leurs estimations à partir des échantillons sont significativement différentes, il faut considérer deux cas de figure selon la taille des échantillons comparés :

Cas où n_1 et $\mathbf{n_2}$> 30

La statistique utilisé est la même que pour le cas où les variances sont connues.

On calcule

$$Z_{obs} = \frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{1.32}$$

On détermine Z_{tab} lue sur la table de la loi normale centrée réduite pour un risque d'erreur fixé, et on décide que:

1. si $Z_{obs} > Z_{tab}$ l'hypothèse H_0 est rejetée au risque d'erreur α: les deux échantillons sont extraits de deux populations ayant des espérances respectivement μ_1 et μ_2.

2. si $Z_{obs} \leq Z_{tab}$ l'hypothèse H_0 est acceptée: les deux échantillons sont extraits de deux populations ayant même espérance μ.

3.3.3 Tests de comparaison de deux variances

Soit X, une variable aléatoire observée sur deux populations suivant une loi normale et deux échantillons indépendants extraits de ces deux populations.

On fait l'hypothèse que les deux échantillons proviennent de deux populations dont les variances sont égales.

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases} \tag{1.33}$$

Dans l'échantillon E_1 de taille n_1 (resp. l'échantillon E_2 de taille n_2), on estime la variance σ_1^2 (resp. σ_2^2) par:

$$s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n_1}(x_i - \overline{x})^2 \text{ et } s_2^2 = \frac{1}{n-1}\sum_{i=1}^{n_2}(x_i - \overline{x})^2$$

On calcule

$$F_{obs} = \frac{s_1^2}{s_2^2} \tag{1.34}$$

Telle que $F_{obs} \geq 1$ sinon on permute les échantillons de sorte que $F_{obs} \geq 1$.

On détermine F_{tab} lue sur la table de Fisher Snédécor pour un risque d'erreur α fixé avec $(n_1 - 1, n_2 - 1)$ degrés de liberté, on cherche F_{tab} tel que $P(F \geq F_{tab}) = \frac{\alpha}{2}$, et on décide que:

1. si $F_{obs} > F_{tab}$ l'hypothèse H_0 est rejetée au risque d'erreur α: les deux échantillons sont extraits de deux populations ayant des variances statistiquement différentes σ_1^2 et σ_2^2.

2. si $F_{obs} \leq F_{tab}$ l'hypothèse H_0 est acceptée: les deux échantillons sont extraits de deux populations ayant même variance σ^2.

3.3.4 Tests de comparaison de deux fréquences

Soit X une variable qualitative prenant deux modalités (succès $X = 1$, échec $X = 0$) observée sur deux populations et deux échantillons indépendants extraits de ces deux populations. On fait l'hypothèse que les deux échantillons proviennent de deux populations dont les probabilités de succès sont identiques.

$$\begin{cases} H_0 : F_1 = F_2 \\ H_1 : F_1 \neq F_2 \end{cases} \tag{1.35}$$

Dans l'échantillon E_1 de taille n_1 on estime la fréquence F_1 par f_1 et dans l'échantillon E_2 de taille n_2 on estime la fréquence F_2 par f_2 et en regroupant les deux échantillons, on peut estimer F par:

$$f = \frac{n_1.f_1 + n_2.f_2}{n_1 + n_2}$$

Remarque: conditions de validité du test: $n_1.f_1 \geq 5; n_1(1-f_1) \geq 5; n_2.f_2 \geq 5; n_2(1-f_2) \geq 5$.

On calcule

$$z_{obs} = \frac{|f_1 - f_2|}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) f\,(1-f)}} \tag{1.36}$$

On détermine z_{tab} pour la loi normale $N(0,1)(\Phi\,(z_{tab}) = 1 - \frac{\alpha}{2})$, et on décide que:

1. si $z_{obs} > z_{tab}$ l'hypothèse H_0 est rejetée au risque d'erreur α: les deux échantillons sont extraits de deux populations ayant des probabilités de succès respectivement F_1 et F_2.

2. si $z_{obs} \leq z_{tab}$ l'hypothèse H_0 est acceptée: les deux échantillons sont extraits de deux populations ayant même probabilité de succès F.

3.4 Test du Chi-deux d'indépendance

Le test du χ^2 d'indépendance a pour objectif d'évaluer si deux variables qualitatives X_1 et X_2 respectivement p et k modalités sont liés, les deux variables étant observées sur un échantillon de taille N.

Les hypothèses du test du χ^2 d'indépendance sont les suivantes :

- H_0 : Les variables X_1 et X_2 sont indépendantes.
- H_1 : Il existe une liaison entre X_1 et X_2 .

3.4.1 Calcul de la statistique de test:

Considérons, le tableau de contingence des effectifs observés suivant :

Variable X_1	Variable X_2	Total
Modalités	$M_1, M_2..M_k$	t
M_1	$e_{11}, e_{12}..e_{1k}$	t_1
M_2	$e_{21}, e_{22}..e_{2k}$	t_2
.	...	...
.	...	...
M_k	$e_{p1}, e_{p2}..e_{pk}$	t_k
Total	$n_1, n_2..n_k$	N

Tab:01 Tableau de contingence pour le test d'indépendance

Le principe du test du χ^2 consiste à calculer, pour chaque case du tableau, l'effectif théorique qui devrait être observé sous l'hypothèse nulle. Sous cette hypothèse, les effectifs sont répartis en proportion égale.

On définit l'effectif théorique E_{ij} associé à la case $\{i, j\}$ du tableau par la quantité suivante:

$$E_{ij} = \frac{n_j \times t_i}{N}$$

Sous l'hypothèse nulle, les effectifs observés et les effectifs théoriques doivent être sensiblement proches, donc la somme de leurs différences devrait être proche de zéro. Aussi, le principe du test du χ^2 se base sur l'évaluation de la somme de ces différences par rapport une valeur seuil. Intuitivement, si cette somme de différences excède une certaine valeur, cela signifie que les effectifs observés et les effectifs théoriques sont différents et par conséquent l'hypothèse peut être remise en cause.

Sous H_0, le test du χ^2 a pour statistique de test:

$$\chi^2_{obs} = \sum_{i=1}^{p}\sum_{j=1}^{k}\frac{(e_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(p-1)(k-1)} \tag{1.37}$$

Condition d'application du test

Le test du χ^2 est sensible aux petits effectifs. Aussi, le test est considéré comme applicable lorsque les effectifs théoriques E_{ij} sont supérieurs ou égaux à 5. En pratique,

si cette condition n'est pas réalisée, la technique consiste à regrouper certaines modalités (ex : regrouper les yeux noirs et les yeux marrons) an de, par construction, augmenter les valeurs des effectifs théoriques.

La région critique conduisant au rejet de H_0 est définie par :

$$\left[\chi^2_{(1-\alpha);(p-1)(k-1)}; +\infty\right[\tag{1.38}$$

Où $\chi^2_{(1-\alpha);(p-1)(k-1)}$ correspond au quantile d'ordre $(1-\alpha)$ de la loi du χ^2 à $(p-1)(k-1)$ degrés de liberté.

Décision:

- Si la valeur de la statistique de test χ^2 est inférieure à la valeur seuil $\chi^2_{(1-\alpha);(p-1)(k-1)}$ alors on accepte l'hypothèse nulle. Les variables X_1 et X_2 sont indépendantes. (c'est-à-dire leurs distributions sont indépendantes).

- Si la valeur de la statistique de test χ^2 est supérieure à la valeur seuil $\chi^2_{(1-\alpha);(p-1)(k-1)}$ alors on rejette l'hypothèse nulle. Il existe une liaison significative entre X_1 et X_2 (c'est-à-dire leurs distributions sont dépendantes).

3.5 Tests non paramétrique

Les résultats de toute analyse statistique dépendent du modèle utilisé, et le choix de celui-ci est donc très important. Pour déterminer si un échantillon peut provenir d'une distribution spécifique, les chercheurs ont développé plusieurs tests statistiques. La plupart d'entre elles sont basées sur les fonctions de distribution empiriques *(EDF)*, la plus ancienne étant la statistique D_n de Kolmogorov-Smirnov *(K-S)* (Kolmogorov 1933). Plus tard, les tests W^2 de Cramer-Von Mises se sont avérés plus puissants que le test *K-S* (D_n) contre une large classe d'hypothèses alternatives. La statistique A^2 d'Anderson-Darling (Anderson et Darling 1954) peut être considérée comme une distribution limite de W^2 et donne plus de poids aux queues que la statistique D_n (voir Darling 1957). Watson (1961a, 1962b) a proposé une nouvelle statistique de test U^2 comme généralisation de la statistique de test W^2 de Cramer-Von Mises. Lorsque les distributions hypothétiques

sont complètement spécifiées, les statistiques *EDF* sont sans distribution, mais dans le cas de paramètres inconnus, leur distribution dépendra non seulement de la taille de l'échantillon, mais aussi de la distribution hypothétique, des paramètres estimés et de la méthode d'estimation des paramètres (voir Lawless 1982). De nombreux auteurs ont reconnu ce fait et ont étendu l'utilisation de différentes statistiques de test à ce cas. En utilisant des méthodes numériques, ils ont développé des statistiques de test modifiées où les paramètres inconnus sont remplacés par leurs estimations.

3.5.1 Test de Kolmogorov-Smirnov

Né en 1933 grâce aux travaux de Andrei Nicolaiévitch Kolmogorov et étendu à la comparaison de deux échantillons en 1939 par ceux de Vladimir Ivanovitch Smirnov. Le test de Kolmogorov-Smirnov est une approch non paramétriques d'ajustement, qui s'étend à la comparaison de deux fonctions de répartition empiriques, et permet alors de tester l'hypothèse que deux échantillons sont issus de la mêmes loi.

On utilise de préférence au test d'ajustement du Chi-deux lorsque le caractère observé peut prendre des valeurs continues.

Ce test non paramétrique effectue une comparaison entre la fonction de répartition empirique et la fonction de répartition théorique de la loi de probabilités considérée.

La statistique D_n du test de Komogorov-Smirnov est le test d'adéquation le plus ancien et le plus utilisé. Elle est définie par

$$D_n = \max\left[D^+; D^-\right] \tag{1.39}$$

D'où

$$D^+ = \max_{1\leq i\leq n}\left[\left(\frac{i}{n}\right) - F\left(t_{(i)}\right)\right] \tag{1.40}$$

Et

$$D^- = \max_{1\leq i\leq n}\left[F\left(t_{(i)}\right) - \left(\frac{i-1}{n}\right)\right] \tag{1.41}$$

3.5.2 Test de Cramer-Von Mises

Né en 1930 suite aux travaux de Harald Cramér et Richard Elder von Mises, le test de Cramér-Von Mises est une approche non paramétrique permettant de tester si une variable continue X suit une loi de distribution fixé.

Le test W_n^2 de Cramér-Von Mises peut êtrevu comme une version plus puissante du test de Kolmogorov-Smirnov. Il est défini par

$$W_n^2 = \sum_{i=1}^{n} \left[F\left(t_{(i)}\right) - \frac{2i-1}{2n} \right]^2 + \frac{1}{12n} \tag{1.42}$$

Partie II

La statistique avec R

```
############################Application sur R#############################
#########################################################################
##################Chapitre 01 : La statistique descriptive#############
> iris
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa
17	5.4	3.9	1.3	0.4	setosa
18	5.1	3.5	1.4	0.3	setosa
19	5.7	3.8	1.7	0.3	setosa
20	5.1	3.8	1.5	0.3	setosa
21	5.4	3.4	1.7	0.2	setosa
22	5.1	3.7	1.5	0.4	setosa
23	4.6	3.6	1.0	0.2	setosa
24	5.1	3.3	1.7	0.5	setosa
25	4.8	3.4	1.9	0.2	setosa
26	5.0	3.0	1.6	0.2	setosa
27	5.0	3.4	1.6	0.4	setosa
28	5.2	3.5	1.5	0.2	setosa
29	5.2	3.4	1.4	0.2	setosa
30	4.7	3.2	1.6	0.2	setosa
31	4.8	3.1	1.6	0.2	setosa
32	5.4	3.4	1.5	0.4	setosa
33	5.2	4.1	1.5	0.1	setosa
34	5.5	4.2	1.4	0.2	setosa
35	4.9	3.1	1.5	0.2	setosa
36	5.0	3.2	1.2	0.2	setosa
37	5.5	3.5	1.3	0.2	setosa
38	4.9	3.6	1.4	0.1	setosa
39	4.4	3.0	1.3	0.2	setosa
40	5.1	3.4	1.5	0.2	setosa
41	5.0	3.5	1.3	0.3	setosa
42	4.5	2.3	1.3	0.3	setosa
43	4.4	3.2	1.3	0.2	setosa
44	5.0	3.5	1.6	0.6	setosa
45	5.1	3.8	1.9	0.4	setosa
46	4.8	3.0	1.4	0.3	setosa
47	5.1	3.8	1.6	0.2	setosa
48	4.6	3.2	1.4	0.2	setosa
49	5.3	3.7	1.5	0.2	setosa
50	5.0	3.3	1.4	0.2	setosa
51	7.0	3.2	4.7	1.4	versicolor
52	6.4	3.2	4.5	1.5	versicolor
53	6.9	3.1	4.9	1.5	versicolor
54	5.5	2.3	4.0	1.3	versicolor
55	6.5	2.8	4.6	1.5	versicolor
56	5.7	2.8	4.5	1.3	versicolor
57	6.3	3.3	4.7	1.6	versicolor
58	4.9	2.4	3.3	1.0	versicolor
59	6.6	2.9	4.6	1.3	versicolor
60	5.2	2.7	3.9	1.4	versicolor
61	5.0	2.0	3.5	1.0	versicolor
62	5.9	3.0	4.2	1.5	versicolor
63	6.0	2.2	4.0	1.0	versicolor
64	6.1	2.9	4.7	1.4	versicolor
65	5.6	2.9	3.6	1.3	versicolor
66	6.7	3.1	4.4	1.4	versicolor
67	5.6	3.0	4.5	1.5	versicolor
68	5.8	2.7	4.1	1.0	versicolor
69	6.2	2.2	4.5	1.5	versicolor
70	5.6	2.5	3.9	1.1	versicolor

71	5.9	3.2	4.8	1.8	versicolor
72	6.1	2.8	4.0	1.3	versicolor
73	6.3	2.5	4.9	1.5	versicolor
74	6.1	2.8	4.7	1.2	versicolor
75	6.4	2.9	4.3	1.3	versicolor
76	6.6	3.0	4.4	1.4	versicolor
77	6.8	2.8	4.8	1.4	versicolor
78	6.7	3.0	5.0	1.7	versicolor
79	6.0	2.9	4.5	1.5	versicolor
80	5.7	2.6	3.5	1.0	versicolor
81	5.5	2.4	3.8	1.1	versicolor
82	5.5	2.4	3.7	1.0	versicolor
83	5.8	2.7	3.9	1.2	versicolor
84	6.0	2.7	5.1	1.6	versicolor
85	5.4	3.0	4.5	1.5	versicolor
86	6.0	3.4	4.5	1.6	versicolor
87	6.7	3.1	4.7	1.5	versicolor
88	6.3	2.3	4.4	1.3	versicolor
89	5.6	3.0	4.1	1.3	versicolor
90	5.5	2.5	4.0	1.3	versicolor
91	5.5	2.6	4.4	1.2	versicolor
92	6.1	3.0	4.6	1.4	versicolor
93	5.8	2.6	4.0	1.2	versicolor
94	5.0	2.3	3.3	1.0	versicolor
95	5.6	2.7	4.2	1.3	versicolor
96	5.7	3.0	4.2	1.2	versicolor
97	5.7	2.9	4.2	1.3	versicolor
98	6.2	2.9	4.3	1.3	versicolor
99	5.1	2.5	3.0	1.1	versicolor
100	5.7	2.8	4.1	1.3	versicolor
101	6.3	3.3	6.0	2.5	virginica
102	5.8	2.7	5.1	1.9	virginica
103	7.1	3.0	5.9	2.1	virginica
104	6.3	2.9	5.6	1.8	virginica
105	6.5	3.0	5.8	2.2	virginica
106	7.6	3.0	6.6	2.1	virginica
107	4.9	2.5	4.5	1.7	virginica
108	7.3	2.9	6.3	1.8	virginica
109	6.7	2.5	5.8	1.8	virginica
110	7.2	3.6	6.1	2.5	virginica
111	6.5	3.2	5.1	2.0	virginica
112	6.4	2.7	5.3	1.9	virginica
113	6.8	3.0	5.5	2.1	virginica
114	5.7	2.5	5.0	2.0	virginica
115	5.8	2.8	5.1	2.4	virginica
116	6.4	3.2	5.3	2.3	virginica
117	6.5	3.0	5.5	1.8	virginica
118	7.7	3.8	6.7	2.2	virginica
119	7.7	2.6	6.9	2.3	virginica
120	6.0	2.2	5.0	1.5	virginica
121	6.9	3.2	5.7	2.3	virginica
122	5.6	2.8	4.9	2.0	virginica
123	7.7	2.8	6.7	2.0	virginica
124	6.3	2.7	4.9	1.8	virginica
125	6.7	3.3	5.7	2.1	virginica
126	7.2	3.2	6.0	1.8	virginica
127	6.2	2.8	4.8	1.8	virginica
128	6.1	3.0	4.9	1.8	virginica
129	6.4	2.8	5.6	2.1	virginica
130	7.2	3.0	5.8	1.6	virginica
131	7.4	2.8	6.1	1.9	virginica
132	7.9	3.8	6.4	2.0	virginica
133	6.4	2.8	5.6	2.2	virginica
134	6.3	2.8	5.1	1.5	virginica
135	6.1	2.6	5.6	1.4	virginica
136	7.7	3.0	6.1	2.3	virginica
137	6.3	3.4	5.6	2.4	virginica
138	6.4	3.1	5.5	1.8	virginica
139	6.0	3.0	4.8	1.8	virginica
140	6.9	3.1	5.4	2.1	virginica
141	6.7	3.1	5.6	2.4	virginica
142	6.9	3.1	5.1	2.3	virginica
143	5.8	2.7	5.1	1.9	virginica
144	6.8	3.2	5.9	2.3	virginica
145	6.7	3.3	5.7	2.5	virginica
146	6.7	3.0	5.2	2.3	virginica
147	6.3	2.5	5.0	1.9	virginica

```
148          6.5         3.0          5.2         2.0  virginica
149          6.2         3.4          5.4         2.3  virginica
150          5.9         3.0          5.1         1.8  virginica
```

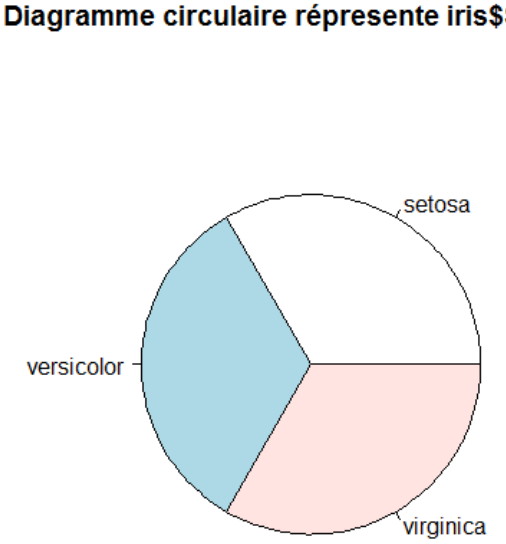
```
> L=iris$Petal.Length
> L
```

```
 [1] 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 1.5 1.6 1.4 1.1 1.2 1.5 1.3 1.4 1.7 1.5 1.7 1.5 1.0 1.7 1.9
[26] 1.6 1.6 1.5 1.4 1.6 1.6 1.5 1.5 1.4 1.5 1.2 1.3 1.4 1.3 1.5 1.3 1.3 1.3 1.6 1.9 1.4 1.6 1.4 1.5 1.4
[51] 4.7 4.5 4.9 4.0 4.6 4.5 4.7 3.3 4.6 3.9 3.5 4.2 4.0 4.7 3.6 4.4 4.5 4.1 4.5 3.9 4.8 4.0 4.9 4.7 4.3
[76] 4.4 4.8 5.0 4.5 3.5 3.8 3.7 3.9 5.1 4.5 4.5 4.7 4.4 4.1 4.0 4.4 4.6 4.0 3.3 4.2 4.2 4.2 4.3 3.0 4.1
[101] 6.0 5.1 5.9 5.6 5.8 6.6 4.5 6.3 5.8 6.1 5.1 5.3 5.5 5.0 5.1 5.3 5.5 6.7 6.9 5.0 5.7 4.9 6.7 4.9 5.7
[126] 6.0 4.8 4.9 5.6 5.8 6.1 6.4 5.6 5.1 5.6 6.1 5.6 5.5 4.8 5.4 5.6 5.1 5.1 5.9 5.7 5.2 5.0 5.2 5.4 5.1
```

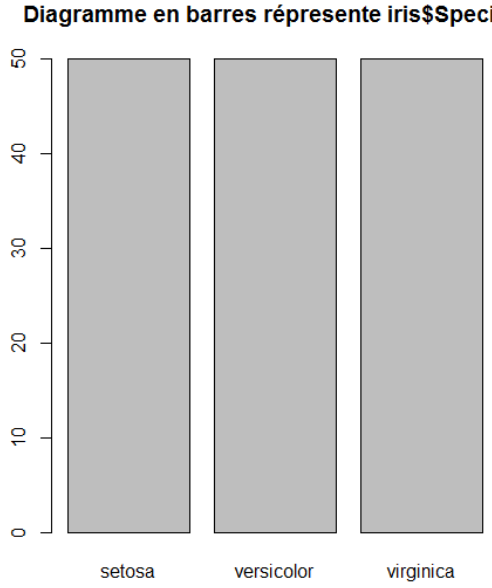
```
> T=table(iris$Species)
> T
    setosa versicolor  virginica
        50         50         50
```

```
##########################################################################
```

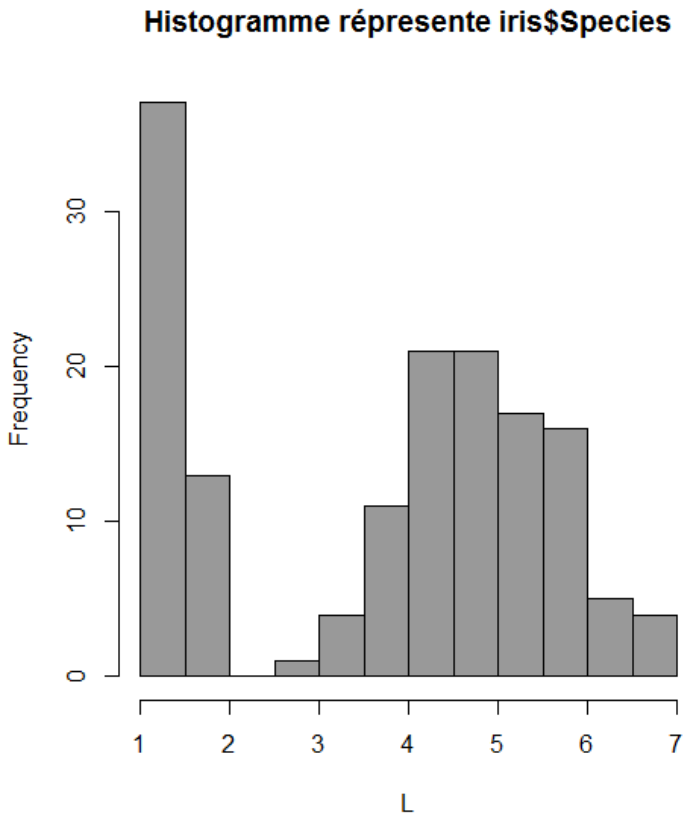
```
#Représentation graphique des séries statistiques
> #Variable qualitative
> #Diagramme circulaire (diagramme en secteurs)
> pie(T,main='Diagramme circulaire répresente iris$Species ')
```




```
> #Diagramme en barres
> barplot(T,main='Diagramme en barres répresente iris$Species ')
```



```
> #Variable quantitative continue
> #Histogramme
> hist(L,col = grey(0.6),main = "Histogramme répresente iris$Species")
```

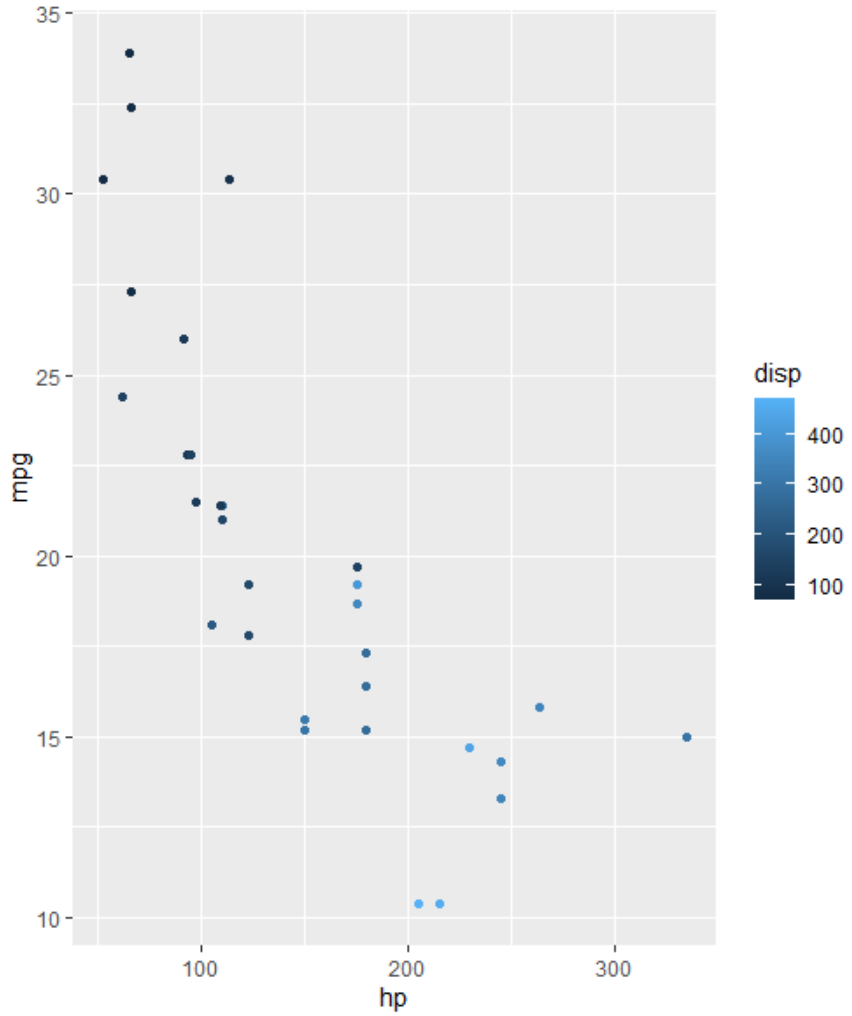



```
> #Représentation graphique avec ggplot2
> summary(mtcars)
```

mpg	cyl	disp	hp	drat	wt
Min. :10.40	Min. :4.000	Min. : 71.1	Min. : 52.0	Min. :2.760	Min. :1.513
1st Qu.:15.43	1st Qu.:4.000	1st Qu.:120.8	1st Qu.: 96.5	1st Qu.:3.080	1st Qu.:2.581
Median :19.20	Median :6.000	Median :196.3	Median :123.0	Median :3.695	Median :3.325
Mean :20.09	Mean :6.188	Mean :230.7	Mean :146.7	Mean :3.597	Mean :3.217
3rd Qu.:22.80	3rd Qu.:8.000	3rd Qu.:326.0	3rd Qu.:180.0	3rd Qu.:3.920	3rd Qu.:3.610
Max. :33.90	Max. :8.000	Max. :472.0	Max. :335.0	Max. :4.930	Max. :5.424

qsec	vs	am	gear	carb
Min. :14.50	Min. :0.0000	Min. :0.0000	Min. :3.000	Min. :1.000
1st Qu.:16.89	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:3.000	1st Qu.:2.000
Median :17.71	Median :0.0000	Median :0.0000	Median :4.000	Median :2.000
Mean :17.85	Mean :0.4375	Mean :0.4062	Mean :3.688	Mean :2.812
3rd Qu.:18.90	3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:4.000	3rd Qu.:4.000
Max. :22.90	Max. :1.0000	Max. :1.0000	Max. :5.000	Max. :8.000

```
> library(ggplot2)
> library(dplyr)
> ggplot(data = mtcars,aes(x = hp, y = mpg, col = disp)) + geom_point()
```




```
> #Représentation graphique avec GGally
> library(GGally)
> data(tips, package="reshape")
> ggpairs(data=tips,variablescolumns=1:3,aes(color="sex"))
```

```
#########################################################################
> #Paramétres de position
> mean(L)
[1] 3.758

>median(L)
[1] 4.35

> mode(L)
[1] "numeric"

> min(L)
[1] 1

> max(L)
[1] 6.9

> quantile(L)
  0%  25%  50%  75% 100%
1.00 1.60 4.35 5.10 6.90

> #Paramétres de dispersion
> range(L)
[1] 1.0 6.9

> var(L)
[1] 3.116278

> sd(L)
[1] 1.765298

#########################################################################

> #manipulation avec dplyr
> library(dplyr)
> data = data.frame(gender = c("M", "M", "F"),age = c(20, 60, 30),height
= c(180, 200, 150))
> data
  gender age height
1      M  20    180
2      M  60    200
3      F  30    150
> mutate (data, height2 = height / 100)
  gender age height height2
1      M  20    180     1.8
2      M  60    200     2.0
3      F  30    150     1.5

> select (data, age)
  age
1  20
2  60
3  30

> filter (data, height > 160)
  gender age height
1      M  20    180
2      M  60    200

> arrange (data, height)
  gender age height
1      F  30    150
2      M  20    180
3      M  60    200
```

```
> summarise(data, average_height = mean(height))
  average_height
1       176.6667

#####################################################################
##################Chapitre 02: Estimation des paramétres#############
#####################################################################

#Estimation par intervalle de confiance
> x = c(1,4,3,5,3)
> x
[1] 1 4 3 5 3

> # Intervalle de confiance de la moyenne
> library(KefiR)
> t.test(x, conf.level=0.99 )
#le paramètre conf.level indique le niveau de confiance (95% par défaut)

      One Sample t-test

data:  x
t = 4.8242, df = 4, p-value = 0.008497
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 0.145989 6.254011
sample estimates:
mean of x
      3.2

> # Intervalle de confiance de la variance
> library(EnvStats)
> varTest(x = x, alternative = "less")
$statistic
Chi-Squared
        8.8

$parameters
df
 4

$p.value
[1] 0.9337024

$estimate
variance
     2.2

$null.value
variance
       1

$alternative
[1] "less"

$method
[1] "Chi-Squared Test on Variance"

$data.name
[1] "x"

$conf.int
     LCL      UCL
 0.00000 12.38176
```

```
attr(,"conf.level")
[1] 0.95

attr(,"class")
[1] "htestEnvStats"

> # Intervalle de confiance de la proportion
> prop.test(x=60,n=100,p= 0.5, alternative = "greater", correct = FALSE)
#Cas des grands échantillons

      1-sample proportions test without continuity correction

data:  60 out of 100, null probability 0.5
X-squared = 4, df = 1, p-value = 0.02275
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
 0.5178095 1.0000000
sample estimates:
  p
0.6

#Estimation par nleqslv
> library(nleqslv)
> target <- function(x){z = x[1]/(x[1]+x[2])
+ y = numeric(2)
+ y[1] <- z*exp(-x[2]*(x[2]+z*(1-exp(-x[1]/z))))-0.00680
+ y[2] <- z/x[2]*(1-exp(-x[2]))-exp(-x[2])*z/x[1]*(1-exp(-x[1]))-3.43164
+ y}
> # Usage
> xstart <- c(1,1)
> target(xstart)
[1]  0.1125757 -3.2318518

> nleqslv(xstart, target, control=list(ftol=.0001, allowSingular=TRUE),j
acobian=TRUE,method="Newton")
$x
[1] 496832.823000     -2.113625

$fvec
[1] 0.08821001 0.01182427

$termcd
[1] 2

$message
[1] "x-values within tolerance 'xtol'"

$scalex
[1] 1 1

$nfcnt
[1] 43

$njcnt
[1] 39

$iter
[1] 39

$jac
              [,1]        [,2]
[1,] -2.541843e-12  0.3066201
[2,]  4.018961e-12 -2.2874116
```

```
#########################################################################
#######################Chapitre 03: Tests d'hypothéses###################
#########################################################################

#Tests de conformité et d'homogénéite
#Comparaison de 2 moyennes
> ech1 <- c(5, 6, 7, 8, 10, 11)
> ech2 <- c(6, 7, 9, 9, 12)
> t.test(x = ech1, y = ech2, alternative = "two.sided", var.equal =TRUE)

        Two Sample t-test

data:  ech1 and ech2
t = -0.54805, df = 9, p-value = 0.597
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.931192  2.397859
sample estimates:
mean of x mean of y
 7.833333  8.600000

#########################################################################
#Comparaison de 2 proportions
> prop.test(c(5, 10),c(100, 80),alternative ="two.sided")
#ne s'applique que dans le cas des grands échantillons

2-sample test for equality of proportions with continuity correction

data:  c(5, 10) out of c(100, 80)
X-squared = 2.3645, df = 1, p-value = 0.1241
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.17037305  0.02037305
sample estimates:
prop 1 prop 2
 0.050  0.125

#########################################################################
#Comparaison de 2 variances
> e <- c(5, 6, 7, 8, 10, 11)
> f <- c(6, 7, 9, 9, 12)
> var.test(x = e, y = f, alternative = "two.sided" )

        F test to compare two variances

data:  e and f
F = 1.0126, num df = 5, denom df = 4, p-value = 0.9813
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1081298 7.4808151
sample estimates:
ratio of variances
          1.012579

#########################################################################
#Test du Chi-deux d'indépendance
> tableau <- matrix(c(20, 30, 26, 24), nrow = 2, byrow = FALSE)
> tableau
     [,1] [,2]
[1,]   20   26
[2,]   30   24

> TEST <- chisq.test(tableau)
> TEST
```

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  tableau
X-squared = 1.0064, df = 1, p-value = 0.3158

##########################################################################
#Test EDF
#Test de Kolmogorov-Smirnov
> X1 <- c(8.43, 8.70, 11.27, 12.92, 13.05, 13.050001, 13.17, 13.44, 13.8
9, 18.90)
> ks.test(X1,"pnorm",mean=13, sd=3)

        Exact one-sample Kolmogorov-Smirnov test

data:  X1
D = 0.28336, p-value = 0.3326
alternative hypothesis: two-sided


#Test de Cramer-Von Mises
> library(goftest)
> set.seed(0)#make this example reproducible
> data <- rnorm(100)#create dataset of 100 random values generated from
a normal distribution
> cvm.test(data, 'pnorm')

        Cramer-von Mises test of goodness-of-fit
        Null hypothesis: Normal distribution
        Parameters assumed to be fixed

data:  data
omega2 = 0.078666, p-value = 0.7007
```

Partie III

La distribution q-exponentielle

CHAPITRE 4

La distribution q-exponentielle

La distribution q-exponentielle a été introduite pour la première fois par Constantino Tsallis en 1988 comme généralisation de la distribution exponentielle standard. Tsallis était un physicien brésilien qui était intéressé à développer un nouveau cadre pour décrire la mécanique statistique non-extensive. Il a observé que de nombreux systèmes complexes présentent un comportement non gaussien, avec des queues lourdes dans leurs distributions de probabilité. Il a suggéré que ce comportement pourrait être décrit par une nouvelle classe de distributions de probabilités, qu'il a appelée distributions de type q.

Depuis son introduction, la distribution q-exponentielle a fait l'objet de nombreuses recherches et a trouvé de nombreuses applications en physique, finance, économie, et d'autres domaines. De nombreux chercheurs ont utilisé la distribution q-exponentielle pour modéliser le comportement de systèmes complexes qui présentent un comportement à forte queue, et il s'est avéré être un outil utile pour comprendre les propriétés statistiques de ces systèmes.

Récemment, il y a eu un intérêt croissant pour la distribution q-exponentielle et d'autres distributions de type q dans le contexte de l'apprentissage profond et de l'intelligence artificielle. Certains chercheurs ont proposé d'utiliser ces distributions pour modéliser l'incertitude dans les modèles d'apprentissage profond, dans le but d'améliorer leur performance sur des applications réelles.

4.1 Présentation du modèle

La distribution q-exponentielle est une distribution de probabilité qui est une généralisation de la distribution exponentielle standard. Il est caractérisé par un paramètre q, qui reflète le degré de non-extensivité ou de non-additivité du système modélisé. La fonction q-exponentielle est définie comme suit

$$f_q(t) = \exp_q(t) = \begin{cases} [1-(1-q)t]^{\frac{1}{1-q}}, \text{si } [1-(1-q)t] > 0 \\ 0, \text{sinon.} \end{cases} \tag{3.1}$$

et l'inverse le q-logarithme est défini par

$$\ln_q(t) = \frac{t^{1-q}-1}{1-q} \tag{3.2}$$

La densité de probabilité de la distribution q-exponentielle est

$$f(t) = \frac{2-q}{\eta} \exp_q\left(-\frac{t}{\eta}\right) = \frac{2-q}{\eta}\left[1-\frac{(1-q)t}{\eta}\right]^{\frac{1}{1-q}} \tag{3.3}$$

où $q < 2$ détermine la paramètre de forme, alors que $\eta > 0$ est le paramètre d'échelle. Dans la limite $q \longrightarrow 1$, Eq. (3.3) deviennent une distribution exponentielle. Lorsque $q < 1$.

La fonction de répartition de la distribution q-exponentielle est

$$F(t) = 1 - \left[\exp_q\left(-\frac{t}{\eta}\right)\right]^{2-q} = 1 - \left[1-\frac{(1-q)t}{\eta}\right]^{\frac{2-q}{1-q}}, t \geq 0. \tag{3.4}$$

Par définition, le taux de hasard

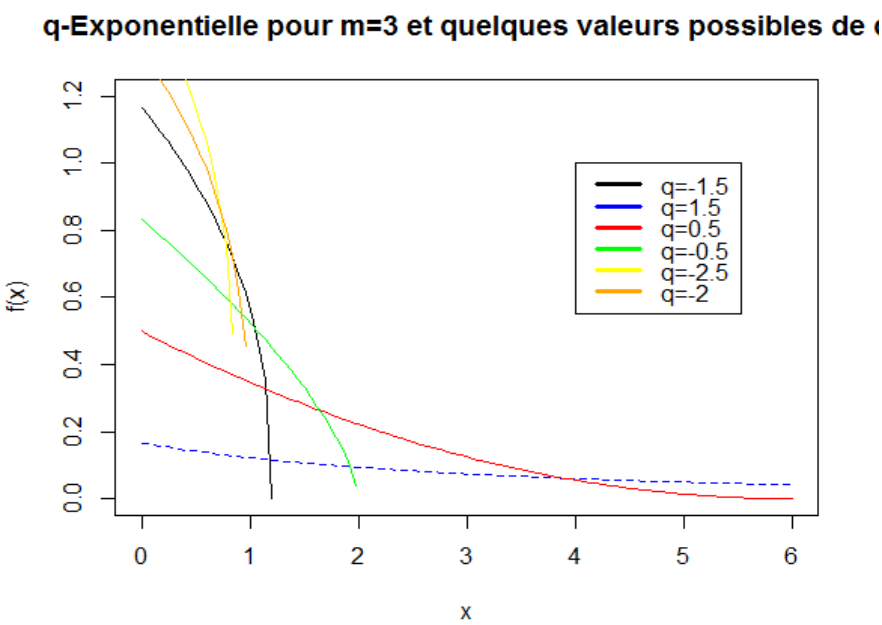
$$h(t) = f(x)/R(t)$$

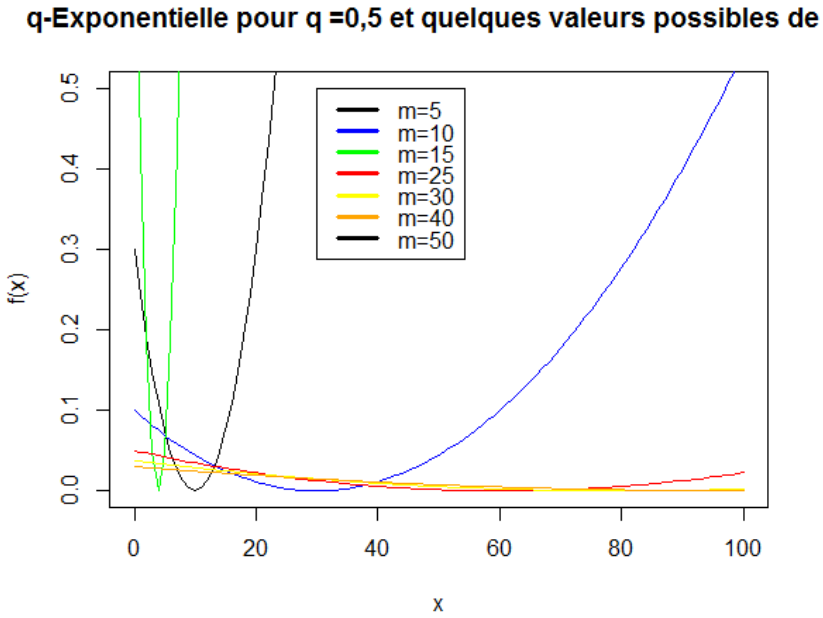
où $R(t)$ est la fonction de fiabilité avec

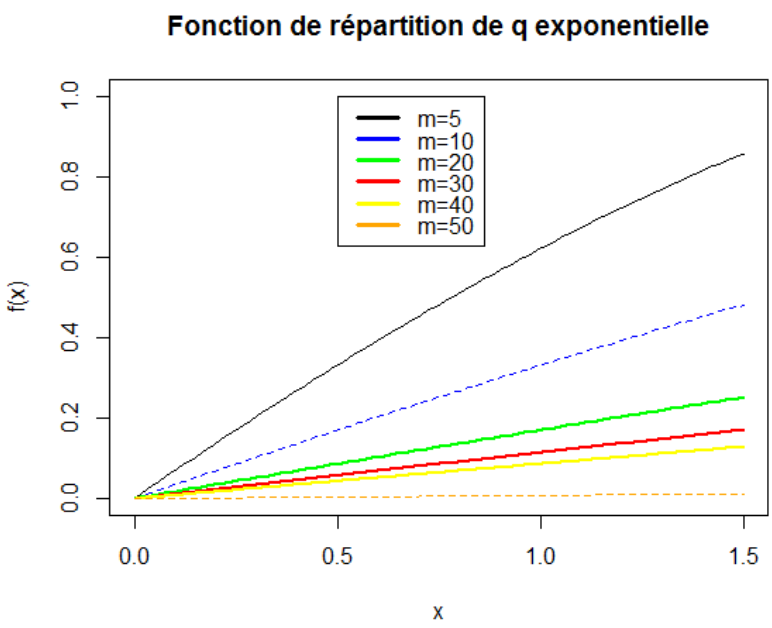
$$R(t) = 1 - F(t)$$

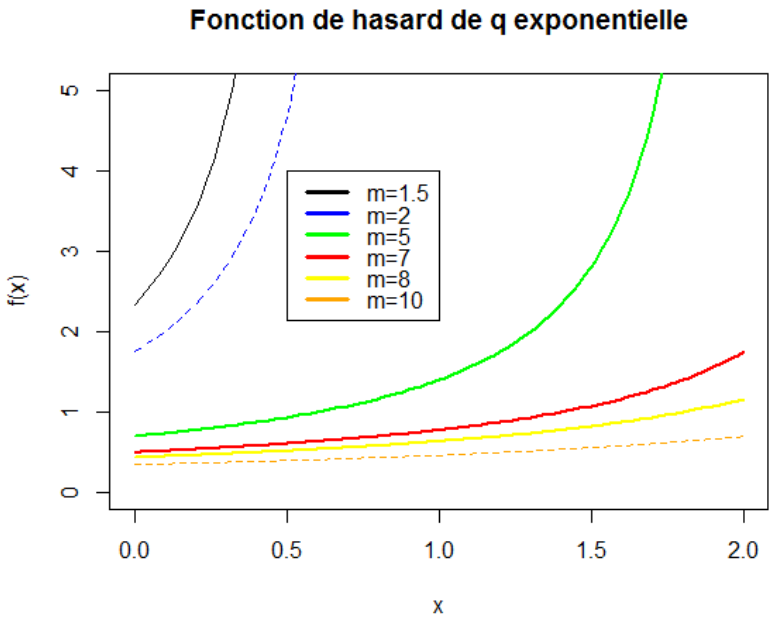
Il s'ensuit donc que

$$h(t) = \frac{(2-q)}{\eta}\left[\exp_q\left(-\frac{t}{\eta}\right)\right]^{-(1-q)} = \frac{(2-q)}{\eta}\left[1-\frac{(1-q)t}{\eta}\right]^{-1}. \tag{3.5}$$


q-Exponentielle pour m=3 et quelques valeurs possibles de
f(x)
x
q=-1.5
q=1.5
q=0.5
q=-0.5
q=-2.5
q=-2


q-Exponentielle pour q =0,5 et quelques valeurs possibles de
f(x)
x
m=5
m=10
m=15
m=25
m=30
m=40
m=50


Fonction de répartition de q exponentielle
m=5
m=10
m=20
m=30
m=40
m=50
f(x)
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
x


Fonction de hasard de q exponentielle
m=1.5
m=2
m=5
m=7
m=8
m=10
f(x)
0
1
2
3
4
5
0.0
0.5
1.0
1.5
2.0
x

Le taux de hasard de la distribution q-exponentiel peut être monotone croissant, monotone décroissant ou constant pour $q < 1, 1 < q < 2$ et $q = 1$, respectivement. En fait, il s'agit d'une caractéristique importante de la distribution q-exponentielle, en particulier dans le contexte de la fiabilité, car elle permet de modéliser chacune des trois phases de la courbe de la baignoire.

Afin de générer des échantillons pseudo-aléatoires qui suivent une distribution q-Exponentielle, Eq. (3.6) peut être utilisé

$$t = \frac{\eta \left[1 - U^{\left(\frac{1-q}{2-q}\right)}\right]}{1-q} \tag{3.6}$$

où U dénote un nombre pseudo-aléatoire uniforme. La formule en Eq. (3.6) est obtenue au moyen de la méthode de transformation inverse.

4.2 Estimation des paramètres

La distribution q-exponentielle est une généralisation de la distribution exponentielle standard qui a un paramètre supplémentaire, q, qui contrôle la forme de la distribution. Pour estimer les paramètres de la distribution q-exponentielle, on utilise couramment l'estimation du maximum de vraisemblance.

La fonction de vraisemblance pour une distribution q-exponentielle avec les paramètres q et η, compte tenu d'un échantillon d'observations $(t_1, t_2, ..., t_n)$ est :

$$L(t \mid q, \eta) = \prod_{i=1}^{n} \frac{2-q}{\eta} \exp\left(-\frac{t_i}{\eta}\right) = \prod_{i=1}^{n} \frac{2-q}{\eta} \left[1 - \frac{(1-q)t_i}{\eta}\right]^{\frac{1}{1-q}} \tag{3.7}$$

La fonction de log-vraisemblance est

$$l(t \mid q, \eta) = n \ln\left(\frac{2-q}{\eta}\right) + \frac{1}{1-q} \sum_{i=1}^{n} \ln\left[1 - \frac{(1-q)t_i}{\eta}\right] \tag{3.8}$$

Les fonctions de scores sont

$$\frac{\partial l}{\partial q} = -\frac{\eta}{2-q} + \frac{1}{(1-q)^2} \sum_{i=1}^{n} \ln\left[1 - \frac{(1-q)t_i}{\eta}\right] + \frac{1}{1-q} \sum_{i=1}^{n} \frac{t_i}{\eta - (1-q)t_i} = 0, \tag{3.9}$$

$$\frac{\partial l}{\partial \eta} = \frac{1}{\eta}\left[-n + \sum_{i=1}^{n} \frac{t_i}{\eta - (1-q)t_i}\right] = 0. \tag{3.10}$$

Pour estimer les paramètres q et η, nous devons trouver les valeurs qui maximisent la fonction de vraisemblance. Ceci peut être fait en utilisant des méthodes d'optimisation numérique telles que la méthode Newton-Raphson ou la méthode de descente de gradient.

Une fois les paramètres q et η estimés, nous pouvons utiliser la distribution q-exponentielle pour modéliser les données et faire des prédictions. Par exemple, nous pourrions utiliser la distribution pour estimer la probabilité d'observer une certaine valeur de x, ou pour générer de nouveaux échantillons à partir de la distribution.

4.3 Différents EDF Tests

Les tests d'adéquation *EDF* mesurent la différence entre la distribution de la fonction hypothétique F et la fonction de distribution empirique F_n et cette quantité est comparée aux valeurs critiques. Lorsque la distribution supposée est spécifiée, les statistiques *EDF* courantes peuvent être appliquées facilement. Mais lorsque les paramètres sont inconnus et doivent être estimés, les valeurs critiques des statistiques modifiées n'étaient pas disponibles dans la littérature statistique jusqu'aux dernières décennies. Grâce à des simulations, certains auteurs ont fourni des tableaux de valeurs critiques pour les modèles classiques et certaines de leurs généralisations (pour plus de détails, voir Lemeshko et Lemeshko 2011*b*). Dans ce travail, en utilisant la méthode de Monte Carlo et le logiciel matlab, nous proposons des tables de valeurs critiques d'adéquation de D_n, A^2 W^2, L_n et U^2 pour le modèle des risques concurrents de Bertholon lorsque les paramètres sont inconnus.

4.3.1 Test de Komogorov-Smirnov D_n

La statistique D_n du test de Komogorov-Smirnov est le test d'adéquation le plus ancien et le plus utilisé. Elle est définie par

$$D_n = \max\left[D^+; D^-\right]$$

D'où

$$D^+ = \max_{1\leq i\leq n}\left[\left(\frac{i}{n}\right) - F\left(t_{(i)}\right)\right]$$

Et

$$D^- = \max_{1\leq i\leq n}\left[F\left(t_{(i)}\right) - \left(\frac{i-1}{n}\right)\right]$$

Avec ti est la statistique d'ordre. Pour le modèle q-exponentiel, la statistique D_n devient:

$$D^+ = \max_{1\leq i\leq n}\left[\left(\frac{i}{n}\right) + \left[\exp_q\left(-\frac{t_{(i)}}{\eta}\right)\right]^{2-q} - 1\right] \tag{3.11}$$

Et

$$D^- = \max_{1\leq i\leq n}\left[1 - \left[\exp_q\left(-\frac{t_{(i)}}{\eta}\right)\right]^{2-q} - \left(\frac{i-1}{n}\right)\right] \tag{3.12}$$

4.3.2 Test de Cramer-Von Mises W^2

La statistique de Cramer-Von Mises W^2 peut être considérée comme la somme des carrés des différences entre la fonction de distribution empirique et la fonction de distribution cumulative théorique. Nous avons la forme suivante

$$W_n^2 = \sum_{i=1}^{n}\left[F\left(t_{(i)}\right) - \frac{2i-1}{2n}\right]^2 + \frac{1}{12n}$$

Ainsi, pour une distribution q-exponentielle, nous obtenons:

$$W_n^2 = \sum_{i=1}^{n}\left[1 - \left[\exp_q\left(-\frac{t_{(i)}}{\eta}\right)\right]^{2-q} - \frac{2i-1}{2n}\right]^2 + \frac{1}{12n} \tag{3.13}$$

CHAPITRE 5

Simulations

5.1 Calcul des valeurs critiques

Le but de ce travail est de fournir des valeurs critiques d'ajustement des statistiques modifiées de Kolmogorov-Smirnov et Cramer-Von-Mises pour la distribution q-exponentielle lorsque les paramètres sont inconnus et remplacés par leurs estimateurs de maximum de vraisemblance. Pour cela, nous utilisons la méthode de simulation de Monte-Carlo et le logiciel R pour générer 1000 échantillons de différentes tailles n. Selon l'hypothèse nulle H_0, selon laquelle un échantillon appartient au modèle q-exponentielle, nous avons calculé les valeurs des différentes statistiques des tests d'ajustement mentionnées ci-dessus. À cette fin, les étapes suivantes sont utilisées pour calculer les valeurs critiques pour chaque statistique :

1. Générer un échantillon aléatoire $X = \left(x_{(1)}, x_{(2)}, ..., x_{(n)}\right)^T$ de n statistique d'ordre d'une distribution uniforme.

2. Générer un échantillon aléatoire $T = \left(t_{(1)}, t_{(2)}, ..., t_{(n)}\right)$ de n statistique d'ordre de la distribution q-exponentielle, à partir de l'équation non linéaire suivante

$$t = \frac{\eta\left[1 - U^{\left(\frac{1-q}{2-q}\right)}\right]}{1-q} \tag{3.14}$$

3. Les paramètres inconnus des échantillons aléatoires générés sont estimés à l'aide de packages BBsolve.

4. Les estimateurs de paramètres inconnus ont été utilisés pour déterminer la fonction de distribution cumulative hypothétique de la distribution q-exponentielle.

5. Les statistiques de tests mentionnées ci-dessus sont calculées pour chaque généré échantillon aléatoire de tailles différentes.

6. Cette procédure a été répétée 1000 fois de façon indépendante. Par conséquent, nous avons obtenu 1000 valeurs pour chaque statistique proposée. Ces valeurs ont été classées à différents niveaux de signification $0.01, 0.02, 0.05$ sont présentés dans le tableau 1.

$N = 1000$		$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
$n = 15$	D_n	0.9922	0.9926	0.9930
	W_n^2	0.9250	0.9257	0.9265
$n = 30$	D_n	0.9962	0.9965	0.9967
	W_n^2	0.9623	0.9628	0.9632
$n = 50$	D_n	0.9976	0.9977	0.9978
	W_n^2	0.9771	0.9773	0.9775
$n = 80$	D_n	0.9979	0.9980	0.982
	W_n^2	0.9856	0.9858	0.9860
$n = 100$	D_n	0.9983	0.9985	0.9988
	W_n^2	0.9880	0.9885	0.9887

Selon le tableau, nous avons remarqué que:

- Pour chaque test statistique, la puissance augmente de façon monotone avec la taille de l'échantillon et le niveau de signification.

- Le test statistique de Kolmogorov-Smirnov D_n est le plus puissant.

- Le test statistique de Cramer-Von-Mises W_n^2 est le moins puissant.

5.2 Etude de puissance

Dans cette section, nous avons effectué une comparaison de puissance entre les statistiques modifiées de Komlmogorv-Smirnov et Cramer-Von-Mises pour la distribution q-exponentielle avec des paramètres inconnus. À cette fin, nous avons simulé 1000 échantillons aléatoires de différentes tailles à partir de chacune des distributions alternatives :

1. La distribution exponentielle, avec la fonction de densité de probabilité

$$F_\theta(t) = 1 - \exp(-\theta t) \tag{3.15}$$

2. La distribution exponentielle exponentiée, avec fonction de densité de probabilité

$$F_{EE}(t) = (1 - \exp(-\theta t))^\alpha \tag{3.15}$$

Les résultats de puissance des statistiques de tests Kolmogorov-Smirnov et Cramer-Von-Mises pour chaque distribution alternative au niveau de signification $\alpha = 0.05$ sont présentés dans le tableau suivant.

n	$Tests$	Exp	EE
15	KS	0.8219	0.8198
	CVM	0.8139	0.8089
30	KS	0.8501	0.8701
	CVM	0.8397	0.8570
50	KS	0.8965	0.8990
	CVM	0.8795	0.8870

Comme le montrent les résultats:

- Toutes les valeurs de puissance de test pour les différentes statistiques sont supérieures à 80%.
- Quelle que soit la taille de l'échantillon, le test de Kolmogorov-Smirnov D_n a la plus grande puissance.

- Lorsque $n < 30$, la distribution exponentielle est la meilleure distribution alternative.

- Quand $n \geq 30$, la distribution exponentielle exponentitée est la meilleure distribution alternative.

Ainsi, les statistiques de test modifiées de Kolmogorov-Smirnov D_n et de Cramer-Von-Mises W_n^2 fournis dans ce travail et leurs valeurs critiques peuvent détecter la différence entre la distribution q-exponentielle et les différentes distributions alternatives avec haute puissance.

CHAPITRE 6

Conclusion

La distribution q-exponentielle est l'une des nombreuses distributions de type q qui ont été développées pour modéliser une variété de phénomènes, y compris la distribution des temps inter-événement dans des systèmes complexes avec un comportement non gaussien, la distribution de la richesse et du revenu, nous avons fourni des valeurs critiques pour les statistiques modifiées de Kolmogov-Smirnov et Cramer-Von-Mises pour ce modèle lorsque les paramètres sont inconnus, en utilisant le logiciel de programmation R qui joue un rôle important dans l'application des mathématiques dans de nombreux domaines, et qu'il nous permet aussi de résoudre des problèmes complexes plus efficacement, de visualiser leurs données plus efficacement.

Les tableaux présentés dans le présent travail peuvent être utilisés pour vérifier si les données de l'échantillon correspondent à ce modèle qui aide les praticiens à choisir le modèle approprié pour leur analyse.

Bibliographie

[1] A. C. S. Vidal, I. D. Linsa , M. J. C. Mouraa, E. L. Droguett; "Reliability data analysis of systems in the wear-out phase using a q-Exponential likelihood", *Elsevier Ltd*, 0951-8320,2020.

[2] A. J. Lemonte; "A new exponential-type distribution with constant, decreasing, increasing, upside-down bathtub and bathtub-shaped failure rate function", *Comput Stat Data Anal,* 62:149–70, 2013.

[3] B. Y. Lemeshko, S. B. Lemeshko; "Models of statistic distributions of nonparametric goodness-of-fit tests in composite hypotheses testing for double exponential law cases", *Communications in Statistics - Theory and Methods*, 40 (16):2879–92, 2011a. Doi:10.1080/03610926.2011.562770.

[4] C. Chesneau; Cours: "Introduction aux tests statistiques avec R", *Université de Caen*, France, 2016.

[5] C. Tsallis; "Possible generalization of Boltzmann-Gibbs statistics", *J. Stat. Phys.*, vol. 52, no. 1–2, pp. 479–487, 1988.

[6] D. A. Darling; "The Kolmogorov-Smirnov, Cramer-von Mises tests", *The Annals of Mathematical Statistics* 28 (4):823–38, 1957. Doi:10.1214/aoms/1177706788.

[7] E. Paradis; "R pour les débutants". *Institut des Sciences de l'Evolution Université Montpellier II* .F-34095 Montpellier cédex 05, France, 2005.

[8] F. Bertrand, M. M. Bertrand; "Initiation à la statistique avec R". 2émé édition, 2010.

[9] J. Larmarange; "Analyse-R-Introduction à l'analyse d'enquêtes avec R et RStudio", 2019. DOI 10.5281/zenodo.2669067.

[10] J. J. Ruch; Cours: "STATISTIQUE: TESTS D'HYPOTHESES", Bordeaux, 2012 - 2013.

[11] K. K. Jose, S. R. Naik; "On the q-Weibull distribution and its applications" *Commun. Stat. - Theory Methods*, vol. 38, no. 6, pp. 912–926, 2009.

[12] L.C. Malacarne, R. S. Mendes, E. K. Lenzi; "q-exponential distribution in urban agglomeration". *Phys Rev E Stat Nonlinear Soft Matter Phys,* 65(1):1–3, 2002.

[13] L. Shuyan; "Notes de cours Statistique avec le logiciel R", 2013/2014.

[14] M. Xu, E. L. Droguett, I. D. Lins, M. C. Moura; "On the q-Weibull distribution for reliability applications: an adaptative hybrid artificial bee colony algorithm for parameter estimation", *Reliab Eng Syst Saf* , 2017.

[15] P. Dalgaard; "Introductory Statistics with R", *Springer*, 2002.

[16] S. Baillargeon; Cours: "Présentation de R", 2021.

[17] S. Chouia, N. Seddik-Ameur; "Different EDF goodness-of-fit tests for competing risks models", *Communications in Statistics - Simulation and Computation*, 2021. DOI: 10.1080/03610918.2021.1938119.

[18] S. Chouia; Cours: "L'inference statistiques", *Université Badji Mokhtar Annaba*, Algérie.

[19] S. Picoli, R. S. Mendes, and L. C. Malacarne; "q-exponential, Weibull, and q-Weibull distributions: An empirical analysis", *Phys. A Stat. Mech. its Appl.*, vol. 324, no. 3–4, pp. 678–688, 2003.

[20] S. Picoli Jr., R. S. Mendes, L. C. Malacarne, and R. P. B. Santos; "q-distributions in complex systems: a brief review" *Brazilian J. Phys.*, vol. 39, no. 2A, pp. 468–474, 2009.

[21] V. Goulet; "Introduction à la programmation en R", *Université Laval avec la collaboration de Laurent Caron*, cinquième édition, 2016.

[22] https://beginr.u-bordeaux.fr

[23] https://www.statology.org/cramer-von-mises-test-r/

[24] https://www.normalesup.org/~carpenti/Notes/Normalite/normalite.html