وزارة التعليم العالي و البحث العلمي

Université 20 Aout 1955 de Skikda

Faculté des Sciences

Département de Mathématiques

جامعة 20 أوت 1955 ، سكيكدة

كلـــــــية العلـــــوم

قسم الرياضيات

N^0 : U.S/F.S/D.M/……/2022

Faculté des Sciences
Département de Mathématiques

Mémoire

Présenté en vue de l'obtention du diplôme de
Master en Mathématiques

Tests basés sur la fonction de répartition empirique

Option : Commande Optimale et Systèmes Dynamiques

Par :

DAHDOUH Nourhane

Encadré par : TILBI Djahida M.C.B U.SKIKDA

Devant le jury :

Président : LALLOUCHE Abdallah M.C.B U. SKIKDA
Examinatrice : KIMOUCHE Karima M.C.A U. SKIKDA

Année : 2021/2022

Remerciements

A la fin de ce travail, je ne manque d'adresser mes sincères Remerciements à mon ***Dieu*** *le grand créateur qui ma a guidé dans mes pats pour arriver à ce niveau.*

La réalisation de ce travail n'aurait pu être menée à terme sans le support constant de mon encadreur Docteur ***TILBI Djahida****. Je désire lui adresser un merci tout particulier, ses précieux commentaires et ses conseils pertinents m'ont grandement aidé tout au long des différentes étapes inhérentes au processus de recherche et à l'élaboration de ce mémoire.*

Nombreuses sont les personnes qui m'ont aidé à réaliser ce travail, auxquelles je dois avec plaisir, présenter mes remerciements.

Je voudrais également remercier les membres de jury, pour avoir bien voulu lire, commenter et débattre mon travail.

Je remercie toute personne, qui de prés ou de loin ayant généreusement contribuée à l'élaboration de ce travail.

En fin, un grand merci à mon père, ma mère, mes sœurs et mes frères pour leur amour, leur conseils ainsi que leur soutiens inconditionnel, qui m'a permis de réaliser ce mémoire.

DAHDOUH NOURHANE

Dédicace

A l'homme de ma vie, mon exemple éternel, mon soutien moral et source de joie et de bonheur, celui qui s'est toujours sacrifié pour me voir réussir, à toi

Mon chér père ***OUAHID****.*

A la lumière de mes jours, la source de mes efforts, la flamme de mon cœur, ma vie et mon bonheur, maman que j'adore.

Ma chère mère ***LAILA****.*

A celle qui m'a soutenu et a fait mes pas avec moi, et ses encouragements constants .

Ma tante ***SAMIA***

A mes Frères ***ISLEM****,* ***MOUHAMED EL-AMINE,*** *vous êtes deux joyaux précieux, que allah vous protège*

A ma sœur ***RANIA*** *Aucune dédicace ne peut exprimer mon amour de t'avoir comme sœur.*

Je vous aime beaucoup.

A mon ***grand-père****, que Dieu prolonge sa vie, A l'âme de* ***Ma grand-mère****, qui Dieu lui fasse miséricorde*

A mes amies ***BOUCHRA, RANIA*** *et* ***SOUMIA*** *qui m'ont beaucoup aidé durant ces années d'études. Vous êtes pour moi des sœurs et des amies sur qui je peux compter. En témoignage de l'amitié qui nous unit et des souvenirs de tous les moments que nous avons passés ensemble a toute année d'études, et je vous souhaite une vie pleine de santé et de bonheur.*

DAHDOUH NOURHANE

ملخص

تعتمد اختبارات الملائمة المستخدمة بشكل شائع على وظائف التوزيعات التجريبية حيث تتم مقارنة المسافات بين التوزيعات الافتراضية التجريبية والتوزيعات النظرية بالقيم الحرجة . الغرض من هذا العمل هو توفير أحجام عينات مختلفة ، وجداول لقيم التعديل الحرجة لاختبار كولموغوروف سميرنوف D_n، اختبار كرامر- فون ميزس W^2، واختبار أندرسون- دارلينج A^2، واختبار واتسون U^2 ، واختبار لياو شيموكاوا L_n لنموذج رايلي المعمم المهم في الإحصاء والبحوث التشغيلية ، يتم تطبيقه في العديد من المجالات مثل الصحة والزراعة والبيولوجيا والعلوم الأخرى. يتم التحقق من قوة هذه الاختبارات باستخدام بعض البدائل مثل الأسي ، ويبل ، ويبل العكسي ، وتوزيع ويبل الأسي ، والتوزيعات الأسية ـ الأسية EE. يتم إجراء جميع الحسابات باستخدام البرنامج R وطريقة مونت كارلو.

الكلمات المفتاحية: توزيع رايلي المعمم ، اختبار كولموغوروف سميرنوف D_n، اختبار كرامر- فون ميزس W^2، واختبار أندرسون- دارلينج A^2، واختبار واتسون U^2 ، واختبار لياو شيموكاوا L_n.

RÉSUMÉ

Les tests d'ajustement couramment utilisés sont basés sur les fonctions de répartitions empiriques où les distances entre les distributions hypothétiques empiriques et théoriques sont comparées aux valeurs critiques. le but de ce travail est de fournir pour différentes tailles d'échantillons, des tableaux de valeurs critiques d'ajustement du test Kolmogorov-Smirnov $\boldsymbol{D_n}$, test de Cramer-Von Mises $\boldsymbol{W^2}$, test d'Anderson-Darling $\boldsymbol{A^2}$, test de Watson $\boldsymbol{U^2}$ et test de Liao et Shimokawa $\boldsymbol{L_n}$ pour le modéle de Rayleigh généralisé qui est important en statistique et en recherche opérationnelle, il est appliqué dans plusieurs domaines tels que la santé, l'agriculture, la biologie et d'autre sciences. La puissance des ces tests est étudiée à l'aide de certaines alternatives telles que l'exponentielle, Weibull, Weibull inverse, Weibull exponentielle et les distributions exponentielles-exponentielles $\boldsymbol{EE}$. Tous les calculs sont effectués à l'aide du logiciel $\boldsymbol{R}$ et la méthode de Monte Carlo.

Mots clés : Distribution Rayleigh généralisée, test Kolmogorov-Smirnov $\boldsymbol{D_n}$, test de Cramer-Von Mises $\boldsymbol{W^2}$, test d'Anderson-Darling $\boldsymbol{A^2}$, test de Watson $\boldsymbol{U^2}$, test de Liao et Shimokawa $\boldsymbol{L_n}$.

ABSTRACT

Commonly used fit tests are based on empirical distributions functions where the distances between hypothetical empirical and theoretical distributions are compared to critical values.The prupose of this work is to provide for different sample sizes, tables of critical adjustment values of the Kolmogorov-Smirnov test $\boldsymbol{D_n}$ Cramer-Von Mises test $\boldsymbol{W^2}$, Anderson-Darling test $\boldsymbol{A^2}$, Watson test $\boldsymbol{U^2}$, and Liao and Shimokawa test $\boldsymbol{L_n}$ for the generalized Rayleigh model which is important in statistics and operational research, it is applied in several fields such as health, agriculture, biology and other sciences. The power of these tests is investigated using some alternatives such as exponential, Weibull, inverse Weibull, exponential Weibull and exponential-exponential $\boldsymbol{EE}$ distributions. All calculations are performed using the software $\boldsymbol{R}$ and the Monte Carlo method.

Key words : Generalized Rayleigh distribution, Kolmogorov-Smirnov test $\boldsymbol{D_n}$, Cramer-Von Mises test $\boldsymbol{W^2}$, Anderson-Darling test $\boldsymbol{A^2}$, Watson test, Liao and Shimokawa test $\boldsymbol{L_n}$.

TABLE DES MATIÈRES

LISTE DES ABRÉVIATIONS

Symbole	Description
KS	Kolmogorov-Smirnov
LS	Liao et Shimokawa
W	Watson
FDC	fonction de distribution cumulée
AD	Anderson-Darling
FDE	fonction de distribution empirique
C-vM	Cramer-von-Mises
RG	Rayleigh généralisée
v.a	variable aléatoire
i.e	in essence
TLC	Théorème de la limite centrale
EMV	Estimateur du maximum de vraisemblance

INTRODUCTION GÉNÉRALE

Les résultats de toute analyse statistique dépendent du modèle utilisé, le choix de celui-ci est donc très important. Pour déterminer si un échantillon pourrait provenir d'une distribution spécifique, les chercheurs ont développé plusieurs statistiques de tests. La plupart d'entre elles sont basées sur les fonctions de distribution empiriques (FDE), la plus ancienne étant la statistique de Kolmogorov-Smirnov $\boldsymbol{D_n}$ (Kolmogorov 1933). Par contre, Le test de Cramér-von Mises $\boldsymbol{W^2}$ peut être vu comme une version plus puissante du test de Kolmogorov-Smirnov. La statistique d'Anderson-Darling $\boldsymbol{A^2}$ (Anderson et Darling 1954) peut être considérée comme une distribution limite de $\boldsymbol{W^2}$ et elle donne plus de poids aux queues que la statistique $\boldsymbol{D_n}$ (voir Darling 1957). Watson (1961a, 1962b) a proposé une nouvelle statistique de test $\boldsymbol{U^2}$ comme généralisation de la statistique de test de Cramer-Von Mises $\boldsymbol{W^2}$.

Les statistiques FDE sont diffusées gratuitement mais dans le cas des paramètres inconnus, leur distribution dépendra non seulement de la taille de l'échantillon mais aussi sur la distribution hypothétique, les paramètres estimés et la méthode d'estimation des paramètres (voir Lawless 1982). De nombreux auteurs ont reconnu ce fait et étendu l'utilisation de statistiques de tests différents dans ce cas. En utilisant des méthodes numériques, ils ont développé un test modifié statistique où les paramètres inconnus sont remplacés par leurs estimateurs.

Grâce à des simulations informatiques, des tableaux détaillés de valeurs critiques pour les tests statistiques $\boldsymbol{A^2}$ sont disponibles pour les distributions normale, uniforme, log-normale, exponentielle de Weibull, de valeur extrême du type I, de Pareto généralisée et logistique (voir Stephens 1970, 1974, 1977, 1979), pour les distributions de Weibull à deux et trois paramètres (Evans, Johnson et Green 1989), pour la distribution exponentielle généralisée (Hassan 2005), pour la distribution généralisée de Frechet (Abd-Elfattah, Fergany, et Omima 2010), pour la distribution gamma exponentielle (Shawky et Bakoban, 2009), pour la double distribution exponentielle (Lemeshko et Lemeshko 2011a), pour la distribution Topp-Leone (Al-Zahrani 2012). Plus tard, Liao et Shimokawa (1999a,

1999b) ont développé et appliqué une nouvelle statistique $\boldsymbol{L_n}$ pour tester la qualité de l'ajustement de la valeur extrême de type I et de Weibull à **2** paramètres distributions avec des paramètres estimés. Plusieurs études de puissance ont mentionné que les essais d'FDE sont plus puissants que d'autres tests d'ajustement pour une large gamme de taille d'échantillons et en particulier que la statistique de test d'Anderson-Darling $\boldsymbol{A^2}$ est le test FDE le plus puissant dans une variété de situations.

La distribution généralisée de Rayleigh a été introduite par Surles et Padgett (2001). A l'origine, Mudholkar et Srivastava (1993), Mudholkar et al. (1995) ont proposé plusieurs distributions appelées les distributions Burr, dont la distribution de Rayleigh généralisée (RG) est un cas particulier de celles de Burr Type $\boldsymbol{X}$. Selon les valeurs des paramètres, Kundu et Raqab (2005) ont montré que la densité de probabilité de cette distribution a différentes formes lui permettant de décrire beaucoup plus de données réelles.

Ce mémoire est organisé comme suit :

Le premier chapitre est consacré aux rappels sur les notions fondamentales concernant les variables aléatoires, les tests d'hypothèse, l'estimation et la distribution de Rayleigh généralisée(RG).

Dans **le deuxième chapitre** nous construisons les tests d'ajustement des statistiques modifiés $\boldsymbol{D_n}$, $\boldsymbol{A_n^2}$, $\boldsymbol{W_n^2}$, $\boldsymbol{L_n}$ et $\boldsymbol{U_n^2}$ basés sur la FDE pour le modèle de Rayleigh généralisé.

Dans le troisième chapitre, nous utilisons le logiciel R et les méthodes de Monte-Carlo pour fournir des tableaux de valeurs critiques des statistiques modifiées pour ce modèle.

Enfin, la puissance de ces statistiques est étudiée à l'aide de distributions alternatives (Weibull et exponentielle...) est présentée dans **le quatrième chapitre**.

CHAPITRE 1

PRÉLIMINAIRE

1.1 Loi Normale

Soient deux réels $\boldsymbol{m}$ et $\boldsymbol{\sigma}$. On suppose $\boldsymbol{\sigma > 0}$ on dit que la variable aléatoire réelle continue $\boldsymbol{X}$ suit la loi normale de paramètres $\boldsymbol{m}$ et $\boldsymbol{\sigma}$ lors qu'elle admet pour densité de probabilité la fonction :

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

La loi normale de paramètre $\boldsymbol{m}$ et $\boldsymbol{\sigma}$ est notée $\boldsymbol{N(m, \sigma)}$ telle que $\boldsymbol{m}$ est la moyenne et d'écart-type $\boldsymbol{\sigma}$.

Cas Particulier :

Si $\boldsymbol{m = 0}$ et $\boldsymbol{\sigma = 1}$, la loi normale est dite centrée réduite

$$\boldsymbol{X \to N(0, 1)}$$

On dit que la variable aléatoire réelle continue $\boldsymbol{X}$ suit la loi normale centrée réduite si elle admet pour densité de probabilité de fonction :

$$\boldsymbol{f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R}.}$$

1.2 Loi uniforme

Soit un univers des possibles $\boldsymbol{\Omega}$, intervalle de $\mathbb{R}$, et $\boldsymbol{a}$ et $\boldsymbol{b}$ deux réels tels que $\boldsymbol{a < b}$.

On appelle loi uniforme sur $\boldsymbol{[a, b]}$ la loi de probabilité dont la densité $\boldsymbol{f}$ est la fonction constante définié par :

$$\boldsymbol{f(x) = \begin{cases} 0, & si \quad x > a \\ \frac{1}{b-a}, & si \quad a \leq x \leq b \\ 0, & si \quad x > b \end{cases}}$$

La fonction de répartition $\boldsymbol{F}$ est définié par :

$$\boldsymbol{F(x)} = \begin{cases} \boldsymbol{0}, & \boldsymbol{si} \quad \boldsymbol{x > a} \\ \boldsymbol{\frac{x-a}{b-a}}, & \boldsymbol{si} \quad \boldsymbol{a \leq x \leq b} \\ \boldsymbol{0}, & \boldsymbol{si} \quad \boldsymbol{x > b} \end{cases}$$

1.3 Fonction de répartition empirique

Soit $\boldsymbol{X_1, X_2, ...X_n, ...}$ une suite de variables aléatoires réelles i.i.d. de fonction de répartition $\boldsymbol{F}$. On rappelle que, pour tout $\boldsymbol{x} \in \mathbb{R}$:

$$\boldsymbol{F(x)} = \mathbb{P}(\boldsymbol{X_i \leq x}).$$

On rappelle fonction de répartition empirique associée au $\boldsymbol{n}$ échantillon $\boldsymbol{X_1, X_2, ...X_n}$ la fonction :

$$\boldsymbol{F_n(x)} = \boldsymbol{\frac{1}{n} \sum_{i=0}^{n} 1_{X_i \leq x}} \ .$$

Propriétés :

•$\boldsymbol{F_n}$ est croissante, continue à droite, $\boldsymbol{\lim_{x \to -\infty} F_n(x) = 0}$ et $\boldsymbol{\lim_{x \to +\infty} F_n(x) = 1}$.

•$\boldsymbol{nF_n(x)}$ suit une loi binomiale de paramètre $\boldsymbol{(n, F(x))}$.

•$\boldsymbol{E(F_n(x)) = F(x)}$, pour tout $\boldsymbol{x}$, $\boldsymbol{F_n(x)}$ est un estimateur sans biais de $\boldsymbol{F(x)}$.

•$\boldsymbol{Var(nF_n(x)) = nF(x)(1 - F(x))}$

$$\boldsymbol{Var(F_n(x)) = \frac{F(x)(1-F(x))}{n} \to 0}$$

•Par l'inégalité de Tchebichev,

$$\boldsymbol{\forall \varepsilon > 0, \mathbb{P}(|F_n(x) - F(x)| \geq \varepsilon) \leq \frac{1}{\varepsilon^2} Var(F_n(x)) \to 0,}$$
$$\boldsymbol{|F_n(x) - F(x)| \overset{prob}{\to} 0}$$

•On déduit du TLC que pour tout $\boldsymbol{x}$ tel que $\boldsymbol{F(x)(1 - F(x)) \neq 0}$

$$\boldsymbol{\sqrt{n}(F_n(x) - F(x)) \overset{loi}{\to} \mathcal{N}(0, F(x)(1 - F(x)))}.$$

1.4 Echantillonnage

Un échantillon statistique est constitué d'un nombre limité d'individus tirés au sort dans la population étudiée.

C'est le tirage au sort qui assure la représentativité.

Un échantillon de taille $\boldsymbol{n}$ d'une variable aléatoire $\boldsymbol{X}$ est obtenu en répétant $\boldsymbol{n}$ fois l'épreuve qui donne $\boldsymbol{X}$.

Notation : $(\boldsymbol{X_1, X_2, ..., X_n})$

Une réalisation particulière : $(\boldsymbol{x_1, x_2, ..., x_n})$

Définition

Considérons une population d'où l'on extrait un échantillon d'effectif $\boldsymbol{n}$ dont les éléments sont $\boldsymbol{x_i}$. La statistique descriptive associe à cet échantillon une valeur centrale, la moyenne empirique

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

et une valeur de dispersion, la variance empirique

$$\boldsymbol{s^2} = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{(x_i - \overline{X})^2}.$$

La loi de probabilité associée à cette population(en toute rigueur inconnue)possède une moyenne $\boldsymbol{\mu}$ et une variance $\boldsymbol{\sigma^2}$ définie en probabilité dont les valeurs empiriques correspondantes donnent une idée. Le problème est que, si on avait choisi un autre échantillon, on aurait trouvé des valeurs différentes.

Ceci conduit à considérer la moyenne empirique et la variance empirique comme des variables aléatoires possédant une loi de probabilité, une moyenne et une variance. On ne peut continuer le raisonnement qu'en supposant que les variables qui constituent l'échantillon sont indépendantes.

Sous cette condition, on peut calculer la moyenne (ou espérance) et la variance de la moyenne empirique et de la variance empirique. On obtient :

$$\boldsymbol{E(\overline{X}) = \mu}, \qquad \boldsymbol{V(\overline{X}) = \frac{\sigma^2}{n}},$$

$$\boldsymbol{E(s^2) = \frac{n-1}{n}\sigma^2}, \qquad \boldsymbol{V(s^2) = \frac{1}{n}(E(x^4) - \frac{n-1}{n-3}\sigma^4)}.$$

L'écart-type de la moyenne empirique vaut $\frac{\sigma}{\sqrt{n}}$. Si $\boldsymbol{n}$ devient grand d'après le théorème de la limite centrale la moyenne suit une loi normale caractérisée par la moyenne $\boldsymbol{\mu}$ et cet écart-type. Ce résultat reste valable quelle que soit la taille de l'échantillon lorsque la loi de probabilité assignée à la population est normale. Dans ce dernier cas, particulièrement important en pratique, on montre également que $\frac{ns^2}{\sigma^2}$ suit une loi de $\boldsymbol{\chi^2}$.

Preuve

Les variables aléatoires $\overline{\boldsymbol{X}}$ et $\boldsymbol{s^2}$ sont indépendantes, $\overline{\boldsymbol{X}}$ suit une loi de probabilité $\boldsymbol{N(\mu, \frac{\sigma^2}{n})}$ et $\frac{ns^2}{\sigma^2}$ est la somme de $\boldsymbol{(n-1)}$ carrés de variables aléatoires indépendantes de la loi $\boldsymbol{N(0,1)}$.
On a :

$$\boldsymbol{s^2} = \frac{\boldsymbol{z}}{\boldsymbol{n}}$$

avec :

$$\boldsymbol{z = \sum_{i=1}^{n}[x_i - \overline{X}]^2 = \sum_{i=1}^{n}[x_i^2 - 2x_i\overline{X} + \overline{X}]^2 = \sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i \frac{1}{n}\sum_{i=1}^{n} x_i + n(\frac{1}{n}\sum_{i=1}^{n} x_i)^2,}$$

$$\boldsymbol{= \sum_{i=1}^{n} x_i^2 - \frac{1}{n}[\sum_{i=1}^{n} x_i]^2.}$$

On pose

$$\boldsymbol{w_i = x_i^2} \text{ et } \boldsymbol{w_1 = \frac{\sum_{i=1}^{n} x_i}{\sqrt{n}}}, \text{ donc } \boldsymbol{z = \sum_{i=1}^{n} w_i^2 - w_1^2 = \sum_{i=1}^{n} w_i^2},$$

alors

$$\boldsymbol{ns^2\sigma^2 = \frac{\sum_{i=2}^{n} w_i^2}{\sigma^2} = \sum_{i=2}^{n}(\frac{w_i}{\sigma^2})^2.}$$

Et comme

$$\boldsymbol{s^2 \rightsquigarrow N(0, \sigma^2)}\text{,alors } \boldsymbol{\frac{w_i}{\sigma} \rightsquigarrow N(0,1)},$$

donc la variable aléatoire $\boldsymbol{\frac{ns^2}{\sigma^2}}$ admet une loi de probabilité de $\boldsymbol{\chi^2}$ à $\boldsymbol{n-1}$ degrés de liberté.

1.4.1 Le théorème de la limite centrale

Soit $\boldsymbol{X_1, X_2...}$ un ensemble de variables aléatoires définies sur le même espace de probabilité, suivant la même loi $\boldsymbol{D}$ et indépendantes. Supposons que l'espérance $\boldsymbol{\mu}$ et l'écart-type $\boldsymbol{\sigma}$ de $\boldsymbol{D}$ existent et soient finis $\boldsymbol{(\sigma \neq 0)}$.

Considérons la somme $\boldsymbol{S_n = X_1 + ... + X_n}$. Alors l'espérance de $\boldsymbol{S_n}$ est $\boldsymbol{n\mu}$ et son écart-type vaut $\boldsymbol{\sigma\sqrt{n}}$. De plus, pour parler de manière informelle, la loi de $\boldsymbol{S_n}$ tend vers la loi normale $\boldsymbol{N(n\mu, \sigma^2 n)}$ quand $\boldsymbol{n}$ tend vers l'infini.

Afin de clarifier cette idée de convergence, nous allons poser

$$\boldsymbol{Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}},$$

alors la loi de $\boldsymbol{Z_n}$ converge vers la loi normale centrée réduite $\boldsymbol{N(0,1)}$ lorsque n tend vers l'infini (il s'agit de la convergence en loi). Cela signifie que si $\boldsymbol{\phi}$ est la fonction de répartition de $\boldsymbol{N(0,1)}$, alors pour tout réel $\boldsymbol{z}$:

$$\boldsymbol{\lim_{n\to\infty} P(Z_n \leq z) = \phi(z)}$$

ou, de façon équivalente :

$$\boldsymbol{\lim_{n\to\infty} P\left(\frac{\overline{X}_n - \mu}{\sigma\sqrt{n}} \leq z\right) = \phi(z)}$$

où

$$\overline{\boldsymbol{X}}_{\boldsymbol{n}} = \frac{\boldsymbol{S_n}}{\boldsymbol{n}} = \frac{\boldsymbol{X_1 + ... + X_n}}{\boldsymbol{n}}.$$

Application à la statistique mathématique

Ce théorème de probabilités possède une interprétation en statistiques mathématiques. Cette dernière associe une loi de probabilité à une population. Chaque élément extrait de la population est donc considéré comme une variable aléatoire et, en réunissant un nombre $\boldsymbol{n}$ de ces variables supposées indépendantes, on obtient un échantillon. La somme de ces variables aléatoires divisée par $\boldsymbol{n}$ donne une nouvelle variable nommée la moyenne empirique. Celle-ci, une fois réduite, tend vers une variable normale réduite lorsque $\boldsymbol{n}$ tend vers l'infini.

1.5 Estimation

Tout d'abord, une estimation, qu'elle soit statistique ou pas, concerne une information inconnue. Toute estimation a donc un caractère incertain.

Il faut bien noter que la valeur à estimer existe, même si elle est inconnue, et c'est pourquoi une estimation ne doit pas être confondue avec une prévision d'une valeur inconnue parce que future.

En statistiques mathématiques on cherche à estimer la valeur inconnue d'une statistique d'une population.

On utilise une statistique d'un échantillon prélevé au hasard ou par d'autres méthodes aléatoires.

On peut donner une estimation : ponctuelle et par l'intervalle de confiance.

La qualité de l'estimation dépend :

1-De la méthode de prélèvement de l'échantillon : quand l'échantillon est prélevé au hasard ou par une méthode aléatoire l'estimation sera aléatoire, c'est-à-dire que les différentes valeurs possibles soient assorties de probabilités.

2-De l'information statistique : disponible dans la population d'où est extrait l'échantillon.

Définition :

Soient $\boldsymbol{x_1, ..., x_i, ..., x_n}$ les $\boldsymbol{n}$ valeurs prises par la variable aléatoire $\boldsymbol{X}$ dans un échantillon de taille $\boldsymbol{n}$ prélevé dans la population-mère.

On appelle estimateur de $\boldsymbol{\theta}$, et l'on note $\boldsymbol{T_n}$, la fonction qui aux valeurs $\boldsymbol{x_i}$ de l'échantillon fait correspondre la valeur du paramètre $\boldsymbol{\theta}$. On note la valeur numérique de cette estimation par :

$$\boldsymbol{\theta = T_n(x_1, ..., x_n)}.$$

Par définition, $\boldsymbol{T_n}$ est une fonction des réalisations d'une variable aléatoire.

Exemple :

On observe un phénomène de production de pièces manufacturées. Chaque pièce est associée à une mesure (un indicateur de qualité par exemple). Comme on ne peut pas vérifier chaque mesure, on procède à un échantillonnage qui nous fournit donc un échantillon. Supposons que la connaissance de la nature de cet inclicateur nous permet de faire l'hypothèse qu'il obéit à une loi de probabilité normale. Le problème est maintenant, au vue de l'échantillon $\{x_i\}$, de proposer une valeur pour la moyenne de cette loi normale. Il faut procéder à une estimation du paramètre vrai μ qui se traduit par la valeur $\hat{\mu}$. Il y a une infinité de manières possibles parmi lesquelles on peut citer

- $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- $\hat{\mu} = médiane\{x_i\}$
- $\hat{\mu} = mode\{x_i\}$

Quel est le meilleur estimateur de la moyenne ? Existe-t-il ?

La réponse est simple, il n'en existe pas. Alors comment comparer les estimateurs. Pour cela, on se sert de plusieurs critères, le plus souvent liés au bon sens :

Le biais : Un estimateur T_n d'un paramètre θ est dit sans biais, si son l'espérance mathématique est la vrai valeur du paramètre θ

$$E(T_n) = \theta \text{ ou } E(\hat{\theta}_n) = \theta$$

On appelle biais d'un estimateur T_n de θ, l'écart :

$$\text{biais } (T_n) = E(T_n) - \theta$$

Un estimateur T_n de θ est asymptotiquement sans biais $\Longleftrightarrow$

$$\lim_{n\to\infty} E(T_n) = \theta$$

La convergence : Un estimateur T_n d'un paramètre θ est dit convergent s'il tend en probabilité vers θ

$$T_n \longrightarrow \theta \Longleftrightarrow \lim_{n\to\infty} P(|T_n - \theta| > \varepsilon) = 0$$

Théorème :

Un estimateur T_n d'un paramètre θ dont l'espérance mathématique tant vers θ et la variance tend vers θ lorsque $n \longrightarrow \infty$, est un estimateur convergent

$$E(T_n) = \theta \text{ et } V(T_n) = 0 \text{ pour } n \longrightarrow \infty \text{ alors } T_n \longrightarrow \theta$$

On en déduit l'inégalité de **Bienaymé-Tchebychev :**

$$\boldsymbol{P(|T_n - \theta| > \varepsilon) \leq \frac{V(T_n)}{\varepsilon^2}}$$

ou

$$\boldsymbol{P(|T_n - \theta| > \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}}$$

Définition un estimateur efficace :

La variance d'un estimateur représente sa précision. Pour tous les estimateurs (ayant même moyenne), il est possible de trouver celui dont la précision sera la meilleure, i.e. dont la variance sera la plus faible. On parle alors d'estimateur à variance minimum.

Lorsque l'on compare deux estimateurs, on ne dira également que $\boldsymbol{T_n}$ est plus efficace que $\boldsymbol{T_n^*}$ si $\boldsymbol{V(T_n) < V(T_n^*)}$

Remarque : Si deux estimateurs sans biais et convergents du paramètre $\boldsymbol{\theta}$, alors le plus précis donc le meilleur est celui qui possède la variance la plus faible.

1.5.1 Estimation ponctuelle :

C'est la valeur observée dans l'échantillon, ou une valeur unique calculée à partir de la valeur observée dans l'échantillon, dans le voisinage de laquelle la valeur inconnue de la statistique qu'on cherche à estimer une très grande probabilité de se trouver .

Estimation d'une moyenne :

Soit donc une variable aléatoire $\boldsymbol{X}$ suivant une loi normale de moyenne $\boldsymbol{\mu}$ inconnue et d'écart-type $\boldsymbol{\sigma}$. On dispose d'un échantillon de $\boldsymbol{n}$ réalisations de cette variable aléatoire la moyenne empirique

$$\boldsymbol{\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i} \text{ est un estimateur de } \boldsymbol{\mu}$$

Estimation d'une variance :

Nous n'aborderons que le cas de l'estimation de la variance $\boldsymbol{\sigma^2}$ d'une v.a. $\boldsymbol{X}$ normale de moyenne $\boldsymbol{\mu}$ à partir d'un échantillon de $\boldsymbol{n}$ valeurs.

*Si $\boldsymbol{\mu}$ est connue, la variance de la v.a. $\boldsymbol{X}$ est

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

*Si $\boldsymbol{\mu}$ est inconnue, la variance de la v.a. $\boldsymbol{X}$ est

$$\boldsymbol{S_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{n}{n-1} S_n^2}.$$

Estimation d'une proportion :

On veut estimer la proportion $\boldsymbol{p}$ d'individus ayant ou non un caractère donné $\boldsymbol{A}$ dans une population de taille $\boldsymbol{n}$. Le nombre d'individus ayant un caractère donné $\boldsymbol{A}$ est $\boldsymbol{np}$, on tire un $\boldsymbol{n}$ -échantillon $\boldsymbol{X_1, X_2, ..., X_n}$ de cette population.

On considère

$$\boldsymbol{F_n = \frac{1}{n} \sum_{i=1}^{n} X_i},$$

$\boldsymbol{F_n}$ est un estimateur naturel de $\boldsymbol{p}$ (fréquence de l'échantillon).

$$\boldsymbol{F_n \rightsquigarrow N(p, \frac{pq}{n})} \text{ ou } \boldsymbol{q = 1 - p}$$

1.5.2 Estimation par intervalle de confiance :

Cette nouvelle approche est souvent préférée dans la pratique car elle introduit la notion d'incertitude. On cherche à déterminer l'intervalle $\boldsymbol{[a, b]}$ centré sur la valeur numérique estimée du paramètre inconnu $\boldsymbol{\theta}$ contenant la valeur vraie avec une probabilité $\boldsymbol{\alpha}$ fixée a priori. Cette probabilité permet de s'adapter aux exigences de l'application.

$$\boldsymbol{P[a < \theta < b] = 1 - \alpha}.$$

L'intervalle $\boldsymbol{[a, b]}$ est appelé intervalle de confiance et $\boldsymbol{1 - \alpha}$ est le coefficient de confiance $\boldsymbol{(\alpha \in]0, 1[)}$. Une estimation par intervalle de confiance sera d'autant meilleure que l'intervalle sera petit pour un coefficient de confiance grand.

La donnée de départ, outre l'échantillon, sera la connaissance de la loi de probabilité du paramètre à estimer. Comme il n'existe pas de résolution générale de ce problème, nous allons aborder successivement les cas les plus fréquents(estimation d'une proportion, d'une moyenne, d'une variance de loi normale).

a) Estimation d'une proportion :

Soit une population dont les individus possèdent un caractère $\boldsymbol{A}$ avec une probabilité $\boldsymbol{p}$. On cherche à déterminer cette probabilité inconnue en prélevant un échantillon de taille $\boldsymbol{n}$ dans cette population. On constate que $\boldsymbol{x}$ parmi les $\boldsymbol{n}$ individus possèdent le caractère $\boldsymbol{A}$. Que peut-on en déduire ?, la proportion $\boldsymbol{f_n = \frac{x}{n}}$ approxime la valeur vraie $\boldsymbol{p}$, mais avec quelle confiance ?

Soit $\boldsymbol{F_n = \frac{x}{n}}$;$\boldsymbol{F_n}$ est une v.a. construite par la somme de $\boldsymbol{n}$ variables aléatoires et de même paramètre $\boldsymbol{p}$. C'est donc, d'après le théorème central limite, une v.a. dont la loi de probabilité tend vers une loi normale de moyenne $\boldsymbol{p}$ et d'écart-type $\boldsymbol{\sqrt{\frac{p(1-p)}{n}}}$. Cette approximation est valable uniquement si la taille de l'échantillon est suffisamment grande (i.e. $\boldsymbol{n > 30}$ en pratique).

Construisons l'intervalle de confiance autour $\boldsymbol{p}$:

$$\boldsymbol{P(|f_n - p| < t) = 1 - \alpha},$$

où $\boldsymbol{\alpha}$ est le risque (a priori, on construit un intervalle symétrique). $\boldsymbol{f_n}$ est une réalisation d'une v.a. $\boldsymbol{N(p, \sqrt{\frac{p(1-p)}{n}})}$. Donc on peut par normalisation et centrage obtenir une nouvelle v.a.$\boldsymbol{U}$

$$\boldsymbol{u_\alpha = \frac{f_n - p}{\sqrt{\frac{p(1-p)}{n}}} \rightsquigarrow N(0,1)}.$$

On en déduit donc l'intervalle de confiance sous la forme :

$$\boldsymbol{P[a < \theta < b] = P\left[f_n - u_\alpha\sqrt{\frac{p(1-p)}{n}} < p < f_n + u_\alpha\sqrt{\frac{p(1-p)}{n}}\right] = 1 - \alpha}.$$

La valeur $\boldsymbol{t = \sqrt{\frac{p(1-p)}{n}}}$ est donc un résultat de calcul. La valeur de $\boldsymbol{u_\alpha}$ sera lue sur une table de loi normale $\boldsymbol{N(0,1)}$.

Il existe par ailleurs différentes manières pour approximer la valeur de $\boldsymbol{p}$:

• Soit par la proportion $\boldsymbol{f_n}$:

$$\boldsymbol{P[a < \theta < b] = P\left[f_n - u_\alpha\sqrt{\frac{f_n(1-f_n)}{n}} < p < f_n + u_\alpha\sqrt{\frac{f_n(1-f_n)}{n}}\right] = 1 - \alpha}.$$

• Soit par majoration : en effet, quelle que soit la valeur de $\boldsymbol{p}$, le produit $\boldsymbol{p(1-p)}$ est majore par $\frac{1}{4}$,

$$\boldsymbol{P[a < \theta < b] = P\left[f_n - \frac{u_\alpha}{2\sqrt{n}} < p < f_n - \frac{u_\alpha}{2\sqrt{n}}\right] \geq 1 - \alpha}.$$

b) Estimation d'une moyenne :

• Cas des grands échantillons $(\boldsymbol{n > 30})$

Théorème : Quelle que soit la loi de $\boldsymbol{X}$, $\boldsymbol{U = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n-1}}}}$ suit sensiblement la loi normale centrée réduite.

Utilisation : La valeur de $\boldsymbol{u_\alpha}$ sera lue sur une table de loi normale $\boldsymbol{N(0,1)}$ tel que

$$\boldsymbol{P[-u_\alpha < U < u_\alpha] = 1 - \alpha.}$$

On en déduit comme intervalle de confiance de la moyenne $\boldsymbol{\mu}$ de la population, au risque $\boldsymbol{\alpha}$:

$$\boldsymbol{\left]\bar{x} - u_\alpha\frac{s}{\sqrt{n-1}}; \bar{x} + u_\alpha\frac{s}{\sqrt{n-1}}\right[}.$$

•Cas des petits échantillons :

Théorème : Si $\boldsymbol{X}$ suit une loi normale, $\boldsymbol{\overline{X}}$ suit une loi normale $\boldsymbol{N(\mu, \frac{\sigma}{\sqrt{n}})}$. Si $\boldsymbol{\sigma}$ est connu, l'intervalle de confiance de $\boldsymbol{\mu}$ se construit alors comme dans le cas des grands échantillons.

Mais en général $\boldsymbol{\sigma}$ est inconnu et estimé par $\boldsymbol{s}$. Dans le cas des petits échantillons, en remplaçant $\boldsymbol{\sigma}$ par $\boldsymbol{s}$ on modifie la nature de la loi suivie par $\overline{\boldsymbol{X}}$.

Théorème : La variable aléatoire $\boldsymbol{T = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}$ suit la loi de Student à $\boldsymbol{n-1}$ degrés de liberté.

Utilisation : Une loi de Student est une loi de probabilité continue dont la densité est une fonction paire et qui dépend d'un paramètre appelé nombre de degrés de liberté (d.d.l). Le coefficient de risque $\boldsymbol{\alpha}$ étant choisi et le nombre de degrés de liberté étant connu. On en déduit comme intervalle de confiance de la moyenne $\boldsymbol{\mu}$ de la population :

$$\left]\boldsymbol{\bar{x} - t_\alpha \frac{s}{\sqrt{n-1}}; \bar{x} + t_\alpha \frac{s}{\sqrt{n-1}}}\right[.$$

c) Estimation d'une variance :

Théorème : Si $\boldsymbol{X}$ suit une loi normale, la variable aléatoire$\boldsymbol{Y^2 = \frac{n-1}{\sigma^2}s^2}$ suit la loi du $\boldsymbol{\chi^2}$ à $\boldsymbol{n-1=\nu}$ degrés de liberté.

Utilisation :

• Si $\boldsymbol{n \leq 31}$

Le coefficient de risque $\boldsymbol{\alpha}$ étant choisi et le nombre de degré de liberté étant connu, tels que

$$\boldsymbol{P\left]a < Y^2 < b\right[= 1-\alpha},$$

avec $\rightsquigarrow \boldsymbol{P[Y^2 \geq b] = \frac{\alpha}{2}}$ et $\boldsymbol{P[Y^2 \geq a] = 1-\frac{\alpha}{2}}$.

La seule valeur connue de $\boldsymbol{S^2}$ étant $\boldsymbol{s^2}$, on obtient comme intervalle de confiance de la $\boldsymbol{\sigma^2}$ au risque $\boldsymbol{\alpha}$:

$$\left]\boldsymbol{\frac{n-1}{b}s^2; \frac{n-1}{a}s^2}\right[.$$

• Si $\boldsymbol{n > 31}$

Le théorème cité est vrai quelque soit $\boldsymbol{n}$. Mais les tables du $\boldsymbol{\chi^2}$ s'arrêtent habituellement au degrés de liberté $\boldsymbol{\nu = 30}$. On ne peut pas donc les utiliser si $\boldsymbol{n > 31}$.

Théorème : Si $\boldsymbol{Y^2}$ est une variable aléatoire qui suit une loi du $\boldsymbol{\chi^2}$ à $\boldsymbol{\nu}$ degrés de liberté et si $\boldsymbol{\nu = 30}$, alors la variable aléatoire $\boldsymbol{U = \sqrt{2Y^2} - \sqrt{2\nu - 1}}$ suit sensiblement la loi réduite de Gauss.

Utilisation :

Ici on a : $\boldsymbol{U = \sqrt{\frac{2(n-1)}{\sigma^2}S^2} - \sqrt{2n-3}}$. Après avoir choisi $\boldsymbol{\alpha}$ on détermine $\boldsymbol{u_\alpha}$ tel que $\boldsymbol{P[-u_\alpha < U < u_\alpha] = 1 - \alpha}$, et on en déduit l'intervalle de confiance de la $\boldsymbol{\sigma^2}$:

$$\left]\frac{2(n-1)s^2}{\sqrt{2n-3}+u_\alpha}; \frac{2(n-1)s^2}{\sqrt{2n-3}-u_\alpha}\right[.$$

1.5.3 Estimation du maximum de vraisemblance

Soit $\boldsymbol{X}$ une variable aléatoire de densité de probabilité $\boldsymbol{f(x,\theta)}$ connue analytiquement mais dont l'un des paramètres $\boldsymbol{\theta}$ est inconnu (numériquement). Le problème consiste donc à construire une expression analytique fonction des réalisations de cette variable dans un échantillon de taille $\boldsymbol{n}$, permettant de trouver la valeur numérique la plus vraisemblable pour le paramètre $\boldsymbol{\theta}$.

Si $\boldsymbol{\{x_1, ..., x_n\}}$ sont des réalisations indépendantes de la variable aléatoire, on peut dire que $\vec{\boldsymbol{x}} = \begin{pmatrix} \boldsymbol{x_1} \\ \vdots \\ \boldsymbol{x_2} \end{pmatrix}$, est une réalisation d'un vecteur aléatoire $\vec{\boldsymbol{X}} = \begin{pmatrix} \boldsymbol{X_1} \\ \vdots \\ \boldsymbol{X_2} \end{pmatrix}$.
Dont les composantes $\boldsymbol{X_i}$ sont indépendantes deux à deux.

L'approche retenue consiste à chercher la valeur de $\boldsymbol{\theta}$ qui rend le plus probable les réalisations que l'on vient d'obtenir. La probabilité d'apparition a priori de l'échantillon en question peut alors être caractérisée par le produit des probabilités d'apparition de chacune des réalisations (puisque celles-ci sont supposées indépendantes deux à deux)

$$\boldsymbol{P(\vec{X} = \vec{x}) = \prod_{i=1}^{n} f(x_i, \theta)}$$

La méthode du maximum de vraisemblance consiste à rechercher la valeur de $\boldsymbol{\theta}$ qui rend cette probabilité maximale. Comme nous l'avons vu plus haut, le produit des valeurs $\boldsymbol{f(x_i,\theta)}$ est aussi noté $\boldsymbol{L(x_1, ..., x_n, \theta)}$ et appelé fonction de vraisemblance. La valeur $\hat{\boldsymbol{\theta}}$ qui rend maximum la fonction de vraisemblance $\boldsymbol{L}$ est donc la solution de :

$$\boldsymbol{\frac{\partial lnL}{\partial \theta} = 0 \Longrightarrow \hat{\theta} : \frac{\partial^2 lnL}{\partial \theta^2} < 0.}$$

L'emploi du logarithme sur la fonction permet de passer de la maximisation d'un produit à celle d'une somme, le résultat restant le même car la fonction logarithme est monotone strictement croissante.

1.6 La distribution de Rayleigh généralisée

la distribution de Rayleigh généralisée à deux paramètres est un membre particulier de la distribution de Weibull généralisée , proposée à l'origine par Mudholkar et Srivastava (1993), dans ce travail, nous préférons également appeler la distribution de type $\boldsymbol{X}$ de Burr à deux paramètres

comme la distribution de Rayleigh généralisée (RG). Pour $\boldsymbol{\alpha > 0}$ et $\boldsymbol{\lambda > 0}$, la distribution RG à deux paramètres a la fonction de distribution :

$$\boldsymbol{F(x, \alpha, \lambda) = (1 - e^{-(\lambda x)^2})^{\alpha}, \quad x > 0.}$$

Ici $\boldsymbol{\alpha}$ est le paramétre d'échelle et $\boldsymbol{\lambda}$ le paramétre de forme.

1.6.1 La fonction de densité

La distribution RG a la fonction de densité :

$$\boldsymbol{f(x, \alpha, \lambda) = 2\alpha\lambda^2 x e^{-(\lambda x)^2}(1 - e^{-(\lambda x)^2})^{\alpha-1}, \quad x > 0.}$$

1.6.2 La fonction de survie

La fonction de survie de la distribution de Rayleigh généralisée, est définie par :

$$\boldsymbol{s(x, \alpha, \lambda) = 1 - (1 - e^{-(\lambda x)^2})^{\alpha}, \quad x > 0.}$$

1.6.3 La fonction de hasard

La fonction de hasard (le taux de défaillance, de panne ou de risque) de la distribution de Rayleigh généralisée est définie par :

$$\boldsymbol{h(x, \alpha, \lambda) = \frac{f(x,\alpha,\lambda)}{s(x,\alpha,\lambda)} = \frac{2\alpha\lambda^2 x e^{-(\lambda x)^2}(1-e^{-(\lambda x)^2})^{\alpha-1}}{1-(1-e^{-(\lambda x)^2})^{\alpha}}.}$$

1.7 Les estimateurs du maximum de vraisemblance de $RG(\alpha, \lambda)$

Soit $\boldsymbol{X_1 ..., X_n}$ échantillon aléatoire de taille n de $\boldsymbol{RG(\alpha, \lambda)}$ et dans le cas où $\boldsymbol{\alpha}$ et $\boldsymbol{\lambda}$ sont inconnus, alors la fonction log-vraisemblance $\boldsymbol{L(\alpha, \lambda)}$ peut s'écrire :

$$\boldsymbol{L(\alpha, \lambda) = C + n \ln \alpha + 2n \ln \lambda + \sum_{i=1}^{n} \ln x_i - \lambda^2 \sum_{i=1}^{n} x_i^2 + (\alpha - 1) \sum_{i=1}^{n} \ln(1 - e^{-(\lambda x)^2})} \quad (1.1)$$

Les équations normales deviennent :

$$\boldsymbol{\frac{\partial L}{\partial \alpha} = \frac{n}{\alpha} + \sum_{i=1}^{n} \ln(1 - e^{-(\lambda x)^2}) = 0} \quad (1.2)$$

$$\boldsymbol{\frac{\partial L}{\partial \lambda} = \frac{2n}{\lambda} - 2\lambda \sum_{i=1}^{n} x_i^2 + 2\lambda(\alpha - 1) \sum_{i=1}^{n} \frac{x^2 e^{-(\lambda x)^2}}{1 - e^{-(\lambda x)^2}} = 0} \quad (1.3)$$

A partir de (1.2), nous obtenons le EMV de $\boldsymbol{\alpha}$ en fonction de $\boldsymbol{\lambda}$, disons $\boldsymbol{\hat{\alpha}(\lambda)}$, comme :

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{\lambda}) = -\frac{\boldsymbol{n}}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \ln(1 - e^{-(\lambda x)^2})} \tag{1.4}$$

En substituant $\hat{\boldsymbol{\alpha}}(\boldsymbol{\lambda})$ dans (1.1), on obtient le profil log-vraisemblance de $\boldsymbol{\lambda}$ comme :

$$g(\lambda) = c + n \ln\left(-\sum_{i=1}^{n} \ln(1 - e^{-(\lambda x)^2})\right) + 2n \ln \lambda + \sum_{i=1}^{n} \ln x_i$$

$$-\lambda^2 \sum_{i=1}^{n} x_i^2 - \left(-\sum_{i=1}^{n} \ln(1 - e^{-(\lambda x)^2})\right) \tag{1.5}$$

où : $\boldsymbol{c = n \ln 2}$

Par conséquent, le EMV de $\boldsymbol{\lambda}$, disons $\hat{\boldsymbol{\lambda}}$, peut être obtenu en maximisant (1.5) par rapport à $\boldsymbol{\lambda}$.

1.8 Tests d'hypothèse

Un test d'hypothèse (ou test statistique) est une démarche qui a pour but de fournir une règle de décision permettant, sur la base de résultats d'échantillon, de faire un choix entre deux hypothèses statistiques(rejeter ou à accepter).

Hypothèse nulle (H_0) et hypothèse alternative (H_1)

Soit $X_1 ..., X_n$ un échantillon aléatoire de la distribution $F(x) = \mathbb{P}(X \leq x)$.

$$\begin{cases} H_0 : F(x) = F_0(x) \\ H_1 : F(x) \neq F_0(x) \end{cases}$$

Où $F_0(x)$ est une fonction de distribution inconnue.

L'hypothèse selon laquelle on fixe à priori un paramètre de la population à une valeur particulière s'appelle l'hypothèse nulle et est notée H_0. N'importe quelle autre hypothèse qui diffère de l'hypothèse H_0 s'appelle l'hypothèse alternative (ou contre-hypothèse) et est notée H_1 .

C'est l'hypothèse nulle qui est soumise au test et toute la démarche du test s'effectue en considérant cette hypothèse comme vraie.

Dans notre démarche, nous allons établir des règles de décision qui vont nous conduire à l'acceptation ou au rejet de l'hypothèse nulle H_0. Toutefois cette décision est fondée sur une information partielle, les résultats d'un échantillon. Il est donc statistiquement impossible de prendre la bonne décision à coup sûr. En pratique, on met en oeuvre une démarche qui nous permettrait, à long terme de rejeter à tort une hypothèse nulle vraie dans une faible proportion de cas. La conclusion qui sera déduite des résultats de l'échantillon aura un caractère probabiliste :on ne pourra prendre une décision qu'en ayant conscience qu'il y a un certain risque qu'elle soit erronée. Ce risque nous

est donné par le seuil de signification du test.

Seuil de signification du test

Le risque, consenti à l'avance et que nous notons $\boldsymbol{\alpha}$ de rejeter à tort l'hypothèse nulle $\boldsymbol{H_0}$ alors qu'elle est vraie, s'appelle le seuil de signification du test et s'énonce en probabilité ainsi

$$\boldsymbol{\alpha = P(rejeter H_0 | H_0 vraie)}.$$

A ce seuil de signification, on fait correspondre sur la distribution d'échantillonnage de la statistique une région de rejet de l'hypothèse nulle (appelée également région critique). L'aire de cette région correspond à la probabilité $\boldsymbol{\alpha}$. Si par exemple , on choisit $\boldsymbol{\alpha = 0.05}$, cela signifie que l'on admet d'avance que la variable d'échantillonnage peut prendre, dans $\mathbf{5\%}$ rejet de $\boldsymbol{H_0}$, bien que $\boldsymbol{H_0}$ soit vraie et ceci uniquement d'après le hasard de l'échantillonnage.

Sur la distribution d'échantillonnage correspondra aussi une région complémentaire, dite région d'acceptation de $\boldsymbol{H_0}$ (ou région de non-rejet) de probabilité $\boldsymbol{1 - \alpha}$.

Remarques

•Les seuils de signification les plus utilisés sont $\boldsymbol{\alpha = 0.05}$ et $\boldsymbol{\alpha = 0.01}$, dépendant des conséquences de rejeter à tort l'hypothèse $\boldsymbol{H_0}$.

•La statistique qui convient pour le test est donc une variable aléatoire dont la valeur observée sera utilisée pour décider du « rejet » ou du « non-rejet » de $\boldsymbol{H_0}$. La distribution d'échantillonnage de cette statistique sera déterminée en supposant que l'hypothèse $\boldsymbol{H_0}$ est vraie.

1.8.1 Test paramétrique

Un test paramétrique est un test pour lequel on fait une hypothése sur la forme des données sous $\boldsymbol{H_0}$(normale, Poisson, ...). Les hypothéses du test concernant alors les paramétres gouvernant cette loi.

1.8.2 Test non paramétrique

Test statistique qui ne repose pas sur l'hypothèse que les variables utilisées suivent une distribution prédéterminée. Ils peuvent être utilisés lorsque l'on dispose d'un très petit nombre d'observations. Les tests non paramétriques sont traditionnellement utilisés pour les variables qualitatives.

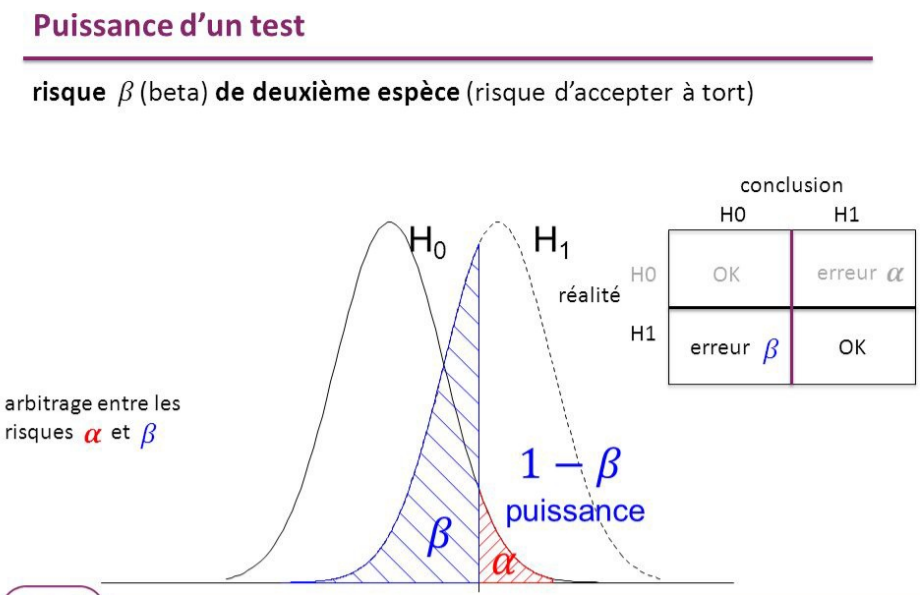
Remarques

• Les tests non-paramétriques sont plus robustes que les tests paramétriques. En d'autres termes, peuvent être utilisés dans un plus grand nombre de situations.

•Les tests paramétriques sont, eux, plus puissants en général que leurs équivalents non-paramétriques. En d'autres termes, un test paramétrique sera plus apte à aboutir à un rejet de $\boldsymbol{H_0}$, si ce rejet est justifié. La plupart du temps, la p-value calculée par un test paramétrique sera inférieure à la p-value calculée par un équivalent non-paramétrique exécuté sur les mêmes données.

1.9 La puissance d'un test

La puissance statistique d'un test est en statistique la probabilité de rejeter l'hypothèse nulle (par exemple l'hypothèse selon laquelle les groupes sont identiques au regard d'une variable) sachant que l'hypothèse nulle est incorrecte (en réalité les groupes sont différents). On peut l'exprimer sous la forme $\mathbf{1-\beta}$ où $\boldsymbol{\beta}$ est le risque de $\mathbf{2}^{\boldsymbol{éme}}$ espèce c'est-à-dire le risque de ne pas démontrer que deux groupes sont différents alors qu'ils le sont dans la réalité.



CHAPITRE 2

STATISTIQUES DES TESTS D'AJUSTEMENT

Les tests d'adéquation d'FDE mesurent la différence entre la fonction de distribution hypothétique $\boldsymbol{F}$ et la fonction de distribution empirique $\boldsymbol{F_n}$ et cette quantité est comparée à des valeurs critiques. Lorsque la distribution supposée est spécifiée, les statistiques communes d'FDE 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 avant les dernières décennies. A travers 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 2011b). Dans ce travail, en utilisant la méthode de Monte Carlo et le logiciel R, nous proposons des tables de valeurs critiques d'ajustement de $\boldsymbol{D_n}$, $\boldsymbol{A^2}$, $\boldsymbol{U^2}$, $\boldsymbol{W^2}$ et $\boldsymbol{L_n}$ pour le modèle de rayleigh généralisé lorsque les paramètres sont inconnus.

2.1 Statistique du test de Komogorov-Smirnov D_n

Le test d'ajustement le plus populaire est le test de Kolmogorov-Smirnov KS. La statistique de ce test est définie comme :

$$\boldsymbol{D_n = \max[D^+; D^-]},$$

où

$$\boldsymbol{D^+ = \max_{1\leq i\leq n}\left(\frac{i}{n} - F(x_i)\right)} \tag{2.1}$$

et

$$\boldsymbol{D^- = \max_{1\leq i\leq n}\left(F(x_i) - \frac{i-1}{n}\right)} \tag{2.2}$$

où $\boldsymbol{x_i}$ est un échantillon aléatoire ordonné et $\boldsymbol{F}$ est la fonction de distribution cumulée de la distribution spécifiée.

Pour le modèle de RG, la statistique $\boldsymbol{D_n}$ devient comme suit :

$$\boldsymbol{D^+ = \max_{1\leq i\leq n}\left(\frac{i}{n} - (1 - e^{-(\hat{\lambda}x_i)^2})^{\hat{\alpha}}\right)} \tag{2.3}$$

et

$$\boldsymbol{D^- = \max_{1\leq i\leq n}\left((1 - e^{-(\hat{\lambda}x_i)^2})^{\hat{\alpha}} - \frac{i-1}{n}\right)} \tag{2.4}$$

où $\hat{\boldsymbol{\alpha}}$ et $\hat{\boldsymbol{\lambda}}$ sont les estimateur des paramètres $\boldsymbol{\alpha}$ et $\boldsymbol{\lambda}$ successivement.

2.2 Statistique du test de Cramer-von Mises W^2

La statistique de Cramer-Von Mises $\boldsymbol{W^2}$ peut être considérée comme la somme des différences au carré entre la fonction de distribution empirique (FDE) et la fonction de distribution cumulée théorique (FDC).

Les tests Cramer-von Mises se sont avérés plus puissants que KS tester par rapport à une large classe d'hypothèses alternatives. $\boldsymbol{W^2}$ est définie par :

$$\boldsymbol{W^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left(F(x_i) - \frac{2i-1}{2n}\right)^2} \tag{2.5}$$

Donc, pour la distribution RG, on obtient :

$$\boldsymbol{W_n^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left((1 - e^{-(\hat{\lambda}x_i)^2})^{\hat{\alpha}} - \frac{2i-1}{2n}\right)^2} \tag{2.6}$$

2.3 Statistique du test Anderson-Darling A^2

La statistique de test AD a été développée par Anderson et Darling (1954) comme une distribution limite du test de C-v M comme dans $\boldsymbol{n \to \infty}$. La statistique de ce test est donnée par :

$$\boldsymbol{A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left(\log(F(x_i)) + \log(1 - F(x_i))\right)} \tag{2.7}$$

On obtient la statistique de test pour RG comme suit :

$$\boldsymbol{A_n^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left(\log((1 - e^{-(\hat{\lambda}x_i)^2})^{\hat{\alpha}}) + \log(1 - (1 - e^{-(\hat{\lambda}x_i)^2})^{\hat{\alpha}})\right)} \tag{2.8}$$

Région critiques :

Les valeurs critiques du test AD dépendent de la distribution spécifique testée. Des valeurs tabulées et des formules ont été publiées pour quelques distributions spécifiques (normale, log-normale, exponentielle, Weibull, logistique, valeur extrême de type **1** et autres). Le test est un test unilatéral et l'hypothèse selon laquelle la distribution est d'une forme spécifique est rejetée si la statistique de test, $\boldsymbol{A^2}$, est supérieure à la valeur critique.

Notez que pour une distribution donnée, la statistique AD peut être multipliée par une constante (qui dépend généralement de la taille de l'échantillon $\boldsymbol{n}$). Ces constantes sont données dans les divers articles de Stephens. Dans l'exemple de sortie ci-dessous, les valeurs des statistiques de test sont ajustées. Sachez également que différentes constantes (et donc des valeurs critiques) ont été publiées. Vous avez juste besoin de savoir quelle constante a été utilisée pour un ensemble donné de valeurs critiques (la constante nécessaire est généralement donnée avec les valeurs critiques).

2.4 Statistique du test Watson U^2

Une version modifiée du test de C-v M est le test de Watson $\boldsymbol{U^2}$.

La statistique de test de Watson a été développée pour les distributions cycliques. La statistique de Watson $\boldsymbol{U^2}$ (Watson 1961a, 1962b), peut être utilisée pour tester l'ajustement des données distribuées sur la sphère car la valeur de cette statistique ne dépend pas de le point arbitraire choisi pour commencer à cumuler la densité de probabilité. Il peut également être utilisé pour observations sur une ligne. La forme simple de cette statistique s'écrit :

$$\boldsymbol{U^2 = W^2 + \sum_{i=1}^{n}\left(\frac{F(x_i)}{n} - \frac{1}{2}\right)^2} \tag{2.9}$$

La forme explicite de cette statistique pour le modèle de RG est :

$$\boldsymbol{U_n^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left((1 - e^{-(\hat{\lambda}x_i)^2})^{\hat{\alpha}} - \frac{2i-1}{2n}\right)^2 + \sum_{i=1}^{n}\left(\frac{(1 - e^{-(\hat{\lambda}x_i)^2})^{\hat{\alpha}}}{n} - \frac{1}{2}\right)^2} \tag{2.10}$$

2.5 Statistique du test Liao et Shimokawa L_n

Liao et Shimokawa (1999a, 1999b) ont proposé une nouvelle statistique, notée $\boldsymbol{L_n}$, qui mesure la moyenne de toutes les distances pondérées sur toute la gamme de la variable. Ils ont proposé un

calcul expression de $\boldsymbol{L_n}$ comme suit :

$$\boldsymbol{L_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\max_i \left(\frac{i}{n} - F(x_i), F(x_i) - \frac{i-1}{n} \right)}{\sqrt{F(x_i)[1 - F(x_i)]}}} \tag{2.11}$$

Pour la distribution de RG, $\boldsymbol{L_n}$ devient :

$$\boldsymbol{L_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\max_i \left(\frac{i}{n} - (1 - e^{-(\hat{\lambda} x_i)^2})^{\hat{\alpha}}, (1 - e^{-(\hat{\lambda} x_i)^2})^{\hat{\alpha}} - \frac{i-1}{n} \right)}{\sqrt{(1 - e^{-(\hat{\lambda} x_i)^2})^{\hat{\alpha}} [1 - (1 - e^{-(\hat{\lambda} x_i)^2})^{\hat{\alpha}}]}}} \tag{2.12}$$

CHAPITRE 3

CALCUL DES VALEURS CRITIQUES

Le but de ce travail est de fournir des valeurs critiques d'ajustement des statistiques modifiées $\boldsymbol{D_n}$, $\boldsymbol{A_n^2}$, $\boldsymbol{W_n^2}$, $\boldsymbol{U_n^2}$ et $\boldsymbol{L_n}$ pour la distribution de Rayleigh généralisée lorsque les paramètres sont inconnus et remplacés par leurs estimateurs du maximum de vraisemblance sur les données non groupées. Pour cela, nous utilisons Monte Carlo méthode de simulation et logiciel R pour générer $\mathbf{10,000}$ échantillons de différentes tailles $\boldsymbol{n}$.

Sous l'hypothèse nulle $\boldsymbol{H_0}$ selon laquelle un échantillon $\boldsymbol{X = X_1, X_2, ..., X_n}$ appartient à modèle de Rayleigh généralisé, nous avons calculé les valeurs des différentes statistiques de tests d'ajustement mentionnées ci-dessus. À cette fin, les étapes suivantes sont utilisées pour calculer les valeurs critiques pour chaque statistique des tests d'ajustement à différents niveaux de signification $\boldsymbol{\alpha = 0.01, 0.05, 0.10, 0.15}$ et $\mathbf{0.25}$ et tailles d'échantillon $\boldsymbol{n = 10, 20, 50}$ et $\mathbf{100}$:

Étape 1 : Générer $\boldsymbol{n}$ variables aléatoires $\boldsymbol{U(0,1)}$ indépendantes $\boldsymbol{U_1, U_2, ..., U_n}$ ·

Étape 2 : Pour des valeurs données des paramètres $\boldsymbol{\alpha}$ et $\boldsymbol{\lambda}$, on pose $\boldsymbol{x_i = F^{-1}(U_i)}$.

Alors, $\boldsymbol{(x_1, x_2, ..., x_n)}$ est l'échantillon requis de taille $\boldsymbol{n}$ de la distribution RG.

Étape 3 : Utiliser l'échantillon généré pour estimer les paramètres inconnus en utilisant les estimateurs du maximum de vraisemblance donnés par (1.4) et (1.5).

Étape 4 : Les estimateurs de paramètres inconnus ont été utilisés pour déterminer la distribution cumulative hypothétique fonction de la distribution de RG.

Étape 5 : Les tests statistiques $\boldsymbol{D_n, L_n, W_n^2, U_n^2}$ et $\boldsymbol{A_n^2}$ mentionnés ci-dessus sont calculés pour chaque génération échantillon aléatoire de différentes tailles.

Étape 6 : Cette procédure a été répétée $\mathbf{10,000}$ fois indépendamment. Par conséquent, nous avons obtenu $\mathbf{10,000}$ valeurs pour chaque statistique de test proposée. Ces valeurs ont été classées à différents niveaux de signification $\mathbf{0.01, 0.05, 0.10, 0.15}$ et $\mathbf{0.25}$ sont présentés dans **le tableau 1**.

Tableau 1 : Valeurs critiques pour les tests KS, C-vM, AD, W et LS

Taille	tests	Niveau de signification α				
de l'échantillon n	statistiques	0.01	0.05	0.10	0.15	0.25
10	D_n	0.0000	0.0004	0.0017	0.0049	0.0111
	W_n^2	0.0003	0.0054	0.0182	0.0306	0.0593
	A_n^2	0.0145	0.0384	0.0706	0.1359	0.1986
	U_n^2	0.0002	0.0029	0.0099	0.0178	0.0352
	L_n	0.0125	0.0345	0.0522	0.0965	0.1002
20	D_n	0.0000	0.0004	0.0011	0.0045	0.0100
	W_n^2	0.0003	0.0048	0.0133	0.0304	0.0565
	A_n^2	0.0131	0.0347	0.0657	0.1297	0.1836
	U_n^2	0.0002	0.0026	0.0072	0.0174	0.0332
	L_n	0.0111	0.0318	0.0452	0.0865	0.0987
50	D_n	0.0000	0.0003	0.0009	0.0038	0.0079
	W_n^2	0.0002	0.0043	0.0126	0.0289	0.0563
	A_n^2	0.0120	0.0259	0.0561	0.1132	0.1755
	U_n^2	0.0001	0.0023	0.0067	0.0163	0.0321
	L_n	0.0101	0.0245	0.0398	0.0792	0.0923
100	D_n	0.0000	0.0001	0.0006	0.0030	0.0067
	W_n^2	0.0002	0.0039	0.0120	0.0284	0.0530
	A_n^2	0.0106	0.0250	0.0524	0.1078	0 .1585
	U_n^2	0.0001	0.0020	0.0063	0.0157	0.0298
	L_n	0.0097	0.0231	0.0341	0.0762	0.0878

Remarque

D'après le tableau, nous avons remarqué que :

•Pour chaque test statistique, la puissance augmente de manière monotone à mesure que la taille de l'échantillon augmente et que le niveau de signification augmente.

•Le test statistique de **Anderson-Darling** A_n^2 est le plus puissant parmi les tests d'ajustement proposés.

•Le test statistique de **Komogorov-Smirnov** D_n est le moins puissant parmi les tests d'ajustement proposés.

CHAPITRE 4

ÉTUDE DE PUISSANCE

Dans ce chapitre, nous avons effectué une comparaison de puissance entre $\boldsymbol{D_n, L_n, W_n^2, U_n^2}$ et $\boldsymbol{A_n^2}$ statistiques pour le modèle de RG avec des paramètres inconnus. Pour cela, nous avons simulé $\mathbf{10,000}$ échantillons aléatoires de différentes tailles $\boldsymbol{n = 10, 20, 50}$ et $\mathbf{100}$, pour chaque test au niveau de signification $\boldsymbol{\alpha = 0.05}$ et de chacune des distributions alternatives :

4.1 Distributions alternatives

4.1.1 Distribution exponentielle

Dans la théorie des probabilités et les statistiques, la distribution exponentielle est une distribution de probabilité continue qui concerne souvent le temps jusqu'à ce qu'un événement spécifique se produise. C'est un processus dans lequel les événements se produisent de manière continue et indépendante à un rythme moyen constant. La distribution exponentielle a la propriété clé d'être sans mémoire. La variable aléatoire exponentielle peut être soit plus de petites valeurs, soit moins de grandes variables.

Par exemple, le montant d'argent dépensé par le client lors d'un voyage au supermarché suit une distribution exponentielle.

Définition :

La variable aléatoire continue, disons $\boldsymbol{X}$, est dite avoir une distribution exponentielle, si elle a **la fonction de densité** de probabilité suivante :

$$\boldsymbol{f_X(x, \lambda)} = \begin{cases} \boldsymbol{\lambda e^{-\lambda x}}, & \boldsymbol{x > 0} \\ \boldsymbol{0}, & \boldsymbol{x \leq 0} \end{cases}$$

Où $\boldsymbol{\lambda > 0}$ s'appelle le taux de distribution.

Et sa **Fonction de répartition** est :

$$\boldsymbol{F_{Exp}(x,\lambda)} = \begin{cases} \mathbf{1 - e^{-\lambda x},} & \boldsymbol{x > 0} \\ \mathbf{0,} & \boldsymbol{x \leq 0} \end{cases}$$

4.1.2 Distribution de Weibull

La distribution de Weibull est une distribution de probabilité continue utilisée pour analyser les données de durée de vie, modéliser les temps de défaillance et accéder à la fiabilité des produits. Il peut également contenir une vaste gamme de données provenant de nombreux autres domaines tels que l'économie, l'hydrologie, la biologie, les sciences de l'ingénieur. Il s'agit d'une valeur extrême de distribution de probabilité fréquemment utilisée pour modéliser la fiabilité, la survie, la vitesse du vent et d'autres données.

La seule raison d'utiliser la distribution de Weibull est sa flexibilité. Parce qu'il peut simuler diverses distributions comme les distributions normales et exponentielles. La fiabilité de la distribution de Weibull est mesurée à l'aide de paramètres.

Distribution de Weibull à deux paramètres

Avec deux paramètres, La formule de la fonction de densité de probabilité de Weibull $\boldsymbol{Wei(\gamma,\alpha)}$ est :

$$\boldsymbol{f(x,\gamma,\alpha)} = \frac{\gamma}{\alpha}(\frac{x}{\alpha})^{\gamma-1} e^{-(\frac{x}{\alpha})^{\gamma}}, \quad \boldsymbol{x > 0}.$$

Et sa **fonction de répartition** est :

$$\boldsymbol{F_{Wei}(x,\gamma,\alpha)} = \mathbf{1} - e^{-(\frac{x}{\alpha})^{\gamma}}, \quad \boldsymbol{x > 0}.$$

où $\boldsymbol{\gamma > 0}$ est le paramètre de forme et $\boldsymbol{\alpha > 0}$ le paramètre d'échelle de la distribution.

4.1.3 Distribution de Weibull inverse

La distribution de Weibull inverse a la capacité de modéliser les taux d'échecs qui sont les plus importants dans les domaines de la fiabilité et de l'étude biologique.
La fonction de densité de probabilité de la distribution de Weibull inverse $\boldsymbol{Weiin(\alpha,\gamma)}$ est donnée par :

$$\boldsymbol{f(x,\gamma,\alpha)} = \gamma\alpha^{\gamma}x^{-(\gamma+1)}e^{-(\frac{\alpha}{x})^{-\gamma}}, \quad \boldsymbol{x > 0}.$$

Et sa **fonction de répartition** est :

$$\boldsymbol{F_{Weiin}(x,\gamma,\alpha)} = e^{-\alpha(\frac{1}{x})^{-\gamma}}, \quad \boldsymbol{x > 0}.$$

4.1.4 Distribution de Weibull exponentielle

La distribution de Weibull exponentielle $\boldsymbol{WeiExp(\alpha, \gamma, \lambda)}$ est définie de la manière suivante. Sa fonction de répartition est donnée par :

$$\boldsymbol{F_{EW}(x, \alpha, \gamma, \lambda) = (1 - e^{-\lambda x^{\gamma}})^{\alpha}, \quad x > 0, \alpha, \lambda, \gamma > 0.}$$

Et donc sa fonction de densité de probabilité (fdp) est de la forme :

$$\boldsymbol{f_{EW}(x, \alpha, \gamma, \lambda) = \alpha\gamma\lambda^{\gamma}x^{\gamma-1}(1 - e^{-\lambda x^{\gamma}})^{\alpha}, \quad x > 0.}$$

4.1.5 Distribution Exponentielle Exponentielle

La fonction de répartition $\boldsymbol{F_{EE}}$ et la fonction de densité probabilité $\boldsymbol{f_{EE}}$ de la distribution Exponentielle Exponentielle $\boldsymbol{EE(\alpha, \lambda)}$ sont données par :

$$\boldsymbol{F_{EE}(x, \alpha, \lambda) = (1 - e^{-\lambda x})^{\alpha}, \quad x > 0.}$$

Et

$$\boldsymbol{f_{EE}(x, \alpha, \lambda) = \alpha\lambda(1 - e^{-\lambda x})^{\alpha-1}e^{-\lambda x}, \quad x > 0.}$$

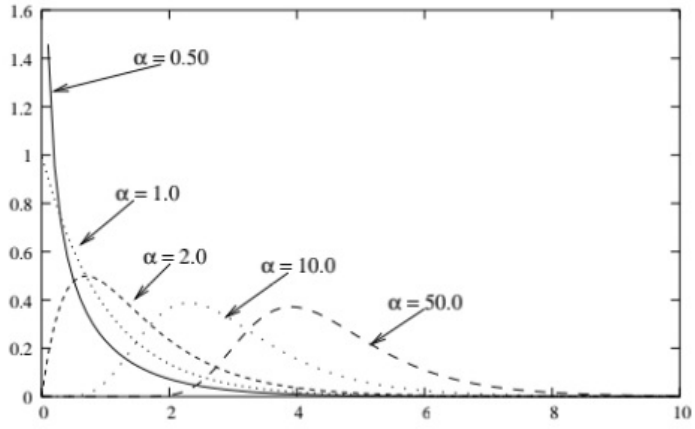


Figure 1: Different density functions of the exponentiated exponential distributions

pour $\boldsymbol{\alpha, \lambda > 0}$. Ici $\boldsymbol{\alpha}$ et $\boldsymbol{\lambda}$ sont respectivement les paramètres de forme et d'échelle. Notez que le la distribution exponentielle exponentielle est un membre particulier de la distribution de Weibull exponentielle. Elle est à deux paramètres peut-être utilisée assez efficacement dans l'analyse de plusieurs données de durée de vie, en particulier en place de gamma à deux paramètres ou de distribution de Weibull à deux paramètres.

4.2 Essais de puissance

A titre d'illustration, nous avons estimé la puissance des tests d'ajustements lorsque les alternatives sont les distributions exponentielles, Weibull, Weibull inverse, Weibull exponentielle et les distributions exponentielles-exponentielles $\boldsymbol{EE}$. Nos résultats sont résumés dans le tableau **2**. La puissance a été déterminée en générant $\mathbf{10,000}$ échantillons aléatoires de taille $\boldsymbol{n} = \mathbf{10}$, **20**, **50** et **100** à partir de chacune des alternatives et pour chaque test au niveau de signification $\boldsymbol{\alpha} = \mathbf{0.05}$.

Tableau 2 : Puissance des tests statistiques pour la distribution RG où Exp, Wei, Weiin, WeiExp et EE sont les distributions alternatives.

Alternatives	tests	Taille de l'échantillon n			
	statistiques	10	20	50	100
Exponentielle $Exp(1)$	D_n	1.0000	1.0000	1.0000	1.0000
	W_n^2	0.1016	0.3653	0.9119	0.9997
	A_n^2	0.4158	0.7834	0.9976	1.0000
	U_n^2	0.1004	0.3202	0.9164	0.9786
	L_n	0.1059	0.3728	0.9993	1.0000
Weibull $Wei(1,2)$	D_n	1.0000	1.0000	1.0000	1.0000
	W_n^2	0.0995	0.3542	0.9080	0.9998
	A_n^2	0.0644	0.0603	0.0539	0.0495
	U_n^2	0.0244	0.0282	0.0393	0.0450
	L_n	0.0159	0.0228	0.0324	0.0445
Weibull inverse $Weiin(1,2)$	D_n	0.9249	0.9992	1.0000	1.0000
	W_n^2	0.1059	0.3101	0.9463	0.9459
	A_n^2	0.8286	0.9324	0.9981	1.0000
	U_n^2	0.1083	0.3089	0.9059	0.9228
	L_n	0.1055	0.3076	0.9034	0.9210
Weibull exponentielle $WeiExp(1,2,3)$	D_n	0.9999	0.9996	1.0000	0.9999
	W_n^2	0.1035	0.3588	0.9997	0.9995
	A_n^2	0.9999	0.9998	0.9992	0.8853
	U_n^2	0.1030	0.3438	0.9127	0.9960
	L_n	0.1011	0.3298	0.9037	0.9860
Exponentielle Exponentielle $EE(1,2)$	D_n	1.0000	1.0000	1.0000	1.0000
	W_n^2	0.1055	0.3676	0.9087	0.9994
	A_n^2	0.0592	0.0726	0.0639	0.0597
	U_n^2	0.0548	0.0526	0.0611	0.0684
	L_n	0.0539	0.0523	0.0601	0.0672

Remarque :

Les résultats de puissance des tests statistiques D_n, L_n, W_n^2, U_n^2 et A_n^2 pour chaque distribution alternative à niveau de signification $\alpha = 0.05$ sont présentés dans le tableau **2**.

D'après le tableau, nous remarquons que :

•les valeurs de puissance de test pour les différentes statistiques, sont indiquant que le modèle de Rayleigh généralisé est distinct des distributions concurrentes de toutes les tailles d'échantillon.

•La puissance de la statistique de test augmente à mesure que la taille de l'échantillon augmente.

•les statistiques de test modifiées $\boldsymbol{D_n, A_n^2, U_n^2, W_n^2}$ et $\boldsymbol{L_n}$ fournies dans ce travail et leurs valeurs critiques peuvent détecter la différence entre le modèle RG et différentes alternatives avec haute puissance.

CONCLUSION

Nous avons fourni des valeurs critiques pour les statistiques $\boldsymbol{D_n, L_n, W_n^2, U_n^2}$ et $\boldsymbol{A_n^2}$ pour le modèle de Rayleigh généralisé lorsque les paramètres sont inconnus. Les tableaux **1** et **2** donnés dans ce manuscrit 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] Abd-Elfattah, A. M., H. A. Fergany, and A. M. Omima. 2010. Goodness-Of- Fit test for the generalized Fr-echet distribution. Australian Journal of Basic and Applied Science 4 (2) :286–301.

[2] Anderson, T. W., and D. A. Darling. 1954. A test of goodness of fit. Journal of the American Statistical Association 49 (268) :765–9. doi :10.1080/01621459.1954.10501232.

[3] Al-Zahrani, B. 2012. Goodness-of-Fit for the Topp-Leone distribution with unknown parameters. Applied Mathematical Sciences 6 (128) :6355–63.

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

[5] D. Tilbi, N. Seddik-Ameur (2017).Chi-squared goodness-of-fit tests for the generalized Rayleigh distribution, Journal of Statistical Theory and Practice 11 (4), 594-603.

[6] Evans, J. W., R. A. Johnson, and D. W. Green. 1989. Two- and three-parameter Weibull goodness-of-fit tests. Madison : U.S. Department of agriculture Forest Service, Forest Products Laboratory.

[7] Gupta, R. D. and Kundu, D. (2001), "Exponentiated exponential family ; an alternative to gamma and Weibull", Biometrical Journal, vol. 43, 117 - 130.

[8] Kolmogorov, A. N. 1933. Sulla determinazione empirica di una legge di distribuzione [On the empirical determination of a distribution law]. Giornale dell'Istituto Italiano degli Attuari 4 (1) :83–91.

[9] Kundu, D. and Raqab, M.Z. (2005). Generalized Rayleigh distribution : different methods of estimation. Computational Statistics and Data Analysis, 49, 187-200.

[10] Lemeshko, B. Y., and S. B. Lemeshko. 2011b. Construction of statistic distribution models for nonparametric goodness-of-fit tests in testing composite hypotheses. The computer approach. Quality Technology Quantitative Management 8(4) :359–73. doi :10.1080/16843703.2011.11673263.

[11] Liao, M., and T. Shimokawa. 1999a. A new goodness-of-fit test for Type-I extreme-value and 2-parameter Weibull distributions with estimated parameters. Journal of Statistical Computation and Simulation 64 (1) :23–48. doi :10.1080/00949659908811965.

[12] Liao, M., and T. Shimokawa. 1999b. Goodness-of-fit test extreme-value and 2-parameter Weibull distributions.IEEE Transactions on Reliability 48 (I) :79–86. doi :10.1109/24.765931.

[13] Mudholkar, G.S.and Srivastava, D.K. (1993). "Exponentiated Weibull family for analyzing bathtub failure-rate data". IEEE Transactions on Reliability. 42, 299- 302.

[14] Raqab, M. Z. and Ahsanullah, M. (2001), “Estimation of the location and scale pa rameters of generalized exponential distribution based on order statistics”, Journal of Statistical Computation and Simulation, vol. 69, 109 - 124.

[15] Stephens, M. A. 1970. Use of the Kolmogorov-Smirnov, Cramervon Mises and related statistics without extensive tables. Journal of the Royal Statistical Society. Series (B) 32 (1) :115–22.

[16] Stephens, M. A. 1974. EDF statistics for goodness of fit and some comparisons. Journal of the American Statistica Association 69 (347) :730–7. doi :10.1080/01621459.1974.10480196.

[17] Stephens, M. A. 1977. Goodness-of-fit for the extreme value distribution. Biometrika 64 (3) :583–8. doi :10.1093/biomet/ 64.3.583.

[18] Stephens, M. A. 1979. EDF tests of fit for the logistic distribution. Technical report No. 275, Department of statistics, Stanford university, California, USA.

[19] Watson, G. S. 1961a. Goodness-of-fit tests on a circle I. Biometrika 48 (1-2) :109–14. doi :10.2307/2333135.

[20] Watson, G. S. 1962b. Goodness-of-fit tests on a circle. II. Biometrika 49 (1-2) :57–63. doi :10.1093/biomet/49.1-2.57.