RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
Ministère de L'enseignement Supérieur et de la Recherche Scientifique
UNIVERSITÉ IBN KHALDOUN TIARET
FACULTÉ DE MATHÉMATIQUES ET DE L'INFORMATIQUES
Département de Mathématiques

MÉMOIRE DE MASTER

Spécialité :
« Mathématiques »
Option :
«Analyse fonctionnelle et applications »

Présenté Par :

MEFTAH Ikhlas & OUIS Khaoula

Intitulé :

Tests d'hypothèses : Principes et méthodes

Soutenu publiquement le 06 / 06 / 2024
à Tiaret devant le jury composé de :

M. MAAZOUZ Kadda	MCA	U. Ibn Khaldoun Tiaret	Président
M. BENALLOU Mohamed	MCB	U. Ibn Khaldoun Tiaret	Encadreur
M. BEDDANI Hamid	MCA	ESGEE Oran	Examinateur

Année universitaire :2023/2024

الحمد لله الذي بنعمته تتم الصالحات.

الحمد له حمدا كثيرا على توفيقه لهذا

و الصلاة والسلام على أشرف المرسلين

سيدنا محمد و على اله و صحبه أجمعين

Remerciements

Nos remerciements et nos profondes gratitudes vont à notre Encadreur Monsieur BENALLOU MOHAMED pour ses précieux conseils et pour tout le soutien et l'orientation. D'avoir bien voulu diriger notre travail, d'avoir donner le meilleur de son savoir, de son aide, surtout d'avoir fait preuve de beaucoup de patience, son aide durant toute la période du travail.

Nous tenons aussi à remercier les membres du jury pour leur précieux temps accordé à l'étude de notre mémoire.

Nous remercions nos enseignants pour leurs efforts , nos parents et nos proches pour l'amour et le soutien contant qu'ils nous ont témoigné tout au long de notre parcours. Merci à toutes et touts nos amis pour leurs encouragements.

Résumé

Le test d'hypothèse est un outil fondamental utilisé dans la recherche scientifique pour valider ou rejeter des hypothèses sur les paramètres d'une population à partir d'un échantillon de données. Il fournit un cadre structuré pour évaluer la signification statistique d'une hypothèse et tirer des conclusions sur la véritable nature d'une population. Les tests d'hypothèses sont largement utilisés dans des domaines tels que **la biologie**, **la psychologie**, **l'économie** et **l'ingénierie** pour déterminer l'efficacité de nouveaux traitements, explorer les relations entre les variables et prendre des décisions fondées sur des données. Cependant, malgré son importance, le test d'hypothèse peut être un sujet difficile à comprendre et à appliquer correctement. Dans ce mémoire, nous présentons une introduction aux tests d'hypothèses, y compris leur objectif, les types de tests, les étapes à suivre, les erreurs courantes et les meilleures pratiques. Que vous soyez un débutant ou un chercheur expérimenté, ce travail vous servira de guide précieux pour maîtriser les tests d'hypothèses.

Table des matières

Table des figures

Introduction

0.1 Historique

Dans la pratique on est souvent amené à prendre diverses décisions au sujet d'une population et ce à partir de l'information que donne un échantillon. On appelle de telles décisions des décisions statistiques. On veut par exemple décider si un nouveau médicament est efficace, si une méthode pédagogique est meilleure qu'une autre, si une pièce de monnaie est bien équilibrée, etc...
Une hypothèse statistique, notée H, est une proposition logique contenant les caractéristiques d'une ou plusieurs populations données, (comme la forme éventuelle de la distribution ou l'égalité entre des lois) ou des valeurs pour des paramètres. Nous sommes maintenant en cours de déterminer s'il existe suffisamment de preuves pour étayer ou rejeter cette hypothèse, c-à-d tester l'hypothèse, donc on peut dire que le test d'hypothèse est un outil statistique couramment utilisé dans la recherche, il consiste à formuler une hypothèse sur un paramètre de la population, à collecter des données et à les analyser pour déterminer la probabilité que l'hypothèse soit vraie. Il s'agit d'un élément essentiel de la méthode scientifique, qui est utilisé dans un grand nombre de domaines. Le processus de vérification des hypothèses implique généralement deux hypothèses : l'hypothèse nulle (H_0) et l'hypothèse alternative (H_1). L'hypothèse nulle est une affirmation selon laquelle il n'y a pas de différence significative entre deux variables ou pas de relation entre elles, tandis que l'hypothèse alternative suggère la présence d'une relation ou d'une différence. Un test statistique de l'hypothèse H_0 contre l'hypothèse H_1 est une démarche qui a pour but de fournir une règle de décision permettant de faire un choix entre les deux hypothèses sur la base des réalisations de l'échantillon. Les chercheurs collectent des données et effectuent des analyses statistiques pour déterminer si l'hypothèse nulle peut être rejetée en faveur de l'hypothèse alternative. Évidemment, il ne doit pas exister d'événement réalisant les hypothèses H_0 et H_1 simultanément. Ce qui s'écrit $H_0 \cap H_1 = \varnothing$. La région critique d'un test est l'ensemble des valeurs observées pour les-

quelles l'hypothèse nulle H_0 est rejetée. les valeurs limites de cette région constituent les valeurs critiques. La région d'acceptation de H_0 est le complément de la région critique, autrement dit elle est formée par l'ensemble des valeurs observées pour lesquelles l'hypothèse nulle H_0 est acceptée. Que l'on rejette ou que l'on accepte une hypothèse nulle, donc quelle que soit la décision, on prend le risque de commettre l'une des erreurs suivantes :
1. L'hypothèse H_0 est rejetée à tort.
2. L'hypothèse H_0 est retenue de façon injustifiée.
Nous sommes donc face à deux erreurs possible dont nous discuterons en détail plus tard.
Les tests d'hypothèses sont utilisés pour prendre des décisions basées sur des données, et il est important de comprendre les hypothèses sous-jacentes et les limites du processus. Il est essentiel de choisir les tests statistiques et les tailles d'échantillons appropriés pour garantir la précision et la fiabilité des résultats, et il peut s'agir d'un outil puissant permettant aux chercheurs de valider leurs théories et de prendre des décisions fondées sur des données probantes.
La théorie des tests est l'une des deux branches de la statistique mathématique. Elle se subdivise en deux volets principaux, les tests paramétriques et les tests non-paramétriques. Ces derniers n'imposent aucune forme à la loi de probabilité des phénomènes étudiés contrairement au cas paramétrique qui requiert un modèle à fortes contraintes (comme la normalité des distributions, l'égalité des moyennes,. . .). Parmi les tests les plus usuels en statistique, on peut citer le test de normalité d'une population, les tests d'égalité des paramètres, les tests de corrélation, etc. La littérature statistique abonde de types de tests statistiques. Notre choix s'est porté sur les tests paramétriques dans le cas d'une population normale, on va étudier ici les tests de conformités et les tests d'homogénéités, puis on va aborder quelque tests non-paramétriques et plus particulièrement les tests de kolmogorov-sminrov et les tests de Khi-deux, vu leur importance pratique (domaine médical, économique, social,. . .).

0.2 L'organisation du mémoire

Ce mémoire s'organise en trois chapitres.
Le premier chapitre est divisé en deux parties, la première traite les notions de base en statistique et les définitions nécéssaires, à savoir, les variables aléatoires, espace probabilisé, les lois de probabilité, ect ..., qui sont nécessaires pour l'analyse statistique des données, dans la seconde partie nous discutons sur les tests statistiques, on rappelle un certain nombre de généralités autour des tests d'hypothèses. On présente alors toutes les notions principales qui dépendent d'un test comme hypothèse, erreur, risque, région d'acceptation et de rejet,..., de plus on montre la démarche d'un test qui nous conduit à prendre une décision concernant l'hypothèse posée. L'objectif étant d'être capable de bien formuler un test.

Le deuxième chapitre concerne les tests paramétriques, nous aborderons dans ce chapitre les deux types : (test de conformité et test de homogénéité).
Le troisième chapitre est consacré au deuxième méthode de test, qui est les test non paramétrique, nous nous limiterons ici au test de kolmogorov-sminrov et le test du Khi-deux.
Enfin, nous avons collecté les tableaux statistiques pour les lois utilisées en annexe.

Chapitre 1

Généralités

Ce chapitre permet de reprendre certaines notions de base de l'inférence statistique, et les principe général d'un test d'hypothèse.

1.1 Notions de base

Nous allons maintenant donner quelques définitions et informations de base dont nous aurons besoin pour les testes d'hypothèses.

1.1.1 Population :

Une population est un ensemble d'individus ou d'éléments partageant une ou plusieurs caractéristiques qui servent à les regrouper. On parle ainsi de population humaine, statistique, biologique, civile. En statistique descriptive, une population est un ensemble fini d'objets (les individus ou unités statistiques) sur lesquels une étude se porte et dont les éléments répondent à une ou plusieurs caractéristiques communes.

1.1.2 Échantillon :

Un échantillon est un ensemble d'éléments choisis aléatoirement pour représenter une population étudiée statistiquement. Les éléments peuvent être des objets, comme les pièces prélevées dans une ligne de production pour vérifier leur conformité, des informations, comme les mesures d'épaisseur en divers points d'une plaque, des êtres vivants dans le cas de la surveillance sanitaire, ou des humains comme dans le cas d'un sondage d'opinion.

Exemple 1.1.1 *Supposons que nous ayons un groupe de* 1000 *élèves dans une école et que nous voulions savoir quel pourcentage d'élèves aiment les mathématiques. Au lieu d'interroger*

tous les étudiants, nous pouvons prendre un petit échantillon de 100 *étudiants et les interroger sur leurs intérets. Si nous constatons que* 30 *élèves aiment les mathémtiques, nous pouvons utiliser ce nombre pour estimer la proportion d'élèves qui aiment les mathématiques dans l'ensemble de l'école.*

1.1.3 Espace probabilisé :

Une expérience est appellé "aléatoire" s'il est impossible de prévoir son résultat sûr à l'avance, et si elle est répétée dans des conditions identiques. On appelle ensemble associé à une expérience aléatoire l'énsemble fondamentale Ω ={tous résultat possibles de cette expérience}. Un espace de probabilité ou espace probabilisé est la donnée d'une probabilité à tout événement, il permet la modélisation quantitative de l'expérience aléatoire étudiée. Formellement, c'est un triplet $(\Omega, F, \mathbb{P})$, F est un ensemble des événements ou tribu sur Ω, et $\mathbb{P}$ est une probabilité sur F.

Exemple 1.1.2 *Lancement d'une pièce de monnaie équilibrée. Les résultats possibles sont "pile"ou"face" et chaque résultat a une probabilité de* $\frac{1}{2}$.

1.1.4 Variable aléatoire :

On distingue divers types de variables selon la nature des données. Ainsi, une variable peut être qualitative ou quantitative ; une variable qualitative peut être nominale ou ordinale, alors qu'une variable quantitative peut être continue ou discrète. Une variable qualitative est dite ordinale si ses valeurs peuvent être ordonnées, c'est-à-dire classées sans ambiguïté de la plus petite à la plus grande, par exemple des qualificatifs comme « souvent » ou « parfois », ou des mentions comme « bien » et « très bien ». Elle est dite nominale si ces valeurs ne peuvent pas être ordonnées, du moins a priori, par exemple les caractéristiques socio-économiques comme la profession, le sexe, la nationalité.
D'autre part une variable aléatoire quantitative est une fonction entre l'ensemble des éventualités, c'est-à-dire l'ensemble des résultats possibles d'une expérience aléatoire et les nombres réels telle que pour chaque événement élémentaire il y a un et un seul nombre réel qui lui est associé. Une variable aléatoire est généralement noté par une lettre de la fin de l'alphabet en majuscule comme par exemple X, T, W, etc. Cela est une convention généralement acceptée et comme toutes les conventions il y a certaines exceptions. La définition des évènements sur l'ensemble des nombres réels est facilitée par les relations d'ordre entre les nombres ($=$, $<$, $\leq$, $\neq$, $\geq$, $>$).

– On peut ainsi définir l'événement "le résultat est 7" par $X = 7$ ou "le résultat est de moins de 4" par $X < 4$, etc.

– L'ensemble des nombres réels que la variable aléatoire peut prendre s'appellele support et on le note S_X.
– Lorsque l'ensemble des résultats possibles S_X de la **v.a.** est fini ou dénomébrable, on dit que la variable aléatoire est discrète.

– Lorsque les résultats possibles d'une **v.a**. est un intervalle de l'ensemble des nombres réels, on dit que la **v.a**. est continue.
– Il y a deux facettes à la notion de variable aléatoire quantitative, la fonction qui fait l'association et l'expérience aléatoire sur les nombres.

Variable aléatoire discrète

Lorsqu'une variable aléatoire est discrète, il suffit de connaître la probabilité de chaque évènement de la forme $X = x$ pour chaque valeur x possible pour être en mesure d'évaluer la probabilité d'un évènement quelconque. On peut donc dire que la v.a. est entièrement définie par son support S_x, et l'ensemble des probabilités associées. Soit X une variable alèatoire de support S_x et notons $f(x)$ la fonction qui permet de calculer la probabilité de chaque résultat possible de la variable aléatoire : $f(x) = \Pr(X = x)$, on dit que f est la loi de probabilité de la variable aléatoire ou sa fonction de masse.
On note la loi de probabilité simplement par f lorsqu'il n'y a pas d'ambiguité possible et par f_x lorsqu'il peut y avoir plusieurs variables aléatoires dans un même contexte.

Exemple 1.1.3 *On considère l'expérience aléatoire consistant à lancer un dé équilibré, et observer le nombre de points sur la face visible. On veut la loi de probabilité de cette variable aléatoire.*
L'ensemble S est les 6 résultats possibles (les six faces du dé) tandis que la variable aléatoire qui donne le nombre de points sur la face visible du dé prend les valeurs de 1 à 6, $S_x = \{1, 2, 3, 4, 5, 6\}$. Si on veut par exemple calculer la probabilité d'obtenir un 2, on doit avoir la fonction de masse de la variable aléatoire qui donne le nombre de points sur la face du dé visible : $X \equiv$«le nombre de points sur le dé ». On peut déterminer cette fonction de masse par un argument d' équiprobabilité : $f(x) = 1/6$ pour $x = 1, 2, 3, 4, 5, 6$. Cela veut dire que $\Pr(X = 2) = f(2) = 1/6$ et ainsi de suite pour toutes les valeurs.

Proposition 1.1.1 *Soit X une variable aléatoire de support S_X et A un événement défini sur ce support alors :*

$$\begin{aligned} \Pr(A) &= \sum_{x \in A} \Pr(X = x) \\ &= \sum_{x \in A} f_X(x). \end{aligned}$$

On peut donc calculer une probabilité quelconque en se basant sur la loi de probabilité (fonction de masse). En fait on applique le principe des événements disjoints pour une fonction de probabilité : la loi de probabilité est une fonction de probabilité donc cette propriété s'applique. Pour obtenir la probabilité d'un événements quelconque défini sur $\mathbb{R}$ il suffit de prendre chaque élément du support qui est dans l'événement puis de faire la somme des valeurs pour la fonction de masse.
Si une variable aléatoire a un support donné par $S_X = \{4, 16, 64, 256\}$, alors pour calculer la probabilité $\Pr(X > 16)$, il suffit de trouver les valeurs du support satisfaisant cet événement (64 et 256) puis d'y appliquer la fonction de masse :

$$\begin{aligned} \Pr(X &> 16) = \Pr(X = 64 \text{ ou } X = 256) \\ &= f_X(64) + f_X(256). \end{aligned}$$

Espérance et variance d'une v.a. discrète

Une variable aléatoire discrète est entièrement définie par sa fonction de masse. L'information est cependant très dense et il est difficile de comprendre le comportement de la variable aléatoire en considérant toute l'information. Il est plus facile de se baser sur des mesures ponctuelles, pour décrire certaines caractéristiques des variables aléatoires et visualiser un angle à la fois. Il y a plusieurs angles différents qui contiennent tous des éléments d'information pertinent pour l'interprétation. Les deux principales caractéristiques abordées ici sont la notion de "centre" et de dispersion des valeurs du support. La première caractéristique est la moyenne ou l'espérance.

Définition 1.1.1 *[2]Soit X une variable aléatoire de support S_X son espérance est définie par :*

$$\mathbb{E}(X) = \sum_{x \in S_X} x f_X(x).$$

On note aussi ce paramètre μ ou μ_X.

La notion de moyenne n'est pas suffisante pour donner une idée du comportement de la variable aléatoire : la notion de variation est très importante c'est-à-dire dans quelle mesure il y aura des valeurs plus ou moins éloignées de la moyenne. La variance permet de mesurer l'écart entre les différentes valeurs possibles c'est un indice de la dispersion des valeurs autour de la moyenne.

Définition 1.1.2 *[2]Soit X une variable aléatoire, sa variance est définie par :*

$$\sigma_X^2 = \sum_{x \in S_X} (x - \mu)^2 f_X(x).$$

On peut conclure que :

$$\sigma_X^2 = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right]$$

L'écart type, noté σ_X est donné par la racine de la variance, $\sigma_X = \sqrt{\sigma_X^2}$.

Quelques variables aléatoires discrètes :

La variable aléatoire discrète la plus simple est appelée variable de **Bernoulli,** notée $\mathcal{B}(p)$. Celle-ci peut prendre deux états, qu'il est toujours possible de coder 1 et 0, avec les probabilités p et $1-p$. Une interprétation simple concerne un jeu de dé dans lequel on gagnerait 100 dinars en tirant le six ($p = 1/6$). Sur une séquence de parties, la moyenne des gains tend vers p lorsque le nombre de parties tend vers l'infini, alors cette variable a pour espérence p et la variance égale à $p(1-p)$.
Si on considère qu'une partie est constituée par tirages au lieu d'un seul, le total des gains est une réalisation d'une variable **binomiale** notée $\mathcal{B}(n,p)$ qui peut prendre toutes les valeurs entières de 0 à n , alors si X est une variable $\mathcal{B}(n,p)$ on a $\Pr(X = k) = C_n^k(p)^k(1-p)^{n-k}$. Cette variable a pour moyenne le produit np, et la variance égale à $np(1-p)$.

Remarque 1.1.1 *On considère qu'une très bonne approximation de La loi binomiale $\mathcal{B}(n,p)$ par la loi normale $\mathcal{N}(np, np(1-p))$ lorsque $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$.*

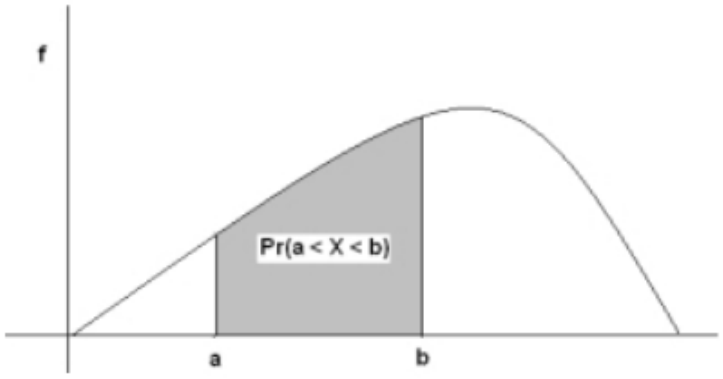
Variable aléatoire continue

Une variable continue peut prendre des valeurs réelles dans un intervalle. Cela veut dire qu'il y a toujours une infinité non dénombrable de valeurs possibles dans le support. Une conséquence immédiate est que pour une variable aléatoire continue la probabilité d'un point est nulle : $\Pr(X = x) = 0$, et cela pour chaque point dans l'intervalle. On ne peut alors évaluer une probabilité que sur un intervalle de valeurs c'est-à-dire pour un événement du type $a \leq X \leq b$. La loi de probabilité est alors la fonction qui permet de calculer les probabilités pour la variable aléatoire, qui est la fonction de densité.

La fonction de densité
La densité ou loi de probabilité d'une variable aléatoire continue est la fonction f non négative telle que pour tout $a < b$,

$$\Pr(a \leq X \leq b) = \int_a^b f(x)dx.$$

L'opérateur $\int$ est un outil mathématique pour calculer la surface sous une courbe entre deux bornes. On peut donc dire que le calcul d'une probabilité pour une variable aléatoire continue consiste à calculer une surface sous une courbe entre deux points comme l'illustre le graphique suivant :



(1.1.1)

Fig 1.1.1 : Calcul de probabilité

On aura toujours une surface totale de 1 sous l'ensemble de la courbe pour satisfaire la propriété $\Pr(S_X) = 1$.

Exemple 1.1.4 *Considérons la fonction de densité suivante :*

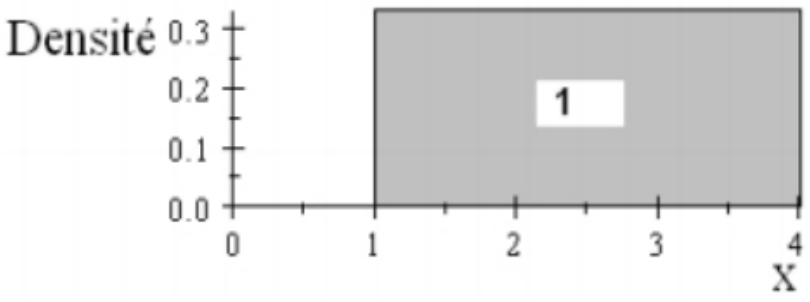
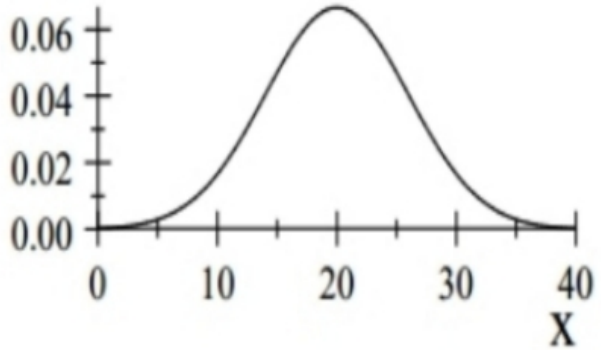


Fig 1.1.2 : *Exemple de densité*

La fonction prend la valeur $1/3$ si $1 < x < 4$ et 0 sinon. Le support de la v.a. est de $(1, 4)$ et on a effectivement une densité puisque la surface sous la courbe au total donne 1 et que la fonction est non négative.
Pour évaluer la probabilité que la variable alétoire prenne une valeur entre 2 et 4 il faut évaluer la surface sous la courbe. La surface est un rectangle de base 2 et de hauteur $1/3$ et ainsi

$$\Pr(2 \leq X \leq 4) = 2 \times \frac{1}{3} = \frac{2}{3}$$

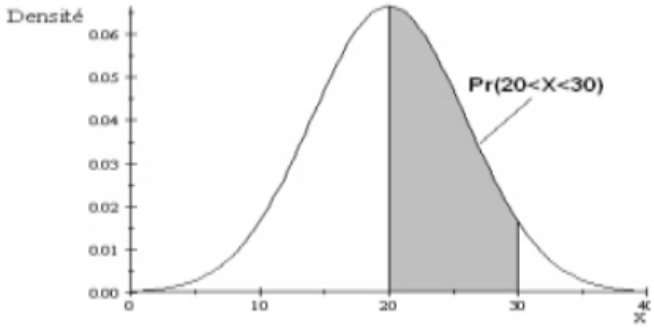
Exemple 1.1.5 *Considérons la densité donnée par la fonction suivante*:



(1.1.2)

Fig 1.1.3 : *La densité de la loi normale.*

On cherche la probabilité que la variable aléatoire soit entre 20 *et* 30. *Il sufit de calculer la surface sous la densité entre ces deux valeurs :*



(1.1.3)

Fig 1.1.4 : *Probabilité entre deux bornes.*

la fonction de répartition :

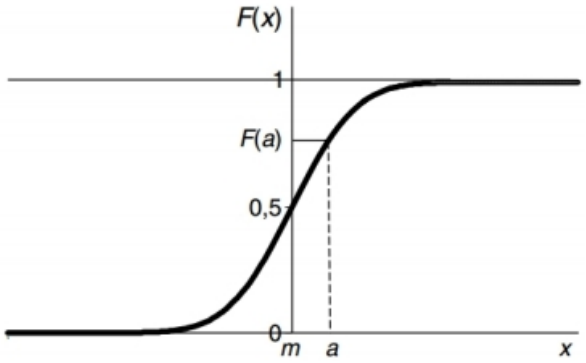
La fonction de répartition F définisse la loi de probabilité d'une variable aléatoire continue X, elle exprime la probabilité que X n'excède pas la valeur x :

$$F(x) = \Pr(X \leq x).$$

De même, la probabilité que X soit entre a et b, $(b > a)$ vaut :

$$\Pr(a < X < b) = F(b) - F(a).$$

Elles donnent lieu aux représentations graphiques suivantes :



(1.1.4)

Fig 1.1.5 : La fonction de répartition.

Les notions d'espérance et de variance sont basées sur les mêmes idées pour les variables aléatoires discrètes mais les calculs demandent l'utilisation de l'intégrale.

Définition 1.1.3 *[3]Soit X une v.a. continue de support S_X, l'espérance est donnée par*

$$\mathbb{E}(X) = \int_{S_X} x f(x).$$

C'est le centre de masse de la fonction de probabilité c'est-à-dire de la densité.

Définition 1.1.4 *[3]Soit X une v.a. continue, sa variance est définie par :*

$$\mathbb{V}(X) = \int_{S_X} (x - \mathbb{E}(X))^2 f(x).$$

1.1.5 Quelques lois de probabilité continues

Nous discuterons uniquement des lois dont nous aurons besoin dans ce travail, qui sont : Loi normale, Loi du khi-deux, Loi de Student et Loi de Fisher-Snedecor.

Loi normale ou loi de Gauss[9]

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

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

Une telle variable aléatoire est alors dite variable gaussienne. Une loi normale sera notée de la manière $\mathcal{N}(\mu, \sigma^2)$ car elle dépend de deux paramètres μ (la moyenne) et σ (l'écart-type). Ainsi si une variable aléatoire X suit $\mathcal{N}(\mu, \sigma^2)$ alors :

$$\mathbb{E}(X) = \mu \text{ et } \mathrm{Var}(X) = \sigma^2.$$

Lorsque la moyenne μ vaut 0, et l'écart-type vaut 1, la loi sera notée $\mathcal{N}(0, 1)$ et sera appelée loi *normale standard* ou la loi *normale centrée et réduite*. Seule la loi $\mathcal{N}(0, 1)$ est tabulée car les autres lois (c'est à-dire avec d'autres paramètres) se déduise de celle-ci à l'aide du théorème suivant :

Théoréme 1.1.1 *[5]Si Y suit $\mathcal{N}(\mu, \sigma^2)$ alors $Z = \frac{Y-\mu}{\sigma}$ suit $\mathcal{N}(0, 1)$.*

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

$$\Phi(x) = \Pr(Z \leq x),$$

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

$$\Phi(-x) = 1 - \Phi(x)$$

$$\Phi(0) = 0.5,\ \Phi(1.645) \approx 0.95,\ \Phi(1.960) \approx 0.9750.$$

Pour $|x| < 2$, une approximation de Φ peut être utilisée, il s'agit de son développement de Taylor à l'ordre 5 au voisinage de 0 :

$$\Phi(x) \approx 0.5 + \frac{1}{\sqrt{2\pi}}(x - \frac{x^3}{6} + \frac{x^5}{40}).$$

Inversement, à partir d'une probabilité, on peut chercher la borne pour laquelle cette probabilité est effective, on notera $z_{\alpha/2}$ nombre pour lequel :

$$\Pr(Z > z_{\alpha/2}) = \alpha/2$$

lorsque la variable aléatoire suit la loi normale standard.

α	0.01	0.02	0.05	0.10
$z_{\alpha/2}$	2.58	2.33	1.96	1.645
coefficient de sécurité c	99%	98%	95%	90%

Remarque 1.1.2 *Pour le tableau complet de la loi normale standard, [voir l'Annexe].*

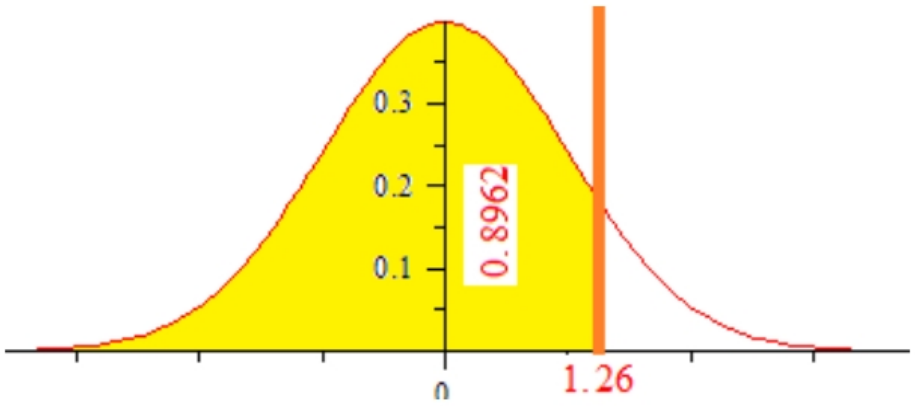
A l'aide des propriétés de la loi normale standard, on remarque que le nombre $z_{\alpha/2}$ vérifie également :

$$\Pr(Z > z_{\alpha/2}) = \Pr(Z < -z_{\alpha/2})$$

et

$$\Pr(|Z| > z_{\alpha/2}) = 1 - \Pr(-z_{\alpha/2} < Z < z_{\alpha/2})$$

Exemple 1.1.6 *On suppose que* Z *suit la loi* $\mathcal{N}(0,1)$*, et on veut trouver* $\Pr(Z \leq 1.26)$*. Puisque* 1.26 *peut s'écrire sous la forme* $1.26 = 1.20 + 0.06$*, on trouve* $\Pr(Z \leq 1.26)$ *à l'intersection de la ligne* $\ll 1.2 \gg$ *et de la colonne* $\ll 0.06 \gg$ *de la table de la loi normale standard, [voir l'Annexe]. On obtient* $\Pr(Z \leq 1.26) = \Phi(1.26) = 0.8962$*. Bref, la surface à gauche de* 1.26 *sous la densité de la loi* $\mathcal{N}(0,1)$ *est égale à* 0.8962*.*



(1.1.5)

Fig 1.1.6 : *Exemple pour l'utilisation de la table de la loi normale.*

Exemple 1.1.7 *On suppose que* Z *suit la loi* $\mathcal{N}(0,1)$ *et on veut trouver* $\Pr(Z \leq -0.94)$*. En utilisant le fait que la densité de la loi normale est symétrique et en procédant comme à l'exemple précédent, on obtient* $\Pr(Z \leq -0.94) =$ *surface à gauche de* $-0.94 =$ *surface à droite de* $0.94 = 1-$ *surface à gauche de* $0.94 = 1 - 0.8264 = 0.1736$*.*

Exemple 1.1.8 *On suppose que* X *suit la loi* $\mathcal{N}(18,4)$*, c'est-à-dire la loi normale avec moyenne* 18 *et variance* 4*, donc écart-type* 2*, et on veut trouver* $\Pr(16.72 \leq X \leq 18.94)$*. D'abord on se ramène à la loi* $\mathcal{N}(0,1)$*, puis on procède comme aux exemples précédents. On obtient :*

$$\Pr(16.72 \leq X \leq 18.94) = \Pr\left(\frac{16.72 - 18}{2} \leq Z \leq \frac{18.94 - 18}{2}\right) = 0.6808 - 0.2611 = 0.4197$$

Remarque 1.1.3 *La somme de deux variables gaussiennes indépendantes est elle-même une variable gaussienne (stabilité) :*
Soient X et Y deux variables aléatoires indépendantes suivant respectivement les lois $\mathcal{N}(\mu_1, \sigma_1^2)$ et $\mathcal{N}(\mu_2, \sigma_2^2)$. Alors, la variable aléatoire $X+Y$ suit la loi normale $\mathcal{N}(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$.

Loi du χ^2(khi-deux)[9]

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

1. La densité de la loi du $\chi^2(v)$ est $f_v(x) = \begin{cases} \frac{1}{2^{v/2}\Gamma(v/2)} x^{(v/2)-1} e^{-x/2} & \text{pour } x > 0 \\ 0 & \text{sinon} \end{cases}$.

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

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

3. La somme de deux variables aléatoires indépendantes suivant respectivement $\chi^2(v_1)$et $\chi^2(v_2)$ suit aussi une loi du χ^2avec $v_1 + v_2$ degrés de liberté.

Loi de Student[9]

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

$$T = \frac{Z}{\sqrt{\frac{Q}{v}}}$$

suit une loi appelée **loi de Student** à v degrés de liberté, notée $St(v)$.

1. La densité de la loi de Student à v degrés de liberté est

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

2. L'espérance n'est pas définie pour $v = 1$ et vaut 0 si $v \geq 2$. Sa variance n'existe pas pour $v \leq 2$ et vaut $v/(v-2)$ pour $v \geq 3$.

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

Remarque 1.1.4 *: pour $v = 1$, la loi de Student s'appelle loi de Cauchy, ou loi de Lorentz.*

Loi de Fisher-Snedecor[9]

On dit qu'une variable aléatoire X suit la loi de **Fisher-Snedecor** de paramètres $m \geq 1$ et $n \geq 1$ si elle admet une densité qui vaut :

$$f(x) = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} n^{n/2} m^{m/2} \frac{x^{(m/2)-1}}{(mx+n)^{\frac{m+n}{2}}}.$$

Signification :
La loi de Fisher-Snedecor de paramètres m et n est la loi du quotient normalisé de deux variables aléatoires qui suivent une loi du χ^2 à respectivement m et n degrés de liberté :

$$F_{m,n} = \frac{\chi^2(m)/m}{\chi^2(n)/n}.$$

1.2 Généralités sur les tests

Un test d'hypothèse est un procédé d'inférence permettant de contrôler (accepter ou rejeter) à partir de l'étude d'un ou plusieurs échantillons aléatoires, la validité d'hypothèses relatives à une ou plusieurs populations. Les méthodes de l'inférence statistique nous permettent de déterminer, avec une probabilité donnée, si les différences constatées au niveau des échantillons peuvent être imputables au hasard ou si elles sont suffisamment importantes pour signifier que les échantillons proviennent de populations vraisemblablement différentes.

1.2.1 Principe des tests

a. Méthodologie

Le principe des tests d'hypothèse est de poser une hypothèse de travail et de prédire les conséquences de cette hypothèse pour la population ou l'échantillon. On compare ces prédictions avec les observations et l'on conclut en acceptant ou en rejetant l'hypothèse de travail à partir de règles de décisions objectives.
Définir les hypothèses de travail, constitue un élément essentiel des tests d'hypothèses de même que vérifier les conditions d'application de ces dernières. Différentes étapes doivent être suivies pour tester une hypothèse :

(1) définir l'hypothèse nulle, notée H_0, à contrôler ;

(2) choisir une statistique pour contrôler H_0 ;

(3) définir la distribution de la statistique sous l'hypothèse « H_0 est réalisée» ;

(4) définir le niveau de signification du test α et la région critique associée ;

(5) calculer, à partir des données fournies par l'échantillon, la valeur de la statistique ;

(6) prendre une décision concernant l'hypothèse posée .

b. Hypothèse nulle - hypothèse alternative

L'hypothèse nulle notée H_0 est l'hypothèse que l'on désire contrôler : elle consiste à dire qu'il n'existe pas de différence entre les paramètres comparés ou que la différence observée n'est pas significative et est due aux fluctuations d'échantillonnage. Cette hypothèse est formulée dans le but d'être rejetée.
L'hypothèse alternative notée H_1 est la "négation" de H_0, elle est équivalente à dire « H_0 est fausse ». La décision de rejeter H_0 signifie que H_1 est réalisée ou H_1 est vraie.

Remarque 1.2.1 *Il existe une dissymétrie importante dans les conclusions des tests. En effet, la décision d'accepter H_0 n'est pas équivalente à « H_0 est vraie et H_1 est fausse ». Cela traduit seulement l'opinion selon laquelle, il n'y a pas d'évidence nette pour que H_0 soit fausse. Un test conduit à rejeter ou à ne pas rejeter une hypothèse nulle jamais à l'accepter d'emblée.*

La nature de H_0 détermine la façon de formuler H_1 et par conséquent la nature **unilatérale** ou **bilatérale** du test.

– On parle de test **bilatéral** lorsque l'hypothèse alternative se "décompose en deux parties". Par exemple si H_0 consiste à dire que la population estudiantine avec une fréquence de fumeurs p est représentative de la population globale avec une fréquence de fumeurs p_0, on pose alors :$H_0 : p = p_0$ et $H_1 : p \neq p_0$. Le test sera bilatéral car on considère que la fréquence p peut être supérieure ou inférieure à la fréquence p_0.

– On parle de test **unilatéral** lorsque l'hypothèse alternative se "compose d'une seule partie". Par exemple si l'on fait l'hypothèse que la fréquence de fumeurs dans la population estudiantine p est supérieure à la fréquence de fumeurs dans la population p_0, on pose alors $H_0 : p = p_0$ et $H_1 : p > p_0$. Le test sera unilatéral car on considère que la fréquence p ne peut être que supérieure à la fréquence p_0. Il aurait été possible également d'avoir :$H_0 : p = p$ et $H_1 : p < p_0$.

c. Statistique et niveau de signification

Une statistique S est une fonction des variables aléatoires représentant l'échantillon. Le choix de la statistique dépend de la nature des données, du type d'hypothèse que l'on désire contrôler, des affirmations que l'on peut admettre concernant la nature des populations étudiées. La valeur numérique de la statistique obtenue pour l'échantillon considéré permet de distinguer entre H_0 vraie et H_0 fausse.

Connaissant la loi de probabilité suivie par la statistique S sous l'hypothèse H_0, il est possible d'établir une valeur seuil, S_{seuil} de la statistique pour une probabilité donnée appelée le niveau de signification α du test. La région critique $R_c = f(S_{seuil})$ correspond à l'ensemble des valeurs telles que : $\Pr(S \in R_c) = \alpha$.
Selon la nature unilatérale ou bilatérale du test, la définition de la région critique varie.

Test	Unilatéral $H_0 : t = t_0$		Bilatéral $H_0 : t = t_0$
Hypothèse alternative	$H_1 : t > t_0$	$H_1 : t < t_0$	$H_1 : t \neq t_0$
Niveau de signification	$\Pr(S > S_{seuil}) = \alpha$	$\Pr(S < S_{seuil}) = \alpha$	$\Pr(\|S\| > S_{seuil}) = \alpha$

d. Règle de décision

Il existe deux stratégies pour prendre une décision en ce qui concerne un test d'hypothèse : la première stratégie fixe à priori la valeur du seuil de signification α et la seconde établit la valeur de la probabilité critique α_{obs} à posteriori.

Règle de décision 1 :
Sous l'hypothèse « H_0 est vraie » et pour un seuil de signification α fixé
- si la valeur de la statistique S_{obs} calculée appartient à la région critique alors l'hypothèse H_0 est rejetée au risque d'erreur α et l'hypothèse H_1 est acceptée ;
- si la valeur de la statistique S_{obs} n'appartient pas à la région critique alors l'hypothèse H_0 ne peut être rejetée.

Remarque 1.2.2 *Le choix du niveau de signification ou risque α est lié aux conséquences pratiques de la décision ; en général on choisira $\alpha = 0,05,\ 0,01$ ou $0,001$.*

Règle de décision 2 :
La probabilité critique α telle que $\Pr(S \geq S_{obs}) = \alpha_{obs}$ est évaluée
- si $\alpha_{obs} \geq \alpha$ l'hypothèse H_0 est acceptée car le risque d'erreur de rejeter H_0 alors qu'elle est vrai est trop important ;
- si $\alpha_{obs} < \alpha$ l'hypothèse H_0 est rejetée car le risque d'erreur de rejeter H_0 alors qu'elle est vrai est très faible.

1.2.2 Risques d'erreur

Définition 1.2.1 *[1]On appelle risque d'erreur de première espèce la probabilité de rejeter H_0 et d'accepter H_1 alors que H_0 est vraie. Ceci se produit si la valeur de la statistique de*

test tombe dans la région de rejet alors que l'hypothèse H_0 est vraie. La probabilité de cet évènement est le niveau de signifiation α. On dit aussi que le niveau de signification est la probabilité de rejeter l'hypothèse nulle à tort.

Remarque 1.2.3 *La valeur du risque α doit être fixée a priori par l'expérimentateur et jamais en fonction des données. C'est un compromis entre le risque de conclure à tort et la faculté de conclure. A vouloir commettre moins d'erreurs la région critique diminue lorsque α décroît, et donc on rejette moins fréquemment H_0.*

Exemple 1.2.1 *Si l'on cherche à tester l'hypothèse qu'une pièce de monnaie n'est pas « truquée », nous allons adopter la règle de décision suivante :*

$$H_0 : \textit{la pièce n'est pas truquée,}$$

acceptée si $X \in [40, 60]$, et rejetée si $X \notin [40, 60]$ donc soit $X < 40$ ou $X > 60$, avec X « nombre de faces » obtenus en lançant 100 fois la pièce. Le risque d'erreur de première espèce est :

$$\alpha = 1 - [\Pr(\mathcal{B}(100, 1/2) \in [40, 60])] .$$

Définition 1.2.2 *[1]On appelle risque d'erreur de seconde espèce, notée β la probabilité de rejeter H_1 et d'accepter H_0 alors que H_1 est vraie.*

Ceci se produit si la valeur de la statistique de test ne tombe pas dans la région de rejet alors que l'hypothèse H_1 est vraie.

Remarque 1.2.4 *Pour quantifier le risque β, il faut connaître la loi de probabilité de la statistique sous l'hypothèse H_1.*

Exemple 1.2.2 *Si l'on reprend l'exemple précédent de la pièce de monnaie, et que l'on suppose la probabilité d'obtenir face est de $0,6$ pour une pièce truquée. En adoptant toujours la même règle de décision :*

$$H_0 : \textit{la pièce n'est pas truquée,}$$

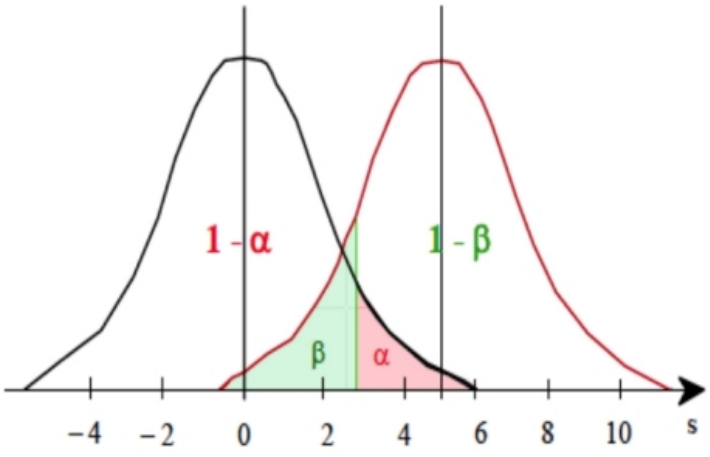
acceptée si $X \in [40, 60]$, et rejetée si $X \notin [40, 60]$ donc soit $X < 40$ ou $X > 60$, avec X « nombre de faces » obtenus en lançant 100 fois la pièce. Le risque de seconde espèce est :

$$\beta = \Pr(\mathcal{B}(100, 0, 6) \in [40, 60]).$$

1.2.3 Puissance d'un test

Rappelons que les tests ne sont pas faits pour « démontrer » H_0 mais pour « rejeter » H_0. L'aptitude d'un test à rejeter H_0 alors qu'elle est fausse constitue la puissance du test. On appelle puissance d'un test, la probabilité de rejeter H_0 et d'accepter H_1 alors que H_1 est vraie. Sa valeur est $1-\beta$. La puissance d'un test est fonction de la nature de H_1, un test unilatéral est plus puissant qu'un test bilatéral. Elle augmente avec taille de l'échantillon étudié, et diminue lorsque α diminue. La robustesse d'une technique statistique représente sa sensibilité à des écarts aux hypothèses faites.
Les différentes situations que l'on peut rencontrer dans le cadre des tests d'hypothèse sont résumées dans le tableau suivant :

Décision Réalité	H_0 vraie	H_1 vraie
H_0 acceptée	correct	manque de puissance risque de seconde espèce β
H_1 acceptée	rejet à tort risque de première espèce α	puissance du test $1-\beta$



(1.2.1)

Fig 1.2.1 : Puissance du test

Chapitre 2

Tests paramétriques

Lorsque l'on réalise des comparaisons de population ou que l'on compare une population à une valeur théorique, il existe deux grandes familles de tests : les tests paramétriques, et les tests non paramétriques.
Les tests paramétriques fonctionnent en supposant que les données que l'on a à disposition suivent un type de loi de distribution connu (en général la loi normale).
Pour calculer le risque α du test statistique, il suffit de calculer la moyenne et l'écart-type de l'échantillon afin d'accéder à la loi de distribution de l'échantillon.
La loi de distribution étant ainsi parfaitement connue, on peut calculer le risque α en se basant sur les calculs théoriques de la loi gaussienne. Ces tests sont en général très fins, mais ils nécessitent que les données suivent effectivement la loi de distribution supposée, contrairement au tests non paramétriques que nous verrons au chapitre trois.
Il existe plusieurs types des tests paramétriques, dans ce chapitre, nous nous limiterons à deux types importants : le test de conformité et le test d'homogénéité.

2.1 Test de conformité

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.
On suppose dans cette section que les échantillons sont issus d'une loi normale ou peuvent être approximés par une loi normale, et nous présenterons les trois cas les plus courants.

2.1.1 Test sur la moyenne d'une population normale

Considérons un caractère quantitatif représenté par une variable aléatoire X de loi normale $\mathcal{N}(\mu, \sigma^2)$ et un échantillon$(X_1, X_2, ..., X_n)$ de taille n de X, avec μ inconnue, et μ_0 est

une valeur donnée par l'énoncé ou provenant de connaissances théoriques. On veut tester l'hypothèse $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$.
On sait que une l'estimateur ponctuelle de μ est la moyenne empirique donnée par :

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{2.1.1}$$

Nous distinguerons ici les deux cas de la variance.

$1^{\underline{ier}}$ cas : Variance connue

Sous l'hypothèse H_0 la variable aléatoire $\overline{X_n}$ suit une loi $\mathcal{N}(\mu_0, \sigma^2/n)$ et par conséquent la statistique :

$$Z = \frac{\overline{X_n} - \mu_0}{\sqrt{\sigma^2/n}},$$

suit une loi normale centrée réduite $\mathcal{N}(0,1)$.
Si on considère le test bilatéral suivant :

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

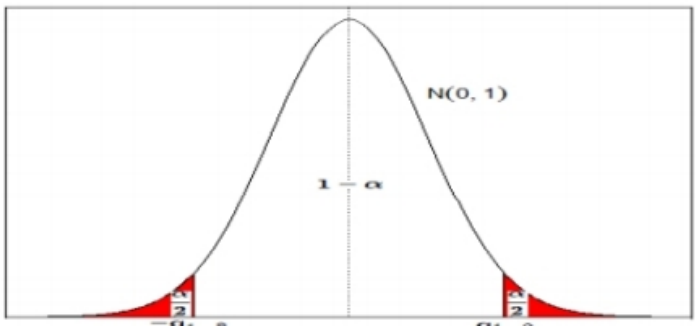
Pour un risque d'erreur α fixé on a donc :

$$\Pr(|Z| \leq q_{1-\alpha/2}) = 1 - \alpha,$$

avec $q_{1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0,1)$; et donc la région de rejet est :

$$]-\infty, -q_{1-\alpha/2}[\cup]q_{1-\alpha/2}, +\infty[.$$

On calcule alors pour les valeurs de l'échantillon Z et on accèpte ou on rejette H_0 suivant la valeur trouvée, au risque α.



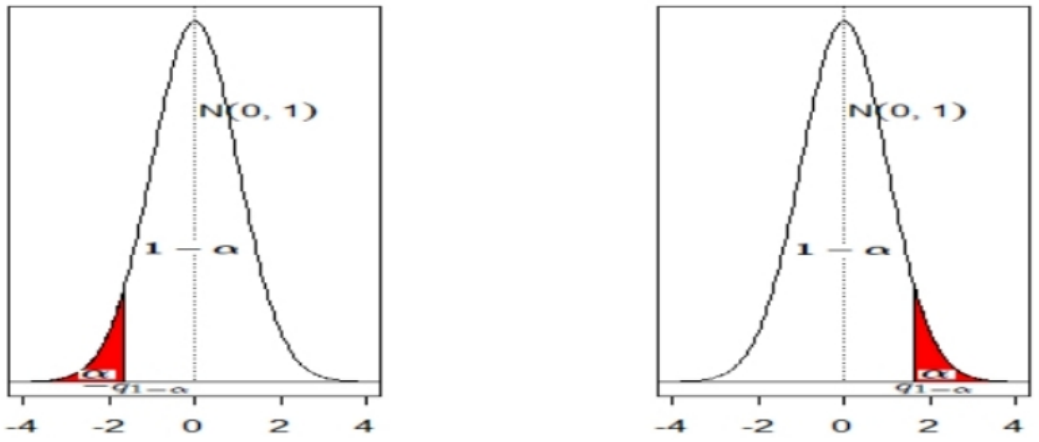
(2.1.2)

Fig 2.1.2 : Test bilatéral normal.

Si on considère un test unilatéral et une hypothèse alternative $H_1 : \mu > \mu_0$ par exemple, on obtient pour un risque d'erreur α,

$$\Pr(Z \leq q_{1-\alpha}) = 1 - \alpha$$

avec $q_{1-\alpha}$ le quantile d'ordre $1-\alpha$ de la loi $\mathcal{N}(0,1)$; et donc la région de rejet est $]q_{1-\alpha}, +\infty[$.



(2.1.3)

Fig 2.1.3 : Région d'acceptation et de rejet du test unilatéral à droite et à gauche.

$2^{\underline{ème}}$ cas : Variance inconnue

On suppose que l'on a un échantillon qui suit une loi normale $\mathcal{N}(\mu, \sigma^2)$ où la variance est maintenant inconnue. On veut tester $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$, dans le cas bilatéral. Sous l'hypothèse H_0 la variable aléatoire $\overline{X_n} = \frac{1}{n}\sum_{k=1}^{n} X_k$ suit une loi $\mathcal{N}(\mu_0, \sigma^2/n)$. Comme la variance est inconnue, on l'estime par la variance empirique :

$$\overline{S'_n}^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X_n})^2, \tag{2.1.4}$$

alors la variable :

$$T = \frac{\overline{X_n} - \mu_0}{\sqrt{\overline{S'_n}^2/n}}$$

suit une loi de Student à $n-1$ degrés de liberté, et pour un risque d'erreur α fixé on a donc :

$$\Pr(|T| \leq t_{1-\alpha/2}) = 1 - \alpha,$$

avec $t_{1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n-1$ degrés de liberté, et donc la région de rejet est :

$$]-\infty, -t_{1-\alpha/2}[\cup]t_{1-\alpha/2}, +\infty[$$

On calcule alors pour les valeurs de l'échantillon T et on accepte ou on rejette H_0 suivant la valeur trouvée, au risque α.

Théoréme 2.1.1 *[5]Lorsque le nombre d'observations est grand (supérieur à 30), on peut approcher la loi de T par la loi normale .*

Exemple 2.1.1 *Avec le médicament A, la durée moyenne de disparition de la douleur était $30mn$. On a administré le médicament B à 12 malades et on trouve la moyenne et l'écart type empiriques sont respectivement $\overline{X} = 26,75$ et $\overline{S_n} = 6,08$. Est-ce-que le médicament B est efficace que A ou non ? On réalise le test de la moyenne unilatéral :*

$$H_0 : \mu = \mu_0 \text{ contre } H_1 : \mu \geq \mu_0.$$

Au niveau de signification $\alpha = 0.05$, $T = -1.85 < 1.79 = t_{1-\alpha}$ (lue dans la table de Student), [voir l'Annexe].
Donc on accepte l'hypothèse H_0, et on conclut que le médicament B est efficace.

2.1.2 Test sur la variance d'une population normale

Pour faire ce test on se met dans les deux cas de la moyenne.

$1^{\underline{ier}}$ cas : Moyenne connue

On suppose que l'on a un échantillon qui suit une loi normale $\mathcal{N}(\mu, \sigma^2)$, où la moyenne μ est connue. On veut tester :

$$H_0 : \sigma^2 = \sigma_0^2 \text{ contre } H_1 : \sigma^2 \neq \sigma_0^2. \tag{2.1.5}$$

σ_0^2 étant une valeur donnée par l'énoncé ou provenant de connaissances théoriques.
On utilise l'estimateur de la variance $\overline{S_n}^2$ telle que :

$$\overline{S_n}^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \mu)^2. \tag{2.1.6}$$

Sous l'hypothèse H_0 la statistique :

$$V = \frac{n\overline{S_n}^2}{\sigma_0^2} = \sum_{k=1}^{n}\left(\frac{X_k - \mu}{\sigma_0}\right)^2,$$

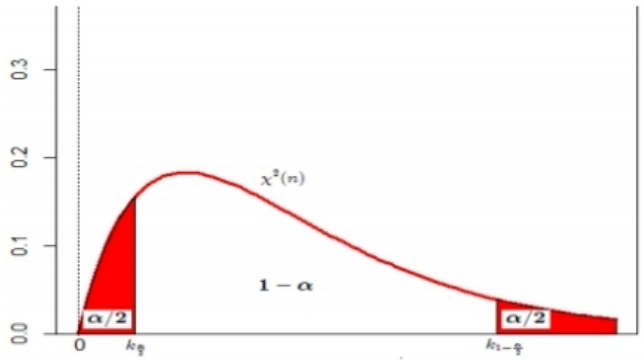
suit une loi du χ^2 à n degrés de libertés.
Si, on considère le test bilatéral, pour un risque d'erreur α fixé on a donc :

$$\Pr(\chi^2_{\alpha/2}(n) \leq V \leq \chi^2_{1-\alpha/2}(n)) = 1 - \alpha,$$

avec $\chi^2_{\alpha/2}(n)$ et $\chi^2_{1-\alpha/2}(n)$ les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ de la loi $\chi^2(n)$. Donc la région de rejet est :

$$[0, \chi^2_{\alpha/2}(n)[\cup]\chi^2_{1-\alpha/2}(n), +\infty[.$$

On calcule alors pour les valeurs de l'échantillon, V , et on accepte H_0 ou on rejette au risque α suivant la valeur trouvée.



(2.1.7)

Fig 2.1.7 : Région d'acceptation et de rejet du test bilatéral de la variance (connue).

Remarque 2.1.1 *Si on a une hypothèse alternative $H_1 : \sigma^2 > \sigma_0^2$ on fera un test unilatéral, et obtient au risque α :*

$$\Pr(V \leq \chi^2_{1-\alpha}(n)) = 1-\alpha,$$

donc la région de rejet est $]\chi^2_{1-\alpha}(n), +\infty[$.

$2^{\underline{ème}}$ cas : Moyenne inconnue

Dans ce cas on va utiliser l'estimateur $\overline{X_n}$ (2.1.1) de la moyenne μ, et l'estimateur $\overline{S'_n}^2$(2.1.4) de la variance.
Sous H_0(2.1.5) la variable de décision :

$$U = \frac{(n-1)\overline{S'_n}^2}{\sigma_0^2} = \sum_{k=1}^{n}\left(\frac{X_k - \overline{X_n}}{\sigma_0}\right)^2,$$

suit une loi du χ^2 à $n-1$ degrés de libertés. Pour un risque d'erreur α fixé on a donc :

$$\Pr\left[\chi^2_{\alpha/2}(n-1)) \leq U \leq \chi^2_{1-\alpha/2}(n-1)\right] = 1-\alpha,$$

avec $\chi^2_{\alpha/2}(n-1)$ et $\chi^2_{1-\alpha/2}(n-1)$ les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ de la loi $\chi^2(n-1)$. Donc la région de rejet est :

$$[0, \chi^2_{\alpha/2}(n-1)[\cup]\chi^2_{1-\alpha/2}(n-1), +\infty[.$$

On calcule alors pour les valeurs de l'échantillon, U et on accepte H_0 ou on rejette au risque α suivant la valeur trouvée.

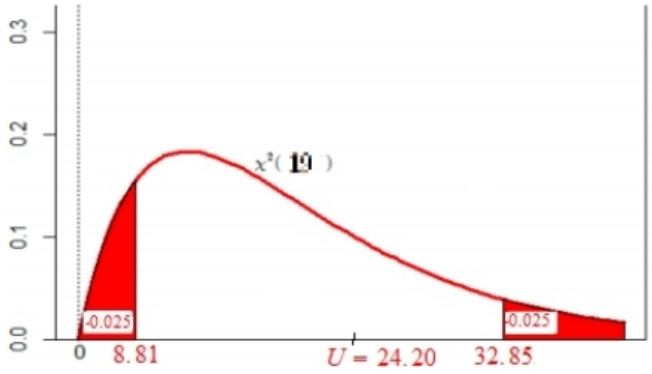
Exemple 2.1.2 *Une usine fabrique des boites de conserve de poids avec une précision* $\sigma_0^2 = 10$. *On veut savoir si la chaîne de production est déréglée (précision différente de* 10). *Voici les poids d'une série de vingt boites de conserve.*

173.6 168.8 171.1 168.8 170.0 166.8 165.6 168.1 171.5 165.1
165.4 163.5 169.9 174.4 171.8 166.0 174.6 174.5 166.4 173.8

On réalise le test de $H_0 : \sigma^2 = 10$ *contre* $H_1 : \sigma^2 \neq 10$. *Au niveau de signification* $\alpha = 0.05$. *Après le calcul on trouve :*

$$U = 24.20.$$

Sur le tableau de la loi $\chi^2(k)$ *et sur la ligne* $k = 19$ *on trouve les deux quantiles* $\chi^2_{0.025}(19) = 8.81$ *et* $\chi^2_{0.975}(19) = 32.85$. *[Voir l'Annexe]. Donc on accepte l'hypothèse* H_0, *et la chaîne de production n'est pas déréglée.*



(2.1.8)

Fig 2.1.8 : *Exemple d'utilisation du tableau de la loi Khi-deux.*

2.1.3 Test de la fréquence

On dispose d'une population dans laquelle chaque individu présente ou non un certain caractère Δ, et un échantillon aléatoire de taille n extrait de cette population.
Soit la v.a X qui prend 1 si l'individu possède le caractère Δ et 0 sinon, la proportion d'individus présentant le caracère Δ étant notée $p(0 < p < 1)$, alors la variable $\sum_{i=1}^{n} X_i$ suit la loi binomiale $\mathcal{B}(n, p)$ qu'on peut approximer, si n est assez grand, à une loi normale $\mathcal{N}(np, np(1-p))$, d'où la fréquence empirique $f = \frac{1}{n}\sum_{i=1}^{n} X_i$ calculée à partir de l'échantillon est considérée comme une réalisation d'une v.a. qui est un estimateur de p suit la loi $\mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$.

Maintenant, on se propose de comparer la proportion p par une valeur donnée p_0, c-à-d on veut tester $H_0 : p = p_0$ contre $H_1 : p \neq p_0$
Sous H_0, la statistique :

$$K = \frac{f - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \rightsquigarrow \mathcal{N}(0,1).$$

Pour α fixé et dans le cas bilaéral la région de rejet est :

$$]-\infty;\ p_0 - q_{1-\alpha/2}\frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}[\cup]\ p_0 + q_{1-\alpha/2}\frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}};\ \ +\infty[,$$

avec $q_{1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$.

Exemple 2.1.3 *On sait que la grippe touche* 30% *d'une population lord d'une épidémie. Pour tester l'efficacité d'un vaccin antigrippal, on vaccine* 300 *personnes. A la fin de la saison grippale, on dénombre* 50 *ont atteintes par la grippe. On désire tester l'efficacité du vaccin. Au niveau de signification* $\alpha = 0.05$.
On dispose le test unilatéral $H_0 : p = 0.3$ *contre* $H_1 : p < 0.3$. *Après le calcul on trouve :*

$$K = -5.12 < -1.64 = q_{0.95},$$

donc on rejette l'hypothèse H_0*, c'est à dire le vaccin est efficace.*

2.2 Test d'homogénéité

Les tests d'homogénéité destinés à comparer deux populations à l'aide d'un nombre équivalent d'échantillons (tests d'égalité ou d'homogénéité) par un paramètre comme la moyenne, la variance ou la fréquence observée, par exemple, losque l'on veut verifier l'efficacité d'un traitement médical. Dans ce cas, on considère deux groupes de patients, l'un recevant le traitement et l'autre un placebo. Si on note respectivement μ_T et μ_P les positions des populations sous traitement et sous placebo, on pose les hypothèses de test suivantes :

$$H_0 : \mu_T = \mu_P \text{ contre } H_1 : \mu_T \neq \mu_P.$$

Donc le problème consiste à comparer la position de deux échantillons.
On remarque que l'hypothèse nulle traduit toujours l'absence d'effet (c'est à dire un effet et nul), qui conduit à un test de conformités par changement de variable.
Nous introduisons ci-dessous les trois tests suivants :

2.2.1 Test d'égalite des moyennes de deux populations normales

Soient les deux échantillons $(X_1, ..., X_n)$ et $(Y_1, ..., Y_m)$ indépendants de tailles respectivement n et m issus des v.a X et Y de la loi $\mathcal{N}(\mu_X, \sigma_X^2)$ et $\mathcal{N}(\mu_Y, \sigma_Y^2)$ (respectivement), On se donne donc l'hypothèse nulle $(H_0 : \mu_X = \mu_Y)$.

Remarque 2.2.1 *Si $n = m$, le Le test se ramène à une test de conformité à une moyenne nulle de l'échantillon $(Z_1, ..., Z_n)$, avec $Z_i = X_i - Y_i$.*

$1^{\underline{ier}}$ cas. Variances connues

On suppose que les variances sont connues, et on veut tester $H_0 : \mu_X = \mu_Y$ contre $H_1 : \mu_X \neq \mu_Y$. On sait que la variable aléatoire $\overline{X_n} = \frac{1}{n}\sum_{k=1}^{n} X_k$ suit une loi $\mathcal{N}(\mu_X, \sigma_X^2/n)$ et la variable aléatoire $\overline{Y_m} = \frac{1}{m}\sum_{k=1}^{m} Y_k$ suit une loi $\mathcal{N}(\mu_Y, \sigma_Y^2/m)$, et alors sous l'hypothèse H_0 la variable aléatoire $\overline{X_n} - \overline{Y_m}$ suit une loi $\mathcal{N}(0, \sigma_X^2/n + \sigma_Y^2/m)$, par conséquent la statistique :

$$U = \frac{\overline{X_n} - \overline{Y_m}}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} \rightsquigarrow \mathcal{N}(0, 1).$$

Pour le cas bilatéral et un risque d'erreur α fixé on a donc

$$\Pr(|U| \leq q_{1-\alpha/2}) = 1 - \alpha$$

avec $q_{1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0, 1)$; et donc la région de rejet est :

$$]-\infty, -q_{1-\alpha/2}[\cup]q_{1-\alpha/2}, +\infty[.$$

On calcule U pour les valeurs de l'échantillon et on accepte ou on rejette H_0 suivant la valeur trouvée.
Si on considère un test unilatéral et une hypothèse alternative $H_1 : \mu_1 > \mu_2$ par exemple, on obtient pour un risque d'erreur α :

$$\Pr(Z \leq q_{1-\alpha}) = 1 - \alpha$$

avec $q_{1-\alpha}$ le quantile d'ordre $1 - \alpha$ de la loi $\mathcal{N}(0, 1)$; et donc la région de rejet est :

$$]q_{1-\alpha}, +\infty[$$

$2^{\underline{ème}}$ cas. Variance inconnue

Comme les variances sont inconnues, nous l'estimons par les variances empiriques corrigées :

$$\overline{S'_n}^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X_n})^2 \text{ et } \overline{S'_m}^2 = \frac{1}{m-1}\sum_{k=1}^{m}(Y_k - \overline{Y_m})^2. \tag{2.2.1}$$

On distingue les deux cas :
a) n, m sont supérieures à 30, alors sous l'hypothèse H_0 la variable aléatoire :

$$Z = \frac{\overline{X_n} - \overline{Y_m}}{\sqrt{\overline{S'_n}^2/_n + \overline{S'_m}^2/_m}},$$

approximativement suit la loi normale standard, et nous suivrons les mêmes étapes précédentes.
b) n, m sont inférieures à 30, on les estimes les variances par la variance empirique commune donnée par :

$$S_c^2 = \frac{(n-1)\overline{S'_n}^2 + (m-1)\overline{S'_m}^2}{n+m-2},$$

et alors sous l'hypothèse H_0 la variable aléatoire :

$$T = \frac{\overline{X_n} - \overline{Y_m}}{S_c\sqrt{\frac{1}{n} + \frac{1}{m}}},$$

suit une loi de Student à $n + m - 2$ degrés de liberté.

Exemple 2.2.1 *On dispose de deux échantillons de tubes, construits suivant deux procédés de fabrication A et B de tailles $n_A = 5$ et $n_B = 4$, et que les écarts-types sont $\sigma_A = 1mm$ et $\sigma_B = 0.45mm$, et les moyennes empiriques $\overline{x}_A = 52.38mm$ et $\overline{x}_B = 51.5mm$ Peut-on affirmer qu'il y a une différence significative entre les procédés de fabrication A et B ? On veut tester au niveau de signification $\alpha = 0.05$, $H_0 : \mu_A = \mu_B$ contre $H_1 : \mu_A \neq \mu_B$. Comme $U = 1.76$ et $q_{1-\alpha/2} = 1.96$, alors -$q_{1-\alpha/2} \leq U \leq q_{1-\alpha/2}$, alors on ne rejette pas l'hypothèse H_0 et on constate qu'il n'y a pas une différence entre les procédés de fabrications A et B.*

2.2.2 Test d'égalité des variances de deux populations normales

Avec les mêmes notations que précédemment on teste

$$H_0 : \sigma_X^2 = \sigma_Y^2 \quad \text{contre} \quad H_1 : \sigma_X^2 \neq \sigma_Y^2$$

On considère

$$\overline{S_n}^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \mu_X)^2 \quad \text{et } \overline{S_m}^2 \quad = \frac{1}{m}\sum_{k=1}^{m}(Y_k - \mu_Y)^2,$$

ainsi que la statistique :

$$Z = \frac{\overline{S_n}^2}{\overline{S_m}^2}.$$

Sous l'hypothèse H_0 la statistique Z suit une loi de Fisher-Snedecor $F(n,m)$ à n et m degrés de liberté, telles que :

$$\frac{n\overline{S_n}^2}{\sigma_X^2} \rightsquigarrow \chi^2(n) \text{ et } \frac{m\overline{S_m}^2}{\sigma_Y^2} \rightsquigarrow \chi^2(m),$$

Pour un risque d'erreur α fixé on a une région de rejet

$$[0, F_{\alpha/2}(n,m)[\cup]F_{1-\alpha/2}(n,m), +\infty[,$$

où les quantiles sont déterminées à l'aide de la loi précédente.

2.2.3 Test de comparaison de deux proportions

Soient $\left(X_1^{(1)}, \cdots, X_{n_1}^{(1)}\right)$ et $\left(X_1^{(2)}, \cdots, X_{n_2}^{(2)}\right)$ deux échantillons de loi Bernoulli $\mathcal{B}(p_1)$ (resp $\mathcal{B}(p_2)$), où p_1 et p_2 sont deux proportions inconnus dans $]0;1[$. On veut comparer p_1 et p_2 à partir de ces deux échantillons, c-à-d on veut tester $H_0 : p_1 = p_2$ contre $H_1 : p_1 \neq p_2$. On a $T_1 = \sum_{i=1}^{n_1} X_i^{(1)}$ $\left(\text{respectivement } T_2 = \sum_{i=1}^{n_2} X_i^{(2)}\right)$ suit une loi Binomiale $\mathcal{B}(n_1, p_1)$ (resp $\mathcal{B}(n_2, p_2)$). On suppose que les tailles d'échantillons sont sufisamment grandes pour que l'on puisse approximer la loi binomiale par la loi normale. Alors T_1 et T_2 suivent des lois normales $\mathcal{N}(n_1p_1; n_1p_1(1-p_1))$ et $\mathcal{N}(n_2p_2; n_2p_2(1-p_2))$ respectivement.
Les fréquences empiriques $f_1 = \frac{T_1}{n_1}$ et $f_2 = \frac{T_2}{n_2}$ suivent les lois $\mathcal{N}(p_1; p_1(1-p_1)/n_1)$ et $\mathcal{N}(p_2; p_2(1-p_2)/n_2)$, donc $f_1 - f_2$ suit $\mathcal{N}(p_1 - p_2; p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2)$.
Sous H_0, on peut donc poser $p = p_1 = p_2$. Alors $f_1 - f_2$ suit $\mathcal{N}(0; p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right))$, et $\frac{f_1 - f_2}{\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ suit $\mathcal{N}(0;1)$.
Comme p est inconnu, on remplace par son estimateur $\widehat{p} = \frac{T_1+T_2}{n_1+n_2}$. Donc finalement, sous H_0, la variable aléatoire :

$$U = \frac{f_1 - f_2}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \rightsquigarrow \mathcal{N}(0;1)$$

Exemple 2.2.2 *On veut tester l'impact de l'assiduité aux travaux dirigés dans la réussite à l'examen de statistique.*

	groupe 1	*groupe* 2
Nbre d'heures en TD	18 *h*	30 *h*
Nbre d'étudiants	180	150
Nbre d'étudiants ayant réussi à l'examen	126	129

Qu'en concluez-vous ?
On choisit un test unilatéral car on suppose que la réussite est meilleure avec plus d'heures

de TD. Ainsi on teste l'hypothèse : $H_0 : p_1 = p_2$ *contre* $H_1 : p_1 < p_2$.
Calculs :

$$U = \frac{f_1 - f_2}{\sqrt{\hat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = -3,45 \text{ avec } \hat{p} = 0,773$$

Décision :

Avec $\alpha = 0,05$*, la valeur théorique, lue dans la table de la loi centrée réduite, vaut* $-z_\alpha = -1,64$ *(il s'agit d'un test unilatéral). Comme* $U < -z_\alpha$*,* H_0 *est rejetée au risque d'erreur* $0,05$*. Donc le taux de réussite est significativement plus grand lorsque l'assiduité aux TD est plus élevé .*

Chapitre 3

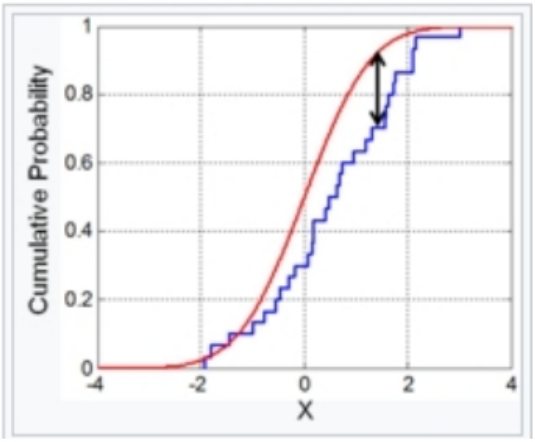
Test non paramétrique

Contrairement aux tests paramétriques qui nécessitent que les données soient issues d'une distribution paramétrée, les tests non paramétriques ne font aucune hypothèse sur la distribution sous-jacente des données. On les qualifie souvent de tests distribution free. L'étape préalable consistant à estimer les paramètres des distributions avant de procéder au test d'hypothèse proprement dit n'est plus nécessaire (test de conformité en loi). En contrepartie, ils sont moins puissants que les tests paramétriques lorsque les hypothèses sur les données peuvent être validées.
Lorsque les données sont quantitatives, les tests non paramétriques transforment les valeurs en rangs. L'appellation tests de rangs est souvent rencontrée. Lorsque les données sont qualitatives, seuls les tests non paramétriques sont utilisables.
Dans ce chapitre, nous aborderons les deux méthodes de tests non paramétriques les plus courantes : le test de **Kolmogorov-Smirnov** qui sera utilisé dans le cas où la variable aléatoire est continue avec ces deux applications : test d'ajustement et test d'homogénéité, et le test de **Khi-deux** utilisé dans le cas où la variable aléatoire quantative discrète ou qualitative, et on va voir ces applications dans l'ajustement, l'homogénéité et l'indépendance.

3.1 Test de Kolmogorov-Smirnov

En statistiques, le test de Kolmogorov-Smirnov utilisé sur une variable continue est un test d'hypothèse non paramétriques utilisés pour déterminer si un échantillon suit bien une loi donnée connue par sa fonction de répartition continue (test d'ajustement), ou bien si deux échantillons suivent la même loi (test d'homogénéité).
On modélise une loi donnée par sa fonction de répartition F (voir courbe rouge sur la figure ($Fig.$3.1)). La fonction de répartition associe à tout x (en abscisses), la probabilité qu'une variable suivant cette loi soit plus petite que x (en ordonnées). Le test consiste alors à comparer la fonction de répartition de la loi (fonction de répartition théorique) à la fonction

de répartition empirique F_n (en bleu). Cette dernière est une fonction de répartition obtenue en attribuant la probabilité $1/n$ à chacun des n éléments de l'échantillon. On mesure alors la différence entre les deux courbes. Si cette différence est inférieure à un seuil, alors on décide que l'échantillon provient bien de cette loi donnée. Plus précisément, on rejette l'hypothèse que l'échantillon provienne de la loi s'il y a une valeur de x pour laquelle $|F_n(x) - F(x)|$ est grande.



(3.1.1)

Fig 3.1.1 : Fonction de répartition empirique F_n (en bleu) et fonction de répartition théorique F (rouge).

Pour traiter ce problème, on utilisera la notion de fonction de répartition empirique.

Fonction de répartition empirique

On s'intéresse à l'estimation de la loi d'une variable aléatoire ainsi qu'aux problèmes de tests associés. Pour traiter ces questions, nous allons chercher à estimer la fonction de répartition de cette variable. Nous sommes donc confrontés à un problème de statistique non paramétrique.
On considère un n-échantillon $(X_1, ..., X_n)$ d'une variable aléatoire X. On note F la fonction de répartition de X, c'est-à-dire :

$$\forall t \in \mathbb{R},\ F(t) = \Pr(X \leq t) = \Pr(X_i \leq t).$$

C'est cette fonction F que nous allons chercher à estimer en introduisant la fonction de répartition empirique.

Définition 3.1.1 *[6]La fonction de répartition empirique associée à cet échantillon est la*

fonction :

$$\begin{aligned} \mathbb{R} &\rightarrow [0,1] \\ t &\mapsto F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq t\}}. \end{aligned}$$

Pour tout $t \in \mathbb{R}$, la variable aléatoire $nF_n(t)$ suit la loi binomiale $\mathcal{B}(n, F(t))$.

Pour représenter la fonction F_n, on introduit la statistique d'ordre $(X_{(1)}, ..., X_{(n)})$ associée à l'échantillon $(X_1, ..., X_n)$ définie par :

$$\{X_{(1)}, ..., X_{(n)}\} = \{X_1, ..., X_n\} \text{ et } X_{(1)} \leq \cdots \leq X_{(n)}.$$

On a alors :

$$\forall t \in \mathbb{R}, F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_{(i)} \leq t\}}$$

On déduit le résultat suivant.

Proposition 3.1.1 *[4]La fonction F_n est une fonction en escalier, croissante, continue à droite et admettant une limite à gauche. Elle est discontinue aux points $(X_{(i)})_{1 \leq i \leq n}$ et constante sur $[X_{(i)}, X_{(i+1)}[$ pour $i \in \{1, ..., n-1\}$.*

Remarque 3.1.1 *La fonction F_n est un estimateur naturel de F, on a pour tout $t \in \mathbb{R}$, $F_n(t)$ est un estimateur sans biais et fortement consistant de $F(t)$. Par ailleurs, on a*

$$\forall t \in \mathbb{R}\ ,\ \sqrt{n}(F_n(t) - F(t)) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}(0, F(t)(1 - F(t)).$$

Dans la suite, nous allons nous intéresser non pas à la convergence ponctuelle simple de F_n vers F mais à la convergence uniforme. Notons que la discontinuité de F_n présente des inconvénients théoriques évidents dans l'optique d'estimer F. Néanmoins, comme elle est constante par morceaux, elle est simple à construire en pratique. Maintenant nous aurons besoin du **Théorème de Glivenko-Cantelli** suivant :

Théoréme 3.1.1 *Soit $(X_i)_{i \geq 1}$ une suite de variables aléatoires de fonction de répartition F. On pose :*

$$\forall t \in \mathbb{R},\ F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq t\}}$$

Alors on a :

$$\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \to 0\ p.s.$$

Nous aborderons ici un test d'ajustement de Kolmogorov à une loi et dans le même esprit, nous construirons également un test d'homogénéité de Kolmogorov-Smirnov.

3.1.1 Test d'ajustement de Kolmogorov

La statistique introduite dans le théorème de Glivenko-Cantelli nous permettra de construire un test d'ajustement à une loi.

Proposition 3.1.2 *Soit $(X_{(1)}, ..., X_{(n)})$ un n-échantillon issu de X. On note F la fonction de répartition de X et F_n la fonction de répartition empirique. Si F est continue alors la loi de*

$$D(F_n, F) = \sup_{t\in\mathbb{R}} |F_n(t) - F(t)|$$

ne dépend pas de F.

Le résultat principal de cette section est le suivant.

Théoréme 3.1.2 *[3]On suppose que l'on a les mêmes hypothèses que ci-dessus. Alors la variable aléatoire $\sqrt{n}D(F_n, F)$ converge en loi, vers une loi limite qui ne dépend pas de F et dont la fonction de répartition est égale à :*

$$\forall t \geq 0, F_{KS}(t) = 1 - 2\sum_{k=1}^{+\infty} (-1)^k \exp(-2k^2t^2).$$

On a donc, pour $\lambda > 0$,

$$\Pr(\sqrt{n}D(F_n, F) \leq \lambda) \to 1 - 2\sum_{k=1}^{+\infty} (-1)^k \exp(-2k^2\lambda^2).$$

Nous pouvons donc construire un test d'ajustement à une loi, dit test d'ajustement de Kolmogorov. On peut d'abord remarquer que si on réordonne de manière croissante l'échantillon, $(X_{(1)}, ..., X_{(n)})$ alors $F_n(X_{(j)}) = j/n$ et

$$D(F_n, F) = \max_{1\leq j\leq n} \max\left(\left[\frac{j}{n} - F(X_{(j)})\right], \left[F(X_{(j)}) - \frac{j-1}{n}\right]\right).$$

Si on veut tester que la loi de l'échantillon a pour fonction de répartition F_0, c'est-à-dire $H_0 : F = F_0$ contre $H_1 : F = F_1$, on commence par réordonner l'échantillon, puis on calcule $D(F_n, F)$, en remarquant que sous H_0, on a $D(F_n, F) = D(F_n, F_0)$, et on cherche (dans le table de Kolmogorov, voir l'annexe) le quantile $k_{1-\alpha}$ de la loi de Kolmogorov. On accepte alors H_0 si $D(F_n, F_0) < k_{1-\alpha}$. Ce test est asymptotiquement de niveau α et sa puissance tend vers 1 quand n tend vers $+\infty$.
On peut généraliser le test de Kolmogorov au cas de deux échantillons afin de comparer leurs distributions. Le test s'appelle alors test d'homogénéité de Kolmogorov-Smirnov.

3.1.2 Test d'homogénéité de Kolmogorov-Smirnov

Dans le même esprit, nous allons construire un test d'homogénéité. On observe deux échantillons de taille respective n et m, $(X_1, ..., X_n)$ et $(Y_1, ..., Y_m)$. On veut tester si les deux échantillons sont issus d'une même loi (éventuellement inconnue). On note F la fonction de répartition de chacune des variables X_i et G la fonction de répartition de chacune des variables Y_i. On veut tester $H_0 : F = G$ contre $H_1 : F \neq G$.
Pour cela, on introduit les fonctions de répartitions empiriques :

$$\forall t \in \mathbb{R},\ F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq t\}} \text{ et } G_m(t) = \frac{1}{m}\sum_{i=1}^{n} 1_{\{Y_i \leq t\}},$$

et on utilise la distance de K-S d'homogénéité suivante :

$$D_{n,m} = \sup_{t\in\mathbb{R}} |F_n(t) - G_m(t)|$$

On a le résultat suivant.

Théoréme 3.1.3 *[5]Avec les hypothèses données ci-dessus on a, sous "$H_0 : F = G$" :*

$$\Pr\left(\sqrt{\frac{nm}{n+m}} D_{n,m} \leq \lambda\right) \to 1 - 2\sum_{k=1}^{+\infty} (-1)^k \exp(-2k^2\lambda^2).$$

Ceci permet alors de construire un test sur le même modèle que ci-dessus.

Exemple 3.1.1 *Un test a été étalonné sur une population A de manière que sa distribution suive une loi normale de moyenne* 13 *et d'écart type* 3*, sur un échantillon de taille* 10 *issu d'une population B, on a observé les valeurs suivantes :*

$$8.43, 8.70, 11.27, 12.92, 13.05, 13.05, 13.17, 13.44, 13.89, 18.90.$$

Ces valeurs sont-elles compatibles avec l'hypothèse selon laquelle la variable sous-jacente est distribuée selon une loi normale de moyenne 13 *et d'écart type* 3 *?*

On calcule, pour chaque valeur observée x_i, $F_{10}(x_i)$ et $F(x_i)$, avec :
$F_{10}(x_i) = \frac{1}{10}\sum_{k=1}^{10} \mathbf{1}_{(x_k \leq x_i)}$ et $F(x_i) = \frac{1}{3\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x_i - 13}{3})^2}$.

On calcule enfin les écarts entre $F_{10}(x_i)$ et $F(x_i)$. La valeur D est le maximum de ces écarts :

x_i	$F_{10}(x_i)$	$F(x_i)$	$Ecart-$	$Ecart+$
8.43	0.1	0.0638	0.0638	0.0362
8.7	0.2	0.0759	0.0241	0.1241
11.27	0.3	0.2821	0.0821	0.0179
12.92	0.4	0.4894	0.1894	0.0894
13.05	0.6	0.5066	0.1066	0.0934
13.17	0.7	0.5226	0.0774	0.1774
13.44	0.8	0.5583	0.1417	0.2417
13.89	0.9	0.6166	0.1834	0.2834
18.9	1	0.9754	0.0754	0.0246

On retrouve ainsi $D = 0.2834$, et pour un risque $\alpha = 0.05$ la valeur significative dans le table de Kolmogorov-sminrov (voir l'Annexe) $S = 0.369$ pour $n = 10$.
$D < S$, donc, nous acceptons que l'échantillon suit bien la loi $\mathcal{N}(13; 9)$.

3.1.3 Test de Lilliefors

Le test de **Lilliefors** est une variante tdu test de Kolmogorov-Smirnov, il a été introduit en 1967 par Hubert Lilliefors. C'est une approche non paramétrique visant à tester si une variable continue X suit une loi normale de paramètres μ et σ^2 inconnus et qui sont alors estimés par leurs contre parties empiriques $\overline{x_n}$ et $\overline{s_n}^2$ respectivement.
Statistique de Lilliefors :

$$L_n = \max_{1\leq i\leq n} \max \left\{ \left[\Phi(z_{(i)}) - \frac{i}{n} \right], \left[\Phi(z_{(i)}) - \frac{i-1}{n} \right] \right\},$$

où Φ désigne la fonction de répartition de la loi normale centrée réduite et $z_{(i)}$ est la valeur ordonnée de z_i, où $z_i = \frac{x_i - \overline{x_n}}{\overline{s_n}}$; $i = 1, 2, \cdots, n$, où $\overline{x_n}$ est la moyenne empirique et $\overline{s_n}$ est l'écart type empirique.
– La région critique : on rejette H_0 si $L_n > D_{crit}$ (D_{crit} la valeur critique de test Lilliefors).

Exemple 3.1.2 *Sur un échantillon de taille 10, on a observé les valeurs suivantes d'une VD numérique :*

$$8, 9, 9, 10, 10, 10, 11, 13, 14, 14.$$

Est-il légitime de supposer que la distribution de la VD dans la population parente suit une loi normale ?

Ici, la moyenne et l'écart type de la distribution théorique sont estimés à partir des 10 observations. Le test de Lilliefors est donc préférable au test de Kolmogorov-Smirnov.
La statistique de test se calcule de la même façon que celle du test de Kolmogorov-Smirnov. On trouve $L_{10} = 0.2451$, avec $D_{crit} = 0.258$ (la valeur critique de test Lilliefors). Donc, on accepte l"hypothèse.

3.2 Tests de Pearson (Khi-deux)

Les tests du Khi-deux sont basés sur la statistique de $\mathcal{X}^2$ proposée par **Karl Pearson**. L'objectif de ces tests est principalement de comparer des distributions entre elles. Ces tests peuvent être appliqués à des variables de nature qualitative ou quantitative.
Dans le cas où la variable aléatoire est discrète ou qualitative Le khi-deux est une statistique permettant de comparer les effectifs (fréquences) observés dans un échantillon avec des fréquences théoriques qui découlent des hypothèses statistiques. On s'intéresse dans ce module à quatre situations dans lesquelles la statistique est applicable pour effectuer un test d'hypothèse.

- Ajustement :On suppose que la loi de probabilité de la variable aléatoire qualitative (ou quantitative avec peu de modalités) est connue et on veut vérifier c'est le cas. C'est le cas classique du lancer d'un dé. On suppose que chaque face a une probabilité identique et on veut vérifier si le dé est équilibré.
- Indépendance : On mesure deux variables aléatoires qualitatives dans une population et on veut savoir si ces variables sont indépendantes c'est-à-dire si la connaissance d'une des v.a. peut influencer la loi de probabilité de l'autre. C'est le cas lorsqu'on veut vérifier si la satisfaction (en quelques catégories) par rapport au service de transport en commun est indépendant de la fréquence d'utilisation (en quelques catégories) de ces transports. Il n'y a qu'une petite nuance entre l'homogénéité et l'indépendance.
- Homogénéité : La variable aléatoire qualitative provient de k populations et on veut vérifier si la loi de probabilité est la même dans chaque population. On a donc k échantillons et on mesure la même caractéristique dans chacune d'elles. C'est le cas lorsqu'on veut savoir si la satisfaction (en quelques catégories) par rapport au service de transport en commun est semblable entre trois villes.
- Égalité de proportions : On est dans le contexte d'un test d'homogénéité mais la variable n'a que deux modalités que l'on peut qualifier de "succès" ou d'"échec" et il n'y a que deux populations. Le fait de se demander si les deux populations ont la même distribution pour la variable mesurée c'est la même chose que de vérifier si les deux proportions de succès sont identiques. Cela mérite une section particulière puisque c'est le seul test du khi-deux qui peut se décliné en unilatéral ou bilatéral. On utilise ce test lorsqu'on veut savoir si le

taux de réussite chez les hommes dans un programme d'administration est le même que le taux de réussite chez les femmes.

Statistique du test

L'idée des tests du khi deux est de comparer les valeurs observées et les valeurs moyennes qu'on observerait si l'hypothèses nulle est vraie. Considérons le cas d'un test visant à vérifier si un dé est équilibré c'est-à-dire si chacune des faces avait la même probabilité (1/6). Si on lance le dé 500 fois on devrait retrouver en moyenne $500 * 1/6 = 250/3 = 83.333$ fois la valeur "1" et 83.333 fois la valeur "2", etc. Supposons qu'on observe 90 valeur "1" sur les 500 lancers, 74 fois la valeur "2", 68 fois la valeur "3", 105 fois la valeur "4", 85 fois la valeur "5" et finalement $500 - (90 + 74 + 68 + 105 + 84) = 79$ fois la valeur "6". On cherche à établir si la différence entre les valeurs observées et les valeurs théoriques est importante ou simplement due à une variation aléatoire. Posons n_i la valeur observée pour le nombre de fois que le "i" est sorti et T_i la valeur moyenne attendue. Si on fait simplement la différence entre les deux on obtient toujours 0 :

$$\sum n_i - T_i = \sum n_i - \sum T_i = n - n = 0,$$

ce qui n'est pas particulièrement pratique. Pour éviter ça la statistique du khi deux est donnée par :

$$\frac{\sum (n_i - T_i)^2}{T_i},$$

soit la différence relative. Dans tous les cas le principe est le même, seule la formulation des fréquences théoriques diffèrent selon les hypothèses.

3.2.1 Test d'ajustement du khi-deux

Le test d'ajustement du khi-deux permet de vérifier qu'une variable qualitative ou quantitative discrète mesurée dans une population suit une loi de probabilité théorique connue. Considérons un dé à six faces et supposons que l'on veuille vérifier s'il est bien équilibré.

– On peut effectuer un test pour chaque face séparément ou utiliser la loi de probabilité de la variable aléatoire qui donne le nombre de points sur la face visible du dé. Dans ce cas il suffit de confronter les hypothèses :

$$\begin{aligned} H_0 &: \pi_i = \frac{1}{6}, \text{ pour chaque } i = 1, \cdots, 6 \\ H_1 &: \pi_i \neq \frac{1}{6}, \text{ pour au moins un } i. \end{aligned}$$

On peut tester l'ensemble des faces en une seule opération à l'aide d'un test d'ajustement du khi-deux.

– On cherche à déterminer s'il y a une différence dans le nombre de créations d'entreprises dans l'année (les saisons plus spécifiquement). Les hypothèses à confronter sont

$$\begin{aligned} H_0 \ : \quad & \pi_i = \frac{1}{4}, \text{ pour } i = \text{"été", "printemps", "automne", "hiver"} \\ H_1 \ : \quad & \pi_i \neq \frac{1}{4}, \text{ pour au moins un } i. \end{aligned}$$

où π est la probabilité de créer une entreprise.
Soit X une v.a. discrète de support S_X et loi de probabilité

$$f(x_i) = \pi_i \text{ pour } x_i \in S_X,$$

et considérons les hypothèses statistiques :

$$\begin{aligned} H_0 \ : \quad & \pi_i = \pi_{i0}, \text{ pour chaque } i \\ H_1 \ : \quad & \pi_i \neq \pi_{i0}, \text{ pour au moins un } i. \end{aligned}$$

où π_{i0} sont des constantes connues.
Le test d'ajustement du khi-deux de niveau α pour confronter ces hypothèses est de rejeter H_0 si

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

où

$$\begin{aligned} n_i &= np_i \\ T_i &= n\pi_{i0} \end{aligned}$$

et $\mathcal{X}^2_{k-1;\alpha}$ est le point critique de niveau α pour une loi khi-deux de paramètre $k-1$.
Conditions d'application : Le test approximatif est valide si :
a. $T_i \geq 1$, pour chaque i.
b. Il y a un maximun de 20% des valeurs T_i qui sont moins grandes que 5.

Remarque 3.2.1 *Les deux conditions d'application sont connues comme étant la règle de Cochran.*

Exemple 3.2.1 *Dans le but de vérifier si un dé est bien équilibré une machine "lance" le dé* 1000 *fois et on observe le nombre de points sur la face visible du dé. Les résultats sont donnés dans le tableau suivant :*

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

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

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

$$\begin{aligned} H_0 &: \pi_i = 1/6, \text{ pour chaque } i = 1, \cdots, 6 \\ H_1 &: \pi_i \neq 1/6, \text{ pour au moins un } i. \end{aligned}$$

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

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

On observe :

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

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

Exemple 3.2.2 *Une étude sur la création d'entreprises vise à vérifier s'il y a une variabilité au cours de l'année. On observe* 52 *créations d'entreprises en* 2007 *et la distribution selon les saisons est la suivante :*

Saison	*Été*	*Automne*	*Hiver*	*Printemps*
Créations	10	21	8	13

Faire un test au niveau 10% *pour vérifier s'il y a une fluctuation dans l'année.*

On veut confronter les hypothèses :

$$\begin{aligned} H_0 &: \quad \pi_i = \frac{1}{4}, \text{ pour } i = \text{"été", "printemps", "automne", "hiver"} \\ H_1 &: \quad \pi_i \neq \frac{1}{4}, \text{ pour au moins un } i. \end{aligned}$$

où π_i est la probabilité de création de l'entreprise à la saison i. Le test de niveau α est de rejeter H_0 si

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

k étant le nombre de saisons soit 4. On obtient :

$$T_i = 52 * \frac{1}{4} = 13,$$

pour chaque saison et ainsi les conditions d'application du test d'ajustement sont respectées. Selon l'échantillon on observe :

$$\begin{aligned}\mathcal{X}^2 &= \frac{(10-13)^2}{13}+\frac{(21-13)^2}{13}\\ &\quad +\frac{(8-13)^2}{13}+\frac{(13-13)^2}{13}\\ &= 7.5385\end{aligned}$$

tandis que le point critique est :

$$\mathcal{X}^2_{3;0.1} = 6.2514$$

On rejette alors H_0 au niveau 10% et on peut dire qu'il y a une différence selon les saisons.

3.2.2 Test d'indépendance pour deux variables discrètes

Lorsque deux variables dicrètes ou qualitatives sont mesurées sur les mêmes individus on est en présence d'une population et de deux mesures. Il est alors intéressant de vérifier si ces variables aléatoires sont indépendantes c'est-à-dire si elles ont une influence l'une sur l'autre. La notion même de dépendance doit être définie. Intuitivement, il y a indépendance entre deux v.a. si le fait de connaître le résultat d'une ne donne aucune information sur le résultat de la deuxième. Plus précisément, il y a indépendance entre deux v.a. X et Y si :

$$\Pr(X = x \text{ et } Y = y) = \Pr(X = x) \times \Pr(Y = y),$$

ce qui revient à dire que

$$\Pr(X = x \mid Y = y) = \Pr(X = x)$$

et

$$\Pr(Y = y|X = x) = \Pr(Y = y).$$

Les hypothèses statistiques à confronter pour X et Y deux variables aléatoires qualitatives ou quantitatives discrètes sont :

$$\begin{aligned}H_0 &: \Pr(X = x \text{ et } Y = y) = \Pr(X = x) \times \Pr(Y = y), \text{ pour tout } x,\ y\\ H_1 &: \Pr(X = x \text{ et } Y = y) \neq \Pr(X = x) \times \Pr(Y = y), \text{ pour au moins un } x,\ y\end{aligned}$$

Cette formulation de l'indépendance étant un peu rébarbative on écrit généralement les hypothèses :

$$\begin{aligned}H_0 &: X \text{ et } Y \text{ sont indépendantes}\\ H_1 &: X \text{ et } Y \text{ sont dépendantes}\end{aligned}$$

sous entendu que cela correspond à la formulation ci-haut.
Pour effectuer le test d'indépendance on utilise la statistique du khi-deux. Cette dernière est assez complexe à calculer c'est pourquoi on passe par le tableau de contingence des observations et le tableau des valeurs attendues ou théoriques. Il est alors plus facile de calculer la valeur de la statistique.

Tableau de contingence
Lorsque deux v.a. sont discrètes, il est possible de représenter les résultats d'un échantillon de taille n par un tableau de contingence :

$X \backslash Y$	mod 1	$\cdots$	mod j	$\cdots$
mod 1	n_{11}		n_{1j}	
$\vdots$				
mod i			n_{ij}	
$\vdots$				

où n_{ij} est le nombre de sujets pour lesquels la v.a. X a la modalité i et la v.a. Y a la modalité j . En plus de ces informations il est intéressant de mettre dans le tableau les marginales pour la v.a. X et la v.a. Y , c'est-à-dire les fréquences par variable aléatoire :

$X \backslash Y$	mod 1	$\cdots$	mod j	$\cdots$	
mod 1	n_{11}		n_{1j}		$n_{1.}$
$\vdots$					
mod i			n_{ij}		$n_{i.}$
$\vdots$					
	$n_{.1}$		$n_{.j}$		n

où n et la taille d'échantillon, $n_{i.}$ est la fréquence de la modalité i de la v.a. X et $n_{.j}$ est la fréqence de la modalité j de la v.a. Y . On a $n_{1.}/n$ une estimation de la probabilité que la v.a. X prenne la modalité 1, $n_{.1}/n$ une estimation de la probabilité que la v.a. Y prenne la modalité 1 et n_{11}/n une estimation de la probabilité que les v.a. X et Y prennent les modalités 1 et 1 respectivement.

Statistique du khi-deux
S'il y a indépendance on devrait avoir

$$\frac{n_{ij}}{n} \approx \frac{n_{i.}}{n} \times \frac{n_{.j}}{n}$$

Posons $T_{ij} = \frac{n_{i.} \times n_{.j}}{n}$ la fréquence attendue pour les modalités i et j s'il y avait indépendence. La statistique pour le test du khi deux est donnée par :

$$\mathcal{X}^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(n_{ij} - T_{ij})^2}{T_{ij}},$$

où k est le nombre de modalités de X et m est le nombre de modalités de Y. Cette statistique est une mesure de la dépendance entre les v.a. X et Y .
Le test d'hypothèses pour confronter

$$\begin{aligned} H_0 &: X \text{ et } Y \text{ sont indépendantes} \\ H_1 &: X \text{ et } Y \text{ sont dépendantes} \end{aligned}$$

est de rejeter H_0 si $\mathcal{X}^2 \geq \mathcal{X}^2_{(k-1)(m-1);\alpha}$, c'est-à-dire si la statistique est plus grande que le point critique de niveau α d'une loi khi deux à $(k-1)(m-1)$ degrés de liberté.

Conditions d'application : Ce test approximatif est valide si (règle de Cochran)

1. $T_{ij} \geq 1$ pour tout i et j
2. Il n'y a pas plus de 20% des valeurs T_{ij} plus petites que 5.

Exemple 3.2.3 *Pour cibler la clientèle d'un nouveau produit de consommation, une entreprise fait un sondage auprès de* 321 *personnes. L'intérêt dans le produit est noté par "aucun intérêt", "un intérêt mineur" ou un "intérêt important". La situation familiale (au moins un enfant à charge : oui ou non) est notée également. On cherche à vérifier si l'intérêt dans le produit dépend de la situation familiale. Les résultat sont les suivants :*

Enfant	*aucun*	*mineur*	*important*
oui	10	12	3
non	7	38	9

On a donc 79 *personnes qui répondent. On veut vérifier s'il y a un lien entre les deux mesures au niveau* 5%.

On cherche à confronter les hypothèses H_0 : indépendance entre la v.a. famille et intérêt dans le produit et H_1 : dépendance entre la v.a. famille et intérêt dans le produit. Le niveau est fixé à 5% . Le test est de rejeter H_0 si
$\mathcal{X}^2 = \sum_{i=1}^{3}\sum_{j=1}^{2} \frac{(n_{ij}-T_{ij})^2}{T_{ij}} \geq \mathcal{X}^2_{(k-1)(m-1);\alpha} = \mathcal{X}^2_{2;0.05} = 5.9915$.On obtient le tableau de contingence suivant :

n_{ij}	aucun	mineur	important	
oui	10	12	3	25
non	7	38	9	54
	17	50	12	79

et le tableau des fréquences théoriques :

n_{ij}	aucun	mineur	important	
oui	5.400	15.823	3.800	25
non	11.620	34.177	8.203	54
	17	50	12	79

Il y a une cellule sur 6 qui contient une valeur attendue plus petite que 5. Cela corre-spond à $1/6 * 100 = 16.667\%$ des valeurs attendues, soit moins de 20%. Le test est donc valide.
La statistique observée est :

$$\begin{aligned} \mathcal{X}^2 &= \frac{(10-5.4)^2}{5.4} + \cdots \\ &= 7.401 \end{aligned}$$

Comme la statistique est plus grande que le point critique on rejette H_0, donc il existe une dépendence entre les variables.

Exemple 3.2.4 *Un chercheur veut vérifier si deux universités ont un même barême, pour l'attribution des cotes. Pour ce faire il choisit un échantillon de* 21000 *étudiants provenant des deux université et il regarde les cotes attribuées aux étudiants de* 2001 *:*

Cote	*A*	*B*	*C*	*D*	*E*
Université I	605	1400	1789	300	70
Université II	2014	4178	8032	2005	607

En fait, on cherche à vérifier si la répartition des cotes est dépendante des universités c'est-à-dire si les variables "université" et "cote" sont des v.a. indépendantes au niveau 5%

Les hypothèses statistiques sont

$$\begin{aligned} H_0 &: \text{les v.a sont indépendantes} \\ H_1 &: \text{les v.a sont dépendantes} \end{aligned}$$

et le test du khi-deux est utilisé. Le test est de rejeter H_0 si

$$\mathcal{X}^2 = \sum_{i=1}^{3}\sum_{j=1}^{2} \frac{(n_{ij} - T_{ij})^2}{T_{ij}} \geq \mathcal{X}^2_{(k-1)(m-1);\alpha} = \mathcal{X}^2_{4;0.05} = 9.94.$$

On obtient le tableau de contingence suivant :

Cote	*A*	*B*	*C*	*D*	*E*	
Université I	605	1400	1789	300	70	4164
Université II	2014	4178	8032	2005	607	16836
	2619	5578	9821	2305	677	21000

Les fréquences théoriques sont données par :

Cote	*A*	*B*	*C*	*D*	*E*	
Université I	519.310	1106.038	1947.364	457.049	134.239	4164
Université II	2099.690	4471.962	7873.636	1847.951	542.761	16836
	2619	5578	9821	2305	677	21000

La statistique observée est 236.808.
On rejette donc H_0 au niveau 5% et on peut dire qu'il ya une dépendance.

3.2.3 Test d'homogénéite pour k populations

Considérons une variable sur une échelle nominale qui est mesurée dans k populations. Dans la population j cette variable a une loi de probabilité donnée par $f_j(x_i)$ pour chaque x_i dans le support (commun à toutes les populations). Une question intéressante est de vérifier si ces populations sont régies par la même loi de probabilité. Les hypothèses statistiques sont :

$$\begin{aligned} H_0 &: f_j(x_i) = f_l(x_i), \text{ pour tous } j, l \text{ et pour chaque } i \\ H_1 &: f_j(x_i) \neq f_l(x_i), \text{ pour un certain } (j, l) \text{ et un certain } i \end{aligned}$$

c'est-à-dire que les k lois de probabilité sont identiques contre l'hypothèse alternative qu'il y a au moins une loi de probabilité qui est différente des autres. Ce contexte est assez fréquent comme l'illustre les exemples suivants :
Le test est similaire au test du khi-deux pour deux variables dans une population (test d'indépendance) : la statistique est la même en considérant qu'il y a une variable qui indique la population mais qu'elle est fixée. On a alors le schéma suivant :

$X \backslash$POP	POP1	$\cdots$	POPj	$\cdots$	
mod 1	n_{11}		n_{1j}		$n_{1.}$
$\vdots$					
mod i			n_{ij}		$n_{i.}$
$\vdots$					
	$n_{.1}$		$n_{.j}$		n

et le test est de rejeter H_0 si

$$\mathcal{X}^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(n_{ij} - T_{ij})^2}{T_{ij}} \geq \mathcal{X}^2_{(k-1)(m-1);\alpha}$$

Les valeurs n_{ij} et T_{ij} sont telles que définies dans la subsection précédante.

Exemple 3.2.5 *On veut comparer le salaire des femmes et des hommes dans une grande entreprise. On prend un échantillon de* 120 *hommes et un échantillon de* 150 *femmes puis on note le salaire selon "faible", "moyen" et "élevé". Il y a donc* 2 *populations (hommes et femmes) puis une variable sur une échelle à tout le moins nominale (catégorie de salaire). On veut savoir si les femmes ont un salaire équivalent aux hommes avec un niveau de* 5%. *On observe*

	faible	*moyen*	*élevé*
Homme	10	70	40
Femme	30	60	60

On veut comparer les hypothèses

$$\begin{aligned} H_0 &: \text{les distributions sont identiques} \\ H_1 &: \text{les distributions sont différentes} \end{aligned}$$

avec un niveau de $\alpha = 5\%$.
Le test consiste à rejeter H_0 si $\mathcal{X}^2 \geq \mathcal{X}^2_{2;0.05} = 5.99$.
On observe $\mathcal{X}^2 = 11.58$ et on rejette donc H_0 au niveau 5% et on peut dire que les hommes et les femmes sont traités différemment au niveau du salaire.

Remarque 3.2.2 *Le test sur l'indépendance et sur l'homogénéité sont en fait identique la différence réside dans le fait que dans le test d'indépendance on mesure deux vari-ables tandis que dans le test d'homogénéité on mesure une seule variable mais dans plusieurs populations.*

3.2.4 Test sur deux proportions

Le même test du khi deux peut être utilisé pour vérifier l'égalité de deux proportions (deux échantillons)

$$\begin{aligned} H_0 &: \pi_1 = \pi_2 \\ H_1 &: \pi_1 \neq \pi_2 \end{aligned}$$

puisque si on pose X la v.a. qui donne la population (1 ou 2) et Y la v.a. qui donne succès ou échec, cela correspond aux hypothèses

$$\begin{aligned} H_0 &: \text{les v.a sont indépendantes} \\ H_1 &: \text{les v.a sont dépendantes} \end{aligned}$$

Lors d'un sondage électoral au début de la compagne, 635 personnes sur 2031 étaient en faveur d'un certain candidat tandis qu'à une semaine des élections, 327 sur 1289 étaient en faveur du même candidat. Peut-on dire à un niveau de 10% que l'opinion a changer ?
Pour répondre à cette question on veut confronter les hypothèses

$$\begin{aligned} H_0 &: \pi_1 = \pi_2 \\ H_1 &: \pi_1 \neq \pi_2 \end{aligned}$$

et le test de niveau 10% pour confronter ces hypothèses est de rejeter H_0 si
$\mathcal{X}^2 \geq \mathcal{X}^2_{1;0.1} = 2.7055$. Or on observe :

	Début	*Fin*
favorable	635	327
non favorable	1396	962

et la statistique du khi-deux telle que calculée par EXCEL est On rejette H_0 au niveau 10%. On doit donc conclure que la proportion de personne favorable au candidat a changé au cours de la campagne électorale.

Remarque 3.2.3 *Il est possible de faire un test unilatéral en utilisant une statistique du khi deux. Pour faire le test, il faut diviser le niveau de signification par* 2 *lorsqu'on utilise ce dernier pour prendre la décision ou utiliser le point critique* $\mathcal{X}^2_{(k-1)(m-1);2\alpha}$

Exemple 3.2.6 *On pense qu'il y a une proportion plus grande de personnes qui utilisent le transport en commun à Tiaret qu'à Rélizane. Pour valider cette hypothèse un échantillon de* 350 *résidents de Tiaret est constitué et sur ce nombre il y en a* 155 *qui utilisent régulièrement le transport en commun. À Rélizane, sur un échantillon de* 500 *personnes il y en a* 260 *qui utilisent régulièrement le transport en commun.*
Peut-on dire avec un niveau de 10% *que les habitants de Rélizane utilisent plus le transport en commun que les habitants de Tiaret ?*

On veut confronter les hypothèses

$$\begin{aligned} H_0 &: \pi_R = \pi_T \\ H_1 &: \pi_R > \pi_T \end{aligned}$$

où π représente la probabilité d'utiliser régulièrement les transports en commun.
Le test unilatéral de niveau 10% est de rejeter H_0 si

$$\mathcal{X}^2 > \mathcal{X}^2_{1;0.2} = 1.6424.$$

Or on observe

Transport	Tiaret	Rélizane
Oui	155	260
Non	195	240

Le tableau des fréquences théoriques est donné par :

Transport	Tiaret	Rélizane	
Oui	170.88	244.12	415
Non	179.12	255.88	435
	350	500	850

et la statistique du khi-deux $\mathcal{X}^2 = 4.9021$,On rejette H_0 et on peut dire que Rélizane utilise plus les transports en commun que Tiaret au niveau 10%.

Remarque 3.2.4 *Les tests du khi-deux permettent de faire des tests d'hypothèses sur des variables qualitatives ou quantitatives discrètes avec peu de modalités. Le principe est toujours le même : on compare les fréquences observées et les fréquences théoriques selon les hypothèses.*

Conclusion

Ce que nous avons présenté dans ce mémoire sont les bases du travail dans ce domaine et quelques méthodes importantes les plus utilisées dans la réalité appliquée, que ce soit de type paramétrique ou non paramétrique, avec quelques exemples illustratifs, alors ce travail n'est qu'une tentative étudié d'une manière objective et scientifique. L'une des perspectives de recherche que l'on peut ajouter prochainement est l'application (avec comparaison) des tests d'homogénéité sur des données réelles de notre environnement..

Annexe

1. Table de la Loi Normale Centré Réduite[11]

Soit la variable, à valeurs réelles, $X \rightsquigarrow \mathcal{N}(0;1)$, de densité f. la table donne, pour différentes valeur u positives :

$$F(u) = \Pr(X < u) = \int_{-\infty}^{u} f(x)dx$$

Exemple : $F(1.35) = \Pr(X < 1.35) = 0.91149$.

	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	.0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8254	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	.0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	6,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

Table de la loi Normale centré réduite.

2.Table de la Loi de Khi-deux[11] :

Loi du khi-deux avec k degrés de liberté

Quantiles d'ordre $1-\gamma$

	γ										
k	0.995	0.990	0.975	0.950	0.900	0.500	0.100	0.050	0.025	0.010	0.005
1	0.00	0.00	0.00	0.00	0.02	0.45	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	1.39	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	2.37	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	3.36	7.78	9.94	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	4.35	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	5.35	10.65	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	6.35	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	7.34	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	8.34	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	9.34	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	10.34	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	11.34	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	12.34	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	13.34	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.27	7.26	8.55	14.34	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	15.34	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	16.34	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.87	17.34	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.81	10.12	11.65	18.34	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	19.34	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	20.34	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	21.34	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	22.34	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	23.34	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	24.34	34.28	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	25.34	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	26.34	36.74	40.11	43.19	46.96	49.65
28	12.46	13.57	15.31	16.93	18.94	27.34	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	28.34	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	29.34	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	39.34	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	49.33	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	59.33	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	69.33	85.53	90.53	95.02	100.42	104.22
80	51.17	53.54	57.15	60.39	64.28	79.33	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	89.33	107.57	113.14	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	82.36	99.33	118.50	124.34	129.56	135.81	140.17

Si k est entre 30 et 100 mais n'est pas un multiple de 10, on utilise la table ci-haut et on fait une interpolation linéaire. Si $k > 100$ on peut, grâce au théorème limite central, approximer la loi $\chi^2(k)$ par la loi $N(k, 2k)$.

Table de la loi de khi-deux

Fractiles de la loi du χ^2 (ν)

Cette table donne les fractiles F_p de la loi de khi-deux
à ν degrés de liberté : P = Probabilité ($\chi^2 < F_p$)

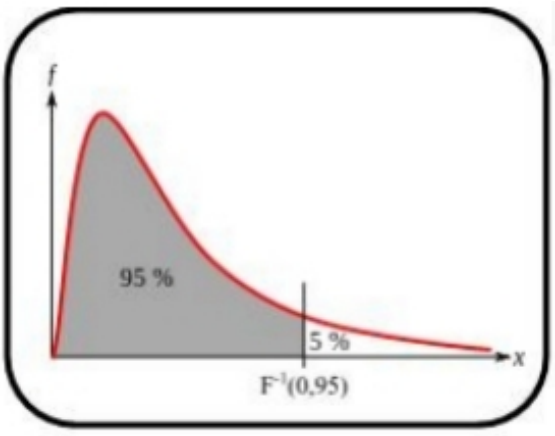
P / ν	0.01	0.02	0.05	0.1	0.15	0.2	0.25	0.3	0.4	0.5	0.6	0.7	0.75	0.8	0.85	0.9	0.95	0.98	0.99	0.9
1	0.000	0.001	0.004	0.016	0.036	0.064	0.102	0.148	0.275	0.455	0.708	1.074	1.323	1.642	2.072	2.706	3.841	5.412	6.635	10.
2	0.020	0.040	0.103	0.211	0.325	0.446	0.575	0.713	1.022	1.386	1.833	2.408	2.773	3.219	3.794	4.605	5.991	7.824	9.210	13.
3	0.115	0.185	0.352	0.584	0.798	1.005	1.213	1.424	1.869	2.366	2.946	3.665	4.108	4.642	5.317	6.251	7.815	9.837	11.345	16.
4	0.297	0.429	0.711	1.064	1.366	1.649	1.923	2.195	2.753	3.357	4.045	4.878	5.385	5.989	6.745	7.779	9.488	11.668	13.277	18.
5	0.554	0.752	1.145	1.610	1.994	2.343	2.675	3.000	3.656	4.351	5.132	6.064	6.626	7.289	8.115	9.236	11.070	13.388	15.086	20.
6	0.872	1.134	1.635	2.204	2.661	3.070	3.455	3.828	4.570	5.348	6.211	7.231	7.841	8.558	9.446	10.645	12.592	15.033	16.812	22.
7	1.239	1.564	2.167	2.833	3.358	3.822	4.255	4.671	5.493	6.346	7.283	8.383	9.037	9.803	10.748	12.017	14.067	16.622	18.475	24.
8	1.647	2.032	2.733	3.490	4.078	4.594	5.071	5.527	6.423	7.344	8.351	9.524	10.219	11.030	12.027	13.362	15.507	18.168	20.090	26.
9	2.088	2.532	3.325	4.168	4.817	5.380	5.899	6.393	7.357	8.343	9.414	10.656	11.389	12.242	13.288	14.684	16.919	19.679	21.666	27.
10	2.558	3.059	3.940	4.865	5.570	6.179	6.737	7.267	8.295	9.342	10.473	11.781	12.549	13.442	14.534	15.987	18.307	21.161	23.209	29.
11	3.053	3.609	4.575	5.578	6.336	6.989	7.584	8.148	9.237	10.341	11.530	12.899	13.701	14.631	15.767	17.275	19.675	22.618	24.725	31.
12	3.571	4.178	5.226	6.304	7.114	7.807	8.438	9.034	10.182	11.340	12.584	14.011	14.845	15.812	16.989	18.549	21.026	24.054	26.217	32.

P / ν	0.01	0.02	0.05	0.1	0.15	0.2	0.25	0.3	0.4	0.5	0.6	0.7	0.75	0.8	0.85	0.9	0.95	0.98	0.99	0.9
13	4.107	4.765	5.892	7.041	7.901	8.634	9.299	9.926	11.129	12.340	13.636	15.119	15.984	16.985	18.202	19.812	22.362	25.471	27.688	
14	4.660	5.368	6.571	7.790	8.696	9.467	10.165	10.821	12.078	13.339	14.685	16.222	17.117	18.151	19.406	21.064	23.685	26.873	29.141	
15	5.229	5.985	7.261	8.547	9.499	10.307	11.037	11.721	13.030	14.339	15.733	17.322	18.245	19.311	20.603	22.307	24.996	28.259	30.578	
16	5.812	6.614	7.962	9.312	10.309	11.152	11.912	12.624	13.983	15.338	16.780	18.418	19.369	20.465	21.793	23.542	26.296	29.633	32.000	
17	6.408	7.255	8.672	10.085	11.125	12.002	12.792	13.531	14.937	16.338	17.824	19.511	20.489	21.615	22.977	24.769	27.587	30.995	33.409	
18	7.015	7.906	9.390	10.865	11.946	12.857	13.675	14.440	15.893	17.338	18.868	20.601	21.605	22.760	24.155	25.989	28.869	32.346	34.805	
19	7.633	8.567	10.117	11.651	12.773	13.716	14.562	15.352	16.850	18.338	19.910	21.689	22.718	23.900	25.329	27.204	30.144	33.687	36.191	
20	8.260	9.237	10.851	12.443	13.604	14.578	15.452	16.266	17.809	19.337	20.951	22.775	23.828	25.038	26.498	28.412	31.410	35.020	37.566	
21	8.897	9.915	11.591	13.240	14.439	15.445	16.344	17.182	18.768	20.337	21.992	23.858	24.935	26.171	27.662	29.615	32.671	36.343	38.932	
22	9.542	10.600	12.338	14.041	15.279	16.314	17.240	18.101	19.729	21.337	23.031	24.939	26.039	27.301	28.822	30.813	33.924	37.659	40.289	
23	10.196	11.293	13.091	14.848	16.122	17.187	18.137	19.021	20.690	22.337	24.069	26.018	27.141	28.429	29.979	32.007	35.172	38.968	41.638	
24	10.856	11.992	13.848	15.659	16.969	18.062	19.037	19.943	21.652	23.337	25.106	27.096	28.241	29.553	31.132	33.196	36.415	40.270	42.980	
25	11.524	12.697	14.611	16.473	17.818	18.940	19.939	20.867	22.616	24.337	26.143	28.172	29.339	30.675	32.282	34.382	37.652	41.566	44.314	
26	12.198	13.409	15.379	17.292	18.671	19.820	20.843	21.792	23.579	25.336	27.179	29.246	30.435	31.795	33.429	35.563	38.885	42.856	45.642	
27	12.878	14.125	16.151	18.114	19.527	20.703	21.749	22.719	24.544	26.336	28.214	30.319	31.528	32.912	34.574	36.741	40.113	44.140	46.963	
28	13.565	14.847	16.928	18.939	20.386	21.588	22.657	23.647	25.509	27.336	29.249	31.391	32.620	34.027	35.715	37.916	41.337	45.419	48.278	
29	14.256	15.574	17.708	19.768	21.247	22.475	23.567	24.577	26.475	28.336	30.283	32.461	33.711	35.139	36.854	39.087	42.557	46.693	49.588	

3. Table de la Loi Fisher-Snedecor :

Soit F une variable dont la loi est celle de Fisher Snedecor de n_1, n_2 degrés de liberté, de

densité ρ , La table donne, pour divers couples (n_1, n_2), les valeurs u pour lesquelles $\Pr(F < u) = G(u) = 0.95$:



Définition du 95 centile d'une loi de Fisher-Snedecor

n_2 \ n_1	1	2	3	4	5	6	8	12	24	>25
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.08	2.64	2.53	2.38	2.20	2.00	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83	1.61	1.25
>120	3.84	2.99	2.60	2.37	2.21	2.10	1.94	1.75	1.52	1.00

Table de la Loi Fisher-Snedecor

4.Table de Student[11] :

Loi de Student avec k degrés de liberté
Quantiles d'ordre $1-\gamma$

Exemple 3.2.7 *Trouvons le quantile d'ordre* 0.975 *de la loi de Student avec* 18 *degrés de liberté. On pose* $1-\gamma = 0.975$. *On a donc* $\gamma = 1-0.975 = 0.025$. *Dans la table, le quantile d'ordre* 0.975 *de la loi de Student avec* 18 *degrés de liberté se trouve donc à l'intersection de la ligne* $\ll k = 18 \gg$ *avec la colonne* $\ll \gamma = 0.025 \gg$. *On obtient la valeur* 2.101. *Ce quantile est habituellement dénoté* $t_{18,0.025}$. *On a donc* $t_{18,0.025} = 2.101$.

	γ										
k	0.25	0.20	0.15	0.10	0.05	0.025	0.010	0.005	0.0025	0.0010	0.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792

Table de student

5. Table de Kolmogorov-sminrov pour un échantillon[11] :

N Taille de l'échantillon	Niveau de signification de D				N Taille de l'échantillon	Niveau de signification de D			
	bilatéral		unilatéral			bilatéral		unilatéral	
	0,05	0,01	0,05	0,01		0,05	0,01	0,05	0,01
5	0,565	0,669	0,509	0,627	24	0,269	0,323	0,242	0,301
6	0,521	0,618	0,468	0,577	25	0,264	0,317	0,238	0,295
7	0,486	0,577	0,436	0,538	26	0,259	0,311	0,233	0,290
8	0,457	0,543	0,410	0,507	27	0,254	0,305	0,229	0,284
9	0,432	0,514	0,388	0,480	28	0,250	0,300	0,225	0,279
10	0,410	0,490	0,369	0,457	29	0,246	0,295	0,221	0,275
11	0,391	0,468	0,352	0,437	30	0,242	0,290	0,18	0,270
12	0,375	0450	0,338	0,419	31	0,238	0,285	0,214	0,266
13	0,361	0,433	0,326	0,404	32	0,234	0,281	0,211	0,262
14	0,349	0,418	0,314	0,390	33	0,231	0,277	0,208	0,258
15	0,338	0,404	0,304	0,377	34	0,227	0,273	0,205	0,254
16	0,328	0,392	0,295	0,366	35	0,224	0,269	0,202	0,251
17	0,318	0,381	0,286	0,355	36	0,221	0,265	0,199	0,247
18	0,309	0,371	0,279	0,346	37	0,218	0,62	0,197	0,244
19	0,301	0,363	0,271	0,337	38	0,215	0,258	0,194	0,241
20	0,294	0,356	0,265	0,329	39	0,213	0,255	0,192	0,238
21	0,287	0,344	0,259	0,312	40	0,210	0,252	0,189	0,235
22	0,281	0,337	0,253	0,314	>40	1,36/	1,63/	1,22/	1,52/

6. Table de Kolmogorov-sminrov pour deux échantillons, test bilatéral et unilatéral :

Niveau de signification 5%, valeur au-dessus de la diagonale.

Niveau de signification 1%, valeur en-dessous de la diagonale.

n1 n2	5	6	7	8	9	10	11	12
5		0,800	0,800	0,750	0,778	0,800	0,709	0,717
6	1,000		0,714	0,708	0,722	0,667	0,652	0,667
7	1,000	0,857		0,714	0,667	0,657	0,623	0,631
8	0,875	0,833	0,857		0,639	0,600	0,602	0,625
9	0,889	0,883	0,778	0,764		0,589	0,596	0,583
10	0,900	0,800	0,757	0,750	0,700		0,545	0,550
11	0,818	0,818	0,766	0,727	0,707	0,700		0,545
12	0,833	0,833	0,714	0,708	0,694	0,667	0,652	
13	0,800	0,769	0,714	0,692	0,667	0,646	0,636	0,608
14	0,800	0,762	0,786	0,679	0,667	0,643	0,623	0,619
15	0,800	0,767	0,714	0,675	0,667	0,667	0,618	0,600
16	0,800	0,750	0,688	0,688	0,653	0,625	0,602	0,604
17	0,800	0,716	0,706	0,647	0,647	0,624	0,588	0,583
18	0,788	0,778	0,690	0,653	0,667	0,600	0,596	0,583
19	0,747	0,728	0,684	0,645	0,626	0,595	0,584	0,570
20	0,800	0,733	0,664	0,650	0,617	0,650	0,577	0,583

n1 n2	13	14	15	16	17	18	19	20
5	0,692	0,657	0,733	0,800	0,647	0,667	0,642	0,650
6	0,667	0,643	0,533	0,625	0,667	0,667	0,614	0,600
7	0,615	0,643	0,590	0,571	0,571	0,571	0,571	0,564
8	0,596	0,571	0,558	0,625	0,566	0,611	0,539	0,550
9	0,556	0,556	0,556	0,542	0,536	0,556	0,520	0,517
10	0,569	0,529	0,533	0,525	0,524	0,511	0,495	0,550
11	0,524	0,532	0,509	0,506	0,497	0,490	0,488	0,486
12	0,519	0,512	0,517	0,500	0,490	0,500	0,474	0,483
13		0,489	0,492	0,486	0,475	0,470	0,462	0,462
14	0,571		0,467	0,473	0,467	0,460	0,455	0,450
15	0,590	0,586		0,475	0,455	0,456	0,446	0,450
16	0,582	0,563	0,554		0,456	0,444	0,437	0,437
17	0,576	0,563	0,557	0,526		0,435	0,437	0,429
18	0,585	0,556	0,544	0,535	0,536		0,415	0,422
19	0,559	0,556	0,533	0,526	0,514	0,515		0,421
20	0,550	0,543	0,533	0,525	0,515	0,506	0,492	

Table des valeurs critiques de D_{n_1,n_2} pour le test de Kolmogorov-Smirnov à deux échantillons. **Test bilatéral**

n1 n2	5	6	7	8	9	10	11	12
5		0,800	0,714	0,675	0,667	0,700	0,636	0,600
6	1,000		0,667	0,625	0,611	0,600	0,576	0,667
7	0,857	0,833		0,607	0,571	0,571	0,571	0,547
8	0,875	0,833	0,750		0,556	0,550	0,545	0,500
9	0,800	0,778	0,746	0,750		0,556	0,525	0,528
10	0,800	0,733	0,714	0,700	0,678		0,518	0,500
11	0,800	0,742	0,714	0,693	0,636	0,627		0,485
12	0,800	0,750	0,690	0,667	0,639	0,617	0,583	
13	0,769	0,692	0,692	0,644	0,624	0,600	0,601	0,590
14	0,729	0,714	0,714	0,643	0,635	0,600	0,584	0,560
15	0,800	0,700	0,667	0,625	0,622	0,600	0,576	0,567
16	0,738	0,688	0,652	0,688	0,604	0,588	0,568	0,562
17	0,741	0,667	0,647	0,625	0,601	0,582	0,556	0,549
18	0,722	0,722	0,659	0,611	0,611	0,578	0,545	0,556
19	0,737	0,675	0,647	0,612	0,579	0,547	0,545	0,535
20	0,750	0,667	0,650	0,625	0,578	0,600	0,536	0,533

n1 n2	13	14	15	16	17	18	19	20
5	0,615	0,600	0,667	0,600	0,588	0,578	0,589	0,600
6	0,590	0,571	0,567	0,563	0,649	0,511	0,561	0,550
7	0,549	0,571	0,533	0,526	0,513	0,516	0,519	0,514
8	0,519	0,517	0,500	0,563	0,500	0,500	0,487	0,500
9	0,504	0,500	0,511	0,479	0,484	0,500	0,468	0,467
10	0,492	0,486	0,500	0,475	0,465	0,456	0,447	0,500
11	0,469	0,474	0,461	0,455	0,455	0,444	0,440	0,436
12	0,455	0,464	0,467	0,458	0,441	0,444	0,434	0,433
13		0,428	0,446	0,438	0,434	0,423	0,421	0,415
14	0,560		0,438	0,428	0,420	0,413	0,414	0,407
15	0,574	0,529		0,421	0,412	0,411	0,400	0,417
16	0,538	0,536	0,500		0,401	0,403	0,395	0,400
17	0,534	0,525	0,514	0,511		0,386	0,390	0,383
18	0,526	0,519	0,511	0,493	0,490		0,389	0,378
19	0,526	0,508	0,498	0,497	0,489	0,468		0,379
20	0,519	0,507	0,500	0,488	0,479	0,472	0,450	

Table des valeurs critiques de D_{n_1,n_2} pour le test de Kolmogorov-Smirnov à deux échantillons. **Tes unilatéral**

7. Table de Lilliefors[11] :

$n \backslash^{\alpha}$	0.01	0.05	0.10	0.15	0.20
4	0.417	0.381	0.352	0.319	0.300
5	0.405	0.337	0.315	0.299	0.285
6	0.364	0.319	0.294	0.277	0.265
7	0.348	0.300	0.276	0.258	0.247
8	0.331	0.285	0.261	0.244	0.233
9	0.311	0.271	0.249	0.233	0.223
10	0.294	0.258	0.239	0.224	0.215
11	0.284	0.249	0.230	0.217	0.206
12	0.275	0.242	0.223	0.212	0.199
13	0.268	0.234	0.214	0.202	0.190
14	0.261	0.227	0.207	0.194	0.183
15	0.257	0.220	0.201	0.187	0.177
16	0.250	0.213	0.195	0.182	0.173
17	0.245	0.206	0.189	0.177	0.169
18	0.239	0.200	0.184	0.173	0.166
19	0.235	0.195	0.179	0.169	0.163
20	0.231	0.190	0.174	0.166	0.160
25	0.203	0.180	0.165	0.153	0.149
30	0.187	0.161	0.144	0.136	0.131
OVER 30	$\frac{1.031}{\sqrt{n}}$	$\frac{0.886}{\sqrt{n}}$	$\frac{0.805}{\sqrt{n}}$	$\frac{0.768}{\sqrt{n}}$	$\frac{0.736}{\sqrt{n}}$

Bibliographie

[1] AGRESI, A. (1994). A Handbook of Small Data Sets. Journal of the American Statistical Association, 89(428), 1569-1569.

[2] AKAKPO, N. (2017). Tests statistiques. Notes de cours issues du module 4M018 Statistique Appliquée. Polycopié disponible sur la page web de l'auteur.

[3] Bagdonavicius, V , Kruopis, J., & Nikulin, M. (2011). Non-parametric tests for complete sample. ISTE-Wiley : Hoboken.

[4] DíAgostino, R. B. (1986). Goodness-of-Öt-techniques (Vol. 68). CRC press.

[5] Gaudoin, O., & B... GUIN, M. (2009). Principes et méthodes statistiques. 2ème Année, INP Grenoble.

[6] Gibbons, J. D., & Chakraborti, S. (2014). Nonparametric Statistical Inference : Revised and Expanded. CRC press.

[7] GUEDIDA, H. (2016). Test statistique de comparaison, Mémoire de master, spécialité statistique. Université de Biskra.

[8] GUESMIA, N. (2018). Tests d'ajustement à une distribution basés sur la fonction de répartition empirique, Mémoire de master, spécialité statistique. Université de Biskra.

[9] Lejeune, M. (2010). Théorie de l'estimation paramétrique ponctuelle. La théorie et ses applications ($pp.91-133$). Springer, Paris.

[10] Meraghni, Dj. (2017). Cours de première MASTER. Université Mohamed Khider de Biskra.

[11] https ://fr.wikipedia.org/wiki/Hémoglobine_glyquée.