République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Mohamed Khider, Biskra

Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie

Département de Mathématiques

Mémoire présenté pour obtenir le diplôme de

Master en **Mathématiques Appliquées**

Option : Statistique

Par Mme. TORKI Wafa Rahma

Titre :

Introduction aux tests non-paramétriques et applications

Devant le Jury :

Mr.	Brahimi Brahim	Prof.	U. Biskra	Prèsident
Mr.	Cherfaoui Mouloud	Prof.	U. Biskra	Rapporteur
Mme.	Chine Amel	MCA	U. Biskra	Examinatrice

Soutenu Publiquement le 28/06/2022

Dédicace

*J*e dédie ce modeste travail

Ma mère qui m'a initié à la vie, qui m'apprit la modestie,

L'esprit de mon défunt Père,

Mon Mari : Abdelhadi Rochdi,

Ma Belle-mére : Fatima soualhi ,

Ma soeur : Manel ,

Mes belles-soeurs : Assia et Sabrina,

Mes chers frères : Nourddine, Wassim

Toute la famille : Torki, Abdelhadi, Salhi,

Toutes mes amies,

A tous mes professeurs et tous ceux qui ont contribué à mon éducation et toute la promotion "Master" Statistique 2022.

$\mathcal{R}$emerciements

$\mathcal{T}$out d'abord je tiens à remercier Dieu de m'avoir donné le courage, la volonté et la santé pour mener à bien ce travail.

$\mathcal{J}$e remercie chaleureusement toute ma famille et mes amis pour leur soutien, leur encouragement et leur amour illimité.

$\mathcal{J}$e tiens à remercier sincèrement mon encadreur "Cherfaoui Mouloud", qui s'est toujours montrer à l'écoute tous au long de la réalisation de ce mémoire, ainsi pour son grand soutien scientifique et moral, pour les suggestions et les encouragements.

$\mathcal{J}$e remercie aussi l'ensemble des enseignants et professeurs du département de mathématiques.

$\mathcal{U}$n grand merci à mes parents pour leurs grands sacrifices et leur dévouement pour mon bonheur.

$\mathcal{E}$nfin, je remercie mes amies "Nour, Khadija, Marwa et Sofia" et toutes les personnes qui ont contribué de près ou de loin pour la réalisation de ce travail.

Rèsumè du mèmoire

Résumé : L'objectif principal de ce mémoire est d'identifier les différentes conditions d'applicabilité de certains tests paramétriques puis d'exposer leur alternative, à savoir les tests non paramétriques, en cas de non-vérification d'une ou de plusieurs des conditions en question. Enfin, dans le but d'illustrer les étapes de la mise en oeuvre de ces tests, des exemples d'application numérique ont été présentés.

Mots clés : Tests statistiques, Tests (non)paramétriques, risque de décision, la Statistique du test.

Abstract : The main objective of this thesis is to identify the different conditions of applicability of certain parametric tests then to expose their alternative, namely the non-parametric tests, in the event of non-verification of one or more of the conditions in question. Finally, in order to illustrate the steps of the implementation of these tests, examples of numerical application have been presented.

Keywords : Statistical tests, (non)parametric tests, decision risk, test statistics.

Table des matières

Table des figures

Liste des tableaux

Introduction

Pour faire la généralisation de la population à partir de l'échantillon, des tests statistiques sont utilisés. Un test statistique est une technique formelle qui s'appuie généralement sur la distribution de probabilité pour tirer la conclusion concernant le caractère raisonnable de l'hypothèse. Ces tests hypothétiques liés aux différences situations sont classés en tests paramétriques et non paramétriques.

En effet lorsque on réalise des comparaisons de population ou que Lorsque on compare une population à une valeur théorique, le but des tests paramétriques est de montrer une égalité sur certaines paramétriques, et bien sûr si les conditions d'application du test sont vérifiées. On cite comme exemple de test paramétrique : test de Student pour la moyenne, ou test de Ficher pour la comparaison de deux variances, ... Par ailleurs, les tests non paramétriques permettent sans aucune hypothèse praticulièrement sur la loi de probabilité (distribution free) et la variable aléatoire impliquée de livrer des conclusions intéressantes.

Notre objectif dans ce mémoire est d'exposé dans un premier certaines test paramétrique en faisons l'accent sur les conditions de leurs applicabilité. Par la suite, on s'intéressera aux tests non paramétriques qui peuvent être une bonne alternative aux tests paramétriques lorsque les conditions d'applicabilités de ces derniers ne sont pas toutes vérifiées.

Pour répondre à notre objectf nous avons organisé le présent mémoire comme suit :

En plus de la présente introduction, le mémoire est constitué en ordre de deux chapitres, d'une conclusion générale et d'une liste bibliographique. Où le premier chapitre est consacré à quelques tests paramétriques très usuels dans la pratique. Tandis que le deuxième chapitre, nous avons exposé un ensemble de tests non paramétrique.

Chapitre 1

Introduction à la théorie de test d'hypothèses

1.1 Introduction et la notion de tests statistiques

Les tests statistiques sont des méthodes de la statistique inférentielle qui, comme l'estimation, permettent d'analyser des données obtenues par tirages au hasard. Ils consistent à généraliser les propriétés constatées sur des observations à la population d'où ces dernières sont extraites, et à répondre à des questions concernant par exemple la nature d'une loi de probabilité, la valeur d'un paramètre ou l'indépendance de deux variables aléatoires.

Qu'est-ce qu'un test statistique ?

- Les tests statistiques font partie de la statistique inférentielle.
- Au contraire de la statistique descriptive, on utilise des lois de probabilités afin de prendre une décision dans une situation faisant intervenir une part de hasard.

- Un test statistique est une procédure de décision entre deux hypothèses. Il s'agit d'une démarche consistant à rejeter ou à ne pas rejeter une hypothèse statistique, appelée hypothèse nulle, en fonction d'un jeu de données (échantillon).

Pourquoi faire des tests statistiques ?

Les tests statistiques (ou tests d'hypothèses) ont plusieurs objectifs :

- Aider à la validation d'hypothèses
- Permettre de tirer des conclusions claires, mathématiquement rigoureuses à partir des données.
- Ils permettent de réduire la subjectivité, en rendant les choix plus objectifs et plus transparent pour pouvoir les critiquer.

Il serait important de chercher à présenter en détail l'ensemble des tests statistiques, mais la littérature est très abondante sur le sujet. Pour cela, dans ce chapitre nous allons se limiter aux tests classiques les plus simples et les plus usuels dans la pratique. En effet, Les tests présentés, concernes les tests à un seul échantillon (moyenne et variance), tests de comparaison de deux échantillons (moyenne et variance) et enfin l'analyse de la variance à un seul facteur....

1.2 Hypothèses de test

Dans ce passage nous énonçons (ou rappelons) un certain nombre de généralités autour des tests d'hypothèses, l'objectif étant d'être capable de bien formuler un test.

- Une hypothèse statistique est un énoncé (une affirmation) concernant les caractéristiques d'une population.
- L'hypothèse que nous voulons vérifier sera appelée hypothèse nulle et est notée

H_0. Par exemple, l'hypothèse selon laquelle on fixe à priori un paramètre de la population à une valeur particulière.

- N'importe quelle autre hypothèse qui diffère de l'hypothèse H0 s'appelle l'hypothèse alternative(ou contre-hypothèse) et est notée H_1.

On distingue deux types d'hypothèses :

- L'hypothèse nulle : $H_0 : \theta = \theta_0$ où θ est un paramètre inconnu.
- L'hypothèse alternative (ou contre hypothèse) : H_1, qui peut prendre l'une des formes suivantes :

  $H_1 : \theta = \theta_0$ (test bilateral).

  $H_1 : \theta < \theta_0$ (test unilateral a gauche).

  $H_1 : \theta > \theta_0$ (test unilateral a droite).

Erreurs de décision

Il est à noter que, l'aspect aléatoire de l'échantillon (observations) peut nous faussé la décision finale (rejeter ou non l'hypothèse H_0). On effet, lorsque on rejette H_0 alors que H_0 est vraie, on commet une erreur. On a donc une probabilité α (car lorsque H_0 est vraie, on a $P\left(U \notin]-\mu_\alpha, \mu_\alpha[\right) = \alpha$) de se tromper : α est appelée ***erreur de première espèce***.

Une autre situation où on peut commettre une erreur de décision est bien que celle lorsque on ne rejette pas H_0 alors que H_0 est fausse. Dans ce cas, on a une probabilité β de se tromper : β est appelée ***erreur de deuxième espèce***. Cette probabilité est difficilement calculable car dans la plupart des temps, on ne connaît pas la loi de U lorsque H_0 est fausse. La valeur $1 - \beta$ est appelée la ***puissance du test***.

Finalement, ces déférentes situations peuvent être résumées par le schéma suivant :

		Réalité	
		H_0	H_1
Décision	H_0	$1-\alpha$	α
	H_1	β	$1-\beta$

Les deux risques peuvent se de finir ainsi :

- α=P(rejeter H_0 sachant que H_0 est vrai) : probabilité de commettre une erreur de première espèce (probabilité de conclure à une différence alors qu'il n'y en a pas),
- β=P(ne pas rejeter H_0 sachant que H_1 est vrai) : probabilité de commettre une erreur de deuxième espèce (c'est la probabilité de ne pas conclure à une différence alors qu'il y en a),
- $P = 1-\beta$: Puissance statistique du test (probabilité à conclure à une difference alors qu'il y en a).

Les étapes d'un test

1. Construire les hypothèses.
2. Considérer un échantillon de taille n.
3. Déterminer les risques d'erreur.
4. Choisir le test adapté : chaque test a ses conditions d'application.
5. Calculer le P grâce au test et l'interpréter.

Dans un problème statistique, chaque situation correspond à un test approprié. Un test est soit paramétrique soit non paramétrique. Un test est soit libre d'une distribution soit non lié à une distribution. Chaque test possède une efficacité est une robustesse. Il faut utiliser le bon test à la bonne place.

Tests paramétriques Les tests paramétriques sont des tests statistiques qui se basent sur des distributions statistiques supposées dans les données. Par conséquent, certaines conditions de validité doivent être vérifiées pour que le résultat d'un test paramétrique soit fiable. Par exemple, le test t de Student pour deux échantillons indépendants n'est fiable que si les données associées à chaque échantillon suivent une distribution normale et si les variances des échantillons sont homogènes (pour plus de détail sur ce test, voir la suite du présent chapitre).

1.3 Tests de conformité pour une moyenne

Considérons un caractère quantitatif représenté par une variable aléatoire X d'espérance mathématique μ, d'écart-type σ, et un échantillon $X_1,\ X_2, ...,\ X_n$ de taille n de X. La moyenne et la variance corrigée d'échantillon sont données respectivement par :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ et } \hat{\sigma}_c^2 = \frac{n}{n-1}\hat{\sigma}^2, \text{ avec } \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2.$$

1.3.1 Cas d'un petit échantillon gaussien

Dans ce test deux cas sont envisageable. En effet, on peut distinguer le cas où l'écart-type est une quantité bien connue et le cas où l'écart-type n'est connue qu'approximativement à travers son estimateur.

Cas σ connu

Il s'agit de faire un choix entre plusieurs hypothèses possibles sur μ sans disposer d'informations suffisantes pour que ce choix soit sûr. On met en avant deux

hypothèses privilégiées : l'hypothèse nulle H_0 et l'hypothèse alternative H_1. Par exemple, on testera

$$H_0 : \ ''\mu = \mu_0'' \text{ contre } H_1 : \ ''\mu \neq \mu_0'',$$

avec μ_0 fixé arbitrairement. On veut savoir si l'on doit rejeter H_0 ou pas.

La résolution du présent problème consiste, en résumé, à réaliser les étapes suivantes :

1. Utilise une variable aléatoire dont on connaît la loi de probabilité lorsque H_0 est vraie. Par exemple, on prend $U = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$, en raison que lorsque H_0 est vraie, $U = \frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$ suit la loi $N(0,1)$, et cela le fait que l'échantillon est issue d'une variable aléatoire d'une loi normale $X \rightsquigarrow N(\mu, \sigma^2)$.

2. Fixe une valeur $\alpha \in]0,1[$. En général, on prend α (le risque) petit, le plus souvent
$$\alpha \in \{0.10, 0.05,\ 0.01,\ 0.01,\ 0.001\}.$$

3. Quantifier un réel u_α, tel que $P(-u_\alpha < U < u_\alpha) = 1 - \alpha$. Ce réel u_α peut être extrait de la table de la loi normale centrée et réduite.

4. Comparer la moyenne empirique $\overline{X}$ de l'échantillon à la moyenne théorique $\mu = \mu_0$, sachant que l'hypothèse H_0 signifiera que les différences observées sont seulement dues aux fluctuations d'échantillonnage (i.e. ne sont pas significatives). En fin, on décide ce qui suit :
   - On ne rejettera pas H_0 si les différences observées ne sont pas significatives, c'est-à-dire si U est "petite", ce que l'on peut formuler par $-u_\alpha < U < u_\alpha$, ou encore $|U| < u_\alpha$.
   - On rejettera H_0 si les différences observées sont significatives, ce que l'on

peut formuler par $U < -u_\alpha$ ou $U > u_\alpha$, c'est-à-dire $|U| > u_\alpha$. Par construction de u_α, on a $P(U > u_\alpha) = P(U < -u_\alpha) = \frac{\alpha}{2}$, soit encore $P(|U| > u_\alpha) = \alpha$ i.e. $P(U \notin]-u_\alpha, u_\alpha[) = \alpha$.

En pratique, on calcule $u = \frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$ et on décide

- de rejeter H_0 si $u \notin]-\mu_\alpha, \mu_\alpha[$, car si H_0 était vraie, l'événement $U \notin]-\mu_\alpha, \mu_\alpha[$ aurait une probabilité forte de se réaliser ; on pourra dire que la valeur observée $\overline{X}$ n'est pas conforme à la valeur théorique μ_0 mais on ne pourra pas donner de valeurs acceptable de μ ;
- de ne pas rejeter H_0 si $u \in]-\mu_\alpha, \mu_\alpha[$, car si H_0 était vraie, l'événement $U \notin]-\mu_\alpha, \mu_\alpha[$ aurait une probabilité faible de se réaliser ; on pourra dire que la valeur observée $\overline{X}$ est conforme à la valeur théorique μ_0 et que la valeur μ_0 ne peut être rejeter.

Attention : d'autres valeurs μ_0', μ_0'', ... peuvent également convenir.

Les différents tests usuels (formulation et décision) correspondant à la présente situation peuvent être résumer comme suit :

Test (bilatéral) $H_0 : \ ''\mu = \mu_0''$ **contre** $H_1 : \ ''\mu \neq \mu_0''$,

On calcule $u = \frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$. On détermine u_α, à partir de la table de la loi normale, tel que $P(-u_\alpha < U < u_\alpha) = 1 - \alpha$, et on décide que :

- Si $u \in]-u_\alpha, u_\alpha[$, alors on ne peut rejeter H_0 ;
- Si $u \notin]-u_\alpha, u_\alpha[$, alors on rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) $H_0 : \ ''\mu = \mu_0''$ **contre** $H_1 : \ ''\mu > \mu_0''$,

On calcule $u = \frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$. On détermine u_α, à partir de la table de la loi normale, tel que $P(U \geq u_\alpha) = 1 - \alpha$ et on décide que :

- Si $u < u_\alpha$, alors on ne peut rejeter H_0 ;
- Si $u \geq u_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) $H_0 : \ ''\mu = \mu_0''$ **contre** $H_1 : \ ''\mu < \mu_0''$,

On calcule $u = \frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$. On détermine u_α, à partir de la table de la loi normale, tel que $P(U < u_\alpha) = 1 - \alpha$ et on décide que :

- Si $u > -u_\alpha$, alors on ne peut rejeter H_0 ;
- Si $u \leq -u_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

Cas σ inconnu

Par définition, on sait que $T = \frac{\overline{X}-\mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$ suit la loi de Student à $n-1$ degrés de liberté. Alors, les différents tests précédents (bilatéral et unilatéral) se font comme suit :

Test (bilatéral) $H_0 : \ ''\mu = \mu_0''$ **contre** $H_1 : \ ''\mu \neq \mu_0''$,

On calcule $t = \frac{\overline{X}-\mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$. On détermine t_α sur la table de Student pour un degré de liberté $n-1$ tel que $P(-t_\alpha < T < t_\alpha) = 1 - \alpha$ et on décide que :

- Si $t \in]-t_\alpha, t_\alpha[$, alors on ne peut rejeter H_0 ;
- Si $t \notin]-t_\alpha, t_\alpha[$, alors on rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) $H_0 : \ ''\mu = \mu_0''$ **contre** $H_1 : \ ''\mu > \mu_0''$,

On calcule $t = \frac{\overline{X}-\mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$. On détermine t_α tel que $P(T \geq t_\alpha) = 1 - \alpha$ et on décide que :

- Si $t < t_\alpha$, alors on ne peut rejeter H_0 ;
- Si $t \geq t_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) $H_0 : \ ''\mu = \mu_0''$ **contre** $H_1 : \ ''\mu < \mu_0''$,

On calcule $t = \frac{\overline{X}-\mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$. On détermine t_α tel que $P(T < t_\alpha) = 1 - \alpha$ et on décide que :

- Si $t > -t_\alpha$, alors on ne peut rejeter H_0 ;
- Si $t \leq -t_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

1.3.2 Cas d'un grand échantillon

Dans cette situation ($n > 30$), on se basons sur le TCL, on sait que la variable aléatoire $U = \frac{\overline{X}-\mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$ suit approximativement une loi normale centrée et réduite ($U \rightsquigarrow N(0,1)$).

Test (bilatéral) $H_0 : \ ''\mu = \mu_0''$ **contre** $H_1 : \ ''\mu \neq \mu_0''$,

On calcule $u = \frac{\overline{X}-\mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$. On détermine u_α tel que $P(-u_\alpha < U < u_\alpha) = 1-\alpha$, et on décide que :

- Si $u \in]-u_\alpha, u_\alpha[$, alors on ne peut rejeter H_0 ;
- Si $u \notin]-u_\alpha, u_\alpha[$, alors on rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) $H_0 : \ ''\mu = \mu_0''$ **contre** $H_1 : \ ''\mu > \mu_0''$,

On calcule $u = \frac{\overline{X}-\mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$. On détermine u_α tel que $P(U \geq u_\alpha) = 1-\alpha$ et on décide que :

- Si $u < u_\alpha$, alors on ne peut rejeter H_0 ;
- Si $u \geq u_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) $H_0 : \ ''\mu = \mu_0''$ **contre** $H_1 : \ ''\mu < \mu_0''$,

On calcule $u = \frac{\overline{X}-\mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$. On détermine u_α tel que $P(U < u_\alpha) = 1-\alpha$ et on décide que :

- Si $u > -u_\alpha$, alors on ne peut rejeter H_0 ;
- Si $u \leq -u_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

1.4 Test de Khi-deux pour une variance

Considérons un caractère quantitatif représenté par une variable aléatoire X de loi normale $N(\mu, \sigma^2)$ et un échantillon $X_1,\ X_2, ...,\ X_n$ de taille n de X. La moyenne de l'échantillon est $\overline{X}$ et sa variance corrigée est $\hat{\sigma}_c^2 = \frac{n}{n-1}\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

Par définition la variable aléatoire définie par $Y^2 = \frac{(n-1)\hat{\sigma}_c^2}{\sigma^2}$ suit la loi de $Khi-2$ à $n-1$ degrés de liberté.

Les différents tests simples de conformité de la variance (formulations et décisions) sont résumés dans ce qui suit :

Test (bilatéral) H_0 : "$\sigma^2 = \sigma_0^2$" **contre** H_1 : "$\sigma^2 \neq \sigma_0^2$"

On calcule $y^2 = \frac{n-1}{\sigma^2}\hat{\sigma}_c^2$, on détermine a_α et b_α tel que $P(Y^2 \geq a_\alpha) = 1 - \frac{\alpha}{2}$ et

$P(Y^2 \geq b_\alpha) = \frac{\alpha}{2}$ et ensuite on décide de :

- ne peut rejeter H_0 si $y^2 \in]a_\alpha, b_\alpha[$;
- rejette H_0 avec une probabilité α de se tromper si $y^2 \notin]a_\alpha, b_\alpha[$.

Test (unilatéral) H_0 : "$\sigma^2 = \sigma_0^2$" **contre** H_1 : "$\sigma^2 > \sigma_0^2$"

On calcule $y^2 = \frac{n-1}{\sigma^2}\hat{\sigma}_c^2$, on détermine b_α tel que $P(Y^2 \geq b_\alpha) = \alpha$ et on décide :

- Si $y^2 < b_\alpha$, de ne peut rejeter H_0 ;
- Si $y^2 \geq b_\alpha$, de rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) H_0 : "$\sigma^2 = \sigma_0^2$" **contre** H_1 : "$\sigma^2 < \sigma_0^2$"

On calcule $y^2 = \frac{n-1}{\sigma^2}\hat{\sigma}_c^2$, on détermine a_α tel que $P(Y^2 \geq a_\alpha) = 1 - \alpha$ et on décide :

- Si $y^2 > a_\alpha$, de ne peut rejeter H_0 ;
- Si $y^2 \leq a_\alpha$, de rejette H_0 avec une probabilité α de se tromper.

Dans les différents tests présenté dans les sections précédentes on n'a considéré qu'un seul échantillon, pour lequel on s'intéresse si l'un de ses caractères (moyenne, variance) est conforme à une quantité fixée arbitrairement (cette dernière quantité représente généralement une norme du phénomène étudié). Cependant, dans la pratique, il est possible de disposer de deux populations P_1 et P_2 ou voir même

plus de deux populations, dont on étudie un même caractère et pour lesquels on désir comparer ces populations quant à ce caractère, et donc à savoir si elles sont homogènes ou non. Dans la section suivante, nous allons présenter quelques tests paramétriques dont l'objectif est de test l'homogénéité de variance et de moyennes de deux populations indépendantes. Par la suite, nous considérons l'ANOVA à un seul facteur qui nous permet de vérifier l'homogénéité de plusieurs populations (> 2) vis à vis leurs moyennes.

1.5 Comparaison de deux variances

Soient X et Y deux variables aléatoires indépendantes représentant le même caractère quantitative dans chacune des populations P_1 et P_2. On suppose que X et Y suivent des lois normales respectivement, $N(\mu_1; \sigma_1^2)$ et $N(\mu_2; \sigma_2^2)$.

De P_1, on extrait un échantillon $X_1,\ X_2, ...,\ X_{n_1}$ de taille n_1 de X et de P_2, on extrait un échantillon $Y_1,\ Y_2, ...,\ Y_{n_2}$ de taille n_2 de Y.

Les moyennes empiriques des deux échantillons sont alors

$$\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \qquad \overline{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i;$$

et leurs variances corrigées sont :

$$\hat{\sigma}_{c,1}^2 = \frac{n_1}{n_1 - 1}\hat{\sigma}_1^2 \text{ avec } \hat{\sigma}_1^2 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i^2 - \overline{X}^2,$$

$$\hat{\sigma}_{c,2}^2 = \frac{n_2}{n_2 - 1}\hat{\sigma}_2^2 \text{ avec } \hat{\sigma}_2^2 = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i^2 - \overline{Y}^2.$$

On veut réaliser le test bilatéral suivant :

$$H_0 : \text{"}\sigma_1^2 = \sigma_2^2\text{" contre } H_1 : \text{"}\sigma_1^2 \neq \sigma_2^2\text{".}$$

Les étapes de la réalisation de ce test peuvent être résumées comme suit :

1. On calcule la réalisation $f_c = \frac{\hat{\sigma}_{c,1}^2}{\hat{\sigma}_{c,2}^2}$. Si nécessaire, on permute les échantillons de sorte que $f_c \geq 1$ $\left(\text{c'est-à-dire } f_c = \frac{\max\{\sigma_{c,1}^2, \sigma_{c,2}^2\}}{\min\{\sigma_{c,1}^2, \sigma_{c,2}^2\}}\right)$.
2. Sachant que sous l'hypothèse H_0, la statistique (variable aléatoire) $F = \frac{\hat{\sigma}_{c,1}^2}{\hat{\sigma}_{c,2}^2}$ suit une loi de Fisher à $(n_1 - 1; n_2 - 1)$ degrés de liberté, alors à partir de la table de Fisher on détermine f_α tel que :

$$P(F \geq f_\alpha) = \frac{\alpha}{2}(\text{ ou encore } P(F \leq f_\alpha) = 1 - \frac{\alpha}{2}).$$

3. La règle de décision se fait comme suite :
   - Si $f_c < f_\alpha$, alors on ne peut rejeter H_0 (H_0 est vraie).
   - Si $f_c \geq f_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

Avec le même raisonnement on va trouver la zone de non rejet de l'hypothèse nulle dans les tests unilatéral. Les résultats des différents tests sont résumés dans le tableau suivant :

Hypothèses	Zone de non-rejet H_0
$H_0 : \text{"}\sigma_1^2 = \sigma_2^2\text{"}$ contre $H_1 : \text{"}\sigma_1^2 \neq \sigma_2^2\text{"}$	$[1; f(n_1 - 1, n_2 - 1, 1 - \frac{\alpha}{2})]$
$H_0 : \text{"}\sigma_1^2 = \sigma_2^2\text{"}$ contre $H_1 : \text{"}\sigma_1^2 > \sigma_2^2\text{"}$	$[1; f(n_1 - 1, n_2 - 1, 1 - \alpha)]$
$H_0 : \text{"}\sigma_1^2 = \sigma_2^2\text{"}$ contre $H_1 : \text{"}\sigma_1^2 < \sigma_2^2\text{"}$	$[1; f(n_2 - 1, n_1 - 1, 1 - \alpha)]$, avec $f_c = \frac{\hat{\sigma}_{c,2}^2}{\hat{\sigma}_{c,1}^2}$

tel que $f(n, m, 1-\alpha)$ est lu dans la table de loi Fisher-Snedecor $(1-\alpha)$ à colonne n, ligne m, de plus on ne rejettera pas H_0 si f_c appartient à la zone de non-rejet de H_0 et on rejettera H_0 sinon.

1.6 Test de Student pour deux échantillons

Dans cette section, nous allons intéresser à l'homogénéité de deux populations par rapport à la moyenne. Notons que, le test de comparaison de deux moyennes dépend de la distribution des échantillons dont on dispose. Dans le cadre de ce document, nous allons se focalisé sur le cas où les deux échantillons sont de grand taille issues d'une loi quelconque et le cas où les deux échantillons sont gaussien et de petite taille.

1.6.1 Cas des grands échantillons

Soient X et Y des variables aléatoires indépendantes représentant le caractère qualitative étudié dans chaque population. On suppose que X et Y suivent une loi quelconque de moyennes respectives μ_1 et μ_2 et d'écart-types respectifs σ_1 et σ_2. On extrait un échantillon $X_1,\ X_2, ...,\ X_{n_1}$ de taille $n_1 > 30$ de X et un échantillon $Y_1,\ Y_2, ...,\ Y_{n_2}$ de taille $n_2 > 30$ de Y.

Soit la statistique

$$U = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \tag{1.1}$$

et u sa réalisation.

Test (bilatéral) $H_0 :\ ''\mu_1 = \mu_2''$ **contre** $H_1 :\ ''\mu_1 \neq \mu_2''$,

Sous l'hypothèse H_0, la statistique U définie par (1.1) suit approximativement la loi normale centrée réduite $N(0, 1)$.

On calcule $u = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{\acute{\sigma}_{c,1}^2}{n_1} + \frac{\acute{\sigma}_{c,2}^2}{n_2}}}$, et on détermine u_α, sur la table de la loi normale, tel que :

$$P(-u_\alpha < U < u_\alpha) = 1 - \alpha,$$

c'est-à-dire

$$P(U < u_\alpha) = 1 - \frac{\alpha}{2},$$

et on décide de :

- Ne pas rejeter H_0 si $u \in]-u_\alpha, u_\alpha[$;
- Rejeter H_0, avec une probabilité α de se tromper, si $u \notin]-u_\alpha, u_\alpha[$.

Test (unilatéral) de $H_0 :\ ''\mu_1 = \mu_2''$ **contre** $H_1 :\ ''\mu_1 > \mu_2''$,

Sous l'hypothèse H_0, la statistique U suit approximativement la loi normale $N(0,1)$.

On calcule $u = \frac{\overline{x}-\overline{y}}{\sqrt{\frac{\hat{\sigma}^2_{c,1}}{n_1}+\frac{\hat{\sigma}^2_{c,2}}{n_2}}}$, et on détermine u_α, sur la table de la loi normale, tel que : $P(U \geq u_\alpha) = 1 - \alpha$ et on décide de :

- Ne pas rejeter H_0 si $u < u_\alpha$;
- Rejeter H_0, avec une probabilité α de se tromper, si $u \geq u_\alpha$.

Test (unilatéral) $H_0 :\ ''\mu_1 = \mu_2''$ **contre** $H_1 :\ ''\mu_1 < \mu_2''$,

Sous l'hypothèse H_0, la statistique U suit approximativement la loi normale $N(0,1)$.

On calcule $u = \frac{\overline{x}-\overline{y}}{\sqrt{\frac{\hat{\sigma}^2_{c,1}}{n_1}+\frac{\hat{\sigma}^2_{c,2}}{n_2}}}$, et on détermine u_α, sur la table de la loi normale, tel que $P(U < u_\alpha) = 1 - \alpha$ et on décide de :

- Ne pas rejeter H_0 si $u > -u_\alpha$;
- Rejeter H_0, avec une probabilité α de se tromper, si $u \leq -u_\alpha$.

La démarche et les résultats des trois tests ci-dessus restent valable si on remplace σ_1^2 ou σ_2^2 par leurs estimations $\hat{\sigma}^2_{c,1}$, le fait que $U = \frac{\overline{X}-\overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$ suit aussi une loi normale centrée réduite (on peut le justifier par le TCL).

1.6.2 Cas de petits échantillons

Soient X et Y des variables aléatoires indépendantes représentant le caractère dans chaque population. On suppose que X et Y suivent une loi normal de moyennes respectives μ_1 et μ_2, de variance respectives σ_1^2 et σ_2^2. On extrait un échantillon $X_1,\ X_2, ...,\ X_{n_1}$ de taille $n_1 \leq 30$ de X et un échantillon $Y_1,\ Y_2, ...,\ Y_{n_2}$ de taille $n_2 \leq 30$ de Y.

Test (bilatéral) $H_0 :\ ''\mu_1 = \mu_2''$ **contre** $H_1 :\ ''\mu_1 \neq \mu_2''$,

Afin de réaliser ce test, nous définissons la statistique suivante :

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}. \tag{1.2}$$

Sous l'hypothèse H_0 et l'hypothèse $\sigma_1 = \sigma_2$ la statistique du test définie dans (1.2) suit approximativement la loi de Student à $n_1 + n_2 - 2$ degrés de liberté.

Cependant, dans la pratique on ne sait pas si $\sigma_1^2 = \sigma_2^2 = \sigma^2$ ou non. A cet effet, on doit d'abord tester l'égalité des deux variances, $\sigma_1^2 = \sigma_2^2$ (Voir Section 1.5).

Si cette dernière hypothèse est retenue, alors la valeur commune σ^2 peut être estimer par $\hat{\sigma}_c^2 = \frac{(n_1-1)\sigma_{c,1}^2+(n_2-1)\sigma_{c,2}^2}{n_1+n_2-2}$. Ensuite, on calcule la réalisation de la statistique T, c'est-à-dire $t = \frac{\overline{x}-\overline{y}}{\hat{\sigma}_c \times \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ et on détermine sur la table de la loi de Student la valeurs critique, t_α, du test tel que : $P(-t_\alpha < T < t_\alpha) = 1-\alpha$. Finalement, on décide que :

- On ne peut rejeter H_0 si $t \in]-t_\alpha, t_\alpha[$;
- On rejette H_0 si $\notin]-t_\alpha, t_\alpha[$, avec une probabilité α de se tromper dans la décision.

Test (unilatéral) $H_0 : \ ''\mu_1 = \mu_2''$ **contre** $H_1 : \ ''\mu_1 > \mu_2''$,

Sous l'hypothèse H_0, si $\sigma_1 = \sigma_2$ alors la statistique, T, du test définie dans (1.2) suit approximativement la loi de Student à n_1+n_2-2 degrés de liberté. Ainsi, on détermine t_α sur la table de la loi de Student pour un $n = n_1+n_2-2$ et qui vérifié l'égalité $P(T \geq t_\alpha) = 1 - \alpha$ et on décide :

- Si $t < t_\alpha$, alors on ne peut rejeter H_0 ;
- Si $t \geq t_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper dans la décision.

Test (unilatéral) $H_0 : \ ''\mu_1 = \mu_2''$ **contre** $H_1 : \ ''\mu_1 < \mu_2''$,

Sous l'hypothèse H_0, si $\sigma_1 = \sigma_2$ alors la statistique, T, du test définie dans (1.2) suit encore approximativement la loi de Student à $n_1 + n_2 - 2$ degrés de liberté.

Pour prendre la décision sur le rejet de l'hypothèse H_0, il suffit de déterminer sur la table de Student pour un *ddl* $n = n_1 + n_2 - 2$ la valeur critique t_α tel que $P(T < t_\alpha) = 1 - \alpha$ et on décide :

- Si $t > -t_\alpha$, alors on ne peut rejeter H_0 ;
- Si $t \leq -t_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper dans la décision.

1.6.3 Exemple d'application

Supposons que nous intéressons à la durée de vie de deux types de capteurs dans un réseau de communication. Un échantillonnage simple des durées de vie de quelques capteurs, des deux types, nous a fourni ce qui suit :

C_1	34	23	31	14	21	33	30		
C_2	39	32	16	37	39	32	34	34	25

Question : Si on sait que

✓ Les deux échantillons sont indépendants entre eux,

✓ La durée de vie des deux composants suivent des lois normales,

alors, pour un risque de décision $\alpha = 5\%$, peut-on conclure que l'un des capteurs a une durée de vie moyenne significativement plus longue que l'autre ?

Réponse à la question : Dans cet exemple on a deux échantillon dont la taille des deux est inférieur à 30 (petite taille d'échantillons) alors le test candidat pour répondre à la question est bien que le test de Student pour l'homogénéité de moyennes. Soit les notations suivantes : μ_1 est la durée de vie moyenne du capteur C_1 et μ_2 est la durée de vie moyenne du capteur C_2.

1. Statistiques descriptive (Moyennes est variances) : On a,
   - $\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i = 26.5714$
   - $\overline{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i = 32$
   - $\hat{\sigma}^2_{1,c} = \frac{n_1}{n_1-1}\hat{\sigma}^2_1 = \frac{n_1}{n_1-1}\left[\frac{1}{n_1}\sum_{i=1}^{n_1} X_i^2 - \overline{X}^2\right] = \frac{7}{6}\left[\frac{5272}{7} - 26.5714^2\right] = 54.9524$
   - $\hat{\sigma}^2_{2,c} = \frac{n_1}{n_2-1}\hat{\sigma}^2_2 = \frac{n_1}{n_2-1}\left[\frac{1}{n_1}\sum_{i=1}^{n_1} Y_i^2 - \overline{Y}^2\right] = \frac{9}{8}\left[\frac{9652}{9} - 32^2\right] = 54.5000$
2. Les hypothèses du test : Le test à réaliser dans ce cas est le test unilatéral suivant :
$$H_0\text{"}\mu_1 = \mu_2\text{"}\ \textit{contre}\ H_1\text{"}\mu_1 < \mu_2\text{"}. \tag{1.3}$$
3. Les conditions d'applications : D'après l'énoncé de l'exemple la condition de normalité des deux échantillons et leurs indépendances sont déjà vérifiées,

alors le fais que la taille des échantillons est < 30, alors pour réaliser test (1.3) par le test de Student, on est contraint de vérifier d'abord ce qui suit :

$$H_0\text{"}\sigma^2 = \sigma^2\text{"} \ contre \ H_1\text{"}\sigma^2 \neq \sigma^2\text{"}.$$

(a) Sachant que $\hat{\sigma}^2_{c,1} > \hat{\sigma}^2_{c,2}$ alors, la statistique du test est : $F = \frac{\hat{\sigma}^2_{c,1}}{\hat{\sigma}^2_{c,2}} = \frac{54.9524}{54.5} = 1.0083$,

(b) La valeur critique du test est $f_\alpha = f_{(n_2-1,n_1-1,1-\alpha/2)} = f_{(6,8,0.975)} = 4.65$,

(c) On remarque que $F < f_\alpha$ alors on admet que les deux échantillons ont la même variance avec un risque 5% de se tromper.

(d) Le fait que les deux échantillons ont la même variance, alors on calcule la variance commune :

$$\hat{\sigma}^2_c = \frac{(n_1-1)\hat{\sigma}^2_{c,1} + (n_2-1)\hat{\sigma}^2_{c,2}}{n_1+n_2-2} = \frac{6*54.9524 + 8*54.5}{9+7-2} = 54.6939$$

4. La réalisation de la statistique du test : Revenant maintenant au test (1.3), la statistique de ce dernier est :

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\hat{\sigma}^2_c \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{26.5714 - 32}{\sqrt{54.6939\left(\frac{1}{7} + \frac{1}{9}\right)}} = -1.4566,$$

5. La valeur critique du test sera lue sur la table de la loi de Student et qui est $t_\alpha = t_{(n_1+n_2-2,1-\alpha)} = t_{(14,1-0.05)} = 1.7613$,

6. Discision : On remarque que $|T| < t_\alpha$, alors on ne rejette pas l'hypothèse H_0, c'est-à-dire on admet que les deux capteurs ont la même durée moyenne de vie.

1.7 Analyse de la variance à un facteur

Dans cette section, nous allons intéresser à un cas plus générale pour la comparaison de moyennes et cela lorsque le nombre d'échantillon est supérieur strictement à deux. Plus précisément nous allons intéresser à la technique d'analyse de la variance à un seul facteur qui est la plus adéquate avec la situation.

1.7.1 Position du problème

Supposons que nous ayons 3 forêts contenant un type d'arbre bien déterminé où nous désirons savoir si ces forêts ont une influence sur la hauteur des arbres ou non. À cet effet, nous avons réalisés un recueil de hauteur de six (06) arbres dans chaque forêt, dont les mesures sont rangées dans le tableau suivant.

Nř	forêt 1	forêt 2	forêt 3
1	23.3	18.9	22.5
2	24.4	21.1	22.9
3	24.6	21.1	23.7
4	24.9	22.1	24.0
5	25.0	22.5	24.0
6	26.2	23.5	24.5

TABLE 1.1: Tailles des arbres selon la forêt

Soit les notions et les notations suivantes :

- Les forêts : Variable qualitative contenant trois modalités, appelée facteur.
- Hauteur des arbres : Réponse, notée X, et μ_i la hauteur moyenne des arbres de

la $i^{\text{ème}}$ forêt ($i = \overline{1,3}$).

Répondre à notre objectif consiste à la réalisation du test suivant :

$$H_0 : \ ''\mu_1 = \mu_2 = \mu_3 = \mu'' \text{ contre } H_1 : \ ''\exists\ i, j \in \{1,2,3\} \text{ tel que } \mu_i \neq \mu_j;'' .$$

Pour réaliser ce test nous pourrons le décomposer en trois sous-tests où nous comparons la hauteur moyenne des arbres deux à deux selon les forêts. Mais afin de contourner le problème d'erreur α gonflé, le fait elle ne réalise qu'une seule comparaison à la fois, nous utilisons la technique statistique connue sous le nom d'analyse de variance (en anglais : Analyse Of Variance (ANOVA)) plutôt que des tests de Student t (voir Section 1.6) multiples. Remarquez que l'ANOVA 1 peut aussi être utilisée quand $p = 2$ puisque, elle retourne la même conclusion qu'un test t.

1.7.2 Analyse de la variance à un seul facteur

L'identification de l'$ANOVA$ 1 au sens littéraire peut être résumée dans la définition suivante :

Définition 1.7.1 *(ANOVA 1)*
L'analyse de la variance à un facteur teste l'effet d'un facteur contrôlé A ayant p modalités (groupes) sur les moyennes d'une variable quantitative X.

Les problèmes concernés par la technique $ANOVA$ 1 se présentent, en générale, de la manière suivante :

N	*groupe* 1	*groupe* 2		*groupe* p
1	$X_{1,1}$	$X_{1,2}$	$\cdots$	$X_{1,p}$
2	$X_{2,1}$	$X_{2,2}$	$\cdots$	$X_{2,p}$
3	$X_{3,1}$	$X_{3,2}$	$\cdots$	$X_{3,p}$
4	$X_{4,1}$	$X_{4,2}$	$\cdots$	$X_{4,p}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_j	$X_{n_1,1}$	$X_{n_2,2}$	$\cdots$	$X_{n_p,p}$

,

et le modèle mathématique leurs associés est donné par :

$$X_{ij} = \mu_i + \epsilon_{ij}, \text{ avec } i = \overline{1,n}, j = \overline{1,p} \text{ et } \epsilon_{ij} \rightsquigarrow N(0, \sigma^2), \tag{1.4}$$

où X_{ij} est la $j^{\text{ième}}$ réalisation de la variable quantitative X dans le $i^{\text{ième}}$ échantillon et ϵ_{ij} sont les erreurs de mesure.

Si on retient ce modèle alors le test à réaliser est défini par :

$$H_0 : \ ''\mu_1 = \mu_2 = ... = \mu_p = \mu'' \text{ contre } H_1 : \ ''\exists \ i, j \in \{1, 2, ..., p\} \text{ tel que } \mu_i \neq \mu_j''. \tag{1.5}$$

Dans ce qui suit, nous allons énumérer les étapes de la mise en oeuvre de l'ANOVA 1 qui nous permet de réaliser ce test.

1.7.3 Les étapes de l'$ANOVA$ 1

Afin de réaliser le test définie dans (1.5), trois conditions doit être vérifiées préalablement, à savoir :

- Les p échantillons comparés sont indépendants.

- La variable quantitative étudiée suit une loi normale dans les p populations comparées.
- Les p populations comparées ont même variance : *Homogénéité* des variances ou *homoscédasticité*.

Si ces dernières conditions sont vérifiées alors, on peut utiliser la technique $ANOVA\,1$ pour réaliser le test (1.5), et pour ce faire nous avons besoin des quantités (statistiques) suivantes :

- La moyenne de toutes les observations : $\overline{X} = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j} X_{ij}$ avec $n = \sum_{j=1}^{p} n_j$;
- Moyenne de chaque échantillon : $\overline{X}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij}$, pour $j = \overline{1,p}$;
- Variance de chaque échantillon : $\hat{\sigma}_i^2 = \frac{1}{n_j}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_j)^2$, pour $j = \overline{1,p}$;
- La variance de toutes les observations : $\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \overline{X})^2$ avec $n = \sum_{j=1}^{p} n_j$.

On peut démonter facilement que la variance de toutes les observations est la somme de la variance des moyennes et de la moyenne des variances des p échantillons, c'est-à-dire :

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \overline{X})^2 = \frac{1}{p}\sum_{j=1}^{p}\sigma_i^2 + \frac{1}{p}\sum_{j=1}^{p}(\overline{X}_j - \overline{X})^2, \tag{1.6}$$

ou encore :

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \overline{X})^2 = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_j)^2 + \frac{1}{p}\sum_{j=1}^{p}(\overline{X}_j - \overline{X})^2. \tag{1.7}$$

On multipliant (1.7), par n on obtient :

$$\underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \overline{X})^2}_{SC_{Tot}} = \underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_j)^2}_{SC_{Res}} + \underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(\overline{X}_j - \overline{X})^2}_{SC_{Fac}}, \tag{1.8}$$

où,

SC_{Tot} : est la variation totale qui représente la dispersion des données autour de la moyenne générale.

SC_{Fac} : est la variation due au facteur (variation intergroupes) qui représente la dispersion des moyennes autour de la moyenne générale.

SC_{Res} : est la variation résiduelle (variation intra-groupes) qui représente la dispersion des données à l'intérieur de chaque échantillon autour de sa moyenne.

L'idée la plus naturelle est que le facteur n'a pas d'impact sur le caractère étudié si la variation totale n'est engendrée que par la variation intra-groupes (résiduelle) associée au caractère, c'est-à-dire,

- Si H_0 est vraie, alors la variation SC_{Fac} due au facteur doit être petite par rapport à la variation résiduelle SC_{Res}.
- Par contre, si H_1 est vraie alors la variation SC_{Fac} due au facteur doit être grande par rapport à la quantité SC_{Res}.

Pour comparer ces quantités, Fisher a considéré le rapport des carrés moyens associés au facteur CM_{Fac} et les carrés moyens résiduels CM_{Res}, où

le carré moyen associé au facteur est : $CM_{Fac} = \frac{SC_{Fac}}{p-1}$.

le carré moyen résiduel est : $CM_{Res} = \frac{SC_{Res}}{n-p}$.

Si les 3 conditions d'application d'ANOVA (Indépendance, Normalité et Homogénéité) sont vérifiées et H_0 est vraie, alors

$$F_{obs} = \frac{CM_{Fac}}{SC_{Res}} \rightsquigarrow f_{(p-1,n-p)}.$$

Décision : Pour un seuil de risque donné α les tables de Fisher nous fournissent une valeur critique f_α telle que :

$$P\left(\frac{CM_{Fac}}{SC_{Res}} < f_\alpha\right) = 1 - \alpha,$$

- si $f_{obs} < f_\alpha \implies$ on ne peut pas rejeter H_0 (le facteur n'a aucune influence sur le caractère étudier),
- si $f_{obs} \geq f_\alpha \implies$ on rejette H_0 (le facteur influe sur le caractère étudié),

avec f_{obs} est la réalisation de la variable (statistique) F_{obs}.
Les calculs intermédiaires et les résultats d'une ANOVA 1 sont souvent présentés dans un tableau dont la forme est :

	Somme des carrés	Degrés de libertés	Carré moyen	ratio	Ficher
source de variation	SC	ddl	CM	F_{obs}	c
Inter-groupe (Fac)	SC_{Fac}	$p-1$	CM_{Fac}	$\frac{CM_{Fac}}{CM_{Res}}$	c
Intra-groupe (Rés)	SC_{Res}	$n-p$	CM_{Res}		
Total	SC_{Tot}	$n-1$			

1.7.4 Exemple d'application

Reprenant l'exemple présenté dans la Section 1.7.1. Les étapes qu'on doit suivre pour réaliser le test

$$H_0 : \text{ " } \mu_1 = \mu_2 = \mu_3 = \mu \text{ " contre } H_1 : \text{ " } \exists\, i, j \in \{1, 2, 3\} \text{ tel que } \mu_i \neq \mu_j \text{ ",}$$

à l'aide de la technique ANOVA 1, sont les suivantes :

- Calculer les moyennes des différents échantillons : $\overline{X}_1 = 24.73$, $\overline{X}_2 = 21.53$ et $\overline{X}_3 = 23.60$.
- Calculer la moyenne globale de toutes les observations : $\overline{X} = \frac{1}{n}(n_1\overline{X}_1 + n_2\overline{X}_2 + n_3\overline{X}_3) = 23.2889$.
- Compléter le tableau de l'ANOVA à un seul facteur :

source de variation	Somme des carrés SC	Degrés de libertés ddl	Carré moyen CM	ratio F_{obs}	Ficher c
Inter-groupe	31.5911	2	15.7956	12.02	3.6823
Intra-groupe	19.7067	15	1.3138		
Total	51.2978	17			

- Décision : on constate que $f_{obs} = 12.02 > f_{\alpha} = 3.6823$ (pour un risque de $\alpha = 5\%$), donc les hauteurs moyennes des arbres sont significativement différentes d'une forêt à une autre. Cela signifié que le facteur forêt influe sur la hauteur des arbres.

Conclusion

A partir les différentes notions et différents tests exposés dans ce chapitre on peut conclure que : 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.

Les tests paramétriques d'hypothèses font appel à un certain nombre d'hypothèses concernant la nature de la population dont provient l'échantillon étudié (normalité de la variable, égalité des variances, indépendance, etc.) et qui doivent être vérifié préalablement.

Chapitre 2

Quelques tests non paramétriques

Introduction

Dans ce chapitre après avoir mis aux claires la notion d'un test non paramétrique et sa différentiation à ceux paramétrique nous allons exposer les tests non paramétriques les plus usités dans la pratique.

2.1 Comparaison entre test paramétrique et non paramétrique

Définition 2.1.1 *Le test paramétrique est le test d'hypothèse qui fournit des généralisations pour faire des déclarations sur une caractéristique bien déterminer de la population parente.*

Définition 2.1.2 *Le test non paramétrique est défini comme le test d'hypothèse qui ne repose pas sur des hypothèses sous-jacentes, c'est-à-dire qu'il n'exige pas que la distribution de la population soit indiquée par des paramètres spécifiques.*

Le test repose principalement sur les différences de médianes. Par conséquent, il est également connu sous le nom de test sans distribution. Le test suppose que les variables sont mesurées au niveau nominal ou ordinal. Il est utilisé lorsque les variables indépendantes sont non métriques.

Dans un problème statistique, chaque situation correspond à un test approprié. Un test est soit paramétrique soit non paramétrique. Un test est soit libre d'une distribution soit non lié à une distribution. Chaque test possède une efficacité est une robustesse. Il faut utiliser le bon test à la bonne place. Les Principales différences entre les tests paramétriques et non paramétriques sont résumées dans la table suivante.

Base de comparaison	Test paramétrique	Test non paramétrique
Sens	Les hypothèses sont faites sur le paramètre de population.	Un test statistique utilisé dans le cas de variables indépendantes non métriques.
Base de la statistique du test	Distribution	Arbitraire
Niveau de mesure	Intervalle ou ratio	Nominal ou ordinal
Mesure de tendance centrale	Signifier	Médiane
Informations sur la population	Complètement connu	Indisponible
Applicabilité	Variables	Variables et attributs
Test de corrélation	Pearson	Lancier

TABLE 2.1: Différences entre les tests paramétriques et non paramétriques.

Conditions d'application des tests paramétriques

L'utilisation des tests paramétriques est soumise à des conditions d'application ou d'hypothèses à priori sur la distribution des variables dans les populations de référence, on cite :

1. La normalité des distributions parentes.

2. L'égalité des variances (homoscédasticité des résidus).

3. Les observations au sein d'un échantillon doivent être indépendantes.

Pour vérifier la première condition c'est-à-dire la normalité d'une distribution on dispose de différents tests : Test de Kolmogorov-Smirnov, test de Lilliefors, test de Shapiro-Wilk.

La quasi-totalité des tests que l'on a exposé dans le chapitre précédent supposent que la loi de la variable aléatoire X étudiée est normale dans les populations considérées (hormis pour la conformité ou la comparaison de moyennes sur de grands échantillons). Cette condition n'étant pas toujours satisfaite, nous allons voir dans la suite quelque tests qui sont valables même si la loi de X n'est pas normale.

Par exemple, pour la comparaison de moyennes, lorsque les échantillons peuvent être considérés indépendants, on peut applique le test de Mann et Whitney pour 2 échantillons, celui de Kruskal et Wallis pour un nombre quelconque d'échantillons. Lorsqu'on a affaire à deux échantillons appariés (c'est-à-dire non indépendants), on applique le test de Wilcoxon. Tous ces tests sont dits non paramétriques car ils ne nécessitent pas d'estimation de la moyenne et de la variance. En fait, ils n'utilisent même pas les valeurs x_i recueillies dans les échantillons, mais seulement leur rang dans la liste ordonnée de toutes les valeurs (voir figure 2.1).

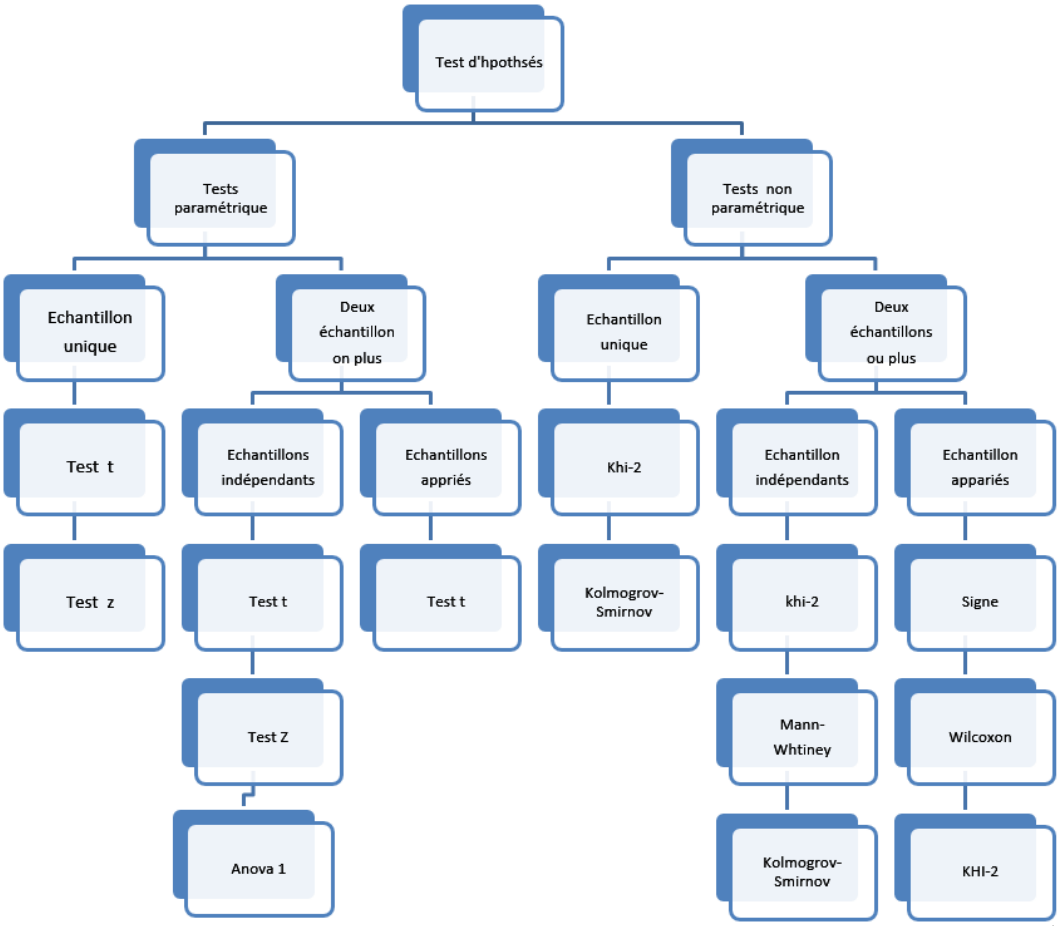


FIGURE 2.1 – Hiérarchie des tests d'hypothèses

2.2 Les tests d'ajustement K-S, Lilliefors et χ_2

Dans cette section nous allons présenter trois tests non paramétriques très usuels pour la vérification à travers d'un échantillon d'une population P l'adéquation de distribution de cette dernière à une distribution bien fixer. Les trois tests en question sont : le test de Kolmogorov-Smirnov (K-S), le test de Lilliefors et le test d'ajustement de Khi-deux.

Ce que nous allons remarquer, que la principale différence entre ces deux tests est que les deux tests Kolmogorov-Smirnov (K-S) et Lilliefors se basent sur la

distribution empirique de l'échantillon tandis que celui de khi-deux, il se base sur la densité empirique.

2.2.1 Test de Kolmogorov-Smirnov (K-S)

L'idée du test est de comparer la fonction de distribution empirique à la fonction de répartition. Le test de $K-S$ permet de tester n'importe quelle distribution.

Soit $X=(X_1,X_2,...X_n)$ un n-échantillon d'une loi P absolument continue par rapport à la mesure de Lebesgue sur $(\mathbb{R},\ \ \beta(\mathbb{R}))$ inconnue, soit $x=(x_1,...,x_n)$ une observation de cet échantillon, et on note F_n la distribution empirique associée à l'échantillon X définie pour $t\in\mathbb{R}$ par :

$$\begin{aligned} F_n(t) &= \frac{1}{n}\sum_{i=1}^{n} 1_{(X_i\leq t)} = \frac{1}{n}\sum_{i=1}^{n} 1_{(X_{(i)}\leq t)}, \\ &= \begin{cases} 0, & si \quad x_{(i)} < t, \\ \frac{i}{n}, & si \quad x_{(i)} \leq t < x_{(n)}, \\ 1, & si \quad x_{(n)} \leq t. \end{cases} \end{aligned} \tag{2.1}$$

où les $x_{(i)}$ sont les statistiques d'ordre de l'échantillon (valeurs de l'échantillon rangées par ordre croissant). En d'autres termes, $F_n(t)$ est la proportion d'éléments de l'échantillon qui sont inférieurs ou égaux à x.

On a :

- Pour chaque $t\in\mathbb{R}$ fixé, $F_n(t)$ est une valeur dans $[0,1]$
- Si on cherche à tester l'hypothèse : $H_0(P=P_0)\Longleftrightarrow H_0(F=F_0)$.

- Le théorème de Glivento–Gantelli(ajouter une reference) donne :

$$\sup_{t\in\mathbb{R}} \left|F_n(t) - F_(t)\right| \underset{n\to+\infty}{\longrightarrow} 0 \quad Ps$$

- La statistique de $K-S$ est défini par la distance en norme infinie entre la fonction de répartition empirique F_n, et la fonction de répartition F, c'est-à-dire :

$$KS = D_{KS}(P, P_0) = \sup_{t\in\mathbb{R}} |F_n(t) - F(t)|,$$

ainsi, si $(x_{(1)}, ..., x_{(n)})$ est la statistique d'ordre associée à l'échantillon X alors la statistique du test de $K-S$ sera donnée par :

$$D_{KS}(P,\ P_n) = \max_{1\leq i\leq n} \left\{ \left| F(x_{(i)}) - \frac{i}{n} \right|, \left| F(x_{(i)}) - \frac{i-1}{n} \right| \right\}.$$

- La région critique : on rejette H_0 si $D_{KS} > d_{n,\alpha}$, où $d_{n,\alpha}$ est le quantile théorique lu à partir de la table de Kolmogorov.

Il est à noter que mathématiquement parlant la statistique de $K-S$ est la distance en norme uniforme entre les deux fonctions de répartition. Par contre au sens graphique, elle représente le plus grand écart vertical en valeur absolue entre la valeur empirique et la valeur théorique de la fonction de répartition (voir Figure 2.2).

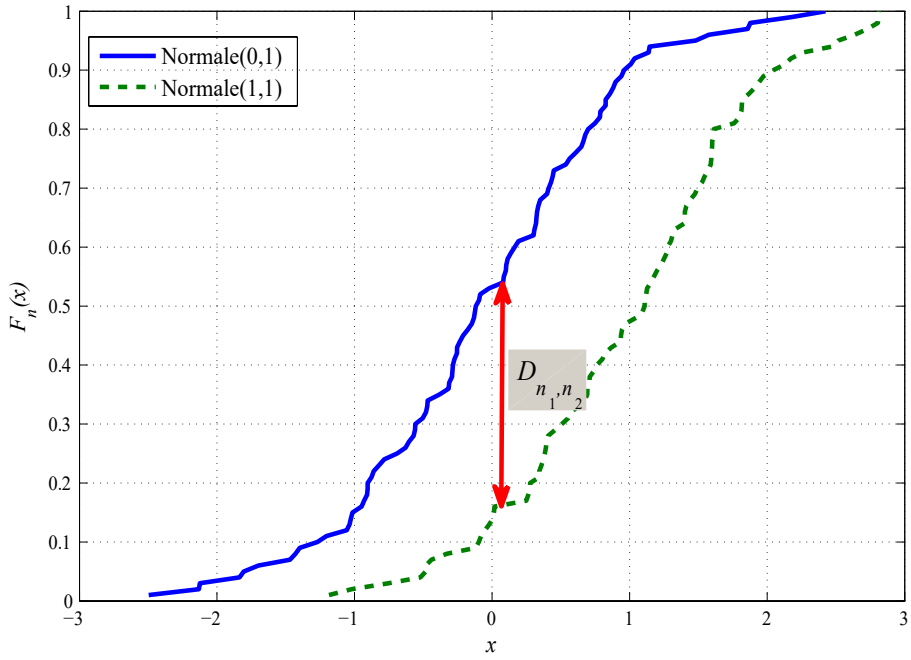


FIGURE 2.2 – La statistique du test de Kolmogorov-Smirnov

Exemple 2.2.1 *Le tableau suivant représente un échantillon de durées de vie en heures de cinq appareils de même type.*

Appareil	*1*	*2*	*3*	*4*	*5*
Durée de vie (heures)	*133*	*169*	*8*	*122*	*58*

Question : *Peut-on admettre, à un seuil* $\alpha = 5\%$*, que la durée de vie de ce type d'appareil suit une loi exponentielle ?*

Réponse : D'après l'énoncé de l'exemple $n = 5$ et le paramètre λ de la loi théorique est inconnu alors on doit l'estimer. Pour cela nous utilisons la moyenne empirique $\bar{X}$ de l'échantillon car $\bar{X}$ est un estimateur de $E(X) = \frac{1}{\lambda}$. Par conséquent, à partir de l'échantillon on trouve que $\bar{X} = 98$ et donc pour tous les calculs nous considérons que $\lambda = \frac{1}{98}$.

• Rappelons que la fonction de répartition théorique de la loi exponentielle est exprimée par :

$$F(x) = 1 - e^{-\lambda x}.$$

• Les différents calculs intermédiaires pour l'application du test KS sur nos données sont rangés dans le tableau suivant :

i	1	2	3	4	5
X_i	8	58	122	133	169
$F(x_i)$	0.078	0.447	0.712	0.743	0.822
$\frac{i}{n}$	0.2	0.4	0.6	0.8	1.0
$\left\|F(X_i) - \frac{i}{n}\right\|$	0.122	0.047	0.112	0.057	0.178
$\frac{i-1}{n}$	0.0	0.2	0.4	0.6	0.8
$\left\|F(X_i) - \frac{i-1}{n}\right\|$	0.078	0.247	0.312	0.143	0.22

• **La réalisation de la statistique** La distance de Kolmogorov-Smirnov est le plus grand des écarts en valeur absolue alors d'après la dernière ligne du tableau précédent la réalisation de la statistique du test est $D_{KS} = 0.312$.

• **Valeur critique :** Une lecture sur la table de Kolmogorov-Smirnov pour $n = 5$ au seuil $\alpha = 0.05$ indique que la valeur critique du test $d_\alpha = 0,565$.

• **Discision** : Puisque $D_{KS} = 0.312 < d_\alpha 0,565$, alors on rejette pas l'hypothèse H_0, c'est-à-dire on admet que la distribution de la durée de vie de l'appareil considéré dans l'exemple suit une loi exponentielle de paramètre $\lambda = \frac{1}{98}$.

Remarque 2.2.1 *On utilise le test de Kolmogorov-Smirnov lorsque la moyenne et l'écart type de la loi théorique sont connus a priori et sont donc fixés indépendamment de l'échantillon.*

2.2.2 Test de Lilliefors

Ce test est une variante du test de Kolmogorov-Smirnov, sous l'hypothèse de normalité, où les paramètres μ, σ de la loi sont estimées à partir des données.

⋆ La statistique du test est :

$$L_n = \sqrt{n}KS = \max_{1\leq i\leq n}\left\{\left|F_0\left(\frac{x_{(i)}-\overline{x}}{s_x}\right)-\frac{i}{n}\right|,\left|F_0\left(\frac{x_{(i)}-\overline{x}}{s_x}\right)-\frac{i-1}{n}\right|\right\}.$$

où $\overline{x}$ est la moyenne empirique et s_x 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).

Remarque 2.2.2 *On utilise le test Lilliefors lorsque la moyenne et l'écart type de la loi normale théorique sont estimés à partir de l'échantillon.*

2.2.3 Test de Khi-deux

Le test du $Khi-2$ utilise des propriétés de la loi Multinomiale. Il permet de juger si une hypothèse concernant la loi de probabilité d'une variable aléatoire est compatible avec la réalisation d'un échantillon ou non.

Dans ce test deux cas sont à distinguer :

1. La fonction de répartition F_0 est entièrement spécifiée, et ses paramètres sont connus.
2. On connaît seulement la forme de la loi de distribution, et ses paramètres sont estimés à partir d'un échantillon.

Soit $X_1, X_2, ..., X_n$ un n-échantillon issu d'une variable aléatoire X dont on désir

tester si la loi de l'échantillon est une loi entièrement spécifiée F_0. Le test de $Khi-2$ se base principalement sur les étapes suivantes :

- On partage le domaine D de la variable X en r classes $c_1, c_2, ..., c_r$ (avec le même principe que dans les statistiques descriptives voir Chapitre 1).
- Quantifier les N_i (pratique) : l'effectif de la classe c_i, $i = \overline{1:r}$.
- Pour $i = \overline{1:r}$, calculer p_i (théorique) : la probabilité de se trouver dans la classe c_i. Elle est déduite à partir de la loi de probabilité F_0.
- Pour $i = \overline{1:r}$, calculer $n_i = np_i$ (théorique) : effectif théorique de la classe c_i, $i = \overline{1:r}$.
- Déterminer la valeur de k_n^2 la réalisation de variable aléatoire K_n^2 définie par :

$$K_n^2 = \sum_{i=1}^{r} \frac{(N_i - n_i)^2}{n_i}. \tag{2.2}$$

Sachant que **Pearson** a démontré que la variable aléatoire K_n^2 suit asymptotiquement un $Khi-2$ à $(r-1)$ degré de liberté, alors la valeur critique du test sera déterminer à partir de le table de $Khi-2$, $\chi^2_{(r-1,\alpha)}$ et la décision sera prise comme suite :

- Si $k_n^2 < \chi^2_{(r-1,\alpha)}$, alors on accepte l'ajustement de la distribution de la variable aléatoire X par la loi choisie F_0.

- Si $k_n^2 \geq \chi^2_{(r-1,\alpha)}$, alors on rejette l'ajustement de la distribution de la variable aléatoire X par la loi choisie F_0.

Lorsque les paramètres de la loi à valider sont estimés à partir de l'échantillon, le degré de liberté de la distribution de $Khi-2$ est alors égal à $(r-q-1)$, q étant le nombre de paramètres estimés c'est-à-dire valeur critique du test est égale au fractile $\chi^2_{(r-q-1,\alpha)}$.

L'application du test $Khi-2$ doit satisfaire les conditions suivantes :

- Le nombre de classes doit être supérieur ou égal à 7.
- L'effectif théorique de chaque classe doit être supérieur ou égal à 5.
- Les effectifs théoriques des k classes doivent être sensiblement égaux.

Exemple 2.2.2 *On veut tester si un dé n'est pas truqué au risque $\alpha = 5\%$. Pour cela on lance ce dé 60 fois et on obtient les résultats suivants*

X_i *(Face)*	*1*	*2*	*3*	*4*	*5*	*6*
N_i	*15*	*7*	*4*	*11*	*6*	*17*
n_i	*10*	*10*	*10*	*10*	*10*	*10*

On a fait figurer dans le tableau la valeur théoriquement espérée n_i du nombre d'apparitions de la face i dans l'hypothèse où le dé n'est pas truqué, ceci afin de faciliter le calcul de la quantité $\chi^2_n(p, \bar{p}_n)$ qui est donc ici égale à

$$\begin{aligned}\sum_{i=1}^{k} \frac{(N_i - n_i)^2}{n_i} &= \frac{(15-10)^2}{10} + \frac{(7-10)^2}{10} + \frac{(4-10)^2}{10} + \frac{(11-10)^2}{10} + \frac{(6-10)^2}{10} + \frac{(17-10)^2}{10} \\ &= 13.6.\end{aligned}$$

Sous l'hypothèse

$$H_0 : "p_1 = \cdots = p_6 = \frac{1}{6}",$$

la variable aléatoire $\chi^2_n(p, p_n)$ a donc pris la valeur $13,6$. Or le seuil de rejet lu dans la table de la loi du $\chi^2_{(k-1,\alpha)}$ est $\chi^2_{(5,0.05)} = 11.07$.

La valeur observée dépassant cette valeur, on est amené à rejeter l'hypothèse H_0 au risque $\alpha = 0.05$. Ceci signifie qu'avec un seuil de risque 5% on admet que le Dé est truqué.

2.3 Test de $khi-2$ d'indépendance

Ce test est utilisé pour étudier la liaison entre deux variables quantitatives dans le même échantillon de taille n . Soient X_1, X_2 deux variables qualitatives telle que :

- X_1 est à valeur dans $\{a_1, ..., a_m\}$,
- X_2 est à valeur dans $\{b_1, ..., b_l\}$.

Sous H_0, la distribution de X_1 devrait être indépendante de celle de X_2. Par contre, si la distribution de X_1 est liée à celle de X_2, on rejette H_0 au profit de H_1, les deux variables X_1 et X_2 sont liées. Ainsi, on cherche à tester :

$$\begin{cases} H_0 : & \text{Les variables } X_1, X_2 \text{ sont indépendantes,} \\ H_1 : & \text{Les variables sont liées.} \end{cases}$$

- La statistique de $khi-2$ d'indépendance est donnée comme suit :

$$Q_{ind} = \sum_{i=1}^{m} \sum_{j=1}^{l} \frac{\left(N_{ij} - \frac{N_{i\bullet} \times N_{\bullet j}}{n}\right)^2}{\frac{N_{i\bullet} \times N_{\bullet j}}{n}}. \tag{2.3}$$

  $N_{i\bullet}$: Nombre de X_1 de valeurs a_i $(i = 1, ..., m)$.

  $N_{\bullet j}$: Nombre de X_2 de valeurs b_j $(j = 1, ..., l)$.

  N_{ij} : Nombre de (X_1, X_2) de valeurs (a_i, b_j).

- La région critique : alors la règle de décision du test d'indépendance de $Khi-2$ sera sous la forme suivante,

  ⋆ Si $Q_{ind} < \chi^2_{(m-1)(l-1)}(1-\alpha)$, alors les deux variables en question sont indépendantes,

  ⋆ Si $Q_{ind} > \chi^2_{(m-1)(l-1)}(1-\alpha)$, alors les deux variables en question sont liées,

où $\chi^2_{(m-1)(l-1)}(1-\alpha)$ le fractile d'ordre $1-\alpha$ d'une loi de $Khi-2$ à $(m-1)(l-1)$ *ddl*

Exemple 2.3.1 *Supposons qu'on veut comparer l'action de deux types de levures (A et B) sur une pâte à gâteaux. Pour cela, on prélève, pour chacune des levures, un échantillon aléatoire de gâteaux. L'aptitude des pâtes à lever est définie par les critères suivants : Moyenne, Bonne et Très bonne. Les résultats constatés sont rassemblés dans le tableau suivant :*

Aptitude à lever	*Moyenne*	*Bonne*	*Très bonne*
A	*41*	*16*	*63*
B	*22*	*27*	*51*

Question : *Au risque de 5%, peut-on conclure à une différence d'activité des deux types levures ?*

Réponse : Le test statistique adéquat pour déterminer s'il y a une différence d'activité des deux levures ou non est bien que le test d'indépendance de χ^2. Sous l'hypothèse que l'aptitude à lever et indépendante de la levure utilisée, les résultats obtenus par le calcul des effectifs espérés

$$n_{ij} = \frac{N_{i\bullet} \times N_{\bullet j}}{n},$$

sont résumés dans le tableau suivant :

			Aptitude à lever			Total
			Moyenne	Bonne	Très Bonne	
Levure	*A*	N_{ij}	41	16	63	120
		n_{ij}	34.4	23.5	62.2	120
		$e_i = N_{ij} - n_{ij}$	6.6	-7.5	0.8	
	B	N_{ij}	22	27	51	100
		n_{ij}	28.6	19.5	51.8	100
		$e_i = N_{ij} - n_{ij}$	-6.6	7.5	-0.8	
Total	Total		43	63	114	220

Pour répondre à notre objectif, il suffit de comparé la réalisation de la statistique du test avec la valeur critique qui est le quantile d'ordre $\alpha = 5\%$ d'une loi de *Khi-Deux* à $(r-1)*(k-1)$ degré de liberté, avec $r = 2$ (nombre de levures) et $k = 3$ (nombre de critères).

$$K_n^2 = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{\left(N_{ij} - \frac{n_{i\bullet}\times n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}\times n_{\bullet j}}{n}} = \sum_{i=1}^{r}\sum_{j=1}^{k} e_{ij}^2/n_{ij} \approx 8.1, \tag{2.4}$$

et la valeur critique du test est donnée par :

$$\chi^2_{((r-1)(k-1),\alpha)} = \chi^2_{((2-1)(3-1),0.05)} = 5.991 \tag{2.5}$$

De (2.4) et (2.5), on conclut qu'il y a une différence entre l'activité des deux levures.

2.4 Test de Wilcoxon-Mann-Whitney

Comparer deux échantillons non appariés (respectivement, appariés) pour une variable étudiée : test de Wilcoxon-Mann-Whitney C'est l'homologue non paramétrique du test t de Student non appariés(respectivement, appariés) en paramétrique. Ce test permet de verifier, pour une variable quantitative et au risque d'erreur α (0.05 bien souvent), si deux échantillons non apparies sont issus d'une même population ou bien de deux populations différentes vis-a-vis de la variable étudiée.

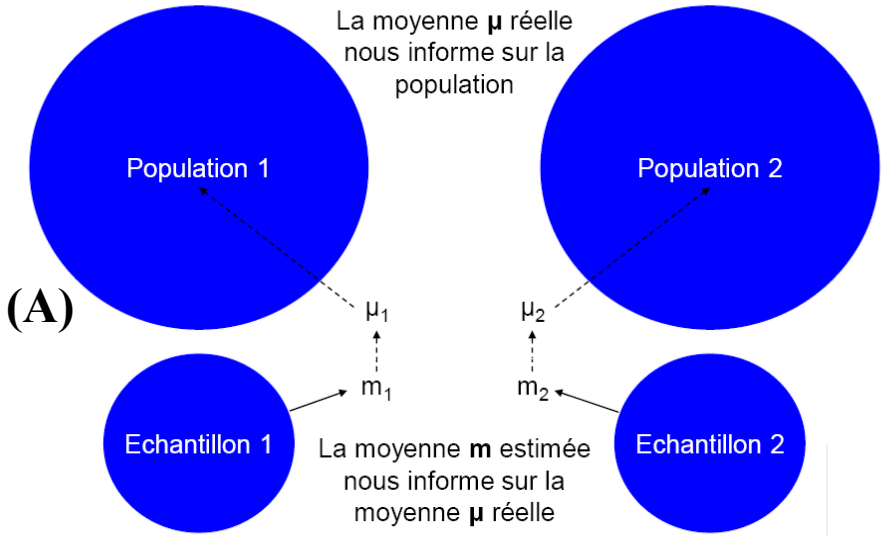
$$\begin{cases} H_0 : & \text{les deux échantillons appartiennent à la même population,} \\ H_1 : & \text{Les deux échantillons appartiennent à deux différentes populations.} \end{cases}$$

Remarque 2.4.1 *Il est important de remarquer qu'ici on test des hypothèses* H_0

et H_1 qui portent sur l'ensemble de la population et non pas sur un paramètre bien spécifique de moyenne par exemple (l'équivalent de ce test en paramétrique qui n'est autre que le test de comparaison de moyennes de Student).

L'équivalent paramétrique de ce test répond à la question : "Peut-on déceler une différence entre les paramètres de moyennes estimées m_1 et m_2 à partir de deux échantillons non apparies et si oui, alors les moyennes réelles μ_1 et μ_2 sont issues de deux populations distinctes".

L'équivalent paramétrique du test de Wilcoxon-Mann-Whitney fonctionne tel qu'il est illustrer dans la figure 2.3 suivantes.



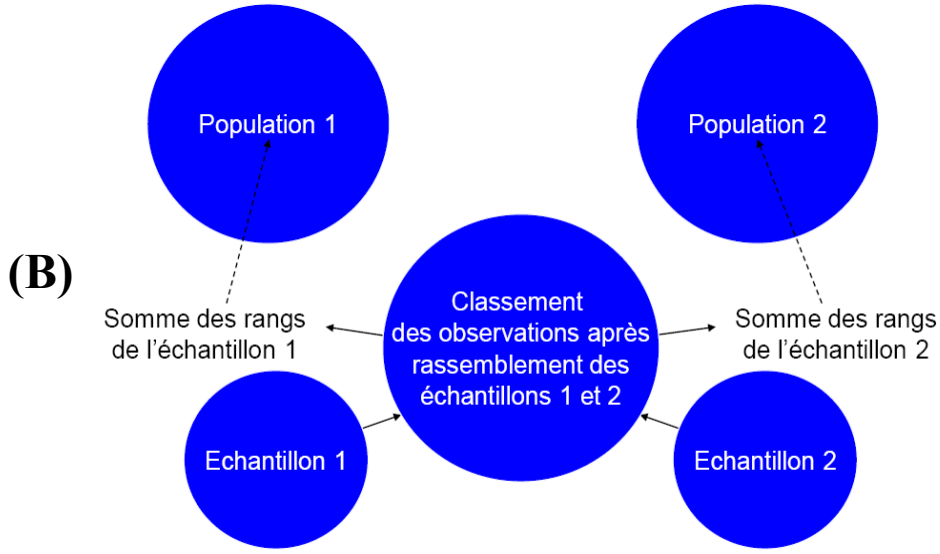


FIGURE 2.3 – Fonctionnement de l'équivalent du test de Wilcoxon-Mann-Whitney en paramétrique : (A) cas paramétrique et (B) cas non paramétrique.

Dans le cadre paramétrique du test de comparaison de moyenne de Student, si H_0 n'est pas rejetée (les deux moyennes sont égales), on peut conclure que les deux échantillons sont identiques et sont issus d'une même population, mais seulement après avoir vérifié l'hypothèse d'égalité des variances (pour plus de détails voir Chapitre 1).

Pourquoi : Le fais que les deux échantillons ayant la même moyenne ne signifier pas qu'ils sont issus d'une même population car ils peuvent provenir de deux population distincte dont les moyennes sont les mêmes mais leurs variances sont totalement différentes.

Certainement le test de Student est le plus populaire des tests non paramétriques. Il recouvre en réalité deux formulations qui sont équivalentes (elles peuvent se déduire l'une de l'autre), d'une part le test de Wilcoxon, d'autre part le test de Mann-Whitney.

On considère deux échantillons indépendants $(X_1, ..., X_{n_1})$ de la fonction de répartition F_1 et $(Y_1, ..., Y_{n_2})$ aussi de fonction de répartition F_2. Supposons que F_1 et F_2 sont continues et on veut tester ce qui suit :

$$\begin{cases} H_0 : "F_1(X) = F_2(X+\theta), \quad \theta = 0" & \text{Distributions identiques,} \\ H_1 : "F_1(X) = F_2(X+\theta), \theta \neq 0" & \text{Distributions différentes,} \end{cases}$$

où θ est un paramètre de translation c'est-à-dire il représente le décalage entre les fonctions de répartitions.

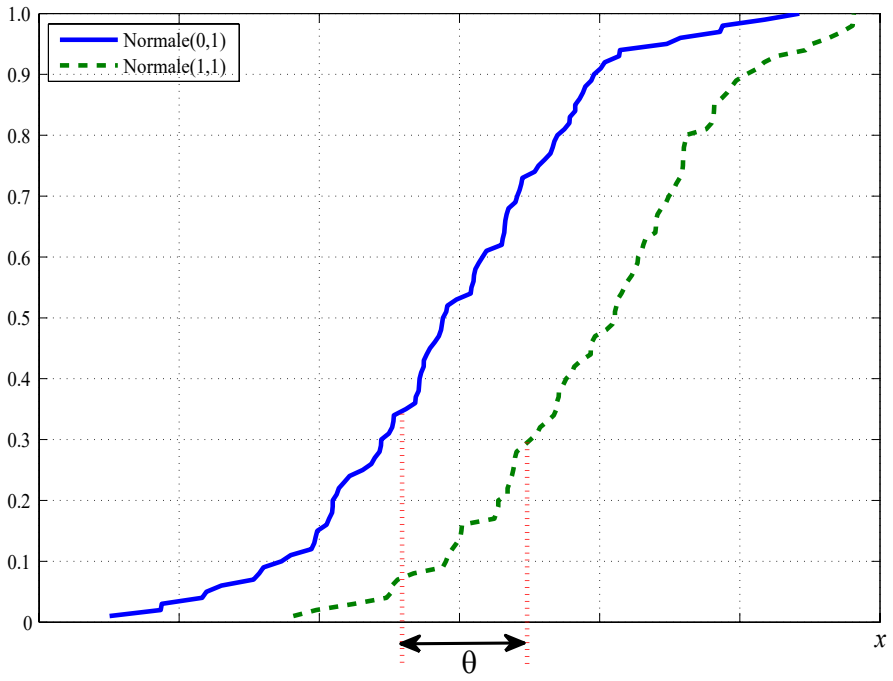


FIGURE 2.4 – Différentiation selon le paramètre de localisation.

La procédure du Wilcoxon-Mann-Whitney se base sur les quatre étapes suivantes :

1. Regrouper les deux échantillons et Ordonner les observations par ordre croissant.
2. Affecter un rang à chaque observation.
3. Calculer W la somme des rangs d'un échantillon (en général pour des raisons

calculatoire on choisit celui de plus petite taille) défini par :

$$W = \sum_{i=1}^{N} R(X_i),$$

avec $R(X_i)$ est le rang de la i^{eme} observation de X et $N = n_1 + n_2$.

4. Calculer la statistique U_{n_1,n_2}.

La statistique de Mann et Whitney utilise la somme des rangs uniquement. Nous retrouvons bien l'idée de décalage entre les distributions basé sur leur localisation, pour le test :

$$\begin{cases} H_0 : \theta = 0, \\ H_1 : \theta \neq 0. \end{cases} \Leftrightarrow \begin{cases} H_0 : F_1(X) = F_2(X + \theta), \\ H_1 : F_1(X) = F_2(X + \theta). \end{cases}$$

Nous calculons les quantiles :

$$U_1 = W_1 - \frac{n_1(n_1+1)}{2}, \quad U_2 = W_2 - \frac{n_2(n_2+1)}{2},$$

avec W_1 (respectivement, W_2) est le cumule des rangs des observations de l'échantillon X (respectivement, Y) dans l'échantillon regrouper. Le principe du test de Mann-Whitney correspond à la plus petite quantité,

$$U = min(U_1, U_2).$$

Ainsi, sous l'hypothèse que H_0 est vraie, l'espérance et la variance de U s'écrivent :

$$E(U) = \frac{n_1 n_2}{2}, \quad V(U) = \frac{(n_1 + n_2 + 1)n_1 n_2}{12}. \tag{2.6}$$

Traitement des ex-aequo (principe des rangs moyens)

Dans certaines situations on trouve des ex-aequo dans les valeurs. Pour surpasser ces situations deux approches sont possibles. La méthode des rangs aléatoires attribue aléatoirement les rangs aux observations confondues. Dans ce cas, la modification des tables et lois asymptotiques existantes n'est pas nécessaire. Cependant, la puissance du test est faible. La deuxième approche dite la méthode des rangs moyens procède de la manière suivante : les observations qui possèdent des valeurs identiques se voient attribuer la moyenne de leurs rangs (voir l'exemple 2.5.1).

Cas de grands échantillons Lorsque les échantillons atteignent une taille suffisamment élevés (> 20), la loi de la statistique U converge vers la loi normale de moyenne $E(U)$ et de variance $V(U)$ définis dans (2.6). Pour un test $H_0(F_x = F_y)$, nous pouvons donc définir la statistique centrée réduite

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{1}{12}(n_1 + n_2 + 1)n_1 n_2}} \sim N(0, 1).$$

La région critique du test au niveau de signification α est :

$$|Z| > z_{1-\frac{\alpha}{2}},$$

où $z_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée réduite.

Remarque 2.4.2 *Ce type de test est fortement recommandé quand les échantillons sont de petites tailles, car la variance est de plus en plus variable, même pour des échantillons issus de la même population $N(m, \sigma)$. Par conséquent, le risque de ne pas satisfaire la condition d'égalité des variances en paramétrique*

augmente. S'il est moins précis que son homologue paramétrique, il est aussi plus robuste car permet de déceler n'importe quel type de différence entre deux échantillons.

Exemple 2.4.1 *(Test de Wilcoxon) Disposons de deux échantillons (mâle et femelles) d'une espèce de souris, dont on a mesuré le poids (en g) chez l'individu adulte :*

X (échantillon femelle, $n_1 = 5$)	*2 4 7 9 11*
Y (échantillon mâle, $n_2 = 6$)	*3 5 6 8 13 15*

Question : *On veut tester, à un seuil $\alpha = 5\%$, si le poids moyen des mâles et le poids moyen des femelles sont les mêmes ou non ?*

Réponse :

• Les hypothèses du test sont :

$$\left\{ \begin{array}{l} H_0 : \text{le mâle et la femelle ont le même poids} \\ H_1 : \text{le mâle et la femelle ont des poids différents} \end{array} \right\}$$

• Calcul des rangs pour chaque échantillon et de la statistique du test : En regroupant les deux échantillons on aura

Z_i	2	3	4	5	6	7	8	9	11	13	15
X_i ou Y_i	x_1	y_1	x_2	y_2	y_3	x_3	y_4	x_4	x_5	y_5	y_6
Rang	1	2	3	4	5	6	7	8	9	10	11

D'après ce tableau on a $R(x_i) = 1, 3, 6, 8, 9$ alors

$$W_x = \sum R(x_i) = 1 + 3 + 6 + 8 + 9 = 27,$$

et

$$Z = \frac{U - \frac{n_1(n_1+n_2+1)}{2}}{\sqrt{\frac{1}{12}(n_1+n_2+1)n_1n_2}} = -0.5477$$

• Décision : On remarque que $|Z| = 0.5477 < Z_{0.975} = 1.96$ alors on ne rejette pas l'hypothèse $H_0(F_x = F_y)$, autrement dit, on admet que les souris femelles et les souris mâles de l'espèce étudiées ont le même poids moyen.

Exemple 2.4.2 *(Test de Mann-Whitney) Etant donné les deux échantillons suivant :*

X $(n_1 = 8)$	*14*	*25*	*30*	*32*	*40*	*41*	*43*	*45*
Y $(n_2 = 7)$	*12*	*16*	*19*	*20*	*24*	*27*	*35*	

Le regroupement des deux échantillon en ordre croissant nous fournit ce qui suit :

Z_i	12	14	16	19	20	24	25	27	30	32	35	40	41	43	45
X_i ;Y_i	y_1	x_1	y_2	y_3	y_4	y_5	x_2	y_6	x_3	x_4	y_7	x_5	x_6	x_7	x_8
Rang	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Ainsi, un calcul direct de U_{xy} donne :

U_{xy} =6+2+1+1=10

Et ceci le fait que, 14 précède 6 valeurs de X ;

25 précède 2 valeurs de X ;

30 précède 1 valeur de X ;

32 précède 1valeur de X

Ce qui fait au total :10.

Les rangs des Y (soulignés ci-dessus) étant :

2 7 9 10 12 13 14 15

on a alors $S_y = 82$ d'ou :

$$U_{xy} = n_2 n_1 + \frac{n_2(n_2+1)}{2} - S_y = 56 + \frac{72}{2} - 82 = 10.$$

$$Z = \frac{U_{xy} - \frac{n_1(n_1+n_2+1)}{2}}{\sqrt{\frac{1}{12}(n_1+n_2+1)n_1 n_2}} = -2.0831$$

• Décision : On remarque que $|Z| = 2.0831 > Z_{0.975} = 1.96$ alors on rejette l'hypothèse $H_0(F_x = F_y)$, autrement dit, on admet que les deux échantillon ne sont pas issue d'une même population.

2.5 Test de Kruskal-Wallis pour $K \geq 2$

Ce test vérifie si plusieurs échantillons ($k > 2$) appartiennent à la même population. Il s'agit de l'homologue non paramétrique de l'analyse de variance à un facteur mais avec le sérieux avantage de ne pas tenir compte de la loi de distribution de la variable étudiée (l'$ANOVA$ suppose qui les données sont gaussiennes) ni de l'égalité des variances entre échantillons.

Ce test est une généralisation du test de Wilcoxon-Mann-Whitney pour le cas de comparaisons de moyennes de plus de deux échantillons et par conséquent il fonctionne avec le même principe de remplacement des valeurs de la variable d'étude par leurs rangs respectifs. Comme pour le test de Wilcoxon-Mann-Whitney, les échantillons sont indépendants et chaque échantillon peut avoir un nombre différent d'observations. L'objectif du test de Kruskal-Wallis est comparer la distribution d'une variable quantitative X entre K groupes indépendantes, c'est-à-dire à tester :

$$\begin{cases} H_0 : F_1(X) = F_2(X+\theta) = ... = F_n(X+\theta), \theta = 0 & \text{distributions identiques,} \\ H_1 : \exists i \neq j, \quad F_i(X) = F_j(X+\theta), \quad \theta \neq 0 & \text{distributions différentes.} \end{cases}$$

Les étapes de la mise en oeuvre du test de Kruskal-Wallis sont comme suit :

1. Combiner tous les échantillons en un seul et affecter un rang à chaque observation.
2. Pour chaque échantillon, calculer les rangs r_{ij} de ses observations (r_{ij} rang de l'observation X_{ij} parmi les n observations).
3. Le calcul des rangs moyens pour chaque groupe par la relation suivante :

$$\overline{w}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} r_{ij}, \quad \text{pour } i = \overline{1:K}.$$

4. Le calcul de la moyenne globale des rangs par :

$$\overline{w} = \frac{1}{k} \sum_{j=1}^{k} \overline{w}_i.$$

5. La statistique du test sous H_0 est définie de la manière suivante :

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} n_i (\overline{w}_i - \overline{w})^2 \sim \chi^2_{k-1}.$$

6. La région critique du test, au seuil α, est donnée par :

$$H > \chi^2_{k-1,1-\alpha},$$

avec $\chi^2_{k-1,1-\alpha}$ est le quantile d'ordre $1-\alpha$ d'un khi-deux à $K-1$ degré de

liberté.

Exemple 2.5.1 *Dans une agglomération, on a relevé le taux d'Azote dans l'air à intervalle de temps régulier pendant trois mois consécutifs. On a observé les résultats suivants :*

Mai	*41*	*36*	*12*	*18*	*28*	*23*	*19*	*8*
Juin	*39*	*23*	*21*	*37*	*20*	*12*	*13*	
Juillet	*48*	*35*	*61*	*79*	*63*	*16*	*80*	

On veut tester l'hypothèse d'égalité du taux mensuel moyen d'Azote.

Dans cet exemple, il s'agit d'un test d'homogénéité de moyennes sur trois échantillons dont on ne connaît pas la distribution. Pour répondre à notre objectif on fait donc recours au test de Kruskal-Wallis.

On peut présenter les résultats comme ceci :

Rang	Taux	Groupe
1	8	$x_{1,1}$
2.5	12	$x_{1,2}$
2.5	12	$x_{2,1}$
4	13	$x_{2,2}$
5	16	$x_{3,1}$
6	18	$x_{1,3}$
7	19	$x_{1,4}$
8	20	$x_{2,3}$
9	21	$x_{2,4}$
10.5	23	$x_{1,5}$
10.5	23	$x_{2,5}$

Rang	Taux	Groupe
12	28	$x_{1,6}$
13	35	$x_{3,2}$
14	36	$x_{1,7}$
15	37	$x_{2,6}$
16	39	$x_{2,7}$
17	41	$x_{1,8}$
18	48	$x_{3,3}$
19	61	$x_{3,4}$
20	63	$x_{3,5}$
21	79	$x_{3,6}$
22	80	$x_{3,7}$

Noter que les valeurs 12 et 23 apparaissent deux fois : aux rangs 2 et 3 pour la valeur 12 et aux rangs 10 et 11 pour la valeur 23. Les rangs sont remplacés par leur moyenne (respectivement 2.5 et 10.5). On calcule ensuite les sommes ri des rangs pour chacun des trois groupes.

Par exemple pour r_1 les calculs se font comme suit :

$$r_1 = 1 + 2.5 + 6 + 7 + 10.5 + 12 + 14 + 17 = 70$$

De même, on trouve $r_2 = 65$ et $r_3 = 118$.

La variable de décision est

$h = \frac{12}{n(n+1)} \left(\sum_{i=1}^{k} \frac{r_i^2}{n_i} \right) - 3(n+1)$

avec $n = \sum_{i=1}^{k} n_i = 8 + 7 + 7 = 22$ et on trouve $h = 7.013$.

La statistique h suit une loi du χ^2 à $k-1 = 2$ degrés de liberté. La valeur critique au seuil 5% est $u_\alpha \chi^2_{2,0.05} = 5.99$. Comme h est supérieure à la valeur critique, alors on rejette l'hypothèse H_0 avec un risque de $\alpha = 5\%$ de se tromper, c'est-à-dire on admet qu'il y a au moins deux échantillons parmi les trois qui n'ont pas la même moyenne.

2.6 Tests de Cramer-Von Mises

En plus des tests exposés dans la Section 2.2, ll existe d'autres tests non paramétrique permettant de comparer des distributions de deux échantillons où le test de Cramer-Von Mises fait partie. Le test de Cramer-Von Mises repose sur la somme des carrés des écarts en valeurs absolue ,S_d^2, entre les fonctions de répartition emériques des deux échantillons. La statistique de ce test est

$$T = \frac{n_1 n_2 S_d^2}{n},$$

avec n_1 et n_2 représentent le nombres d'observations du premier échantillon et du deuxième échantillon respectivement et $n = n_1 + n_2$.

Pour un test bilatéral, on rejette H_0 au niveau de signification 5% (resp. 1%) si T est supérieur à 0.461 (resp. 0.743).

L'avantage du test de Cramer-Von Mises par rapport au test classique de Kolmogorov-Smirnov est que :

1. Le test de Cramer-Von Mises est souvent plus puissant que le test de Kolmogorov-Smirnov
2. Le test de Cramer-Von Mises il est plus facile à utiliser grâce à la bonne approximation qui évite le recours à des tables.

2.7 Avantages et inconvénients des tests statistiques non paramétriques

Les avantages des tests non paramétriques peuvent être résumés en :

– Pas d'hypothèses sur la forme de la distribution de la population de l'échantillon.
– S'emploient même pour des échantillons de taille très faible ($n = 6$).
– Peut s'utiliser dans le cas de variables qualitatives (ordinales et nominales).
– Méthodes intuitives.
– Robustesse du test.

Les inconvénients des tests non paramétriques sont donnés comme suit :

– Perte de puissance pour mettre en évidence un effet faible sur la variable étudiée.

– Ils sont moins précis que leurs homologues paramétriques.

Conclusion

Les tests non paramétriques exposés dans le présent chapitre nous permettent de conclure que ces derniers sont moins efficaces que les tests paramétriques. En effet, puisqu'ils sont soumis à des conditions contraignantes, qu'il faut bien vérifier préalables avant leurs mise en oeuvre les rendent plus précises et justifiables par la théorie.

Cependant, lorsque les conditions d'application d'un test paramétrique ne sont pas vérifiées ou sont impossible à les vérifiées (exemple de petits échantillons), on optera pour ceux de non paramétrique car la quasi-totalité de ces derniers sont exempte de conditions.

Conclusion générale

Dans ce mémoire notre objectif et de distinguer et de mettre en évidence la différence entre les deux catégories des tests statistiques existantes dans la littérature à savoir : les tests paramétriques et les tests non paramétriques.

Pour répondre à notre objectif, après avoir introduit une brève description de la notion d'un test statistique (hypothèses, risque et erreurs, ...), nous avons présenté un échantillon de tests paramétriques les plus usités dans la pratique où nous avons mis l'accent sur les conditions de leurs mise en IJuvre. Par la suite, nous avons présenté quelques tests non paramétriques. Enfin, les tests en question (paramétriques et non paramétriques) ont été illustrés à travers des exemples numériques.

Par ailleurs, L'analyse du principe et de l'efficacité des tests paramétriques et non paramétrique nous permet de conclure que : D'une part, pour effectuer un test statistique, si les informations sur la population sont complètement connues, sous forme de paramètres, alors le test adéquat dans ce genre de situation est bien que le test paramétrique tandis que, si on ne dispose pas de connaissances sur la population et qu'il est nécessaire de tester l'hypothèse sur le test effectué alors il faut choisir un test non paramétrique. D'autre part, faire un choix entre le test paramétrique et le test non paramétrique n'est pas une tâche facile pour un chercheur effectuant une analyse statistique.

Bibliographie

[1] T. Bulle, (1989) Comparaison de populations : Tests non paramétriques et analyse de variance. Edition Fenixx Réédition Numérique.

[2] M. Cherfaoui, (2016/2017) Polycopié de cours : Statistiques Appliquées à l'Expérimentation En Sciences Biologique, 3 année licence biologie. Département SNV, Université de Biskra.

[3] Khac Vo, (1985) Estimation et tests paramétriques et non paramétriques : cours et exercices avec solutions. Edition Ellipses-Marketing.

[4] G. Pupion et P.C Pupion, (1998) Tests non paramétriques avec applications à l'économie et à la gestion. Edition Economica.

[5] P.C Pupion et G. Pupion, (2010) Méthodes statistiques applicables aux petits échantillons : Avec applications en sciences sociales, économie et gestion. Edition Hermann.

[6] P. Sprent, (2010) Pratique des statistiques non paramétriques. Edition INRA, collection Techniques et pratiques.

Résumé

L'objectif principal de ce mémoire est d'identifier les différentes conditions d'applicabilité de certains tests paramétriques puis d'exposer leur alternative, à savoir les tests non paramétriques, en cas de non-vérification d'une ou de plusieurs des conditions en question. Enfin, dans le but d'illustrer les étapes de la mise en oeuvre de ces tests, des exemples d'application numérique ont été présentés.

Mots clés : Tests statistiques, Tests (non)paramétriques, risque de décision, la Statistique du test.

Abstract

The main objective of this thesis is to identify the different conditions of applicability of certain parametric tests then to expose their alternative, namely the non-parametric tests, in the event of non-verification of one or more of the conditions in question. Finally, in order to illustrate the steps of the implementation of these tests, examples of numerical application have been presented.

Keywords : Statistical tests, (non)parametric tests, decision risk, test statistics.