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 en vue de l'obtention du Diplôme **:**

MASTER en Mathématiques

Option : **Statistique**

Par

Bouden Ahlem

Titre :

Tests de comparaison de deux distributions

Membres du Comité d'Examen :

Dr.	BENAMEUR Sana	UMKB	Président
Pr.	MERAGHNI Djamel	UMKB	Encadreur
Dr.	BERKANE Hassiba	UMKB	Examinateur

Septembre 2020

Dédicace

Je dédie cet humble travail

A mes chers parents, pour tous leurs sacrifices, leur amour, leur tendresse, leur soutien et leurs prières durant toutes mes études.

A ma chère sœur Naima pour son encouragement permanent et son soutien moral.

A mes chers frères, Rabeh et Nacer Eddine, pour leur appui et encouragement.

A toute ma famille pour leur soutien tout au long de mon parcours universitaire.

A toutes mes amies, particulièrement Halima Ben Hamza et Maroua Attaf.

REMERCIEMENTS

Je remercie tout d'abord ‹‹ALLAH›› de m'avoir donné le courage d'entamer et de finir ce mémoire dans de bonnes conditions.

Je remercie le professeur MERAGHNI Djamel, d'avoir encadré ce travail ainsi que les membres du jury de soutenance.

Enfin, je voudrais exprimer ma gratitude aux personnes suivantes

Soumia Aboud et Fouzia Mansouri

pour leur aide à répondre à mes différentes questions.

Table des matières

Table des figures

Liste des tableaux

Introduction

Les statistiques développent des techniques et des méthodes qui permettent d'analyser les données issues de l'observation, afin de cerner les caractéristiques de la population concernée et d'indentifier un modèle capable d'engendrer ces données. En statistique, à travers un ensemble de données, on effectue un test entre deux hypothèses, et au final on rejette ou non rejette une hypothèse statistique, appelée hypothèse nulle.

Pour effectuer un test d'hypothèses, on commence par faire une hypothèse sur la population considérée ou un de ses paramètres. Cette hypothèse est appelée "hypothèse nulle H_0", et c'est l'hypothèse que l'on désire contrôler : elle consiste à dire qu'il nexist 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 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 que H_1 est vraie.

Il existe deux types de tests : test paramétrique si la répartition de la population est connue, et test non paramétrique si la répartition de la population est inconnue. Lorsque l'on dispose de deux échantillons, la questions se pose de savoir s'ils proviennent de la même population on verra qu'avant d'effectuer un test, il faudra déterminer si les deux échantillons sont indépendants ou appariés.

• **Données indépendantes :** les observations sont indépendantes à l'intérieur de chaque échantillon et d'un échantillon à l'autre. Exemple : dosage d'un produit chez deux groupes de patientes ayant reçu une molécule ou un placebo.

• **Données appariées :** les même individus sont soumis à deux mesures successives d'une même variable. Exemple : dosage d'un produit avant et après un traitement chez les même individus.

En plus d'une introduction et d'une conclusion, ce mémoire se compose de deux chapitres.

- Le premier chapitre : on passe en revue un certain nombre de points des plus importants sur les tests : principe des tests, catégories de tests et éléments d'un test (hypothèses (nulle et alternative), risques (première et deuxième espèces), variable de décision,...).

- Le deuxième chapitre : on présente de façon détaillée les tests de comparaison de deux distributions : tests paramétriques et tests non paramétriques.

Tests paramétriques : comparaison de deux moyennes et de deux variances dans deux cas : "cas d'échantillons indépendants" et "cas d'échantillons appariés".

Tests non paramétriques :

Cas d'échantillons indépendants : on a par exemple le test de Kolmogrov-Smirnov et celui de Wilcoxon-Mann-Whitney.

Cas d'échantillons appariés : on a le test des signes et le test des rangs signés de Wilcoxon.

Dans ce chapitre, on utilisera les logiciels R et RStudio, Excel 2007 et PASW Statistics 18 pour des applications numériques.

Chapitre 1

Généralités sur les tests d'hypothèses

Un test d'hypothèse est un procédure statistique de décision (rejeter ou accepter une hypothèse) qui repose sur l'étude d'un ou plusieurs échantillons aléatoires pour valider des hypothèses liées à un ou plusieurs groupes.

1.1 Principe des tests

Le principe des tests d'hypothèses est de poser une hypothèse de travail et de prédire ses conséquences sur l'échantillon. On compare ces prédictions avec les observations et l'on conclut en acceptant ou en rejetant l'hypothèse de travail. Définir les hypothèses de travail, constitue un élément important dans le choix des hypothèses ainsi que la vérification des conditions de leur application.

1.2 Eléments d'un test

1.2.1 Hypothèses

Les hypothèses a tester notées par H_0 et H_1 sont respectivement appellées hypothèse nulle et hypothèse alternative.

Dans le cas d'un test paramétrique on distingue deux types d'hypothèses : les hypothèses simples et les hypothèses composées (multiples).

Une hypothèse H est dite simple si elle a la forme" $\theta = a$" où $a \in \Theta$, avec Θ désignant l'ensemble de toutes les valeurs de θ.

Une hypothèse H est dite composée ou multiple si elle a la forme "$\theta \in A$" où $A \subset \Theta$ ayant deux élements ou plus.

– Une hypothèse H est dite **bilatérale** si elle est de la forme $\theta \neq a$.

– Une hypothèse H est dite **unilatérale à gauche** si elle est de la forme $\theta < a$.

– Une hypothèse H est dite **unilatérale à droite** si elle est de la forme $\theta > a$.

1.2.2 Risques

Le risque est la probabilité de l'erreur qu'on commet lors de la prise de décision. On a deux types de risque :

1. **Risque de première espèce :** probabilité, notée α, de rejeter H_0 alors qu'elle est vraie.

$$\alpha := P(\text{rejeter } H_0/H_0 \text{ est vraie}).$$

   La valeur maximale que peut prendre le risque de première espèce est appelé niveau ou seuil de signification du test.

2. **Risque de deuxième espèce** : probabilité, notée β, d'accepter H_0 alors qu'elle est fausse.

$$\beta := P(\text{accepter } H_0/H_0 \text{ est fausse})$$

On résume les quatre situations possibles dans le tableau 1.1.

Décision (accepter)\vraie	H_0	H_1
H_0	$1-\alpha$	β
H_1	α	$1-\beta$

TAB. 1.1 – Tableau de décision

Remarque 1.2.1 *Les seuils de signification les plus utilisés sont* $\alpha = 0.10,\ 0.05$ *et* 0.01.

1.2.3 Variable de décision

C'est une statistique qui apporte le plus d'informations sur le problème posé. C'est une variable aléatoire, définie indépendamment des données observées. On l'appelle aussi "statistique de test". La statistique qui convient pour le test est celle dont la valeur observée est utilisée pour décider du ‹‹rejet›› ou du ‹‹non - rejet›› de H_0. La distribution d'échantillonnage de cette statistique est déterminée en supposant que l'hypothèse H_0 est vraie.

1.2.4 Région de rejet - Région d'acceptation

La région du rejet d'un test est l'ensemble des observations pour lesquelles l'hypothèse nulle H_0 est ecartée au profit de l'alternative H_1. On l'applle aussi " Région critique du test " et on la note généralement par W. Elle est définie par l'équation

$$P(W/H_0) = \alpha.$$

Le complémentaire de W est la région d'acceptation du test, on la note par $\overline{W}$. D'après le tableau 1.1, on a

$$P(W/H_0) = \alpha,\ P(\overline{W}/H_0) = 1 - \alpha,\ P(\overline{W}/H_1) = \beta \text{ et } P(W/H_1) = 1 - \beta.$$

1.2.5 Puissance d'un test

Définition 1.2.1 *On appelle puissance d'un test, la probabilité de rejeter H_0 lorsqu'elle est fausse. En d'autres termes, c'est la probabilité d'accepter H_1 alors qu'elle est vraie. On la note par π. Sa valeur est $1 - \beta$.*

$$\pi := P(\textit{rejeter } H_0/H_0 \textit{ est fausse}) = P(W/H_1) = 1 - \beta.$$

1.2.6 Notion de p-valeur

Les logiciels de calcul statistique expriment souvent le résultat d'un test en fournissant une grandeur appelée p-valeur (p-value en anglais). C'est une probabilité critique qu'on note par α_{obs}. Si on désigne par S la statistique du test et par S_{obs} sa valeur échantionnale, alors pour un test unilatéral à gauche

1.4.1, on a

$$\alpha_{obs} = p - value := P(S \leq S_{obs}).$$

Dans le cas d'un test unilatéral à droite 1.4.1, on a

$$\alpha_{obs} = p - value := P(S \geq S_{obs}),$$

et pour un test bilatéral 1.4.1, on a

$$\alpha_{obs} = p - value := P(|S| \geq |S_{obs}|).$$

1.2.7 Règle de décision

Pour un niveau de signification $0 < \alpha < 1$ fixé, il existe deux façons de prendre la décision.

- **Régle de décision 1 :** Si S_{obs} appartient à la région critique W alors l'hypothèse H_0 est rejetée au risque d'erreur α. Sinon elle ne peut être rejetée.
- **Régle de décision 2 :** On compare entre α_{obs} et α et on décide.

Remarque 1.2.2 *Les meilleures décisions sont obtenues à travers les tests les puissants. Le lemme de Neyman-Pearson permet de construire les tests uniformément les plus puissants. Pour les détails sur ce lemme et ses applications, voir [11], page 330.*

1.3 Demarche d'un test

Différentes étapes doivent être suivies pour tester une hypothèse :

1. Formuler les hypothèses H_0 et H_1.
2. Fixer le niveau de signification α.
3. Déterminer la région critique W.
4. Etablir la règle de décision.

5. Calculer la valeur empirique (observé) de la variable de décision.

6. Prendre la décision : rejeter ou accepter H_0.

1.4 Catégories des tests

1.4.1 Test paramétrique

Un test est dit paramétrique si la distribution de la population est connue. L'objet du test est de vérifier certaines hypothèses sur un ou plusieurs paramètres de cette distribution. Les tests paramétriques les plus usuels sont :

· test bilatéral : test de H_0 : "$\theta = \theta_0$" contre H_1 : "$\theta \neq \theta_0$".

· tests unilatéraux : test de H_0 : "$\theta \leq \theta_0$" contre H_1 : "$\theta > \theta_0$" et test de H_0 : "$\theta \geq \theta_0$" contre H_1 : "$\theta < \theta_0$".

Les tests paramétriques les plus connues sont le test de conformité et le test d'égalité.

1.4.2 Test non paramétrique

Un test est dit non paramétrique si la distribution de la population est inconnue. L'objet du test est de décider, au vu de l'observation d'un échantillon si cette population suit une loi de probabilité spécifiée par l'hypothèse nulle.

Les tests non paramétriques les plus répandus sont les tests d'ajustement, les tests d'indépendance et les tests de comparaison de distributions.

Chapitre 2

Comparaison de deux distributions

En comparant des distributions de même qualité entre deux populations, les techniques statistiques utilisées dépendent du type de caractère étudié, qualitatif ou quantitatif, des tailles des échantillons et de leur indépendance ou non. Pour le caractère qualitatif, et des tailles d'échantillons sufisamment grandes ($n > 30$), on utilise des tests de khi-deux. Pour le caractère quantitatif, lorsque les distributions sont supposées normales, il suffit de faire une comparaison entre deux méthodes basées sur la loi de Student.

Les distributions ne peuvent pas être considérées comme normales, lorsque les tailles d'échantillons sont petites ($n < 30$), il est préférable d'utiliser des tests non paramétriques.

2.1 Tests paramétriques

Soient X et Y deux populations d'espérances μ_1 et μ_2 et de variances σ_1^2 et σ_2^2 respectivement. On extrait un échantillon $(X_1, ..., X_{n_1})$, de taille $n_1 \geq 1$, de X et un autre $(Y_1, ..., Y_{n_2})$, de taille $n_2 \geq 1$, de Y. On suppose que les distributions de X et Y sont gaussiennes (ou peuvent être approximées par une loi normale) :

$$X \sim \mathcal{N}(\mu_1, \sigma_1^2) \text{ et } Y \sim \mathcal{N}(\mu_2, \sigma_2^2).$$

2.1.1 Test de comparaison de deux moyennes

On se propose de tester si la moyenne μ_1 de la première population est égale ou non la moyenne μ_2 de la deuxième population. En d'autres termes, on veut tester l'hypothèse H_0 contre l'hypothèse H_1 :

$$\begin{cases} H_0 & \mu_1 = \mu_2 \\ H_1 & \mu_1 \neq \mu_2 \end{cases} \qquad \text{(cas bilatéral).}$$

En considérant la variable aléatoire $Z = X - Y$, ce test revient à vérifier si la moyenne $\mu_Z := E(Z) = \mu_1 - \mu_2$ est nulle ou non. On désigne par

$$\overline{X} := \frac{1}{n_1}\sum_{k=1}^{n_1} X_k \text{ et } \overline{Y} := \frac{1}{n_2}\sum_{k=1}^{n_2} Y_k,$$

les moyennes empiriques de X et Y respectivement. Alors la moyenne empirique de la variable Z est $\overline{Z} = \overline{X} - \overline{Y}$. Elle est de moyenne $\mu_{\overline{Z}} = \mu_1 - \mu_2$ et de variance

$$\sigma^2_{\overline{Z}} := V(\overline{Z}) = V(\overline{X}) + V(\overline{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2},$$

car $\overline{X}$ et $\overline{Y}$ sont indépendantes. On a donc

$$\overline{Z} \sim \mathcal{N}(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}).$$

1. Cas d'échantillons indépendants

– **Variances connues :** sous l'hypothèse H_0, on a

$$D := \frac{\overline{Z}}{\sigma_{\overline{Z}}} = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1).$$

Pour un risque d'erreur α fixé, on a

$$P(|D| \leq z_{1-\frac{\alpha}{2}}) = 1 - \alpha,$$

où $z_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$. La région de rejet est

$$]-\infty, -z_{1-\frac{\alpha}{2}}[\cup]z_{1-\frac{\alpha}{2}}, +\infty[.$$

On calcule la valeur de D et on accepte ou on rejette H_0, au risque α, suivant la valeur trouvée.

– **Variances inconnues :** en pratique, on ne connait pas les valeurs des variances théoriques σ_1^2 et σ_2^2. On les approxime, à partir des données, par leurs estimateurs non biaisés respectifs

$$\overline{S}_{n_1}^2 := \frac{1}{n_1-1}\sum_{k=1}^{n_1}(X_k-\overline{X})^2 \text{ et } \overline{S}_{n_2}^2 := \frac{1}{n_2-1}\sum_{k=1}^{n_2}(Y_k-\overline{Y})^2,$$

qui représentent les variances empiriques X et Y.

- **cas 1.** $n_1 > 30$ et $n_2 > 30$: sous l'hypothèse H_0 la variable aléatoire

$$T_1 := \frac{\overline{Z}}{\sqrt{\dfrac{\overline{S}_{n_1}^2}{n_1} + \dfrac{\overline{S}_{n2}^2}{n_2}}} \rightsquigarrow \mathcal{N}(0,1).$$

Pour un risque d'erreur α fixé, on a

$$P(|T_1| \leq z_{1-\frac{\alpha}{2}}) = 1-\alpha,$$

et la région de rejet est donc

$$]-\infty, -z_{1-\frac{\alpha}{2}}[\cup]z_{1-\frac{\alpha}{2}}, +\infty[.$$

On calcule alors la valeur de T_1 et on accepte ou on rejette H_0, au risque α, suivant la valeur trouvée.

- **Cas 2.** $n_1 \leq 30$ ou $n_2 \leq 30$ et $\sigma_1^2 = \sigma_2^2$: sous l'hypothèse H_0, la variable aléatoire

$$T_2 := \frac{\overline{Z}}{\sqrt{\overline{S}_{n_1,n_2}^2(\dfrac{1}{n_1}+\dfrac{1}{n_2})}},$$

suit la loi de Student à $n_1 + n_2 - 2$ degrés de liberté, où

$$\overline{S}^2_{n_1,n_2} := \frac{(n_1 - 1)\overline{S}^2_{n_1} + (n_2 - 1)\overline{S}^2_{n_2}}{n_1 + n_2 - 2},$$

est une estimation de σ_1^2 (ou σ_2^2) meilleure que $\overline{S}^2_{n_1}$ (ou $\overline{S}^2_{n_2}$). Pour un risque d'erreur α fixé, on a

$$P(|T_2| \leq t_{1-\frac{\alpha}{2}}) = 1 - \alpha,$$

avec $t_{1-\frac{\alpha}{2}}$ désignant le quantile d'ordre $1 - \alpha/2$ de la loi ci-dessus. La région de rejet est

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

On calcule alors la valeur de T_2 et on accepte ou on rejette H_0, au risque α, suivant la valeur trouvée.

- **Cas 3.** $n_1 \leq 30$ ou $n_2 \leq 30$ et $\sigma_1^2 \neq \sigma_2^2$: on estime la variance de $\overline{Z}$ par (voir [9], page 15)

$$\hat{\sigma}_Z^2 := \frac{\overline{S}^2_{n_1}}{n_1} + \frac{\overline{S}^2_{n_2}}{n_2}.$$

Sous l'hypothèse H_0, la variable aléatoire

$$T_3 := \frac{\overline{Z}}{\hat{\sigma}_Z},$$

suit la loi de Student à ϑ degrés de liberté, où ϑ est l'entier le plus proche de

$$\frac{\left(\overline{S}^2_{n_1}/n_1 + \overline{S}^2_{n2}/n_2\right)^2}{\overline{S}^4_{n_1}/\left(n_1^2(n_1 - 1)\right) + \overline{S}^4_{n2}/\left(n_2^2(n_2 - 1)\right)}.$$

Pour α fixé, on a

$$P(|T_3| \leq t'_{1-\frac{\alpha}{2}}) = 1 - \alpha,$$

et la région critique est

$$]-\infty, -t'_{1-\frac{\alpha}{2}}[\cup]t'_{1-\frac{\alpha}{2}}, +\infty[,$$

où avec $t'_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\alpha/2$ de la loi ci-dessus. On calcule la valeur de T_3 et on accepte ou on rejette H_0, au risque α, suivant la valeur trouvée.

2. Cas d'échantillons appariés

Dans ce cas, les deux échantillons de X et Y sont de tailles égales $(n_1 = n_2 = n \geq 1)$. On extrait un échantillon $(Z_1, ..., Z_n)$, de taille n, de la variable aléatoire Z. Les valeurs z_i sont donc obtenues par différence des paires de valeurs c-à-d

$$z_i = x_i - y_i, \; i = 1, ..., n.$$

On fait donc un test de Student sur la moyenne μ_Z de Z. Les estimations sans biais respectives de la moyenne de Z et sa variance sont les moyenne et variance empiriques

$$\overline{Z} := \frac{1}{n}\sum_{i=1}^{n} Z_i \text{ et } \overline{S}_Z^2 := \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \overline{Z})^2.$$

La statistique du test s'écrit

$$T = \frac{\overline{Z}}{\overline{S}_Z/\sqrt{n}}.$$

Sous H_0, elle suit la loi de Student à $(n-1)$ degrés de liberté. On rejette l'hypothèse nulle si

$$|T| \geq t_{1-\frac{\alpha}{2}}(n-1).$$

2.1.2 Test de comparaison de deux variances

Le test de comparaison de Fisher compare les variances de deux sous populations. On veut tester les hypothèses

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

On considère le cas où $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, et $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$.

1. Cas d'échantillons indépendants

– **Moyennes connues :** dans ce cas on définit les estimateurs respectifs suivants pour les variances σ_1^2 et σ_2^2 :

$$\overline{U}_{n_1}^2 := \frac{1}{n_1}\sum_{k=1}^{n_1}(X_k-\mu_1)^2 \text{ et } \overline{U}_{n_2}^2 := \frac{1}{n_2}\sum_{k=1}^{n_2}(Y_k-\mu_2)^2.$$

Alors

$$\frac{n_1\overline{U}_{n_1}^2}{\sigma_1^2} \sim \chi^2(n_1) \text{ et } \frac{n_2\overline{U}_{n2}^2}{\sigma_2^2} \sim \chi^2(n_2),$$

et sous l'hypothèse H_0, la variable aléatoire $F_1 := \overline{U}_{n_1}^2/\overline{U}_{n2}^2$ suit la loi $\mathcal{F}(n_1,n_2)$. On rejette H_0, au seuil de risque α, si

$$F_1 \leq f_{\frac{\alpha}{2}}(n_1,n_2) \text{ ou } F_1 \geq f_{1-\frac{\alpha}{2}}(n_1,n_2),$$

où $f_{\frac{\alpha}{2}}(n_1,n_2)$ et $f_{1-\frac{\alpha}{2}}(n_1,n_2)$ sont définis par $P(Z < f_{\frac{\alpha}{2}}(n_1,n_2)) = P(Z > f_{1-\frac{\alpha}{2}}(n_1,n_2)) = \alpha/2$.

– **Moyennes inconnues :** dans ce cas on utilise $\overline{S}_{n_1}^2$ et $\overline{S}_{n_2}^2$ comme estimations de σ_1^2 et σ_2^2 respectivement. On a

$$\frac{(n_1-1)\overline{S}_{n_1}^2}{\sigma_1^2} \sim \chi^2(n_1-1) \text{ et } \frac{(n_2-1)\overline{S}_{n_2}^2}{\sigma_2^2} \sim \chi^2(n_2-1),$$

et sous l'hypothèse H_0, la variable aléatoire $F_2 := \overline{S}_{n_1}^2/\overline{S}_{n_2}^2$ suit la loi $\mathcal{F}(n_1-1,n_2-1)$. La région de rejet est

$$\left]0, f_{\frac{\alpha}{2}}(n_1-1,n_2-1)\right[\cup \left]f_{1-\frac{\alpha}{2}}(n_1-1,n_2-1),+\infty\right[.$$

Proposition 2.1.1 *Le quantile d'ordre α de la loi $\mathcal{F}(a,b)$ est égal à l'inverse du quantile d'ordre $1-\alpha$ de la loi $\mathcal{F}(b,a)$, c-à-d $f_\alpha(a,b) = 1/f_{1-\alpha}(b,a)$.*

2. Cas d'échantillons appariés

On considère n mesures des variables X et Y sur les mêmes individus. Le test peut s'écrire

$$\begin{cases} H_0 & \sigma_1^2/\sigma_2^2 = 1 \\ H_1 & \sigma_1^2/\sigma_2^2 \neq 1 \end{cases}$$

Pour les variables $U = X + Y$ et $V = X - Y$, on a la relation suivante

$$r_{uv} = 0 \Leftrightarrow \frac{\sigma_1}{\sigma_2} = 1,$$

où r_{uv} est le coefficient de corrélation de Pearson. Alors le test d'égalité des variances est équivalent au test d'hypothèses suivant :

$$\begin{cases} H_0 & r_{uv} = 0 \\ H_1 & r_{uv} \neq 0 \end{cases}$$

La statistique du test s'écrit

$$T := \frac{\hat{r}_{uv}}{\sqrt{\dfrac{1 - \hat{r}_{uv}^2}{n-2}}},$$

où $\hat{r}_{uv}$ est l'estimation de r_{uv} sur l'échantillon. Sous H_0, T suit la loi de Student à $(n-2)$ ddl. La région critique du test, au risque α, s'écrit

$$|T| \geq t_{1-\frac{\alpha}{2}}(n-2).$$

2.2 Tests non paramétriques

Soient X et Y deux populations de fonctions de répartition respectives F et G. On considère deux échantillons $(X_1, ..., X_{n_1})$ de X et $(Y_1, ..., Y_{n_2})$ de Y de fonctions de répartition empiriques F_{n_1} et G_{n_2} respectivement.

2.2.1 Cas d'échantillons indépendants

On suppose que les deux échantillons sont indépendants.

1. Test de Kolmogorov-Smirnov

Le test de Kolmogrov-Smirnov de l'hypothèse H_0 : "$F(x) = G(x)$" contre H_1 : "$F(x) \neq G(x)$" est défini par la statistique

$$D_{n_1,n_2} := \sup_{x \in \mathbb{R}} \left| F_{n_1}(x) - G_{n_2}(x) \right|.$$

Il consiste à rejeter l'hypothèse H_0, au risque α, si $D_{n_1,n_2} \geq d_\alpha(n_1, n_2)$. Les seuils $d_\alpha(n_1, n_2)$ sont tabulés pour différentes valeurs des tailles n_1 et n_2 et du niveau de signification α. Ils sont aussi disponibles dans les logiciels statistiques.

Remarque 2.2.1 *Pour faire un test unilatéral* ($H_0 : F(x) = G(x)$ *contre* $H_1 : F(x) \geq G(x)$) *on utilise la statistique de test*

$$D^+_{n_1,n_2} := \sup_{x\in\mathbb{R}}(F_{n_1}(x) - G_{n_2}(x)).$$

2. Test de Cramer-von Mises

Le test de Cramer-von Mises est aussi basé sur la fonction de répartition. Mais, contrairement au test de Kolmogorov-Smirnov, qui se concentre sur l'écart maximal, on utilise la somme des carrés des différence des écarts calculés sur la totalité des observations. La statistique de ce test est donnée par

$$CvM_{n_1,n_2} := \frac{n_1 n_2}{(n_1+n_2)^2} \sum_{i=1}^{n_1+n_2} [F_{n_1}(x_i) - G_{n_2}(x_i)]^2.$$

La décision consiste à rejeter l'hypothèse H_0 si $CvM_{n_1,n_2} \geq c_\alpha(n_1, n_2)$. Les valeurs critiques $c_\alpha(n_1, n_2)$ sont tabulées pour différentes tailles n_1 et n_2 et différents risques α. On peut également les obtenir grâce aux logiciels statistiques.

3. Test de Wilcoxon-Mann-Whitney

On veut tester H_0 : "$F(x) = G(x)$" contre H_1 : "$F(x) \geq G(x)$".

Test de Wilcoxon

Ce test est basé sur la rangement en ordre croissant des observations $(X_1, X_2, ..., X_{n_1}, Y_1, Y_2, ..., Y_{n_2})$. La statistique correpondante est la somme des rangs de X :

$$W_{n_1,n_2} = \sum_{i=1}^{n_1} R(X_i),$$

où $R(X_i)$ le rang de la $i^{ème}$ observation de X dans l'échantillon complet. On a

$$E(W_{n_1,n_2}) = \frac{n_1(n_1+n_2+1)}{2} \text{ et } V(W_{n_1,n_2}) = \frac{n_1 n_2(n_1+n_2+1)}{12}.$$

On en déduit que la statistique

$$Z := \frac{W_{n_1,n_2} - \dfrac{n_1(n_1+n_2+1)}{2}}{\sqrt{\dfrac{n_1 n_2(n_1+n_2+1)}{12}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1), \text{ quand } n \to \infty.$$

Par conséquent, on accepte H_0, au risque α, si $|Z| \leq z_{1-\frac{\alpha}{2}}$, où $z_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale centrée réduite.

Test de Mann-Whitney

Ce test est basé sur la détermination des couples (X_i, Y_j) pour lesquels $X_i \neq Y_j$. Il est défini par la statistique

$$U_{n_1,n_2} := \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} 1_{(X_i > Y_j)}.$$

C'est le nombre de couples (X, Y) tels que $X > Y$. Sous l'hypothèse H_0, on a

$$E(U_{n_1,n_2}) = \frac{n_1 n_2}{2} \text{ et } V(U_{n_1,n_2}) = \frac{n_1 n_2(n_1+n_2+1)}{12}.$$

On en déduit que la statistique

$$Z := \frac{U_{n_1,n_2} - \dfrac{n_1 n_2}{2}}{\sqrt{\dfrac{n_1 n_2(n_1+n_2+1)}{12}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1), \text{ quand } n \to \infty.$$

Par conséquent, la décision est de rejeter H_0, au risque α, si $|Z| > z_{1-\frac{\alpha}{2}}$. En pratique, on utilise ce résultat lorsque $n_1, n_2 \geq 8$.

Remarque 2.2.2 *Les statistiques des tests de Wilcoxon et Mann-Whitney sont liées par la relation suivante :*

$$W_{n_1,n_2} = U_{n_1,n_2} + \frac{n_1\,(n_1+1)}{2}.$$

2.2.2 Cas d'échantillons appariés

1. Test des signes

Le test des signes est une méthode statistique pour tester les différences constantes entre des paires d'observations. On utilise les signes ($+$ et $-$) au lieu de données quantitatives. L'hypothèse nulle peut s'écrire $p_{-} = p_{+} = 1/2$ avec p_{+} la probabilité d'observer une différence positive et p_{-} celle d'observer une différence négative. On pose

$$S^{+} := \sum_{i=1}^{n} 1_{]0,+\infty[}(Y_i - X_i) \text{ et } S^{-} := \sum_{i=1}^{n} 1_{]-\infty,0[}(Y_i - X_i).$$

Sous l'hypothèse H_0 et pour n paires d'observations, le nombre de différences positives (ou négatives) est une variable aléatoire binomiale $N \sim \mathcal{B}(n, 1/2)$. Le test permet de comparer, grâce à cette distribution, le nombre observé de signes plus (ou moins) et le nombre attendu $E(N) = n/2$. Quand certaines différences sont nulles, les paires d'observations correspondantes sont écartées de l'analyse et la valeur de n est réduite en conséquence.

Il est possible d'obtenir directement la probabilité critique ($p - value$) du test bilatéral en calculant la quantité

$$p = 2P[\mathcal{B}(n, 1/2) \geq \max(S, n - S)].$$

On rappelle que la probabilité cumulée d'une loi binomiale s'écrit

$$\begin{aligned} P[\mathcal{B}(n,\pi) \geq s] &= \sum_{j=s}^{n} \binom{n}{j} \pi^j (1-\pi)^{n-j} \\ &= \sum_{j=s}^{n} \frac{n!}{j!(n-j)!} \pi^j (1-\pi)^{n-j}. \end{aligned}$$

2. Test des rangs signés de Wilcoxon

Ce test est utilisé comme alternative au test de Student lorsque la distribution ne présente pas de normalité. Pour calculer sa statistique, on range par ordre croissant les valeurs absolues des différences $|Z_1|, ..., |Z_n|$. On note r_i le rang de $|Z_i|$ et on désigne par T^{+} la somme des rangs des diférences positives et par T^- la somme des rangs des différences négatives. La région critique du test, au risque α, s'écrit

$$T \leq T_\alpha(n),$$

où les valeurs du seuil $T_\alpha(n)$ sont tabulées selon les valeurs de α et n.

Remarque 2.2.3

1. *On a*
$$T^+ + T^- = \frac{n(n+1)}{2}.$$

2. *Sous l'hypothèse H_0, les variables T^+ et T^- suivent la même loi de probabilité.*

Proposition 2.2.1 *Sous l'hypothèse H_0, on a*

$$E(T^+) = \frac{n(n+1)}{4} \text{ et } V(T^+) = \frac{n(n+1)(2n+1)}{24}.$$

De plus,
$$Z := \frac{T^+ - \dfrac{n(n+1)}{4}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1), \text{ quand } n \to \infty.$$

En pratique, on applique ce résultat lorsque $n > 30$. Si $n \leq 30$ la statistique du test ne suit pas une loi usuelle.

2.3 Exemples d'application

Dans cette partie, on traite quelques exemples de comparaison de deux distributions. On considère les deux types de tests : paramétriques et non paramétriques. On utilise les programmes suivants : R, RStudio et Excel, PASW Statistics 18.

2.3.1 Tests paramétriques

Test de comparaison de deux moyennes - commande R : t.test

Exemple 2.3.1 ([15]) *On vérifie s'il y a une différence significative entre deux appareils de mesure A et B (présentés dans le tableau 2.1), au niveau de signification* 5%, *dans les deux cas suivants.*

1. *On considère que les* 2 *échantillons sont indépendants.*
2. *On suppose maintenant que ce sont les mêmes feuilles qui sont analysées par deux méthodes.*

A	**B**	**A**	**B**	**A**	**B**
12.0	10.1	24.5	23.6	69	64.8
12.1	13.5	5.0	4.9	11.8	12.0
7.5	8.5	47.9	47.8	20.0	17.5
8.0	9.6	43.1	46.7	57.6	55.2
16.0	16.8	38.2	38.3	15.0	14.8

TAB. 2.1 – Exemple sur le test de comparaison de deux moyennes

1- Cas d'échantillons indépendants

A l'aide du logiciel Excel, on calcule les effectif, moyenne et écart type dans chaque groupe (2.2). On a $n_1 = 15$, $\overline{x} = 25.85$ et $\overline{S}_{n_1} = 20.223$ (resp. $n_2 = 15$, $\overline{y} = 25.61$ et $\overline{S}_{n_2} = 19.502$).

L'écart entre les moyennes est $\overline{z} = \overline{x} - \overline{y} = 0.240$. L'estimation de la variance est $\overline{S}^2_{n_1,n_2} = 394.635$. D'où l'estimation de l'écart type de $\overline{Z}$ est $\hat{\sigma}_{\overline{Z}} = \sqrt{394.635\,(1/15 + 1/15)} = 7.254$. Finalement, on calcule la statistique du test $T_2 = 0.240/7.254 = 0.033$. Le nombre de ddl est $n_1 + n_2 - 2 = 28$. Au risque $\alpha = 5\%$, on compare T_2 avec le seuil $t_{0.975}(28) = 2.048$ et on conclut que les données sont compatibles avec l'hypothèse d'égalité des moyennes.

On aboutit à la même conclusion si on utilise la probabilité critique du test ($p - value = 0.974 > 0.05 = \alpha$). On résume tous les résultats dans le tableau 2.2.

	A	B
moyenne	25.847	25.607
variance	408.958	380.311
observations	15	15
variance pondérée	394.635	
$\overline{Z}$	0.240	
sigma($\overline{Z}$)	7.254	
statistique T_2	0.033	
ddl	28	
valeur critique de t	2.048	
p-value	0.974	

TAB. 2.2 – Test de comparaison de 2 appareils de mesure A et B sur des feuilles-Echantillons indépendants

La boîte à moustaches de la figure 2.1 permet d'avoir une idée à la fois sur des paramètres de localisation et de dispersion des variables dans chaque sous groupe. De plus, on peut détecter visuellement les éventuels points atypiques.

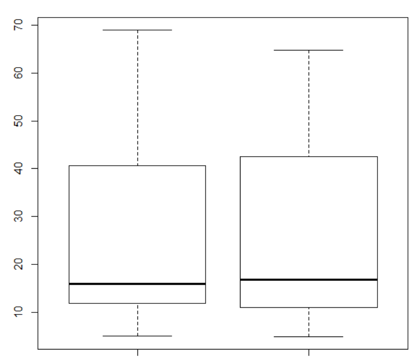

FIG. 2.1 – Boîtes à moustaches du test de 2 appareills de mesure A et B sur des feuilles.

2- Cas d'échantillons appariés

La répésentation graphique des appareils de mesures A et B est donnée dans la figure 2.2.

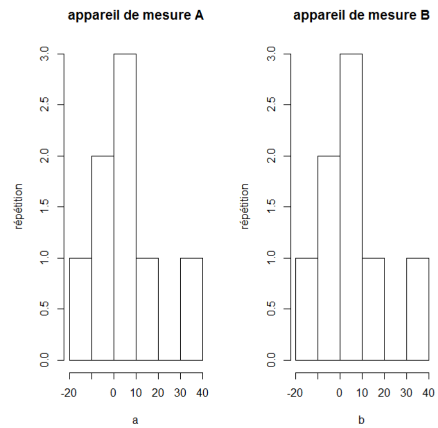

FIG. 2.2 – La répésentation graphique des appareills de mesures A et B.

Pour trouver "la différence significative entre les appareills de mesure A et B". On construit la variable Z, ensuite, on vérifie que sa moyenne s'écarte significativement de la valeur 0.

On dispose de $n = 15$ différences Z obtenues à partir des données du tableau 2.1. Leurs moyenne et écart type sont $\overline{z} = 0.24$ et $s_{\overline{z}} = 1.93$. La statistique du test est $T = 0.480$. Au risque $\alpha = 5\%$, le seuil critique est $t_{0.975}(14) = 2.145$.

On a $T < t_{0.975}(14)$ ou $p - value = 0.639 > 0.05 = \alpha$. Alors, on conclut qu'il n'y a pas de différence significative entre les appareills de mesures A et B. On résume tous les résultats dans le tableau 2.3.

observations	15
$\overline{Z}$	0.240
sigma($\overline{Z}$)	1.930
statistique T	0.480
ddl	14
valeur critique de t	2.145
p-value	0.639

TAB. 2.3 – Test de comparaison de 2 appareills de mesure A et B sur des feuilles-Echantillons appariés

Test de comparaison de deux variances - commande R : var.test

Exemple 2.3.2 ([16]) *Les mesures A et B sont des variables aléatoires suivant des lois normales de moyennes μ_1 et μ_2 et de variances σ_1^2 et σ_2^2 toutes inconnues. Tester au seuil de* 5% *l'égalité de ces deux dernières dans le cas d'échantillons indépendants et appariés.*

A	63.47	62.52	69.44	63.00	77.70	76.35	70.78	80.2	59.6	56.68
B	68.32	73.45	77.89	72.44	63.74	76.25	65.21	53.54	72.36	71.19

TAB. 2.4 – Exemple sur le test de comparaison de deux variances

1. Cas d'échantillons indépendants

L'outil Excel fournit directement les variances conditionnelles (tableau 2.5) :

	moyenne	variance	écart type	n	ddl	F_2	$f_{0.025}(9,9)$	$f_{0.975}(9,9)$	p-value
A	67.974	66.30	8.142	10	9	1.30	0.248	4.030	0.702
B	69.439	50.99	7.141	10	9				

TAB. 2.5 – Test de comparaison de 2 mesures A et B-Echantillons indépendants

On constate que la valeur de $\overline{S}_{n_1}^2$ est supérieure à celle de $\overline{S}_{n_2}^2$. Pour comparer les 2 variances σ_1^2 et σ_2^2, on forme le rapport $F_2 = \overline{S}_{n_1}^2/\overline{S}_{n_2}^2$ et on compare sa valeur, égale à 1.30, aux fractiles de la loi de Fisher à $(9,9)$ ddl qui sont égaux à 0.248 et 4.03. On constate que $f_{0.025}(9,9) < F_2 < f_{0.975}(9,9)$ ou que $p - value = 0.702 > 0.05 = \alpha$. Donc, on accepte l'hypothèse d'égalité des variances, au risque 5%.

2. Cas d'échantillons appariés

En utilisant Excel, on obtient les résultats du tableau 2.6. Pour les variables synthétiques U et V, on calcule leur corrélation $\hat{r}_{uv} = 0.152$. La valeur de la statistique du test est $T = 0.434$. Au risque $\alpha = 5\%$, on compare cette dernière avec la valeur critique $t_{0.975}(8) = 2.306$. Puisque $T < t_{0.975}(8)$, alors on accepte le fait que les deux variances sont égales.

A	63.47	62.52	69.44	63	77.7	76.35	70.78	80.2	59.6	56.68
B	68.32	73.45	77.89	72.44	63.74	76.25	65.21	53.54	72.36	71.19
u=x+y	131.79	135.97	147.33	135.44	141.44	152.6	135.99	133.74	131.96	127.87
v=x-y	-4.85	-10.93	-8.45	-9.44	13.96	0.1	5.57	26.66	-12.76	-14.51
n	ddl	corrélation	T	$t_{0.975}(8)$						
10	8	0.152	0.434	2.306						
10	8									

TAB. 2.6 – Comparaison de test de 2 mesure A et B-Echantillons appariés

La figure 2.3 représente les histogrammes A et B.

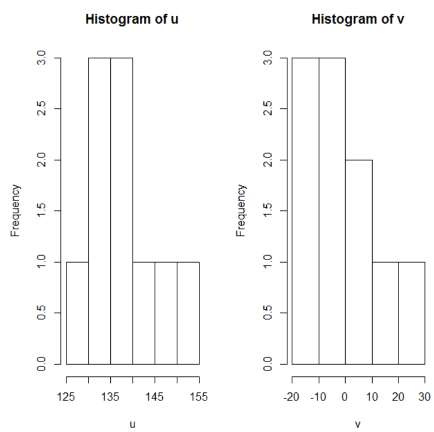

FIG. 2.3 – La répésentation graphique des mesures A et B.

2.3.2 Tests non paramétriques

Cas d'échantillons indépendants

Test de Kolmogorov-Smirnov - commande R : ks.test

Exemple 2.3.3 ([10], page 20) *Dans cet exemple, on étudie le degré de différence de concentration entre les femmes et les hommes selon l'âge. Pour cela, on applique le test de Kolmogorov-Smirnov avec un risque* $\alpha = 5\%$.

femme (X)	17	15	13	17	20	14	22	25	14
homme (Y)	15	14	13	15	18	19	20	24	

TAB. 2.7 – Exemple sur le test de kolmogrov-Smirnov

La distance $D_{n_1,n_2} = 0.167$ et la valeur critique du test à 5% est $d_{0.05}(9,8) = 0.639$. Puisque $D_{n_1,n_2} < d_{0.05}(9,8)$, alors on conclut, au risque 5%, que le degré de différence de concentration entre les femmes et les hommes est le même.

D'autre part, la fonction *ks.test* du logiciel R donne une p-value dépassant 0.999. Cette dernière étant supérieure au seuil α, on peut confirmer la conclusion ci-dessus. Le code R de cet exemple est le suivant :

femme= c(17,15,13,17,20,14,22,25,14)

homme= c(15,14,13,15,18,19,20,24)

ks.test(femme,homme)

On remarque qu'on trouve le même resultat on utilisant la réprésentation graphique de la figure 2.4.

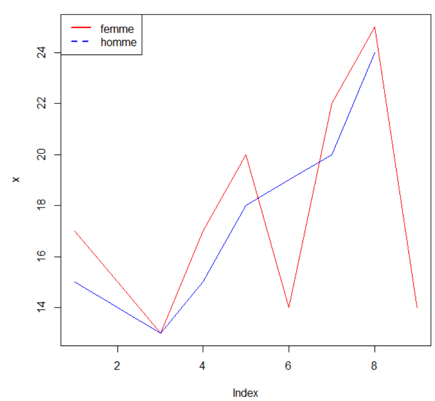

FIG. 2.4 – Les fonctions de répartition empirique de X et Y.

Test de Cramer-von Mises(CvM) - commande R : cvm_test

Exemple 2.3.4 ([10], page 24) *On souhaite comparer, au risque $\alpha = 5\%$, l'indice de masse corporelle des lycéens masculins d'une classe de terminale selon le niveau de leur activité sportive : journalier ($n_1 = 7$ élèves) ou jamais ($n_2 = 9$ élèves).*

journalier	12.4	30.5	33.0	24.8	17.3	22.2	28.6		
jamais	17.0	23.0	23.4	33.3	25.8	27.1	30.0	32.0	18.6

TAB. 2.8 – Exemple de test de Cramer-Van Mises

La valeur de la statistique de ce test est $CvM_{7,9} = 0.039$ et sa valeur critique est $c_\alpha(7,9) = 0.778$. Les détails des calculs par le logiciel Excel sont résumés dans le tableau 2.9. Puisque $CvM_{7,9} < c_\alpha(7,9)$, alors on conclut, au risque 5%, de compatibilité des données avec l'hypothèse H_0.

Le code R de cet exemple est le suivant :

library(twosamples)

journalier= c(12.4,30.5,33,24.8,17.3,22.2,28.6)

jamais= c(17,23,23.4,33.3,25.8,27.1,30,32,18.6)

cvm_test(journalier,jamais)

La $p-value = 0.821$ est supérieure au seuil $\alpha = 0.05$, la conclusion ci-dessus est confirmée. On obtient le même résultat en utilisant la fonction de répartition empirique (figure 2.5).

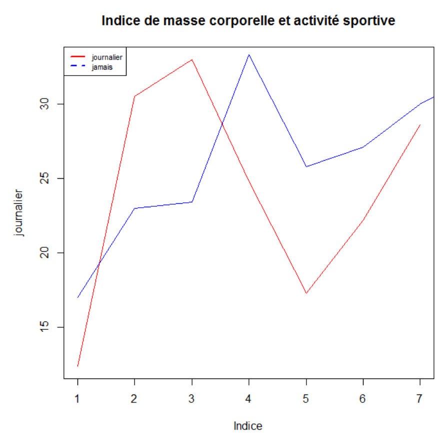

FIG. 2.5 – Indice de masse corporelle selon l'activité sportive - Fonctions de répartition.

x	journalier	jamais	F_{n_1}	G_{n_2}	écart	écart^2
12.4	1		0.143	0.000	0.143	0.020
17		1	0.143	0.111	0.032	0.001
17.3	1		0.286	0.111	0.175	0.030
18.6		1	0.286	0.222	0.063	0.004
22.2	1		0.429	0.222	0.206	0.043
23		1	0.429	0.333	0.095	0.009
23.4		1	0.429	0.444	−0.016	0.000
24.8	1		0.571	0.444	0.127	0.016
25.8		1	0.571	0.555	0.016	0.000
27.1		1	0.571	0.667	−0.095	0.009
28.6	1		0.714	0.667	0.0477	0.002
30		1	0.714	0.778	−0.063	0.004
30.5	1		0.857	0.778	0.079	0.006
32		1	0.857	0.889	−0.032	0.001
33	1		1	0.889	0.111	0.012
33.3		1	1	1	0	0
total	7	9			somme	0.159

Tab. 2.9 – Test de Cramer-von Mises : comparaison des indices de masse corporelle

Test de Wilcoxon-Mann-Whitney

Exemple 2.3.5 ([14]) *Un enseignant a posé deux séries de questions à deux groupes d'étudiants. le premier groupe est formé par les étudiants assis sur des sièges impairs et le deuxième par ceux qui sont assis sur des sièges pairs. Au risque 5%, tester si les deux séries de question sont équivalentes ou non. Les données, présentées dans le tableau 2.10, sont représentées graphiquement par la figure 2.6.*

groupe 1	50	40	10	55	30	22
groupe 2	28	30	45	24	45	36

TAB. 2.10 – Exemple de Test de Wilcoxon-Mann-Whitney

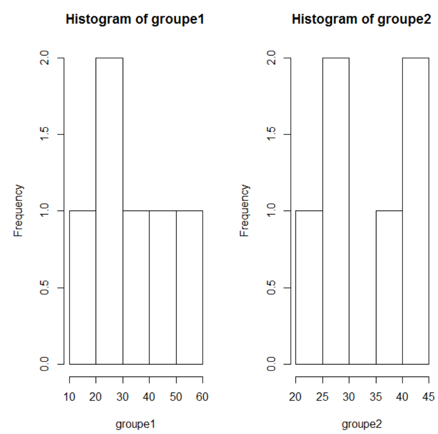

FIG. 2.6 – La répésentation graphique de groupe 1 et groupe 2.

Test de Wilcoxon - commande R : wilcox.test

Le logiciel Excel donne les résultats résumés dans le tableau 2.11.

X	Y	X ordonné	rang(X)
50	28	10	11
40	30	22	8
10	45	24	1
55	24	28	12
30	45	30	5.5
22	36	30	2
		somme rang(X)	39.5

TAB. 2.11 – Test de Wilcoxon-Comparaison de deux séries de question

La somme des rangs de la variable X représentant le premier groupe d'étudiants est $W_{n_1,n_2} = 39.5$, sa moyenne et variance sont toutes les deux égales à 39. La valeur de la statistique du test est $Z = 0.08$ et le seuil critique, au risque 5%, est $z_{0.975} = 1.96$. On constate que $|Z| < 1.96$, donc les deux séries sont équivalentes.

Le code R de cet exemple est le suivant :

groupe1= c(50,40,10,55,30,22)

groupe2= c(28,30,45,24,45,36)

wilcox.test(groupe1,groupe2)

La $p - value = 1 > \alpha$ permet de confirmer la conclusion ci-dessus. On btient le même résultat grâce à la fonction de répartition empirique de la figure 2.7.

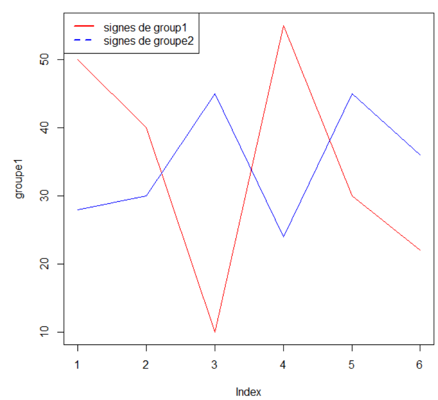

FIG. 2.7 – Les signes de groupe1 et de groupe2.

Test de Mann-Whitney

En utilisant PASW Statistics 18, on obtient les résultats du tableau 2.12.

groupe	n	somme des rangs	moyenne des rangs
groupe 1	6	39.5	6.58
groupe 2	6	38.5	6.42
total	12		

TAB. 2.12 – Test de Mann-Whitney -Comparaison de deux séries de question

La statistique du test est $U_{n_1,n_2} = \sum_{i=1}^{6}\sum_{j=1}^{6} 1_{(X_i>Y_j)} = 6+4+0+6+1.5+0 = 17.5$ avec $E(U_{n_1,n_2}) = 18$

et $V(U_{n_1,n_2}) = 39$. La variable $Z = (U_{n_1,n_2} - E(U_{n_1,n_2})) / \sqrt{V(U_{n_1,n_2})} = -0.08$. Au risque 5%, on a $|Z| = 0.08 < 1.96 = z_{0.975}$, donc les deux séries en question sont équivalentes.

Le code R de cet exemple est le suivant :

x= c(50,40,10,55,30,22)

y= c(28,30,45,24,45,36)

wilcox.test(x,y)

On a $p - value = 1 > 0.05 = \alpha$, donc on conclut le même résultat au risque $\alpha = 5\%$.

Cas d'échantillons appariés

Test des signes - commande R : binom.test

Exemple 2.3.6 *([13])On prend* 15 *jeunes diplômés recrutés par une entrepricse. Les augmentations salariales de la première année sont relativement homogènes et liées à leurs diplômes. Puis, on regarde les augmentations de l'année suivante. Comparez les augmentations de salaire pour les deux années, au risque* 5%.

	année 1	année 2		année1	année2
Bob	250	800	Jim	260	210
Gus	250	900	Ida	220	140
Ali	360	650	Guy	250	150
Léo	300	220	Zoé	300	150
Joe	210	150	Roy	200	90
Luc	250	120	Eva	240	120
Lou	200	110	Tom	210	100
Léa	210	190			

TAB. 2.13 – Exempele de Test des signes-Première et la deuxième année.

En utilisant Excel, on obtient les résultats suivants (tableau 2.14).

année 1	année 2	signe (écart)	année 1	année 2	signe (écart)
250	800	$-$	260	210	$+$
250	900	$-$	220	140	$+$
360	650	$-$	250	150	$+$
300	220	$+$	300	150	$+$
210	150	$+$	200	90	$+$
250	120	$+$	240	120	$+$
200	110	$+$	210	100	$+$
210	190	$+$			

TAB. 2.14 – Augmentations de salaire pour la première et la deuxième années

On obtient le reste des résultats en utilisant PASW Statistics 18 (tableau 2.15).

	n
différences négatives	3
différences positives	12
Tiesc	0
total	15

TAB. 2.15 – Test des signes- Echantillons appariés

On a

$$\begin{aligned} P[\mathcal{B}(n,1/2) \geq \max(S, n-S) \max(12, 15-12) = 12] \\ &= P[\mathcal{B}(n,1/2) = 12] + P[\mathcal{B}(n,1/2) = 13] + P[\mathcal{B}(n,1/2) = 14] + P[\mathcal{B}(n,1/2) = 15] \\ &= 0.028 + 0.003 + 0 + 0 = 0.031. \end{aligned}$$

La probabilité critique du test est égale à $p = 2 \times 0.031 = 0.062$. Au risque 5%, on a $p = 0.062 < \pi = 1/2$, alors on rejette l'hypothèse d'égalité des augmentations des salaires des deux années.

Le code R de cet exemple est le suivant :

année1<- c(250, 250,360,300,210,250,200,210,260,220,250,300,200,240,210)

année2<- c(800,900,650,220,150,120,110,190,210,140,150,150,90,120,100)

diff<- (année1-année2)

nmoins <- sum(diff<0)

nplus <- sum(diff>0)

binom.test(min(nplus,nmoins),nplus+nmoins)

On a $p - value = 0.035 < 0.05 = \alpha$, ce qui permet de faire la même conclusion que ci-dessus. Ainsi, on rejette H_0 et on conclut que les augmentations salariales sont différentes. Il est à noter qu'on trouve le même résultat en utilisant la réprésentation graphique de la figure 2.8.

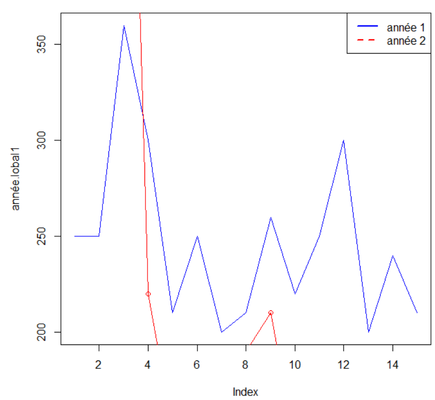

Fig. 2.8 – La fonction de répartition des augmentations de salaire pour la première et la deuxième année.

Test des rangs signés de Wilcoxon - commande R : wilcox.test

Exemple 2.3.7 ([10], page 132) *On veut savoir si un entraînement régulier a un effet sur la tension artérielle. On a recueilli la tension systolique de $n = 8$ personnes, on leur a fait suivre un programme d'entraînement spécifique pendant 6 mois, puis on leur a mesuré la tension (tableau 2.16).*

	avant	après
1	160	125
2	120	135
3	125	120
4	146	130
5	130	128
6	120	124
7	158	150
8	150	153

TAB. 2.16 – Exemple de test des rangs signés de Wilcoxon

En utilisant Excel, on obtient les résultats du tableau 2.17.

avant	après	écart	$\lvert$écart$\rvert$	rang($\lvert$écart$\rvert$)	rang signé
160	125	35	35	8	8
120	135	-15	15	6	-6
125	120	5	5	4	4
146	130	16	16	7	7
130	128	2	2	1	1
120	124	-4	4	3	-3
158	150	8	8	5	5
150	153	-3	3	2	-2

TAB. 2.17 – Evolution de la tension artérielle après entraînement - Test des rangs signés

Par le biais du logiciel PASW Statistics 18, on a les résultats qu'on résume dans la tableau 2.18.

		n	somme des rangs	moyenne des rangs
avant−après	rangs négatifs	3	11	3.67
	rangs positifs	5	25	5
	égalités	0		
	total	8		

	avant-après
Z	-0.98
$p-value$	0.327

TAB. 2.18 – Test des rangs signés de Wilcoxon- Echantillons appariés

La statistique des rangs signés de Wilcoxon est $T^+ = 8+4+7+1+5 = 25$ et $T^- = 11$ alors $T = 11$.

Dans la table des seuils critiques, pour un test bilatéral à $\alpha = 5\%$, on lit $T_{0.05}(8) = 4$, pour un échantillon de taille $n = 8$. Puisque on a $T > T_{0.05}(8)$, alors on conclut que la pression artérielle est la même après 6 mois d'entraînement.

Le code R de cet exemple est le suivant :

avant<- c(160,120,125,146,130,120,158,150)

après<- c(125,135,120,130,128,124,150,153)

wilcox.test(avant,après,paired=T,correct=FALSE,exact=FALSE)

On note que la $p-value = 0.327 > \alpha$, d'où on décide d'accepter l'hypothèse H_0, comme ci-dessus. Par conséquent, on conclut que cet entraînement n'a pas d'effet sur la tension systolique. On obtient le même résultat grâce la fonction de répartition empirique de la figure 2.9.

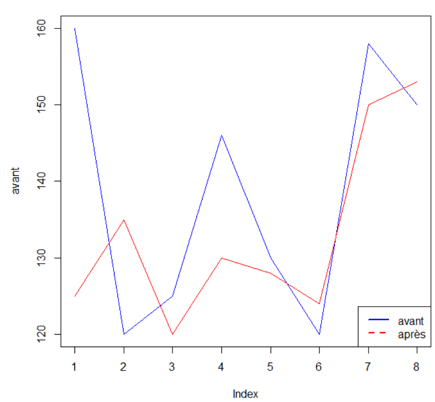

FIG. 2.9 – La fonction de répartition empirique de test des rangs signés de Wilcoxon.

Conclusion

En conclusion, les tests statistiques peuvent tourner autour de caractéristiques inconnues de la communauté sont appelés les tests paramétriques, et ils pouvent être liés à d'autres choses qui peuvent être descriptives, et dans ce cas, est appelé le tests non paramétriques.

Toute décision (rejet ou acceptation) se fait par le biais de tests d'hypothèses statistiques qui dépendent des probabilités et des distributions d'échantillonnage, qui jouent un rôle dans la prévision et la prise de décisions.

– Il existe deux méthodes pour prendre une décision dans les tests statistiques :
– calculer la statistique du test et la comparer à la valeur critique dans le tableau correspondant.
– calculer la valeur de probabilité, communément appelée $p - value$, puis la comparer au niveau de signification (risque) α du test. Une $p - value$ inférieure à α conduit au rejet de l'hypothèse H_0 nulle alors qu'une $p - value$ supérieure à α est synonime du non rejet de H_0.

Bibliographie

[1] Benelmir Imane. (2019). Cours de première année MASTER. Université Mohamed Khider, Biskra.

[2] De Micheaux, P. L., Drouilhet, R., & Liquet, B. (2011). Le logiciel R : Maitriser le langage-Effectuer des analyses statistiques. Springer Science & Business Media.

[3] Desgraupes. B. (2017).Méthodes statistiques : Introduction aux tests d'hypothèses. Université Paris Ouest Nanterre la Défense.

[4] Immediato, H. (2001). Cours de statistique. Chapitre, 3, 23.

[5] Jean-Jacques Ruch. (2013). Staistique : Tests d'hypothèses.

[6] Lejeune, M. (2004). Statistique : La théorie et ses applications.

[7] Mercadier. C. Cours de Statistique non paramétrique. MASTER SITN Lyon 1.

[8] Olivier, G. Statistique inférentielle avancées. Notes de cours.

[9] Rakotomalala, R. (2013). Comparaison de populations : Tests paramétriques. Université Lumière, Lyon 2.

[10] Rakotomalala, R. (2008). Comparaison de populations : Tests non paramétriques. Université Lumière, Lyon 2.

[11] Saporta, G. (2006). Probabilités, analyse des données et statistique. Editions Technip.

[12] Thibault, L. Les tests statistiques.

[13] Lien : http ://www.jybaudot.fr/Inferentielle/signes.html.

[14] Lien : https ://www.youtube.com/watch ?v=o6t14Tezr6Q&t=11s.

[15] Lien : https ://www.youtube.com/watch ?v=wlyIjGjlPZw&t=555s.

[16] Lien : https ://www.youtube.com/watch ?v=zyKndxrnL1Y&t=4s.

Annexe A : Quelques éléménts du logiciel *R*

Le programme R est un package de programmation spécialisé dans l'analyse et la représentation de données. Ce programme dépend du langage de programmation dévloppé par Ross Ihaka et Robert Gentleman. Le programme a été lancé en 1992.

Le programme statistique R contient une grande variété d'applications statistiques telles que :

-Les tests d'hypothèses statistiques.

-Il existe de nombreux packages spécialisés qu'on peut télécharger et utiliser dans le programme.

Pour le test de comparaison de deux distributions, on a téléchargé et installé les packages suivants : *nortest* et *twosamples.*

Test paramétrique :

Test de comparaison de deux moyennes

Cas d'échantillons indépendants

t.test(x,y,var.equal = T)

Cas d'échantillons appariés

t.test(x,y,paired = T)

#Test de comparaison de deux variances

Cas d'échantillons indépendants

var.test(x,y)

Cas d'échantillons appariés

var.test(u,v)

Test non paramétrique

Cas d'échantillons indépendants

#Test de Kolmogorov-Smirnov

Ks.test(x,y)

#Test de Cramer-Van Mises

cvm_test(x,y)

#Test de Wilcoxon-Mann-Whitney

wilcox.test(x,y)

Cas d'échantillons appariés

#Test des signes

binom.test(min(nplus,nmoins),nplus+nmoins)

#Test des rangs signés de Wilcoxon

wilcox.test(x,y,paired=T,correct=FALSE,exact=FALSE)

Annexe B : Abréviations et Notations

Les différentes abréviations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous :

Notation		Signification
H_0	:	hypothèse nulle
H_1	:	hypothèse alternative
θ	:	paramètre inconnu
Θ	:	ensemble des valeurs de θ
α	:	risque de premièr espèce
β	:	risque de deuxième espèce
W	:	région critique ou région de rejet
$\overline{W}$	:	région d'acceptation
π	:	fonction puissance
p-$value$	:	probabilité critique
S	:	statistique du test
S_{obs}	:	valeur échantionnale
$E(X)$	:	moyenne (espérance) d'une variable aléatoire X
$V(X)$	:	variance de X
S^2	:	variance empirique
$X_{(i)}$	:	$i^{ème}$ statistique d'ordre
c-à-d	:	c'est à dire
$\mathcal{N}(0,1)$	:	loi normale standard

f_{obs}	:	fractile observé
f	:	densité de probabilité
$\max(A)(ou\ \min)$	:	maximum de A(ou le minimum)
ddl	:	dégré de liberté
$t_{(n_1,n_2)}$	:	loi de student à n_1, n_2 ddl
χ_n^2	:	loi de Chi-deux à n ddl
$\mathcal{F}(n_1, n_2)$	:	loi de Ficher à (n_1, n_2) ddl
r_{uv}	:	coefficient de corrélation de Pearson
$\hat{r_{uv}}$	:	estimateur de r_{uv}
KS	:	Kolmogorov-Smirnov
D_{n_1,n_2}	:	statistique de Kolmogorov-Smirnov
$\sup(A)$	:	borne supérieure de A
$\overline{X}$	:	moyenne empirique
CM	:	Cramer-von Mises
CvM_{n_1,n_2}	:	statistique de Cramer-von Mises
$R(X_i)$	:	rang de la $i^{ème}$ observation de X
$\xrightarrow{\mathcal{L}}$	:	converge en loi
$n \to \infty$	:	n tend vers l'infini
U_{n_1,n_2}	:	statistique de Mann-Whitney
W_{n_1,n_2}	:	statistique de Wilcoxon
1	:	indicatrice
P	:	probabilité
$!$	:	factorielle
$\mathcal{B}(n,p)$	:	loi binomiale de paramètres n et p
F_n	:	fonction de répartition empirique

ملخص

في الإحصاء , اختبار المقارنة لاثنين من التوزيعات هو اختبار فرضية تستخدم لتحديد ما إذا كانت عينتان من المجتمع نفسه او من مجتمعين مختلفين.

الهدف من هذا العمل هو تبرير صلاحية هذا النوع من الاختبار المعلمي واللامعلمي ووضعه موضع التنفيذ على امثلة بسيطة.

Abstract

In statistics, the comparison test of two distributions is a hypothesis test used to determine whether two samples are from the same population or from two different populations. The goal of this work is to justify the validity of this type of tests, parametric and nonparametric, and to put it into practice on simple examples.

Résumé

En statistique, le test de comparaison de deux distributions est un test d'hypothèses utilisé pour déterminer si deux échantillons sont de la même population ou de deux populations différentes. Le but de ce travail est de justifier la validité de ce type de tests, paramétrique et non paramétrique et de le mettre en pratique sur des exemples simples.