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

GUETTAF Fatima Zohra Noudjoud

Titre :

Sur quelques tests d'indépendance et applications

Membres du Comité d'Examen :

Pr.	CHERFAOUI MOULOUD	UMKB	Président
Dr.	CHINE AMEL	UMKB	Encadreur
Dr.	SOLTANE LUIZA	UMKB	Examinateur

Juin 2024

Dédicace

Je dédie cet humble travail à :

À l'âme de mon cher père,

Je te dédie, cher papa, ce mémoire, fruit de mes efforts et de tes sacrifices. Je serai toujours la fille dont tu serais fier.

À ma chère maman,

Source de tendresse qui m'accueille avec un sourire et me dit au revoir avec une prière... Maman chérie, mon soutien et la lumière de mon chemin, qui m'a appris la persévérance et l'ambition, source d'espoir et d'inspiration.

À mes chers frères : Laid et Islem

Vous êtes mes compagnons de route. Merci d'être toujours là pour moi, pour vos bons sentiments et votre soutien constant.

À toute ma famille et à mes amis,

Merci pour votre grand amour pour moi. J'apprécie cet amour immense et j'espère vous rendre toujours fiers de moi.

À tous ceux qui m'ont appris et aidé,

J'apprécie vos efforts et vos sacrifices pour mon éducation et mon orientation. Merci d'avoir cru en mes capacités et de m'avoir aidé à atteindre cette réalisation.

Enfin,

Je dédie ce mémoire à tous ceux qui ont cru en moi et m'ont soutenue, et à tous ceux qui ont contribué à mon parcours de réussite. Je vous remercie tous du fond du cœur.

Avec tout mon amour et ma reconnaissance,

Noudjoud

.

REMERCIEMENTS

Avant tout, je tiens à exprimer ma profonde gratitude à "*Allah*" Tout-Puissant, qui m'a donné la force et la volonté de réaliser ce travail.

J'adresse mes plus sincères remerciements à mes parents, pour leur soutien indéfectible et leurs sacrifices immenses pour mon éducation et mon bien-être.

Je suis particulièrement reconnaissante à ma directrice de mémoire, Madame Dr. **Amel Chine**, pour ses précieux conseils, son suivi constant de mon travail et sa patience exemplaire tout au long de ce parcours.

Je tiens également à remercier chaleureusement les membres du jury Pr. **CHERFAOUI MOULOU** et Dr. **SOLTANE LUIZA** d'avoir accepté d'évaluer et de juger mon travail.

Ma reconnaissance s'étend à tous les enseignants du département de mathématiques qui ont contribué à ma formation.

Enfin, je n'oublierai jamais mes collègues et amis, grâce à qui ma vie universitaire a été agréable et enrichissante. Merci à tous ceux que j'ai omis de mentionner.

Avec toute ma gratitude et mon respect,

Noudjoud

.

Table des matières

Table des figures

Liste des tableaux

Introduction

L'analyse statistique constitue un pilier fondamental de la recherche scientifique, permettant aux chercheurs de quantifier, d'interpréter et de tirer des conclusions significatives à partir des données complexes entre les variables. Parmi les outils statistiques indispensables, les tests d'indépendance se distinguent comme des instruments puissants pour évaluer l'association ou la relation entre deux variables. Ce mémoire propose un voyage complet pour comprendre quelques-uns des tests d'indépendance les plus couramment utilisés, en mettant en lumière leurs applications pratiques dans une variété de contextes scientifiques.

L'histoire des tests d'indépendance remonte au début du XXe siècle, avec les travaux pionniers de Karl Pearson et de Ronald Aylmer Fisher. Pearson a développé le test du khi-deux, un outil statistique robuste pour analyser l'association entre deux variables catégorielles. Fisher, quant à lui, a introduit le test exact de Fisher, une variante du test du khi-deux adaptée aux petits échantillons. Au fil du temps, d'autres tests d'indépendance ont été développés pour répondre à des besoins spécifiques, tels que le test de Kendall pour mesurer la concordance entre deux classements et le test de Spearman pour évaluer la corrélation entre deux variables ordinales...

Les tests d'indépendance trouvent leur application dans une multitude de domaines scientifiques, allant des sciences sociales aux sciences médicales en passant par la psychologie. En sciences sociales, ces tests permettent d'examiner l'association entre des variables telles que le niveau d'éducation et le revenu et l'appartenance..etc. En

psychologie, les tests d'indépendance s'avèrent cruciaux pour évaluer la relation entre des variables comme les traits de personnalité et les performances scolaires, l'efficacité d'une thérapie et l'état de santé mentale...etc. Dans le domaine médical, ces tests jouent un rôle déterminant dans l'analyse de l'association entre des facteurs de risque et l'apparition de maladies, l'efficacité de traitements médicaux et l'évolution de pathologies, ou encore l'impact de facteurs socio-économiques sur la santé.

Dans ce mémoire, nous concentrerons sur cinq tests d'indépendance majeurs : le test de coefficient de corrélation de Pearson, le test de Spearman, le test de Kendall, le test d'indépendance du khi-deux et le test exact de Fisher. Nous explorerons les fondements théoriques de chaque test, en décrivant leurs hypothèses nulles et alternatives, leurs statistiques de test et leurs critères de décision. De plus, nous illustrerons l'application pratique de ces tests à travers des exemples concrets issus de divers domaines de recherche.

Ce mémoire est rédigé en deux chapitres :

• **Chapitre** 1 : Variables quantitatives : mesure de corrélation et tests d'indépendance. Ce chapitre présente les concepts de base de la corrélation et des tests d'indépendance pour les variables quantitatives. Il détaille le calcul des coefficients de corrélation (Pearson, Spearman et Kendall) et explique les différents tests d'indépendance, tels que le test du coefficient de corrélation de Pearson, test du coefficient de corrélation de Spearman et le test du coefficient de corrélation de Kendall. Des exemples concrets illustreront l'application de ces tests dans divers domaines.

• **Chapitre** 2 : Variables qualitatives : tests d'indépendance. Ce chapitre se concentre sur les variables qualitatives et explore les méthodes de mesure de l'association et les tests d'indépendance adaptés à ce type de données. Le coefficient d'association de Phi de Pearson (ϕ), ainsi que le test d'indépendance du khi-deux et le test exact de Fisher seront présentés en détail. Enfin, des exemples concrets illustreront l'application de ces tests dans divers domaines en utilisant le langage de programmation R.

Chapitre 1

Variables quantitatives : mesure de corrélation et tests d'indépendance

La corrélation est une technique statistique utilisée pour décrire la liaison et la dépendance entre deux variables quantitatives. Elle permet de déterminer si les variables sont liées, la direction de cette relation (positive ou négative) et la force de cette relation. Elle est mesurée par le coefficient de corrélation.

Coefficient de corrélation : Le coefficient de corrélation est une mesure numérique qui quantifie la force et la direction de la corrélation entre deux variables. Il existe plusieurs types de coefficients de corrélation, les plus courants étant : coefficient de corrélation de Pearson, coefficient de corrélation de Spearman et coefficient de corrélation de Kendall.

1.1 Coefficient de corrélation de Pearson

Le coefficient de corrélation de Pearson, développé par Karl Pearson, est un outil statistique permettant d'évaluer la force et la direction de la relation linéaire entre deux variables quantitatives.

Son calcul nécessite au préalable le calcul de la covariance entre ces deux variables.

Rappel de la définition de la covariance entre X et Y :

Définition 1.1.1 *La covariance est égale à l'espérance du produit des variables centrées.*

$$COV(X,Y) = E([X - E(X)][Y - E(Y)]).$$

On peut aussi l'écrire comme l'espérance du produit des variables, moins le produit des espérances

$$COV(X,Y) = E(XY) - E(X)E(Y).$$

Définition 1.1.2 *Le coefficient de corrélation linéaire de Pearson entre deux variables X et Y est une normalisation de leur covariance par le produit de leur écarts-types, sa formule est :*

$$R(X,Y) = \frac{COV\,(X,Y)}{\sqrt{V\,(X) * V\,(Y)}}, \tag{1.1}$$

$$= \frac{COV\,(X,Y)}{\sigma_X.\sigma_Y}, \tag{1.2}$$

où :

$COV\,(X,\ Y)$: *la covariance entre les variables X et Y.*

σ_X : *l'écart-type de la variable X.*

σ_Y : *l'écart-type de la variable Y.*

Interprétation du coefficient de corrélation de Pearson :

Le coefficient de corrélation de Pearson est compris entre -1 et 1. Son interprétation dépend de sa valeur :

– Le coefficient $R > 0$ indique une relation positive.

– Le coefficient $R < 0$ indique une relation négative.

– Le coefficient $R = 0$ indique l'absence de relation (les variables sont indépendantes et non liées).

– Le coefficient $R = +1$ décrit une corrélation positive parfaite, et $R = -1$ décrit une corrélation négative parfaite. Les points sont alignés.

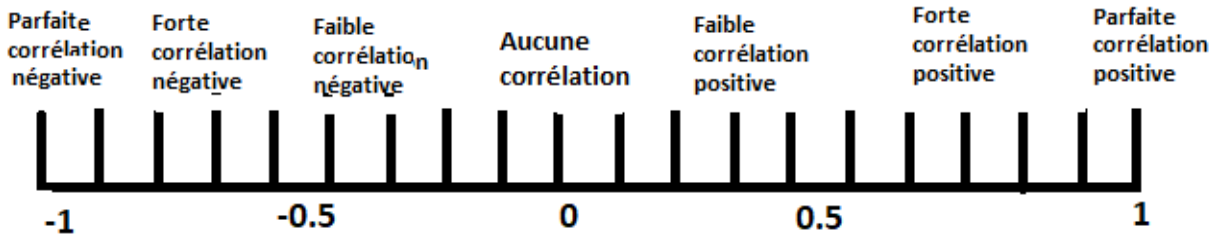


FIG. 1.1 – Interprétation du coefficient de corrélation.

Propriétés 1.1.1 • *Si X et Y sont indépendants, alors $R = 0$, mais la réciproque est généralement fausse. On équivalence si le vecteur (X, Y) est un vecteur gaussien.*

• *Le coefficient de corrélation est indépendant des unités de mesure des variables.*

• *La corrélation d'une variable avec elle même est* $R(X, X) = 1$.

1.1.1 Estimation du coefficient de corrélation de Pearson

Définition 1.1.3 *Soit un échantillon de n observations d'un couple (X, Y). Le coefficient de corrélation empirique de Pearson, noté $\hat{R}$, est défini par :*

$$\hat{R} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

La formule du coefficient de corrélation de Pearson peut également être exprimée

sous la forme suivante :

$$\hat{R} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \times \sqrt{\sum_{i=1}^{n} \bar{y}^2 - n\bar{y}^2}}, \tag{1.3}$$

où :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Propriété 1.1.1 *Le coefficient de corrélation empirique est un estimateur baisé et convergent, avec :*

$$E(\hat{R}) = R - \frac{R(1-R^2)}{2n},$$

et

$$V(\hat{R}) = \frac{(1-R^2)^2}{n}.$$

Remarque 1.1.1 *Le signe de R indique donc le sens de la liaison tandis que la valeur absolue de R indique l'intensité de la liaison.*

1.1.2 Test de significativité du coefficient de Pearson

Après le calcul du coefficient de corrélation de Pearson $\hat{R}$ estimé sur un échantillon, il faut déterminer si le coefficient de corrélation R est significativement différent de 0.

Conditions d'application du test

Avant d'effectuer le test de signification du coefficient de corrélation de Pearson, il est important de vérifier que les conditions suivantes sont remplies :

- Les deux variables sont quantitatives.
- Les données sont issues d'un échantillon aléatoire et indépendant.
- La relation entre les variables est approximativement linéaire.
- Distribution conditionnelle normale et de variance constante.

Hypothèses : Sont :

$$\begin{cases} H_0 : R = 0, \text{ « absence de liaison linéaire entre } X \text{ et } Y \text{»}. \\ H_1 : R \neq 0, \text{ « existence d'une liaison linéaire entre } X \text{ et } Y \text{ (positive ou négative)»}. \end{cases}$$

Statistique du test sous H_0 : Cette statistique est définie par :

$$T = \frac{\hat{R}\sqrt{n-2}}{\sqrt{1-\hat{R}^2}}. \tag{1.4}$$

Suit la loi de Student de $(n-2)$ degrés de liberté.

Région critique : la région critique du test au risque α (généralement $\alpha = 0,05$ ou $\alpha = 0,01$) est définie par :

$$|T| > t_{1-\frac{\alpha}{2}}(n-2),$$

où $t_{1-\frac{\alpha}{2}}(n-2)$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi de Student à $(n-2)$ degrés de liberté. Il s'agit d'un test bilatéral.

Décision :

- Si la valeur absolue de la statistique du test ($|T|$) est supérieure au quantile critique $t_{1-\frac{\alpha}{2}}(n-2)$, on rejette l'hypothèse nulle H_0 au risque α et on conclut qu'il existe une relation linéaire significative entre les variables X et Y.
- Si la valeur absolue de la statistique du test ($|T|$) est inférieure ou égale au quantile critique $t_{1-\frac{\alpha}{2}}(n-2)$, on ne rejette pas l'hypothèse nulle H_0 et on ne peut pas conclure qu'il existe une relation linéaire significative entre les variables X et Y.

1.1.3 Exemple

Une étude a été menée pour examiner la relation entre le nombre d'heures de sommeil par nuit (Variable X) et la performance académique des étudiants (Variable Y). Les données pour 5 étudiants sont les suivantes :

Étudiant	1	2	3	4	5
Heures de sommeil (heure)	7	6	8	5	9
Performance académique (note)	18	16	17	15	19

TAB. 1.1 – Table présente heures de sommeil et performance académique

Les données sont représentées graphiquement dans la figure suivante :

FIG. 1.2 – Représentation du la relation entre les heures de sommeil et les notes académiques

D'après la figure 1.2 on peut voir une liaison linéaire positive entre le nombre d'heures de sommeil par nuit et la performance académique des étudiants.

- **Etape 1 :** Premièrement on calcule le coefficient de corrélation de Pearson $\hat{R}$ entre X et Y. par la formule 1.3 et on obtient :

$$\hat{R} = 0.9.$$

- **Etape 2 :** Après, le calcul de $\hat{R}$ entre X et Y. On tester sa significativité au risque $\alpha = 5\%$ où $n = 10$, et $\hat{R} = 0.9$.
  - **Hypothèses** :

$$\begin{cases} H_0 : \text{« absence de corrélation entre les heures de sommeil et les notes académiques»}. \\ H_1 : \text{« existence de corrélation entre les heures de sommeil et les notes académiques»}. \end{cases}$$

  - **Statistique du test** : sous l'hypothèse nulle et par la formule 1.4 on obtient :

$$t = \frac{0.9\sqrt{5-2}}{\sqrt{1-0.9^2}} = 3.\,576\,2.$$

  - **Degré de liberté** : égale à :

$$ddl = 5 - 2 = 3.$$

  donc le quantile d'ordre $1-\frac{\alpha}{2}(n-2)$ de la loi de Student égale à $t_{0.95}(3) = 3.182$.
- **Etape 3 :** Décision, on compare la valeur de t avec $t_{0.975}(3)$, on observe que :

$$t = 3.\,576\,2 > t_{1-\frac{\alpha}{2}}(n-2) = t_{0.975}(3) = 3.182.$$

Alors on rejette l'hypothèse H_0, c'est à dire il existe une relation linéaire entre le nombre d'heures de sommeil et la performance académique des étudiants.

Application sous R :

```
x <- c(7,6,8,5,9)
y <- c(8,6,7,5,9)
cor(x,y,method="pearson")
[1] 0.9
cor.test(x,y,method="pearson")
Pearson's product-moment correlation
data : x and y
t = 3.5762, df = 3, p-value = 0.03739
alternative hypothesis : true correlation is not equal to 0
95 percent confidence interval :
0.08610194 0.99343752
sample estimates :
cor
0.9
```

Commentaire : On remarque que p-value $= 0.03739 < \alpha = 0.05$. Alors on rejette l'hypothèse H_0, c'est à dire il existe une relation linéaire entre le nombre d'heures de sommeil et la performance académique des étudiants.

1.2 Coefficient de corrélation ρ de Spearman

Le coefficient de corrélation de rang de Spearman, ou ρ de Spearman, est une mesure non paramétrique de corrélation de rang. Il permet d'analyser les liaisons monotones (croissantes ou décroissantes) entre les rangs des observations des variables. Il a été développé par Charles Spearman. Autrement dit, le coefficient de Spearman est identique à la corrélation de Pearson appliquée aux rangs des données.

Soit une série de n observations $\{(x_i, y_i)\}_{1\leq i\leq n}$d'un couple (X, Y). Les rangs observés de X et de Y sont définis par :

$$r_i = rang(x_i) \text{ et } s_i = rang(y_i) \text{ pour } i = 1, ..., n.$$

Définition 1.2.1 *Le coefficient de Spearman ρ est calculé comme suit :*

$$\rho = Corr(r_i, s_i) = \frac{COV\,(r_i, s_i)}{\sqrt{V\,(r_i) * V\,(s_i)}}.$$

Interprétation du coefficient de corrélation de Spearman :

Comme la valeur de ρ est comprise entre -1 et 1 nous avons les interprétations suivantes :

- Si $\rho = 1$: relation monotone croissante parfaite.
- Si $\rho = 0$: absence de relation monotone.
- Si $\rho = -1$: relation monotone décroissante parfaite.
- Valeur proche de 0 : relation faible. Valeur proche de 1 ou -1 : relation forte.

1.2.1 Estimation du coefficient de corrélation de Spearman

Le coefficient de corrélation empirique de Spearman, noté $\hat{\rho}$, est défini par :

$$\hat{\rho} := \frac{\sum_{i=1}^{n}(r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{n}(r_i - \bar{r})^2 \sum_{i=1}^{n}(s_i - \bar{s})^2}},$$

où :

$\bar{r} = \frac{1}{n}\sum_{i=1}^{n} r_i$: Moyenne des rangs observés pour la variable X.

$\bar{s} = \frac{1}{n}\sum_{i=1}^{n} s_i$: Moyenne des rangs observés pour la variable Y.

Pour simplifier les calcules, posons $D_i = r_i - s_i$, alors :

$$\hat{\rho} := 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} D_i^2,$$

tel que :

D_i : La différence entre les deux rangs de chaque observation.

n : le nombre d'observations.

Preuve. En l'absence d'ex-aequo, Il suffit de démontrer que :

$$\hat{\rho} = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} D_i^2.$$

Si l'on pose $D_i = r_i - s_i$ différence des rangs d'un même objet selon les deux classements,

on a :

$$\sum_{i=1}^{n} r_i s_i = -\frac{1}{2} \sum_{i=1}^{n} (r_i - s_i)^2 + \frac{1}{2} \sum_{i=1}^{n} r_i^2 + \frac{1}{2} \sum_{i=1}^{n} s_i^2,$$

mais

$$\sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} s_i^2 = \frac{n(n+1)(2n+1)}{6},$$

somme des carrés des nombres entiers, d'où :

$$\hat{\rho} = -\frac{6}{n(n^2-1)} \sum_{i=1}^{n} D_i^2 + \frac{\frac{n(n+1)(2n+1)}{6} - (\frac{n+1}{2})^2}{\frac{n^2-1}{12}},$$

Le deuxième terme vaut 1 après calcul et on a la formule pratique :

$$\hat{\rho} = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} D_i^2. \tag{1.5}$$

■

Remarque 1.2.1 *En cas d'exaequo (égalité) dans les données, la moyenne des rangs est calculée.*

1.2.2 Test de significativité du coefficient de Spearman

Modèle :

Soit un échantillon aléatoire et indépendant (i.i.d) de taille n, noté $(X_1, Y_1), ..., (X_n, Y_n)$, tiré d'une distribution bidimensionnelle (X, Y).

Hypothèses :

$$\begin{cases} H_0 : \hat{\rho} = 0. \\ H_1 : \hat{\rho} \neq 0. \end{cases}$$

Statistique du test :

Le coefficient de corrélation de Spearman ρ sert de statistique de test pour évaluer la dépendance entre X et Y. Il est calculé par la formule suivante :

$$T = \frac{\hat{\rho}}{\sqrt{\frac{1-\hat{\rho}^2}{n-2}}} = \frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}}.$$

La statistique du test T suit une loi de Student à $(n-2)$ degrés de liberté.

Région critique : la région critique du test au risque α (généralement $\alpha = 0,05$ ou $\alpha = 0,01$) s'écrit :

$$|T| \succ t_{1-\frac{\alpha}{2}}(n-2),$$

où $t_{1-\frac{\alpha}{2}}(n-2)$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi de Student à $(n-2)$degrés de liberté. Il s'agit d'un test bilatéral.

Décision :

– Si $|T| > t_{1-\frac{\alpha}{2}}(n-2)$, on rejette H_0 au risque d'erreur α.

– Sinon, on ne rejette pas H_0 et on ne peut pas conclure que X et Y sont dépendants.

1.2.3 Exemple

Une étude a été réalisée pour examiner la relation entre le nombre d'heures d'exercice par semaine (X) et le niveau de forme physique (en poids) (Y) pour un groupe de personnes. Les données pour 5 personnes sont présentées dans la table 1.2

Heures d'exercice par semaine (heures) X	1	3	5	2	4
Niveau de condition physique (points) Y	50	52	65	57	70
$r_i = rang(x_i)$	1	3	5	2	4
$s_i = rang(y_i)$	1	2	4	3	5
$D_i = r_i - s_i$	0	1	1	-1	-1
D_i^2	0	1	1	1	1

TAB. 1.2 – Table présente les heures d'exercice par semaine et le niveau de condition physique

– **Etape 1 :** Premièrement on calcule le coefficient de corrélation de Spearman $\hat{\rho}$ entre X et Y. par la formule 1.5 et on obtient :

$$\hat{\rho} = 1 - \frac{6 * 4}{5(25-1)} = 0.8.$$

– **Etape 2 :** Après, le calcul de $\hat{\rho}$ entre X et Y. On teste sa significativité au risque $\alpha = 5\%$ où $n = 5$, et $\hat{\rho} = 0.8$.

Hypothèses :

$$\begin{cases} H_0 : \text{le NH d'exercice par semaine et le niveau de forme physique sont indépendants.} \\ H_1 : \text{le NH d'exercice par semaine et le niveau de forme physique sont dépendants.} \end{cases}$$

Où le NH est le nombre d'heures.

– **Statistique du test** : sous l'hypothèse nulle et par la formule 1.2.2 on obtient :

$$t = \frac{0.8\sqrt{5-2}}{\sqrt{1-0.8^2}} = 2.309\,4.$$

– **Degré de liberté** : égale à :

$$ddl = 5 - 2 = 3.$$

donc le quantile d'ordre $1-\frac{\alpha}{2}(n-2)$ de la loi de Student égale à $t_{0.975}(3) = 3.182$.

– **Etape 3 : Décision** **:** on compare la valeur de t avec $t_{0.975}(3)$, on observe que :

$$t = 2.309\,4 < t_{1-\frac{\alpha}{2}}(n-2) = t_{0.975}(3) = 3.182.$$

Alors on accepte l'hypothèse H_0, c'est à dire il n'existe pas une relation entre le nombre d'heures d'exercice par semaine (X) et le niveau de forme physique (en poids) d'heures de sommeil et la performance académique des étudiants.

Application sous R :

```
x <- c(1,3,5,2,4)
y <- c(50,52,65,57,70)
cor(x,y,method="spearman")
[1] 0.8
cor.test(x,y,method="spearman")
Spearman's rank correlation rho
data : x and y
S = 4, p-value = 0.1333
alternative hypothesis : true rho is not equal to 0
sample estimates :
rho
0.8
```

Commentaire : On remarque que p-value $= 0.1333 > \alpha = 0.05$. Alors on accepte l'hypothèse H_0.

1.3 Coefficient de corrélation τ de Kendall

Le tau de Kendall est défini pour mesurer la liaison non linéaire entre deux variables. Il donne une mesure de la corrélation entre les rangs des observations. On peut exprimer le tau de Kendall de deux manières différentes, soit en fonction des observations, ou en fonction de la concordance.

Définition 1.3.1 *Soit (X_1, Y_1) un vecteur aléatoire et (X_2, Y_2)un vecteur indépendant mais de même loi que (X_1, Y_1). Le tau de Kendall noté τ est définit par :*

$$\begin{aligned}\tau_{XY} &= P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0], \\ &= 2P[(X_1 - X_2)(Y_1 - Y_2) > 0] - 1.\end{aligned}$$

1.3.1 Estimation du coefficient de corrélation de Kendall

Définition 1.3.2 *Le tau de Kendall empirique noté τ est définit par :*

$$\hat{\tau} := \frac{(n_c - n_d)}{N} = \frac{(n_c - n_d)}{\frac{1}{2}n(n-1)};\ \tau \in [-1, 1],$$

où :

n_c *: est le nombre de paires concordantes,*

n_d *: est le nombre de paires discordantes,*

N *: est le nombre total de paires.*

Interprétation :

Comme la valeur τ est comprise entre -1 et 1 nous avons les interprétations suivantes :

- Si tous les paires sont concordantes, alors $\tau = 1$.
- Si les deux classements de paires sont totalement indépendants, alors $\tau = 0$.
- Si tous les paires sont discordantes, alors $\tau = -1$.

1.3.2 Coefficient de Kendall tau b (τ_b) :

Le coefficient de Kendall τ est utilisé pour mesurer la concordance entre les rangs des observations de deux variables. Bien qu'utile dans de nombreux cas, ce coefficient peut être perturbé par la présence d'ex-aequo (égalité des rangs) dans les données.

Pour améliorer la précision du coefficient de Kendall en cas d'ex-aequo, Pablo Kendall lui-même a proposé le coefficient de Kendall corrigé (τ_b). Ce variant prend en compte le nombre d'observations répétées (ex-aequo) dans chaque catégorie des variables.

Formule du coefficient de Kendall corrigé : définie par :

$$\tau_b = \frac{(n_c - n_d)}{N^*},$$

où :

N^* : est un coefficient de correction calculé selon la formule suivante :

$$N^* = \sqrt{(\frac{n(n-1)}{2} - T)(\frac{n(n-1)}{2} - U)},$$

où

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

et

$$U = \frac{1}{2}\sum_{i=1}^{n} u_i(u_i - 1),$$

où :

t_i : c'est le nombre de sujet constituent la $i^{éme}$ classe d'éx-aequos de X.

u_i : c'est le nombre de sujet constituent la $i^{éme}$ classe d'éx-aequos de Y.

1.3.3 Test de significativité du coefficient de Kendall

Modèle :

Soit $(X_1, Y_1), ..., (X_n, Y_n)$ un échantillon aléatoire et indépendant (i.i.d) de taille n, tiré d'une distribution bidimensionnelle (X, Y).

Hypothèses :

$$\begin{cases} H_0 : \tau = 0, \text{ ce qui signifie qu'il n'ya pas de concordance entre les variables } X \text{ et } Y. \\ H_1 : \tau \neq 0, \text{ ce qui signifie qu'il existe une concordance entre les variables } X \text{ et } Y. \end{cases}$$

Statistique du test : elle est définie par :

$$U = \frac{\hat{\tau}}{\sqrt{\frac{2(2n+5)}{9n(n-1)}}} = 3\hat{\tau}\sqrt{\frac{n(n-1)}{2(2n+5)}}.$$

Sous l'hypothèse nulle H_0, la statistique U de Kendall suit une loi normale centrée et réduite.

Région critique :

La région critique du test au risque α (généralement $\alpha = 0,05$ ou $\alpha = 0,01$) est définie par :

$$|U| > u_{1-\frac{\alpha}{2}},$$

où $u_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la distribution normale standard.

Décision :

- Si $|U| > u_{1-\frac{\alpha}{2}}$, on rejette H_0 au risque α et on conclut qu'il existe une concordance significative entre les variables X et Y.
- Si $|U| < u_{1-\frac{\alpha}{2}}$, on accepte H_0 et on ne peut pas conclure qu'il existe une concordance significative entre les variables X et Y.

1.3.4 Exemple

Nous avons les valeurs indiquées dans le tableau suivant :

X	10	14	16	11	13	17	12	15	19
Y	63	75	57	81	54	60	34	64	58

TAB. 1.3 – Données ordinales distinctes

Existe-t-il une corrélation statistiquement significative entre les deux variables X et Y ?

Solution :

- **Etape 1 :** Premièrement on calcule le coefficient de corrélation de kendall $\hat{\tau}$ entre X et Y. par la formule 1.3.2 et on a :

$n_c = 16, n_d = 20, n = 9.$

$$\hat{\tau} = \frac{(16 - 20)}{\frac{1}{2} * 9(9 - 1)} = -0.111\,11.$$

- **Etape 2 :** Après, le calcul de $\hat{\tau}$ entre X et Y. On teste sa significativité au risque $\alpha = 5\%$ où $n = 9$, et $\hat{\tau} = -0.111\,11$.
  - **Hypothèses :**

$$\begin{cases} H_0 : \text{il n' y a pas une relation entre } X \text{ et } Y. \\ H_1 : \text{il y a une relation entre } X \text{ et } Y. \end{cases}$$

  - **Statistique du test** : sous l'hypothèse nulle et par la formule 1.3.3 on obtient :

$$u = 3 * (-0.111\,11) * \sqrt{\frac{9 * 8}{2(2 * 9 + 5)}} = -0.417\,02.$$

  - **Valeur critique :** la valeur critique est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la distribution normale standard. égale à $u_{0.975} = 1.96$.
- **Etape 3 Décision :** on compare la valeur de $|u|$ avec $u_{0.975}$, on observe que :

$$|u| = 0.417\,02 < u_{1-\frac{\alpha}{2}} = u_{0.975} = 1.96.$$

Alors on accepte H_0, et nous concluons qu'il n'y a pas de corrélation significative entre X et Y.

Application sous R :

```
x <- c(10,14,16,11,13,17,12,15,19)
y <- c(63,75,57,81,54,60,34,64,58)
cor(x,y,method="kendall")
[1] -0.1111111
cor.test(x,y,method="kendall")
Kendall's rank correlation tau
data : x and y
T = 16, p-value = 0.7614
alternative hypothesis : true rho is not equal to 0
sample estimates :
tau
-0.1111111
```

Commentaire : On remarque que p-value $= 0.7614 > \alpha = 0.05$. Alors on accepte l'hypothèse H_0.

1.3.5 Relation avec le ρ de Spearman et Le τ de Kendall

Le τ de Kendall et le ρ de Spearman sont des mesures utilisées pour la caractérisation d'une liaison non linéaire, mais la seule différenciation entre les deux coefficients est que le τ de Kendall peut considérer comme une probabilité et le ρ de Spearman s'interprété comme une proportion de variance expliquée.

Il y a cependant une relation entre les valeurs de ces deux coefficients. En effet,

$$-1 \leq 3\tau - 2\rho \leq 1,$$

et lorsque n est assez grand, et τ, ρ pas trop proches de 1 :

$$\rho = \frac{3}{2}\tau.$$

Finalement, si (X, Y) suit une loi normale bivariée, alors

$$\tau = \frac{2}{\pi} \arcsin \rho.$$

Chapitre 2

Variables qualitatives : tests d'indépendance

Dans ce chapitre, nous concentrerons sur l'étude des variables qualitatives en abordant trois tests statistiques importants :

Le coefficient phi (ϕ) de Pearson : Ce coefficient mesure la force et la direction de l'association entre deux variables qualitatives nominales à deux catégories.

Le test du khi-deux : Ce test permet de déterminer s'il existe une association statistique entre deux variables qualitatives nominales.

Le test exact de Fisher : Ce test est utilisé lorsque le nombre total d'observations est petit et que le test du khi-deux n'est pas approprié.

2.1 Coefficient Phi de Pearson

Développé par Karl Pearson en 1900, le coefficient phi (ϕ) de Pearson, également appelé coefficient de contingence des moindres carrés, est une approche non paramétrique consistant à fournir une mesure d'association entre deux variables qualitatives binaires X et Y.

2.1.1 Calcul du coefficient Phi de Pearson

Le coefficient Phi de Pearson est calculé à l'aide de la formule suivante :

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}},$$

où a, b, c et d représentent les effectifs observés dans le tableau de contingence 2×2, comme l'indique le tableau 2.1 :

Y vs X	1	0
1	a	b
0	c	d

TAB. 2.1 – Tableau de contingence

Le coefficient (ϕ) peut être calculé à l'aide de la formule suivante :

$$\phi = \sqrt{\frac{Q_{ind}}{N}},$$

où :

Q_{ind} : est la statistique du khi-deux.

N : est le nombre total d'observations.

Interprétation du coefficient Phi (ϕ)

- Le coefficient (ϕ) varie entre -1 et 1.
- Plus le coefficient Phi est proche de -1 ou 1, plus l'association entre les variables est forte.
- Si $\phi = 0$ indique une situation d'indépendance.

Propriétés 2.1.1 • *Le coefficient ϕ est similaire au coefficient de corrélation de Pearson dans son interprétation.*

• *Le coefficient ϕ ne suppose pas une distribution normale des données.*

• *Le coefficient ϕ est principalement utilisé avec des données qualitatives binaires.*

Exemple 2.1.1 *Supposons que nous voulions étudier la relation entre le sexe de l'étudiant (homme ou femme) et sa propension à choisir une spécialité en sciences à l'université (oui ou non).les données sont données dans le tableau.*

Sexe de l'étudiant	Oui	Non	Total
Homme	20	30	50
Femme	30	20	50
Total	50	50	100

TAB. 2.2 – Tableau croisé des fréquences : Sexe de l'étudiant et choix de spécialité en sciences

où :

Oui : choisir une spécialité en sciences.

Non : Ne choisit pas une spécialité en sciences.

Calcul du coefficient phi de Pearson : À partir de la formule 2.1.1, Nous trouvons :

$$\phi = \frac{20 * 20 - 30 * 30}{\sqrt{(20+30)(30+20)(20+30)(30+20)}} = -0.2.$$

Interprétation des résultats :

Un coefficient phi de Pearson $\phi = -0.2$ indique une faible association négative entre le sexe de l'étudiant et sa propension à choisir une spécialisation en sciences.

2.2 Test d'indépendance du khi-deux

Définition 2.2.1 *Le test d'indépendance entre deux variable, ou test du khi-deux donne la possibilité de vérifier si les données provenant d'un échantillon aléatoire permettent de conclure à l'indépendance entre deux variables qualitatives dans la population d'où a été tiré cet échantillon.*

2.2.1 Tableau de contingence et tableau de contingence théorique

Les données de ce test sont présentées dans le tableau de contingences

Définition 2.2.2 *Une table de contingence est un tableau de comptage croisant les modalités de* 2 *variables. Lorsque deux v.a sont discrètes, il est possible de représenter les résultats d'un échantillon de taille* n *par un tableau de contingence, comme l'indique la table 2.3.Pour le tableau de contingence théorique, Posant* $T_{ij} = \frac{n_{i.} \times n_{.j}}{n}$ *la fréquence attendue pour les modalités* i *et* j *s'il y avait indépendance, comme l'indique la table 2.4*

$X \backslash Y$	mod 1	...	mod j	...	Total
mod 1	n_{11}	...	n_{1j}	...	$n_{1.}$
$\vdots$		...		...	$\vdots$
mod i	n_{i1}	...	n_{ij}	...	$n_{i.}$
$\vdots$	$\cdots$	...		...	$\vdots$
Total	$n_{.1}$	...	$n_{.j}$	...	$n_{..} = n$

TAB. 2.3 – Forme générique d'une table de contingence

$X \backslash Y$	mod 1	...	mod j	...	Total
mod 1	$\frac{n_1 \times n_1}{n}$	...	$\frac{n_1 \times n_j}{n}$	...	$n_{x=1}$
$\vdots$		...		...	$\vdots$
mod i	$\frac{n_1 \times n_1}{n}$	...	$\frac{n_i \times n_j}{n}$	...	$n_{x=i}$
$\vdots$		...		...	$\vdots$
Total	$n_{y=1}$	...	$n_{y=j}$	...	n

TAB. 2.4 – Forme générique d'une table de contingence théorique

Et on a :

n_{ij} : le nombre de sujets pour lesquels la v.a X a la modalité i et la v.a Y a la modalité j.

n : la taille d'échantillon.

$n_{i.}$: la fréquence de la modalité i de la v.a X.

$n_{.j}$: la fréquence de la modalité j de la v.a Y.

2.2.2 Condition de l'utilisation de ce test

Le test d'indépendance du khi-deux peut être utilisé dans les situations suivantes :

– Deux variables catégorielles nominales.

– Deux ou plusieurs catégories (modalités) pour chaque variable.

– Indépendance des observations.

– Taille d'échantillon relativement grande et aléatoire.

– Les occurrences attendues doivent être supérieures ou égales à 5.

2.2.3 Etapes d'un test d'indépendance de khi-deux

On résume ce test dans les étapes suivantes :

1. **Etape 1 : Hypothèses** :

   L'hypothèse nulle (H_0) et l'hypothèse alternative (H_1) du test d'indépendance du khi-deux peuvent être exprimées de la manière suivante :

$$\begin{cases} H_0 : X \text{ et } Y \text{ sont indépendants.} \\ H_1 : X \text{ et } Y \text{ sont dépendants.} \end{cases}$$

2. **Etape 02 : Statistique du test** :

   La statistique du test khi-deux d'indépendance est calculée en comparant les effectifs observés (fréquences) aux effectifs théoriques (attendus) sous l'hypothèse d'indépendance. La formule de la statistique du test est la suivante :

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

   où :

   n_{ij} : est l'effectif observé pour la modalité i de X et la modalité j de Y.

   T_{ij} : est l'effectif théorique (attendu) pour la modalité i de X et la modalité j de Y sous l'hypothèse d'indépendance.

3. **Etape 03 : Valeur critique** :

   La valeur critique pour le test d'indépendance du khi-deux est obtenue à partir d'une table de distribution du khi-deux, en fonction du nombre de degrés de liberté ν et du niveau de signification α. Le nombre de degrés de liberté pour le test d'indépendance du khi-deux est calculé par la formule suivante :

$$\nu = (k-1)(m-1),$$

où :

k : est le nombre de modalités de la variable X.

m : est le nombre de modalités de la variable Y.

donc la valeur critique ce test est le fractile d'ordre $(1-\alpha)$ et de degrés de liberté $\nu = (k-1)(m-1)$. Noté $\chi^2_{\nu,(1-\alpha)}$

$$\chi^2_{\nu,(1-\alpha)} = \chi^2_{(k-1)(m-1),(1-\alpha)},$$

et la région critique est :

$$Q_{ind} > \chi^2_{\nu,(1-\alpha)}.$$

4. **Etape 04 : Décision** : Pour rejeter l'hypothèse H_0 ou l'accepter en suivant la règle suivante :

– Si la valeur calculée du χ^2 est supérieure à la valeur critique, c'est à dire si $Q_{ind} > \chi^2_\nu(1-\alpha)$ alors on rejette l'hypothèse nulle H_0 et on conclut que les deux variables sont dépendantes.

– Sinon, on ne rejette pas H_0 et on ne peut pas conclure que les variables sont dépendantes.

2.2.4 Exemple

Les données suivantes dans les tableaux 2.5 représente le suivi des Algériens de l'actualité Corona via les pages Facebook par sexe et niveau d'éducation.

```
> #create table
> data <- matrix(c(5, 6, 15, 77,5, 5, 17, 90), ncol=2, byrow=TRUE)
> colnames(data) <- c("Homme","femme")
> rownames(data) <- c("niveau moyen","niveau secondaire","niveau universitair
e","education supérieur")
> data <- as.table(data)
> #Perform Chi-Square Test of Independence
> chisq.test(data)

        Pearson's Chi-squared test

data:  data
X-squared = 12.308, df = 3, p-value = 0.006399

Warning message:
In chisq.test(data) : Chi-squared approximation may be incorrect
> #create table
> data <- matrix(c(11, 15, 77,10, 17, 90), ncol=2, byrow=TRUE)
> colnames(data) <- c("Homme","femme")
> rownames(data) <- c("niveau secondaire","niveau universitaire","education supér
ieur")
> data <- as.table(data)
> #Perform Chi-Square Test of Independence
> chisq.test(data)

        Pearson's Chi-squared test

data:  data
X-squared = 101.77, df = 2, p-value < 2.2e-16

>
```

FIG. 2.1 – Résultats du test de d'indépendance de khi-deux en R

	Homme	femme	Total
Niveau moyen	5	5	10
Niveau secondaire	6	5	11
Niveau universitaire	15	17	32
Éducation supérieur	77	90	167
Total	103	117	220

TAB. 2.5 – Table présente le suivi des Algériens de l'actualité Corona via les pages Facebook par sexe et niveau d'éducation.

Solution :

Application sous R :

Commentaire sur les résultats de l'image :

Test initial :

$\chi^2 = 12,308$, ddl $= 3$, p-value $= 0,006399$

Décision : On a p-value $= 0,006399 < 0.05$ alors on rejette l'hypothèse H_0 et qu'il existe une association significative entre le sexe de l'individu et son niveau d'éducation.

Avertissement : Approximation du khi-deux potentiellement incorrecte due à un petit nombre d'observations dans certaines cellules.

– Le test ne remplissait pas l'une des conditions du test Khi-deux.

Test après fusion du niveau d'enseignement supérieur :

$\chi^2 = 101.77$, ddl $= 2$, p-value $< 2,2e-16$

Décision : Forte association significative entre le sexe et le niveau d'éducation (p-value < 0.05).

Remarque 2.2.1 *Toutes les conditions du Khi-deux doivent être vérifiées avant sa mise en œuvre.*

2.3 Test exact de Fisher

Développé par Ronald Aylmer Fisher en 1922, le test exact de Fisher s'impose comme une méthode statistique rigoureuse pour évaluer l'indépendance entre deux variables catégorielles binaires. Il constitue une alternative robuste au test du khi-deux, particulièrement pertinent lorsque les conditions d'utilisation de ce dernier ne sont pas réunies.

Définition 2.3.1 *Le test exact de Fisher vise à évaluer s'il existe une association entre deux variables catégorielles en comparant la probabilité observée de la distribution des fréquences dans le tableau de contingence à la probabilité attendue sous l'hypothèse d'indépendance entre les variables.*

Le tableau de contingence présente les effectifs observés pour chaque combinaison des niveaux des deux variables catégorielles X et Y :

Y vs X	1	0
1	a	b
0	c	d

TAB. 2.6 – Tableau des effectifs croisés de X et Y

2.3.1 Conditions pour réaliser test exact de Fisher

Le test exact de Fisher est particulièrement adapté aux situations suivantes :

Petits effectifs : Lorsque les effectifs des échantillons sont réduits, ne permettant pas l'utilisation du test du khi-deux.

Tableaux déséquilibrés : Lorsque la distribution des fréquences dans le tableau de contingence est déséquilibrée, le test exact de Fisher offre une meilleure précision que le test du khi-deux.

Conditions du test du khi-deux non remplies : De manière générale, le test exact de Fisher peut être utilisé comme alternative au test du khi-deux dans toute situation où les conditions d'utilisation de ce dernier ne sont pas respectées.

2.3.2 Réalisation du test exact de Fisher

L'idée du test exact de Fisher consiste à comparer la distribution observée avec l'ensemble des combinaisons possibles issues d'une distribution aléatoire. Le test exact de Fisher est initialement développé pour les tableaux de taille 2×2, mais sa version étendue sera abordée ultérieurement.

Formule générale du test exact de Fisher :

On se retrouve donc dans le cadre d'une distribution hypergéométrique et la formule

générale pour la statistique du test exact de Fisher est donnée par :

$$p' = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}.$$

Calcul de la valeur p :

La valeur p du test exact de Fisher est calculée à l'aide de logiciels statistiques spécialisés, car le processus de calcul implique des équations complexes basées sur les distributions de probabilité. Ces logiciels fournissent directement la valeur p après l'entrée des données du tableau de contingence.

Interprétation de la valeur p

- Si p-value < 0.05 : rejeter l'hypothèse nulle H_0 (indépendance des variables).
- Si p-value > 0.05 : ne pas rejeter l'hypothèse nulle H_0.

2.3.3 Exemple

Nous voulons déterminer s'il existe une association statistiquement significative entre le tabagisme et le fait d'être un athlète professionnel. Fumer ne peut être que « oui » ou « non » et être un athlète professionnel ne peut être « oui » ou « non ». Les deux variables d'intérêt sont les variables qualitatives et nous avons collecté des données sur 14 personnes.

Solution :

Application sous R :

```
> dat <- data.frame(
+    "ne_fume_pas" = c(7, 0),
+    "fumez_oui" = c(2, 5),
+    row.names = c("Athlète", "Non-athlète"),
+    stringsAsFactors = FALSE)
> colnames(dat) <- c("Ne_fume_pas", "Fumer")
> dat
             Ne_fume_pas Fumer
Athlète                7     2
Non-athlète            0     5
> test <- fisher.test(dat)
> test

        Fisher's Exact Test for Count Data

data:  dat
p-value = 0.02098
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.449481      Inf
sample estimates:
odds ratio
       Inf

> test$p.value
[1] 0.02097902
>
```

FIG. 2.2 – Résultats du test exact de Fisher en R

Décision : D'après l'image que nous avons on a :

p-value= 0.02098 $<$ 0.05, rejeter l'hypothèse nulle signifie qu'il existe une relation significative entre les deux variables catégoriques (habitudes de fumer et être athlète ou non).

Conclusion

En conclusion, ce mémoire nous a conduits dans un voyage fascinant à travers le monde des tests d'indépendance, ces outils statistiques puissants qui permettent aux chercheurs de plonger dans les profondeurs des relations et des associations entre deux variables.

En explorant certains des tests d'indépendance les plus courants, tels que les tests de corrélation (Pearson, Spearman, Kendall) et les tests d'indépendance (khi-deux, exact de Fisher), nous avons pu mettre en lumière leurs applications pratiques dans divers domaines scientifiques, allant des sciences sociales et psychologiques aux sciences médicales.

Cependant, notre voyage dans le monde des tests d'indépendance ne s'arrête pas là. Il existe encore de nombreux tests avancés à explorer et d'autres applications pratiques à découvrir.

En effet, ces dernières années, nous avons assisté à l'émergence de nombreux tests de pointe qui offrent des solutions plus précises et plus puissantes pour mesurer l'association entre les variables, en particulier face à la complexité et aux diversités croissantes des données.

Parmi ces tests, on peut citer :

Le test de Cramer's V : Il est utilisé pour mesurer l'association entre deux variables qualitatives lorsque le nombre de catégories de l'une des variables est faible.

Le test de Goodman's Lambda : Il offre une alternative au test du khi-deux pour comparer des proportions multiples dans les tableaux de fréquences.

Le test de Tau-c : Il mesure l'association entre deux variables ordinales en tenant compte des égalités entre les valeurs.

Les développements ne se sont pas limités à de nouveaux tests, mais nous avons également constaté une expansion notable des applications des tests d'indépendance. Ils ont été appliqués avec succès dans de nouveaux domaines scientifiques, tels que :

La science des données : Pour comprendre les relations entre les caractéristiques des données et extraire des connaissances d'ensembles de données massifs.

L'analyse financière : Pour évaluer les risques d'investissement et prendre des décisions financières éclairées.

L'intelligence artificielle : Pour développer des systèmes intelligents capables d'apprendre à partir des données, d'identifier des modèles et de prédire le comportement.

Mais le voyage de découverte ne s'arrête pas là. Avec l'évolution des méthodes statistiques et l'accumulation de connaissances scientifiques, de nouveaux tests sophistiqués et leurs applications innovantes apparaîtront sans aucun doute à l'avenir.

Bibliographie

[1] Abann, A. (2024). Tau de Kendall. https ://youtu.be/No61a5yayks ?feature=shared. Univ-Ouargla.

[2] Benaichouhe, O., & Boussersoub, H. (2020). Le rôle de Facebook dans la sensibilisation à la santé sur Corona Covid 19. Journal de l'Autonomisation Sociale, page 288-309.

[3] Djabrane, Y. (2021). Cours SNP master 2 chapitre 3. Tests non paramétriques. Univ-Biskra.

[4] Djabrane, Y. (2023). Cours SNP master 2 chapitre 4. Mesures d'association. Univ-Biskra.

[5] Hurlin, C., & Mignon, V. (2015). Statistique et probabilités en économie-gestion. Dunod.

[6] Lejeune, M. (2010). Statistique la théorie et ses applications deuxième édition. Springer-Verlag France, Paris.

[7] Rakotomalala, R. (2012). Analyse de corrélation Étude des dépendances-Variables quantitatives. Université Lumière Lyon 2, Version 1.0.

[8] Rakotomalala, R. (2020). Etude des dépendances-Variables qualitatives. Tableau de contingence et mesures d'association. Université Lumière. Lyon 2.

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

[10] Soetewey, A. (2020). Test exact de Fisher dans R : test d'indépendance pour un petit échantillon. https ://statsandr.com/blog/fisher-s-exact-test-in-r-independence-test-for-a-small-sample.

[11] Youhou, M. (2013). Le test exact de Fisher. https ://lemakistatheux.wordpress.com.

Annexe A : Logiciel R

2.4 Qu'est-ce-que le langage R ?

• Le langage **R** est un langage de programmation et un environnement mathématique utilisés pour le traitement de données. Il permet de faire des analyses statistiques aussi bien simples que complexes comme des modèles linéaires ou non-linéaires, des tests d'hypothèse, de la modélisation de séries chronologiques, de la classification, etc. Il dispose également de nombreuses fonctions graphiques très utiles et de qualité professionnelle.

• **R** a été créé par Ross Ihaka et Robert Gentleman en 1993 à l'Université d'Auckland, Nouvelle Zélande, et est maintenant développé par la R Développement Core Team.

L'origine du nom du langage provient, d'une part, des initiales des prénoms des deux auteurs (Ross Ihaka et Robert Gentleman) et, d'autre part, d'un jeu de mots sur le nom du langage S auquel il est apparenté.

entleman) et, d'autre part, d'un jeu de mots sur le nom du langage S auquel il est apparenté.

Calcul des coefficients de corrélation et des tests d'hypothèse dans R :

R fournit des fonctions dédiées au calcul des coefficients de corrélation et à la réalisation de tests d'hypothèse, telles que :

cor() : Calcul du coefficient de corrélation (Pearson, Kendall ou Spearman) pour

quantifier la force et la direction de la relation entre deux variables numériques.

cor.test() : Effectue un test d'hypothèse pour déterminer si la corrélation observée entre deux variables est statistiquement significative.

chisq.test() : Réalise le test du chi-carré pour évaluer l'association entre deux variables catégorielles.

fisher.test() : Offre une alternative au test du chi-carré lorsque les effectifs dans les catégories du tableau sont petits.

Annexe B : Abréviations et Notations

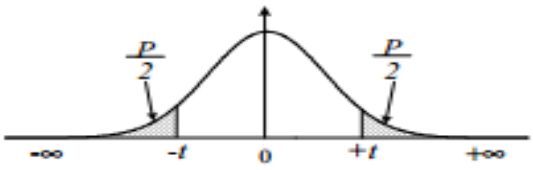
Les différentes abréviations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous :

$E(.)$	:	Espérance mathématique.
$V(.)$	:	Variance mathématique.
$Cov(X, Y)$	:	Covariance mathématique du couple (X, Y).
$\bar{x}$, $\overline{y}$	:	Moyenne empirique de X et Y respectivement.
σ_X, σ_Y	:	Écart-type de X et Y respectivement.
R	:	Le coefficient de corrélation de Pearson.
$\hat{R}$	:	Estimateur de R.
H_0	:	Hypothèse nulle.
H_1	:	Hypothèse alternative.
α	:	Risque d'erreur.
iid	:	Indépendantes et identiquement distribuées.
ddl	:	Degré de liberté.
T	:	Statistique.
ρ	:	Le coefficient de corrélation de Spearman.
$\hat{\rho}$	:	Le coefficient de corrélation empirique de Spearman.
r_i, s_i	:	Rang des observations de X et Y respectivement.

$\overline{r}$, $\overline{s}$	:	Moyenne de rang des observations de X et Y respectivement.
D_i	:	Différence entre les rangs des observations de X et Y.
τ	:	Le coefficient de corrélation de Kendall.
$\hat{\tau}$	:	Le coefficient de corrélation empirique de Kendall.
τ_b	:	Le coefficient de Kendall corrigé.
U	:	Statistique du test de Kendall.
ϕ	:	Le coefficient Phi de Pearson.
Q_{ind}	:	La statistique du test khi-deux d'indépendance.
$v.a$	:	Variable aléatoire.
T_{ij}	:	L'effectif théorique.
ν	:	Degrés de liberté pour le test d'indépendance du khi-deux.
p'	:	La statistique du test exact de Fisher.

Annexe C : Table de la loi Student

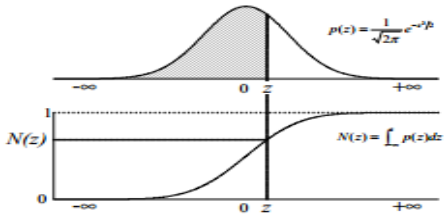
Table de la loi Student



r	P= 0,90	P= 0,80	P= 0,70	P= 0,60	P= 0,50	P= 0,40	P= 0,30	P= 0,20	P= 0,10	P= 0,05	P= 0,01	P= 0,005
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	63,657	127,321
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	9,925	14,089
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	5,841	7,453
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	4,604	5,598
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	4,032	4,773
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,707	4,317
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	3,499	4,029
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	3,355	3,833
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	3,250	3,690
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	3,169	3,581
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	3,106	3,497
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	3,055	3,428
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	3,012	3,372
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,977	3,326
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,947	3,286
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,921	3,252
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,898	3,222
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,878	3,197
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,861	3,174
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,845	3,153
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,831	3,135
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,819	3,119
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,807	3,104
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,797	3,091
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,787	3,078
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,779	3,067
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,771	3,057
28	0,127	0,256	0,389	0,530	0,683	0,855	1,056	1,313	1,701	2,048	2,763	3,047
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,756	3,038
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,750	3,030
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,704	2,971
80	0,126	0,254	0,387	0,526	0,678	0,846	1,043	1,292	1,664	1,990	2,639	2,887
120	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,658	1,980	2,617	2,860
∞	0,126	0,253	0,385	0,524	0,675	0,842	1,036	1,282	1,645	1,960	2,576	2,808

Annexe D : Table de la loi normale centrée et réduite

Table de la loi normale centrée et réduite



La table ci-dessous présente les valeurs pour z positif. Pour z négatif la valeur est $N(z) = 1-N(-z)$

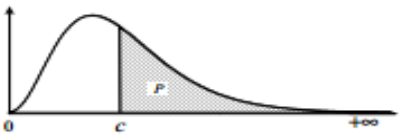
z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,500000	0,503989	0,507978	0,511966	0,515953	0,519939	0,523922	0,527903	0,531881	0,535856
0,1	0,539828	0,543795	0,547758	0,551717	0,555670	0,559618	0,563559	0,567495	0,571424	0,575345
0,2	0,579260	0,583166	0,587064	0,590954	0,594835	0,598706	0,602568	0,606420	0,610261	0,614092
0,3	0,617911	0,621720	0,625516	0,629300	0,633072	0,636831	0,640576	0,644309	0,648027	0,651732
0,4	0,655422	0,659097	0,662757	0,666402	0,670031	0,673645	0,677242	0,680822	0,684386	0,687933
0,5	0,691462	0,694974	0,698468	0,701944	0,705401	0,708840	0,712260	0,715661	0,719043	0,722405
0,6	0,725747	0,729069	0,732371	0,735653	0,738914	0,742154	0,745373	0,748571	0,751748	0,754903
0,7	0,758036	0,761148	0,764238	0,767305	0,770350	0,773373	0,776373	0,779350	0,782305	0,785236
0,8	0,788145	0,791030	0,793892	0,796731	0,799546	0,802337	0,805105	0,807850	0,810570	0,813267
0,9	0,815940	0,818589	0,821214	0,823814	0,826391	0,828944	0,831472	0,833977	0,836457	0,838913
1,0	0,841345	0,843752	0,846136	0,848495	0,850830	0,853141	0,855428	0,857690	0,859929	0,862143
1,1	0,864334	0,866500	0,868643	0,870762	0,872857	0,874928	0,876976	0,879000	0,881000	0,882977
1,2	0,884930	0,886861	0,888768	0,890651	0,892512	0,894350	0,896165	0,897958	0,899727	0,901475
1,3	0,903200	0,904902	0,906582	0,908241	0,909877	0,911492	0,913085	0,914657	0,916207	0,917736
1,4	0,919243	0,920730	0,922196	0,923641	0,925066	0,926471	0,927855	0,929219	0,930563	0,931888
1,5	0,933193	0,934478	0,935745	0,936992	0,938220	0,939429	0,940620	0,941792	0,942947	0,944083
1,6	0,945201	0,946301	0,947384	0,948449	0,949497	0,950529	0,951543	0,952540	0,953521	0,954486
1,7	0,955435	0,956367	0,957284	0,958185	0,959070	0,959941	0,960796	0,961636	0,962462	0,963273
1,8	0,964070	0,964852	0,965620	0,966375	0,967116	0,967843	0,968557	0,969258	0,969946	0,970621
1,9	0,971283	0,971933	0,972571	0,973197	0,973810	0,974412	0,975002	0,975581	0,976148	0,976705
2,0	0,977250	0,977784	0,978308	0,978822	0,979325	0,979818	0,980301	0,980774	0,981237	0,981691
2,1	0,982136	0,982571	0,982997	0,983414	0,983823	0,984222	0,984614	0,984997	0,985371	0,985738
2,2	0,986097	0,986447	0,986791	0,987126	0,987455	0,987776	0,988089	0,988396	0,988696	0,988989
2,3	0,989276	0,989556	0,989830	0,990097	0,990358	0,990613	0,990863	0,991106	0,991344	0,991576
2,4	0,991802	0,992024	0,992240	0,992451	0,992656	0,992857	0,993053	0,993244	0,993431	0,993613
2,5	0,993790	0,993963	0,994132	0,994297	0,994457	0,994614	0,994766	0,994915	0,995060	0,995201
2,6	0,995339	0,995473	0,995604	0,995731	0,995855	0,995975	0,996093	0,996207	0,996319	0,996427
2,7	0,996533	0,996636	0,996736	0,996833	0,996928	0,997020	0,997110	0,997197	0,997282	0,997365
2,8	0,997445	0,997523	0,997599	0,997673	0,997744	0,997814	0,997882	0,997948	0,998012	0,998074
2,9	0,998134	0,998193	0,998250	0,998305	0,998359	0,998411	0,998462	0,998511	0,998559	0,998605

Pour les valeurs de z supérieures à 3 :

z	3,0	3,1	3,2	3,3	3,4	3,5	3,6	3,8	4,0	4,5
N(z)	0,998650	0,999032	0,999313	0,999517	0,999663	0,999767	0,999841	0,999928	0,999968	0,999997

Annexe E : Table de la loi du khi-deux

Table de la loi du χ^2



ν	P= 0,990	P= 0,975	P= 0,950	P= 0,900	P= 0,800	P= 0,700	P= 0,500	P= 0,300	P= 0,200	P= 0,100	P= 0,050	P= 0,025	P= 0,001
1	0,000	0,001	0,004	0,016	0,064	0,148	0,455	1,074	1,642	2,706	3,835	5,879	10,828
2	0,020	0,051	0,103	0,211	0,446	0,713	1,386	2,408	3,219	4,605	5,210	10,597	13,816
3	0,115	0,216	0,352	0,584	1,005	1,424	2,366	3,665	4,642	6,251	11,345	12,838	16,266
4	0,297	0,484	0,711	1,064	1,649	2,195	3,357	4,878	5,989	7,779	13,277	14,860	18,467
5	0,554	0,831	1,145	1,610	2,343	3,000	4,351	6,064	7,289	9,236	15,086	16,750	20,515
6	0,872	1,237	1,635	2,204	3,070	3,828	5,348	7,231	8,558	10,645	16,812	18,548	22,458
7	1,239	1,690	2,167	2,833	3,822	4,671	6,346	8,383	9,803	12,017	18,475	20,278	24,322
8	1,646	2,180	2,733	3,490	4,594	5,527	7,344	9,524	11,030	13,362	20,090	21,955	26,124
9	2,088	2,700	3,325	4,168	5,380	6,393	8,343	10,656	12,242	14,684	21,666	23,589	27,877
10	2,558	3,247	3,940	4,865	6,179	7,267	9,342	11,781	13,442	15,987	23,209	25,188	29,588
11	3,053	3,816	4,575	5,578	6,989	8,148	10,341	12,899	14,631	17,275	24,725	26,757	31,264
12	3,571	4,404	5,226	6,304	7,807	9,034	11,340	14,011	15,812	18,549	26,217	28,300	32,909
13	4,107	5,009	5,892	7,042	8,634	9,926	12,340	15,119	16,985	19,812	27,688	29,819	34,528
14	4,660	5,629	6,571	7,790	9,467	10,821	13,339	16,222	18,151	21,064	29,141	31,319	36,123
15	5,229	6,262	7,261	8,547	10,307	11,721	14,339	17,322	19,311	22,307	30,578	32,801	37,697
16	5,812	6,908	7,962	9,312	11,152	12,624	15,338	18,418	20,465	23,542	32,000	34,267	39,252
17	6,408	7,564	8,672	10,085	12,002	13,531	16,338	19,511	21,615	24,769	33,409	35,718	40,790
18	7,015	8,231	9,390	10,865	12,857	14,440	17,338	20,601	22,760	25,989	34,805	37,156	42,312
19	7,633	8,907	10,117	11,651	13,716	15,352	18,338	21,689	23,900	27,204	36,191	38,582	43,820
20	8,260	9,591	10,851	12,443	14,578	16,266	19,337	22,775	25,038	28,412	37,566	39,997	45,315
21	8,897	10,283	11,591	13,240	15,445	17,182	20,337	23,858	26,171	29,615	38,932	41,401	46,797
22	9,542	10,982	12,338	14,041	16,314	18,101	21,337	24,939	27,301	30,813	40,289	42,796	48,268
23	10,196	11,689	13,091	14,848	17,187	19,021	22,337	26,018	28,429	32,007	41,638	44,181	49,728
24	10,856	12,401	13,848	15,659	18,062	19,943	23,337	27,096	29,553	33,196	42,980	45,559	51,179
25	11,524	13,120	14,611	16,473	18,940	20,867	24,337	28,172	30,675	34,382	44,314	46,928	52,620
26	12,198	13,844	15,379	17,292	19,820	21,792	25,336	29,246	31,795	35,563	45,642	48,290	54,052
27	12,879	14,573	16,151	18,114	20,703	22,719	26,336	30,319	32,912	36,741	46,963	49,645	55,476
28	13,565	15,308	16,928	18,939	21,588	23,647	27,336	31,391	34,027	37,916	48,278	50,993	56,892
29	14,256	16,047	17,708	19,768	22,475	24,577	28,336	32,461	35,139	39,087	49,588	52,336	58,301
30	14,953	16,791	18,493	20,599	23,364	25,508	29,336	33,530	36,250	40,256	50,892	53,672	59,703
40	22,164	24,433	26,509	29,051	32,345	34,872	39,335	44,165	47,269	51,805	63,691	66,766	73,402
60	53,540	57,153	60,391	64,278	69,207	72,915	79,334	86,120	90,405	96,578	112,329	116,321	124,839
120	86,923	91,573	95,705	100,624	106,806	111,419	119,334	127,616	132,806	140,233	158,950	163,648	173,617

الملخص

تعد اختبارات الاستقلال أدوات إحصائية جوهرية تستخدم لتقييم وجود ارتباط أو علاقة بين متغيرين. وتتنوع هذه الاختبارات لتشمل اختبار كاي تربيع، اختبار فيشر الدقيق، اختبار بيرسون، اختبار سبيرمان، واختبار كيندال، ولكل منها خصائص وتطبيقات مميزة حسب طبيعة المتغيرات(كمية أو نوعية). تُلعب هذه الاختبارات دورًا هامًا في مجالات علمية متعددة، مثل علم الاجتماع، علم النفس، الطب، الاقتصاد، والتسويق.

تُمكّن هذه الاختبارات الباحثين من تحليل البيانات بدقة وموثوقية، واستخلاص نتائج ذات صلة، ممّا يُساهم في تقدم المعرفة العلمية.

الكلمات المفتاحية: الارتباط ، معامل الارتباط، اختبار الاستقلالية،المتغير الكمي،المتغير النوعي،الفرضيات، إحصائية الاختبار، القرار،درجة الحرية.

Résumé

Les tests d'indépendance sont des outils statistiques essentiels utilisés pour évaluer l'existence d'une association ou d'une relation entre deux variables. Ces tests comprennent le test du khi-deux, le test exact de Fisher, le test de Pearson, le test de Spearman et le test de Kendall, chacun ayant ses propres caractéristiques et applications en fonction de la nature des variables (quantitative ou qualitative). Ils jouent un rôle crucial dans divers domaines scientifiques tels que la sociologie, la psychologie, la médecine, l'économie et le marketing.

Ces tests permettent aux chercheurs d'analyser les données avec précision et fiabilité, d'extraire des conclusions pertinentes et de contribuer ainsi à l'avancement des connaissances scientifiques.

Mots clés: **corrélation, coefficient de corrélation, test d'indépendance, variable quantitative, variable qualitative, statistiques de test, hypothèses, décision, degré de liberté.**

Abstract

Independence tests are essential statistical tools used to assess the existence of an association or relationship between two variables. These tests include the chi-square test, Fisher's exact test, Pearson's test, Spearman's test, and Kendall's tau test, each with its own characteristics and applications depending on the nature of the variables(quantitative or qualitative). They play a crucial role in various scientific fields such as sociology, psychology, medicine, economics, and marketing.

These tests allow researchers to analyze the data accurately and reliably, to extract relevant conclusions and thus contribute to the advancement of scientific knowledge.

Key words : **correlation, correlation coefficient, independence test, quantitative variable, qualitative variable, test statistics, hypotheses, decision, degree of freedom.**