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

ATTAF Maroua

Titre :

Quelques Mesures d'Association et Applications

Membres du Comité d'Examen :

Pr.	BENATIA Fateh	UMKB	Président
Pr.	YAHIA Djabrane	UMKB	Encadreur
Dr.	ROUBI Afaf	UMKB	Examinatrice

Septembre 2020

Dédicace

Je dédie ce humble travail.

À mon chère père Messaoud et ma chère mère Nabiha

Pour leur amour inestimable, leur confiance, leur soutien, leurs sacrifices.

À mes chères sœurs Nadjoua et Hadjer.

À mon chère frère Abd Elhamid.

À mes cousines : Islam, Saida, Amel, Zoulikha, Djena, Khadija, et Chaima.

À mes tantes Firouz et Nadia.

Une dédicace spéciale

À mon oncle Pr. Attaf Abdallah et sa femme Mme Saidi Hanane.

À tout ma famille ATTAF.

À tous mes amies sans exception :

Achouak,Halima, Ahlem, Sihem.

À tous mes collègues qui ont été à cotés pendant ces années d'études.

À tous les personnes qui m'ont aidé pour l'élaboration de ce travail

Je te dis Merci et Merci tous.

À tous ceux qui connaissent Maroua Attaf.

REMERCIEMENTS

Je tiens tout d'abord à remercier "Allah" le tout puissant et miséricordieux

Qui ma donné la force et la patience d'accomplir ce modeste travail.

J'adresse mes remerciement les plus sincères à mon encadreur Pr.Yahia Djabrane

Qui m'a encadré tout au long de cette année de la confiance qu'il m'a accordée en me confiant ce travail

Ainsi que ses conseils scientifiques et son encouragement.

Je tiens à remercier avec ma plus grande gratitude les membres du jury :

Mr.Benatia F. et Mme.Roubi A.

Pour l'ntérêt qu'ils ont porté à ce travail en acceptant de l'examiner.

J'adresse mes remerciement les plus sincères aussi

À mon oncle Pr. Attaf Abdallah et à sa femme Mme. Saidi H.

Pour leurs grand soutiens, leurs conseils, leurs encouragements, et leurs efforts avec moi.

Je n'oublie pas de remercier l'ensemble des enseignants du département de Mathématiques, faculté

des Sciences Exactes et des Sciences de la Nature et de la vie.

Enfin, je remercie tous ceux qui m'ont aidé de prés ou de loin à la réalisation

De ce mémoire.

Table des matières

Table des figures

Liste des tableaux

Introduction

L'objectif de la statistique est de collecter, traiter et analyser des données à l'aide d'un ensemble de méthodes et techniques, y compris des méthodes statistiques qui concerne la quantification des phénomènes aléatoires dans divers domaines. Par exemple, en épidémiologie, on cherche à savoir si l'exposition à un facteur de risque entraîne l'apparition d'une maladie, en sociologie, on cherche à savoir s'il y a un lien entre la profession du père et la filière choisie par un étudiant à l'université...etc. En particulier on essaie de savoir l'existence d'une liaison entre deux ou plusieurs variables et c'est le sujet du présent mémoire. Donc l'idée est basée sur la notion de corrélation, d'association et de dépendance. Les paramètres de mesure variant selon la nature des variables étudiées (quantitative, qualitatives).

En statistique et en probabilités, la corrélation entre plusieurs variables aléatoires ou statistiques est une notion de liaison qui contredit leur indépendance. Cette corrélation est très souvent réduite à la corrélation linéaire entre variables quantitatives, c'est-à-dire l'ajustement d'une variable par rapport à l'autre par une relation affine obtenue par régression linéaire. Pour cela, on calcule un coefficient de corrélation linéaire. La valeur absolue de ce coefficient, toujours comprise entre 0 et 1, ne mesure pas l'intensité de la liaison mais la prépondérance de la relation affine sur les variations internes des variables. Un coefficient nul n'implique pas indépendance, car d'autres types de corrélation sont possibles.

Dans ce mémoire, nous avons essayé d'éclairer d'une façon générale le concept des mesures d'association, leur nécessité et leurs utilités, et proposer quelques exemples d'application. Les tests de corrélation sont utilisés pour évaluer une association (dépendance) entre deux variables. Le calcul du coefficient de corrélation peut être effectué en utilisant différentes méthodes : phi de Pearson, tau de Kendall, rho de Spearman... Ces méthodes et mesures d'association sont bien décrites dans ce mémoire.

Ce travail est composé en deux chapitre comme suit : Le premier chapitre est consacré à une étude simultanée de deux variables quantitatives, en introduisant le graphique appelé nuage de points qui donné une aidé générale sur la dépendance de deux variables étudiées, et un ensemble des mesures de corrélation simple à calculer et peuvent être facilement interprétés, c'est ce qu'on va le détailler.

Dans le deuxième chapitre, le cas d'étude simultané de deux variables qualitatives, on adopte une mesure différente de tous celles qui précédent qui est l'écart à l'indépendance cette dernière conçue pour étudier et analyser des tableaux appelés couramment tableau de contingence ou tableau croisés c'est ce qui nous allons tenter de le préciser.

Les études de simulation sont réalisées en utilisant le logiciel statistique R.

Chapitre 1

Corrélation et liaison entre variables quantitatives

En statistique, le terme population s'applique à tout objet statistique étudié, c'est donc l'ensemble sur lequel porte notre étude. On cherche souvent à décrire une population donnée. Nous nous intéressons à la caractéristique des unités qui peuvent prendre différentes valeurs.

Ainsi, toute application X définie par :

$$\begin{array}{c} \Omega \xrightarrow{X} D \\ \omega \longmapsto x(\omega) \end{array}$$

Est appelle variable, caractère, indicateur ou paramètre. L'ensemble D est celle des valeurs du caractère X (c'est ce qui est mesuré ou observé sur les individus de la population).

– Si $D \subseteq \mathbb{N}$, X est une variable statistique entière ou discrète.

– Si $D \subseteq \mathbb{R}$, X est une variable réelle ou continue.

– Si D n'est pas mesurable alors X est une variable qualitative, dont les réponses

possibles associées sont appelées "modalités".

– Dans le cas où X est à une variable entière ou continue, X est dite variable quantitative.

1.1 Liaison entre deux variables quantitatives

Nous savons que, deux variables X et Y sont indépendantes (entre elles) ou bien elles ne sont pas. Lorsqu'elles sont liées l'une à l'autre on s'intéresse souvent à l'étude de cette liaison, dans ce cadre il y a deux procédures pour facilité l'étude. Le premier par l'analyse graphique à travers le nuage de points pour mettre une visualisation de type de la liaison qui peut exister entre deux variables. Le second par des indicateurs numériques pour essayer de quantifier ce que l'on voit.

1.1.1 L'analyse graphique

Les distributions statistiques à deux dimensions peuvent être représentées graphiquement sous forme de nuage de points dans un plan. Le principe est que chaque couple (X, Y) est représenté par un point (en abscisse la variable X, en ordonnée la variable Y), l'ensemble de points forme un nuage de points dont la forme permette de caractériser le type de liaison.

– Liaison linéaire positive : les deux variables X et Y varient dans le même sens.

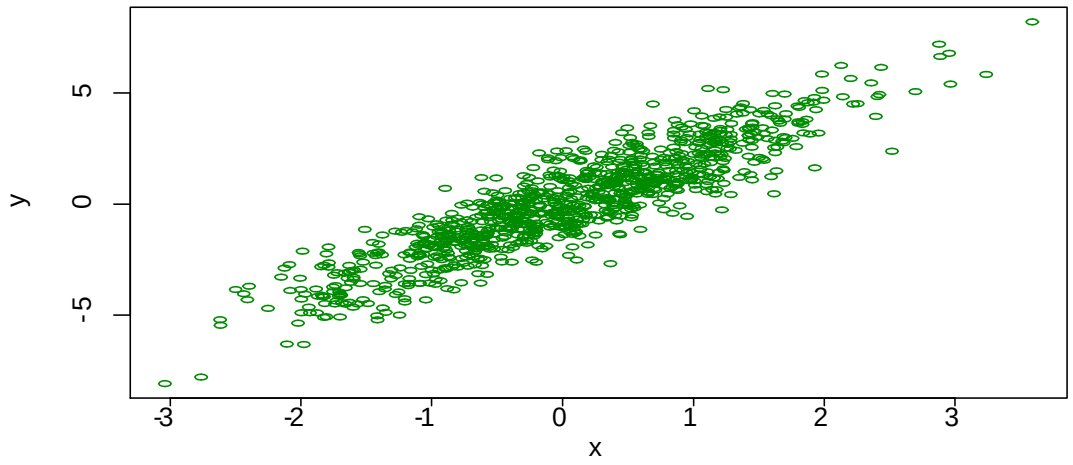

FIG. 1.1 – Liaison linéaire positive

– Liaison linéaire négative : les deux variables X et Y varient en sens inverse.

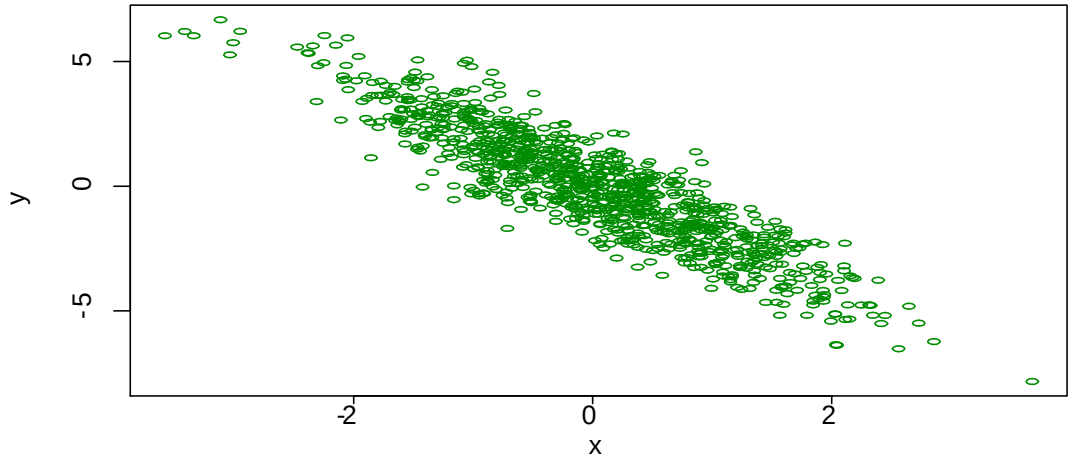

FIG. 1.2 – Liaison linéaire négative

– Liaison non-linéaire monotone positive : les deux variables X et Y varient dans le même sens mais la pente est différente selon le niveau de X.

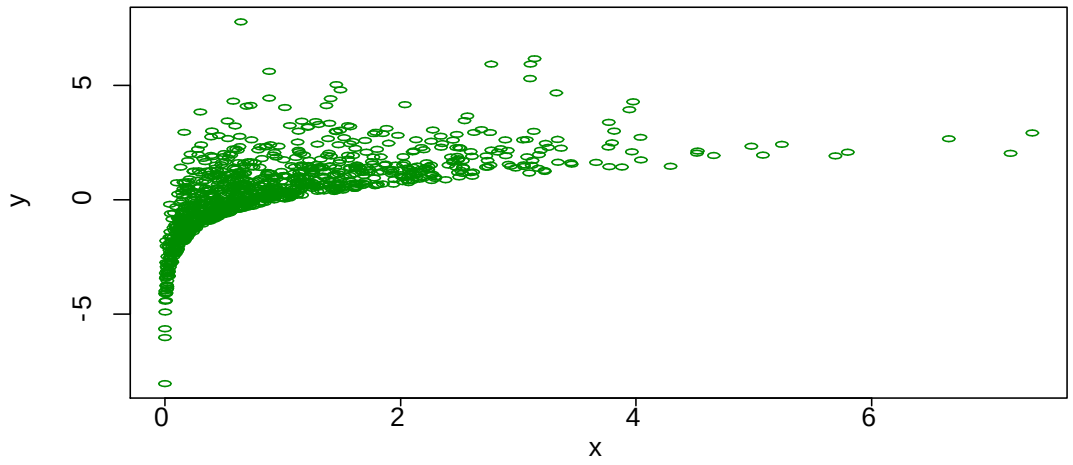

FIG. 1.3 – Liaison non-linéaire monotone positive

– Liaison non-linéaire non-monotone : il y a une relation fonctionnelle entre X et Y, mais la liaison n'est pas monotone.

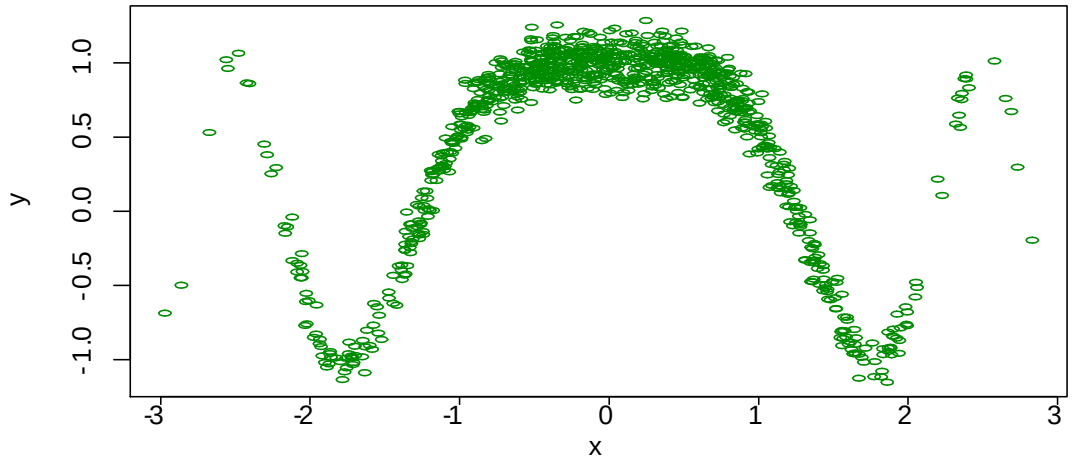

FIG. 1.4 – Liaison non-linéaire non-monotone

– Absence de liaison : La valeur de X ne donne indication sur la valeur de Y.

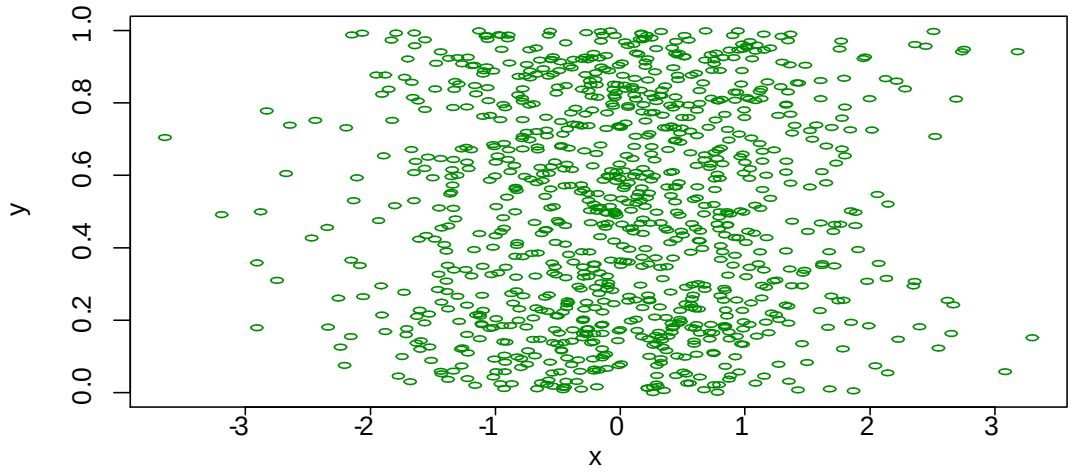

FIG. 1.5 – Absence de liaison

Exemple 1.1.1 *On souhaite étudier l'existence d'une liaison entre la taille et le poids de 17 individus, Les données sont représentées graphiquement dans la figure suivante :*

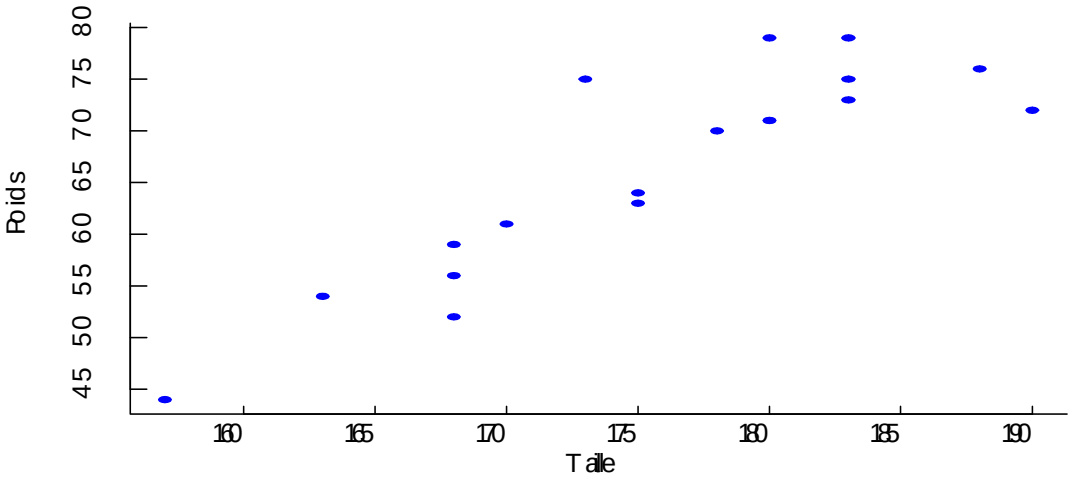

FIG. 1.6 – Représentation simultanée du taille et poids des individus

La figure (1.6) montre que la taille et le poids ont tendance à varier dans le même sens, Il s'agit donc d'une liaison linéaire positive.

Sous R :

Taille=c(188,173,178,183,168,168,163,180,183,175,175,183,157,190,170,180,165)

Poids=c(76,75,70,75,56,52,54,71,79,64,63,73,44,72,61,79,59)

plot(Taille,Poids)

1.1.2 Indicateurs de liaison linéaire

Pour avoir un accès pratique à la quantification de sens et d'intensité de liaison peuvent exister entre les variables. Les indicateurs le plus utilisée sont la covariance et le coefficient de corrélation de Pearson.

La covariance

Définition 1.1.1 *Soient X et Y deux variables, la covariance de X et Y est le nombre*

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

$$= E(XY) - E(X)E(Y)$$

Définition 1.1.2 *Soit $\{(x_1, y_1)(x_2, y_2), \ldots, (x_n, y_n)\}$ un échantillon de n observations d'un couple (X, Y) .la covariance empirique est définie par :*

$$\hat{S}_{XY} := \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n}$$

où $\bar{x}$ est la moyenne empirique de X et $\bar{y}$ celle de Y.

Propriétés 1.1.1 *La covariance possède les propriétés suivantes :*

1. *Symétrie :* $COV(X,Y) = COV(Y,X)$.

2. *Distributivité :* $COV(X,Y+Z) = COV(X,Y) + COV(X,Z)$.

3. *Covariance avec une constante :* $COV(X,a) = 0$.

4. *Covariance avec une transformation affine :* $COV(X,a+bY) = bCOV(X,Y)$.

5. *Variance de la somme de deux variables aléatoires.* $V(X+Y) = V(X)+V(Y)+2COV(XY)$.

6. *Lorsque* X *et* Y *sont indépendantes,* $COV(X,Y) = 0$ *, mais la réciproque est généralement fausse, on peut donner le contre exemple suivantes : Soit* $X \rightsquigarrow N(0,1)$ *et* $Y = X^2$ *alors*

$$COV(X,Y) = E(XX^2) - E(X)E(X^2) = E(X^3) = 0$$

Donc $COV(X,Y) = 0$ *au contraire* X *et* Y *ne sont pas indépendantes.*

Interprétation

Nous allons interpréter la valeur de covariance de la façon suivante :

– $COV(X,Y) > 0$: la liaison est positive.

– $COV(X,Y) = 0$: absence de liaison monotone.

– $COV(X,Y) < 0$: la liaison est négative.

Exemple 1.1.2 *Reprenons notre exemple, la valeur de covariance entre la taille et le poids est égale à* $+83.67$*, Ce résultat confirme quantitativement ce que le nuage de points de la figure (1.6) nous suggère visuellement, à savoir une liaison positive entre la taille et le poids d'une personne.*

Remarque 1.1.1 *Lorsque nous essayons d'interpréter la valeur d'une covariance le problème qui se pose est l'affectation de ce valeur par un changement d'unités, pour remédier ce problème on doit proposer une mesure normalisée qui s'appelle le coefficient de corrélation linéaire de Pearson.*

Le coefficient de corrélation linéaire de Pearson

Définition 1.1.3 *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 :*

$$r(X,Y) = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}; \quad r\epsilon\,[-1,1]$$

où $\sigma_Z = \sqrt{V(Z)}$ *pour* $Z = X$ *et* Y.

Propriétés 1.1.2 *La corrélation possède les propriétés suivantes :*

1. *Le coefficient est indépendant des unités de mesure.*
2. *La corrélation d'une variable avec elle même est* $r(X,X) = 1$.
3. *Le coefficient n'est applicable que dans le cas gaussien ou linéaire.*
4. X *et* Y *sont indépendants, alors* $r = 0$. *La réciproque est fausse, sauf Lorsque le couple de variables* (X,Y) *suit une loi normale bi-variée, nous avons l'équivalence.*
5. *Son carré que l'on appelle coefficient de détermination permet de mesurer le pourcentage de* Y *qui soit expliquée par* X *(en régression par exemple).*

Interprétation

Selon la valeur du coefficient r, nous avons les Interprétations suivantes :

– Si r est proche de 1, il existe une forte liaison linéaire positive entre X et Y.

– Si r est proche de 0, il n'y a pas de liaison linéaire entre X et Y.

– Si r est proche de -1, il existe une forte liaison linéaire négative entre X et Y.

Le signe de r indique donc le sens de la liaison tandis que la valeur absolue de r indique l'intensité de la liaison.

Exemple 1.1.3 *On reprend l'exemple de la liaison entre la taille et le poids. La valeur de corrélation est égale à* $+0.88$, *Ce qui indique qu'il y a une liaison positive et très forte reliant la taille et le poids d'une personne.*

Estimation du coefficient de corrélation

Définition 1.1.4 *Soit un échantillon de n observations d'un couple (X, Y), la corrélation empirique est définie par :*

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

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

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

Sa valeur ajusté est donnée par :

$$\hat{r}_{aj} = \sqrt{1 - \frac{n-1}{n-2}(1-\hat{r}^2)}.$$

Test paramétrique sur le coefficient de corrélation linéaire

Après le calcul du coefficient de corrélation $\hat{r}$ estimé sur un échantillon de taille n, il faut déterminer si le coefficient de corrélation r est significativement différent de 0 le test est de l'hypothèse $H_0 : r = 0$, ‹‹absence de liaison linéaire entre X et Y››, contre $H_1 : r = 0$, ‹‹existence d'une liaison entre X et Y››.

La statistique de test sous H_0 est :

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

Qui suit une loi de Student à $(n-2)$ degrés de liberté. La région critique du test au risque α est défini par :

$$R.C = \left\{ |T| \ > T_{1-\frac{\alpha}{2}}(n-2) \right\}$$

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.

Exemple 1.1.4 *Nous revenons à l'exemple précédent du même échantillon, mais nous voulons savoir maintenant s'il existe une liaison entre le poids et l'âge des individus, les données sont dans le tableau (1.1) :*

poids	76	75	70	75	56	52	54	71	79	64	63	73	44	72	61	79	59
Âge	41	42	32	39	30	33	26	30	53	32	47	34	23	36	31	29	28

TAB. 1.1 – Tableau représente l'âge et le poids des individus

On pose les hypothèses de test :

H_0 : absence de liaison entre l'âge et le poids.

H_1 : existence d'une liaison entre l'âge et le poids.

Pour un risque d'erreur $\alpha = 5\%$, la statistique de test sous l'hypothèse nulle $T = 2.84$, le quantile d'ordre $1 - \frac{\alpha}{2}$ correspondant à 15 degré de liberté est $T(0.975, 15) = 2.13$. Puisque $2.84 > 2.13$, on peut rejeter donc H_0 et affirme qu'il existe une liaison entre le poids et l'âge d'une personne.

1.2 Autres mesures de corrélation

D'une part, on peut voir le coefficient de corrélation de Pearson de différentes manières pour obtenir des informations supplémentaire comme le phi de Pearson. D'autre part, il ne caractérise qu'une liaison linéaire pour l'étude de liaison non linéaire, plusieurs mesures on été proposées, certaines sont basées sur les rangs des observations comme le rho de Spearman, d'autres sont basées sur la notion de paires concordances et discordances comme le tau de Kendall.

1.2.1 Phi de Pearson

Le coefficient de corrélation phi de Pearson permet de mesurer l'intensité de la liaison entre deux variables binaires (codées 0 ou 1). Le calcul est réalisé à travers le coefficient de Pearson sur les variables binaires ou sur un tableau composé de deux lignes et deux colonnes comme suit :

$Yvs.X$	1	0
1	a	b
0	c	d

TAB. 1.2 – Tableau générique 2 * 2

Sa formule est :

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

Interprétation

Le coefficient phi est similaire au coefficient de corrélation de Pearson dans son interprétation. Alors il varie entre -1 et 1. Plus il est proche de ces bornes plus la liaison est forte entre les deux variables. Un coefficient de 0 indique une situation d'indépendance.

Remarque 1.2.1 *On utilise souvent le codage* 1 *de la modalité qui nous intéresse et* 0 *à la second modalité. De plus, ce codage détermine le signe de* ϕ*, mais il n'a pas d'incidence sur la valeur absolue du coefficient.*

Exemple 1.2.1 *On veut étudier la liaison entre les caractères : ‹‹être fumeur›› (plus de* 20*cigarettes par jour, pendant* 10 *ans) et ‹‹avoir un cancer de la gorge››, sur une population de* 1000*personnes, les résultats sont dans le tableau (1.3) :*

Observé	cancer	non cancer
fumeur	342	258
non fumeur	158	242

TAB. 1.3 – Tableau représente «être fumeur» et «avoir un cancer de la gorge»

Après le calcul nous obtenons le résultat $\phi = 0.17$, Il s'agit donc, d'une faible dépendance positive entre ‹‹être fumeur›› et ‹‹avoir un cancer de la gorge››.

1.2.2 Concept de concordance

Soit $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ un échantillon de n observations d'un couple (X, Y). Il existe $\mathbb{C}_n^2 = \frac{n!}{2(n-1)!}$ pairs de distributions distinctes de couples (x_i, y_i) et (x_j, y_j) qui sont dites concordantes où discordantes selon :

1. concordantes : $(x_i - x_j)(y_i - y_j) > 0$ i.e. $(x_i < x_j$ et $y_i < y_j)$ où $(x_i > x_j$ et $y_i > y_j)$.

2. discordantes : $(x_i - x_j)(y_i - y_j) < 0$ i.e. $(x_i < x_j$ et $y_i > y_j)$ où $(x_i > x_j$ et $y_i < y_j)$.

Définition 1.2.1 *La fonction de concordance est la différence entre la probabilité de concordance et celle de discordance entre deux couples* (X_1, Y_1) *et* (X_2, Y_2) *. Elle est donnée par :*

$$Q := P\left[(X_1 - X_2)(Y_1 - Y_2) > 0\right] - P\left[(X_1 - X_2)(Y_1 - Y_2) < 0\right].$$

Comme les variables aléatoires sont continues, alors

$$P\left[(X_1 - X_2)(Y_1 - Y_2) < 0\right] = 1 - P\left[(X_1 - X_2)(Y_1 - Y_2) > 0\right].$$

Donc

$$Q = 2P\left[(X_1 - X_2)(Y_1 - Y_2) > 0\right] - 1,$$

où

$$P\left[(X_1 - X_2)(Y_1 - Y_2) > 0\right] = P\left[X_1 < X_2, Y_1 < Y_2\right] + P\left[X_1 > X_2, Y_1 > Y_2\right].$$

1.2.3 Rho de Spearman

Le coefficient de corrélation rho de Spearman permet d'analyser les liaisons non linéaires entre les rangs des observations des variables. La valeur de ce coefficient notée ρ est équivalente au coefficient de corrélation de Pearson. Il a été développé par Spearman. Autrement dit, le coefficient rho de Spearman est la corrélation de Pearson appliquée sur les rangs.

Définition 1.2.2 *Soit une série de n observations* $\{(x_i, y_i)\}_{1 \leq i \leq n}$ *d'un couple* (X, Y)*, on note :*

– *Les rangs de X et de Y observés :*

$$r_i = rang(x_i) \qquad\qquad s_i = rang(y_i)$$

– Les moyennes des rangs observés $\bar{r}$ et $\bar{s}$:

$$\bar{r} = \frac{1}{n}\sum_{i=1}^{n} r_i \qquad\qquad \bar{s} = \frac{1}{n}\sum_{i=1}^{n} s_i$$

– Les variances des rangs observés S_r^2 et S_s^2 :

$$S_r^2 = \frac{1}{n}\sum_{i=1}^{n} (r_i - \bar{r}) \qquad\qquad S_s^2 = \frac{1}{n}\sum_{i=1}^{n} (s_i - \bar{s})$$

– La covariance des rangs observés S_{rs} :

$$S_{rs} = \frac{1}{n}\sum_{i=1}^{n} (r_i - \bar{r})(s_i - \bar{s})$$

Alors le coefficient rho de Spearman devient :

$$\rho := \frac{\sum_{i=1}^{n} (r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{n} (r_i - \bar{r})^2 \sum_{i=1}^{n} (s_i - \bar{s})^2}} = \frac{S_{rs}}{S_r S_s}.$$

Interprétation

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

1. Si tous les classements des paires sont concordantes, $\rho = 1$.

2. Si tous les classements des paires sont totalement indépendants, $\rho = 0$.

3. Si tous les classements des paires sont discordantes, $\rho = -1$.

Remarque 1.2.2 *Lorsqu'il n'y a pas d'ex-æquo (i.e. observations du mêmes rang) dans les données, alors le coefficient rho de Spearman est donné par la formule suivante :*

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

où d_i désigne l'écart entre les rangs : $d_i = r_i - s_i$.

1.2.4 Tau 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 1.2.3 *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 est définit par :*

$$\rho_{XY} := P\left[(X_1 - X_2)(Y_1 - Y_2) > 0\right] - P\left[(X_1 - X_2)(Y_1 - Y_2) < 0\right].$$

Définition 1.2.4 *Soit une série de n observations $\{(x_i, y_i)\}_{1 \leq i \leq n}$ d'un couple (X, Y). Le tau de Kendall est définit par :*

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

où

$$\begin{cases} n_c : \textit{nombre de paires concordantes,} \\ n_d : \textit{nombre de paires discordantes,} \\ n : \textit{nombre total de paires.} \end{cases}$$

Interprétation

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

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

Remarque 1.2.3 *On présence d'ex-æquo, Il faut utiliser l'indice tau-b(noté τ_b). Il est considère comme une extension du tau de Kendall pour des données ordinales non distinctes, sa formule est :*

$$\tau_b := \frac{n_c - n_d}{\left(\sqrt{\left[n\left(n-1\right) - E_X\right]\left[n\left(n-1\right) - E_Y\right]}\right)^{1/2}},$$

où $E_Z = \sum_{g=1}^{G_Z} e_g(e_g - 1)$ pour $Z = X$ et Y ,avec :

$$\begin{cases} e_g : \text{le nombre d'occurrence pour chaque valeur de } Z, \text{ pour } Z = X \text{ et } Y \\ G_Z : \text{le nombre des valeurs non distinctes de } Z, \text{ pour } Z = X \text{ et } Y \end{cases}$$

Remarque 1.2.4 *Le tau de Kendall et le rho 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 tau de Kendall peut considérer comme une probabilité.*

Exemple 1.2.2 *Dans cet exemple, on fait passer des tests d'intelligence à 8 couples de vrais jumeaux. Le but est devoir s'il y a une liaison entre les tests de celui qui est né en premier et ceux de celui qui est né en second. Les données se trouvent dans le tableau (1.4), les scores plus élevés correspondant à de meilleurs résultats aux tests.*

couple de jumeaux	1	2	3	4	5	6	7	8
Né le premier	90	75	99	60	72	83	83	90
Né en second	88	79	98	66	64	83	88	98

TAB. 1.4 – Tableau représente les tests d'intelligence de vrais jumeaux

Les données sont représentées graphiquement dans la figure suivante :

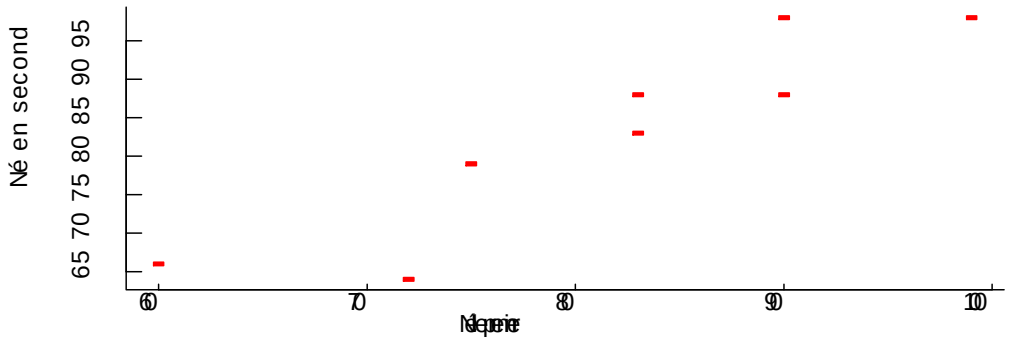

FIG. 1.7 – Représentation simultanée des tests d'intelligence de vrais jumeaux

À partir la figure (1.7) on peut voir une liaison entre les résultats du test d'intelligence d'un couple de jumeaux. De plus, la valeur de rho de Spearman $\rho = 0.93$ et tau de Kendall $\tau = 0.85$ affirme numériquement ce que l'on voit.

Chapitre 2

Mesures d'association entre variables qualitatives

2.1 Liaison entre deux variables qualitatives

On veut à présent étudier l'association entre deux variables qualitatives et pour quantifier celle-ci on exploite le tableau de contingence.

2.1.1 L'analyse d'un tableau de contingence

Un tableau de contingence est une méthode de représentation de données issues d'un comptage. De plus, il est considéré comme le moyen qui permet de présenter simultanément et de manière croisée deux variables observées sur la même population. Autrement dit, un tableau de contingence est une manière de résumer la relation entre deux variables qualitatives X et Y possédantes respectivement p et q modalités d'écrivant un ensemble de n individus. Un tableau de contingence a la structure suivante :

Y / X	y_1	y_2	$\cdots$	y_j	$\cdots$	y_q	$n_{i.}$
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1q}	$n_{1.}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2q}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{is}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_p	n_{p1}	n_{p2}	$\cdots$	n_{pj}	$\cdots$	n_{pq}	$n_{p.}$
$n_{.j}$	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.q}$	n

TAB. 2.1 – Tableau de contingence de deux variables qualitatives X (p modalités) et Y (q modalités)

On a n_{ij} est le nombre d'individus ayant comme attribut la $i^{éme}$ modalité de X et la $j^{éme}$ de Y i.e. (l'effectif de la case correspondant à la $i^{éme}$ ligne et la $j^{éme}$ colonne du tableau).

Les $n_{i.}$ et les $n_{.j}$ s'appellent respectivement marges en linges (le nombre d'individus ayant comme attribut la $i^{éme}$ modalité de X i.e. (la somme de la $i^{éme}$ ligne) et marges en colonnes (le nombre d'individus ayant comme attribut la $j^{éme}$ modalité de Y i.e. (la somme de la $j^{éme}$ colonne) elles sont calculées comme suit :

$$\begin{cases} n_{i.} = \sum_{j=1}^{q} n_{ij} \\ n_{.j} = \sum_{i=1}^{p} n_{ij} \end{cases}$$

n est le nombre total d'individu étudié i.e. (somme générale du tableau)

$$n = \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij} = \sum_{i=1}^{p} n_{i.} = \sum_{j=1}^{q} n_{.j}$$

Exemple 2.1.1 *Voyons sur cet exemple la construction d'un tableau de contingence*

croisant la couleur des yeux et de cheveux de 592 *femmes d'une population (exemple proposé par Cohen* (1988) *et repris par Lebart et al.* (1995)*).*

	Brun	Châtain	Roux	Blond	Total
Marron	68	119	26	7	220
Noisette	15	54	14	10	93
Vert	5	29	14	16	64
Bleu	20	84	17	94	215
Total	108	286	71	127	592

TAB. 2.2 – Tableau de contingence, répartition de 592 femmes suivant les couleurs des yeux et des cheveux

À travers le tableau (2.2) nous pouvons dire qu'il y a 68 femmes de cheveux brun et de yeux marron (case n_{11}),qu'il y a 64 femmes de yeux verts (case $n_{.3}$), qu'il y a 286 femmes de cheveux châtain (case $n_{.2}$) , et qu'il y a en tout 592 femmes (case n).

Il est clair que, le tableau de contingence donne le résultat d'un traitement des données, mais ne permet pas de répondre directement à des questions du type "la proportion des femmes aux cheveux blonds est plus élevée chez les femmes aux yeux bleus qu'aux yeux bruns ?", c'est à dire de comparer les proportions des femmes de tel ou tel type. Alors on construit généralement deux types de tableau indiquant les pourcentages en lignes ou en colonnes s'appellent respectivement tableau des profils en lignes ou en colonnes.

2.1.2 Profils lignes et profils colonnes

On appelle tableau des profiles-lignes, le tableau des fréquences conditionnelles $\frac{n_{ij}}{n_{i.}}$, et le tableau des profiles-colonnes, le tableau des fréquences conditionnelles $\frac{n_{ij}}{n_{.j}}$.

Remarque 2.1.1 *L'interprétation des deux tableaux est évidement différente, car les rapports ne sont pas effectués sur une même base de référence.*

Exemple 2.1.2 *Construction des profils en lignes et on colonnes d'un tableau de contingence croisant la couleur de yeux et de cheveux des femmes.*

Profils en lignes

	Brun	Châtain	Roux	Blond	Total
Marron	31%	54%	12%	3%	100%
Noisette	16%	58%	15%	11%	100%
Vert	8%	45%	22%	25%	100%
Bleu	9%	39%	8%	44%	100%
Total	18%	48%	12%	22%	100%

TAB. 2.3 – Tableau des profils-lignes

À partir le tableau (2.3) nous observons par exemple que la proportion de femmes de couleur de cheveux roux est 12% mais qu'elle est plus élève dans les femmes d'yeux vert (22%) que de yeux bleu (8%).

Profils en colonnes

	Brun	Châtain	Roux	Blond	Total
Marron	63%	42%	37%	6%	37%
Noisette	14%	19%	20%	8%	16%
Vert	5%	10%	20%	13%	11%
Bleu	19%	29%	24%	74%	36%
Total	100%	100%	100%	100%	100%

TAB. 2.4 – Tableau des profils-colonnes

Par le tableau (2.4) nous pouvons dire de même que les femmes de yeux vert ne totalise que 11% de l'ensemble des femmes mais on y trouve 20% de cheveux roux et seulement 5% de cheveux brun de l'ensemble des femmes.

Lorsque l'on commente les résultats on remarque qu'une même case du tableau de contingence peut être décrite de deux façons différentes. Si l'on prend la case n_{22}, elle indique que les 54 femmes de couleur de cheveux châtain et de couleur d'yeux noisette représentent 58% des femmes d'yeux noisette et 19% de couleur de cheveux châtain parmi 592 femmes. De plus, la fréquence marginale en ligne ou en colonne soit systématiquement égale à 100%.

2.1.3 Calcul des effectifs théoriques et des écarts à l'indépendance

Nous avons vu qu'il y a différentes formes d'organisation dans un tableau de contingence, aussi il y a une autre manière d'étude basé sur le calcul des effectifs théoriques n^*_{ij} a partir des effectifs marginaux $n_{i.}$ et $n_{.j}$, sont données par la formule :

$$\frac{n_{i.}.n_{.j}}{n}$$

Cet effectif théorique serait obtenu s'il existe une indépendance entre les deux modalités de deux variables, lorsque il existe des écarts entre les deux effectifs théoriques et observées on peut le calculer par :

$$D_{n_{ij}} = n_{ij} - n^*_{ij}$$

Exemple 2.1.3 *En reprenant notre exemple pour la construction des effectifs théoriques et l'écart à l'indépendance du tableau de contingence croisant la couleur d'yeux et de cheveux des femmes.*

Les effectifs théoriques

	Brun	Châtain	Roux	Blond	Total
Marron	40.14	106.28	26.39	47.20	220.01
Noisette	16.97	44.93	11.15	19.95	93
Vert	11.68	30.92	7.68	13.73	64.01
Bleu	39.22	103.87	25.79	46.12	215
Total	108.01	286	71.01	127	592

TAB. 2.5 – Tableau des effectifs théoriques

Écart à l'indépendance

	Brun	Châtain	Roux	Blond	Total
Marron	27.86	12.72	−0.39	−40.20	−0.01
Noisette	−1.97	9.07	2.85	−9.95	0
Vert	−6.68	−1.92	6.32	2.27	−0.01
Bleu	−19.22	−19.87	−8.79	47.88	0
Total	−0.01	0	−0.01	0	−0.02

TAB. 2.6 – Tableau d'écart à l'indépendance

Il y a un grand nombre de tests permettant de mesurer le degré de significativité de la relation entre deux variables qualitatives, le test le plus fréquemment utilisé et le mieux adapté à la plupart des situations est le test du Khi-deux. Le principe de ce test est de mesurer l'écart entre les effectifs observées n_{ij} et les effectifs théoriques n^*_{ij} que l'on obtiendrait si les deux variables sont indépendantes sous l'hypothèse nulle.

La statistique de Khi-deux définie par la formule :

$$\chi^2 = \sum \frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}}.$$

Exemple 2.1.4 *Reprenons notre exemple pour la détermination de valeur du Khi-deux du tableau de contingence croisant la couleur d'yeux et de cheveux des femmes :*

	Brun	Châtain	Roux	Blond	Total
Marron	19.34	1.52	0.01	34.24	55.11
Noisette	0.23	1.83	0.73	4.96	7.75
Vert	3.82	0.12	5.20	0.38	9.52
Bleu	9.42	3.80	3.00	49.71	65.93
Total	32.81	7.27	8.94	89.29	138.31

TAB. 2.7 – Tableau de calcul de la statistique du khi-deux

À partir le tableau (2.7) nous pouvons déterminer la valeur du Khi-deux, qu'est égale à 138.31.

2.2 Test d'indépendance du Khi-deux

Pour réaliser le test d'indépendance il faut, d'abord, poser l'hypothèse nulle H_0 :

"Il n'y a pas de relation entre les deux variables qualitatives"

Ensuite, déterminer la valeur de Khi-deux du tableau étudié, puis définit le nombre de degré de liberté qui dépend du nombre de lignes r et de colonnes s du tableau de contingence donné par $ddl = (r-1) * (s-1)$. Et après, fixer un risque d'erreur $\alpha \in [0, 1]$, qui est la probabilité de rejeter l'hypothèse nulle.

Enfin, déterminer la valeur $\chi^2_{(1-\alpha)}(r-1)(s-1)$ et on applique la règle de décision suivante :

Si $\chi^2 > \chi^2_{(1-\alpha)}(r-1)(s-1)$, alors l'hypothèse nulle est rejetée avec un risque d'erreur α.

Exemple 2.2.1 *Dans l'exemple précédent, pour un risque d'erreur* $\alpha = 5\%$*, la valeur de Khi-deux correspondant à un degré de liberté est* $Khi - deux(9, 0.95) = 16.9$ *puisque* $\chi^2 = 138.31 > 16.9$*. On peut rejeter* H_0 *et affirme avec un risque d'erreur* 5% *que la couleur des yeux et des cheveux sont liées.*

Conclusion

Un des principaux intérêts en statistique et probabilité l'étude des mesures de dépendance, notamment dans le cas entre deux variables grâce à beaucoup de domaine d'application, Elle permet de mettre en évidence la présence ou l'absence d'une éventuelle relation des variables sous études. De plus à la description d'un ensemble des données, elle peut encore détermine la force et la forme de la liaison entre les variables sous études.

Aujourd'hui l'analyse des mesures d'association est nécessaire dans tous les secteurs de l'activité humaine pour confirmer ou contredire une relation entre différente phénomènes. Certainement, les mesures d'association sont des outils essentiels pour l'interprétation et la compréhension des phénomènes complexe, dont en cherche à quantifier ou à mesurer la qualité d'ajustement, la liaison, ainsi que la dépendance entre une ou plusieurs variables présentant le phénomène étudier ou le modèle correspondant. Comme exemple, en épidémiologique, en cherche les mesures d'association entre une maladie et un facteur de risque.

Ce mémoire donne donc une aidé générale sur ces différentes mesures d'association, leurs interprétations, leurs applications et les différences entre elles.

Bibliographie

[1] Bailly, P., Carrère, C. (2015). Statistiques descriptives : Théorie et applications, PUG, coll., p. 165-167.

[2] .Bernard, P.M., Lapointe, C. (1995). Mesures statistiques en épidémiologie, Presses de l'Université du Québec, Sainte-Foy, page 89.

[3] Cohen, J. (1988). Statistical power analysis for the behavioral sciences, 2nd edn. Á/L.

[4] Frechet, M. (1934). Sur l'usage du soi-disant coefficient de corrélation, Rapport pour la 22e session de l'IIS à Londres, Bulletin de l'IIS.

[5] Lebart, L., Morineau, A., Piron, M. (1995). Statistique exploratoire multidimensionnelle (Vol. 3). Paris : Dunod.

[6] Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London. 58 : 240–242.

[7] Olivier, L., Michel, R., Spiegel, A., Boutin, JP. (2003). Les mesures d'association en épidémiologie. Med Trop, 63 : 75-78.

[8] Rakotomalala, R. (2011). Etude des dépendances-Variables qualitatives Tableau de contingence et mesures d'association. Support de cours. Université Lumière Lyon, 2.

[9] Rakotomalala, R. (2015). Analyse de corrélation. Cours statistique à l'université de lumière Lyon, 2, 89.

[10] Rousson, V. (2013). Statistique appliquée aux sciences de la vie. Springer.

Annexe A : Logiciel R

2.3 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 Developement 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.

Annexe B : Abréviations et Notations

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

$E(.)$	:	Espérence mathématique.
$V(.)$	:	Variance mathématique.
$Cov(X,Y)$	:	Covariance mathématique du couple (X,Y).
$:=$	:	Égalité par définition.
$\overline{x},\overline{y}$	:	Moyenne empirique de X et Y respectivement.
$\hat{S}_{XY}$	:	Covariance empirique.
σ_X,σ_Y	:	Écart-type de X et Y respectivement.
$r(X,Y)$	:	Corrélation mathématique du couple(X,Y).
$\hat{r}_{XY}$	:	Corrélation empirique.
H_0	:	Hypothèse nulle.
H_1	:	Hypothèse alternative.
T	:	Statistique de test student.
$R.C$	:	Région critique.

ϕ	:	Coefficient phi.
Q	:	Fonction de concordance.
i.e.	:	C'est-à dire.
ρ	:	Rho 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.
S_r^2, S_s^2	:	Covariance de rang des observations de X et Y respectivement.
d_i	:	Différence entre les rangs des observations de X et Y.
τ	:	Tau de Kendall.
n_c, n_d	:	Nombre de paires concordantes, discordantes.
τ_b	:	Tau-b.
e_g	:	Nombre d'occurrence pour chaque valeur de X ou Y
p, q	:	Nombre de modalités.
r, s	:	Nombre de lignes et de colonnes du tableau.
n_{ij}	:	Nombre des individus.
$n_{i.}, n_{.j}$	:	Marge lignes et colonnes respectivement.
n_{ij}^*	:	Effectif théorique.
$D_{n_{ij}}$	:	Écart entre l'effectif théorique et observée.
χ	:	Khi-deux.
α	:	Risque d'erreur.

ملخص

في هذه المذكرة، حاولنا إلقاء الضوء على مفهوم مقاييس الارتباط، ضرورتها واستخداماتها بشكل عام. باستخدام برنامج التحليل الإحصائي "**R**" نقترح بعض التطبيقات لتوضيح مختلف النتائج النظرية التي تمت دراستها.

Résumé

Dans ce mémoire, nous avons essayé d'éclairer d'une façon générale le concept des mesures d'association, leur nécessité et leurs utilités. À l'aide du programme d'analyse statistique «R», nous proposons quelques applications pour clarifier les différents résultats théoriques étudiés.

Abstract

In this memory, we have tried to shed light on the concept of association measures, their necessity and their uses in general. Using the statistical analysis program "R", we propose some applications to clarify the different theoretical results studied.