République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITE MOHAMED KHIDER, BISKRA

FACULTE des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE

DEPARTEMENT DE MATHEMATIQUES

Mémoire présenté par

KARA Lobna

En vue de l'obtention du Diplôme de

MASTER en Mathématiques

Option : **Statistique**

Titre

Inférence statistique sur le coefficient de corrélation linéaire

Membres du Comité d'Examen

Pr.	Benatia Fatah	UMKB	Président
Pr.	Meraghni Djamel	UMKB	Encadreur
Dr.	Chine Amel	UMKB	Examinatrice

Juin 2022

Dédicace

À mon cher père et ma chère mère

pour tout leurs sacrifices, leur amour, leur tendresse, leur soutien et leur prières tout au long de mes études

À mes chers frères et sœurs

À tout ma famille

À tous mes collèges qui m'ont accompagné tout au long de mon parcours universitaire

À tous ceux que j'aime et tous ceux qui m'aiment

Merci d'être toujours là pour moi

Lobna Kara

REMERCIEMENTS

Au debut et avant tout, je tiens à remercier Allah le tous puissant qui m'a donné la force d'accomplir ce modeste travail

mes sicéres remerciment a mon encadreur Pr.Meraghni Djamel qui m'a guidé et aidé a realiser ce travail, pour ces conceils et son encouragement

je lui exprime toute ma gratitude et respect

je tiens a remercier les membres du jury Pr.Benatia Fatah et Dr.Chine Amel pour l'intérêt qu'ils ont porté à ce travail

je veux remercier aussi tous les enseignants du deparement de Mathématiques qui m'ont aidé tout au long mon parcours universitaire

Lobna Kara

Table des matières

Table des figures

Liste des tableaux

Introduction

En statistique, le coefficient de corrélation linéaire, aussi appelé coefficient de corrélation de Pearson et généralement noté ρ, est une mesure d'association. Il permet d'étudier et de quantifier le lien entre différentes variables aléatoires quantitatives et de déterminer la distribution et la direction d'un nuage de points représentant les observations d'un couple de variables aléatoires.

Ce mémoire est composé de deux chapitres. Le premier chapitre introduit le concept de couple aléatoire (continu et discret), avec les définitions et propriétés fondamentales, telles que la loi conjointe, les lois marginales, conditionnelles ainsi que les paramètres statistiques (espérance et matrice de covariance). Le deuxieme chapitre présente la méthode de calcul et d'estimation d'un coeffitient de corrélation linéaire et son application en regression linéaire. Finalement, discute trois types de tests liés aux coefficients de corrélation linéaire. Le test de significativité qui permet de déterminer si ce coefficient est nul ou non, le test de conformité à une valeur hypothétique et le test de comparaison de deux coefficients de corrélation linéaires.

Enfin, je tiens à noter que ce travail s'inspire essentiellement des références [3], [4] et [6], et que les études de simulation, avec calculs et graphes, sont réalisées en utilisant le logiciel de traitement statistique R [2].

Chapitre 1

Couples aléatoires

Dans ce chapitre, on généralise les définitions et propriétés relatives à une variable aléatoire (v.a) réelle au cas où on dispose de deux v.a. On parle alors de couple aléatoire réel ou de vecteur aléatoire de dimension 2.

1.1 Définitions

Soit un $(\Omega, \mathcal{F}, P)$ un espace de probabilités.

1.1.1 Variable aléatoire

Une v.a réelle est une application mesurable X définie sur l'ensemble fondamental Ω à valeurs dans $\mathbb{R}$:

$$\begin{array}{rccc} X: & \Omega & \rightarrow & \mathbb{R} \\ & \omega & \mapsto & X(\omega) \end{array}$$

Selon l'ensemble des valeurs $X(\Omega)$, on distingue deux types de v.a.

1. v.a disrète : concerne les sujet de dénombrement. Les v.a discrètes usuelles sont Bernouilli (succès-échec), binomiale (nombre de succès lors d'un certain nombre

d'épreuves indépendantes), Poisson (nombre d'arrivées pendant une certaine période),...

2. v.a continue : traite les problèmes de mesure. Les plus connues sont les variabes de Gauss (normale), de Pearson (Khi-deux), de Student, de Fisher ainsi que la v.a exponentielle.

Exemple 1.1.1

1. *Nombre d'étudiants absents pendant un mois : Poisson.*
2. *Temps d'attente dans un ceratin service administratif : exponentielle.*

1.1.2 Couple aléatoire

Un couple aléatoire réel (X, Y) est une application mesurable de Ω dans $\mathbb{R}^2$. Il s'écrit sous la forme d'un vecteur colonne $(X, Y)^t$ dont les composantes X et Y sont des v.a réelles définies sur l'ensemble Ω.
La notion de couple est nécessaire dans l'étude du comportement simultané des deux v.a X et Y.

Remarque 1.1.1

1. *Cette définition se généralise au cas où l'ensemble d'arrivée est $\mathbb{R}^n$, avec $n \geq 2$. On parle alors de vecteur aléatoire réel de dimension n : $(X_1, X_2, ..., X_n)^t$.*
2. *Lorsque les v.a X et Y sont toutes les deux discrètes, on dira que le couple (X, Y) est discret. Si les deux v.a X et Y sont continues, on dira que (X, Y) est continu.*

Exemple 1.1.2

1. *Etude de la relation entre le nombre b'absences d'un étudiant et la nombre de modules acquis.*
2. *Etude de l'influence de la taille d'une personne sur son poids.*

1.2 Loi de probabilité

Soit (X, Y) un couple aléatoire. Sa loi de probabilité est déterminée par la fonction de répartition. Dans le cas continu, cette loi peut aussi être obtenue par la fonction de densité et dans le cas discret on utilise les probabilités individuelles.

1.2.1 Fonction de répartition

Définition 1.2.1 (fonction de répartition conjointe) *On appelle fonction de répartition conjointe de (X, Y), qu'on note F, la fonction définie sur $\mathbb{R}^2$par*

$$F(x, y) := P(X \leqslant x, Y \leqslant y).$$

Propriété 1.2.1 *La fonction de répartition d'un couple aléatoire (X, Y) jouit des propriétés suivantes :*

- $\forall (x, y) \in \mathbb{R}^2$, $0 \leq F(x, y) \leq 1$.
- F est croissante par rapport à chacune des variables x et y.
- $\lim\limits_{x,y \to +\infty} F(x, y) = 1$ et $\lim\limits_{x,y \to -\infty} F(x, y) = 0$.
- F est continuité à droite en tout point (x_0, y_0) de $\mathbb{R}^2$: $\lim\limits_{(x,y) \overset{\geq}{\to} (x_0,y_0)} F(x, y) = F(x_0, y_0)$.

1.2.2 Masse ou densité de probabilité

Cas discret

Dans ce cas, le couple (X, Y) prend des valeurs $\{(x_i, y_j),\ 1 \leq i, j \leq n\}$, où $n \in \mathbb{N}^*$. La masse de probabilité conjointe, que l'on notera par P, donne les probabilitées associées à ces couples de valeurs. Donc, pour tout couple (x_i, y_j), on a

$$p_{ij} = P(x_i, y_j) := P(X = x_i, Y = y_j).$$

Elle vérifie les deux conditions suivantes :

- $p_{ij} \geq 0,\ i, j = 1, ..., n.$
- $\sum_{i=1}^{n} \sum_{j=1}^{n} p_{ij} = 1.$

Cas continu

Dans ce cas, la fonction de répartition F est dérivable par rapport aux deux variables. La densité de probabilité conjointe, notée par f, est alors définie sur $\mathbb{R}^2$ par

$$f(x, y) := \frac{\partial^2 F(x, y)}{\partial x \partial y}.$$

Elle vérifie les deux conditions suivantes :

- $f(x, y) \geq 0,\ \forall (x, y) \in \mathbb{R}^2.$
- $\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x, y)dxdy = 1.$

Remarque 1.2.1 *La fonction de répartition du couple (X, Y) se calcule par*

$$F(x, y) = \begin{cases} \sum^{x}\sum^{y} P(u, v) & si \quad (X, Y) \text{ est discret,} \\ \\ \displaystyle\int_{-\infty}^{x}\int_{-\infty}^{y} f(u, v)dudv & si \quad (X, Y) \text{ est continu.} \end{cases}$$

1.2.3 Lois marginales

Soit (X, Y) un couple aléatoire de fonction de répartition F. Quand on s'intéresse à un évènement relatif à l'une des deux v.a X ou Y indépendamment de l'autre, on définit ce que l'on appelle lois marginales.

Définition 1.2.2 (fonctions de répartition marginales) *Les fonction de répartitions marginales de X et Y sont respectivement définie sur $\mathbb{R}$ par*

$$F_X(x) := P(X \leqslant x, Y \in \mathbb{R}) = \lim_{y \to +\infty} F(x, y),$$

et

$$F_Y(y) := P(X \in \mathbb{R}, Y \leq y) = \lim_{x \to +\infty} F(x, y).$$

Les masses et densités marginales sont alors égales à

$$P_X(x_i) = \sum_{j=1}^{n} P(x_i, y_j) \text{ et } P_Y(y_j) = \sum_{i=1}^{n} P(x_i, y_j),$$

et

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y)dy \text{ et } f_Y(y) = \int_{-\infty}^{+\infty} f(x, y)dx.$$

1.2.4 Lois conditionnelles

On s'intéresse maintenant à la loi de probabilité de l'une des deux v.a sachant que l'autre est connue.

Cas discret

La loi de X sachant Y est définie par

$$p_{i|j} = P(X = x_i \mid Y = y_j) := \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)}, \; i = 1, ..., n.$$

De même, la loi de Y sachant X est définie par

$$p_{j|i} = P(Y = y_j \mid X = x_i) := \frac{P(X = x_i, Y = y_j)}{P(X = x_i)}, \; j = 1, ..., n.$$

Remarque 1.2.2 *On vérifie que ce sont bien des lois de probabilité. En effet,*

1. *Sur* $\Omega_X = \{x_i, i = 1, ..., n\}$, *on a, pour* j *fixé,*

$$\sum_{i=1}^{n} P(X = x_i \mid Y = y_j) = \frac{1}{P(Y = y_j)} \sum_{i=1}^{n} P(X = x_i, Y = y_j) = 1.$$

2. Sur $\Omega_Y = \{y_j, j = 1, ..., n\}$*, on a, pour i fixé,*

$$\sum_{j=1}^{n} P(Y = y_j \mid X = x_i) = \frac{1}{P(X = x_i)} \sum_{j=1}^{n} P(X = x_i, Y = y_j) = 1.$$

Cas continu

Les lois de X sachant Y et de Y sachant X sont définies par les densités conditionnelles respectives

$$f_{X|Y}(x) := \frac{f(x,y)}{f_Y(y)} \text{ et } f_{Y|X}(y) := \frac{f(x,y)}{f_X(x)}.$$

Remarque 1.2.3 *On vérifie que ce sont bien des densités de probabilité. En effet, on a*

$$\int_{-\infty}^{\infty} f_{X|Y}(x)dx = \frac{1}{f_Y(y)} \int_{-\infty}^{\infty} f(x,y)dx = 1.$$

De même pour l'autre densité conditionnelle.

1.2.5 Indépendance

Définition 1.2.3 *Deux v.a* X *et* Y *sont dites indépendantes si tout événement relatif à* X *est indépendant de tout événement défini par* Y*.*

Remarque 1.2.4 *L'indépendance se traduit par l'égalité entre la loi conditionnelle et la loi marginale, c-à-d*

$$p_{i|j} = P(X = x_i) \text{ ou } p_{j|i} = P(Y = y_j),$$

ou

$$f_{X|Y}(x) = f_X(x) \text{ ou } f_{Y|X}(y) = f_Y(y).$$

En utilisant les définitions des lois conditionnelles, on obtient la condition qui permet de définir l'indépendance entre deux v.a X *et* Y *:*

$$p_{ij} = P(X = x_i)P(Y = y_j) \text{ ou } f(x,y) = f_X(x)f_Y(y).$$

1.3 Paramètres statistiques

Soit $V = (X, Y)^t$ un couple aléatoire.

1.3.1 Espérance

On appelle espérance ou moyenne de V, $E(V)$ en abrégé, le vecteur des moyennes des deux v.a X et Y. On la note généralement par μ :

$$\mu = E(V) := \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix},$$

où, pour une v.a Z, on a

$$\mu_Z = E(Z) := \begin{cases} \sum_i z_i P(Z = z_i) & \text{si} \quad Z \text{ est discrète,} \\ \int z f_Z(z) dz & \text{si} \quad Z \text{ est continue.} \end{cases}$$

Remarque 1.3.1

1. *L'espérance est linéaire :* $E(U + \lambda V) = E(U) + \lambda E(V)$, $\lambda \in \mathbb{R}$.
2. *Si* X *et* Y *sont indépendantes, alors*

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

1.3.2 Matrice de covariance

La matrice de covariance de V, $Cov(V)$ en abrégé, est une matrice carrée symétrique d'ordre 2, généralement notée par Σ :

$$\Sigma = Cov(V) := \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix}, \tag{1.1}$$

où

$$\sigma_{XY} = cov(X,Y) := E\{(X - E(X))(Y - E(Y))\} = E(XY) - E(X)E(Y),$$

représente la covariance entre les deux v.a X et Y, et

$$\sigma_Z^2 = \sigma_{ZZ},$$

représente la variance d'une v.a Z (sa racine carrée σ_Z est appelée écart type de Z).

Remarque 1.3.2

1. *Σ est une matrice définie positive.*
2. *$\Sigma = E(VV') - \mu\mu'$.*
3. *Si X et Y sont indépendantes, alors Σ est une matrice diagonale. En effet, on a*

$$\sigma_{XY} = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0.$$

La réciproque n'est pas toujours vraie.

1.3.3 Matrice de corrélation

La matrice de corrélation de V, $Cor(V)$ en abrégé, est une matrice carrée symétrique d'ordre 2, définie par

$$Cor(V) := \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \tag{1.2}$$

où

$$\rho = cor(X,Y) := \frac{\sigma_{XY}}{\sigma_X \sigma_Y}, \tag{1.3}$$

représente le coefficient de corrélation linéaire des deux v.a X et Y.

Remarque 1.3.3

1. *$Cor(V)$ est une matrice définie positive.*
2. *Si X et Y sont indépendantes, alors $Cor(V)$ est égale à la matrice identité. La réciproque n'est pas toujours vraie.*
3. *En termes du coefficient de corrélation, on a*

$$\Sigma = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}. \tag{1.4}$$

Remarque 1.3.4 *On peut aussi définir les paramètres statistiques par rapport aux lois conditionnelles.*

Définition 1.3.1 (espérance conditionnelle) *On appelle espérance de Y sachant que $X = x$ et on note $E(Y \mid X = x)$ la quantité*

$$E(Y \mid X = x) := \begin{cases} \sum\limits_{j\in J} y_j P(Y = y_j \mid X = x) = \sum\limits_{j\in J} y_j P_{j|i}, & \textit{cas discret} \\ \int\limits_{\mathbb{R}} y f_{Y|X}(x,y)dy, & \textit{cas continu} \end{cases}.$$

Définition 1.3.2 (variance conditionnelle) *On appelle variance de Y sachant que $X = x$ et on note $V(Y \mid X = x)$ la quantité*

$$V(Y \mid X = x) := E\left\{[Y - E(Y \mid X = x)]^2 \mid X = x\right\}.$$

Pour plus de détails sur l'epérance et la variance conditionnelles, voir [6, pages 71-75].

$Y \backslash X$	-2	0	2	marge
-1	0.1	0.2	0.1	0.4
2	0.2	0.2	0.2	0.6
marge	0.3	0.4	0.3	1

TAB. 1.1 – Distribution de probabilité d'un couple aléatoire avec ses marginales

	-2	0	2
$X\|Y=-1$	$1/4$	$1/2$	$1/4$
$X\|Y=2$	$1/3$	$1/3$	$1/3$

TAB. 1.2 – Distribution conditionnelle de X sachant Y

1.4 Exemples

1.4.1 Couple discret

Cet exemple se trouve dans [3, page 109]. Soit $V = (X, Y)$ un couple aléatoire dont la distribution est donnée par le tableau 1.1.

Lois conditionelles

Les Lois conditionelles sont résumées dans les tableaux 1.2 et 1.3.

Paramètres statistiques

D'après la difinition de l'espérance d'un couple aléatoire, on a

$$\mu = E(V) = \begin{pmatrix} 0 \\ 0.8 \end{pmatrix}.$$

On a

$$Var(X) = 2.4,\ Var(Y) = 3.44 \text{ et } Cov\,(X, Y) = 0.$$

Alors ,d'après (1.1) et (1.2), les matrices de covariance et corrélation de V sont respecti-

	-1	2
$Y\|X=-2$	$1/3$	$2/3$
$Y\|X=0$	$1/2$	$1/2$
$Y\|X=2$	$1/3$	$2/3$

TAB. 1.3 – Distribution conditionnelle de Y sachant X

vement

$$\Sigma = \begin{bmatrix} 2.4 & 0 \\ 0 & 3.44 \end{bmatrix} \text{ et } Cor(V) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Remarque 1.4.1

Bien que $Cov(X,Y) = Cor(X,Y) = 0$, les deux v.a ne sont pas indépendantes (voir les lois marginales et conditionnelles). On a ici un contre exemple à la remarque 3 de (1.3.2) (ou la remarque 2 de (1.3.3)).

1.4.2 Couple continu

Exemple 1 : loi normale bidimensionnelle $V \sim \mathcal{N}_2(\mu, \Sigma)$ ([6, page 90])

$$f(x,y) = \frac{1}{2\pi\sqrt{\det \Sigma}} \exp\left\{-\frac{1}{2}(v-\mu)^t \Sigma^{-1}(v-\mu)\right\}$$

En utilisant la forme (1.4) , on a $\det \Sigma = \sigma_X^2\sigma_Y^2(1-\rho^2)$ et

$$\Sigma^{-1} = \frac{1}{\sigma_X^2\sigma_Y^2(1-\rho^2)} \begin{bmatrix} \sigma_Y^2 & -\rho\sigma_X\sigma_Y \\ -\rho\sigma_X\sigma_Y & \sigma_X^2 \end{bmatrix}.$$

La densité de probabilité s'écrit alors

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \times$$
$$\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}.$$

Si les v.a X et Y sont centrées et réduites, alors $V \sim \mathcal{N}_2(0, I_2)$ et la densité∘ conjointe devient

$$f(x,y) = \frac{1}{2\pi}\exp\left\{-\frac{1}{2}\left[x^2+y^2\right]\right\}.$$

Dans ce cas particulier, les densités marginales sont

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \text{ et } f_Y(y) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\},$$

et les densités conditionnelles sont

$$f_{X|Y}(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \text{ et } f_{Y|X}(y) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\}.$$

On constate que $f_{X|Y}(x) = f_X(x)$ et $f_{Y|X}(y) = f_Y(y)$ sur $\mathbb{R}$. Ceci entraîne que

$$f(x,y) = f_X(x) f_Y(y)$$

et par conséquent les deux v.a X et Y sont indépendantes. D'où le résultat de la Proposition (1.4.1).

Puisque X et Y sont de même distribution $\mathcal{N}(0,1)$, alors l'espérance et les matrices de covariance et de corrélation de V sont

$$E(V) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ et } Cov(V) = Cor(V) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Proposition 1.4.1 *Soit (X,Y) un couple aléatoire Gaussien. Alors les deux v.a X et Y sont indépendantes si et seulement si elles sont non corrélées.*

Exemple 2 : cet exemple est extrait du cours [1]. On suppose que la densité de probabilité du couple V est définie sur $\mathbb{R}^2$ par

$$f(x,y) = \begin{cases} 2\exp\{-(x+y)\} & \text{si} \quad 0 \le x \le y \\ 0 & \text{sinon} \end{cases}$$

La densité marginale de X est définie, pour $x \geq 0$, par

$$\begin{aligned} f_X(x) &= \int_x^\infty f(x,y)dy = 2\exp(-x)\int_x^\infty \exp(-y)dy \\ &= -2\exp(-x)\exp(-y)\ |_x^\infty) = 2\exp(-2x), \end{aligned}$$

c-à-d

$$f_X(x) = 2\exp(-2x)\mathbf{1}_{[0,+\infty[}(x).$$

Par le même raisonnement, on obtient la densité marginale de Y qui est définie pour $y \geq 0$ par

$$\begin{aligned} f_Y(y) &= \int_0^y f(x,y)dx = 2\exp(-y)\int_0^y \exp(-x)dx \\ &= 2\exp(-y)(-\exp(-x)\ |_0^y) = 2\exp(-y)(1-\exp(-y)), \end{aligned}$$

c-à-d

$$f_Y(y) = 2\exp(-y)(1-\exp(-y))\mathbf{1}_{[0,+\infty[}(y).$$

La densité conditionnelle de Y (sachant $X = x \geq 0$) est

$$f_{Y|X}(y) = \frac{2\exp(-(x+y))\mathbf{1}_{\{0\leq x\leq y\}}}{2\exp(-2x)} = \exp(x-y)\mathbf{1}_{\{0\leq x\leq y\}}.$$

De la même façon, on trouve la densité conditionnelle de X (sachant $Y = y \geq 0$) est

$$f_{X|Y}(x) = \frac{2\exp(-(x+y))\mathbf{1}_{\{0\leq x\leq y\}}}{2\exp(-y)(1-\exp(-y))} = \frac{\exp(-x)}{(1-\exp(-y))}\mathbf{1}_{\{0\leq x\leq y\}}.$$

Remarque 1.4.2

- *On constate que $f_X(x)$ est la densité de la loi exponentielle de paramètre 2, c-à-d la v.a $X \sim \mathcal{E}(2)$.*

- *Le produit des densités marginales est*

$$f_X(x)f_Y(y) = 4\exp(-(2x+y))(1-\exp(-y))\mathbf{1}_{[0,+\infty[^2}(x,y).$$

Il est différent de la densité conjointe $f(x,y)$, alors les v.a X et Y ne sont pas indépendantes.

Espérance de V : elle est égale à

$$E(V) = \begin{pmatrix} E(X) \\ E(Y) \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ \frac{3}{2} \end{pmatrix}.$$

En effet, puisque $X \sim \mathcal{E}(2)$ alors $E(X) = 1/2$. Pour la v.a Y, on a

$$\begin{aligned} E(Y) &= \int_{\mathbb{R}} y f_Y(y)dy = 2\int_0^\infty y(\exp(-y) - \exp(-2y))dy \\ &= 2\left[\int_0^\infty y\exp(-y)dy - \int_0^\infty y\exp(-2y)dy\right] =: 2\left[I_1 - I_2\right]. \end{aligned}$$

Les valeurs

$$I_1 = 1 \text{ et } I_2 = \frac{1}{4},$$

sont obtenues par des intégrations par parties. D'où $E(Y) = 3/2$.

Matrice de covariance de V

On a $X \sim \mathcal{E}(2) \longrightarrow Var(X) = 1/2^2 = 1/4$; $Var(Y) = 2$ et $Cov(X,Y) = 1/4$. D'où

$$Cov(V) = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & 2 \end{bmatrix}.$$

Matrice de correlation de V

D'après 1.3 on a $\rho = \frac{1}{4} / \left(\frac{1}{2} \times \sqrt{2}\right) = \sqrt{2}/4$. Donc de 1.2, on a

$$Cor(V) = \begin{bmatrix} 1 & \frac{\sqrt{2}}{4} \\ \frac{\sqrt{2}}{4} & 1 \end{bmatrix}.$$

Chapitre 2

Tests sur le coefficient de corrélation linéaire

2.1 Coéfficient de corrélation linéaire

On rappelle que le coefficient de corrélation linéaire ρ entre X et Y, défini par la realtion (1.3) , est égal au rapport de leur covariance sur le produit de leurs écarts types. On l'appelle aussi coefficient de corrélation de Pearson.

Remarque 2.1.1
1. *On a* $-1 \leq \rho \leq 1$.
2. *Si* X *et* Y *sont indépendantes alors* $\rho = 0$. *On dit alors que les deux v.a sont non corrélées.*

2.1.1 Interprétation du coefficient de corrélation linéaire

ρ mesure le degré de la relation linéaire entre deux variables quantitatives X et Y.

- $\rho > 0$: les valeurs de Y ont tendance à croître quand les valeurs de X augmentent.
- $\rho < 0$: les v.a X et Y varient en sens inverses.
- $|\rho| = 1$: il existe une relation linéaire (exacte) entre X et Y. Cette liaison est d'autant plus forte que $|\rho|$ est proche de 1.

- $\rho = 0$: absence de liaison linéaire entre les v.a X et Y.

2.1.2 Coefficient de corrélation linéaire empirique

Le coefficient ρ peut être estimé par son équivalent empirique R, obtenu à partir d'un échantillon $(X_1, Y_1), ..., (X_n, Y_n)$, de taille $n \geq 1$, du couple aléatoire (X, Y) :

$$R := \frac{S_{XY}}{S_X S_Y},$$

où

$$S_{XY} := \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}),$$

représente la covariance empirique et

$$S_X^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \text{ et } S_Y^2 := \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2,$$

sont les variances empiriques respectives de X et Y, avec

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n}X_i \text{ et } \overline{Y} := \frac{1}{n}\sum_{i=1}^{n}Y_i,$$

étant les moyennes empiriques de X et Y respectivement.

Remarque 2.1.2 *1. Les v.a R, S_{XY}, S_X^2 et S_Y^2 sont les estimateurs naturels des paramètres ρ, σ_{XY}, σ_X^2 et σ_Y^2 respectivement.*

2. R est un estimateur asymptotiquement non biaisé et convergent de ρ. On a

$$E(R) = \rho - \frac{\rho(1-\rho^2)}{2n} \text{ et } Var(R) = \frac{(1-\rho^2)^2}{n}.$$

3. Le coefficient de corrélation empirique R vérifient les mêmes propriétés que son équivalent théorique ρ, avec la même interprétation.

4. La valeur observée de la v.a R est

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

2.1.3 Régression linéaire

Les valeurs $\{(x_1, y_1), ..., (x_n, y_n)\}$, correspondants à l'échantillon $(X_1, Y_1), ..., (X_n, Y_n)$, forment un nuages de points dans le plan. La droite de régression est la droite qui passe globalement proche de l'ensemble des points du nuage. En d'autres termes, elle passe à plus faible distance de chaque point du nuage en moyenne. Son équation s'écrit sous la forme

$$y = \beta_0 + \beta_1 x.$$

Le modèle de régression s'écrit

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ i = 1, ..., n. \tag{2.1}$$

- Y est appelé le caractére expliqué et X le caractére explicatif.
- ε_i est l'erreur du modèle. C'est une v.a appelée aussi résidu.
- β_0, β_1sont des constantes appelées les coefficients du modèle.

Hypothèses du modèle (2.1)

- $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$ et $Cov(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$.
- $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$: normalité des résidus.
- Les v.a ε_i sont indépendantes.
- $Cov(X_i, \varepsilon_i) = 0$: non corrélation entre X et ε.

Droite de régression au sens des moindres carrés

Les coefficients β_0 et β_1 étant inconnues, l'objectif est donc de les estimer. Pour cela, on applique la méthode des moindres carrés dont le principe est de minimiser la distance (au carré)

$$d(y_i, \hat{y}_i) = f(\beta_0, \beta_1) := \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \tag{2.2}$$

avec $\hat{y}_i := E(Y_i | X_i = x_i) = \beta_0 + \beta_1 x_i$. On cherche donc les valeurs de β_0 et β_1, qu'on note $\widehat{\beta}_0$ et $\widehat{\beta}_1$, solutions de

$$\left(\widehat{\beta}_0, \widehat{\beta}_1\right) = \arg\min(f(\beta_0, \beta_1)).$$

En dérivant la fonction (2.2) par rapport à β_0 et β_1 et en annulant les dérivées partielle obtenues, on obtient les solutionset des systémes aux equation on trouve les solutions

$$\widehat{\beta}_1 = \frac{s_{XY}}{s_X^2}, \text{ et } \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}.$$

Ces valeurs peuvent être considérées comme des réalisations de deux v.a

$$\widehat{\beta}_1 = \frac{S_{XY}}{S_X^2}, \text{ et } \widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X}.$$

Ces dernières sont dites estimateurs des coefficients β_0 et β_1 du modèle de régression (2.1) .

Remarque 2.1.3

On obtient les mêmes estimateurs $\widehat{\beta}_0$ et $\widehat{\beta}_1$ en appliquant la méthode du maximum de vraisemblance.

Proposition 2.1.1

1. *Propriétés des estimateurs $\widehat{\beta}_0$ et $\widehat{\beta}_1$: ils sont sans biais et convergents. On a*

$$E(\widehat{\beta}_0) = \beta_0 \;\; et \;\; E(\widehat{\beta}_1) = \beta_1,$$

et

$$Var(\widehat{\beta}_0) = \hat{\sigma}^2_{\widehat{\beta}_0} = \frac{\sigma^2}{n}\left(1 + \frac{\left(\sum\limits_{i=1}^{n} x_i\right)^2}{n^2 s_X^2}\right) \;\; et \;\; Var(\widehat{\beta}_1) = \hat{\sigma}^2_{\widehat{\beta}_1} = \frac{\sigma^2}{n} \times \frac{1}{s_X^2}.$$

2. *Distributions d'échantillonage : on a*

$$\frac{\widehat{\beta}_0 - \beta_0}{\hat{\sigma}_{\widehat{\beta}_0}} \sim \mathcal{T}_{n-2} \ et \ \frac{\widehat{\beta}_1 - \beta_1}{\hat{\sigma}_{\widehat{\beta}_1}} \sim \mathcal{T}_{n-2}.$$

Ces distributions vont nous permettre de réaliser une inférence statistique sur les coefficients β_0 et β_1 du modèle de regression.

Preuve. Pour la démonstration des deux propositions (et d'autres propriétés des estimateurs $\widehat{\beta}_0$ et $\widehat{\beta}_1$), se reférer à [6, pages 390-394]. ■

2.2 Tests sur le coefficient de corrélation

2.2.1 Test de la significativité de ρ

Principe du test

Ce test est appelé test de la nullité du coefficient de corrélation. On se fixe un niveau de signification (ou risque d'erreur) $0 < \alpha < 1$ et on décide, après observation d'un échantillon, de rejeter ou non la nullité de ρ. Les valeurs usuelles de α sont 0.01, 0.05 ou 0.10.

- Si $\rho = 0$ alors il n'y a pas de liaison linéaire entre X et Y.
- Si $\rho \neq 0$ alors X et Y sont liées.

Conditions d'application

Il faut que les deux v.a X et Y soient normalement distribuées, c-à-d

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

Hypothèses de test

L'hypothèse nulle, notée $\mathcal{H}_0$, dit que le coefficient de corrélation est égal à zéro et l'hypothèse alternative, notée $\mathcal{H}_1$, dit le contraire. Le test s'écrit

$$\begin{cases} \mathcal{H}_0 : \rho = 0 \\ \mathcal{H}_1 : \rho \neq 0 \end{cases}$$

Statistique du test

On définit la statistique

$$T := \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}.$$

Sous l'hypothese $\mathcal{H}_0$, cette v.a suit la loi de Student à $n-2$ ddl (voir [6, page 395]), c-àd

$$T \sim \mathcal{T}_{n-2}, \text{ sous } \mathcal{H}_0.$$

Région critique et décision

La région critique W (ou région de rejet) du test est définie, au niveau de sigification α, par

$$W := \left\{ |T| > t_{n-2,\alpha/2} \right\},$$

où $t_{n-2,\alpha/2}$ est le quantile d'ordre $(1-\alpha/2)$ de $\mathcal{T}_{n-2}$. Si la valeur t de la statistique T appartient à W alors on décide de rejeter l'hypothèse $\mathcal{H}_0$ au risque α, et on conclut qu'il existe une liaison entre X et Y. Sinon, on ne peut la rejeter.

Remarque 2.2.1 *Tester la significativité de ρ est équivalent à tester la significativité du coefficient β_1 du modèle de régression linéaire (2.1).*

2.2.2 Test de conformité de ρ

Le test s'écrit

$$\begin{cases} \mathcal{H}_0 : \rho = \rho_0 \\ \mathcal{H}_1 : \rho \neq \rho_0 \end{cases}$$

Pour tester la conformité de ρ avec une valeur hypothétique ρ_0 autre que 0, il faut connaître la distribution de T en général. Pour cela, on utilise la transformation de Fisher (voir [5, page 17])

$$F := \frac{1}{2} \ln \frac{1+R}{1-R}. \tag{2.3}$$

La v.a Z est asymptotiquement normale de paramètres

$$E(F) \approx \frac{1}{2} \ln \frac{1+\rho}{1-\rho} \text{ et } Var(F) \approx \frac{1}{n-3}. \tag{2.4}$$

La région critique W du test est définie asymptotiquement par

$$W := \left\{ |Z| > z_{1-\alpha/2} \right\},$$

où

$$Z := \frac{F - E(F)}{\sqrt{Var(F)}} \rightsquigarrow \mathcal{N}(0,1), \tag{2.5}$$

et $z_{1-\alpha/2}$ est le quantile d'ordre $(1-\alpha/2)$ de $\mathcal{N}(0,1)$.

Si la valeur z de la v.a Z appartient à W alors on rejette l'hypothèse $\mathcal{H}_0$ au risque α, et on conclut que ρ ne peut pas être égal à ρ_0. Sinon, on ne peut la rejeter.

Remarque 2.2.2 *A partir des relations* (2.3), (2.4) *et* (2.5), *on peut construire un intervalle d'estimation pour* ρ *de niveau de confiance* $(1-\alpha)$.

2.2.3 Test de comapraison de deux coefficients de corrélation

On dispose de deux couples aléatoires $\left(X^{(1)}, Y^{(1)}\right)$ et $\left(X^{(2)}, Y^{(2)}\right)$ de coefficients de corrélation respectifs ρ_1 et ρ_2. A titre d'exemple, le premier couple peut désigner la taille et le poids chez les hommes et le deuxième la taille et le poids chez les femmes. Le but est de comparer les deux coefficients de corrélation. Pour cela, on extraît deux échantillons indépendants $E_1 = \left\{\left(X_1^{(1)}, Y_1^{(1)}\right), ..., \left(X_{n_1}^{(1)}, Y_{n_1}^{(1)}\right)\right\}$ et $E_2 = \left\{\left(X_1^{(2)}, Y_1^{(2)}\right), ..., \left(X_{n_2}^{(2)}, Y_{n_2}^{(2)}\right)\right\}$ de tailles respectives n_1 et n_2.

Le test s'ecrit

$$\begin{cases} \mathcal{H}_0 : \rho_1 = \rho_2 \\ \mathcal{H}_1 : \rho_1 \neq \rho_2 \end{cases}$$

Soient R_1 et R_2 les coefficients de corrélation empiriques correspondants à E_1 et E_2 respectivement. On définit la statistique

$$D := F_1 - F_2,$$

où F_1 et F_2 sont obtenues suivant (2.3). D'après [5, page 25], sous l'hypothèse nulle $\mathcal{H}_0$, on a

$$D \rightsquigarrow \mathcal{N}\left(0, \frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}\right),$$

La région critique W du test est définie asymptotiquement par

$$W := \left\{|Z| > z_{1-\alpha/2}\right\},$$

où

$$Z := \frac{D}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}.$$

Si la valeur z de la statistique du test Z est dans W, alors on rejette l'hypothèse $\mathcal{H}_0$ et on

variable	moyenne	variance	covariance	corrélation
vitesse (mph)	15.40	27.96	109.95	$\rho = 0.81$
distance (ft)	42.98	664.06		

TAB. 2.1 – Paramètres statistiques empiriques relatifs au couple vitesse-distance

conclut de la non égalité entre les deux coefficients avec un risque d'erreur égal à α. Sinon, on ne peut pas la rejeter (on l'accepte).

2.3 Exemple

Les données de cet exemple se trouvent sous le nom "cars" dans le package "datasets" du logiciel de traitement statistique R [2]. Le but est d'étudier l'impact de la vitesse X (mph) d'une voiture sur la distane d'arrêt Y (ft). Pour cela, un échantillon de 50 voitures différentes est considéré. Les résultats numériques sont résumés dans la tableau 2.1.

De la valeur du coefficient ρ et son signe, on conclut que la liaison entre la vitesse et la distance d'arrêt est positive et assez forte comme le montre la figure 2.1.

La valeur de la statistique du test de significativité de ρ est égale à 9.46 avec une p-value négligeable. Cette dernière signifie que l'hypothèse nulle ($\rho = 0$) doit être rejetée aux niveaux de signification usuels 1%, 5% et 10%. En d'autres termes, la corrélation entre les deux variables ne peut pas être nulle. Ce qui était logiquement attendu.

Code R

```
library (datasets)
data(cars)
df<-cars
X<-cars[,1] # vecteur des vitesses
Y<-cars[,2]  #Y vecteur des distances
Xb<-mean(X)  # moyenne empirique de X
Yb<-mean(Y)  # moyenne empirique de Y
Sx<-var(X)  # variance empirique de X
```

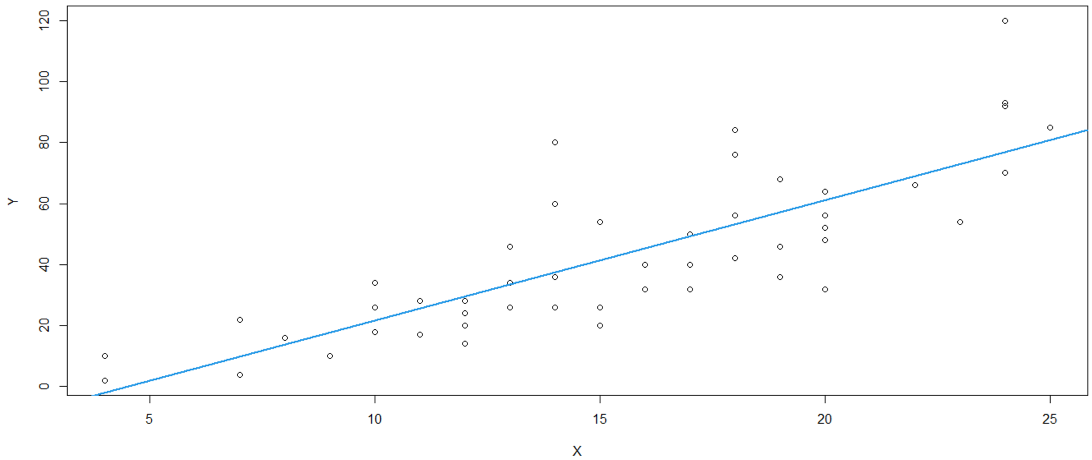

FIG. 2.1 – Nuage de points des vitesses et distances d'arrêt des voitures avec la droite de régression

```
Sy<-var(Y)  # variance empirique de Y
Sxy<-cov(X,Y)  # covariance empirique
R<-cor(X,Y)  # coefficient de corrélation
# resultats
round(Xb,2) ; round(Yb,2) ; round(Sx,2) ; round(Sy,2) ; round(Sxy,2) ; round(R,2)
# Nuage de points et droite de regression
plot(Y~X)
abline(lm(Y~X),col=4,lwd=2)
#Test de significativité du coefficient de corrélation
cor.test(X,Y,method = "pearson")
```

Conclusion

L'étude de la liaison entre deux ou plusieurs caractères est un axe très important de l'analyse statistique. Cette relation est évaluée à l'aide d'une variété de mesures.

Dans ce mémoire, on s'est intéressé au coefficient de corrélation linéaire de deux variables. Plus précisément, on a traité quelques un des tests d'hypothèses statistiques les plus populaires relatifs à ce dernier. Le test de significativité a pour but de vérifier si la valeur du coefficient est nulle ou non permettant ainsi de conclure de l'absence ou l'existence de liaison entre les deux variales. Le test de conformité vise à examiner la valeur théorique inconnue du coefficient par rapport à une valeur hypothétique donnée. Quant au test de comparaison, il permet de juger l'égalité entre deux coefficients de corrélation linéaire ou plus.

Enfin, on note qu'il existe des tests autres que ceux qu'on a discuté dans ce mémoire et que, en plus de la prise de décision, les tests satistiques aident dans la construction d'intervalles de confiance pour le coefficient de corrélation linéaire.

Bibliographie

[1] Hafayed, M. (2021) . Probabilités approfondies. Cours de $1^{ère}$ année master. Université de Biskra.

[2] Ihaka, R. and Gentleman, R. (1996). R : A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics 5, 299-314.

[3] Lecoutre, J.P. (2012). Statistique et probabilités. Dunod, Paris.

[4] Lejeune, M. (2010). Statistique, la théoie et ses applications. Springer.

[5] Rakotomalala, R. (2017) . Analyse de corrélation. Etude de dépendances-Variables quantitatives. Université Lumière Lyon 2.

[6] Saporta, G. (2011). Probabilités, analyse des données et statistique. Technip, Paris

Annexe : Abréviations et Notations

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

Notation	**Significations**
β_0, β_1	coefficient du modèle de regression
$\hat{\beta}_0, \hat{\beta}_1$	estimateurs de β_0 et β_1
c-à-d	c'est-à-dire
$Cor(X, Y)$	corrélation entre X et Y
$Cov(X, Y)$	covariance entre X et Y
ddl	degrée de liberté
F	transformation de Fisher
$F(x, y)$	fonction de répartition conjointe
$f(x, y)$	densité de probabilité conjointe
$F_X(x), F_Y(y)$	fonctions de repartition marginales
$f_{X\|Y}, f_{Y\|X}$	densités conditionelles
$\mathcal{H}_0$	hypothèse nulle
$\mathcal{H}_1$	hypothèse altérnative
μ	espérance du couple
μ_Z	espérance de Z
$\mathcal{N}(0, 1)$	loi normale centré réduite

P_{ij}	masse de probabilité conjointe
$p_X(x_i), p_Y(y_i)$	masses de probabilité marginales
$p_{i\|j}, p_{j\|i}$	masses de probabilité conditionelles
R	coefficient de corrélation empirique
ρ	coefficient de correlation
S_{XY}	covariance empirique entre X et Y
S_X^2, S_Y^2	variances empiriques
σ_Z^2	variance de Z
Σ	matrice de covariance
$\mathcal{T}_\nu$	loi de Student à ν ddl
W	région critique
$\bar{X}, \bar{Y}$	moyenne empirique
v.a	variable aléatoire
z_ω	quantile d'ordre ω de $\mathcal{N}(0,1)$
$:=$	égalité par définition
$\mathbf{1}$	fonction indicatrice

ملخص

يتم قياس الارتباط الخطي بين متغيرين عشوائيين بواسطة معامل بيرسون. قيمة وإشارة هذا الأخير تدلان على التوالي على مدى قوة واتجاه العلاقة الخطية. لإجراء اختبارات إحصائية لفرضيات خاصة بهذا المعامل نستعمل المعامل التجريبي الذي يحسب انطلاقا من عينة تستخرج من الزوج العشوائي. من بين هذه الاختبارات نذكر اختبار الدلالة، اختبار الامتثال واختبار مقارنة معاملين.

الكلمات المفتاحية: اختبار الامتثال؛ اختبار الدلالة؛ اختبار المقارنة؛ الارتباط الخطي؛ التوزيع المشترك؛ الزوج العشوائي؛ معامل بيرسون.

Abstract

The linear relationship between two random variables is measured by means of Pearson correlation coefficient. The value and the sign of the latter respectively indicate on the degree of intensity and the direction of the linear relation. To perform a statistical inference on this coefficient, one uses its empirical counterpart based on a sample from the random couple. Among the statistical tests, one can cite the signification test, the conformity test and the comparison test of two coefficients.

Keywords: Comparison test; Conformity test; Joint distribution; Linear correlation; Pearson coefficient; Random couple; Signification test.

Résumé

La liaison linéaire entre deux variables aléatoires est mesurée à l’aide du coefficient de corrélation de Pearson. La valeur et le signe de ce dernier indiquent respectivement le degré de force et la direction de la relation linéaire. Pour réaliser des tests d’hypothèses statistiques sur ce coefficient, on utilise sur son équivalent empirique basé sur un échantillon du couple aléatoire. Parmi ces tests, on peut citer le test de signification, le test de conformité et celui de comparaison de deux coefficients.

Mots clés : Coefficient de Pearson; Corrélation linéaire; Couple aléatoire; Distribution conjointe; Test de comparaison; Test de conformité; Test de signification.