Association de variables quantitatives

Cas paramétrique gaussien

Frédéric Bertrand[1]

[1]IRMA, Université de Strasbourg
Strasbourg, France

Master 1
2019

Sommaire

1. Coefficient de corrélation simple
   - Introduction
2. Le cas bidimensionnel
   - Loi normale bidimensionnelle
   - Estimation des paramètres
   - Procédure de test
   - Remarques sur les liaisons

Sommaire

3. Le cas général ($n \geqslant 2$)
   - Loi multinormale
   - Estimation
   - Test de l'hypothèse $\rho_{ij} = 0$
   - Test de l'hypothèse $\rho_{ij} = \rho_0$, $\rho_0 \neq 0$

Sommaire

4. Corrélation multiple
   - Définition
   - Estimation
   - Asymptotique
   - Test de l'hypothèse $R(X_1, \boldsymbol{X_2}) = 0$
   - Test de l'hypothèse $R(X_1, \boldsymbol{X_2}) = R_0$, $R_0 \neq 0$

Sommaire

5. Corrélation partielle
   - Définition
   - Estimation
   - Asymptotique
   - Test de l'hypothèse $\rho_{i,j|q+1,\dots,m} = 0$
   - Test de l'hypothèse $\rho_{i,j|q+1,\dots,m} = \rho_0$, $\rho_0 \neq 0$
   - Cas de trois variables réelles

Sommaire

1. Coefficient de corrélation simple
   - Introduction

Introduction

Le coefficient de corrélation simple $\rho(X, Y)$ mesure le degré d'association linéaire entre deux variables aléatoires X et Y.

$$\rho(X, Y) = \mathrm{Cor}(X, Y) = \frac{\mathrm{Cov}[X, Y]}{(\mathrm{Var}[X] \times \mathrm{Var}[Y])^{\frac{1}{2}}}.$$

On a toujours $-1 \leqslant \rho(X, Y) \leqslant 1$.

Introduction

Lorsque $\rho(X, Y) = 0$, on dit que les deux variables aléatoires X et Y sont non corrélées. Attention, ne pas être corrélés ne signifie pas généralement être indépendants. Mais l'indépendance implique toujours l'absence de corrélation.

Exemple

Prenons $X \sim \mathcal{N}(0, 1)$ et $Y = X^2$, on sait alors que $Y \sim \chi^2_1$ une loi du χ^2 à un degré de liberté. Ainsi les propriétés classiques d'une loi du χ^2 impliquent que $\mathbb{E}[Y] = 1$ et $\mathrm{Var}[Y] = 2$.

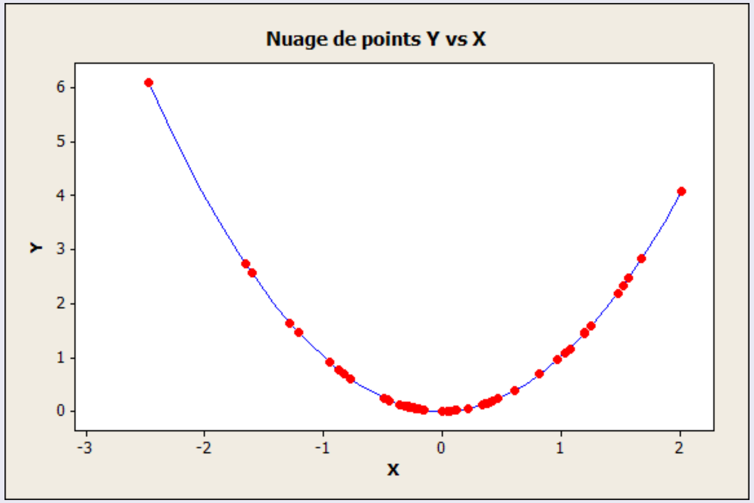
Exemple

Remarquez que nous ne sommes donc pas dans le cas qui sera développé plus tard où le couple (X, Y) suit une loi normale bidimensionnelle. On a alors :

$$\begin{aligned}\rho(X,Y) &= \mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}[X,Y]}{(\mathrm{Var}[X]\times\mathrm{Var}[Y])^{\frac{1}{2}}} = \frac{\mathbb{E}\left[(X-0)(X^2-1)\right]}{(1\times 2)^{\frac{1}{2}}}\\ &= \frac{1}{\sqrt{2}}\mathbb{E}\left[X^3-X\right] = \frac{1}{\sqrt{2}}\left(\mathbb{E}\left[X^3\right]-\mathbb{E}\left[X\right]\right) = 0-0=0,\end{aligned}$$

car X suit une loi normale centrée donc symétrique par rapport à 0. Pourtant il est clair que les deux variables aléatoires X et $Y = X^2$ ne sont pas indépendantes !

Introduction



Exemple

En bleu la relation entre les deux variables aléatoires X et $Y = X^2$, en rouge 50 points qui sont autant de réalisations du couple (X, X^2). On calculera page 27 la valeur de la statistique de corrélation linéaire à partir de l'échantillon formé des points en rouge.

Propriétés

Il faut donc toujours garder en mémoire que le coefficient de corrélation ne mesure que le **degré d'association linéaire** entre deux variables aléatoires. Nous verrons dans les sections suivantes quels outils utiliser lorsque l'on suspecte que l'association n'est pas nécessairement linéaire.

Lorsque $|\rho(X, Y)| = 1$ alors, avec une probabilité de 1, $Y = aX + b$ pour des constantes réelles a, du même signe que $\rho(X, Y)$, et b.

Introduction

Le coefficient de corrélation est invariant par transformation linéaire. Si $\mathrm{Cor}(X, Y) = \rho(X, Y)$ et que l'on pose :

$$\begin{aligned} X' &= aX + b, \\ Y' &= cY + d, \end{aligned}$$

alors $\mathrm{Cor}(X', Y') = \mathrm{Cor}(X, Y) = \rho(X, Y)$.

Ceci a une conséquence extrêmement importante : on peut aussi bien travailler avec les variables brutes qu'avec les variables centrées réduites pour l'étude de la corrélation.

Sommaire

2 Le cas bidimensionnel
- Loi normale bidimensionnelle
- Estimation des paramètres
- Procédure de test
- Remarques sur les liaisons

Supposons que l'on dispose d'un vecteur aléatoire (X, Y) gaussien.

Nous verrons dans la suite, au moment de tester cette hypothèse, qu'un vecteur aléatoire (X, Y) n'est pas nécéssairement gaussien si X et Y suivent des lois normales [a].

a. Voir l'ouvrage de J.-Y. Ouvrard [5], exercice 13.1 page 276, pour un contre-exemple.

L'hypothèse que l'on fait porte sur le couple et est de ce fait plus délicate à tester que simplement tester la normalité de l'échantillon $\boldsymbol{x}$ issu de X et celle de l'échantillon $\boldsymbol{y}$ issu de Y.

Commençons par rappeler ce que l'on entend par la loi d'un vecteur aléatoire gaussien à deux dimensions.

Définition

Loi normale bidimensionnelle
La densité de la loi normale à deux dimensions est :

$$f(x,y) = K \exp\left[-\frac{1}{2}P(x,y)\right]$$

où :

$$K = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho(X,Y)^2}}, \quad P(x,y) = \frac{1}{1-\rho(X,Y)^2} \times \left[\frac{(x-m_X)^2}{\sigma_X^2} - 2\rho(X,Y)\frac{(x-m_X)(y-m_Y)}{\sigma_X\sigma_Y} + \frac{(y-m_Y)^2}{\sigma_Y^2}\right].$$

Définition

Loi normale bidimensionnelle
Cette loi dépend donc des cinq paramètres : m_X, m_Y, σ_X, σ_Y et $\rho(X,Y)$.

Les quatre premiers sont les moyennes et les écarts types des deux lois marginales, la loi de X et celle de Y, qui sont des lois normales, $X \sim \mathcal{N}(m_X, \sigma_X^2)$ et $Y \sim \mathcal{N}(m_Y, \sigma_Y^2)$.

Les lois liées, c'est-à-dire la probabilité d'obtenir une valeur $Y = y$ sachant que l'on a obtenu $X = x$ ou d'obtenir une valeur $X = x$ sachant que l'on a obtenu une valeur $Y = y$ sont également des lois normales dont les paramètres respectifs $\left(m_{Y|X=x}, \sigma^2_{Y|X=x}\right)$ et $\left(m_{X|Y=y}, \sigma^2_{X|Y=y}\right)$ s'expriment ainsi :

$$\begin{aligned}
m_{Y|X=x} &= m_Y + \rho(X,Y)\frac{\sigma_Y}{\sigma_X}(x - m_X), \\
\sigma^2_{Y|X=x} &= \sigma_Y^2\left(1 - \rho(X,Y)^2\right). \\
m_{X|Y=y} &= m_X + \rho(X,Y)\frac{\sigma_X}{\sigma_Y}(y - m_Y), \\
\sigma^2_{X|Y=y} &= \sigma_X^2\left(1 - \rho(X,Y)^2\right).
\end{aligned}$$

On remarque que :

- La variance de chacune des lois liées est constante : elle ne dépend pas de la valeur de X ou de Y prise en compte pour la calculer, c'est-à-dire du nombre x, qui intervient dans $\sigma_{Y|X=x}$, respectivement y, qui intervient dans $\sigma_{X|Y=y}$, pour lequel elle est calculée.

- La moyenne de chacune des lois liées varie linéairement avec la variable liée. On peut donc parler de deux droites de régression :

$$\begin{aligned} D_{YX}: \quad y - m_Y &= \rho(X,Y)\frac{\sigma_Y}{\sigma_X}(x - m_X), \\ D_{XY}: \quad x - m_X &= \rho(X,Y)\frac{\sigma_X}{\sigma_Y}(y - m_Y). \end{aligned}$$

- $\rho(X, Y)$ est le coefficient de corrélation linéaire :

$$\begin{aligned} \rho(X,Y) &= \frac{\mathrm{Cov}[X,Y]}{(\mathrm{Var}[X] \times \mathrm{Var}[Y])^{\frac{1}{2}}} \\ &= \frac{\mathbb{E}[(X - m_X) \times (Y - m_Y)]}{\left[\mathbb{E}\left[(X - m_X)^2\right] \times \mathbb{E}\left[(Y - m_Y)^2\right]\right]^{\frac{1}{2}}}. \end{aligned}$$

$\rho(X, Y) = 0$ est une condition **nécessaire et suffisante** pour que les deux variables aléatoires X et Y soient indépendantes car ici on a une **hypothèse** de normalité bidimensionnelle **supplémentaire**, voir page 7 pour la situation générale.

Si $\rho(X, Y) = \pm 1$, les deux droites sont confondues et il y a une relation fonctionnelle entre X et Y. Plus précisement, $(X - m_X)$ et $(Y - m_Y)$ sont colinéaires, c'est-à-dire qu'il existe λ et μ deux nombres réels tels que $\lambda(X - m_X) + \mu(Y - m_Y) = 0$. On reconnaît dans la formule précédente l'équation d'une droite. Ainsi si $\rho(X, Y) = \pm 1$, les deux variables X et Y sont liées par une relation fonctionnelle linéaire.

La valeur absolue de $\rho(X, Y)$ est toujours inférieure ou égale à 1. Plus $\rho(X, Y)$ est proche de cette valeur, plus la liaison entre X et Y est serrée. Si l'on considère un échantillon $((x_1, y_1), \ldots, (x_n, y_n))$ de réalisations du couple (X, Y), la dispersion des couples de points (x_i, y_i) autour des droites est d'autant plus faible que les deux droites sont plus proches l'une de l'autre.

- $\rho(X, Y)^2$ est égal au coefficient de détermination R^2 que l'on trouve dans le contexte de la régression linéaire simple.

Estimation des paramètres

On considère un $n-$échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ indépendant et identiquement distribué suivant la loi de (X, Y), qui est une loi normale bidimensionnelle.

Les estimateurs des moyennes m_X et m_Y et des écarts types σ_X et σ_Y que nous allons utiliser sont :

$$\begin{aligned}
\widehat{m_X} &= \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \\
\widehat{m_Y} &= \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \\
\widehat{\sigma_X^2} &= \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2,
\end{aligned}$$

$$\widehat{\sigma^2_Y} = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2,$$

$$\widehat{\sigma_X} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = \sqrt{\widehat{\sigma^2_X}},$$

$$\widehat{\sigma_Y} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = \sqrt{\widehat{\sigma^2_Y}}.$$

Par convention et par soucis d'alléger les notations, on note $\overline{X}$ à la place de $\widehat{m_X}$.
Les quatre premiers estimateurs ci-dessus sont des estimateurs **sans biais** et convergents :

$$\mathbb{E}\left[\widehat{m_X}\right] = m_X \quad \mathrm{Var}\left[\widehat{m_X}\right] = \frac{\sigma^2_X}{n},$$

$$\mathbb{E}\left[\widehat{m_Y}\right] = m_Y \quad \mathrm{Var}\left[\widehat{m_Y}\right] = \frac{\sigma^2_Y}{n},$$

$$\mathbb{E}\left[\widehat{\sigma^2_X}\right] = \sigma^2_X \quad \mathrm{Var}\left[\widehat{\sigma^2_X}\right] = \frac{2\sigma^4_X}{n-1},$$

$$\mathbb{E}\left[\widehat{\sigma^2_Y}\right] = \sigma^2_Y \quad \mathrm{Var}\left[\widehat{\sigma^2_Y}\right] = \frac{2\sigma^4_Y}{n-1}.$$

Les deux derniers sont **biaisés** et convergents :

$$\mathbb{E}\left[\widehat{\sigma_X}\right] = \sqrt{\frac{2}{n-1}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\sigma_X = \left(1 - \frac{3}{4n} + o\left(\frac{1}{n}\right)\right)\sigma_X,$$

$$\mathrm{Var}\left[\widehat{\sigma_X}\right] = \frac{1}{n-1}\left[n - 1 - \frac{2\Gamma\left(\frac{n}{2}\right)^2}{\Gamma\left(\frac{n-1}{2}\right)^2}\right]\sigma^2_X,$$

$$\mathbb{E}\left[\widehat{\sigma_Y}\right] = \sqrt{\frac{2}{n-1}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\sigma_Y = \left(1-\frac{3}{4n}+o\left(\frac{1}{n}\right)\right)\sigma_Y,$$

$$\mathrm{Var}\left[\widehat{\sigma_Y}\right] = \frac{1}{n-1}\left[n-1-\frac{2\Gamma\left(\frac{n}{2}\right)^2}{\Gamma\left(\frac{n-1}{2}\right)^2}\right]\sigma_Y^2,$$

où $o(1/n)$ est une fonction qui tend vers 0 plus vite que $1/n$ et $\Gamma(x)$ est une fonction mathématique définie à l'aide d'une intégrale !

Il serait possible de corriger le biais de l'estimateur $\widehat{\sigma_X}$, en posant : $\widehat{\sigma_X}_{\text{corrigé}} = \sqrt{\frac{n-1}{2}}\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\widehat{\sigma_X}$.

Néanmoins l'obtention d'un intervalle de confiance reste délicate puisque la loi de cet estimateur n'est pas classique.

P. Chapouille, dans son livre [2], propose la formule approximative :

$$\widehat{\sigma_X}_{\text{corrigé}} \approx \widehat{\sigma_X}\sqrt{\frac{2n-2}{2n-3}} \approx \widehat{\sigma_X}\left(1+\frac{1}{4n-6}\right).$$

La situation est beaucoup plus compliquée que pour les estimateurs ci-dessus.

On comprend ainsi pourquoi on préfère estimer la variance d'une loi normale à la place de son écart type.
« Heureusement », les deux paramètres qui ont été choisis pour définir une loi normale sont sa moyenne et sa variance.

Au passage on remarque que la racine carrée d'un estimateur sans biais n'est pas nécessairement sans biais.
Le coefficient de corrélation linéaire $\rho(X, Y)$ est un rapport.
Pour estimer ce rapport nous allons estimer le numérateur $\mathrm{Cov}\,[X, Y]$ et le dénominateur $(\mathrm{Var}\,[X] \times \mathrm{Var}\,[Y])^{\frac{1}{2}}$.

En ce qui concerne le dénominateur, nous verrons qu'il suffit d'utiliser les estimateurs $\widehat{\sigma_X}$ et $\widehat{\sigma_Y}$ ci-dessus. Un estimateur du numérateur peut être défini par la quantité :

$$\begin{aligned}\widehat{\mathrm{Cov}\,[X, Y]} &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})\,(Y_i - \overline{Y})\\ &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i \times Y_i) - \frac{n}{n-1}\overline{X} \times \overline{Y}.\end{aligned}$$

C'est un estimateur sans biais et convergent de $\mathrm{Cov}\,[X, Y]$:

$$\mathbb{E}\left[\widehat{\mathrm{Cov}\,[X, Y]}\right] = \mathrm{Cov}\,[X, Y]\,.$$

Notez la similitude avec l'estimateur corrigé de la variance. Ici aussi apparaît au dénominateur le nombre $n-1$.

L'estimateur du coefficient de corrélation $\rho(X, Y)$, que nous allons utiliser, est $\widehat{\rho(X, Y)}$ défini par :

$$\begin{aligned}\widehat{\rho(X, Y)} &= \frac{\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(X_i - \overline{X}) \times (Y_i - \overline{Y})}{\left(\widehat{\sigma_X^2} \times \widehat{\sigma_Y^2}\right)^{\frac{1}{2}}}\\ &= \frac{\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(X_i - \overline{X}) \times (Y_i - \overline{Y})}{\widehat{\sigma_X} \times \widehat{\sigma_Y}}\end{aligned}$$

$$= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})\times(Y_i-\overline{Y})}{\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2\times\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\overline{Y})^2\right)^{\frac{1}{2}}}$$

$$= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})\times(Y_i-\overline{Y})}{\frac{1}{n-1}\left(\sum_{i=1}^{n}(X_i-\overline{X})^2\times\sum_{i=1}^{n}(Y_i-\overline{Y})^2\right)^{\frac{1}{2}}}$$

$$= \frac{\sum_{i=1}^{n}(X_i-\overline{X})\times(Y_i-\overline{Y})}{\left(\sum_{i=1}^{n}(X_i-\overline{X})^2\times\sum_{i=1}^{n}(Y_i-\overline{Y})^2\right)^{\frac{1}{2}}}$$

$$= \frac{\sum_{i=1}^{n}(X_i\times Y_i)-\overline{X}\times\overline{Y}}{\left(\sum_{i=1}^{n}X_i^2-\overline{X}^2\right)^{\frac{1}{2}}\left(\sum_{i=1}^{n}Y_i^2-\overline{Y}^2\right)^{\frac{1}{2}}}.$$

L'étude des propriétés de cet estimateur est plus complexe puisqu'il s'agit d'un rapport de variables aléatoires qui ne sont pas indépendantes. Il s'agit d'un estimateur **biaisé** et convergent de $\rho(X, Y)$:

$$\begin{aligned}\mathbb{E}\left[\widehat{\rho(X,Y)}\right] &= \rho(X,Y) - \frac{\rho(X,Y)(1-\rho(X,Y)^2)}{2n}\\ \mathrm{Var}\left[\widehat{\rho(X,Y)}\right] &= \frac{(1-\rho(X,Y)^2)^2}{n} + o\left(\frac{1}{n}\right),\end{aligned}$$

où $o(1/n)$ est une fonction qui tend vers 0 lorsque n tend vers $+\infty$ plus vite que $1/n$.

La distribution asymptotique de $\rho(X, Y)$ est :

$$\sqrt{n}(\widehat{\rho(X,Y)} - \rho(X,Y)) \approx \mathcal{N}\left(0, \left(1-\rho(X,Y)^2\right)^2\right).$$

Ce résultat n'est pas assez précis pour permettre de résoudre les problèmes qui nous intéressent généralement. On l'utilisera uniquement si l'on ne peut pas faire autrement et si $n \geqslant 30$.

Ainsi lorsque l'on dispose d'un échantillon $(x_1, y_1), \ldots, (x_n, y_n)$ de réalisations du couple (X, Y) on obtient les estimations des paramètres de la loi normale bidimensionnelle suivie par (X, Y) en utilisant les formules ci-dessous qui sont les réalisations des estimateurs ci-dessus sur notre échantillon. On note $\boldsymbol{x}$ l'échantillon $(x_1, \ldots, x_n)$, $\boldsymbol{y}$ l'échantillon $(y_1, \ldots, y_n)$ et $(\boldsymbol{x}, \boldsymbol{y})$ l'échantillon $(x_1, y_1), \ldots, (x_n, y_n)$.

Les estimations des moyennes m_X et m_Y et des écarts types σ_X et σ_Y sont :

$$\begin{aligned}\widehat{m_X}(\boldsymbol{x}) &= \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,\\ \widehat{m_Y}(\boldsymbol{y}) &= \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,\\ \widehat{\sigma_X^2}(\boldsymbol{x}) &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2,\end{aligned}$$

$$\begin{aligned}
\widehat{\sigma_Y^2}(\boldsymbol{y}) &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2,\\
\widehat{\sigma_X}(\boldsymbol{x}) &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2},\\
\widehat{\sigma_Y}(\boldsymbol{y}) &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2}.
\end{aligned}$$

On note $\overline{x}$ bien qu'il s'agisse de la moyenne de l'échantillon $\boldsymbol{x}$ et que de ce fait on pourrait également écrire $\overline{\boldsymbol{x}}$. Il s'agit là aussi d'une convention mais si l'on voulait être cohérent avec les notations utilisées pour les autres estimateurs on devrait écrire $\widehat{m_X}(\boldsymbol{x})$.

L'estimation de la covariance $\mathrm{Cov}\,[X,Y]$ est donnée par :

$$\widehat{\mathrm{Cov}\,[X,Y]}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})\times(y_i-\overline{y}).$$

L'estimation du coefficient de corrélation $\rho(X,Y)$ est donnée par :

$$\begin{aligned}
\widehat{\rho(X,Y)}(\boldsymbol{x},\boldsymbol{y}) &= \frac{\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(x_i-\overline{x})\times(y_i-\overline{y})}{\left(\widehat{\sigma_X^2}(\boldsymbol{x})\times\widehat{\sigma_Y^2}(\boldsymbol{y})\right)^{\frac{1}{2}}}\\
&= \frac{\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(x_i-\overline{x})\times(y_i-\overline{y})}{\widehat{\sigma_X}(\boldsymbol{x})\times\widehat{\sigma_Y}(\boldsymbol{y})}
\end{aligned}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \times (y_i - \overline{y})}{\left(\sum_{i=1}^{n}(x_i - \overline{x})^2 \times \sum_{i=1}^{n}(y_i - \overline{y})^2\right)^{\frac{1}{2}}}$$

$$= \frac{\sum_{i=1}^{n}(x_i \times y_i) - \overline{x} \times \overline{y}}{\left(\sum_{i=1}^{n} x_i^2 - \overline{x}^2\right)^{\frac{1}{2}} \times \left(\sum_{i=1}^{n} y_i^2 - \overline{y}^2\right)^{\frac{1}{2}}}.$$

Exemple

On évalue la formule ci-dessus pour l'échantillon représenté en rouge page 7 :

```
Pearson correlation of Y and X = 0,006
```


Sous l'hypothèse nulle $\mathcal{H}_0$: « X et Y sont indépendantes », la variable :

$$\sqrt{n-2}\,\frac{\widehat{\rho(X,Y)}}{\sqrt{1-\widehat{\rho(X,Y)}^2}}$$

suit une loi de Student T_{n-2} à $n-2$ degrés de liberté.

Ceci permet de réaliser le test suivant :

$$\mathcal{H}_0 : \rho(X, Y) = 0$$

contre

$$\mathcal{H}_1 : \rho(X, Y) \neq 0.$$

Pour prendre une décision associée à un test de niveau $(1 - \alpha)$ %, il suffit donc de trouver la valeur critique d'une loi de Student à $n - 2$ degrés de liberté pour $\alpha/2$.

Il est également possible, en réalisant un test unilatéral, de tester les hypothèses suivantes :

$$\mathcal{H}_0 : \rho(X, Y) = 0$$

contre

$$\mathcal{H}_1 : \rho(X, Y) > 0.$$

ou encore

$$\mathcal{H}_0 : \rho(X, Y) = 0$$

contre

$$\mathcal{H}_1 : \rho(X, Y) < 0.$$

Notons que, si $\rho_0 \neq 0$, la propriété ci-dessus ne permet pas de tester des hypothèses du type :

$$\mathcal{H}_0 : \rho(X, Y) = \rho_0$$

contre

$$\mathcal{H}_1 : \rho(X, Y) \neq \rho_0.$$

Pour procéder à ces types de test, on utilise la transformation de Fisher

Il existe d'autres transformations pour obtenir des intervalles de confiance pour $\rho(X,Y)$, par exemple l'approximation de Ruben qui semble être plus précise mais encore plus complexe que celle de Fisher.

On peut par exemple montrer que, quelque soit la valeur du coefficient de corrélation $\rho(X,Y)$, la variable aléatoire $(R-\rho(X,Y))\sqrt{(n-2)/\left[(1-R^2)(1-\rho(X,Y)^2)\right]}$ suit approximativement une loi $\mathcal{T}_{n-2}$ de Student à $n-2$ degrés de liberté. On retrouve le résultat de la page 56 établi lorsque $\rho(X,Y)=0$. Pour plus de détails voir [1] et [4].

En effet, Fisher, le premier, a démontré que si l'on introduit les deux variables :

$$\begin{aligned} Z(X,Y) &= \frac{1}{2}\ln\left(\frac{1+\rho(X,Y)}{1-\rho(X,Y)}\right) \\ &= \tanh^{-1}\left(\rho(X,Y)\right), \end{aligned}$$

$$\begin{aligned} \widehat{Z(X,Y)} &= \frac{1}{2}\ln\left(\frac{1+\widehat{\rho(X,Y)}}{1-\widehat{\rho(X,Y)}}\right) \\ &= \tanh^{-1}\left(\widehat{\rho(X,Y)}\right), \end{aligned}$$

alors la différence $\widehat{Z(X,Y)}-Z(X,Y)$ suit une loi voisine de la loi normale de moyenne $\rho(X,Y)/(2(n-1))$ et de variance proche de $1/(n-3)$.

On approximera donc dans le cas général la loi de cette différence par une loi $\mathcal{N}(\rho(X,Y)/(2(n-1)), 1/(n-3))$ et sous une hypothèse d'indépendance entre X et Y, qui implique donc $\rho(X,Y)=0$, par une loi $\mathcal{N}(0, 1/(n-3))$.

Pour passer des résultats obtenus pour $Z(X,Y)$ à des résultats concernant $\rho(X,Y)$, on utilise la transformation inverse de celle de Fisher :

$$\rho(X,Y) = \frac{\exp\left(2Z(X,Y)\right)-1}{\exp\left(2Z(X,Y)\right)+1}$$
$$\rho(X,Y) = \tanh(Z(X,Y)).$$

La plupart des logiciels de statistique permettent de calculer les estimations par intervalles de $\rho(X,Y)$. Certains proposent même des intervalles de confiance exacts.

- L'existence de deux droites de régression dans les études de corrélation est embarassante car elle amène à la question suivante : laquelle des deux droites représente-t-elle le mieux la réalité ? On ne peut trancher sans utiliser la notion de causalité : si l'on pense que X est la cause de Y, on retiendra la droite D_{YX} qui est la régression de la variable dépendante par rapport à la variable cause.

Dans les cas où l'on ne peut établir une relation de cause à effet, il n'y a pas lieu de fixer son attention sur l'une des droites de régression plutôt que sur l'autre. Le coefficient de corrélation linéaire $\rho(X,Y)$ est alors le paramètre le plus intéressant ; quelque soit sa valeur, positive ou négative, il ne préjuge en rien d'un quelconque rapport de cause à effet entre les variables étudiées.

Dans tous les cas, l'interprétation de toute étude de liaison mérite beaucoup de réflexions, et seules les raisons physiques ou biologiques et non statistiques pourront permettre de porter des jugements de cause à effet.

- Lorsque l'on trouve une courbe de régression qui serre de près le phénomène étudié, il faut se garder de conclure que cette formule traduit la loi exacte qui gouverne le phénomène.

- Enfin signalons que souvent deux variables X et Y sont fortement corrélées avec une troisième Z, le temps par exemple, et on peut conclure à une corrélation significative entre X et Y, alors qu'à priori il n'y a aucune relation entre ces deux grandeurs si ce n'est leur liaison avec Z. On sent alors la nécessité d'introduire une mesure de liaison qui permettra de connaître l'association de X et Y en éliminant l'influence de Z : il s'agit de la notion de **corrélation partielle** qui sera étudiée aux paragraphes vignette 128, **??** et **??**.

Sommaire

3 Le cas général ($n \geqslant 2$)
- Loi multinormale
- Estimation
- Test de l'hypothèse $\rho_{ij} = 0$
- Test de l'hypothèse $\rho_{ij} = \rho_0$, $\rho_0 \neq 0$

On considère un vecteur gaussien de dimension $p \geqslant 1$: $\boldsymbol{X} = (X_1, \ldots, X_p)$.
Un tel vecteur est entièrement déterminé par la connaissance de

- sa moyenne $\boldsymbol{\mu}(\boldsymbol{X})$
- et de sa matrice de variance-covariance $\boldsymbol{\Sigma}(\boldsymbol{X})$.

Il s'agit d'une extension de ce que nous venons de voir pour le cas bidimensionnel.

En termes plus clairs, il suffit de connaître :

$$\begin{aligned}
\boldsymbol{\mu}(\boldsymbol{X}) &= (\mu_1, \ldots, \mu_p) = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_p]), \\
\boldsymbol{\Sigma}(\boldsymbol{X}) &= \left(\left(\sigma_{i,j} = \mathrm{Cov}\left[X_i, X_j\right]\right)\right)_{1 \leqslant i,j \leqslant p} \\
&= \begin{pmatrix} \mathrm{Var}\left[X_1\right] & \ldots & \mathrm{Cov}\left[X_1, X_p\right] \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}\left[X_p, X_1\right] & \cdots & \mathrm{Var}\left[X_p\right] \end{pmatrix}.
\end{aligned}$$

On introduit alors la matrice des corrélations $\rho(\boldsymbol{X})$ entre les composantes X_i, $1 \leqslant i \leqslant p$ de $\boldsymbol{X}$:

$$\begin{aligned}\rho(\boldsymbol{X}) &= \left(\left(\rho_{i,j} = \mathrm{Cor}\left[X_i, X_j\right]\right)\right)_{1 \leqslant i,j \leqslant p} \\ &= \begin{pmatrix} 1 & \cdots & \dfrac{\mathrm{Cov}\left[X_1, X_p\right]}{\left(\mathrm{Var}\left[X_1\right]\mathrm{Var}\left[X_p\right]\right)^{\frac{1}{2}}} \\ \vdots & \ddots & \vdots \\ \dfrac{\mathrm{Cov}\left[X_p, X_1\right]}{\left(\mathrm{Var}\left[X_p\right]\mathrm{Var}\left[X_1\right]\right)^{\frac{1}{2}}} & \cdots & 1 \end{pmatrix}.\end{aligned}$$

Afin de construire un estimateur simple de la matrice $\rho(\boldsymbol{X})$ on doit disposer d'un n−échantillon $(\boldsymbol{X_1}, \ldots, \boldsymbol{X_n})$ **indépendant et identiquement** distribué du vecteur $\boldsymbol{X}$.

Ce n−échantillon est donc composé de n éléments, qui sont des vecteurs de taille p, que l'on note $(X_{1,1}, \ldots, X_{p,1}), \ldots, (X_{1,n}, \ldots, X_{p,n})$. $X_{i,j}$ est la variable aléatoire associée à la valeur prise par la i−ème composante X_i lors de la réalisation de la j−ème expérience. Par exemple $X_{1,3}$ serait la variable aléatoire associée à la valeur prise par X_1 lors de la troisième expérience et $X_{5,2}$ serait la variable aléatoire associée à la valeur prise par X_5 lors de la seconde expérience.

Un estimateur $\widehat{\boldsymbol{\mu}}$ de la moyenne $\boldsymbol{\mu}$ est :

$$\begin{aligned}\widehat{\boldsymbol{\mu}} &= \frac{\boldsymbol{X_1} + \cdots + \boldsymbol{X_n}}{n} \\ &= \frac{\sum_{i=1}^{n} \boldsymbol{X_i}}{n}.\end{aligned}$$

On note, comme d'habitude, $\widehat{\boldsymbol{\mu}}$ par $\overline{\boldsymbol{x}}$. Il s'agit d'un estimateur sans biais et convergent, comme à la page 27 :

$$\begin{aligned}
\mathbb{E}\left[\widehat{\boldsymbol{\mu}}\right] &= \mathbb{E}\left[\frac{\boldsymbol{X_1}+\cdots+\boldsymbol{X_n}}{n}\right] = \frac{\mathbb{E}\left[\sum_{i=1}^{n}\boldsymbol{X_i}\right]}{n}\\
&= \frac{\sum_{i=1}^{n}\mathbb{E}\left[\boldsymbol{X_i}\right]}{n} = \frac{\sum_{i=1}^{n}\boldsymbol{\mu}}{n} = \boldsymbol{\mu}.\\
\mathrm{Var}\left[\widehat{\boldsymbol{\mu}}\right] &= \frac{\boldsymbol{\Sigma}}{n}.
\end{aligned}$$

Un estimateur $\widehat{\rho(\boldsymbol{X})}$ de la matrice des corrélations $\rho(\boldsymbol{X})$ est alors donné par :

$$\widehat{\rho(\boldsymbol{X})} = \left(\left(\widehat{\rho}_{i,j} = \widehat{\rho_{i,j}} = \widehat{\mathrm{Cor}}\left[X_i, X_j\right]\right)\right)_{1\leqslant i,j\leqslant p}$$

$$= \begin{pmatrix}
1 & \cdots & \dfrac{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{1,i}-\overline{x}_1\right)\times\left(x_{p,i}-\overline{x}_p\right)}{\widehat{\sigma_1}\times\widehat{\sigma_p}}\\
\vdots & \ddots & \vdots\\
\dfrac{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{p,i}-\overline{x}_p\right)\times\left(x_{1,i}-\overline{x}_1\right)}{\widehat{\sigma_p}\times\widehat{\sigma_1}} & \cdots & 1
\end{pmatrix}.$$

On obtient alors la forme explicite suivante :

$$\begin{pmatrix}
1 & \cdots & \dfrac{\sum_{i=1}^{n}\left(x_{1,i}-\overline{x}_1\right)\times\left(x_{p,i}-\overline{x}_p\right)}{\left(\sum_{i=1}^{n}\left(x_{1,i}-\overline{x}_1\right)^2\times\sum_{i=1}^{n}\left(x_{i,p}-\overline{x}_p\right)^2\right)^{\frac{1}{2}}}\\
\vdots & \ddots & \vdots\\
\dfrac{\sum_{i=1}^{n}\left(x_{p,i}-\overline{x}_p\right)\times\left(x_{1,i}-\overline{x}_1\right)}{\left(\sum_{i=1}^{n}\left(x_{p,i}-\overline{x}_p\right)^2\times\sum_{i=1}^{n}\left(x_{1,i}-\overline{x}_1\right)^2\right)^{\frac{1}{2}}} & \cdots & 1
\end{pmatrix}.$$

La distribution asymptotique de ρ_{ij} est :

$$\sqrt{n}(\widehat{\rho_{ij}} - \rho_{ij}) \approx \mathcal{N}\left(0, \left(1 - \rho_{ij}^2\right)^2\right).$$

Ce résultat n'est pas assez précis pour permettre de résoudre les problèmes qui nous intéressent généralement. On l'utilisera uniquement si l'on ne peut pas faire autrement et si $n \geqslant 30$.

On a donc une estimation $\widehat{\rho(\boldsymbol{X})}(\boldsymbol{x})$ de $\rho(\boldsymbol{X})$ à l'aide d'un échantillon $\boldsymbol{x}$ de n réalisations de $\boldsymbol{X}$. En notant $\boldsymbol{x_1}$ l'échantillon formé par les n réalisations, $(x_{1,1}, \ldots, x_{1,n})$, de $X_1, \ldots, \boldsymbol{x_p}$ l'échantillon formé par les n réalisations, $(x_{p,1}, \ldots, x_{p,n})$, de X_p on a :

$$\widehat{\rho(\boldsymbol{X})}(\boldsymbol{x}) = \left(\left(\widehat{\rho(\boldsymbol{X})}(\boldsymbol{x})_{i,j} = \widehat{\rho_{i,j}}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \widehat{\mathrm{Cor}\,[X_i, X_j]}(\boldsymbol{x_i}, \boldsymbol{x_j})\right)\right)_{1 \leqslant i,j \leqslant p}$$

$$= \begin{pmatrix} 1 & \cdots & \dfrac{\frac{1}{n-1}\sum_{i=1}^{n}(x_{1,i} - \overline{x_1}) \times \left(x_{p,i} - \overline{x_p}\right)}{\widehat{\sigma_1}(\boldsymbol{x_1}) \times \widehat{\sigma_p}(\boldsymbol{x_p})} \\ \vdots & \ddots & \vdots \\ \dfrac{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{p,i} - \overline{x_p}\right) \times (x_{1,i} - \overline{x_1})}{\widehat{\sigma_p}(\boldsymbol{x_p}) \times \widehat{\sigma_1}(\boldsymbol{x_1})} & \cdots & 1 \end{pmatrix}.$$

L'estimation $\widehat{\rho}(\boldsymbol{X})(\boldsymbol{x})$ de $\rho(\boldsymbol{X})$ est alors :

$$\begin{pmatrix} 1 & \cdots & \dfrac{\sum_{i=1}^{n}(x_{1,i} - \overline{x}_1) \times \left(x_{p,i} - \overline{x_p}\right)}{\left(\sum_{i=1}^{n}(x_{1,i} - \overline{x}_1)^2 \times \sum_{i=1}^{n}\left(x_{i,p} - \overline{x}_p\right)^2\right)^{\frac{1}{2}}} \\ \vdots & \ddots & \vdots \\ \dfrac{\sum_{i=1}^{n}\left(x_{p,i} - \overline{x_p}\right) \times (x_{1,i} - \overline{x_1})}{\left(\sum_{i=1}^{n}\left(x_{p,i} - \overline{x}_p\right)^2 \times \sum_{i=1}^{n}(x_{i,1} - \overline{x}_1)^2\right)^{\frac{1}{2}}} & \cdots & 1 \end{pmatrix}.$$

On peut maintenant s'intéresser au test d'indépendance des composantes d'un vecteur gaussien.

$$\mathcal{H}_0 : \rho_{ij} = 0$$

contre

$$\mathcal{H}_1 : \rho_{ij} \neq 0.$$

Sous l'hypothèse nulle $\mathcal{H}_0$, alors :

$$t_{ij} = \sqrt{n-2}\frac{\widehat{\rho}_{ij}(\boldsymbol{x_i}, \boldsymbol{x_j})}{\sqrt{1 - \widehat{\rho}_{ij}(\boldsymbol{x_i}, \boldsymbol{x_j})^2}}$$

est la réalisation d'une variable aléatoire T qui suit une loi de Student à $n-2$ degrés de liberté, i.e. $T \sim \mathcal{T}_{n-2}$.
On rejette l'hypothèse nulle $\mathcal{H}_0$ au niveau $(1-\alpha)$ % lorsque $|t_{ij}| > \mathcal{T}_{n-2,\alpha/2}$.

Si $\rho_0 \neq 0$, la propriété de la page 85 ci-dessus ne permet pas de tester des hypothèses du type :

$$\mathcal{H}_0 : \rho_{ij} = \rho_0$$

contre

$$\mathcal{H}_1 : \rho_{ij} \neq \rho_0.$$

Il existe alors deux possibilités :

- Si l'on souhaite faire un test de ce type à la main, on utilise la transformation de Fisher introduite ci-dessus, à la page 56, pour obtenir un intervalle de confiance pour ρ_{ij} de niveau approximatif $(1-\alpha)$ % :

$$\left[\tanh\left(\widehat{z} - \frac{z_{\alpha/2}}{(n-3)^{\frac{1}{2}}}\right), \tanh\left(\widehat{z} + \frac{z_{\alpha/2}}{(n-3)^{\frac{1}{2}}}\right)\right]$$

où $\widehat{z} = \tanh\left(\widehat{\rho}_{ij}(\boldsymbol{x_i}, \boldsymbol{x_j})\right)$ et $z_{\alpha/2}$ est le quantile de la loi $\mathcal{N}(0,1)$.

- Certains logiciels, comme SPSS, R , SAS ou StatXact, proposent des versions *exactes* de l'intervalle de confiance dont une approximation est donnée ci-dessus. En effet la distribution exacte de ρ_{ij} n'est pas une distribution de probabilité classique.

La densité du coefficient de corrélation $\rho(X, Y)$ d'un couple gaussien (X, Y) au point $-1 \leqslant r \leqslant 1$ est

$$\frac{1}{\pi}(n-2)(1-r^2)^{(n-4)/2}(1-\rho^2)^{(n-1)/2}\int_0^{+\infty}\frac{\mathrm{d}\beta}{(\cosh\beta-\rho r)^{n-1}},$$

voir [6] pour plus de détails.

Si historiquement, les approximations présentées ci-dessus étaient les seules possibilités auxquelles l'expérimentateur pouvaient recourir pour obtenir des intervalles de confiance pour ρ_{ij}, il est devenu possible avec l'augmentation des capacités de calcul dont vous disposez désormais de déterminer les intervalles de confiance exacts pour ρ_{ij}. Ceci présente l'avantage de ne pas avoir à utiliser des résultats asymptotiques qui peuvent s'avérer incorrects si la taille de l'échantillon est petite.

Exemple

Considérons deux échantillons $\boldsymbol{x_1}$ et $\boldsymbol{x_2}$.

$\boldsymbol{x_1}$	$\boldsymbol{x_2}$
0,19577	-0,81685
0,92204	0,58257
-0,04690	1,24359
0,45863	0,31088
-0,58454	0,95124
1,51722	0,02028
-0,97862	-0,91823
-0,04557	-0,18670
-0,03707	-0,96033
0,84505	1,11031

Exemple

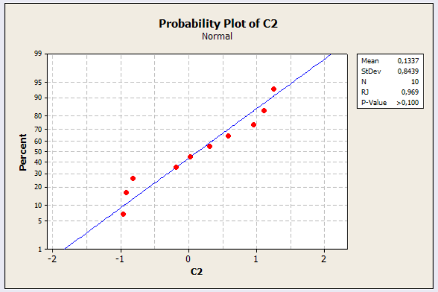
Procédons au test de l'hypothèse $\boldsymbol{x_1}$ est issu d'une variable aléatoire X_1 qui suit une loi normale et au test de l'hypothèse $\boldsymbol{x_2}$ est issu d'une variable aléatoire X_2 qui suit une loi normale.
Compte tenu des effectifs on procède au test de **Shapiro-Wilk** ou plutôt de sa variante disponible dans Minitab : le test de **Ryan-Joiner**.
On commence par tester la normalité du couple (X_1, X_2).

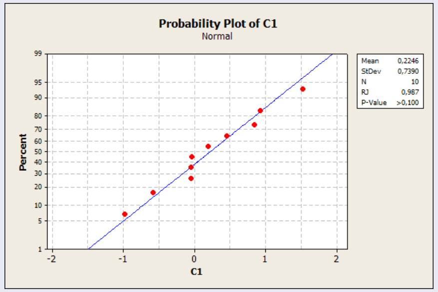
$$\mathcal{H}_0 : (X_1, X_2) \text{ suit une loi normale bidimensionnelle}$$

contre

$$\mathcal{H}_1 : (X_1, X_2) \text{ ne suit pas une loi bidimensionnelle.}$$

Exemple

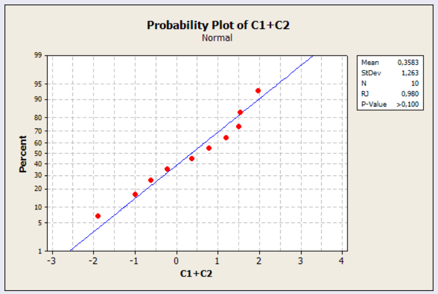


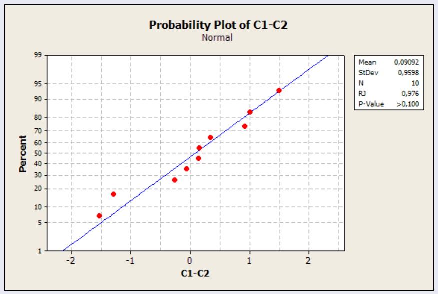


Exemple

Les $p-$valeurs sont toutes les deux supérieures à $0,100$: on ne peut donc pas rejeter l'hypothèse de normalité pour X_1 et X_2. Notons que l'on a uniquement vérifié la normalité de X_1 et celle de X_2. Ceci n'implique pas celle du couple (X_1, X_2). Il s'agit néanmoins d'une **condition nécessaire**. Nous verrons à la page **??** comment vérifier l'hypothèse de normalité du couple (X_1, X_2). Une des premières choses à faire est de tester la normalité de $X_1 + X_2$ et de $X_1 - X_2$, voir [3].

Exemple





Exemple

Les $p-$valeurs sont supérieures à $0,100$, on ne peut donc pas rejeter l'hypothèse de normalité de $X_1 + X_2$ et de $X_1 - X_2$. Attention, nous testions ici la multinormalité de (X_1, X_2) et nous avons procédé à quatre tests de Shapiro-Wilk.

Nous avions, comme d'habitude, fixé un seuil de $\alpha_{global} = 5$ % pour le test global de multinormalité. Pour garantir ce risque global de 5 %, il faut fixer un risque différent de 5 % pour chaque test de Shapiro-Wilk, vous avez déjà rencontré ce problème dans le contexte des comparaisons multiples en analyse de la variance.

Exemple

En effet si le risque pour chacun des tests de Shapiro-Wilk est fixé à $\alpha_{ind} = 5$ %, alors le risque global est au maximum de $\alpha_{global} = 1 - (1 - 0,05)^4 = 0,185 > 0,05$.

Ainsi on doit fixer un seuil de $\alpha_{ind} = 1 - \sqrt[4]{1 - \alpha_{global}} = 1 - \sqrt[4]{0,95} = 0,013$ pour chacun des tests de Shapiro-Wilk, ce qui revient à être moins exigeant pour chacun des tests individuels puisque $0,013 < 0,05$ et donc à accepter la normalité dans plus de cas.

Exemple

Dans cet exemple cette correction ne change rien à la conclusion puisqu'aucune des $p-$valeurs n'étaient comprises entre $0,013$ et $0,05$.
Ainsi on ne peut pas rejeter l'hypothèse nulle $\mathcal{H}_0$ de normalité du couple (X_1, X_2).
Testons maintenant l'hypothèse d'indépendance de X_1 et X_2.

Exemple

Puisque le couple (X_1, X_2) suit une loi normale, l'indépendance de X_1 et X_2 est équivalente à l'absence de corrélation linéaire entre X_1 et X_2. Il suffit donc de tester :

$$\mathcal{H}_0 : \rho(X_1, X_2) = 0$$

contre

$$\mathcal{H}_1 : \rho(X_1, X_2) \neq 0.$$

Exemple

```
Correlations: X1; X2

Pearson correlation of X1 and X2 = 0,270
P-Value = 0,450
```

Exemple

Avec SPSS 13.0 on obtient :

```
Correlations
                              X1     X2
X1    Pearson Correlation     1      ,270
      Sig. (2-tailed)                ,450
      N                       10     10
X2    Pearson Correlation     ,270   1
      Sig. (2-tailed)         ,450
      N                       10     10
```


Exemple

En utilisant le module Tests Exacts de SPSS 13.0 :

```
                     Value    Approx. Sig.    Exact Sig.
Pearson's R          0,270    0,450           0,446
N of Valid Cases     10
```

Notons que dans notre cas la significativité exacte est proche de la significativité approchée. Néanmoins, dans de nombreux cas il n'en va pas de même.

Exemple

Ainsi avec une $p-$valeur de $0,446$, le test n'est pas significatif au seuil $\alpha = 5$ %. On ne peut rejeter l'hypothèse nulle $\mathcal{H}_0$ d'absence de corrélation entre X_1 et X_2. On en déduit que l'on ne peut rejeter l'hypothèse d'indépendance de X_1 et X_2.

Sommaire

4 Corrélation multiple
- Définition
- Estimation
- Asymptotique
- Test de l'hypothèse $R(X_1, \mathbf{X_2}) = 0$
- Test de l'hypothèse $R(X_1, \mathbf{X_2}) = R_0$, $R_0 \neq 0$

Le coefficient de corrélation multiple R est la corrélation maximale possible entre une variable réelle X_1 et toutes les combinaisons linéaires de composantes d'un vecteur aléatoire $\boldsymbol{X_2}$.

Rappelons rapidement ce que l'on entend par **combinaison linéaire**.

Si $\boldsymbol{U}$ et $\boldsymbol{V}$ sont deux vecteurs de $\mathbb{R}^k$, $k \geqslant 1$, une combinaison linéaire de $\boldsymbol{U}$ et $\boldsymbol{V}$ est un vecteur $\mathrm{CL}_{(\boldsymbol{U},\boldsymbol{V})}(\alpha, \beta)$ défini comme la somme $\alpha\boldsymbol{U} + \beta\boldsymbol{V}$ avec $(\alpha, \beta) \in \mathbb{R}^2$.

Cette notion s'étend au cas de n vecteurs : si $\boldsymbol{U_1}, \ldots, \boldsymbol{U_n}$ sont n vecteurs de $\mathbb{R}^k$, $k \geqslant 1$, une combinaison linéaire des $(\boldsymbol{U_i})_{1 \leqslant i \leqslant n}$ est un vecteur $\mathrm{CL}_{(\boldsymbol{U_1},\ldots,\boldsymbol{U_n})}(\alpha_1, \ldots, \alpha_n)$ défini comme la somme $\sum_{i=1}^{n} \alpha_i \boldsymbol{U_i}$ avec $(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$.

Mise en équation : on dispose de $X_1 \in \mathbb{R}$ et $\boldsymbol{X_2} \in \mathbb{R}^{p-1}$ tels que $(X_1, \boldsymbol{X_2}) \in \mathbb{R}^p$ ait une distribution dans $\mathbb{R}^p$ de moyenne $\boldsymbol{\mu}$ et de matrice de variance-covariance $\boldsymbol{\Sigma}$.
La moyenne $\boldsymbol{\mu}$ du couple $(X_1, \boldsymbol{X_2})$ est reliée à la moyenne μ_1 de X_1 et $\boldsymbol{\mu_2}$ de $\boldsymbol{X_2}$ par :

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \boldsymbol{\mu_2} \end{pmatrix}.$$

On adopte une écriture particulière pour $\boldsymbol{\Sigma}$ pour faire ressortir la variance de X_1, notée $\sigma_{11}(X_1)$, la matrice de variance-covariance de $\boldsymbol{X_2}$, notée $\boldsymbol{\Sigma_{22}}(\boldsymbol{X_2})$, et les covariances des composantes de $\boldsymbol{X_2}$ avec X_1, qui sont les composantes du vecteur $\boldsymbol{\Sigma_{21}}(X_1, \boldsymbol{X_2})$.

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11}(X_1) & \boldsymbol{\sigma_{21}}(X_1, \boldsymbol{X_2})^T \\ \boldsymbol{\Sigma_{21}}(X_1, \boldsymbol{X_2}) & \boldsymbol{\Sigma_{22}}(\boldsymbol{X_2}) \end{pmatrix}$$

où $\boldsymbol{\sigma_{21}}(X_1, \boldsymbol{X_2})^T$ est la transposée de $\boldsymbol{\sigma_{21}}(X_1, \boldsymbol{X_2})$. Remarquons que $\mathrm{Var}[X_1] = \sigma_{11}(X_1) = \mathrm{Cov}[X_1, X_1] = \sigma^2_{X_1}$ avec les notations de la page 27.

On montre par un peu de calcul matriciel que la corrélation maximale R vérifie :

$$R^2(X_1, \boldsymbol{X_2}) = \frac{\boldsymbol{\sigma_{21}}(X_1, \boldsymbol{X_2})^T \boldsymbol{\Sigma_{22}}(\boldsymbol{X_2})^{-1} \boldsymbol{\sigma_{21}}(X_1, \boldsymbol{X_2})}{\sigma_{11}(X_1)}.$$

On peut vérifier qu'il s'agit bien d'un nombre réel.

Le cas de la corrélation simple est légèrement différent de celui-ci puisqu'ici on ne peut trouver facilement que la valeur absolue de $R(X_1, \boldsymbol{X_2})$. A nouveau on doit faire le lien entre le coefficient de détermination en régression linéaire multiple et $R^2(X_1, \boldsymbol{X_2})$: ils ont la même valeur.

On se donne, comme précemment, un $n-$échantillon indépendant et identiquement distribué $((X_{1,1}, \mathbf{X_{2,1}}), \ldots, (X_{1,n}, \mathbf{X_{2,n}}))$. En utilisant les estimateurs définis à la page 74, un estimateur du coefficient $R^2(X_1, \mathbf{X_2})$ est donné par la formule suivante :

$$\widehat{R^2(X_1, \mathbf{X_2})} = \frac{\boldsymbol{\sigma}_{21}\widehat{(X_1, \mathbf{X_2})}^T \widehat{\boldsymbol{\Sigma}_{22}(\mathbf{X_2})}^{-1} \boldsymbol{\sigma}_{21}\widehat{(X_1, \mathbf{X_2})}}{\widehat{\sigma_{11}(X_1)}} = \frac{\left(\boldsymbol{\sigma}_{21}\widehat{(X_1, \mathbf{X_2})}\right)^T \left(\widehat{\boldsymbol{\Sigma}_{22}(\mathbf{X_2})}\right)^{-1} \boldsymbol{\sigma}_{21}\widehat{(X_1, \mathbf{X_2})}}{\widehat{\sigma_{11}(X_1)}}.$$

On note désormais $\mathbf{x_1}$ un échantillon de n réalisations de X_1 et $\mathbf{x_2}$ un échantillon de n réalisations de $\mathbf{X_2}$.
Une estimation de $R^2(X_1, \mathbf{X_2})$ est alors :

$$\widehat{R^2(X_1, \mathbf{X_2})}(\mathbf{x_1}, \mathbf{x_2}) = \frac{\left(\boldsymbol{\sigma}_{21}\widehat{(X_1, \mathbf{X_2})}(\mathbf{x_1}, \mathbf{x_2})\right)^T \left(\widehat{\boldsymbol{\Sigma}_{22}(\mathbf{X_2})}(\mathbf{x_2})\right)^{-1} \boldsymbol{\sigma}_{21}\widehat{(X_1, \mathbf{X_2})}(\mathbf{x_1}, \mathbf{x_2})}{\widehat{\sigma_{11}(X_1)}(\mathbf{x_1})}.$$

Sans hypothèse supplémentaire on ne pourrait rien dire [a] sur la distribution de $\widehat{R^2(X_1, \mathbf{X_2})}$ et on ne pourrait donc pas effectuer de test.

a. Le théorème central limite permet d'avoir des informations sur la loi limite de $\widehat{R(X_1, \mathbf{X_2})}$ mais ces renseignements ne sont valables que pour des effectifs très importants, $n \to +\infty$ et de ce fait ne présentent que peu d'intérêt dans la pratique.

Supposons désormais que nos $X_1 \in \mathbb{R}$ et $\mathbf{X_2} \in \mathbb{R}^{p-1}$ soient tels que $(X_1, \mathbf{X_2}) \in \mathbb{R}^p$ ait une distribution **multinormale** dans $\mathbb{R}^p$ de moyenne $\boldsymbol{\mu}$ et de matrice de variance-covariance $\boldsymbol{\Sigma}$. En conservant les notations de la page 105 on peut écrire :

$$(X_1, \mathbf{X_2}) \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \boldsymbol{\mu_2} \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \boldsymbol{\sigma_{21}}^T \\ \boldsymbol{\sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{pmatrix}\right).$$

La distribution asymptotique de $R(X_1, \boldsymbol{X_2})$ est alors identique à celle du coefficient de corrélation simple :

$$\sqrt{n}(\widehat{R(X_1, \boldsymbol{X_2})} - R(X_1, \boldsymbol{X_2})) \approx \mathcal{N}\left(0, \left(1 - R(X_1, \boldsymbol{X_2})^2\right)^2\right).$$

Ce résultat n'est pas assez précis pour permettre de résoudre les problèmes qui nous intéressent généralement. On l'utilisera uniquement si l'on ne peut pas faire autrement et si $n \geqslant 30$.

Toujours sous l'hypothèse de multinormalité du vecteur $(X_1, \boldsymbol{X_2})$ comme précisé à la page 115, on peut connaître la loi exacte d'une fonction de l'estimateur $\widehat{R^2}(X_1, \boldsymbol{X_2})$ sous l'hypothèse nulle « $\mathcal{H}_0$: Le coefficient de corrélation multiple $R(X_1, \boldsymbol{X_2})$ de X_1 et $\boldsymbol{X_2}$ est nul » :

$$\frac{n-p-1}{p} \frac{\widehat{R^2(X_1, \boldsymbol{X_2})}}{1 - \widehat{R^2(X_1, \boldsymbol{X_2})}^2} \sim \mathcal{F}_{p,n-p-1}$$

où $\mathcal{F}_{p,n-p-1}$ est une loi de Fisher à p et $n-p-1$ degrés de liberté.

Si $p = 1$, on retrouve exactement le cas de la corrélation simple. En effet dans cette situation, on étudie la corrélation entre deux variables X_1 et X_2 à valeurs réelles. La formule ci-dessus pour $R^2(X_1, X_2)$ est alors identique à celle de $\rho(X_1, X_2)^2$. De plus, on sait que $\frac{\sqrt{n-2}\widehat{\rho(X_1,X_2)}}{\sqrt{1-\widehat{\rho(X_1,X_2)}^2}}$ suit une loi de Student à $n-2$ degrés de liberté. On montre qu'alors son carré

$$\left(\sqrt{n-2}\frac{\widehat{\rho(X_1,X_2)}}{\sqrt{1-\widehat{\rho(X_1,X_2)}^2}}\right)^2 = (n-2)\frac{\widehat{\rho(X_1,X_2)}^2}{1-\widehat{\rho(X_1,X_2)}^2} = (n-2)\frac{\widehat{R^2(X_1,X_2)}}{1-\widehat{R^2(X_1,X_2)}^2}$$

suit une loi de Fisher à 1 et $n-2$ degrés de liberté.

Ceci est bien conforme au résultat de ce paragraphe puisque si $p = 1$ on a vu que $\frac{n-p-1}{p}\frac{\widehat{R^2(X_1,X_2)}}{1-\widehat{R^2(X_1,X_2)}^2} = (n-2)\frac{\widehat{R^2(X_1,X_2)}}{1-\widehat{R^2(X_1,X_2)}^2}$ suit une loi de Fisher à p et $n-p-1$, c'est-à-dire 1 et $n-2$, degrés de liberté. Réciproquement, si $(n-2)\frac{\widehat{R^2(X_1,X_2)}}{1-\widehat{R^2(X_1,X_2)}^2}$ suit une loi de Fisher à 1 et $n-2$, on sait qu'alors $\sqrt{(n-2)\frac{\widehat{R^2(X_1,X_2)}}{1-\widehat{R^2(X_1,X_2)}^2}} = \frac{\sqrt{n-2}\widehat{\rho(X_1,X_2)}}{\sqrt{1-\widehat{\rho(X_1,X_2)}^2}}$ suit une loi de Student à $n-2$ degrés de liberté.

On se sert de cette propriété pour tester les hypothèses :

$\mathcal{H}_0$: Le coefficient de corrélation multiple $R(X_1, \boldsymbol{X_2})$ de X_1 et $\boldsymbol{X_2}$ est nul

contre

$\mathcal{H}_1$: Le coefficient de corrélation multiple $R(X_1, \boldsymbol{X_2})$ de X_1 et $\boldsymbol{X_2}$ est non nul.

Ces hypothèses sont équivalentes aux hypothèses :

$\mathcal{H}_0$: X_1 et $\boldsymbol{X_2}$ sont indépendants

contre

$\mathcal{H}_1$: X_1 et $\boldsymbol{X_2}$ sont liés.

Pour obtenir des intervalles de confiance pour $R(X_1, \boldsymbol{X_2})$ on peut aussi utiliser la transformation en « argument tangente hyperbolique », i.e. en $\mathtt{tanh}^{-1}$, voir la page 56.

On se place encore sous l'hypothèse de multinormalité du vecteur $(X_1, \boldsymbol{X_2})$ comme précisé à la page 115. Notons que la propriété de la page 118 ci-dessus ne permet pas de tester, si $R_0 \neq 0$, des hypothèses du type :

$$\mathcal{H}_0 : R(X_1, \boldsymbol{X_2}) = R_0$$

contre

$$\mathcal{H}_1 : R(X_1, \boldsymbol{X_2}) \neq R_0.$$

La démarche est alors la même que pour le coefficient de corrélation simple, voir la page 87. Deux approches sont possibles.

- Utiliser la transformation de Fisher introduite ci-dessus, à la page 56, pour obtenir un intervalle de confiance pour $R(X_1, \mathbf{X_2})$ de niveau approximatif $(1-\alpha)$ % :

$$\left[\tanh\left(\widehat{z} - \frac{z_{\alpha/2}}{(n-3)^{\frac{1}{2}}}\right), \tanh\left(\widehat{z} + \frac{z_{\alpha/2}}{(n-3)^{\frac{1}{2}}}\right)\right]$$

où $\widehat{z} = \tanh\left(\widehat{R(X_1, \mathbf{X_2})}(\mathbf{x_1}, \mathbf{x_2})\right)$ et $z_{\alpha/2}$ est le quantile de la loi $\mathcal{N}(0,1)$.

- Utiliser des logiciels, comme SPSS, R , SAS ou StatXact, qui proposent des versions *exactes* de l'intervalle de confiance dont une approximation est donnée ci-dessus. Cette approche est particulièrement intéressante si l'effectif commun n aux échantillons $\mathbf{x_1}$ et $\mathbf{x_2}$ est petit.

Sommaire

5 Corrélation partielle
- Définition
- Estimation
- Asymptotique
- Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = 0$
- Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = \rho_0, \rho_0 \neq 0$
- Cas de trois variables réelles

Pour introduire formellement le concept de corrélation partielle, un petit détour mathématique est nécessaire. Les applications seront généralement beaucoup plus simples.

Considérons deux vecteurs $\mathbf{X_1} \in \mathbb{R}^q$ et $\mathbf{X_2} \in \mathbb{R}^{m-q}$ qui ont une distribution conjointe multinormale :

$$(\mathbf{X_1}, \mathbf{X_2}) \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma_{11}} & \mathbf{\Sigma_{12}} \\ \mathbf{\Sigma_{21}} & \mathbf{\Sigma_{22}} \end{pmatrix}\right).$$

La distribution conditionnelle de $\boldsymbol{X_1}$ sachant $\boldsymbol{X_2}$, on étend les résultats de la page 17 qui concernaient le cas de deux variables réelles au cas de deux vecteurs, est :

$$\boldsymbol{X_1}|\boldsymbol{X_2} \sim \mathcal{N}\left(\boldsymbol{\mu_1} + \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}}^{-1}(\boldsymbol{X_2} - \boldsymbol{\mu_2}), \boldsymbol{\Sigma_{11|2}}\right),$$

où $\boldsymbol{\Sigma_{11|2}} = \boldsymbol{\Sigma_{1,1}} - \boldsymbol{\Sigma_{1,2}}\boldsymbol{\Sigma_{2,2}}^{-1}\boldsymbol{\Sigma_{2,1}}$.

En écrivant $\boldsymbol{\Sigma_{11|2}} = \left(\left(\sigma_{i,j|q+1,...,m}\right)\right)_{1 \leqslant i,j \leqslant p}$, il s'agit d'une matrice carrée de taille p, on définit le coefficient de corrélation partielle entre les composantes i et j de $\boldsymbol{X_1}$ sachant $\boldsymbol{X_2}$ comme :

$$\rho_{i,j|q+1,...,m} = \frac{\sigma_{i,j|q+1,...,m}}{\sigma_{i,i|q+1,...,m}^{\frac{1}{2}}\sigma_{j,j|q+1,...,m}^{\frac{1}{2}}}.$$

Son interprétation pratique est la suivante :
Le coefficient de corrélation partielle entre les composantes i et j du vecteur $\boldsymbol{X_1}$ sachant $\boldsymbol{X_2}$ représente la corrélation entre les composantes i et j après avoir éliminé l'effet des variables de $\boldsymbol{X_2}$ sur les composantes i et j du vecteur $\boldsymbol{X_1}$.

On se donne, comme précemment, un $n-$échantillon indépendant et identiquement distribué $((\boldsymbol{X_{1,1}}, \boldsymbol{X_{2,1}}), \ldots, (\boldsymbol{X_{1,n}}, \boldsymbol{X_{2,n}}))$.

En utilisant les estimateurs définis à la page 74, un estimateur du coefficient $\rho_{i,j|q+1,...,m}$ est donné par la formule suivante :

$$\widehat{\rho_{i,j|q+1,...,m}} = \frac{\widehat{\sigma_{i,j|q+1,...,m}}}{\widehat{\sigma_{i,i|q+1,...,m}}^{\frac{1}{2}}\widehat{\sigma_{j,j|q+1,...,m}}^{\frac{1}{2}}},$$

où $\left(\left(\widehat{\sigma_{i,j|q+1,...,m}}\right)\right)_{1 \leqslant i,j \leqslant p} = \widehat{\boldsymbol{\Sigma_{11|2}}} = \widehat{\boldsymbol{\Sigma_{1,1}}} - \widehat{\boldsymbol{\Sigma_{1,2}}}\widehat{\boldsymbol{\Sigma_{2,2}}}^{-1}\widehat{\boldsymbol{\Sigma_{2,1}}}$.

On note désormais $\boldsymbol{x_1}$ un échantillon de n réalisations de $\boldsymbol{X_1}$ et $\boldsymbol{x_2}$ un échantillon de n réalisations de $\boldsymbol{X_2}$.
Une estimation de $\rho_{i,j|q+1,\ldots,m}$ est alors :

$$\widehat{\rho_{i,j|q+1,\ldots,m}}(\boldsymbol{x_1},\boldsymbol{x_2}) = \frac{\widehat{\sigma_{i,j|q+1,\ldots,m}}(\boldsymbol{x_1},\boldsymbol{x_2})}{\widehat{\sigma_{i,i|q+1,\ldots,m}}(\boldsymbol{x_1},\boldsymbol{x_2})^{\frac{1}{2}}\widehat{\sigma_{j,j|q+1,\ldots,m}}(\boldsymbol{x_1},\boldsymbol{x_2})^{\frac{1}{2}}},$$

où

$$\left(\left(\widehat{\sigma_{i,j|q+1,\ldots,m}}(\boldsymbol{x_1},\boldsymbol{x_2})\right)\right)_{1\leqslant i,j\leqslant p} = \widehat{\boldsymbol{\Sigma}_{11|2}}(\boldsymbol{x_1},\boldsymbol{x_2}) =$$

$$\widehat{\boldsymbol{\Sigma}_{1,1}}(\boldsymbol{x_1},\boldsymbol{x_2}) - \widehat{\boldsymbol{\Sigma}_{1,2}}(\boldsymbol{x_1},\boldsymbol{x_2})\left(\widehat{\boldsymbol{\Sigma}_{2,2}}(\boldsymbol{x_1},\boldsymbol{x_2})\right)^{-1}\widehat{\boldsymbol{\Sigma}_{2,1}}(\boldsymbol{x_1},\boldsymbol{x_2}).$$

La distribution asymptotique d'un coefficient de corrélation partielle est la même que pour une corrélation simple, c'est-à-dire :

$$\sqrt{n}\left(\widehat{\rho_{i,j|q+1,\ldots,m}} - \rho_{i,j|q+1,\ldots,m}\right) \approx \mathcal{N}\left(0,(1-\rho^2_{i,j|q+1,\ldots,m})^2\right).$$

Ce résultat n'est applicable que si $n \geqslant 30$. Ne s'en servir que si l'on ne peut pas faire autrement.

On peut maintenant s'intéresser au test de nullité du coefficient de corrélation partielle des composantes du vecteur gaussien $\boldsymbol{X_1}$ connaissant le vecteur $\boldsymbol{X_2}$.

$$\mathcal{H}_0 : \rho_{i,j|q+1,\ldots,m} = 0$$

contre

$$\mathcal{H}_1 : \rho_{i,j|q+1,\ldots,m} \neq 0.$$

Ces hypothèses sont équivalente à :

$\mathcal{H}_0$: Les composantes $X_{1,i}$ et $X_{1,j}$ de $\boldsymbol{X_1}$ sont indépendantes sans l'effet de $\boldsymbol{X_2}$

contre

$\mathcal{H}_1$: Les composantes $X_{1,i}$ et $X_{1,j}$ de $\boldsymbol{X_1}$ sont liées sans l'effet de $\boldsymbol{X_2}$.

Sous l'hypothèse nulle $\mathcal{H}_0$, alors :

$$t_{i,j|q+1,\ldots,m}(\mathbf{x_1}, \mathbf{x_2}) = \sqrt{n-m+q-2}\,\frac{\widehat{\rho_{i,j|q+1,\ldots,m}}(\mathbf{x_1}, \mathbf{x_2})}{\sqrt{1-\widehat{\rho_{i,j|q+1,\ldots,m}}(\mathbf{x_1}, \mathbf{x_2})^2}}$$

est la réalisation d'une variable aléatoire T qui suit une loi de Student à $n-m+q-2$ degrés de liberté, i.e. $T \sim \mathcal{T}_{n-m+q-2}$.

On rejette l'hypothèse nulle $\mathcal{H}_0$ au niveau $(1-\alpha)$ % lorsque $\left|t_{i,j|q+1,\ldots,m}(\mathbf{x_1}, \mathbf{x_2})\right| > t_{n-m+q-2,\alpha/2}$.

On peut aussi appliquer la transformation z dans ce contexte, voir la page 56.

Si $\rho_0 \neq 0$, la propriété de la page 135 ci-dessus ne permet pas de tester des hypothèses du type :

$$\mathcal{H}_0 : \rho_{i,j|q+1,\ldots,m} = \rho_0$$

contre

$$\mathcal{H}_1 : \rho_{i,j|q+1,\ldots,m} \neq \rho_0.$$

Il existe alors deux possibilités.

- Si l'on souhaite faire un test de ce type à la main, on utilise la transformation de Fisher introduite ci-dessus, à la page 56, pour obtenir un intervalle de confiance pour $\rho_{i,j|q+1,\ldots,m}$ de niveau approximatif $(1-\alpha)$ % :

$$\left[\tanh\left(\widehat{z} - \frac{z_{\alpha/2}}{(n-3)^{\frac{1}{2}}}\right), \tanh\left(\widehat{z} + \frac{z_{\alpha/2}}{(n-3)^{\frac{1}{2}}}\right)\right]$$

où $\widehat{z} = \tanh\left(\widehat{\rho_{i,j|q+1,\ldots,m}}(\mathbf{x_1}, \mathbf{x_2})\right)$ et $z_{\alpha/2}$ est le quantile de la loi $\mathcal{N}(0,1)$.

- Certains logiciels, comme SPSS, R , SAS ou StatXact, proposent des versions *exactes* de l'intervalle de confiance.

Intéressons-nous au cas le plus simple. Soient X_1, X_2 et X_3 trois variables aléatoires réelles telles que la loi jointe de (X_1, X_1, X_3) soit multinormale de paramètres $\boldsymbol{\mu}$ et $\boldsymbol{\Sigma}$.

Ainsi on a ici $m = 3$ et $q = 2$. On montre alors la relation suivante :

$$\rho_{12|3} = \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{1-\rho_{13}^2}\sqrt{1-\rho_{23}^2}}$$

où ρ_{12} est le coefficient de corrélation simple entre les variables X_1 et X_2, ρ_{13} est le coefficient de corrélation simple entre les variables X_1 et X_3 et ρ_{23} est le coefficient de corrélation simple entre les variables X_2 et X_3, voir la page 71.

Dans la plupart des situations expérimentales que vous rencontrerez cette formule sera suffisante. On remarque également que la définition est symétrique en X_1 et X_2, c'est-à-dire :

$$\rho_{12|3} = \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{1-\rho_{13}^2}\sqrt{1-\rho_{23}^2}} = \frac{\rho_{21} - \rho_{23}\rho_{13}}{\sqrt{1-\rho_{23}^2}\sqrt{1-\rho_{13}^2}} = \rho_{21|3}.$$

La corrélation partielle de X_1 et X_2 sachant X_3 est heureusement la même que celle de X_2 et X_1 sachant X_3.

On considère un échantillon $\boldsymbol{x}$ indépendant et identiquement distribué suivant la loi de X_1, X_2, X_3. On note $\boldsymbol{x_1}$ l'échantillon des réalisations de X_1, $\boldsymbol{x_2}$ l'échantillon des réalisations de X_2 et $\boldsymbol{x_3}$ l'échantillon des réalisations de X_3.

On tire de $\boldsymbol{x}$ une estimation de $\rho_{12|3}$ en calculant des estimations de ρ_{12}, ρ_{13}, ρ_{23} comme expliqué à la page 74.

On peut maintenant s'intéresser au test de nullité du coefficient de corrélation partielle de X_1 et X_2 connaissant la variable X_3.

$$\mathcal{H}_0 : \rho_{12|3} = 0$$

contre

$$\mathcal{H}_1 : \rho_{12|3} \neq 0.$$

Ces hypothèses sont équivalente à :

$\mathcal{H}_0$: Les variables X_1 et X_2 sont indépendantes sans l'effet de X_3

contre

$\mathcal{H}_1$: Les variables X_1 et X_2 sont liées sans l'effet de X_3.

Sous l'hypothèse nulle $\mathcal{H}_0$, alors :

$$t_{12|3}(\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}) = \sqrt{n-3}\,\frac{\widehat{\rho_{12|3}}(\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3})}{\sqrt{1-\widehat{\rho_{12|3}}(\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3})^2}}$$

est la réalisation d'une variable aléatoire T qui suit une loi de Student à $n-3$ degrés de liberté, i.e. $T \sim \mathcal{T}_{n-3}$.

On rejette l'hypothèse nulle $\mathcal{H}_0$ au niveau $(1-\alpha)$ % lorsque $\left|t_{12|3}(\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3})\right| > t_{n-3,\alpha/2}$.

On peut aussi appliquer la transformation z dans ce contexte, voir la page 56.

Si $\rho_0 \neq 0$, la propriété du paragraphe ci-dessus ne permet pas de tester des hypothèses du type :

$$\mathcal{H}_0 : \rho_{12|3} = \rho_0$$

contre

$$\mathcal{H}_1 : \rho_{12|3} \neq \rho_0.$$

Il existe alors deux possibilités.

- Si l'on souhaite faire un test de ce type à la main, on utilise la transformation de Fisher introduite ci-dessus, à la page 56, pour obtenir un intervalle de confiance pour $\rho_{12|3}$ de niveau approximatif $(1-\alpha)$ % :

$$\left[\tanh\left(\widehat{z} - \frac{z_{\alpha/2}}{(n-3)^{\frac{1}{2}}}\right), \tanh\left(\widehat{z} + \frac{z_{\alpha/2}}{(n-3)^{\frac{1}{2}}}\right)\right]$$

où $\widehat{z} = \tanh\left(\widehat{\rho_{12|3}}(\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3})\right)$ et $z_{\alpha/2}$ est le quantile de la loi $\mathcal{N}(0,1)$.

- Certains logiciels, comme SPSS, R , SAS ou StatXact, proposent des versions *exactes* de l'intervalle de confiance.

Exemple

Détaillons le traitement de la corrélation partielle dans la situation suivante où l'on dispose de trois variables continues réelles.

- Le BMI X_1.
- Le poids [a] X_2.
- La taille X_3.

a. D'un poids de vue physique, il s'agit de la masse.

Exemple

On rappelle que le BMI d'un individu est un indice qui se calcule à partir de la taille et du poids par la formule suivante :

$$\mathrm{BMI} = \frac{\mathrm{Weight}}{\mathrm{Height}^2}$$

où le poids s'exprime en *kg* et la taille en *m*.
L'échantillon étudié a un effectif $n = 38$ et est exclusivement constitué d'individus de genre féminin.

Exemple

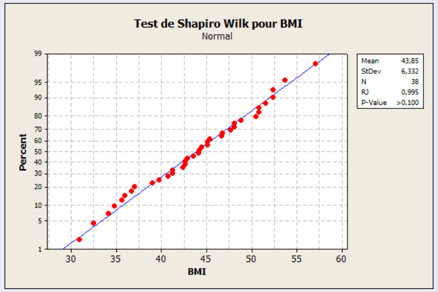
Comme nous l'avons fait remarquer, l'hypothèse nécessaire à l'approche paramétrique de la corrélation simple, multiple ou partielle qui vous a été présentée plus haut sont que la loi du vecteur (X_1, X_2, X_3) est une loi multinormale sur $\mathbb{R}^3$.

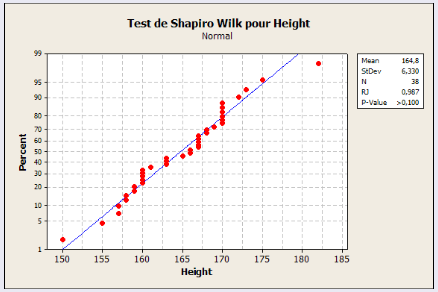
Ceci a pour conséquence qu'il faut que X_1, X_2 et X_3 aient une distribution normale mais comme nous l'avons déjà souligné ce n'est pas suffisant.

Exemple

On doit par exemple aussi tester la normalité de $X_1 + X_2$, $X_2 + X_3$, $X_1 + X_3$ et $X_1 + X_2 + X_3$. Se référer à la page **??** pour un exposé détaillé des tests de multinormalité ou au livre de R. Christensen [3].

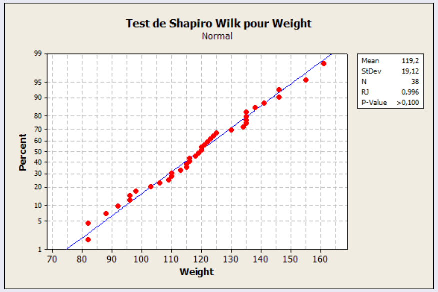
Exemple

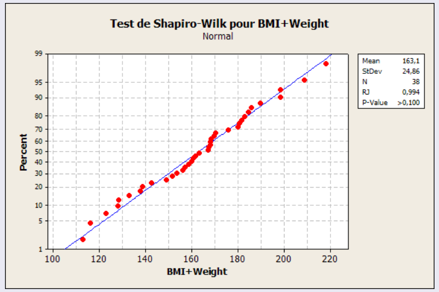




Coefficient de corrélation simple
Le cas bidimensionnel
Le cas général ($n \geqslant 2$)
Corrélation multiple
Corrélation partielle
Définition
Estimation
Asymptotique
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = 0$
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = \rho_0$, $\rho_0 \neq 0$
Cas de trois variables réelles

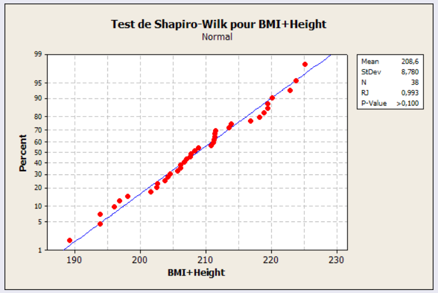
Exemple

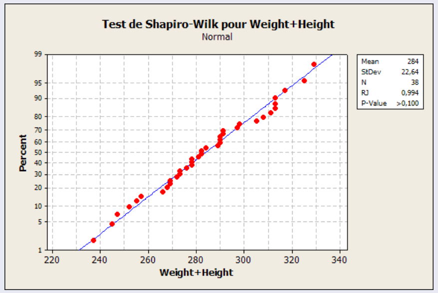




Coefficient de corrélation simple
Le cas bidimensionnel
Le cas général ($n \geqslant 2$)
Corrélation multiple
Corrélation partielle
Définition
Estimation
Asymptotique
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = 0$
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = \rho_0$, $\rho_0 \neq 0$
Cas de trois variables réelles

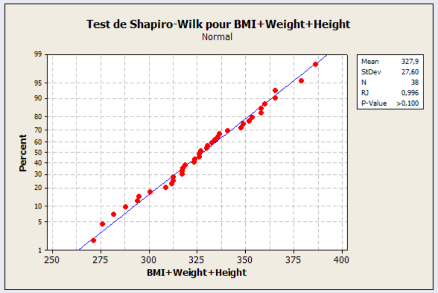
Exemple





Coefficient de corrélation simple
Le cas bidimensionnel
Le cas général ($n \geqslant 2$)
Corrélation multiple
Corrélation partielle
Définition
Estimation
Asymptotique
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = 0$
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = \rho_0$, $\rho_0 \neq 0$
Cas de trois variables réelles

Exemple



Exemple

Les $p-$valeurs sont toutes supérieures à $0,100$, on ne peut donc pas rejeter l'hypothèse de normalité de X_1, X_2, X_3, $X_1 + X_2$, $X_1 + X_3$, $X_2 + X_3$ et de $X_1 + X_2 + X_3$.

Attention, nous testions ici la multinormalité de (X_1, X_2, X_3) et nous avons procédé à sept tests de Shapiro-Wilk. Nous avions, comme d'habitude, fixé un seuil de $\alpha_{global} = 5$ % pour le test global de multinormalité.

Coefficient de corrélation simple
Le cas bidimensionnel
Le cas général ($n \geqslant 2$)
Corrélation multiple
Corrélation partielle
Définition
Estimation
Asymptotique
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = 0$
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = \rho_0$, $\rho_0 \neq 0$
Cas de trois variables réelles

Exemple

Pour garantir ce risque global de 5 %, il faut fixer un risque différent de 5 % pour chaque test de Shapiro-Wilk, nous avons déjà évoqué ce problème à l'exemple page 91. En effet si le risque pour chacun des tests de Shapiro-Wilk est fixé à $\alpha_{ind} = 5$ %, alors le risque global est au maximum de $\alpha_{global} = 1 - (1 - 0,05)^7 = 0,302 \gg 0,05$. Ainsi on doit fixer un seuil de $\alpha_{ind} = 1 - \sqrt[7]{1 - \alpha_{global}} = 1 - \sqrt[7]{0,95} = 0,007$ pour chacun des sept tests de Shapiro-Wilk, ce qui revient à être moins exigeant pour chacun des tests individuels puisque $0,007 < 0,05$ et donc à accepter la normalité dans plus de cas.

Coefficient de corrélation simple
Le cas bidimensionnel
Le cas général ($n \geqslant 2$)
Corrélation multiple
Corrélation partielle
Définition
Estimation
Asymptotique
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = 0$
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = \rho_0$, $\rho_0 \neq 0$
Cas de trois variables réelles

Exemple

Dans cet exemple cette correction ne change rien à la conclusion puisqu'aucune des $p-$valeurs n'étaient comprises entre $0,007$ et $0,05$.
Suite à cette analyse sommaire rien ne permet de rejeter l'hypothèse de multinormalité de (X_1, X_2, X_3).
On peut donc utiliser une approche paramétrique.
Commençons par étudier les coefficients de corrélation simple, dits de Pearson.

Coefficient de corrélation simple
Le cas bidimensionnel
Le cas général ($n \geqslant 2$)
Corrélation multiple
Corrélation partielle
Définition
Estimation
Asymptotique
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = 0$
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = \rho_0$, $\rho_0 \neq 0$
Cas de trois variables réelles

Exemple

```
Correlations: BMI; Weight; Height

            BMI   Weight
Weight    0,878
          0,000

Height   -0,038    0,441
          0,820    0,006

Cell Contents: Pearson correlation
               P-Value
```

Exemple

On constate ainsi que :

- Au niveau $\alpha = 5$ %, on rejette l'hypothèse nulle $\mathcal{H}_0$ d'indépendance des variables BMI et Weight, car $0,000 < 0,05$.
- Au niveau $\alpha = 5$ %, on ne peut pas rejetter l'hypothèse nulle $\mathcal{H}_0$ d'indépendance des variables BMI et Height, car $0,820 > 0,05$.
- Au niveau $\alpha = 5$ %, on rejette l'hypothèse nulle $\mathcal{H}_0$ d'indépendance des variables Height et Weight, car $0,006 < 0,05$.

Exemple

Ces résultats sont en accord avec la formule permettant de calculer le BMI à l'aide de la taille et du poids.

- La liaison est linéaire et positive pour le poids, ce que l'on retrouve bien ici avec la valeur de $\rho_{BW} = 0,878$ et une $p-$valeur de 0,000.

Exemple

- La liaison est négative pour la taille mais, vu sa faiblesse, $\rho_{BH} = -0,038$, on ne peut pas rejeter l'hypothèse d'indépendance entre ces deux variables, $p-$valeur de 0,820. La formule ci-dessus implique une **relation non linéaire** entre le BMI et la taille c'est pourquoi le coefficient de corrélation linéaire ne peut la mettre en évidence.
- La liaison positive entre la taille et la masse, $\rho_{WH} = 0,441$ et $p-$valeur de 0,006, est en accord avec notre connaissance a priori de l'existence d'une relation linéaire positive entre la taille et le poids d'un individu.

Exemple

Passons maintenant au calcul de la corrélation multiple R du BMI sur $(\mathrm{Height}, \mathrm{Weight})$, de la corrélation multiple R du Height sur $(\mathrm{BMI}, \mathrm{Weight})$ et de la corrélation multiple R du Weight sur $(\mathrm{Height}, \mathrm{BMI})$.
Pour effectuer ce calcul on utilise le fait que R^2 est égal au coefficient de détermination obtenu lorsque l'on fait la régression linéaire multiple de BMI sur $(\mathrm{Height}, \mathrm{Weight})$. On obtient par exemple avec Minitab.

Exemple

```
Regression Analysis: BMI versus Weight; Height

The regression equation is
BMI = 87,1 + 0,368 Weight - 0,529 Height

Predictor       Coef    SE Coef       T      P
Constant      87,090      1,899   45,85  0,000
Weight      0,367955   0,004146   88,75  0,000
Height      -0,52855    0,01252  -42,21  0,000

S = 0,432656    R-Sq = 99,6%    R-Sq(adj) = 99,5%
```


Coefficient de corrélation simple
Le cas bidimensionnel
Le cas général ($n \geqslant 2$)
Corrélation multiple
Corrélation partielle
Définition
Estimation
Asymptotique
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = 0$
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = \rho_0, \rho_0 \neq 0$
Cas de trois variables réelles

Exemple

Pour calculer les autres coefficients de corrélation multiple, on procède de même en changeant les rôles de BMI, Height et Weight.

Coefficient de corrélation simple
Le cas bidimensionnel
Le cas général ($n \geqslant 2$)
Corrélation multiple
Corrélation partielle
Définition
Estimation
Asymptotique
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = 0$
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = \rho_0, \rho_0 \neq 0$
Cas de trois variables réelles

Exemple

```
Regression Analysis: Weight versus BMI; Height

The regression equation is
Weight = - 236 + 2,71 BMI + 1,44 Height

Predictor       Coef   SE Coef       T      P
Constant    -236,083     5,253  -44,95  0,000
BMI          2,70570   0,03049   88,75  0,000
Height       1,43599   0,03050   47,09  0,000

S = 1,17324    R-Sq = 99,6%    R-Sq(adj) = 99,6%
```


Coefficient de corrélation simple
Le cas bidimensionnel
Le cas général ($n \geqslant 2$)
Corrélation multiple
Corrélation partielle
Définition
Estimation
Asymptotique
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = 0$
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = \rho_0, \rho_0 \neq 0$
Cas de trois variables réelles

Exemple

```
Regression Analysis: Height versus BMI; Weight

The regression equation is
Height = 164 - 1,86 BMI + 0,686 Weight

Predictor       Coef   SE Coef       T      P
Constant     164,436     0,933  176,17  0,000
BMI         -1,85552   0,04396  -42,21  0,000
Weight       0,68556   0,01456   47,09  0,000

S = 0,810650    R-Sq = 98,4%    R-Sq(adj) = 98,4%
```

Exemple

On a ainsi trouvé successivement :

$$\begin{aligned} \rho_{B/^s(H,W)} &= 99,6\ \% \\ \rho_{W/^s(B,H)} &= 99,6\ \% \\ \rho_{H/^s(B,W)} &= 98,4\ \% \end{aligned}$$

Exemple

Calculons désormais les trois corrélations partielles

- $\rho_{(\text{Height},\text{Weight})|\text{BMI}}$;
- $\rho_{(\text{Weight},\text{BMI})|\text{Height}}$;
- $\rho_{(\text{Height},\text{BMI})|\text{Weight}}$.

Exemple

On applique la formule ci-dessus :

$$\begin{aligned} \rho_{HW|B} &= \frac{\rho_{HW} - \rho_{HB}\rho_{WB}}{\sqrt{1-\rho_{HB}^2}\sqrt{1-\rho_{WB}^2}} \approx 0,992 \\ \rho_{WB|H} &= \frac{\rho_{WB} - \rho_{WH}\rho_{BH}}{\sqrt{1-\rho_{WH}^2}\sqrt{1-\rho_{BH}^2}} \approx 0,998 \\ \rho_{HB|W} &= \frac{\rho_{HB} - \rho_{HW}\rho_{BW}}{\sqrt{1-\rho_{HW}^2}\sqrt{1-\rho_{BW}^2}} \approx -0,990. \end{aligned}$$

Exemple

Ces résultats sont à comparer avec les valeurs de corrélation simple. En particulier la corrélation entre le poids et la taille en éliminant l'effet du poids semble significative. Pour aller plus loin et décider de la significativité de ces corrélations, on peut se référer à une table ou utiliser la loi du coefficient de corrélation partielle énoncée aux pages 134 et 142.
Certains logiciels, comme SPSS, nous renseignent directement sur les $p-$valeurs associées aux corrélations partielles.

Exemple

Partial Correlations

Control Variables			Weight	Height
BMI	Weight	Correlation	1,000	,992
		Significance (2-tailed)	.	,000
		df	0	35
	Height	Correlation	,992	1,000
		Significance (2-tailed)	,000	.
		df	35	0

Exemple

Partial Correlations

Control Variables			Weight	BMI
Height	Weight	Correlation	1,000	,998
		Significance (2-tailed)	.	,000
		df	0	35
	BMI	Correlation	,998	1,000
		Significance (2-tailed)	,000	.
		df	35	0

Exemple

Partial Correlations

Control Variables			BMI	Height
Weight	BMI	Correlation	1,000	-,990
		Significance (2-tailed)	.	,000
		df	0	35
	Height	Correlation	-,990	1,000
		Significance (2-tailed)	,000	.
		df	35	0

Exemple

On constate ainsi que les trois tests sont significatifs au seuil $\alpha = 5$ %.

On ne peut conserver l'hypothèse d'indépendance de deux variables sachant la troisième et ce dans les trois cas qui se présentent ici : $(\mathrm{Height}, \mathrm{Weight})|\mathrm{BMI}$, $(\mathrm{Weight}, \mathrm{BMI})|\mathrm{Height}$ et $(\mathrm{Height}, \mathrm{BMI})|\mathrm{Weight}$.

Exemple

Avez-vous une critique à formuler quant au modèle que nous avons utilisé ici ? Par exemple l'hypothèse de multinormalité est-elle vraisemblable ? Bien que les tests utilisés ne l'infirment pas, notre connaissance de la relation entre le BMI et le couple $(\mathrm{Height}, \mathrm{Weight})$ ne devait-elle pas nous faire exclure ce modèle ?

Exemple

On réalise dans un collège deux tests d'évaluation communs à tous les élèves quel que soit leur âge mais exclusivement de genre masculin. L'un porte sur leur compétence en mathématiques et l'autre en sport, le temps mis à parcourir 100 m en course à pied. Les performances ont été notées sur 100 dans les deux tests puis ramenées à des notes sur 20. On dispose donc de deux variables aléatoires :

- l'une appelée Math (M) et associée à la note en mathématiques qu'a reçu un élève,
- l'autre appelée Sport (S) et associée à la note en sport qu'a reçu un élève.

Exemple

Les données ont été reportées dans le tableau suivant :

Élève	Math	Sport	Élève	Math	Sport
1	7,15	2,69	16	12,29	13,16
2	7,79	10,56	17	11,56	7,92
3	8,57	8,91	18	14,31	9,11
4	7,14	6,04	19	11,33	13,49
5	5,47	7,53	20	9,49	10,99
6	4,43	4,04	21	16,99	16,54
7	4,52	7,45	22	14,86	16,03
8	6,63	4,58	23	18,91	16,41

Exemple

Élève	Math	Sport	Élève	Math	Sport
9	5,43	7,50	24	19,35	14,81
10	7,78	10,18	25	19,24	17,63
11	9,79	13,59	26	14,62	18,78
12	13,40	6,92	27	20,00	18,54
13	8,68	7,51	28	19,19	20,00
14	11,35	9,59	29	19,28	15,64
15	10,62	13,19	30	16,59	19,73

L'échantillon étudié a un effectif $n = 30$ et est exclusivement constitué d'individus de genre masculin.

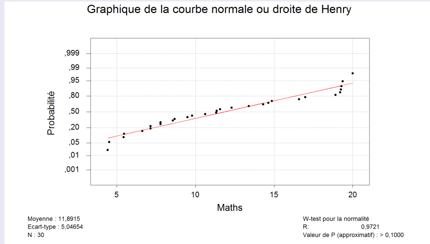
Exemple

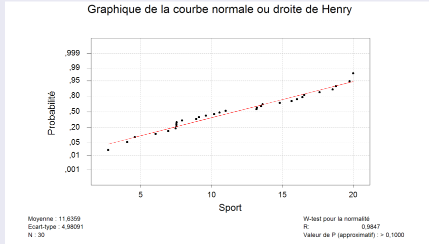
Comme nous l'avons fait remarquer, l'hypothèse nécessaire à l'approche paramétrique de la corrélation simple, multiple ou partielle qui vous a été présentée plus haut est que la loi du vecteur (M, S) est une loi multinormale sur $\mathbb{R}^2$.

Ceci a pour conséquence qu'il faut que M, S aient une distribution normale mais comme nous l'avons déjà souligné ce n'est pas suffisant. On doit par exemple aussi tester la normalité de $\mathrm{M} + \mathrm{S}$ et $\mathrm{M} - \mathrm{S}$. Se référer à la page **??** pour un exposé détaillé des tests de multinormalité ou au livre de R. Christensen [3].

Coefficient de corrélation simple
Le cas bidimensionnel
Le cas général ($n \geqslant 2$)
Corrélation multiple
Corrélation partielle
Définition
Estimation
Asymptotique
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = 0$
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = \rho_0$, $\rho_0 \neq 0$
Cas de trois variables réelles

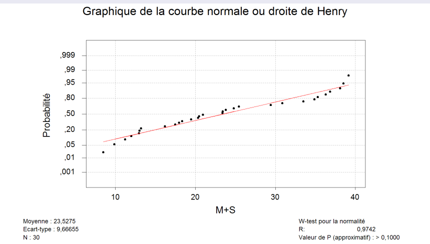
Exemple

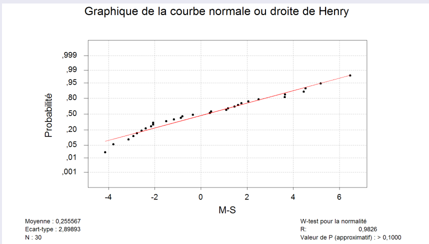




Coefficient de corrélation simple
Le cas bidimensionnel
Le cas général ($n \geqslant 2$)
Corrélation multiple
Corrélation partielle
Définition
Estimation
Asymptotique
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = 0$
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = \rho_0$, $\rho_0 \neq 0$
Cas de trois variables réelles

Exemple





Coefficient de corrélation simple
Le cas bidimensionnel
Le cas général ($n \geqslant 2$)
Corrélation multiple
Corrélation partielle
Définition
Estimation
Asymptotique
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = 0$
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = \rho_0$, $\rho_0 \neq 0$
Cas de trois variables réelles

Exemple

Les $p-$valeurs sont toutes supérieures à $0,100$, on ne peut donc pas rejeter l'hypothèse de normalité de M, S, $\mathrm{M} + \mathrm{S}$ et $\mathrm{M} - \mathrm{S}$.

Attention, nous testions ici la multinormalité de (M, S) et nous avons procédé à quatre tests de Shapiro-Wilk. Nous avions, comme d'habitude, fixé un seuil de $\alpha_{global} = 5$ % pour le test global de multinormalité. Pour garantir ce risque global de 5 %, il faut fixer un risque différent de 5 % pour chaque test de Shapiro-Wilk.

Exemple

En effet si le risque pour chacun des tests de Shapiro-Wilk est fixé à $\alpha_{ind} = 5$ %, alors le risque global est au maximum de $\alpha_{global} = 1 - (1-0,05)^4 = 0,185 > 0,05$. Ainsi on doit fixer un seuil de $\alpha_{ind} = 1 - \sqrt[4]{1-\alpha_{global}} = 1 - \sqrt[4]{0,95} = 0,013$ pour chacun des tests de Shapiro-Wilk, ce qui revient à être moins exigeant pour chacun des tests individuels puisque $0,013 < 0,05$ et donc à accepter la normalité dans plus de cas. Dans cet exemple cette correction ne change rien à la conclusion puisqu'aucune des $p-$valeurs n'étaient comprises entre $0,013$ et $0,05$. Suite à cette analyse sommaire rien ne permet de rejeter l'hypothèse de multinormalité de (M, S).

Exemple

On peut donc utiliser une approche paramétrique et utiliser le coefficient de corrélation de M et S.

```
Corrélation de Pearson de Maths et Sport = 0,833
Valeur de p = 0,000
```

On constate ainsi qu'au seuil $\alpha = 5$ %, on rejette l'hypothèse nulle $\mathcal{H}_0$ d'indépendance des variables Math et Sport, car $0,000 < 0,05$.

Exemple

On décide donc qu'il y a une corrélation significative entre les résultats en mathématiques et en sport. La valeur de l'estimation, $0,833$, du coefficient de corrélation étant positive, l'association est positive, c'est-à-dire, plus on est fort en sport plus on est fort en mathématiques. Qu'en pensez-vous ? **De quel autre facteur faudrait-il tenir compte ?**

Exemple

Il est évident que les résultats des élèves à ces tests dépend de leur âge puisque le même questionnaire est posé à tous les élèves quelque soit leur niveau, donc aussi bien en classe de $6^{\text{ème}}$ qu'en classe de $3^{\text{ème}}$.

Fort heureusement il a été possible de retrouver l'âge, et même une information encore plus précise, la date de naissance, de chacun des élèves qui a participé à l'évaluation. On a alors décidé de coder l'âge comme une variable aléatoire quantitative continue, appelée Age et parfois abrégée en A. Cette décision correspond à l'idée, relativement raisonnable, suivante : le développement d'un enfant ne serait pas exactement le même si celui-ci est né en début ou en fin d'année civile.

Exemple

Élève	Math	Sport	Age	Élève	Math	Sport	Age
1	7,15	2,69	11,72	16	12,29	13,16	14,02
2	7,79	10,56	12,33	17	11,56	7,92	15,02
3	8,57	8,91	11,91	18	14,31	9,11	13,40
4	7,14	6,04	11,92	19	11,33	13,49	13,70
5	5,47	7,53	12,64	20	9,49	10,99	14,14
6	4,43	4,04	12,07	21	16,99	16,54	15,56
7	4,52	7,45	12,12	22	14,86	16,03	16,31
8	6,63	4,58	11,23	23	18,91	16,41	16,30

Exemple

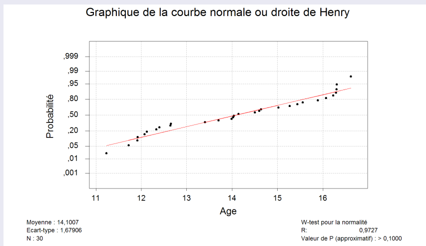
Élève	Math	Sport	Age	Élève	Math	Sport	Age
9	5,43	7,50	12,65	24	19,35	14,81	16,23
10	7,78	10,18	12,40	25	19,24	17,63	15,89
11	9,79	13,59	13,99	26	14,62	18,78	15,44
12	13,40	6,92	14,50	27	20,00	18,54	16,29
13	8,68	7,51	14,04	28	19,19	20,00	16,62
14	11,35	9,59	14,64	29	19,28	15,64	16,07
15	10,62	13,19	14,60	30	16,59	19,73	15,27

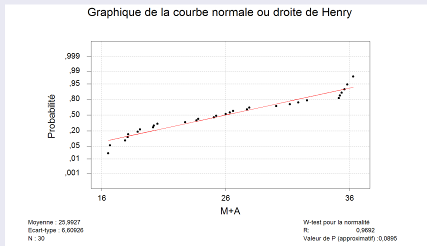
Exemple

Puisque nous introduisons une nouvelle variable aléatoire Age, pour étudier les corrélations du vecteur $(\mathrm{Math}, \mathrm{Sport}, \mathrm{Age})$ nous devons tester la normalité du vecteur $(\mathrm{Math}, \mathrm{Sport}, \mathrm{Age})$.

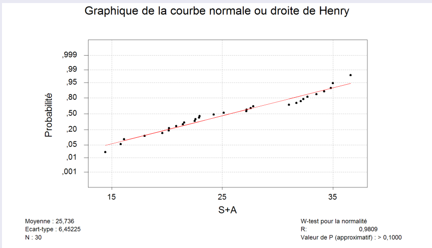
On procède comme précédemment. En s'appuyant sur les résultats des tests de normalité ci-dessus, on voit qu'il ne reste plus qu'à s'intéresser à la normalité de Age, $\mathrm{Math} + \mathrm{Age}$, $\mathrm{Sport} + \mathrm{Age}$ et de $\mathrm{Math} + \mathrm{Sport} + \mathrm{Age}$.

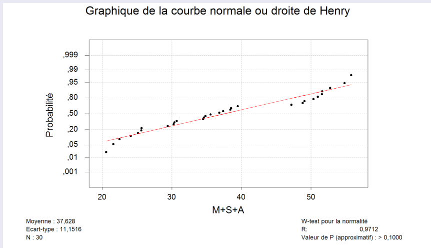
Exemple





Exemple





Coefficient de corrélation simple
Le cas bidimensionnel
Le cas général ($n \geqslant 2$)
Corrélation multiple
Corrélation partielle
Définition
Estimation
Asymptotique
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = 0$
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = \rho_0$, $\rho_0 \neq 0$
Cas de trois variables réelles

Exemple

Aucun de ces tests n'est significatif au seuil 5 % et donc **a fortiori** ils ne seront pas significatifs au seuil inférieur qu'il faudrait fixer pour garantir un risque global de 5 %. En ajoutant ces résultats aux précédents, rien ne permet de rejeter l'hypothèse de normalité du vecteur $(\mathrm{Math}, \mathrm{Sport}, \mathrm{Age})$.

On calcule donc les corrélations simples entre les composantes de ce vecteur.

Coefficient de corrélation simple
Le cas bidimensionnel
Le cas général ($n \geqslant 2$)
Corrélation multiple
Corrélation partielle
Définition
Estimation
Asymptotique
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = 0$
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = \rho_0$, $\rho_0 \neq 0$
Cas de trois variables réelles

Exemple

```
Corrélations : Maths; Sport; Age

          Maths     Sport
Sport     0,833
          0,000

Age       0,909     0,837
          0,000     0,000

Contenu de la cellule : corrélation de Pearson
                        Valeur de p
```


Coefficient de corrélation simple
Le cas bidimensionnel
Le cas général ($n \geqslant 2$)
Corrélation multiple
Corrélation partielle
Définition
Estimation
Asymptotique
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = 0$
Test de l'hypothèse $\rho_{i,j|q+1,\ldots,m} = \rho_0$, $\rho_0 \neq 0$
Cas de trois variables réelles

Exemple

On note une corrélation positive et significative, au seuil $\alpha = 5$ % pour chacune des associations possibles. Ainsi, comme on pouvait le prévoir, plus l'on est âgé, mieux l'on réussit en mathématiques et en sport.

Exerçons-nous encore au calcul de coefficients de corrélation multiple. Ainsi déterminons la corrélation multiple R du Math sur $(\mathrm{Sport}, \mathrm{Age})$, la corrélation multiple R du Sport sur $(\mathrm{Math}, \mathrm{Age})$ et la corrélation multiple R du Age sur $(\mathrm{Math}, \mathrm{Sport})$.

Exemple

Pour effectuer ce calcul on utilise le fait que R^2 est égal au coefficient de détermination obtenu lorsque l'on fait la régression linéaire multiple de Math sur (Sport, Age). On obtient par exemple avec Minitab.

Exemple

```
Regression Analysis: Maths versus Sport; Age

L'équation de régression est
Maths = - 20,9 + 0,244 Sport + 2,13 Age

Régresseur        Coef   Er-T coef      T      P
Constante      -20,919       4,635  -4,51  0,000
Sport           0,2437      0,1412   1,73  0,096
Age             2,1258      0,4189   5,07  0,000

S = 2,071         R-Sq = 84,3%  R-Sq(adj) = 83,2%
```


Exemple

Pour calculer les autres coefficients de corrélation multiple, on procède de même en changeant les rôles de Math, Sport et Age.

Exemple

```
Regression Analysis: Sport versus Maths; Age

L'équation de régression est
Sport = - 12,5 + 0,408 Maths + 1,37 Age

Régresseur        Coef   Er-T coef      T      P
Constante      -12,533       7,565  -1,66  0,109
Maths           0,4077      0,2362   1,73  0,096
Age             1,3701      0,7099   1,93  0,064

S = 2,679         R-Sq = 73,1%  R-Sq(adj) = 71,1%
```

Exemple

```
Regression Analysis: : Age versus Maths; Sport

L'équation de régression est
Age = 10,3 + 0,230 Maths + 0,0885 Sport

Régresseur      Coef   Er-T coef       T      P
Constante    10,3401      0,3338   30,97  0,000
Maths        0,22965     0,04525    5,07  0,000
Sport        0,08848     0,04585    1,93  0,064

S = 0,6807       R-Sq = 84,7%   R-Sq(adj) = 83,6%
```


Exemple

On a ainsi trouvé successivement :

$$\begin{aligned}\rho_{M/^s(S,A)} &= 84,3\ \%\\ \rho_{S/^s(M,A)} &= 73,1\ \%\\ \rho_{A/^s(M,S)} &= 84,7\ \%.\end{aligned}$$

Exemple

Calculons désormais les trois corrélations partielles

- $\rho_{(\text{Maths},\text{Sport})|\text{Age}}$;
- $\rho_{(\text{Maths},\text{Age})|\text{Sport}}$;
- $\rho_{(\text{Sport},\text{Age})|\text{Maths}}$.

Exemple

On applique la formule ci-dessus :

$$\begin{aligned}\rho_{MS|A} &= \frac{\rho_{MS} - \rho_{MA}\rho_{SA}}{\sqrt{1-\rho_{MA}^2}\sqrt{1-\rho_{SA}^2}} \approx 0,833\\ \rho_{MA|S} &= \frac{\rho_{MA} - \rho_{MS}\rho_{SA}}{\sqrt{1-\rho_{MS}^2}\sqrt{1-\rho_{SA}^2}} \approx 0,909\\ \rho_{SA|M} &= \frac{\rho_{SA} - \rho_{MS}\rho_{MA}}{\sqrt{1-\rho_{MS}^2}\sqrt{1-\rho_{MA}^2}} \approx 0,837.\end{aligned}$$

Exemple

Ces résultats sont à comparer avec les valeurs de corrélation simple. En particulier la corrélation entre le poids et la taille en éliminant l'effet du poids semble significative. Pour aller plus loin et décider de la significativité de ces corrélations on peut se réferrer à une table ou utiliser la loi du coefficient de corrélation partielle énoncée aux pages 134 et 142.

Exemple

Certains logiciels, comme SPSS, nous renseignent directement sur les $p-$valeurs associées aux corrélations partielles.

Exemple

```
Correlations
 | ---------------- | ----- | ---------------------- | ----- | ----- |
 | Control Variables |       |                        | Maths | Sport |
 | ---------------- | ----- | ---------------------- | ----- | ----- |
 | Age              | Maths | Correlation            | 1,000 | ,315  |
 |                  |       | ---------------------- | ----- | ----- |
 |                  |       | Significance (2-tailed) | .     | ,096  |
 |                  |       | ---------------------- | ----- | ----- |
 |                  |       | df                     | 0     | 27    |
 |                  | ----- | ---------------------- | ----- | ----- |
 |                  | Sport | Correlation            | ,315  | 1,000 |
 |                  |       | ---------------------- | ----- | ----- |
 |                  |       | Significance (2-tailed) | ,096  | .     |
 |                  |       | ---------------------- | ----- | ----- |
 |                  |       | df                     | 27    | 0     |
 | ---------------- | ----- | ---------------------- | ----- | ----- |
```


Exemple

```
Correlations
 | ---------------- | ----- | ---------------------- | ----- | ----- |
 | Control Variables |       |                        | Age   | Maths |
 | ---------------- | ----- | ---------------------- | ----- | ----- |
 | Sport            | Age   | Correlation            | 1,000 | ,699  |
 |                  |       | ---------------------- | ----- | ----- |
 |                  |       | Significance (2-tailed) | .     | ,000  |
 |                  |       | ---------------------- | ----- | ----- |
 |                  |       | df                     | 0     | 27    |
 |                  | ----- | ---------------------- | ----- | ----- |
 |                  | Maths | Correlation            | ,699  | 1,000 |
 |                  |       | ---------------------- | ----- | ----- |
 |                  |       | Significance (2-tailed) | ,000  | .     |
 |                  |       | ---------------------- | ----- | ----- |
 |                  |       | df                     | 27    | 0     |
 | ---------------- | ----- | ---------------------- | ----- | ----- |
```

Exemple

Correlations

Control Variables			Age	Sport
Maths	Age	Correlation	1,000	,348
		Significance (2-tailed)	.	,064
		df	0	27
	Sport	Correlation	,348	1,000
		Significance (2-tailed)	,064	.
		df	27	0

Exemple

On constate ainsi que les trois tests ne sont pas significatifs au seuil $\alpha = 5$ %.
On conserve l'hypothèse d'indépendance de deux variables sachant la troisième et ce dans les trois cas qui se présentent ici : $(\mathrm{Maths}, \mathrm{Sport})|\mathrm{Age}$, $(\mathrm{Maths}, \mathrm{Age})|\mathrm{Sport}$ et $(\mathrm{Sport}, \mathrm{Age})|\mathrm{Maths}$.

Exemple

On constate que les résultats à l'épreuve de sport et à celle de mathématiques sont indépendants si l'on tient compte de l'âge des élèves, ce qui est bien plus conforme à ce que l'on pouvait penser avant de réaliser cette expérience.

A. Boomsma.
Comparing approximations of confidence intervals for the product-moment correlation coefficient.
Statistica Neerlandica, 31 :179–186, 1977.

P. Chapouille.
Planification et analyse des expériences.
Masson, Paris, 1973.

R. Christensen.
Linear Models for Multivariate, Time Series, and Spatial Data.
Springer Texts in Statistics. Springer-Verlag, 1991.

P. Dagnélie.
Statistique Théorique et Appliquée, volume 2.
De Boeck & Larcier, Bruxelles, 1998.

J.-Y. Ouvrard.
Probabilités, volume 2.
Cassini, Paris, 2000.

E. Weisstein.
Correlation coefficient-bivariate normal distribution from mathworld–a wolfram web resource.
Lien internet : http ://mathworld.wolfram.com/Correlation-CoefficientBivariateNormalDistribution.html.