EPITECH
PROBABILITES ET STATISTIQUES
Cours207
Corrélation

Dominique Neveu

Année 2009-2010

Table des matières

Résumé du cours

Très souvent, il existe en pratique une relation entre deux variables ou plus. Il est souvent utile de pouvoir exprimer cette relation sous la forme d'une formule mathématique. Dans ce chapitre, on passe en revue les différentes méthodes qui permettent d'établir cette relation mathématique. La méthode la plus utilisée est la méthode des moindres carrés. Nous la présentons en détail.

Naturellement, il est nécessaire de prouver que la relation établie entre variables est juste. Pour cela, on dispose du coefficient de corrélation qui mesure la validité de l'ajustement réalisé. On donnera son expression dans le cas général et dans le cas particulier de la corrélation linéaire.

Chapitre 1

Corrélation

1.1 Relation entre variables

1.1.1 Position du problème

Très souvent, il existe une relation entre deux variables ou plus. Par exemple, le poids d'hommes adultes dépend dans une certaine mesure de leur taille, la circonférence d'un cercle dépend de son diamètre, et la pression d'un gaz de sa température et de son volume.
Le but sera donc d'estimer la valeur d'une des variables à l'aide des valeurs de l'autre (ou des autres).

Définition 1 *Soient X et Y deux variables. On cherche à expliquer Y en fonction de X. On dit alors que la variable estimée Y est dépendante. La variable X est dite indépendante.*

Il est souvent utile de pouvoir exprimer la relation entre X et Y sous la forme d'une formule mathématique qui relie les variables.

1.1.2 Données

Pour déterminer la formule mathématique qui relie les variables, il faut d'abord rassembler un ensemble de valeurs de ces variables. Par exemple, si X et Y sont la taille et le poids d'hommes adultes, on recueille sur un échantillon de N individus la valeur des tailles $X_1, X_2, ..., X_N$ et des poids $Y_1, Y_2, ..., Y_N$ correspondants.

1.1.3 Représentation graphique

L'étape suivante est la représentation graphique des données (X_1, Y_1), (X_2, Y_2), ..., (X_N, Y_N), dans un système de coordonnées rectangulaires. On obtient un nuage de points. Chaque point représente un couple de valeurs

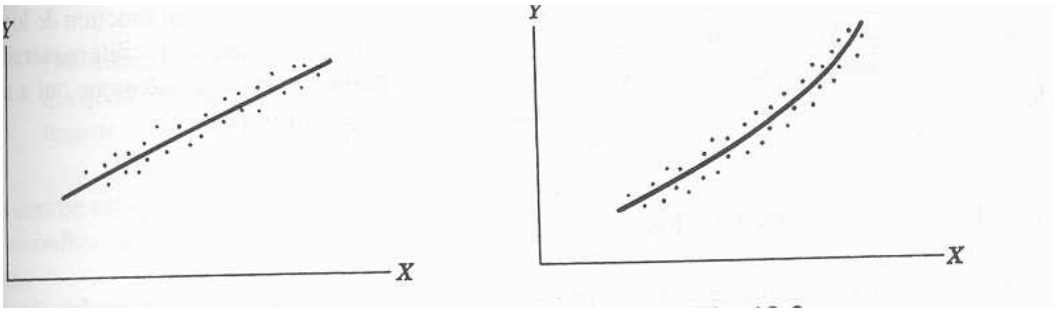

FIG. 1.1 – Courbes d'ajustement

observées de la variable dépendante et de la variable indépendante. Le graphe aide à déterminer s'il existe une relation entre les deux variables et le type d'équation approprié (linéaire, non linéaire).

Définition 2 *A partir de ce nuage de points, il est généralement possible de visualiser une courbe lissée qui ajuste les données. Cette courbe est dite d'**ajustement**.*

Dans le premier graphique de la figure 1.1, les données semblent bien approchées par une droite, et on dit alors qu'il existe une relation linéaire entre les variables. Au contraire, dans le second graphique, la relation qui existe entre les variables ne semble pas être linéaire.

Définition 3 *Le problème général de trouver quelle est la courbe qui s'approche au mieux d'un jeu donné de valeurs s'appelle un problème d'**ajustement**.*

1.1.4 Equations d'ajustement

La liste des courbes généralement utilisées dans les problèmes d'ajustement est donnée ci-dessous. Toutes les lettres autres que X et Y sont des constantes.

Droite

La droite ou polynôme de degré 1 :

$$Y = a + b.X$$

Parabole

La parabole ou courbe quadratique ou polynôme de degré 2 :

$$Y = a + b.X + c.X^2$$

Courbe cubique

La courbe cubique ou polynôme de degré 3 :

$$Y = a + b.X + c.X^2 + d.X^3$$

Polynôme de degré n

$$Y = a_0 + a_1.X + ... + a_n.X^n$$

Fonction hyperbole

$$Y = \frac{1}{a + b.X} \qquad \text{ou} \qquad \frac{1}{Y} = a + b.X$$

Fonction exponentielle

$$Y = Y = a.b^X \qquad \text{ou} \qquad log(Y) = log(a) + log(b).X$$

Courbe géométrique

$$Y = Y = a.X^b \qquad \text{ou} \qquad log(Y) = log(a) + b.log(X)$$

Pour déterminer la courbe à utiliser, on peut tracer les nuages de points pour certaines transformations des variables. Par exemple, si le nuage de points de log(Y) en fonction de X montre une relation linéaire, on pourra retenir l'exponentielle. Si log(Y) est linéaire en fonction de log(X), alors on utilisera une courbe géométrique.

1.1.5 Terme d'erreur

Admettons que l'on cherche à expliquer la variable Y par une relation linéaire avec la variable X. Dans une relation exacte, chaque couple de données (X_i, Y_i) pourrait alors s'écrire sous la forme :

$$Y_i = a + b.X_i$$

où a et b sont les deux constantes qui déterminent la relation linéaire. Ce type de relation exacte est très rare dans la pratique, car il existe des imprécisions dans les valeurs des couples (X_i, Y_i) (erreurs de mesure...). On voit donc que la relation $Y_i = a + b.X_i$ ne s'applique pas exactement et qu'il faut prendre en compte un **terme d'erreur**. La relation correcte s'écrit :

$$Y_i = a + b.X_i + \epsilon_i$$

où ϵ_i est le terme d'erreur qui prend en compte les petites variations aléatoires de X_i et Y_i.

De façon plus générale, lorsque la variable Y est expliquée en fonction de la variable X à l'aide d'une fonction d'ajustement f(.), on prend en compte un terme d'erreur. La relation entre les couples (X_i, Y_i) s'écrit alors :

$$Y_i = f(X_i) + \epsilon_i$$

1.2 Méthodes d'ajustement

1.2.1 Ajustement graphique

On peut utiliser son propre jugement pour tracer une courbe approchée des données. On appelle cette démarche ajustement à main levée. Si l'on connaît la forme de l'équation, on peut alors déduire la valeur des paramètres en prenant les valeurs d'autant de points qu'il y a de constantes. Cette méthode présente le lourd désavantage de dépendre de la personne qui fera l'ajustement.

1.2.2 Méthode des moindres carrés

La méthode des moindres carrés est une des plus utilisées pour calculer les courbes d'ajustement.

Dans la méthode des moindres carrés, on considère la somme des carrés des écarts ϵ_i et on cherche à minimiser cette somme, minimisant ainsi globalement la valeur des écarts. On désigne par Y = f(X) l'équation de la courbe d'ajustement, la fonction f dépendant d'un certain nombre de paramètres recherchés notés $(a_p)_{p=1,...,P}$. Par exemple, pour un ajustement linéaire, la fonction f s'écrira f(X) = a + b.X, avec les paramètres a et b.

La méthode des moindres carrés revient à minimiser la somme suivante :

$$\min_{(a_p)} \sum_{i=1}^{N} [Y_i - f(X_i)]^2 = \min_{(a_p)} \sum_{i=1}^{N} \epsilon_i^2 = \min_{(a_p)} [\epsilon_1^2 + \epsilon_2^2 + ... + \epsilon_N^2]$$

Cette condition permet souvent de calculer les paramètres recherchés (a_p), et donc de déterminer ainsi la fonction d'ajustement f. La courbe obtenue est dite courbe des moindres carrés ou approximation des données au sens des moindres carrés. On parlera donc de droite des moindres carrés, parabole des moindres carrés,...

1.2.3 Droite des moindres carrés

La droite des moindres carrés qui approche le nuage de points (X_1, Y_1), (X_2, Y_2), ...,(X_N, Y_N) a pour équation :

$$Y = a + b.X$$

On peut montrer que la condition de minimum des moindres carrés est équivalente au système de deux équations suivantes :

$$\begin{cases} \sum_{i=1}^{N} Y_i = a.N + b.\sum_{i=1}^{N} X_i \\ \sum_{i=1}^{N} X_i.Y_i = a.\sum_{i=1}^{N} X_i + b.\sum_{i=1}^{N} X_i^2 \end{cases}$$

qui sont les équations de la droite des moindres carrés. Les paramètres a et b sont donc calculés à l'aide des formules suivantes :

$$a = \frac{(\sum Y).(\sum X^2) - (\sum X).(\sum XY)}{N.(\sum X^2) - (\sum X)^2}$$

$$b = \frac{N.(\sum XY) - (\sum X).(\sum Y)}{N.(\sum X^2) - (\sum X)^2}$$

1.2.4 Parabole des moindres carrés

La parabole des moindres carrés qui approche le nuage de points (X_1, Y_1), (X_2, Y_2), ...,(X_N, Y_N) a pour équation :

$$Y = a + b.X + c.X^2$$

On peut montrer que la condition de minimum des moindres carrés est équivalente au système de trois équations suivantes :

$$\begin{cases} \sum Y & = a.N & +b.\sum X & +c.\sum X^2 \\ \sum XY & = a.\sum X & +b.\sum X^2 & +c.\sum X^3 \\ \sum X^2Y & = a.\sum X^2 & +b.\sum X^3 & +c.\sum X^4 \end{cases}$$

appelées équations de la parabole des moindres carrés. On obtient les paramètres a, b et c en inversant ce système linéaire de trois équations à trois inconnues.
On peut également étendre cette technique pour obtenir les équations qui déterminent la courbe des moindres carrés pour un polynôme de degré quelconque.

1.2.5 Ajustement non linéaire par moindres carrés

Il peut arriver que les points représentant une série double ne soient pas alignés, mais soient voisins d'une courbe connue. On se sert alors en général de la méthode des moindres carrés, mais en transformant au préalable une des variables. Ainsi un ajustement linéaire entre Y et X^n donne un ajustement de la forme Y = a + b.X^n ; un ajustement entre Y et log(X) donne Y = a + b.log(X).

1.3 Théorie de la corrélation

Dans ce paragraphe, nous abordons le problème précis de la corrélation dans lequel on cherche à déterminer jusqu'à quel point une équation décrit la relation entre les variables.

1.3.1 Niveau de corrélation

Si toutes les valeurs des variables satisfont exactement à l'équation, on dit que les variables sont **parfaitement corrélées** ou qu'il y a une corrélation parfaite entre elles. Ainsi les circonférences C et les rayons r de tous les cercles sont parfaitement corrélés car C = 2πr. Lorsqu'on lance un dé plusieurs fois de suite, il n'y a pas de relation entre les points (sauf si le dé est pipé), on dit que les points ne sont **pas corrélés**. Des variables telles que le poids et la taille des individus pourraient montrer **quelque corrélation**.

1.3.2 Ecart-type de l'estimation

Définition 4 *On note f(.) la fonction d'ajustement choisie. On appelle écart-type de l'estimation de Y sur X, la quantité suivante :*

$$s_{Y,X} = \sqrt{\frac{\sum (Y - f(X))^2}{N}}$$

C'est une mesure de la dispersion autour de la courbe d'ajustement.

On a vu que la relation entre X et Y comporte un terme d'erreur noté ϵ, on a :

$$Y = f(X) + \epsilon$$

Ainsi, l'expression de l'écart-type de l'estimation s'écrit plus simplement :

$$s_{Y,X} = \sqrt{\frac{\sum \epsilon^2}{N}}$$

L'écart-type de l'estimation a des propriétés analogues à celle de l'écart-type. Par exemple, si on construit des courbes parallèles à la courbe d'ajustement

à des distances respectives de $s_{Y,X}$, $2s_{Y,X}$, $3s_{Y,X}$, on devrait trouver, si N est assez grand, que 68%, 95% et 99,7% des points d'échantillonnage sont inclus entre ces courbes.

1.3.3 Variation expliquée et non expliquée

Définition 5 *La variation totale de Y est définie par la formule :*

$$\sum (Y - \overline{Y})^2$$

C'est la somme des carrés des écarts entre les valeurs de Y et sa moyenne.

Cette variation totale se décompose en deux sommes :

$$\sum (Y - \overline{Y})^2 = \sum (Y - f(X))^2 + \sum (f(X) - \overline{Y})^2$$

Le premier terme $\sum (Y - f(X))^2$ est appelé variation inexpliquée. Le deuxième terme $\sum (f(X) - \overline{Y})^2$ est appelé variation expliquée.
Les écarts de variation expliquée $(f(X) - \overline{Y})$ ont une forme définie, tandis que les écarts de variation inexpliquée $(Y - f(X))$ ont une distribution aléatoire.

1.3.4 Coefficient de corrélation

Le rapport de la variation expliquée sur la variation totale s'appelle le coefficient de détermination. Ce rapport, toujours positif, est noté r^2.

Définition 6 *La quantité r, appelée le coefficient de corrélation, est donnée par :*

$$r = \sqrt{\frac{\sum (f(X) - \overline{Y})^2}{\sum (Y - \overline{Y})^2}}$$

S'il n'y a aucune variation expliquée, ce rapport est égal à 0. S'il n'y a pas de variation inexpliquée, ce rapport est égal à 1. Dans d'autres cas, ce rapport est compris entre 0 et 1. Plus la valeur de r est proche de 1, plus l'ajustement est de bonne qualité. On remarque que r est une quantité sans dimension, c'est à dire qui ne dépend pas de l'unité employée.

1.3.5 Lien avec l'écart-type de l'estimation

L'écart-type de Y est noté s_Y :

$$s_Y = \sqrt{\frac{\sum (Y - \overline{Y})^2}{N}}$$

On peut démontrer que la formule donnant la valeur du coefficient de corrélation r peut aussi s'écrire sous la forme suivante :

$$r = \sqrt{1 - \frac{s_{Y,X}^2}{s_Y^2}}$$

Il faut noter le fait que la valeur de r mesure seulement l'intensité de la liaison avec le type d'équation utilisée. Ainsi, si on utilise une équation linéaire et si la valeur de r est presque nulle, on peut dire tout au plus qu'il n'y a pas de corrélation linéaire entre les variables. Cela ne veut pas dire qu'il n'y a pas de corrélation, car il peut y avoir une forte corrélation non linéaire entre les variables.

1.3.6 Coefficient de corrélation linéaire

Soient X et Y deux variables d'écarts-types s_X et s_Y :

$$s_X = \sqrt{\frac{\sum (X - \overline{X})^2}{N}} \qquad s_Y = \sqrt{\frac{\sum (Y - \overline{Y})^2}{N}}$$

Définition 7 *On appelle covariance de deux variables X et Y, et on note s_{XY}, la quantité suivante :*

$$s_{XY} = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{N}$$

On pose : x = X - $\overline{X}$ et y = Y - $\overline{Y}$. Les formules s'écrivent alors en fonction de x et y :

$$s_X = \sqrt{\frac{\sum x^2}{N}} \qquad s_Y = \sqrt{\frac{\sum y^2}{N}} \qquad s_{XY} = \frac{\sum xy}{N}$$

Si on suppose qu'il existe une relation linéaire entre deux variables, l'équation donnant le coefficient de corrélation peut s'écrire sous la forme :

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)}.\sqrt{(\sum y^2)}}$$

Soit, en fonction de la covariance et des écarts-types :

$$r = \frac{s_{XY}}{s_X.s_Y}$$