Couple de variables aléatoires - Notion d'indépendance.

Préparation au Capes - Université Rennes 1

On considère deux variables aléatoires X et Y. On aimerait connaitre s'il y a influence entre ces deux variables et la quantifier.

Exemple : On peut se poser la question de l'influence des catastrophes météorologiques (tempêtes, ouragans, tsunamis, ...) sur le cours de la bourse. La variable X modélisera alors les catastrophes météorologiques et la variable Y le cours de la bourse.

Hélène Guérin, helene.guerin@univ-rennes1.fr

I - Loi jointe

Définition :

Soient X et Y deux variables aléatoires. La **loi jointe** de (X, Y) est définie par sa **fonction de répartition** $F_{(X,Y)}$:

$$\begin{array}{rcl} F_{(X,Y)} : \mathbb{R} \times \mathbb{R} & \rightarrow & [0, 1] \\ (x, y) & \rightarrow & P(X \leq x, Y \leq y) \end{array}$$

Les **lois marginales** du couple (X, Y) sont la loi $\mathcal{L}(X)$ de X, et la loi $\mathcal{L}(Y)$ de Y.

Attention ! À partir des lois marginales, on ne peut pas connaître la loi du couple.

a - Cas des variables discrètes

Soient X et Y deux variables discrètes, X à valeurs dans $\mathbb{D}_X$ et Y à valeurs dans $\mathbb{D}_Y$. La loi du couple (X, Y) est définie par l'ensemble des probabilités :

$$P(X = x, Y = y) \quad \text{avec } x \in \mathbb{D}_X \text{ et } y \in \mathbb{D}_Y.$$

Dans le cas où les variables sont discrètes et prennent un petit nombre de valeurs, on écrit en général la loi du couple sous la forme d'un tableau :

$Y \backslash X$	$\ldots$	Somme des colonnes
$\vdots$	$P(X = x, Y = y)$	$P(Y = y)$
Somme des lignes	$P(X = x)$	

Exemple 1

1. L'université de Rennes 1 veut évaluer l'effet de l'offre MIPE sur le campus et voir quel système d'exploitation est apprécié des étudiants. Les proportions collectées sont résumées dans un tableau :

Système d'exploitation / Filière	Windows	Mac OS	Linux
Biologie	0.07	0.05	0.02
Droit/Économie	0.08	0.02	0
Informatique	0.25	0.13	0.09
Mathématiques	0.21	0.04	0.04

2. On lance une pièce truquée 3 fois. La probabilité de tomber sur "Pile" est 2/3. Soit X le nombre de "Face" obtenu dans les deux premiers jets et Y le nombre de "Face" obtenu dans les deux derniers jets. Donner la loi de (X,Y) ! !

b- Cas des variables à densité

Définition : La loi du couple de v.a. (X, Y) est dite **à densité** s'il existe une fonction $f_{(X,Y)}$ telle que la fonction de répartition du couple s'écrit

$$F_{(X,Y)}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{(X,Y)}(u, v) dudu,$$

satisfaisant les conditions suivantes :

1. $f_{(X,Y)}(x, y) \geq 0$ pour tout $(x, y) \in \mathbb{R}^2$,
2. $\int \int_{R^2} f_{(X,Y)}(x, y) dxdy = 1$.

On retrouve facilement les lois marginales : les variables X et Y sont des variables continues de densité respectives $f_X(x) = \int_R f_{(X,Y)}(x, y) dy$ et $f_Y(y) = \int_R f_{(X,Y)}(x, y) dx$.

II - Variables aléatoires indépendantes

Définition Deux variables aléatoires X et Y sont indépendantes si pour tout intervalle A et B de $\mathbb{R}$ on a

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B).$$

Proposition

Deux v.a. X et Y sont indépendantes

$\Leftrightarrow$ La fonction de répartition du couple vérifie $\forall(x, y)$

$$F_{(X,Y)}(x, y) = F_X(x)F_Y(y),$$

$\Leftrightarrow$ La transformée de Laplace du couple vérifie pour tout (u, v),

$$L_{(X,Y)}(u, v) = L_X(u)L_Y(v) \quad \text{où } L_{(X,Y)}(u, v) = E[e^{uX+vY}],$$

$\Leftrightarrow$ Dans le cas discret $\forall(x, y)$,

$$P(X = x, Y = y) = P(X = x)P(Y = y),$$

$\Leftrightarrow$ Dans le cas continu, la densité du couple vérifie $\forall(x, y)$

$$f_{(X,Y)}(x, y) = f_X(x)f_Y(y).$$

Définition Les variables aléatoires $X_1, \ldots, X_n$ sont **mutuellement indépendantes** si pour tous intervalles $A_1, \ldots, A_n$ de $\mathbb{R}$ on a

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} P(X_i \in A_i).$$

Si les variables $X_1, \ldots, X_n$ sont mutuellements indépendantes, alors elles sont indépendantes deux à deux.

Attention La réciproque est fausse !

Exemple 2 Une loi binomiale $\mathcal{B}(n,p)$ de paramètres n et p correspond au modèle suivant :

On renouvelle n fois de manière indépendante une épreuve de Bernoulli de paramètre p. On compte le nombre de succès obtenus à l'issue des n épreuves et on appelle S la variable aléatoire correspondant à ce nombre de succès. Donc si on note X_i le résultat du i$^{\text{ème}}$ lancer, les variables $X_1, \ldots, X_n$ sont des variables de Bernoulli de paramètre p, indépendantes. On a $S = X_1 + \ldots + X_n$.

On en déduit facilement que si $S_1 \sim \mathcal{B}(n_1, p)$ et $S_2 \sim \mathcal{B}(n_2, p)$ indépendantes, alors $S_1 + S_2 \sim \mathcal{B}(n_1 + n_2, p)$.

III - La covariance

La covariance permet d'estimer la dépendance entre deux variables aléatoires.

Définition La **covariance** de deux variables X et Y est

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

L'espérance $E(XY)$ est calculée à partir de la loi jointe de (X,Y) :

1. dans le cas discret,

$$E(XY) = \sum_{x \in \mathbb{D}_X, y \in \mathbb{D}_Y} xyP(X = x, Y = y)$$

2. dans le cas continu,

$$E(XY) = \int\int_{R^2} xyf_{(X,Y)}(x,y)dxdy.$$

Remarque Soient X et Y deux v.a. Alors $Var(X+Y) = Var(X) + Var(Y) + 2\,Cov(X,Y)$.

Définition Le **coefficient de corrélation linéaire** est défini pour des variables non constantes par :

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

On a toujours $\rho(X,Y) \in [-1,1]$.

Plus $|\rho(X,Y)|$ est proche de 1 plus la dépendance entre les variables X et Y est forte.

Remarque Si X et Y sont indépendantes, alors $Cov(X,Y) = 0$ et donc $\rho(X,Y) = 0$. On a par conséquent,

$$Var(X+Y) = Var(X) + Var(Y).$$

Attention La réciproque est fausse !

Exemple 3 Une étude médicale sur l'effet du tabac est menée dans un hopital. Les 2278 patients sont divisés en deux groupes : ceux atteints d'un cancer pulmonaire ($X = 1$) et les autres ($X = 0$). Les membres de chaque groupe sont ensuite répartis selon le nombre Y de paquets de cigarettes fumés par jour.

Cancer	Nombre de paquets de cigarettes					Total
pulmonaire	0	1	2	3	4	
0	1247	492	319	58	9	2125
1	66	50	28	6	3	153
Total	1313	542	347	64	12	2278

On souhaite étudier l'association entre cancer pulmonaire et la consommation de cigarette en calculant la covariance.