Chapitre 2
LES NUAGES DE POINTS

2.1 Introduction

Nous consacrons ce chapitre aux variables numériques. Lorsque la variable est continue, il est souvent commode de raisonner en supposant que les valeurs rencontrées sont toutes différentes (éventuellement infiniment voisines mais différentes).

Soit N le cardinal d'une population dont les valeurs observées X_k de la variable sont toutes différentes. La fréquence absolue de la valeur X_k est alors 1 tandis que sa fréquence relative est $1/N$.

Chaque observation est représentée par un point, M_k, d'abscisse X_k. Une statistique à une dimension apparaît comme un nuage de N points répartis sur un axe. Chaque point M_k représente un individu. Tous les points sont affectés de la même masse $m_k = 1/N$ (voir l'annexe page 36).

La moyenne $\overline{X}$, est l'abscisse du centre de masse, G, du nuage de points :

$$\overline{X} = \sum_k \frac{1}{N} X_k = \sum_k \frac{m_k}{m} X_k$$

Le moment d'inertie, du nuage de points par rapport à G est

$$\sum_k m_k \ (GM_k)^2 = \sum_k \frac{1}{N} \ \left(X_k - \overline{X}\right)^2 = \sigma^2$$

c'est le carré de l'écart quadratique moyen, σ.

Considérons une population dont chaque individu est décrit par deux variables X et Y. La population concernée peut être par exemple l'ensemble des ménages hétérosexuels ; la variable X est l'âge de la femme, la variable Y est le nombre d'enfants.

Pour représenter les observations, on utilise un repère orthonormé à deux dimensions. Les unités sur chaque axe sont arbitraires. Un individu est représenté par un point, M, du plan précédent. La variable X est l'abscisse de M, tandis que Y en est l'ordonnée.

L'âge, X, est une variable continue, on peut donc admettre qu'aucun des individus observés n'a le même âge. Par conséquent, même si les nombres d'enfants sont les mêmes, deux individus seront représentés par deux points distincts.

Ce sont de tels nuages de points que nous allons étudier dans ce chapitre.

2.2 Ajustement à un modèle

On peut supposer qu'il existe une relation fonctionnelle de X vers Y, c'est-à-dire que Y prend une valeur et une seule, connue lorsque X est connu (voir la liaison fonctionnelle page 2). Par exemple, on peut tracer point par point le graphe de l'équation horaire d'un mouvement rectiligne. La variable X est la date, t, de l'observation et la

variable Y est l'abscisse, y, du mobile à la date t. La relation $t \mapsto y$ est supposée être une relation fonctionnelle dont nous voulons obtenir le graphe. Admettons que pour des raisons théoriques $y(t)$ soit de la forme $y(t) = a\sin(\omega t)$ sans que nous ne connaissions les coefficients a et ω. Pour ajuster au mieux les observations et la théorie, nous devons disposer d'un indicateur qui apprécie la "distance" entre observation et théorie. Pour cerner les propriétés d'un tel indicateur, donnons nous la courbe théorique, $Y = y(X)$ et un ensemble de points (X_k, Y_k) supposés décrire les observations. L'indicateur doit être 1) un nombre positif, 2) sa nullité doit impliquer que l'ajustement est parfait et 3) il doit croître si on dégrade l'accord entre théorie et expérience en modifiant l'ordonnée d'un point expérimental par exemple.

De nombreux indicateurs peuvent être construits. Parmi ceux-ci,

$$R = \sum_k \left(Y_k - y(X_k)\right)^2 \tag{2.1}$$

On vérifie aisément les trois conditions imposées.

Pour comprendre la signification de R, reprenons l'exemple d'une relation de la forme $y(t) = a\sin(\omega t)$. Toutes les valeurs de t sont possibles, en particulier les valeurs $t_k = X_k$. On ne peut donc prétendre qu'il y aurait erreur à considérer que la date X_k est une date possible. Par contre, l'attribution que nous faisons de donner à $y(X_k)$ la valeur mesurée Y_k est entachée d'une double erreur. En effet $y(t_k) = y(X_k + \delta X_k)$ où δX_k est l'erreur que l'on commet en posant $X_k = t_k$ au moment où l'on mesure $y = Y_k$. En outre, la mesure de y est elle même entachée d'une erreur : $Y_k = y(t_k) + \delta Y_k$. On en déduit $Y_k = y(X_k + \delta X_k) + \delta Y_k$. Admettons que les erreurs soient "petites". Cela signifie que l'on peut remplacer $y(X_k + \delta X_k)$ par un développement du premier ordre : $y(X_k + \delta X_k) \simeq y(X_k) + \dot{y}(X_k) \times \delta X_k$ où $\dot{y}$ est la dérivée de y. On en déduit

$$Y_k = y(X_k) + \varepsilon_k \text{ avec } \varepsilon_k = \dot{y}(X_k) \times \delta X_k + \delta Y_k$$

ε_k est l'erreur (inconnue) sur la mesure de $y(X_k)$. L'indicateur R s'écrit encore sous la forme $R = \sum_k (\varepsilon_k)^2$.

Chercher le meilleur ajustement, c'est chercher la valeur des paramètres a et ω qui minimisent la "distance" de la théorie à l'expérience, c'est-à-dire qui minimisent R.

De façon pratique, on opère de la façon suivante.

On forme R donné par la relation 2.1. Dans le cas particulier considéré, il vient $R = \sum_k \left(Y_k - a\sin(\omega X_k)\right)^2$. Les valeurs de X_k et de Y_k sont connues (ce sont les résultats expérimentaux), par contre a et ω sont inconnus. L'indicateur R est donc une fonction connue des deux variables inconnues a et ω. On pose $R = R(a, \omega)$.

Nous cherchons à déterminer a et ω de telle sorte que R soit minimum. On pose donc les deux équations

$$\frac{\partial R}{\partial a} = 0 \quad \text{et} \quad \frac{\partial R}{\partial \omega} = 0 \tag{2.2}$$

Ces deux équations permettent de déterminer les valeurs numériques de a et ω. Les conditions imposées à R impliquent alors que ces valeurs correspondent à un minimum de R.

La méthode se généralise immédiatement au cas où l'ajustement porte sur un nombre arbitraire de paramètres.

Remarquons que les points expérimentaux étant donnés, la valeur des paramètres théoriques que l'on obtient (a et ω) est modifiée si on utilise un indicateur différent de R.

La méthode décrite ici est appelée ***méthode des moindres carrés*** (c'est la recherche de la moindre valeur de R). Cette méthode permet de proposer des valeurs acceptables pour les divers paramètres théoriques.

Remarquons que les procédés d'ajustement sont très importants car ils permettent aussi bien la vérification d'une théorie que la mesure d'une grandeur.

2.3 Régression linéaire

Bien souvent la théorie est inconnue. Les nuages de points se présentent sous une forme "indescriptible", sans structure apparente. Dans ces conditions, les questions qui se posent sont très élémentaires et les ajustement recherchés sont les ajustements linéaires, les plus simples possibles.

Considérons donc un nuage à deux dimensions, X et Y, formé de N points. Nous posons la question de la "meilleure" droite, susceptible de décrire la relation fonctionnelle $X \mapsto Y$. Posons l'équation de cette droite sous la forme $y = ax + b$. Suivant la méthode générale décrite précédemment nous construisons $R(a,b) = \sum_k (Y_k - aX_k - b)^2$. Nous formons les équations 2.2 :

$$\frac{\partial R}{\partial a} = -2\sum_k X_k (Y_k - aX_k - b) = 0 \quad \text{et} \quad \frac{\partial R}{\partial b} = -2\sum_k (Y_k - aX_k - b) = 0$$

Nous rappelons les définitions suivantes

$$\frac{1}{N}\sum_k X_k = \overline{X}\ , \quad \frac{1}{N}\sum_k Y_k = \overline{Y}\ , \quad \frac{1}{N}\sum_k (X_k)^2 = \overline{X^2} \text{ et } \quad \frac{1}{N}\sum_k X_k Y_k = \overline{XY}$$

Il vient

$$\frac{-1}{2N}\frac{\partial R}{\partial a} = \overline{XY} - a\overline{X^2} - b\overline{X} = 0 \text{ et } \quad \frac{-1}{2N}\frac{\partial R}{\partial b} = \overline{Y} - a\overline{X} - b = 0$$

On en déduit

$$\boxed{a = \frac{\overline{XY} - \overline{X}\cdot\overline{Y}}{\overline{X^2} - \overline{X}^2} = \frac{cov\,[X,Y]}{\sigma_X^2}} \text{ et } \quad \boxed{b = \overline{Y} - a\overline{X}}$$

La droite ainsi construite est la ***droite de régression*** de Y en X(figure 2-1)

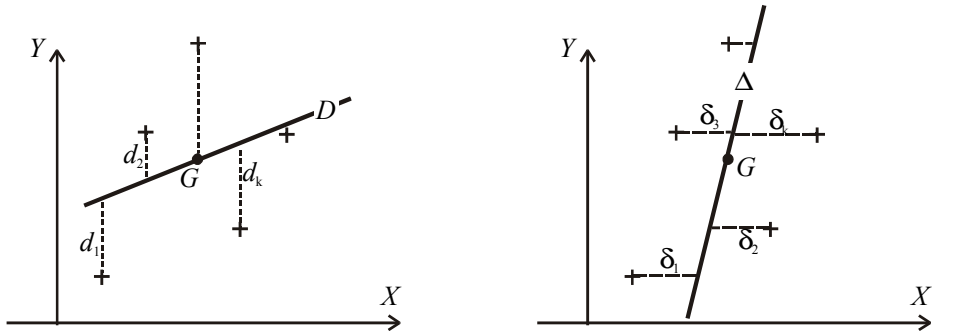
Le mot régression vient d'une étude effectuée par *Francis Galton* (1822-1911) dans la seconde moitié du XIX^{ème} siècle, étude où il montrait que la taille des parents et celle des enfants étaient corrélées mais que la taille des enfants subissait une "régression" vers la normale lorsque les parents étaient de grande taille.

La régression de Y en X ne considère pas les deux variables X et Y de la même manière. Y est considérée comme une fonction de X. La valeur de X est la cause de la valeur observée de Y.

On peut inverser les variables X et Y et obtenir la droite de régression de X en Y (figure 2-1). Celle-ci a pour équation $X = a'Y + b'$ avec

$$a' = \frac{cov\,[X,Y]}{\sigma_Y^2} \text{ et } \quad b' = \overline{X} - a'\overline{Y}$$

Remarquons que **de façon générale les droites de régression passent par le point moyen, point de coordonnées $\overline{X}$ et $\overline{Y}$.**



Droite de régression linéaire de Y en X : D telle que Σd_k^2 *minimum*

Droite de régression linéaire de X en Y : Δ *telle que* $\Sigma \delta_k^2$ *minimum*

Figure 2-1

Nous savons qu'il y a une relation entre le nombre de stages en entreprises et la facilité d'une première embauche. On peut cependant se poser la question de savoir si cette relation dépend du type d'études effectuées. Les tableaux ci-dessous concernent deux populations de dix personnes à qui l'on a demandé le nombre de mois de stages effectués avant leur première embauche (variable X) et le nombre de mois (variable Y) qui sépare la fin de leurs études (ou de leur dernier stage si celui-ci est effectué après la fin des études) de leur première embauche. La population A concerne des étudiants d'une école de commerce, la population B, des étudiants d'une filière scientifique universitaire. Nous avons additionné tous les mois de stage pour construire la variable X, bien que ceux-ci puissent être de natures très variées* : stages optionnels, ou obligatoires, stage de vacances, stage de pré-embauche, stage ANPE, etc..

A : Ecole de commerce

X(en mois) =	0	2	3	3	4	6	7	7	8	9
Y(en mois) =	1	5	0	2	2	6	0	7	2	5

Tableau 1

B : Filière universitaire

X(en mois) =	0	1	2	4	6	6	7	8	8	10
Y(en mois) =	3	9	8	6	2	3	4	6	3	0

Tableau 2

On obtient aisément les valeurs suivantes :

	$\overline{X}$	$\overline{Y}$	$\overline{XY}$	$\overline{X^2}$	$\overline{Y^2}$	$cov\,[X,Y]$	σ_X^2	σ_Y^2	a	a'
A :	4,9	3,0	17	31,7	14,8	2,3	7,69	5,8	0,3	0,4
B :	5,2	4,4	17,9	37,0	26,4	$-4,98$	9,96	7,04	$-0,5$	$-0,7$

où $\overline{X}$, $\overline{Y}$, σ_X et σ_Y sont exprimés en mois tandis que $\overline{XY}$, $\overline{X^2}$, $\overline{Y^2}$ et $cov\,[X,Y]$ sont exprimés en (mois)^2. Les coefficients a et a' sont des nombres purs, sans dimension.

Les nuages de points et les droites de régression sont représentés sur la figure 2-2.

*Ceci explique les faibles corrélations observées (voir la section corrélation page 31). Il faut donc se garder d'interpréter de façon définitive les résultats obtenus ici.

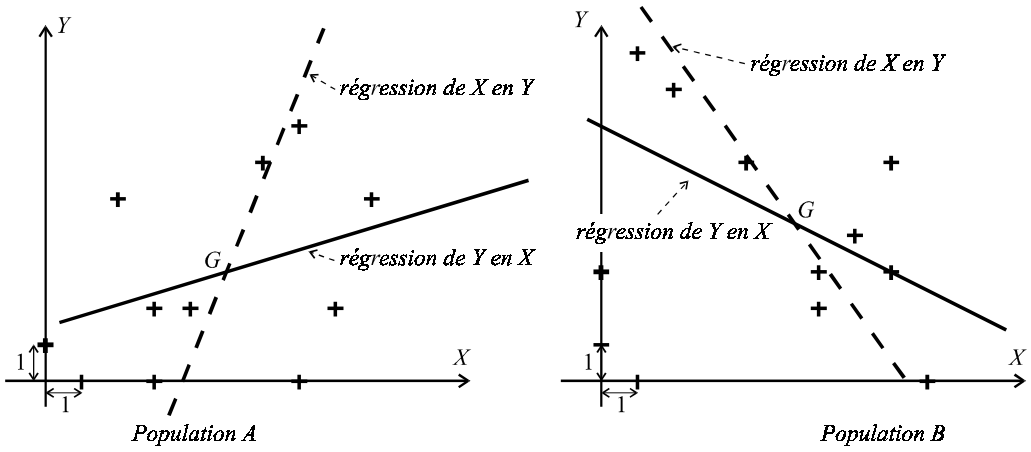


Figure 2-2

On peut se poser la question de savoir si le temps d'attente à la première embauche ne serait pas une fonction de la durée des stages ; tracer la droite de régression de Y en X a donc un sens. Lorsque le marché du travail se dégrade, le temps de recherche d'un premier emploi s'allonge. On peut penser que cette situation favorise la multiplicité des stages. De ce point de vue la régression de X en Y prend son sens.

Que peut-on déduire de ces études de régression (de Y en X)? Dans la population A, les stages apparaissent comme un handicap alors que dans la population B, au contraire, l'augmentation de la durée des stages pendant les études est un élément qui favorise le recrutement rapide des étudiants.

Ce que nous indique l'analyse c'est **une tendance** et seulement une tendance car aucune loi "rigide", qui serait vraie sans exception, ne se dégage. Bien que cette étude soit déjà vieille de plusieurs années, on peut néanmoins chercher une explication à cette "tendance" (voir la note page 24).

Les employeurs connaissent l'école A; ils savent comment l'étudiant est sélectionné et formé. Dans une certaine mesure, le "produit" est uniforme toujours le même et il est apprécié. "*Un stage ailleurs que chez moi est une perte de temps*" pense le recruteur potentiel dont l'intérêt consiste à recruter des étudiants aussi jeunes et malléables que possible. Cette "explication" permet de comprendre pourquoi les stages ne sont pas particulièrement prisés dans l'école A.

L'Université, B, forme des étudiants de niveau, de maturité et de caractère très divers. Des matières fondamentales y sont souvent étudiées. Dans les filières scientifiques les notions de "rendement", de "profit", de "management" ne sont pas au coeur des préoccupations de l'enseignement. Les stages rassurent le recruteur potentiel ; ils apportent la garantie d'une formation restée au contact des réalités de l'entreprise. C'est un élément positif qui favorise le recrutement.

2.4 Description d'un nuage de points

Dans l'étude de la régression de Y en X, on trouve l'idée plus ou moins explicite d'une relation fonctionnelle, voire d'une relation de cause à effet de X à Y. Il est cependant des cas où aucune des deux variables ne se distingue. Prenons pour exemple la ville de Châteauroux divers jours de l'année. En cette ville, il se vend quotidiennement des paires de lunettes de Soleil (variable X) et des kilogrammes de glaces (variable Y). Nul ne pense que la consommation de glaces favorise l'achat de lunettes de Soleil, on ne peut pas croire, non plus, que l'utilisation de lunettes de Soleil provoque une envie irrépressible de glaces. Pourtant ce sont les jours où l'on vend beaucoup de lunettes de Soleil que l'on consomme le

plus de glaces. Bien qu'il n'y ait pas de relation de cause à effet, il existe une ***corrélation*** entre X et Y : les deux variables ne sont pas indépendantes. Cette corrélation dépend d'un facteur extérieur à l'étude : l'ensoleillement.

Pour étudier un nuage de point sans distinguer l'une des variables X ou Y, on pose trois questions :

1. Si le nuage devait être décrit par un seul point, lequel choisirait-on ? Cette question concerne la "position" de la statistique dans le plan X, Y.
2. S'il fallait réduire le nuage à un ensemble de points portés par la même droite, quelle droite choisirait-on ? Cette question concerne la "meilleure" relation linéaire que l'on peut supposer entre X et Y, sans préjuger d'une quelconque causalité. La réponse à cette question donne une "tendance".
3. La "meilleure" relation linéaire étant obtenue, cette relation est-elle floue ou fortement marquée ? La réponse à cette question est fournie par le coefficient de corrélation linéaire.

La première opération à effectuer est de modifier les données de façon à ce qu'elles décrivent des nombres sans dimension physique. Pour cela on effectue la substitution $X \to x = X/U_X$ et $Y \to y = Y/U_Y$ où U_X et U_Y sont des grandeurs de référence de mêmes dimensions que X et Y respectivement. Comment sont choisis U_X et U_Y ? Les unités usuelles peuvent être employées mais elles sont rarement adaptées au problème considéré. Le choix peut être effectué en référence à une donnée extérieure à la statistique (un âge moyen de mortalité pour U_X et le SMIC pour U_Y si l'on étudie la relation entre âge et revenu par exemple) ; ce peut être aussi en référence à la statistique elle-même (voir ci-dessous le paragraphe 2.5.1 page 31).

2.4.1 Position d'un nuage de points

Soit $M(x, y)$ un point du plan. Nous construisons l'indicateur de dispersion par rapport à M que nous notons $\Psi(x, y)$:

$$\Psi(x, y) = \frac{1}{N}\sum_k (\delta_k)^2 \quad \text{avec} \quad \delta_k = \sqrt{(x_k - x)^2 + (y_k - y)^2}$$

δ_k est la distance du point M_k du nuage au point M, tandis que N est le nombre de points du nuage. Nous sommes amenés ici à additionner x_k^2 et y_k^2, il est donc essentiel que ces deux grandeurs soient de mêmes dimensions physiques, le plus simple étant de construire des nombres sans dimensions (ce que nous avons fait).

En affectant chaque point du nuage de la même masse $1/N$, l'indicateur $\Psi(x, y)$ représente l'inertie du nuage par rapport à M.

On vérifie les propriétés qui confèrent à Ψ le statut d'indicateur de dispersion :

• Ψ est positif ou nul ; $\Psi = 0$ signifie que tous les points du nuage sont confondus en M (absence de dispersion)

• Ψ croît lorsqu'on éloigne de M un point du nuage (c'est-à-dire lorsqu'on augmente la dispersion)

Le "meilleur" point M est le point G par rapport auquel la dispersion est minimale (c'est la définition que nous donnons du mot "meilleur"). Les coordonnées (x_G, y_G) du point G, sont solutions des équations :

$$\frac{\partial}{\partial x}\Psi = \frac{2}{N}\sum_k (x_k - x) = 0 \text{ et } \frac{\partial}{\partial y}\Psi = \frac{2}{N}\sum_k (y_k - y) = 0$$

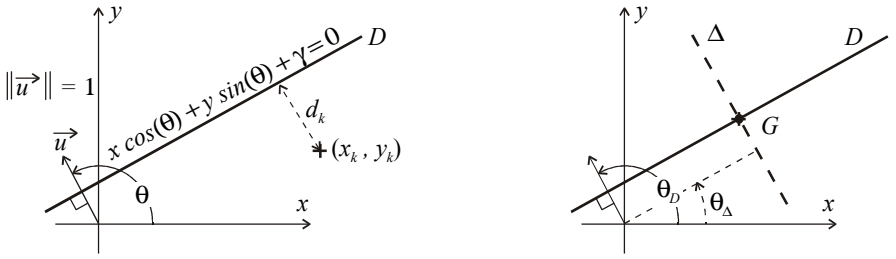
On en déduit que G est le "point moyen" de coordonnées $\boxed{x_G = \overline{x} \text{ et } y_G = \overline{y}}$. La dispersion par rapport à G se calcule facilement ; on trouve $\Psi_G = \frac{1}{N}\sum_k (x_k - \overline{x})^2 + \frac{1}{N}\sum_k (y_k - \overline{y})^2$ soit $\boxed{\Psi_G = \sigma_x^{\,2} + \sigma_y^{\,2}}$

Remarquons que le "meilleur" point est modifié si on choisit un indicateur de dispersion autre que $\Psi(x, y)$. De façon générale, les résultats dépendent des indicateurs choisis. Par la suite, nous ne rappelons pas cette propriété, tandis que nous utilisons les indicateurs usuels sans discussion.

La position de la statistique étant donnée par le point G, nous cherchons la "meilleure" relation linéaire entre X_k et Y_k.

2.4.2 Orientation d'un nuage de points

Entre x_k et y_k, nous cherchons une relation linéaire de la forme $x\cos\theta + y\sin\theta + \gamma = 0$. Une telle relation est l'équation d'une droite : la "meilleure" droite que nous désignons par la lettre D.



a) cas générique.

b) la meilleure droite (D) passe par G, Δ est perpendiculaire à D, au point G.

Figure 2-3

Pour déterminer la "meilleure" droite, nous choisissons un indicateur de dispersion par rapport à D. L'indicateur choisi est l'inertie I du nuage par rapport à D :

$$I = \frac{1}{N}\sum_k (d_k)^2 \text{ avec } \quad d_k = |x_k\cos\theta + y_k\sin\theta + \gamma|$$

Nous démontrons en annexe (page 35) que d_k est la distance à D du point de coordonnées (x_k, y_k). La grandeur I représente donc le moment d'inertie du nuage par rapport à D (en affectant la masse $1/N$ à chaque point).

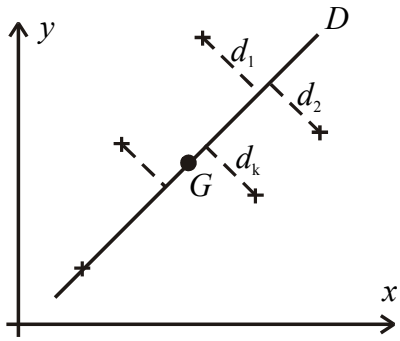
On vérifie que les propriétés qui confèrent à I le statut d'indicateur de dispersion sont effectivement vérifiées.

La "meilleure" droite est obtenue en imposant à I d'être minimal :

$$\frac{\partial}{\partial\theta} I = 0 \text{ et } \quad \frac{\partial}{\partial\gamma} I = 0 = \frac{2}{N}\sum_k (x_k\cos\theta + y_k\sin\theta + \gamma)$$

(voir les figures 2-3 *a*) et 2-4, comparer avec les figures 2-1).

La seconde des équations ci-dessus s'écrit sous la forme $\overline{x}\cos\theta + \overline{y}\sin\theta + \gamma = 0$ où θ et γ correspondent au "meilleur" choix. Cela signifie que **le point G appartient à D**.



La meilleure droite (D) passe par G,
D est telle que $\Sigma\, d_k^{\,2}$ *est minimal*

Figure 2-4

Remplaçons γ par sa valeur, $\gamma = -(\overline{x}\cos\theta + \overline{y}\sin\theta)$ dans l'expression de I. Il vient

$$I = \frac{1}{N}\sum_k \left((x_k - \overline{x})\cos\theta + (y_k - \overline{y})\sin\theta\right)^2$$

Développons cette expression et utilisons les définitions de l'écart quadratique moyen et de la covariance :

$$I = {\sigma_x}^2\cos^2\theta + {\sigma_y}^2\sin^2\theta + 2cov\,[x,y]\sin\theta\cos\theta \tag{2.3}$$

L'angle θ est celui qui minimise I. Les relations suivantes sont donc satisfaites : $\dfrac{dI}{d\theta} = 0$ avec $\dfrac{d^2I}{d\theta^2} > 0$.

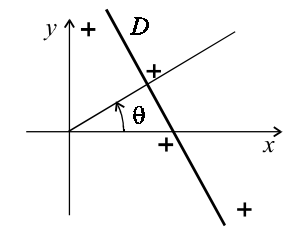
$$\frac{dI}{d\theta} = 0 \Leftrightarrow \boxed{\left({\sigma_x}^2 - {\sigma_y}^2\right)\tan(2\theta) = 2cov\,[x,y]} \tag{2.4}$$

Cette équation admet deux solutions définies modulo π : θ_D et $\theta_\Delta = \theta_D + \pi/2$. A chacune de ces solutions correspond une droite D ou Δ (voir figure 2-3 *b*)) Ces deux droites passent toutes deux par G. Elles sont perpendiculaires. La droite D correspond au minimum de I, noté I_D. Les deux droites étant perpendiculaires, la relation 2.3 implique $\boxed{I_D + I_\Delta = \Psi_G}$.

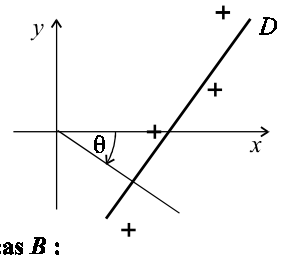
La condition $\dfrac{d^2I}{d\theta^2} > 0$ permet de distinguer θ_D et θ_Δ. La discussion des divers cas possibles se déroule alors de la façon suivante

1. $\sigma_y > \sigma_x \Rightarrow \cos(2\theta_D) > 0$ et $\left\{\begin{array}{lcll} cov\,[x,y] < 0 & \Rightarrow & \tan(2\theta_D) > 0 & \text{cas } A \\ cov\,[x,y] > 0 & \Rightarrow & \tan(2\theta_D) < 0 & \text{cas } B \end{array}\right.$
2. $\sigma_y < \sigma_x \Rightarrow \cos(2\theta_D) < 0$ et $\left\{\begin{array}{lcll} cov\,[x,y] < 0 & \Rightarrow & \tan(2\theta_D) < 0 & \text{cas } C \\ cov\,[x,y] > 0 & \Rightarrow & \tan(2\theta_D) > 0 & \text{cas } D \end{array}\right.$

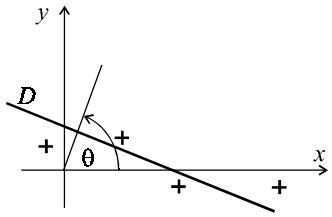
Sur la figure 2-4, nous résumons la discussion précédente.



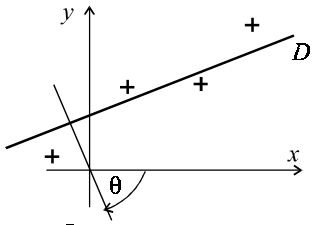
cas *A* :
$\sigma_y > \sigma_x$ et $cov[x,y] < 0 \iff 0 < \theta < \pi/4$



cas *B* :
$\sigma_y > \sigma_x$ et $cov[x,y] > 0 \iff -\pi/4 < \theta < 0$



cas *C* :
$\sigma_y < \sigma_x$ et $cov[x,y] < 0 \iff \pi/4 < \theta < \pi/2$



cas *D* :
$\sigma_y < \sigma_x$ et $cov[x,y] > 0 \iff -\pi/2 < \theta < -\pi/4$

Figure2-5

On remarquera que la droite D est croissante pour $cov\,[x,y] > 0$. On dit que *la corrélation est positive*. Dans le cas contraire elle est dite négative.

2.4.3 Inertie portée par une droite

La tendance étant dégagée, il reste à la qualifier : "Est-elle floue ou fortement marquée ?".

Choisissons une orientation de D et définissons G comme origine. Ces conventions permettent de définir l'abscisse, ℓ, d'un point quelconque sur D. Projetons le nuage de points sur la droite D. Ce faisant nous construisons une statistique à une dimension dont la variable est ℓ.

On vérifie que la moyenne de ℓ est nulle. Toute variable dont la moyenne est nulle est appelée ***"variable centrée"***.

L'écart quadratique moyen de la variable ℓ est σ_ℓ. La grandeur ${\sigma_\ell}^2$ s'interprète comme le moment d'inertie par rapport à G du nuage projeté. C'est aussi le moment d'inertie I_Δ.

On dit que I_Δ **est l'inertie portée par** D. On distinguera l'inertie portée par D (c'est I_Δ) et l'inertie par rapport à D (c'est I_D). L'inertie totale (inertie par rapport à G) est $\Psi = I_\Delta + I_D$ où I_D et I_Δ sont les quantités positives que nous avons définies. Pour apprécier l'inertie portée par D, on introduit le rapport ξ

$$\boxed{\xi = \frac{I_\Delta}{\Psi} = \frac{I_\Delta}{I_\Delta + I_D} = \frac{I_\Delta}{{\sigma_x}^2 + {\sigma_y}^2}}$$

On utilise, de préférence à ξ, le coefficient de corrélation linéaire ρ qui satisfait les relations

$$\rho = \frac{cov\ [X,Y]}{\sigma_X\ \sigma_Y}\ , \quad |\rho| = 2\xi - 1$$

Remarquons que si la droite D est croissante, ρ est positif et que ρ est négatif lorsque D est décroissante : ρ est de même signe que $cov\,[X,Y]$. Nous étudions ρ plus en détail

ci-dessous (paragraphe 2.5, page 31), mais nous en mentionnons la valeur dans les figures qui suivent.

A titre d'exemple, la figure 2-6 résume les propriétés de la population A présentée dans le tableau 1, page 24.

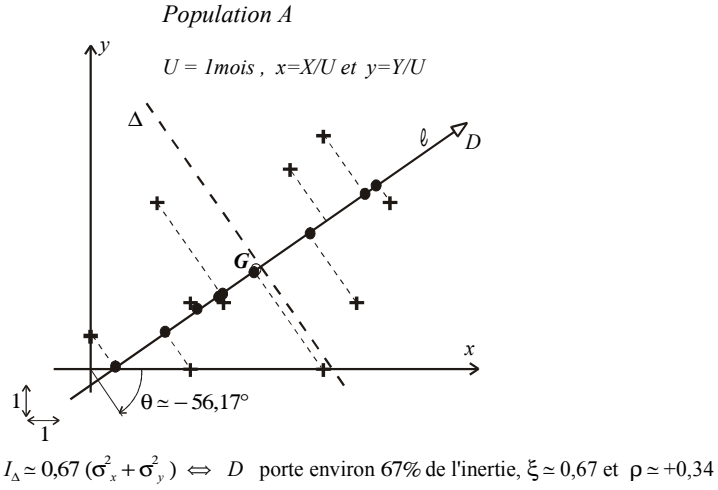


Figure 2-6 : Population A, Tableau 1 (page 24)

La relation $I_D \leq I_\Delta$ est satisfaite car D est choisie de telle sorte que I_D est minimal. On en déduit $0,5 \leq \xi$. La valeur $\xi = 0,5$ est obtenue lorsque $2I_D = \Psi = {\sigma_x}^2 + {\sigma_y}^2$. En utilisant l'expression 2.3 de I_D ainsi que la relation 2.4 on constate que la valeur $\xi = 0,5$ est obtenue pour ${\sigma_x}^2 = {\sigma_y}^2$ et $cov\,[x, y] = 0$ ce qui laisse θ complètement indéterminé.

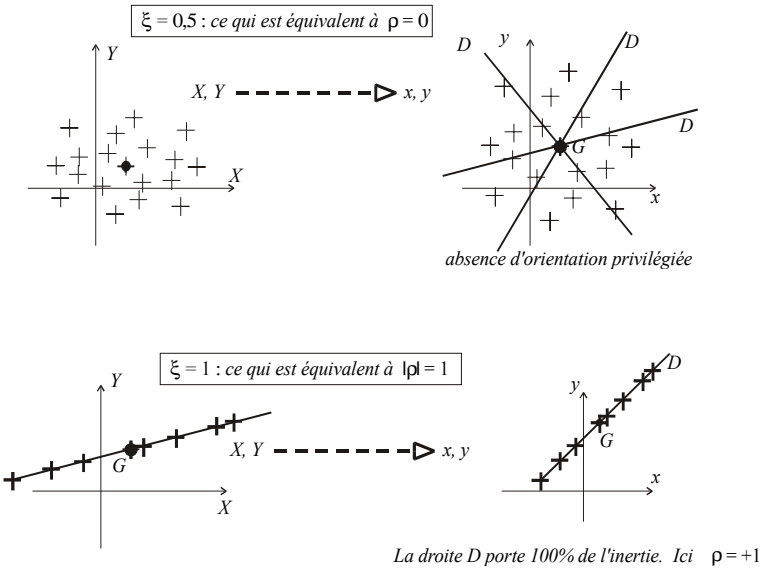


Figure 2-7

On vérifie aisément la relation $\xi \leq 1$. La valeur $\xi = 1$ est atteinte si et seulement si $I_D = 0$. L'inertie I_D est nulle lorsque tous les points du nuage sont alignés ; la droite qui porte ce nuage est alors la droite D. La projection représente donc parfaitement le nuage de points ; la droite D porte alors 100% de l'inertie.

On remarquera que la notion d'inertie portée s'étend à une droite quelconque : c'est l'inertie minimale du nuage de points à une dimension, obtenu en projetant le nuage de points sur la droite. La meilleure droite (la droite D) est la droite du plan qui porte une inertie maximale.

2.4.4 Variables principales

Le plan (x, y) est rapporté au repère $\{O; \vec{u}_x, \vec{u}_y\}$. Nous pouvons le rapporter au repère $\{G; \vec{v}_{x'}, \vec{v}_{y'}\}$, d'origine G, dont les axes sont les droites orthogonales D et Δ.

Effectuons le changement de base

$$\vec{v}_{x'} = \sin\theta\ \vec{u}_x - \cos\theta\ \vec{u}_y \text{ et } \vec{v}_{y'} = \cos\theta\ \vec{u}_x + \sin\theta\ \vec{u}_y \tag{2.5}$$

les nouvelles coordonnées introduites sont alors

$$x' = (x - x_G)\sin\theta - (y - y_G)\cos\theta, \quad y' = (y - y_G)\sin\theta + (x - x_G)\cos\theta$$

On vérifie les relations suivantes :

$$\begin{aligned} (\sigma_{x'})^2 &= (\sigma_x)^2\cos^2\theta + (\sigma_y)^2\sin^2\theta - cov(x,y)\sin(2\theta) \\ (\sigma_{y'})^2 &= (\sigma_x)^2\cos^2\theta + (\sigma_y)^2\sin^2\theta + cov(x,y)\sin(2\theta) \\ cov(x',y') &= 0 \end{aligned}$$

Remarquons que $(\sigma_{x'})^2$ est l'inertie portée par D tandis que $(\sigma_{y'})^2$ est l'inertie portée par Δ.

Lorsque deux variables satisfont la relation $cov(x', y') = 0$, on dit que ce sont deux ***variables orthogonales***. C'est le cas de x' et y' ici.

Dans le repère initial, on définit ***la matrice de covariance*** $C = \begin{bmatrix} (\sigma_x)^2 & cov(x,y) \\ cov(x,y) & (\sigma_y)^2 \end{bmatrix}$.

Cette matrice représente un opérateur linéaire $\hat{C}$. On effectue le changement de base 2.5. La matrice qui représente l'opérateur $\hat{C}$ dans la nouvelle base est notée C'. On obtient l'expression $C' = \begin{bmatrix} (\sigma_{x'})^2 & cov(x',y') \\ cov(x',y') & (\sigma_{y'})^2 \end{bmatrix}$. On vérifie que les vecteurs $\vec{v}_{x'}$ et $\vec{v}_{y'}$ que portent les axes D et Δ sont les vecteurs propres de $\hat{C}$ tandis que l'inertie portée par chaque axe est la valeur propre correspondante.

L'étude de la matrice de covariance, la recherche de ses vecteurs propres et de ses valeurs propres est très importante. C'est une façon très élégante d'introduire l'étude des nuages de points : trop élégante pour mettre en évidence la trivialité des intentions de départ.

2.5 Corrélation

2.5.1 Coefficient de corrélation linéaire

La corrélation entre les ventes quotidiennes de lunettes de Soleil et de glaces permet de mettre en évidence l'existence d'un facteur inconnu *a priori*, influençant les caractères décrits (c'est le Soleil dans ce cas). On mesure l'intérêt des corrélations tant pour les sciences naturelles que pour les sciences humaines. Leur mise en évidence fournit non seulement une connaissance sur les phénomènes mais aussi des indications sur leurs causes. Le point de vue et la méthode d'ajustement linéaire présentés au paragraphe précédent présente donc un grand intérêt. Pour que l'étude soit autonome (et par là même, universelle), il reste cependant à s'affranchir des références extérieures U_X et U_Y qui conduisent aux variables sans dimension x et y.

Reprenons l'étude de la population A précédente (tableau 1 page 24). On considère que l'état du marché du travail, le contexte des études, le dynamisme, la personnalité de l'étudiant, ou toutes autres causes, déterminent son intérêt pour les stages et influent sur le délai pour trouver un premier travail sans qu'il n'y ait de relation de cause à effet mais seulement corrélation. Dans ce contexte, la question est de savoir si les grandes valeurs de X sont, de préférence, associées aux grandes valeurs de Y (ou le contraire). Afin de définir le mot "grand" de la même façon pour X et Y, nous posons $x = X/\sigma_X$ et $y = Y/\sigma_Y$. Ainsi $x > 1$ et $y > 1$ signifient que x et y sont "grands" (voir page 20). Les variables X et Y possèdent en général des dimensions physiques et s'expriment au moyen d'unités de mesure. Par contre, x et y sont des nombres purs sans dimension. Remarquons enfin, que nous avons trouvé, dans la statistique elle-même, les éléments de référence qu'il n'est donc plus nécessaire de demander à l'extérieur comme au paragraphe précédent.

Cette opération ($X \rightarrow x = X/\sigma_X$) consiste à "réduire" la variable : x est une "variable réduite". En utilisant 1.9, on vérifie immédiatement la relation $\boxed{\sigma_x = 1}$. **Une variable réduite est donc une variable sans dimension, dont l'écart quadratique moyen est égale à l'unité.**

Maintenant, la population A est représentée par le tableau modifié, ci-dessous.

A : Ecole de commerce

$x =$	0	0,72	1,1	1,1	1,4	2,2	2,5	2,5	2,9	3,2
$y =$	0,42	2,1	0	0,83	0,83	2,5	0	2,9	0,83	2,1

Nous effectuons une analyse similaire à celle développée au paragraphe précédent en utilisant les mêmes indicateurs de dispersion pour définir le "meilleur" point et la "meilleure" droite susceptibles de représenter le nuage de points.

Le "meilleur" point, celui par rapport auquel la dispersion est minimale, est le point moyen, G, de coordonnées

$$\boxed{x_G = \overline{x} = \frac{\overline{X}}{\sigma_X} \text{ et } y_G = \overline{y} = \frac{\overline{Y}}{\sigma_Y}}$$

Chaque point étant affecté de la masse $1/N$, le point G apparaît comme le centre de masse du nuage.

La "meilleure" droite, celle qui porte la plus grande inertie, est la droite D qui passe par G et dont l'angle directeur, θ, satisfait la relation 2.4. Nous écrivons cette relation sous la forme $\tan(2\theta) = \dfrac{2cov[x,y]}{\sigma_x^2 - \sigma_y^2}$. La réduction des variables conduit aux relations $\sigma_x = 1 = \sigma_y$, et par conséquent $\boxed{\theta = \pm\frac{\pi}{4}}$ La droite D est donc soit parallèle à la première bissectrice soit parallèle à la seconde bissectrice, selon que la corrélation est positive ou négative, le signe de la corrélation étant celui de $cov[x,y]$ (voir la figure 2-5). On pose :

$$\boxed{\rho(X,Y) = cov[x,y] = \frac{cov[X,Y]}{\sigma_X\ \sigma_Y}}$$

ρ est appelé ***coefficient de corrélation linéaire*** de X et Y. Compte-tenu de la relation 1.10, il vient

$$\boxed{-1 \leq \rho \leq 1}$$

En utilisant la discussion résumée sur les figures 2-5, on vérifie la relation $cov[x,y] \times \sin(2\theta) \leq 0$. Dans le cas présent cette relation s'écrit $cov[x,y] \times \sin(2\theta) = -|\rho|$. On

utilise cette relation, l'expression 2.3 de I_D ainsi que la valeur de θ et les relations $\Psi_G = {\sigma_x}^2 + {\sigma_y}^2 = 2 = I_D + I_\Delta$. On démontre alors sans difficulté les propriétés que nous résumons sur les figures 2-7 ainsi que

$$I_D = 1 - |\rho|\,, \quad I_\Delta = 1 + |\rho|\,, \quad \xi = \frac{1+|\rho|}{2}$$

2.5.2 Matrice de covariance

Le plan (x, y) étant rapporté à un repère orthonormé, nous introduisons l'opérateur de covariance. Celui-ci est représenté par la matrice $C_r = \begin{bmatrix} (\sigma_x)^2 & cov\,(x,y) \\ cov\,(x,y) & (\sigma_y)^2 \end{bmatrix} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$. Cette matrice est appelée ***"matrice de covariance des variables réduites"*** x et y

2.5.3 Variables indépendantes

Par mesure de simplicité, nous avons supposé que les valeurs prises par chacune des variables sont toutes différentes. Nous abandonnons, ici, cette hypothèse. Le nombre de couples (X_i, Y_j) est noté n_{ij}. Avec les présentes notations, les variables d'indices différents sont différentes : par exemple $X_j \neq X_k$ pour $j \neq k$ et $Y_m \neq Y_\ell$ pour $m \neq \ell$.

$N_k = \sum_j n_{kj}$ est le nombre de couples ou apparaît la variable X_k. Le nombre total de couples observés est $N = \sum_k N_k$. La variable X_k étant fixée, la proportion de couples pour lesquels $Y = Y_j$ est $\frac{n_{kj}}{N_k}$. Les variables X et Y sont **stochastiquement indépendantes** si cette proportion ne dépend pas de la valeur X_k choisie, ce que l'on peut écrire sous la forme $\frac{n_{kj}}{N_k} = P_j$ (indépendant de l'indice k).

Le tableau des données se présente lors sous la forme suivante où chaque valeur de X correspond à une colonne tandis que chaque valeur de Y correspond à une ligne. A l'intersection de la ligne Y_j et de la colonne X_k figure n_{kj}.

	X_1	X_2	...	X_k	...	Total des lignes
Y_1	$n_{11}= P_1N_1$	$n_{21}= P_1N_2$	...	$n_{k1}= P_1N_k$	...	P_1N
Y_2	$n_{12}= P_2N_1$	$n_{22}= P_2N_2$	...	$n_{k2}= P_2N_k$	...	P_2N
			...		...	
Y_j	$n_{1j}= P_jN_1$	$n_{2j}= P_jN_2$	...	$n_{kj}= P_jN_k$	...	P_jN
			...		...	
Total des colonnes	N_1	N_2	...	N_k	...	$N = N_1+N_2+...$

P_j est la fréquence relative de Y_j, dorénavant nous notons cette quantité sous la forme P_{Y_j}. De même, la fréquence relative de X_k est noté $P_{X_k} = \frac{N_k}{N}$.

De ce qui précède, nous déduisons $n_{kj} = P_{Y_j} N_k = P_{Y_j} \frac{N_k}{N} \times N$ soit $n_{kj} = P_{Y_j} P_{X_k} \times N$. Le tableau précédent s'écrit sous la forme

	X_1	...	X_k	...	Total des lignes
Y_1	$n_{11}= P_{Y_1}P_{X_1}N$	...	$n_{k1}= P_{Y_1}P_{X_k}N$	...	$\mathcal{N}_1= P_{Y_1}\ N$
Y_2	$n_{12}= P_{Y_2}P_{X_1}N$	...	$n_{k2}= P_{Y_2}P_{X_k}N$	...	$\mathcal{N}_2= P_{Y_2}\ N$
		...		...	
Y_j	$n_{1j}= P_{Y_j}P_{X_1}N$	...	$n_{kj}= P_{Y_j}P_{X_k}N$	...	$\mathcal{N}_j= P_{Y_j}\ N$
		...		...	
Total des colonnes	$N_1= P_{X_1}\ N$	...	$N_k= P_{X_k}\ N$	...	$N =\mathcal{N}_1+\mathcal{N}_2+...$ $N = N_1+N_2+...$

En utilisant les définitions il vient
$\boxed{\overline{X} = \sum_k P_{X_k} \times X_k}$, $\boxed{\overline{Y} = \sum_j P_{Y_j} \times Y_j}$ et

$\overline{XY} = \frac{1}{N}\sum_{k\ j} n_{kj}X_kY_j = \sum_{k\ j} P_{X_k}P_{Y_j} \times X_kY_j$ soit $\overline{XY} = \sum_k P_{X_k}\ X_k \times \sum_j P_{Y_j}\ Y_j = \overline{X} \times \overline{Y}$.

On en déduit $\boxed{cov\,[X,Y] = 0}$ et par conséquent $\boxed{(\sigma_{X+Y})^2 = (\sigma_X)^2 + (\sigma_Y)^2}$ ce que nous avons déjà mentionné au paragraphe 1.5.2, ci-dessus. On retiendra la propriété suivante :

$$\boxed{X \text{ et } Y \text{ indépendants } \Rightarrow cov\,[X,Y] = 0 \Longleftrightarrow \rho(X,Y) = 0}$$

La réciproque n'est pas vraie : la nullité de $\rho(X,Y)$ n'implique pas l'indépendance des variables X et Y. Considérons, à titre de contre-exemple, le tableau suivant :

$X_k \rightarrow$ $Y_j \downarrow$	-1	0	1	$\mathcal{N}_j$
-1	1	1	1	3
0	0	1	1	2
1	1	1	1	3
N_k	2	3	3	$N = 8$

Pour $X = -1$, la distribution des valeurs de Y n'est pas la même que pour les autres valeurs de X. Les variables X et Y ne sont donc pas indépendantes. Cependant, on trouve

$\overline{XY} = (1 - 1 - 1 + 1)/8 = 0$, $\overline{X} = (-2 + 3)/8 = 1/8$ et $\overline{Y} = (-3 + 3)/8 = 0$.

On en déduit $cov\,[X,Y] = 0$ et par conséquent $\rho = 0$ bien que les variables ne soient pas indépendantes.

2.6 Conclusion

Le coefficient de corrélation linéaire est souvent considéré comme un indicateur d'indépendance linéaire. Il est cependant dépourvu des propriétés qui lui donneraient ce statut. En effet, pour que $|\rho|$ puisse être considéré comme un indicateur d'indépendance linéaire, il faudrait que la nullité de $|\rho|$ implique l'indépendance, ce qui n'est pas le cas. Par contre, $|\rho| = 1$ implique l'existence d'une relation linéaire entre X/σ_X et Y/σ_Y. Le coefficient de corrélation (ou plus précisément $|1 - |\rho||$) peut donc être considéré comme un indicateur de dépendance linéaire.

Soulignons ici l'importance du vocabulaire : si deux variables ne présentent aucun signe de dépendance (corrélation nulle) cela ne signifie pas qu'elles sont indépendantes (rappelons que de telles variables sont dites "orthogonales").

Disposer de variables indépendantes est une nécessité pour une approche scientifique de nombreuses opérations d'analyse qui commencent par **le choix au hasard d'un**

échantillon. Réaliser un tel échantillon c'est s'assurer que le choix d'un quelconque individu qui figurera dans l'échantillon est indépendant du choix des autres individus. Dans ce but, on numérote chaque individu de la population considérée. Cette numérotation peut être effectivement réalisée, mais on peut aussi se contenter de définir seulement une façon de procéder. C'est généralement le cas lorsqu'on considère des populations nombreuses. A chaque individu est donc affectée une variable X de valeur $1, 2, ...$ ou N si N est le cardinal de la population. Les variables associées à deux individus différents sont différentes.

Extraire un échantillon de cardinal K, de la population considérée revient à choisir K valeurs de la variable X parmi les N valeurs affectées.

Dans le tirage non exhaustif au hasard, la variable X_j qui caractérise le $j^{ème}$ individu choisi est indépendante de la variable X_m qui caractérise le $m^{ème}$ individu de l'échantillon. Dans ces conditions, les lois du hasard pourront s'appliquer et des prédictions précises, concernant la population, pourront être déduites de l'étude de l'échantillon. Nous en verrons des exemples dans la suite de ce cours. Effectuer un tirage au hasard[†] ne consiste donc pas à faire n'importe quoi : à l'évidence, la construction de tables de nombres au hasard ne saurait se faire au hasard...

On trouvera au dernier chapitre une table de nombres au hasard avec son mode d'emploi.

Les méthodes précédentes se généralisent aux populations dont les individus sont décrits par plus de 2 caractères dans le cadre des analyses à variables multiples (dites "analyses multivariées"). Les nuages de points sont alors les nuages d'un espace de dimension n (si le n est le nombre de variables numériques prises en compte).

La matrice de covariance, C, est alors une matrice $n \times n$. Les variables sont notées X_1, X_2, ..., X_n (attention, ici, ce sont les variables elles-mêmes que nous notons X_k et non les valeurs qu'elles sont susceptibles de prendre). On pose $\sigma_{X_k} = \sigma_k$ et $cov\left[X_k, X_j\right] = c_{k,j} = c_{j,k}$. La matrice de covariance s'écrit :

$$\begin{bmatrix} (\sigma_1)^2 & c_{1,2} & ... & c_{1,n} \\ c_{2,1} & (\sigma_2)^2 & ... & c_{2,n} \\ ... & ... & ... & ... \\ c_{n,1} & c_{n,2} & ... & (\sigma_n)^2 \end{bmatrix}$$

La matrice de corrélation s'obtient en posant $\sigma_k = 1$ et $c_{k,j} = \rho\left(X_k, X_j\right)$ dans l'expression formelle de la matrice de covariance. La détermination des axes principaux et des variables principales s'effectue en diagonalisant de telles matrices.

Comme il n'est pas aisé de se représenter un espace de dimension $n > 3$, on projette les n axes et les N points du nuage sur le plan qui porte l'inertie maximale. De tels diagrammes permettent de "visualiser" le nuage de points. Si ce plan porte 90% de l'inertie totale, on peut assimiler le nuage à un nuage plan (à deux dimensions) et ne retenir comme variables significatives que les deux coordonnées de ce plan (ce sont des combinaisons linéaires des variables de départ). On réduit ainsi le nombre de variables sur lesquelles porte l'analyse.

Ces méthodes constituent des outils pour étudier les facteurs qui influencent tel ou tel phénomène, tel ou tel comportement : ce sont des méthodes d'***analyse factorielle***.

Annexe 1 : distance d'un point à une droite.

Le plan est rapporté au repère orthonormé $\left\{\vec{u}_x, \vec{u}_y\right\}$, d'axes Ox, Oy.

[†]Rappelons que, sauf mention contraire, et sans que ce ne soit nécessairement explicité, les échantillons que nous considérons dans ce cours sont tirés au hasard, de façon non exhaustive (avec remise) à partir des populations concernées.

Etant donnée une droite D, d'équation $x\ \cos\theta + y\ \sin\theta + \gamma = 0$ et un point P de coordonnées (x_P, y_P). Nous nous proposons de démontrer que la distance de P à D est

$$d = |x_P\ \cos\theta + y_P\ \sin\theta + \gamma|$$

Considérons une droite, D, qui passe par le point M_0, de coordonnées (x_0, y_0) et qui est orthogonale au vecteur unitaire $\vec{u} = \cos\theta\ \vec{u}_x + \sin\theta\ \vec{u}_y$. Soit M le point courant de cette droite, de coordonnées (x, y).

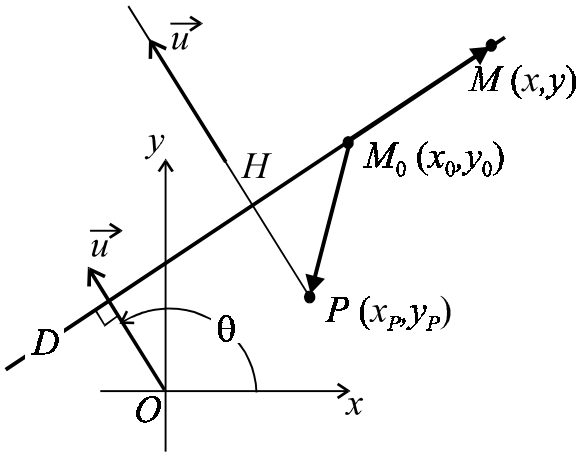


Figure 2-8

L'équation de D s'écrit $\overrightarrow{M_0M} \cdot \vec{u} = 0$. Soit $\cos\theta\ (x - x_0) + \sin\theta\ (y - y_0) = 0$ ou encore

$$x\ \cos\theta + y\ \sin\theta + \gamma = 0 \text{ avec } \gamma = -x_0\ \cos\theta - y_0\ \sin\theta$$

Considérons un point P de coordonnées x_P et y_P. Le point P n'est pas nécessairement sur la droite D : en général $x_P\ \cos\theta + y_P\ \sin\theta + \gamma \neq 0$.

La distance de P à la droite D est $d = PH$ (cf figure 2-8). Pour obtenir PH, nous projetons le vecteur $\overrightarrow{M_0P}$ sur un axe parallèle à $\vec{u}$. Il vient
$d = PH = \left|\overrightarrow{M_0P} \cdot \vec{u}\right| = |(x_P - x_0)\cos\theta + (y_P - y_0)\sin\theta| = |x_P\ \cos\theta + y_P\ \sin\theta + \gamma|$.
C'est ce que nous voulions démontrer.

Annexe 2 : inertie et statistiques

En statistiques, l'étude des nuages de points fait appel à des outils similaires à ceux que l'on emploie pour étudier les corps matériels en mécanique. Nous rappelons ici les principales notions concernant l'inertie et nous présentons les similitudes des deux domaines.

En mécanique, l'inertie est une grandeur qui mesure la répugnance d'un corps à modifier son état de mouvement. Précisons ce concept quelque peu littéraire.

Lorsqu'on applique sur un corps de masse m un ensemble de forces extérieures[‡] dont la résultant est $\overrightarrow{F}$, ce corps subit une accélération $\overrightarrow{a} = \overrightarrow{F}/m$. En fait, c'est le centre

[‡]Les forces internes qui s'exercent entre les diverses parties du corps, les forces de cohésion par exemple, ont une résultante nulle.

de masse du corps qui subit cette accélération (la définition du centre de masse est rappelée ci-dessous). Pour une force donnée, l'accélération est d'autant plus faible que la masse du corps est grande ; la masse mesure l'inertie du corps pour l'expérience considérée.

Considérons un corps solide en rotation autour d'un axe fixe, δ, avec une vitesse angulaire Ω. Appliquons à ce corps des forces extérieures dont le moment résultant par rapport à δ est $\mathcal{M}$. Sous l'effet de $\mathcal{M}$, la vitesse angulaire varie : $\mathcal{M} = I\dfrac{d\Omega}{dt}$ où I est le moment d'inertie du solide par rapport à l'axe de rotation et $\dfrac{d\Omega}{dt}$ l'accélération angulaire du mouvement. Ici, c'est I qui mesure l'inertie du corps pour l'expérience considérée.

La masse, m, du corps étant donnée, le moment d'inertie I ne dépend que de la répartition spatiale de m. Si la matière est située loin de l'axe δ, le moment d'inertie, I, est élevé.

Il est remarquable que ce soit un outil similaire au moment d'inertie de la mécanique qui s'avère utile en statistique. Peut-être est-ce une indication du manque d'originalité de l'esprit humain qui réinvente toujours le même outils dans des situations souvent fort diverses : la linéarité des équations, le rôle de l'oscillateur harmonique et ici les notions liées à l'inertie sont trois exemples de ce manque d'originalité... à moins que nous n'inventions pas les outils mais que nous découvrions les lois de la nature, toujours les mêmes.

Etant donné un corps matériel, on le considère comme la réunion d'un ensemble de petits volumes, assimilés à des points matériels M_k, de masse m_k.

La masse totale du corps est $m = \sum_k m_k$.

Un nuage de points en statistique est un ensemble de N points, M_k, dont les coordonnées sont les valeurs des variables que l'on étudie. Les points peuvent être tous différents, dans ce cas la masse de chaque point est $1/N$. Certains points peuvent être répétés. Le point géométrique M_k doit alors être considéré comme la superposition de n_k points du nuage. A un tel point géométrique est associée la masse $n_k/N = P_k$.

La masse totale du nuage est $\sum_k P_k = 1$.

Le barycentre des points M_k affectés des coefficients m_k est le centre de masse du corps. On le note G. Il est défini par l'une quelconque des deux relations

$$\overrightarrow{OG} = \frac{1}{m}\sum_k m_k \ \overrightarrow{OM_k} \Leftrightarrow \sum_k m_k \ \overrightarrow{GM_k} = \vec{0}$$

ou O est une origine choisie arbitrairement.

En statistiques, le barycentre des points M_k affectés des coefficients P_k est le points moyen du nuage. On le note G. Ses coordonnées sont les moyennes des coordonnées des points M_k :

$$\overrightarrow{OG} = \sum_k P_k \ \overrightarrow{OM_k} \Leftrightarrow \sum_k P_k \ \overrightarrow{GM_k} = \vec{0}$$

ce qui, pour un nuage à trois dimensions, implique

$$X_G = \sum_k P_k \ X_k = \overline{X}, \ Y_G = \sum_k P_k \ Y_k = \overline{Y}, \ Z_G = \sum_k P_k \ Z_k = \overline{Z},$$

Etant donné un point quelconque Q, le moment d'inertie du corps par rapport à Q est la quantité $\Psi(Q) = \sum_k m_k \ \overrightarrow{QM_k}^2$. On peut démontrer que le centre de masse G est le point Q pour lequel $\Psi(Q)$ est minimal.

Etant donné un point quelconque Q, la dispersion du nuage par rapport à Q est défini par la relation $\Psi(Q) = \sum_k P_k \overrightarrow{QM_k}^2$. On peut démontrer que le point moyen G est le point Q pour lequel $\Psi(Q)$ est minimal.

Etant donnée une droite D, on note d_k la distance de M_k à la droite D. Le moment d'inertie du corps par rapport à D est $I_D = \sum_k m_k \left(d_k\right)^2$.

Etant donnée une droite D, on note d_k la distance de M_k à la droite D. En statistique, le moment d'inertie d'un nuage de points par rapport à D est $I_D = \sum_k P_k \left(d_k\right)^2$.

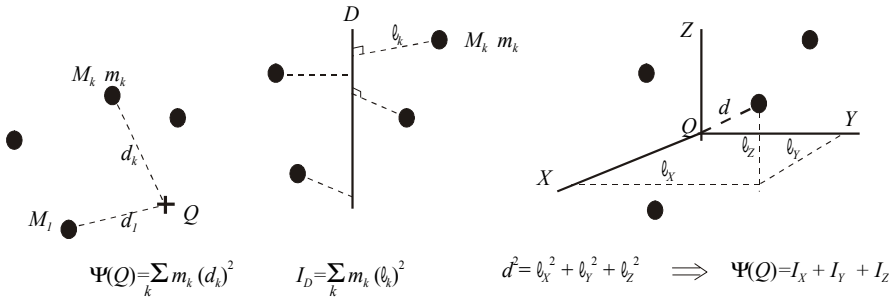


Figure 2-16

N.B. Pour un corps ou un nuage de points statistique à 2 dimensions il vient $I(Q) = I_D + I_\Delta$ où D et Δ sont deux droites orthogonales passant par Q.

Chapitre 3
STATISTIQUES ET PROBABILITÉS

3.1 Introduction

Le paysan :

"Le temps est lourd ; il y aura *probablement* un orage en fin d'après-midi."

Au bistrot :

"Tu gagneras *sans doute* cette tournée au 421."

Au Bac après l'épreuve de philo :

"Je suis *presque certain* d'avoir une bonne note."

Dans bien d'autres circonstances encore les hommes parient sur l'avenir. Mais comment passer d'une estimation intuitive et grossière à une mesure du probable et de l'improbable ?

Si vous avez suivi la façon dont les outils furent construits aux chapitres précédents, la réponse vous vient aux lèvres : "Pour apprécier la probabilité d'un événement, il convient de construire un indicateur, P".

La valeur de l'indicateur doit permettre de reconnaître un événement certain et un événement irréalisable. On convient que P est égal à 1 dans le premier cas et zéro dans le second cas.

D'autre part, considérant une succession d'événements *de plus en plus probables*, la valeur de P doit se rapprocher de façon monotone de la valeur 1. On en déduit la relation à imposer : $0 \leq P \leq 1$.

Mais n'est-il pas choquant d'évoquer des "événements *de plus en plus probables*" sans en avoir défini le sens ? Non. Ce n'est pas plus choquant que d'évoquer la "magnitude des étoiles" avant que ne soient précisés les outils physiques qui en permettent la définition précise : pour définir une grandeur, il faut bien en parler !

C'est le jeu qui est à l'origine du concept et du calcul des probabilités dans sa forme moderne. Le jeu de dés se dit "*alea*" en Latin. Le mot est passé en français. Nous avons déjà utilisé l'adjectif "aléatoire" page 3 pour désigner une variable susceptible de prendre plusieurs valeurs ; nous retrouverons bien des fois encore cette expression.

Si l'histoire commence au $17^{ème}$ siècle avec Pierre de Fermat (1601-1665) et Blaise Pascal (1623-1662), la préhistoire remonte au moins à Jérôme Cardan* (1501-1576) dans un traité écrit vers 1564 ("*De ludo aleae*" : "Du jeu de dés"), perdu, puis retrouvé et enfin publié en 1663 après les travaux de Fermat et Pascal.

Le chevalier de Méré† était un "honnête homme", intéressé par les mathématiques, bel esprit, fort porté sur le jeu. Après bien d'autres questions, vers 1654, il pose à Pascal

*Girolamo Cardano est connu pour la publication de travaux sur les équations de degré 3 et 4, pour la suspension à la Cardan des compas marins ou encore les cardans d'automobiles. Il est beaucoup moins connu pour les travaux qui peuvent lui être incontestablement attribués et concernent des sujets aussi divers que la médecine, la philosophie, la physique, la musique ou les mathématiques (en particulier l'introduction des racines des nombres négatifs, préfiguration des nombres complexes).

†Antoine Gombaud (1607-1684).

le problème suivant. Deux joueurs jouent à "*pile* ou *face*". L'un choisit "*pile*" et l'autre "*face*". A gagné celui qui le premier a vu apparaître cinq fois le côté choisi. L'enjeu est un "pot" où chacun a mis initialement la même somme. Après un certain temps la partie est interrompue pour une raison impérative. Comment doit-on répartir le pot ?

Si 4 fois *pile* et 4 fois *face* sont déjà apparus, c'est assez simple : chacun prend la moitié du pot. Par contre, si 4 fois *face* et une fois *pile* sont déjà apparus, il n'est pas très équitable de se partager le pot à égalité car l'un des joueurs a presque gagné.

Cette question du chevalier, traitée par Fermat et Pascal, reçut deux réponses identiques, obtenues par des raisonnements différents.

C'est au $17^{ème}$ siècle qu'une première définition précise est donnée d'une probabilité comme le rapport du nombre de cas favorables au nombre total de cas possibles (nous y reviendrons). A cette époque, la notion de jeu honnête est déjà mise en évidence depuis longtemps. On ne peut même pas imaginer que, dans l'antiquité la plus lointaine, la notion de jeu truqué n'ait pas été connue. On sait donc que la probabilité de *pile* est égale à la probabilité de *face* dans un jeu non biaisé. Quand le chevalier de Méré pose sa question à Pascal, plusieurs questions simples ont déjà reçu une réponse (la probabilité d'un six et celle d'un double six lorsqu'on lance 2 dés par exemple) mais ici la situation est plus compliquée. Ce n'est pas la première fois qu'un tel problème est posé mais c'est la première fois qu'une solution correcte en est donnée.

Pour toutes ces raisons, c'est au $17^{ème}$ siècle que l'on fait remonter l'origine du calcul des probabilités ; toutefois, c'est au $20^{ème}$ siècle, principalement avec les travaux de Andreï Kolmogorov (1903-1987), qu'apparaît la formulation axiomatique et que se développe ce qui est une branche des mathématiques à part entière.

Pour puissante et élégante que soit la théorie de la mesure, nous ne suivrons pas l'approche axiomatique. Le berger italique qui comptait ses moutons en gravant sur un bâton, un I pour un mouton, un V pour cinq moutons, un X pour dix, n'avait nul besoin d'axiomatique ; l'intendant qui devait additionner XXVII moutons et XXXII moutons pour obtenir LIX moutons a sans doute trouvé dans le système décimal et les chiffres arabes[‡] un progrès important : l'outil s'est construit et amélioré sans la nécessité des axiomes de Peano. Nous ne prétendons pas pour autant que l'axiomatique soit sans intérêt ; c'est tout le contraire, mais à chaque objectif ses moyens. On comprendra donc, compte tenu des choix explicités en introduction, que nous ne privilégions pas ici l'approche axiomatique de la théorie des probabilités.

3.2 Définitions

3.2.1 Probabilités

Pour évaluer la probabilité d'un événement, nous devons construire un indicateur P qui satisfasse la relation $0 \leq P \leq 1$ et qui croisse avec la probabilité[§] de l'événement considéré, de la valeur 0 pour les événements irréalisables à la valeur 1 pour les événements certains.

Etudions un cas simple, le plus simple possible.

Nous considérons ***une épreuve,*** ensemble d'opérations physiques, dont le résultat est ***une éventualité.*** Par exemple, l'épreuve consiste à lancer deux dés, un rouge et un vert. Le résultat de cette épreuve est un couple de valeurs (X, Y) où X est l'indication du dé rouge et Y celle du dé vert. Les nombres X et Y sont des entiers de l'intervalle $[1, 6]$; un couple (X, Y) constitue une éventualité.

[‡]Chiffres persans apportés en occident par les Arabes.

[§]Ici, le mot probabilité est compris dans son acception usuelle, assez vague, que nous essayons de quantifier.

Intéressons-nous à la somme des deux dés : $Z = X + Y$. La réalisation d'une valeur donnée de Z, par exemple $Z = 4$, est un ***événement.*** Plusieurs éventualités permettent de réaliser le même événement. Ici, les trois éventualités $(1,3)$, $(2,2)$ et $(3,1)$ permettent la réalisation de l'événement $Z = 4$.

Une éventualité ne peut être réalisée que d'une seule façon, on dit aussi que c'est un ***événement élémentaire.***

Les résultats possibles de l'épreuve considérée consistent en 36 éventualités et 11 événements. Nous représentons ci-dessous les divers événements et les éventualités qui les constituent.

Evénement : E	Eventualités	Cardinal de E
$Z = 2$	$(1,1)$	1
$Z = 3$	$(1,2)$, $(2,1)$	2
$Z = 4$	$(1,3)$, $(2,2)$, $(3,1)$	3
$Z = 5$	$(1,4)$, $(2,3)$, $(3,2)$, $(4,1)$	4
$Z = 6$	$(1,5)$, $(2,4)$, $(3,3)$, $(4,2)$, $(5,1)$	5
$Z = 7$	$(1,6)$, $(2,5)$, $(3,4)$, $(4,3)$, $(5,2)$, $(6,1)$	6
$7 = 8$	$(2,6)$, $(3,5)$, $(4,4)$, $(5,3)$, $(6,2)$	5
$Z = 9$	$(3,6)$, $(4,5)$, $(5,4)$, $(6,3)$	4
$Z = 10$	$(4,6)$, $(5,5)$, $(6,4)$	3
$Z = 11$	$(5,6)$, $(6,5)$	2
$Z = 12$	$(6,6)$	1
Total = cardinal de Ω		**36**

Les 36 éventualités constituent un ensemble que l'on appelle l'Univers. De façon générale, ***l'Univers est l'ensemble des éventualités,*** on le désigne par la lettre grecque Ω (omega).

Le même exemple, présenté de façon différente, consisterait à disposer de 36 boules dont chacune représente une éventualité (pour savoir laquelle, il suffit de l'écrire sur la boule). Lancer les dés est une opération équivalente au tirage[¶] d'une boule.

La vie n'est pas faite que de jeux ! Parfois, il faut voter. Supposons qu'il y ait 20 millions d'électeurs et deux candidats : Clovis et Clothilde. A chaque électeur sera associé son vote : *Clovis*, *Clothilde* ou encore *abstention.* Il y a aussi "blanc ou nul" mais laissez-nous oublier cette complication pour le moment.

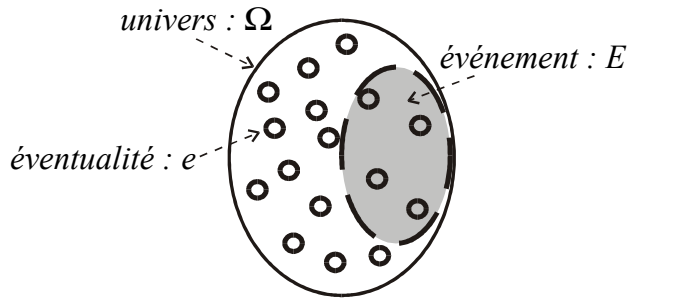
Chaque électeur est représenté par une boule sur laquelle est écrit son vote. Avant le vote, je veux sonder l'opinion pour deviner l'avenir (ou l'influencer). J'admets que l'inscription est marquée sur chaque boule avant le vote. C'est une hypothèse très forte, mal vérifiée car c'est souvent au dernier moment que les décisions se prennent. Cette hypothèse est cependant celle qui est implicitement posée lorsqu'on interprète un sondage comme une prévision.

Dans ces conditions, interroger un électeur c'est tirer une boule ; cette opération constitue l'épreuve. A cette épreuve correspondent 20 millions d'éventualités mais seulement trois événements : *Clovis*, *Clothilde* et *abstention.*

Réaliser une épreuve revient à tirer une boule dans une population de boules, Ω. L'obtention d'une boule désignée à l'avance est une éventualité qui peut se réaliser ou non. Un ensemble de boules constitue un événement, E, qui lui aussi peut se réaliser ou non.

On donne la représentation diagrammatique ci-dessous (figure 3-1)

[¶]Rappelons que, sans mention contraire, et sans que ce ne soit nécessairement explicité, les tirages que nous considérons dans ce cours sont des tirages au hasard, non exhaustifs (avec remise) à partir des populations concernées.



cardinal de $\Omega = N = 16$ *cardinal de E :* $N_E = 4$

Proba $[e] = 1/16$ Proba $[E] = 4/16 = 0{,}25$

Figure 3-1

Les boules sont numérotées mais à cette exception près, elles sont toutes identiques, elles sont brassées par des palettes dans un récipient et de temps en temps l'une de ces boules est prélevée. On peut considérer que l'obtention de la boule n°k est aussi probable que l'obtention d'une autre boule. Les boules sont donc équiprobables. Les éventualités qu'elles décrivent sont équiprobables.

Nous verrons que la loi des grands nombres nous donne un moyen pour tester cette hypothèse ; pour le moment, l'équiprobabilité des éventualités est une hypothèse raisonnable qui repose sur la similitude physique des boules et la nature homogène du brassage.

L'événement E est réalisé si se réalise l'une des éventualités qui composent E.

Dans ces conditions, on définit $P[E] = \dfrac{N_E}{N}$ où N_E est le nombre d'éventualités qui composent l'événement E, tandis que N est le nombre d'éventualités qui composent l'univers, Ω. On dit encore que N_E est le cardinal de E et N, le cardinal de Ω.

Remarquons que P_E possède les propriétés de l'indicateur cherché.

- P n'est pas négatif
- Le sous-ensemble vide de Ω est noté $\varnothing$. C'est un événement irréalisable car le résultat de l'épreuve est nécessairement une éventualité tandis que $\varnothing$ n'en contient aucune ; considéré comme un événement, la définition de P donne $P_\varnothing = 0$.
- L'événement Ω est certainement vérifié car toutes les éventualités appartiennent à Ω. D'autre part, suivant la définition, $P_\Omega = 1$ car dans ce cas $N_E = N_\Omega = N$.
- Il reste à vérifier que P_E croît lorsque la probabilité de E augmente. Peignons en rouge les boules qui représente les événements de E. La probabilité de tirer une boule rouge augmente lorsque le nombre de boules rouges augmente ; dans le même temps $P_E = N_E/N$ augmente.

Nous pouvons donc considérer que P_E est un bon indicateur de la probabilité de voir E se réaliser lors du tirage. On définit donc ***la probabilité*** de l'événement E :

$$\boxed{P[E] = \frac{N_E}{N}} \tag{3.1}$$

A partir de maintenant, le mot "probabilité" désignera le nombre P et non plus une notion vague associée au caractère plus ou moins plausible de la réalisation d'un événement.

L'éventualité e est un cas favorable (pour la réalisation de E) si $e \in E$. On constate que la définition 3.1 de la probabilité peut s'énoncer sous la forme : *la probabilité de E est le rapport du nombre de cas favorables au nombre de cas possibles.* Bien sûr, un tel énoncé n'est valable que si les divers cas sont équiprobables ce qui est réalisé ici.

3.2.2 Opérations sur les événements

Considérons deux événements A et B. Dans la suite nous notons A' et B' leurs complémentaires. Nous rappelons les principales définitions et la façon de les lire :

Complémentaire :

e	$\in$	A'	*ssi*	e	$\notin$	A
e	appartient à	complémentaire de A	si et seulement si	e	n'appartient pas à	A

Réunion :

e	$\in$	$(A$	$\cup$	$B)$	*ssi*	$e \in A$ ou $e \in B$
e	appartient à	A	union	B	si et seulement si	$e \in A$ ou $e \in B$

Intersection :

e	$\in$	$(A$	$\cap$	$B)$	*ssi*	$e \in A$ et $e \in B$
e	appartient à	A	inter	B	si et seulement si	$e \in A$ et $e \in B$

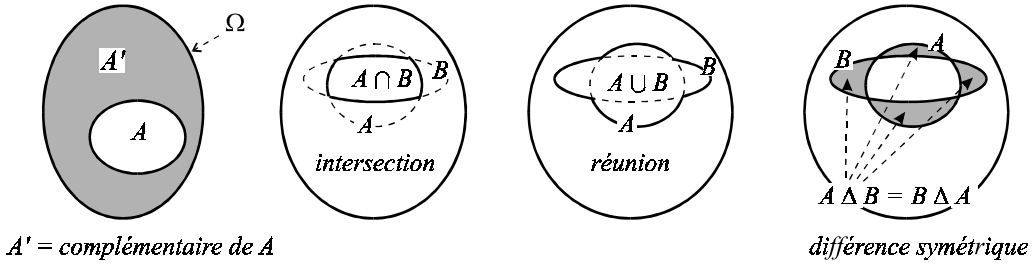


Figure 3-2

Remarquons que $N_A + N_{A'} = N$. on en déduit

$$P_A + P_{A'} = 1$$

Les éventualités qui appartiennent à l'intersection $A \cap B$ assurent la réalisation de A et celle de B simultanément (simultanément signifie ici comme résultat de la même épreuve, sans idée de simultanéité chronologique). Il n'y a pas d'autres éventualités qui assurent cette propriété. On en déduit :

$$\text{Probabilité de } \{A \text{ et } B\} = P_{A \cap B}$$

L'événement $A \cup B$ est l'événement réalisé lorsque se réalise une éventualité qui appartient soit à A, soit à B, soit à A et B le cas échéant. Dans ce cas, on dit que A *ou* B est réalisé ; *ou* a ici un "sens inclusif".

On peut considérer le "sens exclusif" de *ou* : "fromage *ou* dessert". Dans ce cas affirmer que A *ou* B est réalisé signifie que A et B ne peuvent être réalisés simultanément. Les éventualités qui assurent la réalisation de A ou celle de B sans que ne soient réalisées A et B sont les éventualités de $A \Delta B$ où Δ représente la différence symétrique (figure 3-2) ; $A \Delta B = (A \cap B') \cup (B \cap A')$.

$$\begin{aligned} \text{Probabilité de } \{A \text{ ou } B\} \text{ ou } \{A \text{ et } B\} &= P_{A \cup B} \\ \text{Probabilité de } \{A \text{ ou } B\} \text{ sans } \{A \text{ et } B\} &= P_{A \Delta B} \end{aligned}$$

Considérons deux événements, A et B, tels que la réalisation de A nous assure celle de B. Sur le plan logique, nous dirons que A implique B (soit $A \Longrightarrow B$). Les éventualités de A sont donc des éventualités de B. On en déduit $A \subseteq B$:

A	implique	B
A	$\Longrightarrow$	B
A	$\subseteq$	B
A	inclus dans	B

Remarquons que l'implication dont il est question ici n'est pas une implication causale mais une implication logique. Pour préciser la distinction, considérons un exemple. Dans un festival de danses régionales, chaque danseur porte la tenue traditionnelle de sa région d'origine. Etre Breton *implique* le port du chapeau rond. Ici, l'implication est causale car il y a une cause matérielle dont la conséquence est le port du chapeau rond. Par contre, connaissant la règle, porter un chapeau rond *implique* que l'on est Breton. En réalité, l'habit ne fait pas le moine et ce n'est pas la présence du chapeau rond qui détermine le lieu de naissance. Ici, l'implication est logique ; nous sommes en présence d'une *inférence*.

3.2.3 Probabilités composées et probabilités totales

Introduisons la **probabilité de B conditionnée par** A que nous notons $P_{B/A}$. Supposons que l'épreuve se déroule en deux temps et qu'à l'issue du premier temps nous sachions que A sera certainement réalisé à la fin de l'épreuve. A partir de ce moment, tout se passe comme si l'éventualité qui va se réaliser est choisie dans A. La probabilité de B (conditionnée par A) est donc $N_{B\cap A}/N_A = P_{B\cap A}/P_A$. On définit donc

$$P_{B/A} \underset{déf}{=} \frac{P_{B\cap A}}{P_A} \Longleftrightarrow \boxed{P_{B\cap A} = P_A \times P_{B/A}}$$

Les événements A et B sont indépendants si la probabilité de B est indépendante de la réalisation ou non de A. Soit $P_{B/A} = P_{B/A'}$.

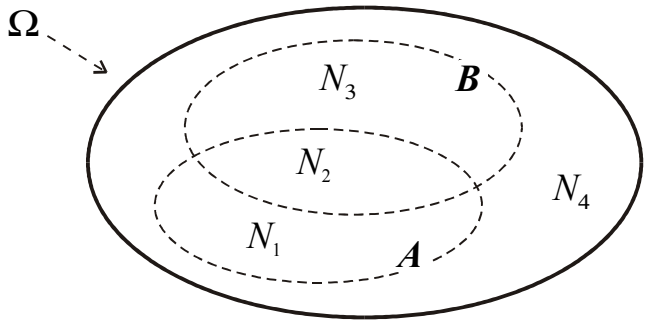


Figure 3-3

Dans les tableaux ci-dessous, nous donnons le cardinal des divers événements représentés sur la figure 3-3 ainsi que l'expression des diverses probabilités

ensemble E :	A	B	A'	$B \cap A$	$B \cap A'$
cardinal de E :	$N_1 + N_2$	$N_3 + N_2$	$N_3 + N_4$	N_2	N_3

$P_B = \dfrac{N_3 + N_2}{N_1 + N_2 + N_3 + N_4}$	$P_{B/A} = \dfrac{N_2}{N_1 + N_2}$	$P_{B/A'} = \dfrac{N_3}{N_3 + N_4}$

La relation $P_{B/A} = P_{B/A'}$ implique $\dfrac{N_2}{N_1+N_2} = \dfrac{N_3}{N_3+N_4}$.
Cependant, $\dfrac{N_2}{N_1+N_2} = \dfrac{N_3}{N_3+N_4}$ implique $\dfrac{N_2}{N_1+N_2} = \dfrac{N_3}{N_3+N_4} = \dfrac{N_2+N_3}{N_1+N_2+N_3+N_4}$, ce qui s'écrit encore sous la forme $P_{B/A} = P_{B/A'} \Longrightarrow P_{B/A} = P_{B/A'} = P_B$. On en déduit $P_{B\cap A} = P_A \times P_{B/A} = P_A \times P_B$ ainsi que $P_{B\cap A} = P_B \times P_{A/B}$ et par conséquent $P_{A/B} = P_A = P_{A/B'}$. Le résultat obtenu est connu comme *l'axiome des probabilités composées*. Dans la présentation que nous avons privilégié, c'est une conséquence des définitions posées. Cet "axiome" se résume ainsi :

$$\boxed{A \textbf{ et } B \textbf{ indépendants} \underset{déf}{\Longleftrightarrow} P_{B/A} = P_{B/A'} \Leftrightarrow P_{A/B} = P_{A/B'} \Leftrightarrow P_{A\cap B} = P_A \times P_B} \tag{3.2}$$

Rappelons que $P_{A\cap B}$ peut aussi s'écrire "Probabilité de $\{A$ et $B\}$".

Considérons maintenant l'événement $\{A$ ou $B\}$ ("ou" inclusif) ; il est possible d'en exprimer la probabilité sous la forme

$$P_{A\cup B} = \frac{N_1+N_2+N_3}{N_1+N_2+N_3+N_4} = \frac{N_1+N_2}{N_1+N_2+N_3+N_4} + \frac{N_2+N_3}{N_1+N_2+N_3+N_4} - \frac{N_2}{N_1+N_2+N_3+N_4}$$

$$\boxed{P_{A\cup B} = P_A + P_B - P_{A\cap B}}$$

Un avatar de cette relation concerne les événements incompatibles, c'est-à-dire ceux qui ne peuvent se réaliser simultanément :

$$\boxed{A \textbf{ et } B \textbf{ incompatibles} \underset{déf}{\Longleftrightarrow} A \cap B = \varnothing \Leftrightarrow P_{A\cap B} = 0 \iff P_{A\cup B} = P_A + P_B}$$

3.3 Probabilités et statistiques

3.3.1 Le double langage

Considérons une épreuve dont le résultat est une éventualité e. Tous les e sont équiprobables et à chaque e est associée la valeur d'une variable X.

Soit $\{X_k\}$ l'ensemble des valeurs possibles et N_k le nombre d'éventualités auxquelles est associée la même valeur X_k de la variable. On pose $N = \sum_k N_k$. Le nombre total d'éventualités est N.

Considérons l'événement E_k caractérisé par la réalisation de l'égalité $X = X_k$ Soit P_k la probabilité de E_k. Par définition, il vient $P_k = proba\,[X = X_k] = \dfrac{N_k}{N}$.

Considérons l'univers Ω : le cardinal de Ω est N. D'un point de vue statistique, la fréquence absolue de X_k est N_k, sa fréquence relative est $\dfrac{N_k}{N} = P_k$: c'est la probabilité de l'événement $\{X = X_k\}$ dans l'épreuve considérée.

Il existe donc une correspondance entre les probabilités de tel ou tel résultat dans l'épreuve considérée et les statistiques concernant l'univers.

Le résultat de l'épreuve n'est pas connu à l'avance. Pour cette raison, la variable X est qualifiée de ***"aléatoire"***. Cependant, si on ne connaît pas la valeur de X *a priori*, on connaît *a priori* la probabilité de l'événement $\{X = X_k\}$: c'est la proportion d'éventualités pour lesquels $X = X_k$ dans la population Ω.

De telles probabilités sont appelées ***probabilités a priori*** car elles sont connues avant le déroulement de l'épreuve (ou, plus précisément, en l'absence de toute information, même parcellaire, sur le résultat ou le déroulement de l'épreuve).

On définit ***l'espérance mathématique*** de la variable X, notée $E(X)$ et la ***variance*** de X, notée $\mathcal{V}(X)$

$$\begin{aligned} E[X] &= \sum_k P_k \times X_k \\ \mathcal{V}(X) &= E\left[\Big(\ X - E[X]\ \Big)^2\right] = \sum_k P_k \times (X_k - E[X])^2 \end{aligned}$$

L'écart-type, σ, est la racine carré de la variance.

Nous verrons† que la moyenne des observations de X (si celles-ci sont assez nombreuses), est probablement très proche de l'espérance mathématique, $E[X]$. Par contre, si on effectue une seule observation, le résultat est susceptible de présenter un écart à cette moyenne, montrant ainsi que le résultat peut varier d'une épreuve à l'autre. Il faut y trouver l'origine des mots "écart-type" et "variance".

Au mot "espérance" est généralement associée une connotation positive dont il faut le dépouiller dans le présent contexte. L'espérance de X est ici la valeur que l'on compte obtenir (que l'on espère obtenir) en effectuant la moyenne de nombreuses observations†.

La moyenne est encore appelée "*moment d'ordre 1*" et la variance "*moment centré d'ordre 2*". De façon générale, on définit le "***moment d'ordre*** N" que l'on note $\mathbb{M}_N$ et le "***moment centré d'ordre*** N" que l'on note $\mathcal{M}_N$.

$$\begin{aligned} \mathbb{M}_N &= E\left[\left(X^N\right)\right] = \sum_k P_k \times (X_k)^N \\ \mathcal{M}_N &= E\left[\left(X - E[X]\right)^N\right] = \sum_k P_k \times (X_k - E[X])^N \end{aligned}$$

Le tableau ci-dessous résume les correspondances entre le vocabulaire des probabilités et celui des statistiques.

Statistiques		**Probabilités**
valeur de la variable : X_k		événement $\{X = X_k\}$
Cardinal de Ω	N	Nb de cas total
fréquence absolue	N_k	Nb de cas favorables
fréquence relative	N_k/N	probabilité
moyenne de X	$\overline{X} = \sum_k P_k X_k = E(X)$	espérance mathématiques
? ?	$\sigma^2 = \sum_k P_k \left(X_k - \overline{X}\right)^2 = \mathcal{V}(X)$	variance
écart quadratique moyen	$\sigma = \sqrt{\mathcal{V}(X)}$	écart-type

La correspondance que nous avons établi entre statistiques et probabilités permet l'emploi de l'un ou l'autre langage. Pour cette raison, nous ferons appel aux deux vocabulaires sans trop nous soucier de rigueur tant qu'aucune ambiguïté n'est à redouter.

Espérance mathématique et écart-type suivent les mêmes lois que moyenne et écart quadratique moyen d'une statistique :

$$\boxed{E(\lambda) = \lambda\ , \quad E(\lambda X) = \lambda E(X)} \tag{3.3}$$

$$\boxed{\mathcal{V}(X) \underset{déf}{=} E([X - E(X)]^2) = E(X^2) - (E(X))^2\ , \quad \mathcal{V}(\lambda X) = \lambda^2 \mathcal{V}(X)} \tag{3.4}$$

†voir la loi des grands nombres page 57.

où $\lambda = cte$.

$$\boxed{E\left(X+Y\right) = E(X) + E(Y)} \tag{3.5}$$

$$\boxed{\mathcal{V}(X+Y) = \mathcal{V}\left(X\right) + \mathcal{V}\left(Y\right) + 2cov\left(X,Y\right)} \tag{3.6}$$

avec

$$\boxed{cov\left(X,Y\right) \underset{déf}{=} E\left(\ \ [X - E\left(X\right)] \times [Y - E(Y)]\ \ \right) = E\left(XY\right) - E(X) \times E(Y)} \tag{3.7}$$

Dans le présent contexte des probabilités, on définit le coefficient de corrélation de deux variables aléatoires de la même manière qu'en statistique :

$$\boxed{\rho \underset{déf}{=} \frac{cov\left(X,Y\right)}{\sqrt{\mathcal{V}(X)}\sqrt{\mathcal{V}(Y)}}} \tag{3.8}$$

De façon générale, si Z est la somme $Z = X_1 + X_2 + ...$ où X_1, X_2, etc sont des variables aléatoires d'espérance $E(X_1)$, $E(X_2)$, etc il vient

$$E(Z) = E(X_1) + E(X_2) + ...$$

En particulier, si les variables aléatoires suivent la même loi de probabilité que la variable X, il vient :

$$\Sigma \underset{déf}{=} \frac{1}{n}\left(X_1 + X_2 + ... + X_n\right) \Rightarrow E\left(\Sigma\right) = E(X) \tag{3.9}$$

3.3.2 Variable aléatoire continue

Nous considérons ici le cas d'une variable numérique, X, susceptible de prendre une valeur x, réelle quelconque, à l'issu d'une épreuve. Le résultat n'est pas déterminé avant l'épreuve, seule est connue la loi de probabilité suivie par X. Cette loi est définie par sa ***fonction de répartition***, $f(x)$ ou sa ***densité de probabilité*** $p(x)$.

Remarquons que les probabilités introduites précédemment (cf. 3.1) étaient nécessairement des fractions rationnelles. Ce n'est plus le cas ici.

Les propriétés satisfaites par $f\left(x\right)$ et $p\left(x\right)$ sont les suivantes :

- $f(x)$ est la probabilité pour que le résultat de l'épreuve satisfasse $X \leq x$. La fonction $f(x)$ est positive et monotone croissante avec $\lim\limits_{x\to-\infty} f(x) = 0$ et $\lim\limits_{x\to\infty} f(x) = 1$.
- $p(x)dx$ est la probabilité de l'événement $x < X \leq x + dx$; ce qui implique $p(x) = \dfrac{df}{dx} \geq 0$ (lorsque $\dfrac{df}{dx}$ est continu).

La loi de probabilité associée à l'épreuve considérée est complètement définie par la connaissance soit de $f\left(x\right)$, soit de $p\left(x\right)$.

Considérons la variable aléatoire $F(X)$ où F est une fonction donnée. Dans le cas d'une variable continue, l'espérance mathématique de $F(X)$ est donnée par la relation

$$E(F) = \int_{-\infty}^{\infty} p(x)\ F(x)\ dx$$

tandis que la variance de F est

$$\mathcal{V}\left(F\right) = \int_{-\infty}^{\infty} p(x)\ \left(F(x) - E(F)\right)^2\ dx$$

en outre, les propriétés 3.3 et 3.4 ci-dessus restent valides dans le cas présent.

3.3.3 Couples de variables aléatoires continues

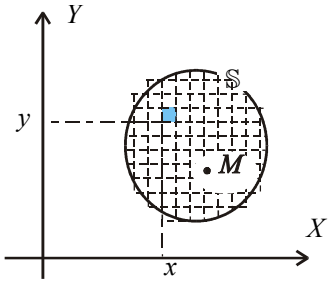
Considérons le cas où le résultat de l'épreuve est un couple de valeurs continues (X, Y). Le résultat de l'épreuve peut donc être représenté par un point M, du plan (X, Y).

Considérons un rectangle élémentaire $X \in (x, x + dx)$ et $Y \in (y, y + dy)$. La probabilité pour que M appartienne à ce rectangle élémentaire est donnée sous la forme :

$$proba\,[X \in (x, x + dx) \text{ et } Y \in (y, y + dy)] = p\,(x, y)\,dxdy$$

$p\,(x, y)$ est la ***densité de probabilité.*** C'est une fonction positive.

Considérons une région, $\mathbb{S}$, du plan (cf. figure 3-4). Cette région peut être considérée comme la réunion de rectangles élémentaires disjoints[‖].



$Proba\,[M \in \mathbb{S}] = \iint_{\mathbb{S}} p(x,y)dxdy$

Figure 3-4

La densité de probabilité $p(x, y)$ est définie par la relation suivante satisfaite pour tout $\mathbb{S}$:

$$\boxed{proba\,[M \in \mathbb{S}] = \iint_{\mathbb{S}} p\,(x, y)\,dxdy}$$

où $p(x, y) \geq 0$.

De façon certaine $M \in \mathbb{R}^2$, c'est-à-dire que M est nécessairement quelque part dans le plan. On en déduit $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p\,(x, y)\,dxdy = 1$.

Toute fonction $p(x, y)$ est la densité de probabilité d'une loi ssi

$$p(x, y) \geq 0 \quad \text{et} \quad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p\,(x, y)\,dxdy = 1$$

La loi de probabilité est alors complètement définie par la densité de probabilité $p\,(x, y)$.

Etant donnée une fonction $F(X, Y)$, on définit l'espérance de F :

$$E\,(F) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(x, y)\; p\,(x, y)\,dxdy$$

Les définitions et les propriétés 3.3 à 3.8 restent valides dans le cas présent.

Supposons que la valeur x de la variable X soit connue. Dans ces conditions, on définit la densité de probabilité de y conditionnée par $X = x$ comme une fonction de y, proportionnelle à $p(x, y)$ que l'on écrit $p_{/x}(y)$. Pour représenter une densité de probabilité,

‖ "*disjoints*" signifie "*dont l'intersection est nulle*"

cette fonction doit satisfaire la relation $\int_{-\infty}^{\infty} p_{/x}(u)\ du = 1$. On définit donc **la densité de probabilité de Y conditionnée par $X = x$** :

$$p_{/x}\,(y) = \frac{p\,(x,y)}{\int_{-\infty}^{\infty} p\,(x,u)\,du}$$

On déduit de cette définition la moyenne de Y conditionnée par X :

$$E_{/X=x}\,(Y) = \int_{-\infty}^{\infty} y \times p_{/x}\,(y) \times dy = \frac{\int_{-\infty}^{\infty} y\ p(x,y)\ dy}{\int_{-\infty}^{\infty} p\,(x,y)\,dy}$$

$E_{/X=x}\,(Y)$ est une fonction de x que nous notons $e(x)$: c'est **la fonction de régression de Y en X.**

Supposons que la grandeur Y soit une fonction de la grandeur X (par exemple $Y = g(X)$). L'épreuve considérée consiste à mesurer Y pour diverses valeurs fixées de X. La valeur, x, de X étant fixée, le résultat est une valeur y qui varie d'une mesure à l'autre compte tenu des erreurs expérimentales. On s'attend à ce que la moyenne des erreurs soit nulle et par conséquent que l'espérance de Y, conditionnée par $X = x$, soit égale à $g(x)$. La fonction de régression de Y en X est donc la fonction $e(x) = g(x)$.

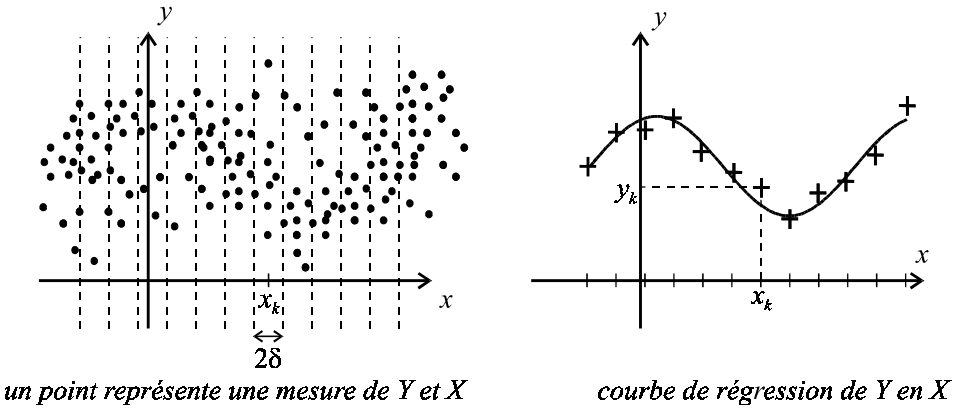


Figure 3-5

Chaque point de la figure 3-5 représente le résultat d'une mesure. Le plan est considéré comme la réunion de bandes étroites, centrées en x_k, de largeur 2δ. La moyenne de Y dans chacune de ces bandes est la valeur y_k. Les points de coordonnées (x_k, y_k) peuvent être considérés pratiquement comme des points du graphe de la fonction de régression de Y en X.

En règle générale cette courbe n'est pas une droite. Cependant, comme nous l'avons signalé au paragraphe 2.3 page 23 le nuage de points de coordonnées (x_k, y_k) se présente souvent sans structure apparente ; l'ajustement recherché est alors un ajustement linéaire qui conduit à la droite de régression de Y en X.

Si la seule valeur de la variable X présente un intérêt ; on en obtient la densité de probabilité $p_m(x)$, appelée ***"densité de probabilité marginale"*** :

$$p_m(x) = \int_{-\infty}^{\infty} p(x,y)\ dy$$

Exemple : Considérons la loi de probabilité caractérisée par la densité

$$p\,(x,y) = Ae^{-x^2/\alpha^2} \times e^{-(y-g(x))^2/\beta^2}$$

Rappelons les relations $\int_{-\infty}^{\infty} e^{-t^2}dt = \sqrt{\pi} =$ et $\int_{-\infty}^{\infty} t^2 e^{-t^2}dt = \frac{1}{2}\sqrt{\pi}$.

- Le coefficient A est défini par la condition :

$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)dxdy = 1$. Il vient $A = \dfrac{1}{\pi\alpha\beta}$.

- La densité de probabilité de y, conditionnée par X est :

$$p_{/x}\,(y) = \frac{p\,(x,y)}{\int_{-\infty}^{\infty} p\,(x,u)\,du} = \frac{e^{-(y-g(x))^2/\beta^2}}{\beta\sqrt{\pi}}$$

- La régression de Y en X donne :

$E_{/X=x}\,(Y) = e(x) = \int_{-\infty}^{\infty} y \times p_{/x}\,(y) \times dy = \int_{-\infty}^{\infty} y \times \dfrac{e^{-(y-g(x))^2/\beta^2}}{\beta\sqrt{\pi}} \times dy = g(x)$. Si on considère que Y est une fonction de X, il est légitime de poser $Y = g(X)$.

- La répartition marginale de X est caractérisée par la densité de répartition :

$p_m(x) = \int_{-\infty}^{\infty} p(x,y)\ dy = \dfrac{e^{-x^2/\alpha^2}}{\alpha\sqrt{\pi}}$

- La moyenne de X est :

$E(X) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\ p\,(x,y)\,dxdy = \int_{-\infty}^{\infty} x\ p_m(x)dx = \int_{-\infty}^{\infty} x\ \dfrac{e^{-x^2/\alpha^2}}{\alpha\sqrt{\pi}}dx = 0.$

- La variance de X est $\mathcal{V}(X) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^2\ p\,(x,y)\,dxdy$

(En effet $E(X) = 0$ ici). Il vient $\mathcal{V}(X) = \int_{-\infty}^{\infty} x^2\ \ p_m(x)dx = \int_{-\infty}^{\infty} x^2\ \dfrac{e^{-x^2/\alpha^2}}{\alpha\sqrt{\pi}}dx = \dfrac{\alpha^2}{2}$.

3.3.4 Variables indépendantes

On considère les variables aléatoires X et Y. L'épreuve consiste à mesurer la valeur de ces variables. Ces variables peuvent être discrètes ou continues. Dans le cas discret, les valeurs possibles sont respectivement notées X_k et Y_k tandis que dans le cas continu, les valeurs possibles sont notées x et y.

Les variables X et Y sont indépendantes lorsque la loi de probabilité de Y conditionnée par X est indépendante de la valeur, x, de X. On démontre que cette propriété est équivalente à la propriété suivante :

Indépendance :

Variables discrètes

Proba $[X = X_k] \underset{déf}{=} P_{X_k}$,
Proba $[Y = Y_j] \underset{déf}{=} P_{Y_j}$

Proba$\{X = X_k$ et $Y = Y_j\} = P_{X_k} \times P_{Y_j}$

Variables continues

Proba $[x, x+dx] \underset{déf}{=} q(x)dx$,
Proba $[y, y+dy] \underset{déf}{=} \pi(y)\ dy$

Proba$\{[x, x+dx]$ et $[y, y+dy]\} = q(x)\ \pi(y)\ dxdy$
soit $p(x,y) = q(x)\ \pi(y) =$ densité de probabilité

Dans le cas de variables discrètes, on définit E_{X_k} comme l'événement $X = X_k$ et E_{Y_j} comme l'événement $Y = Y_j$. On vérifie sans difficulté que l'indépendance des variables X et Y est équivalente à l'indépendance des événements E_{X_k} et E_{Y_j}, telle que nous l'avons définie ci-dessus (cf. 3.2), pour toutes valeurs de X_k et de Y_j.

Un cas particulièrement important concerne les épreuves composées de deux épreuves élémentaires indépendantes. Le résultat de la première épreuve est la variable

X; la seconde épreuve se déroule d'une façon définie *a priori*, toujours la même, quelle que soit la valeur obtenue pour X. L'ensemble des résultats possibles est l'ensemble des couples de valeurs (X_k, Y_j) qui constitue l'univers. Les variables X et Y sont alors stochastiquement indépendantes au sens où cette indépendance a été définie au paragraphe 2.5.3 page 33.

A l'instar de ce que l'on rencontre en statistiques, les variables aléatoires indépendantes présentent un coefficient de corrélation nul mais la réciproque n'est pas vraie. Par conséquent, de 3.5 et 3.6, on déduit

$$\boxed{\begin{array}{l} E\left(X+Y\right) = E(X) + E(Y) \\ \mathcal{V}(X+Y) = \mathcal{V}\left(X\right) + \mathcal{V}\left(Y\right) \end{array}} \tag{3.10}$$

où X et Y sont des variables indépendantes.

Nous considérons maintenant les variables X_1, X_2, etc, X_n. Ici X_k représente une variable aléatoire, ce n'est pas la valeur possible d'une variable X, comme ce l'était précédemment.

Le résultat 3.10 se généralise au cas de la variable $Z = X_1 + X_2 + ... + X_n$ où X_1, X_2, etc sont des variables aléatoires indépendantes qui suivent la même loi de probabilité que la variable X, dont l'écart-type est σ.

Les variables étant supposées indépendantes, il vient

$$\mathcal{V}\left(Z\right) = \mathcal{V}(X_1) + \mathcal{V}(X_2) + ... + \mathcal{V}(X_n) = n\ \mathcal{V}(X) = n\ \sigma_X^2$$

Considérons la variable $\Sigma = \dfrac{1}{n}\left(X_1 + X_2 + ... + X_n\right) = \dfrac{1}{n}Z$, introduite ci-dessus (cf. relation 3.9). Suivant la dernière relation 3.4 il vient

$$\mathcal{V}\left(\Sigma\right) = \frac{1}{n^2}\mathcal{V}\left(Z\right) = \frac{1}{n}\mathcal{V}\left(X\right) \text{ soit } \sigma_\Sigma^2 = \frac{1}{n}\sigma^2 \tag{3.11}$$

3.4 Présentation axiomatique et théorème de Bayes

Dans la présentation axiomatique, on ne tente pas de relier les probabilités à une quelconque épreuve physique. On introduit un ensemble Ω et une famille de parties de Ω, notée $\mathcal{F}$.

La famille $\mathcal{F}$ contient Ω et $\varnothing$. Les ensembles A et B appartenant à $\mathcal{F}$, les ensembles suivants y appartiennent également : A' et B', $A \cap B$, $A \cup B$ et $A \Delta B$. Dans ces conditions on dit que $\mathcal{F}$ forme une algèbre de Boole.

$\mathcal{F}$ est une σ-algèbre, $\mathcal{A}$, si la réunion de toute suite dénombrable de $\mathcal{F}$ appartient à $\mathcal{F}$.

Une mesure sur $(\Omega, \mathcal{A})$ est une application μ, de $\mathcal{A}$ sur l'ensemble des réels telle que $\mu\left(A\right) \geq 0$ pour tout $A \in \mathcal{A}$ et $\mu\left(A \cup B\right) = \mu\left(A\right) + \mu\left(B\right)$ pour tous A et B appartenant à $\mathcal{A}$ tels que $A \cap B = \varnothing$. L'application μ définit ce que l'on appelle **une mesure** sur $\mathcal{A}$.

Lorsque μ est borné ($\mu\left(\Omega\right) < \infty$), on peut poser $\mu\left(\Omega\right) = 1$; dans ces conditions $(\Omega, \mathcal{A}, \mu)$ définit ***un espace de probabilité*** et dans ce cas $\mu\left(A\right)$ est noté $P(A)$.

On démontre alors les propriétés usuelles telles que $A \subseteq B \Rightarrow P\left(A\right) \leq P\left(B\right)$ ou encore $P\left(\varnothing\right) = 0$.

On définit la probabilité de B, conditionnée par A : $P_{B/A} = \dfrac{P\left(B \cap A\right)}{P\left(A\right)}$; on en déduit

$$P(A)\ P_{B/A} = P\left(A \cap B\right) = P(B)\ P_{A/B}$$

Démontrons maintenant le théorème de Bayes, ou théorème des causes. Considérons les ensembles disjoints $A_1, A_2, ...A_k, ...$ dont l'ensemble Ω est la réunion, c'est-à-dire tel que $\Omega = A_1 + A_2 + ...$. On remarquera que l'on utilise le symbole + pour la réunion d'ensembles disjoints. On en déduit $B = B \cap \Omega = B \cap A_1 + B \cap A_2 + ... = \sum_k B \cap A_k$; suivant les axiomes précédents, il vient $P(B) = P\left(\sum_k B \cap A_k\right) = \sum_k P(B \cap A_k) = \sum_k P(A_k)\ P(B/A_k)$.

D'autre part, en utilisant les définitions on trouve $P(B \cap A_j) = P(A_j)\ P(B/A_j) = P(B)\ P(A_j/B)$ et par conséquent $P(A_j/B) = \dfrac{P(B \cap A_j)}{P(B)}$

$$\Omega = A_1 + A_2 + ... \Longrightarrow P(A_j/B) = \frac{P(A_j)\ P(B/A_j)}{\sum_k P(A_k)\ P(B/A_k)}$$

Cette expression constitue le théorème de Bayes.

Pour interpréter ce théorème, supposons que l'événement B puisse être la conséquence d'un événement A_k, avec une certaine probabilité.

Considérons un exemple.

Dans une région, on sait qu'il y a 3 souches différentes de grippe (I, II et III) qui affectent respectivement 10%, 3% et 5% de la population. Parmi les personnes qui sont infectées par la grippe de type I, il y en a 40% qui ont un symptôme de toux, 70% pour le type II et 85 % pour le type III. De plus, on sait que dans la population qui n'est pas infectée par la grippe, il y a 15% des personnes qui ont un symptôme de toux.

Je suis médecin ; un patient se présente à mon cabinet. Il tousse. Quelle est la probabilité pour qu'il ait une grippe de type I ?

Tout d'abord, précisons le rapport entre cet exemple et la théorie.

La population est une population humaine et les individus sont les êtres humains qui habitent la région. *Ce n'est pas très précis mais c'est suffisant ici.*

L'épreuve probabiliste consiste à tirer au hasard un individu dans cette population.

L'ensemble des individus qui ont la grippe de type I, II, ou III constitue l'événement A_1, A_2 ou A_3. Ceux qui n'ont pas la grippe forment l'événement A_4.

Remarquons que tous les individus sont classés dans une catégorie et une seule : $A_i \cap A_j = \varnothing$ et $\sum_k A_k = \Omega$.

Les probabilités des divers événements sont :

$$P(A_1) = 0,1 \quad P(A_2) = 0,03 \quad P(A_3) = 0,05 \quad P(A_4) = 0,82 \quad \text{Total : } 1$$

L'événement B est constitué par l'ensemble des individus qui toussent.

Les probabilités conditionnelles sont

$$P(B/A_1) = 0,4 \quad P(B/A_2) = 0,7 \quad P(B/A_3) = 0,85 \quad P(B/A_4) = 0,15$$

La proportion d'individus qui toussent est donc $P(B) = \sum_k P(A_k)\ P(B/A_k)$

$P(B) = 0,1 \times 0,4 + 0,03 \times 0,7 + 0,05 \times 0,85 + 0,82 \times 0,15 = 0,226$

Parmi les individus qui toussent, la proportion d'individus qui ont la grippe de type I est $P(A_1/B)$. Cette proportion est la probabilité pour que notre patient ait la grippe de type I *en admettant que ce patient soit un individu tiré au hasard dans la population*

concernée. Dans ces conditions, l'application du théorème de Bayes donne $P(A_1/B) = \dfrac{P(A_1)\ P(B/A_1)}{P(B)} = \dfrac{0,1 \times 0,4}{0,226} \simeq 0,18$. La probabilité qu'une grippe de type I soit la cause de la toux est 18%. Considérant l'ensemble des cas possibles, il vient :

$$P(A_1/B) \simeq 0,18 \quad P(A_2/B) \simeq 0,09 \quad P(A_3/B) \simeq 0,18 \quad P(A_4/B) \simeq 0,55$$

S'il faut avoir une opinion (c'est bien sûr le rôle du médecin) celle-ci sera que la toux n'est pas due à la grippe (il se trompera alors presque une fois sur deux!!!).

De la même manière, on peut étudier le cas d'un "patient" qui ne tousse pas (malades potentiels, cible d'un marketing improbable qui faisait rire avec le D^r Knock** mais qui, bien présent aujourd'hui, est devenu inquiétant.).

On rencontre ici un exemple d'aide à la décision apporté par les statistiques. Vous pouvez constater que l'interprétation des faits repose sur certaines approximations et certaines hypothèses soulignées en italique. En particulier, il n'est pas certain que la clientèle d'un médecin donné concerne un échantillon représentatif de la région étudiée.

Si le diagnostique s'appuie sur le théorème de Bayes, il conduit le médecin à se tromper souvent. L'aide apportée ne conduit à aucune certitude. Dans le cas présent, un jeu de pile ou face donnerait à peu près le même résultat. Si le traitement n'est pas nocif, on peut traiter le patient pour une grippe de type I ou de type III. Ce sera inutile dans près de 80% des cas mais ce sera utile parfois. De plus, avec une ordonnance le patient est content !

Enfin, si la grippe de type II est mortelle, peut-on prendre le risque de ne pas la traiter ? Même si c'est inutile dans près de 90% des cas ?

En présence de l'étude précédente, la bonne attitude consiste, bien sûr, à rechercher d'autres symptômes que la toux pour décider si le patient a contracté une grippe. C'est la démarche médicale généralement suivie.

$P(A_k)$ est une "**probabilité *a priori***" (voir ci-dessus page 45).

$P(B/A_k)$ est une *probabilité conditionnelle :* la probabilité de B conditionnée par A_k.

$P(A_k/B)$ est la probabilité pour que A_k soit satisfait sachant que B est satisfait. Une telle probabilité est une "**probabilité *a posteriori***". C'est une probabilité conditionnelle qui est considérée ici comme la probabilité pour que l'événement A_k soit la cause de B : les probabilités $P(A_j/B)$ sont les *probabilité des causes.*

3.5 Conclusion

Etant donnée une population, Ω, il lui correspond une épreuve dans laquelle les probabilités des divers résultats sont les fréquences relatives observées dans Ω. Cette épreuve étant construite, il existe une correspondance entre probabilités et statistiques si bien qu'un double langage est possible pour décrire Ω. Toutefois, les probabilités se réfèrent à une opération "dynamique" dans laquelle la réalité se découvre petit à petit au fur et à mesure de la répétition des épreuves, tandis que les statistiques se réfèrent à une situation statique où l'on peut (en principe) appréhender la réalité "d'un coup d'oeil". Ces deux points de vue sont très différents. Cependant, lorsque la population est très nombreuse, ou lorsqu'elle est définie en compréhension, cette appréhension globale, "d'un seul coup d'oeil" n'est pas possible.

Comment construire cette épreuve qui, pour décrire une population donnée, conduit au langage des probabilités ? Nous avons pris comme exemple le tirage du loto tel que

**Célèbre pièce de Jules Romains, immortalisée par le jeu de Louis Jouvet.

nous le présente la télévision. Cette méthode n'est certainement pas applicable lorsque le nombre d'éventualités équiprobables devient très grand. C'est le cas du contrôle de qualité d'un produit fabriqué à des milliers d'exemplaires ou encore d'un sondage pré-électoral concernant quelques millions d'électeurs. Nous y reviendrons ultérieurement (page 60).