



جامعة الإخوة منتوري قسنطينة I
Frères Mentouri Constantin I University
Université Frères Mentouri Constantine I

DEPARTEMENT BIOCHIMIE

1ere année master biochimie appliquée et biochimie

COURS BIostatistique

RESPONSABLE DU MODULE: DR.ZEGHBID NASSIM LOTFI

Année Universitaire : 2020

Corrélation et régression linéaire

la régression linéaire:

Lorsqu'il existe une relation logique entre deux variables X et Y, il est intéressant de l'exprimer sous la forme d'un modèle mathématique qui sert à estimer la valeur Y correspondant à une valeur donnée de X. C'est ce qu'on appelle l'analyse de régression (ou théorie de la régression). Lorsque le nuage statistique (de points) indique qu'il existe une corrélation entre deux variables, on exprime mathématiquement cette relation par l'équation d'une droite.

$$Y = aX + b$$

On appelle régression linéaire l'ajustement d'une droite au nuage statistique d'une série de couples de données (x_i, y_i) . X (variable indépendante) tandis que Y (variable dépendante).

X = variable explicative / Y = variable expliquée

X = variable indépendante / Y = variable dépendante

Le fait de trouver l'équation de la droite mettant en relation deux variables nous fournira un outil de prévision ou d'estimation. En effet, à partir de cette équation, on pourra estimer ou prévoir les valeurs d'une variable dite dépendante en fonction des valeurs prises par l'autre variable dite indépendante. La méthode pour y parvenir est la suivante : Considérons n couples de données provenant de l'étude de deux variables statistiques X et Y.

La méthode pour y parvenir est la suivante :

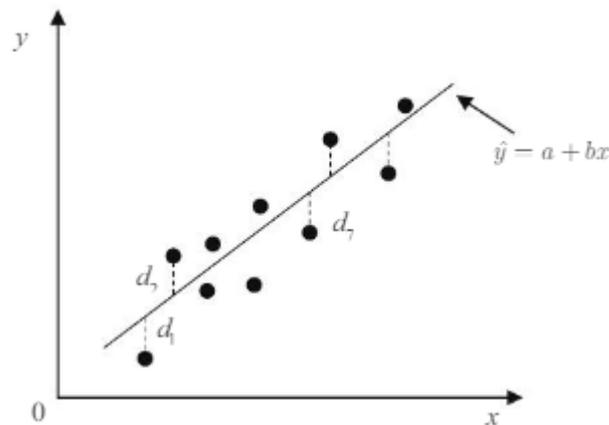
Considérons n couples de données provenant de l'étude de deux variables statistiques X et Y.

X (variable indépendante)	x_1	x_2	x_3	...	x_n
Y (variable dépendante)	y_1	y_2	y_3	...	y_n

On représente ces couples par un nuage statistique. Il s'agit maintenant de trouver une droite notée $Y = aX + b$, pouvant représenter convenablement la relation ou la tendance

se manifestant entre la variable Y (variable dépendant) et la variable X (variable indépendante).

Le graphique suivant illustre la situation :



Exemple :

Etude de la relation entre la tension artérielle et l'âge d'un individu

Objectif On souhaite savoir si, de façon générale, l'âge a une influence sur la tension artérielle et sous quelle forme cette influence peut être exprimée. Le but est d'expliquer au mieux comment la tension artérielle varie en fonction de l'âge et éventuellement de prédire la tension à partir de l'âge.

Population et variables étudiées

Population générale d'individus. Sur cette population, on définit deux variables.

La variable Y : variable tension ; c'est la variable à expliquer, appelée encore variable à régresser, variable réponse, variable dépendante (VD).

La variable X : variable âge ; c'est la variable explicative, appelée également régresseur, variable indépendante (VI).

La méthode de calcul de la régression linéaire et la méthode des moindres carrés

Tout d'abord on cherche l'équation

$$Y = aX + b$$

$$\hat{a} = \frac{n \sum X_i y_i - \sum X_i \sum y_i}{n \sum X_i^2 - (\sum X_i)^2}$$

$$\hat{b} = \bar{y} - \hat{a}\bar{X} = \frac{\sum y_i}{n} - \hat{a} \frac{\sum X_i}{n} = \hat{b}$$

Pour x et y la moyenne arithmétique avec simple donné

N c le nombre de population statistique

Exemple 1 : les données suivante exprime le relation entre la consommation du sucre et le taux de glycémie

y taux de glycémie	10	15	5	4	3	13
X sucre	7	10	4	3	2	8

- 1- Trouvé l'équation de la régression linéaire de y sur x.
- 2- Quelle sera le niveau de la glycémie lorsque la consommation du sucre 20 unité.
(X=20)

Solution

y	x	X y	x²
10	7	70	49
15	10	150	100
5	4	20	16
4	3	12	9
3	2	6	4
13	9	117	81
50	35	375	259

On calcule

$$Y = aX + b$$

$$\hat{a} = \frac{n \sum X_i y_i - \sum X_i \sum y_i}{n \sum X_i^2 - (\sum X_i)^2}$$

$$= \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2} = \frac{6.375 - (35.50)}{6.259 - (35)^2} = \frac{83.33}{54.83} = 1.52$$

$$\begin{aligned} \hat{b} = \bar{y} - \hat{a} \bar{X} &= \frac{\sum y_i}{n} - \hat{a} \frac{\sum X_i}{n} = \hat{b} \\ &= \frac{50}{6} - 1.52 \frac{35}{6} = -32 \end{aligned}$$

Alors l'équation de $Y = aX + b$ est de

$$\hat{y} = -3.2 + 1.52x$$

Estimation du taux de glycémie lorsque la consommation du sucre augmente de ($X=20$) elle devient **27.2**.

$$\hat{y} = -3.2 + 1.52 \cdot 20 = 27.2$$

Exemple 2 : soit les données suivantes qui désignent une relation entre deux variables

$$\sum y = 144 \quad \sum x = 103 \quad \sum xy = 2093$$

$$\sum y^2 = 3012 \quad \sum x^2 = 1531 \quad n=7$$

- 1- Trouver l'équation de la régression linéaire de y sur x .
- 2- Estimation du taux du gras lorsque la consommation du sucre augmente de ($X=10$)

Solution :

On calcule :

$$Y = aX + b$$

$$\hat{a} = \frac{n \sum X_i y_i - \sum X_i \sum y_i}{n \sum X_i^2 - (\sum X_i)^2}$$

$$\begin{aligned} &= \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2} = \frac{7.2093 - (103.144)}{7.1531 - (103)^2} \\ &= \frac{-181}{108} = -1.68 \end{aligned}$$

$$\hat{b} = \bar{y} - \hat{a} \bar{X} = \frac{\sum y_i}{n} - \hat{a} \frac{\sum X_i}{n} = \hat{b}$$

$$= \frac{144}{7} - (-1.68) \frac{103}{7} = 20.57 + 24.72 = 45.29$$

Alors l'équation de $Y = aX + b$ est de

$$\hat{y} = 45.29 - 1.68x$$

Estimation du taux de glycémie lorsque la consommation du sucre augmente de ($X=10$) elle devient **28.49**.

$$\hat{y} = 45.29 - 1.68 \cdot (10) = 28.49$$

2-corrélation :

On dira qu'il y a corrélation, ou dépendance, entre deux variables quantitatives X et Y si elles ont généralement tendance à varier toutes deux dans le même sens ou en sens contraire. Les caractéristiques d'une corrélation entre deux variables X et Y sont :

la forme :

Linéaire : les points du diagramme de dispersion ont tendance à se rapprocher d'une droite. C'est ce type de corrélation que nous étudions (exemples : graphiques 1, 2 et 6).

Non linéaire : les points du diagramme de dispersion ont tendance à se rapprocher d'une courbe (exemples : les graphiques 3 et 4).

le sens :

- positif : les deux variables varient dans le même sens : quand les valeurs de la variable X augmentent, celles de la variable Y augmentent aussi (exemples : les graphiques 1 et 3).

- négatif : les deux variables varient en sens contraires : quand les valeurs de la variable X augmentent, celles de la variable Y diminuent (exemples : les graphiques 2, 4 et 6).

l'intensité :

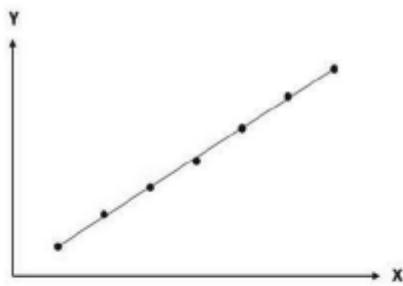
- parfaite : les points du diagramme de dispersion sont parfaitement alignés, dans le cas d'une corrélation linéaire, ou tous situés sur la courbe dans le cas d'une corrélation non linéaire.

- Une dépendance parfaite permet de déterminer, pour chaque valeur de la variable X, la valeur exacte de la variable Y qui lui est associée, et vice-versa

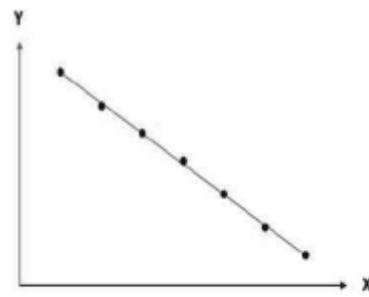
(exemples : les graphiques 1, 2 et 4).

- Une dépendance imparfaite (faible, forte, moyenne) : on constate une tendance moins forte des points du diagramme de dispersion à s'aligner ou à prendre la forme d'une courbe. Dans ce cas, on peut tout au plus estimer approximativement la valeur de la variable Y correspondant à une valeur donnée de la variable X (exemples : le graphique 3 et 6).

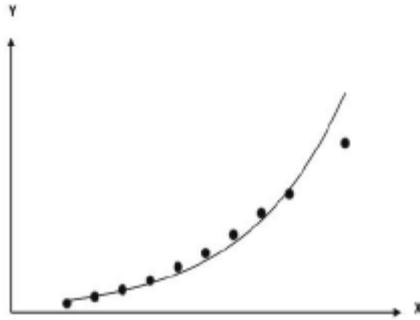
- nulle ou inexistante : les points sont complètement éparpillés dans le plan et ne semblent suivre aucune orientation ni s'approcher d'une droite ou une courbe. On dit alors que les variables sont indépendantes. Il est alors impossible d'estimer la valeur de Y correspondant à une valeur donnée de X, et vice-versa (exemples : le graphique 5).



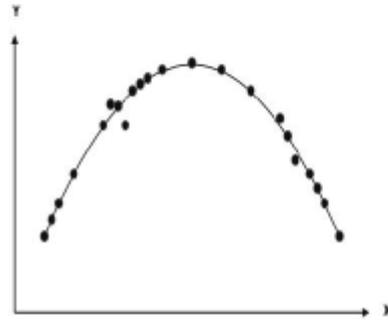
1. Corrélation linéaire positive parfaite



2. Corrélation linéaire négative parfaite



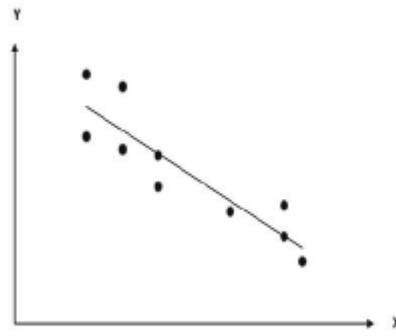
3. Corrélation non linéaire (exponentielle) positive forte



4. Corrélation non linéaire (parabolique ou quadratique) négative parfaite



5. Corrélation nulle



6. Corrélation linéaire négative faible

COEFFICIENT DE CORRÉLATIONS (OU DE PEARSON) :

Le nuage de points permet une analyse qualitative de la tendance à une relation linéaire entre les variables X et Y. Le coefficient de corrélation linéaire, ou coefficient de Pearson noté r , est un nombre sans dimension qui mesure quantitativement l'intensité de la corrélation (ou de la dépendance) linéaire entre les deux variables.

On le calcule à l'aide de la formule suivante :

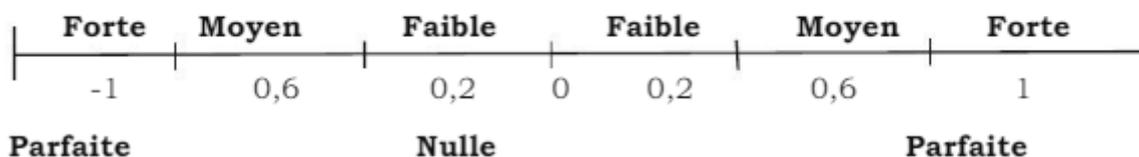
$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}}$$

Interprétation du coefficient de corrélation :

P1. $-1 \leq r \leq 1$.

P2. La corrélation linéaire et parfaite et positive de $r = 1$, parfaite et négative si $r = -1$ et nulle si $r = 0$,

P3. Dans le cas d'une corrélation linéaire positive ($r \geq 0$), plus la valeur de r est pris de 1, plus la corrélation entre X et Y est forte. Il en est de même pour une corrélation linéaire négative : plus la valeur est près de -1, plus la corrélation entre X et Y est forte. Le schéma suivant donne une idée de la force d'une corrélation en sciences humaines.



P4. La valeur $100 r^2$ donne le pourcentage de variation totale de Y qui s'explique par la dépendance de Y par rapport à X.

exemples : soit les données suivant qui montre une relation linéaire en deux variable quantitative

10	7	6	2	X
4	8	10	12	Y

- 1- Trouvé l'équation de la régression linéaire de y sur x.
- 2- Calculez le coefficient de corrélation Pearson.

Solution :

On calcule :

$$Y = aX + b$$

y^2	x_i^2	$x_i y_i$	y	x
144	4	24	12	2
100	36	60	10	6
64	49	56	8	7
16	100	40	4	10
324	189	180	34	25

$$\hat{a} = \frac{n \sum X_i y_i - \sum X_i \sum y_i}{n \sum X_i^2 - (\sum X_i)^2}$$

$$\hat{a} = \frac{720 - 850}{756 - 625} = \frac{130}{131} = -0.99$$

$$\hat{b} = \bar{y} - \hat{a} \bar{x} = \frac{34}{4} + 0.99 \frac{25}{4} = 8.5 \times 6.55$$

$$\hat{b} = 14.7$$

$$\hat{y} = 14.7 - 0.99 X_i$$

coefficient de corrélation Pearson

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

$$\gamma = \frac{n \sum X_i y_i - \sum X_i \sum y_i}{\sqrt{n \sum X_i^2 - (\sum X_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

$$\gamma = \frac{-130}{\sqrt{(756 - 625)(1296 - 1156)}} = \frac{130}{\sqrt{131 \times 140}}$$

$$\gamma = -0.962$$

Interprétation Corrélation fort parfaite

3- Corrélation de Spearman:

Définition et propriété:

Le coefficient de corrélation de Spearman, symbolisé par r_s , mesure le degré de liaison existant entre le classement des éléments selon la variable X et le classement selon Y .il

s'agit en fait d'un coefficient de corrélation de Pearson calculé non pas sur les valeurs de X et Y , mais sur les rangs des valeurs de X et Y .

On peut le calculez avec l'équation

$$\gamma_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

Avec :

Avant de trouve RX et RY on doit ordonné x et y soit de l'inferieure au supérieur / soit de supérieur a l'inferieure pour les deux l'important ces qu'ils soient ordonnés de la même façon

RX correspond au rang de la variable x soit de l'inferieure au supérieur / soit de supérieur a l'inferieure

RY correspond au rang de la variable y soit de l'inferieure au supérieur / soit de supérieur a l'inferieure

A noter que pour les deux variable x ou y le rang doit être de la même façon établie soit du supérieur a l'inferieure soit de l'inferieure au supérieur.

Di Correspond à l' écart = **RX-RY**.

Exemple : les données suivante définie une relation linaire entre deux variable x y

9	6	13	7	16	8	11	10	(X)
4	2	9	1	12	2	7	3	(y)

1-Calculez le coefficient de Spearman.

Solution

D_{i2}	d_i	RY	RX	Y	X
1	1	4	5	1	6
0	0	6	6	2	7
0.25	0.5	2.5	3	2	8
0	0	8	8	3	9
1	1	1	2	4	10
0	0	7	7	7	11
2.25	-1.5	2.5	1	9	13
1	-1	5	4	12	16
5.5					

$$\gamma_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

$$\gamma_s = 1 - \frac{6 \times 5.5}{8(64 - 1)} = 0.9345$$

L'Interprétation du coefficient spearman c la même interprétation du coefficient de corrélation pearson

Exemple : les données suivante définie une relation linéaire entre un traitement x et une maladie y

moyen	faible	bien	moyen	moyen	faible	moyen	bien	bien	bien	module (x)
faible	moyen	bien	bien	faible	moyen	bien	bien	bien	faible	module (y)

1- Calculez le coefficient de Spearman.

d_i^2	d_i	Ry	Rx	y	x	
42.25	-6.5	9	2.5	b	b	
0.25	-0.5	3	2.5	b	b	
0.25	3.5	3	2.5	b	b	
12.25	3	3	6.5	b	b	
9	-2.5	-6.5	9.5	b	m	
6.25	3.5	9	6.5	M	m	
12.25	-0.5	3	6.5	m	m	
0.25	3	3	2.5	f	m	
9	-2.5	6.5	9.5	f	f	
6.25		9	6.5	f	f	
98						

$$\begin{aligned}
 r_s &= 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} \\
 &= 1 - \frac{6 \times 98}{10 \times 99} \\
 &= 1 - 0.594 \\
 &= 0.406
 \end{aligned}$$

Exemple) Calculons en le coefficient de Spearman,

sujet	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X	20	24	24	38	10	26	40	16	24	20	16	36	30	28	32
Y	10	14	20	24	16	14	24	10	18	10	6	4	30	32	24

$$\gamma_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

sujet	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X	20	24	24	38	10	26	40	16	24	20	16	36	30	28	32
Y	10	14	20	24	16	14	24	10	18	10	6	4	30	32	24
rang X'	4,5	7	7	14	1	9	15	2,5	7	4,5	2,5	13	11	10	12
rang Y'	4	6,5	10	12	8	6,5	12	4	9	4	2	1	14	15	12
(X' - Y') ²	0,25	0,25	9	4	49	6,25	9	2,25	4	0,25	0,25	144	9	25	0

le coefficient de corrélation des rangs de spearman est donc

$$1 - \left(6 \frac{0,25+0,25+9+4+49+\dots+0}{15(15^2-1)} \right) \simeq 0,531$$

Le théorème de Bayes

Le théorème de Bayes est une conséquence immédiate des probabilités conditionnelles et des probabilités totales.

Probabilités conditionnelles

Dans une bibliothèque comportant 100 ouvrages, il y en a 40 qui sont écrits en anglais dont 8 portent sur la biologie. Considérons les événements suivants:

■ Exemple

Dans une bibliothèque comportant 100 ouvrages, il y en a 40 qui sont écrits en anglais dont 8 portent sur la biologie. Considérons les événements suivants:

$$A = \text{"le livre est écrit en anglais"}; \quad P(A) = \frac{40}{100};$$

$$B = \text{"le livre porte sur la biologie"};$$

$$A \cap B = \text{"le livre est écrit en anglais et porte sur la biologie"}; \quad P(A \cap B) = \frac{8}{100}.$$

■ Probabilité conditionnelle

$B | A = \text{"le livre porte sur la biologie sachant qu' il est écrit en anglais"}$

$$P(B | A) = \frac{8}{40}$$

(Il s'agit de la fréquence des livres de biologie parmi les livres en langue anglaise.) On a les relations

$$P(B | A) = \frac{8}{40} = \frac{\left(\frac{8}{100}\right)}{\left(\frac{40}{100}\right)} = \frac{P(A \cap B)}{P(A)}$$

Probabilités totales

Considérons une partition A_1, A_2, \dots, A_n de l'ensemble des événements E , c'est-à-dire $P(E) = 1$ et

$A_1 \cup A_2 \cup \dots \cup A_n = E, \quad A_i \cap A_j = \emptyset$ pour $i \neq j$. Alors

$$P(B) = P(A_1) \cdot P(B | A_1) + P(A_2) \cdot P(B | A_2) + \dots + P(A_n) \cdot P(B | A_n)$$

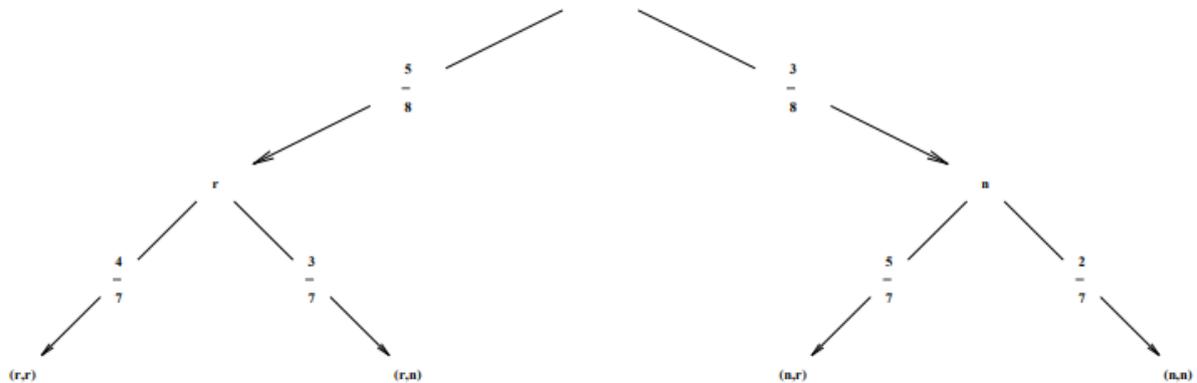
Démonstration

$$P(B) = P(B \cap E) = P(B \cap A_1) + P(B \cap A_2) + \dots + P(B \cap A_n) = \\ P(A_1) \cdot P(B | A_1) + P(A_2) \cdot P(B | A_2) + \dots + P(A_n) \cdot P(B | A_n) \quad \blacksquare$$

■ Problème

Une urne contient 5 boules rouges identiques et 3 boules noires identiques. On effectue des tirages de deux boules sans remise.

- Quelle est la probabilité que la première soit rouge et la deuxième noire ?
- Quelle est la probabilité que l'une des deux boules au moins soit rouge ?
- Sachant que l'une des deux boules au moins est rouge, quelle est la probabilité que l'autre soit noire ?



$$\mathbf{a)} \quad P(r, n) = \frac{5}{8} \cdot \frac{3}{7} = \frac{15}{56}$$

avec $A =$ "la première boule est rouge"; $B =$ "la deuxième est noire"; $A \cap B =$ "la première est rouge et la deuxième est noire"; $B | A =$ "la deuxième est noire sachant que la première est rouge". On retiendra que, dans un arbre, les branches portent des probabilités conditionnelles. La probabilité d'un chemin est égale au produit des probabilités portées par les branches.

$$\begin{aligned}
 \text{b) } P(\text{l' une des deux boules au moins est rouge}) &= \\
 &P(\text{la première est rouge et la deuxième noire}) + \\
 &P(\text{la première est noire et la deuxième rouge}) + \\
 &P(\text{la première est rouge et la deuxième rouge}) = \\
 &P(\text{la première est rouge}) \cdot P(\text{la deuxième est noire} \mid \text{la première est rouge}) + \\
 &P(\text{la première est noire}) \cdot P(\text{la deuxième est rouge} \mid \\
 &\quad \text{la première est noire}) + P(\text{la première est rouge}) \cdot P \\
 &\quad (\text{la deuxième est rouge} \mid \text{la première est rouge}) = \frac{5}{8} \cdot \frac{3}{7} + \frac{3}{8} \cdot \frac{5}{7} + \frac{5}{8} \cdot \frac{4}{7} = \frac{25}{28}
 \end{aligned}$$

$$\text{c) } P(\text{l' autre est noire} \mid \text{l' une des deux boules au moins est rouge}) =$$

$$\frac{P(\text{une boule est rouge et l' autre est noire})}{P(\text{l' une des deux boules au moins est rouge})}$$

$$= \frac{P(r, n) + P(n, r)}{P(r, n) + P(n, r) + P(r, r)} = \frac{\frac{5}{8} \cdot \frac{3}{7} + \frac{3}{8} \cdot \frac{5}{7}}{\frac{5}{8} \cdot \frac{3}{7} + \frac{3}{8} \cdot \frac{5}{7} + \frac{5}{8} \cdot \frac{4}{7}} = \frac{3}{5}$$

→ BAYES

Formule de Bayes

Considérons une partition A_1, A_2, \dots, A_n de l'ensemble des événements E . Alors

$$\begin{aligned}
 P(B) &= P(A_1) \cdot P(B \mid A_1) + P(A_2) \cdot P(B \mid A_2) + \dots + P(A_n) \cdot P(B \mid A_n) \\
 P(A_1 \mid B) &= \frac{P(A_1) \cdot P(B \mid A_1)}{P(B)} \quad \text{probabilité total} \\
 P(A_2 \mid B) &= \frac{P(A_2) \cdot P(B \mid A_2)}{P(B)} \\
 &\dots \\
 P(A_n \mid B) &= \frac{P(A_n) \cdot P(B \mid A_n)}{P(B)} \quad \text{bays}
 \end{aligned}$$

La démonstration a été faite au préalable e sous "Probabilités totales" et "Préparation au théorème de Bayes".

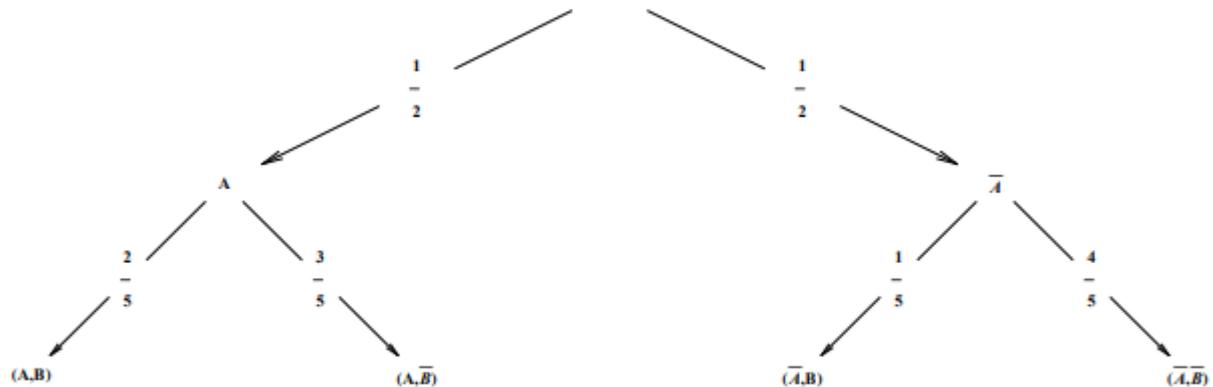
■ Exemple

Dans un laboratoire, on a fait les constats suivants:

si une souris porte l'anticorps A, alors 2 fois sur 5 elle porte aussi l'anticorps B;

si une souris ne porte pas l'anticorps A, alors 4 fois sur 5 elle ne porte pas l'anticorps B.

La moitié de la population porte l'anticorps A.



■ a)

Calculez la probabilité que, si une souris porte l'anticorps B, alors elle porte aussi l'anticorps A.

Pour la partition (A, \bar{A}) , la formule de Bayes donne

$$P(B) = P(A) \cdot P(B | A) + P(\bar{A}) \cdot P(B | \bar{A}) = \frac{1}{2} \cdot \frac{2}{5} + \frac{1}{2} \cdot \frac{1}{5} = \frac{3}{10}$$

$$P(A | B) = \frac{P(A) \cdot P(B | A)}{P(B)} = \frac{\frac{1}{2} \cdot \frac{2}{5}}{\frac{3}{10}} = \frac{2}{3}$$

$$P(\bar{A} | B) = \frac{P(\bar{A}) \cdot P(B | \bar{A})}{P(B)} = \frac{\frac{1}{2} \cdot \frac{1}{5}}{\frac{3}{10}} = \frac{1}{3}$$

Interprétation:

probabilités à priori: $P(A) = \frac{1}{2}$, $P(\bar{A}) = \frac{1}{2}$;

apports d'informations: $P(B | A) = \frac{2}{5}$, $P(B | \bar{A}) = \frac{1}{5}$;

probabilités à postérieures pour les porteuses de l'anticorps B: $P(A | B) = \frac{2}{3}$, $P(\bar{A} | B) = \frac{1}{3}$.

Probabilité des causes :

Les $\frac{2}{3}$ des souris porteuses de l'anticorps B portent aussi l'anticorps A, ce qui dénote une nette incidence de A sur B.

■ b)

Calculez la probabilité que, si une souris ne porte pas l'anticorps B, alors elle ne porte pas l'anticorps A.

Pour la partition (A, \bar{A}) , la formule de Bayes donne

$$P(\bar{B}) = P(A) \cdot P(\bar{B} | A) + P(\bar{A}) \cdot P(\bar{B} | \bar{A}) = \frac{1}{2} \cdot \frac{3}{5} + \frac{1}{2} \cdot \frac{4}{5} = \frac{7}{10}$$

$$P(A | \bar{B}) = \frac{P(A) \cdot P(\bar{B} | A)}{P(\bar{B})} = \frac{\frac{1}{2} \cdot \frac{3}{5}}{\frac{7}{10}} = \frac{3}{7}$$

$$P(\bar{A} | \bar{B}) = \frac{P(\bar{A}) \cdot P(\bar{B} | \bar{A})}{P(\bar{B})} = \frac{\frac{1}{2} \cdot \frac{4}{5}}{\frac{7}{10}} = \frac{4}{7}$$

Interprétation:

probabilités à priori: $P(A) = \frac{1}{2}$, $P(\bar{A}) = \frac{1}{2}$;

apports d'informations: $P(\bar{B} | A) = \frac{3}{5}$, $P(\bar{B} | \bar{A}) = \frac{4}{5}$;

probabilités à postérieures pour les non porteuses de l'anticorps B:

$$P(A | \bar{B}) = \frac{3}{7}, P(\bar{A} | \bar{B}) = \frac{4}{7}.$$

Probabilité des causes :

Les $\frac{4}{7}$ des souris qui ne portent pas l'anticorps B ne portent pas non plus l'anticorps A, ce qui dénote peut-être une légère incidence de \bar{A} sur \bar{B} .

Exemple 5.6 Supposons qu'une population d'adultes soit composée de 30% de fumeurs (A_1) et de 70% de non-fumeurs (A_2). Notons B l'événement "mourir d'un cancer du poumon". Supposons en outre que la probabilité de mourir d'un cancer du poumon est égale à $\Pr(B|A_1) = 20\%$ si l'on est fumeur et de $\Pr(B|A_2) = 1\%$ si l'on est non-fumeur. Le théorème de Bayes permet de calculer

les probabilités a priori, c'est-à-dire la probabilité d'avoir été fumeur si on est mort d'un cancer du poumon. En effet, cette probabilité est notée $\Pr(A_1|B)$ et peut être calculée par

$$\Pr(A_1|B) = \frac{\Pr(A_1)\Pr(B|A_1)}{\Pr(A_1)\Pr(B|A_1) + \Pr(A_2)\Pr(B|A_2)} = \frac{0.3 \times 0.2}{0.3 \times 0.2 + 0.7 \times 0.01} = \frac{0.06}{0.06 + 0.007} \approx 0.896.$$

La probabilité de ne pas avoir été non-fumeur si on est mort d'un cancer du poumon vaut quant à elle :

$$\Pr(A_2|B) = \frac{\Pr(A_2)\Pr(B|A_2)}{\Pr(A_1)\Pr(B|A_1) + \Pr(A_2)\Pr(B|A_2)} = \frac{0.7 \times 0.01}{0.3 \times 0.2 + 0.7 \times 0.01} = \frac{0.007}{0.06 + 0.007} \approx 0.104.$$

Exemple 2 On considère une usine où trois machines fabriquent un même modèle de pièce. 40 % des pièces sont fabriquées par la machine A, qui produit 0,1 % de pièces défectueuses ; 30 % des pièces sont fabriquées par la machine B, plus ancienne, qui produit 0,3 % de pièces défectueuses ; 30 % des pièces sont fabriquées par la machine C, encore plus ancienne, qui produit 0,8 % de pièces défectueuses. On demande la probabilité conditionnelle qu'une pièce ait été fabriquée par la machine C, sachant qu'elle est défectueuse.

Appelons A l'évènement « une pièce prise au hasard a été fabriquée par la machine A », B et C les évènements analogues pour les machines B et C. Appelons D l'évènement « une pièce prise au hasard est défectueuse ». Il faut commencer par traduire les pourcentages en probabilités et en probabilités conditionnelles :

$$P(A) = 0,4, P(B) = 0,3, P(C) = 0,3, \\ P(D|A) = 0,001, P(D|B) = 0,003, P(D|C) = 0,008.$$

On peut alors calculer le dénominateur de la formule de Bayes

$$P(D) = P(A)P(D|A) + P(B)P(D|B) + P(C)P(D|C) \\ = 0,4 \times 0,001 + 0,3 \times 0,003 + 0,3 \times 0,008 = 0,0037.$$

Et on a enfin :

$$P(C|D) = \frac{P(C)P(D|C)}{P(D)} = \frac{0,0024}{0,0037} = 0,65$$

Exercice n° 04 :

Le dépistage de l'hyperthyroïdie s'effectue par un test basé sur le dosage de la THS. Les résultats sont les suivants : chez les malades, 95% des tests sont positifs et chez les non malades, 99 % des tests sont négatifs. Sachant que la prévalence des hyperthyroïdiens dans la population est de 1.5 %. Trouver :

- 1) La probabilité d'être malade sachant que le test est positif.
- 2) La probabilité d'être non malade sachant que le test est négatif.

Solution de l'exo n°4 :

Soient les évènements suivants :

M : « être malade » \bar{M} : « être non malade » T : « test positif » \bar{T} : « test négatif ».

Soient les données suivantes :

$$P(M) = 1.5 \% ; P(T|M) = 95 \% ; P(\bar{T}|\bar{M}) = 99 \%$$

En utilisant le théorème des probabilités totales, des probabilités conditionnelles et la formule de Bayes

$$P(T) = P(M) * P(T|M) + P(\bar{M}) * P(T|\bar{M}) = 0.015 * 0.95 + 0.985 * 0.01 = 0.0241 = 2.41 \%$$

$$P(M|T) = \frac{P(M) * P(T|M)}{P(T)} = \frac{0.015 * 0.95}{0.0241} = 0.5912863071 \cong 59.13 \%$$

$$P(\bar{M}|T) = \frac{P(\bar{M}) * P(T|\bar{M})}{P(T)} = \frac{0.985 * 0.01}{0.0241} = 0.4087136929 \cong 40.87 \% \rightarrow \text{Ces 2 Prob donnent 100\%}.$$

$$P(\bar{T}) = P(M) * P(\bar{T}|M) + P(\bar{M}) * P(\bar{T}|\bar{M}) = 0.015 * 0.05 + 0.985 * 0.99 = 0.9759 \cong 97.59 \%$$

$$P(\bar{M}|\bar{T}) = \frac{P(\bar{M}) * P(\bar{T}|\bar{M})}{P(\bar{T})} = \frac{0.985 * 0.99}{0.9759} = 0.9992315 \cong 99.92315 \%$$

$$P(M|\bar{T}) = \frac{P(M) * P(\bar{T}|M)}{P(\bar{T})} = \frac{0.015 * 0.05}{0.9759} = 0.0007685213649 \cong 0.07685 \% \rightarrow \text{Ces 2 prob} \rightarrow 100 \%$$

Le diagramme arborescent :

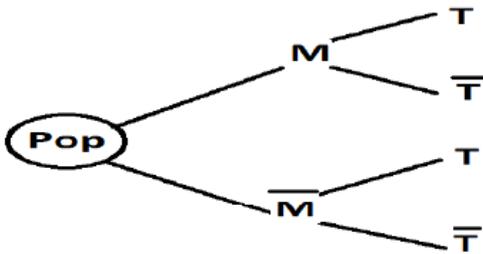


Tableau de Contingence :

Maladie \ Test	T	Non T	Total %
M	1.425	0.075	1.5
Non M	0.985	97.515	98.5
Total %	2.41	97.59	100

Exercice n° 05 :

Dans une population de nouveau-nés, la probabilité de naissance d'un garçon est de 0.52. Dans cette population, 3 % des filles et 2 % de garçons présentent un ictère du nourrisson.

- 1) Calculer la probabilité qu'un nourrisson présente un ictère.
- 2) Quelle est la probabilité qu'un nouveau-né présentant un ictère soit une fille ?

Solution de l'exo n°5 :

Utilisation des probabilités composées, conditionnelles, totales et formule de BAYES.

Les symboles sont : G garçon, F fille, I ictère.

Les données sont : $P(G) = 0.52$; $P(I|F) = 0.03$; $P(I|G) = 0.02$; on tire $P(F) = 0.48$

$$P(I) = P((F \cap I) \cup (G \cap I)) = P(F) * p(I|F) + P(G) * P(I|G) = 0.48 * 0.03 + 0.52 * 0.02 = 0.0248$$

a) Probabilité qu'un nourrisson présente un ictère est 0.0248

$$b) P(F|I) = \frac{P(F \cap I)}{P(I)} = \frac{P(F) * P(I|F)}{P(I)} = \frac{0.48 * 0.03}{0.0248} = 0.580645161 \cong 58 \%$$

$$c) P(G|I) = \frac{P(G \cap I)}{P(I)} = \frac{P(G) * P(I|G)}{P(I)} = \frac{0.52 * 0.02}{0.0248} = 0.419354838 \cong 42 \%$$

Exemple 06 Un laboratoire d'analyses du sang assure avec une fiabilité de 95 % la détection d'une certaine maladie lorsqu'elle est effectivement présente. Cependant, le test indique aussi un résultat faussement « positif » pour 1% des personnes réellement Saines a qui on l'applique (c'est-à-dire qu'une personne saine testée sera déclarée malade une fois sur cent). Si 0,5% de la population porte effectivement la maladie.

quelle est la probabilité qu'une personne soit vraiment malade lorsqu'on la déclare telle sur la base du test?

SOLUTION. Soit D l'événement «la personne soumise au test est porteuse de la maladie» et E l'événement «le résultat du test est positif». La probabilité $P(D|E)$ voulue est donnée par

$$\begin{aligned} P(D|E) &= \frac{P(DE)}{P(E)} \\ &= \frac{P(E|D)P(D)}{P(E|D)P(D) + P(E|D^c)P(D^c)} \\ &= \frac{(.95)(.005)}{(.95)(.005) + (.01)(.995)} \\ &= \frac{95}{294} \approx .323 \end{aligned}$$

