1ère Partie : Probabilités et Statistiques descriptives

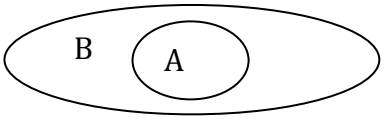
Chapitre 1 : Probabilités

1.1) Probabilités et Ensembles

L'intersection de deux évènements A et B est la partie commune aux deux ensembles (c'est à dire à la fois A et B)

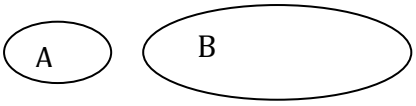
a. Si A est inclus dans B

P(A) : Probabilité de l'événement A P(B) : Probabilité de l'événement B
P(AnB) : Probabilité de l'événement A et B (à la fois A et B), c'est-à-dire A « inter » B.
P(AnB) = P(A) : à la fois A et B



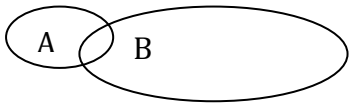
b. Si A et B sont disjoints, c'est à dire incompatibles.

P(AnB) = Ø Ø : représente l'ensemble vide



c. Si A et B ne sont pas disjoints

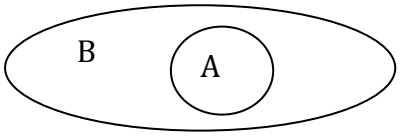
P(AnB) : représente la partie commune aux deux ensembles : A et B



Union de deux évènements A ou B
Il s'agit de la réunion des deux ensembles : A ou B (A « union » B)

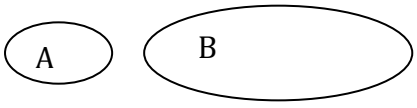
a. Si A est inclus dans B

P(AuB) = B



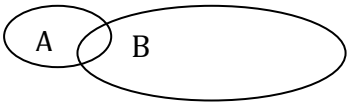
b. Si A et B sont disjoints

P(AuB) = P(A) + P(B) = A ou B



c. Si A et B ne sont pas disjoints

P(AuB) = P(A) + P(B) - P(AnB) = A ou B
On retranche l'intersection P(AnB) afin qu'elle ne soit pas comptée deux fois



1.2) Evènements et probabilités

Une probabilité représente le nombre de cas favorables divisés par le nombre de cas possibles :

Probabilité = Nombre de Cas Favorables
Nombre de Cas Possibles

Soit C un évènement, on note $\overline{C}$ son complémentaire. On a alors : $P(\overline{C}) = 1 - P(C)$

Soient A et B deux évènements, on a : P(AuB) = P(A) + P(B) – P(AnB)

Si A et B sont incompatibles, alors : AnB = Ø : ensemble vide et P(AnB) = 0

P(AuBuC) = P(A) + P(B) + P(C) - P(AnB) - P(AnC) - P(BnC) + P(AnBnC)
Encore une fois, on retranche les intersections entre les deux ensembles (parties communes) car elles ont été comptées plusieurs fois, puis on ajoute l'intersection entre les trois ensembles P(AnBnC) puisqu'elle a été retranchée une fois de trop.

Exemple :
Dans un jeu de 52 cartes, quelle est la probabilité de trouver un roi ou une carte de cœur en retournant une carte au hasard ?

Nombre de cas possibles : 52 cartes
Nombre de cas favorables : Roi ou Carte de cœur = Roi u (Carte de cœur) : Il s'agit d'une « union »
4 Rois dans le jeu et 13 cartes de cœur, une des cartes est un roi de cœur (comptée à la fois dans les Rois et dans les cartes de Cœur).

P(Roi) = $\frac{4}{52}$ P(Cœur) = $\frac{13}{52}$ P(Roi n Cœur) = $\frac{1}{52}$.

P(Roi u Cœur) = P(Roi) + P(Cœur) - P(Roi n Cœur) = $\frac{4}{52} + \frac{13}{52} - \frac{1}{52}$.

Les Combinaisons
On génère des combinaisons pour les cas sans remise (tirage exhaustif) et lorsque l'ordre n'est pas important :
C^2_3 : Nombre de façons d'obtenir différents groupes de deux éléments parmi trois éléments.

Exemple :
Supposons que nous disposons de trois boules blanches, combien de paires de boules blanches pourrions-nous créer à partir de ce groupe de trois boules ?

$C^2_3 = \frac{3!}{(3-2)!\,2!} = 3$

Notation :
$C^k_n = \frac{n!}{(n-k)!\,k!}$ avec ! : le factoriel d'un nombre, c'est-à-dire n ! = n*(n-1)*(n-2)*(n-3)*(n-4)* …* 1

Note : 0! = 1 : le factoriel de 0 est égal à 1.
Et $C^k_n = C^{n-k}_n$

Exemple :
Supposons que nous disposons de trois boules blanches et de quatre boules noires, combien de paires de boules de la même couleur pourrions-nous créer à partir de ces deux groupes de boules ?

Nombre de cas favorables :
Deux boules blanches parmi les (trois) boules blanches :
$C^2_3 = \frac{3!}{(3-2)!\,2!} = 3$

Ou encore deux boules noires parmi les (quatre) boules noires :
$C^2_4 = \frac{4!}{(4-2)!\,2!} = 6$

Nombre de cas possibles :
Deux boules quelconques sélectionnées parmi l'ensemble des sept boules
$C^2_7 = \frac{7!}{(7-2)!\,2!} = 21$

Probabilité de paires de boules de la même couleur créées à partir de ces deux groupes de boules (c'est-à-dire à partir des trois boules blanches et des quatre boules noires) : Deux boules blanches parmi les trois boules blanches ainsi que (Ou) deux boules noires parmi les quatre boules noires.

$$\text{Probabilité} = \frac{\text{Nombre de Cas Favorables}}{\text{Nombre de Cas Possibles}} = \frac{C^2_3 + C^2_4}{C^2_7} = \frac{3+6}{21} = \frac{9}{21}$$

Autre exemple :
Combien de paires de boules de couleurs différentes pourrions-nous créer à partir de ces deux groupes de boules ?

Nombre de cas favorables :
1 boule blanche parmi les (trois) boules blanches :
$C^1_3 = \frac{3!}{(3-1)!\,1!} = 3$

associée à une boule noire parmi les (quatre) boules noires :
$C^1_4 = \frac{4!}{(4-1)!\,1!} = 4$

On associe une boule blanche à une boule noire pour obtenir deux boules de couleurs différentes (paires de boules de couleurs différentes).

Nombre de cas possibles :
Deux boules quelconques sélectionnées parmi l'ensemble des sept boules
$C^2_7 = \frac{7!}{(7-2)!\,2!} = 21$

Nombre de cas favorables : nombre de paires de boules de couleurs différentes créées à partir de ces deux groupes de boules (à partir des trois boules blanches et des quatre boules noires). Une boule blanche doit être associée, à chaque fois, à une boule noire:

$$\text{Probabilité} = \frac{\text{Nombre de Cas Favorables}}{\text{Nombre de Cas Possibles}} = \frac{C^1_3 \times C^1_4}{C^2_7} = \frac{3 \times 4}{21} = \frac{12}{21}$$

1.3) Probabilités conditionnelles et formule des probabilités totales

1. Probabilités conditionnelles

P(A sachant B) = $\frac{P(AnB)}{P(B)}$
Cette probabilité se note P(A/B) ou $P_B(A)$
Et $P(\overline{A}/B) = 1 - P(A/B)$: $P(\overline{A}/B)$ est le complémentaire de P(A/B)

Exemple :
Dans une classe de 25 élèves de 6ème dans un collège, 80% des élèves ont préparé leurs devoirs pour le lendemain. 85% de ceux qui ont préparé et rendu leur devoir auront une bonne note.
85% représente ici une probabilité conditionnelle (probabilité d'obtenir une bonne note sachant que l'élève a rendu son devoir : P(Bonne Note/Devoir Rendu)).
20 élèves sur 25 (80%) ont préparé leur devoir. 17 élèves sur les 20 qui ont rendu leur devoir obtiendront une bonne note ($\frac{17}{20}$ = 85% est donc une probabilité conditionnelle : sachant qu'ils ont rendu leur devoir).

- Si on reprend les données de l'exemple précédent. La probabilité d'avoir une bonne note et d'avoir rendu son devoir : P(AnB) = P(A/B)*P(B) = P(B/A)* P(A)

P(Bonne Note n Devoir Rendu) = P(Bonne Note/Devoir Rendu) * P(Devoir Rendu) = 0,85 * 0,80 = 0,68
68 % des 25 élèves, c'est-à-dire 17 élèves, ont rendu leur devoir et ont obtenu une bonne note.

- Toujours en reprenant les données de l'exemple précédent, la probabilité d'une bonne note sachant que le devoir a été rendu peut aussi être re-calculée : $P(A/B) = P_B(A) = P(AnB)/P(B)$

P(Bonne Note/Devoir Rendu) = $\frac{P(\text{Bonne Note n Devoir Rendu})}{P(\text{Devoir Rendu})}$ = 0,85 = $\frac{0,68}{0,8}$

Indépendance :
Si A et B sont indépendants, cela signifie que les deux événements n'ont aucune influence l'un sur l'autre : P(A/B) = P(A) et P(B/A) = P(B)

Si A et B sont indépendants alors : P(AnB) = P(A/B)*P(B) = P(B/A)* P(A) = P(A)*P(B)
A contrario si P(AnB) ≠ P(A)*P(B), cela signifie forcément que A et B ne sont pas des événements indépendants.

Exemple :
Supposons que la probabilité de voter pour le parti socialiste aux prochaines élections soit de 25% et supposons aussi que la probabilité de voter socialiste dans l'électorat féminin (sachant qu'il s'agit d'une femme) est aussi de 25% :

Cela signifie que P(Socialiste/Femme) = 0,25 et que P(Socialiste) = 0,25 donc :
P(Socialiste/Femme) = P(Socialiste) = 0,25 Il y a bien indépendance entre l'événement voter Socialiste et l'événement être une femme, et le fait d'être une femme ou un homme n'a aucune influence sur le vote pour le parti socialiste : P(Socialiste/Femme) = P(Socialiste/Homme).
Dans ce cas, P(Femme) * P(Vote Socialiste) = P(Femme n Vote Socialiste) = 0,5*0,25=0,125
12,5% de l'électorat total (hommes et femmes confondus) est constitué par des électrices socialistes : P(Femme n Vote Socialiste)

La probabilité d'un Vote Socialiste 'inter' Femme est égale à la probabilité d'être une femme multipliée par la probabilité d'un vote socialiste, puisque les deux événements sont indépendants l'un de l'autre (pas d'influences croisées entre les deux événements), dans notre exemple.

2. Formule des probabilités totales

On peut partitionner un ensemble en plusieurs sous-ensembles. Pour deux événements A et B on a :
$P(A) = P(AnB) + P(An\overline{B})$ où $\overline{B}$ est le complémentaire de B
A est soit associé à B (AnB) soit séparé de B ($An\overline{B}$). L'ensemble A peut être divisé en deux sous-ensembles complémentaires : P(AnB) et $P(An\overline{B})$

Donc $P(A) = P(A/B)*P(B) + P(A/\overline{B})*P(\overline{B})$ en utilisant les probabilités conditionnelles
Puisque $P(AnB) = P(A/B)*P(B)$ et $P(An\overline{B}) = P(A/\overline{B})*P(\overline{B})$

Exemple :
Dans une classe de 25 élèves de 6ème dans un collège, 80% des élèves ont effectivement révisé leur contrôle pour le lendemain. 85% de ceux qui ont révisé leur contrôle auront une bonne note. Mais seulement 20% de ceux qui n'ont pas révisé leur contrôle, auront une bonne note.
P(Révision) = 0,8 P(PasRévision) = 1 – P(Révision) = 0,2
P(Bonne Note/Révision) = 0,85 P(Bonne Note/PasRévision) = 0,2

Quelle est la probabilité qu'un élève de cette classe ait une bonne note ?
Nous pouvons diviser les élèves en deux sous-groupes complémentaires l'un par rapport à l'autre, ceux qui ont révisé leur contrôle et ceux qui n'ont pas révisé leur contrôle.

Nous utiliserons la formule des probabilités totales :
P(Bonne Note) = P(Bonne Note n Révision) + P(Bonne Note n PasRévision)
P(Bonne Note) = P(Bonne Note/Révision) * P(Révision) + P(Bonne Note/PasRévision) * P(PasRévision)
P(Bonne Note) = 0,85 * 0,8 + 0,2 * 0,2
P(Bonne Note) = 0,68 + 0,04 = 0,72

Quelle est la probabilité qu'un élève de cette classe ait une bonne note sachant qu'il a révisé ?
Supposons maintenant que nous connaissons la probabilité de révision, la probabilité de Bonne Note ainsi que P(Bonne Note/PasRévision), nous désirons maintenant estimer P(Bonne Note/Révision) que nous supposerons ne pas connaître, à ce stade :
P(Révision) = 0,8 P(PasRévision) = 1–P(Révision) = 0,2 P(Bonne Note)=0,72
P(Bonne Note/Révision) = ? = **X** : X est une inconnue P(Bonne Note/PasRévision) = 0,2

Nous utiliserons encore une fois la formule des probabilités totales :
P(Bonne Note) = P(Bonne Note/Révision) * P(Révision) + P(Bonne Note/PasRévision) * P(PasRévision)
P(Bonne Note) = **X** * 0,8 + 0,2 * 0,2
0,72 = **X** * 0,8 +0,04 d'où 0,72 – 0,04 = 0,8 * **X** d'où **X** = 0,68/0,8 = **0,85**
Donc P(Bonne Note/Révision) = 0,85

Théorème de Bayes :
Le théorème de Bayes est utile pour inverser le sens de la conditionnalité, pour passer de la probabilité de A sachant B par exemple, à la probabilité de B sachant A.

$$P(\mathbf{A/B}) = \frac{[P(\mathbf{B/A}) * P(A)]}{P(B)} = \frac{P(AnB)}{P(B)}$$

$$P(\mathbf{A/B}) = \frac{[P(\mathbf{B/A}) * P(A)]}{P(B/A)*P(A)+P(B/\overline{A}*P(\overline{A})} = \frac{P(AnB)}{P(B)}$$

Car $P(B) = P(B/A)*P(A)+P(B/\overline{A}*P(\overline{A})$: Formule des probabilités totales

Exemple :
Reprenons les données de l'exemple précédent. Quelle est la probabilité qu'un élève qui a eu une bonne note, ait effectivement révisé son contrôle ?
On doit inverser le sens de la conditionnalité car on connait P(Bonne Note/Révision) et on cherche à calculer P(Révision/Bonne Note).

$$P(\text{Révision/Bonne Note}) = \frac{P(\text{Bonne Note/Révision}) * P(\text{Révision})}{P(\text{Bonne Note})} = \frac{0,85 * 0,80}{0,72} = 0,944$$

1.4) Probabilités et diagnostique

Notre objectif est de détecter la présence d'une maladie chez un sujet (donc de diagnostiquer une maladie) tout en évitant de générer de fausses alarmes, qui sont à l'origine d'erreurs potentielles et de traitements inutiles. Un test efficace doit permettre de bien discriminer entre malades et non malades.

Quelques définitions :
La Prévalence est la probabilité d'être atteint d'une maladie M, dans la population. On la note en général p ou encore P(M).

La sensibilité est la probabilité pour un sujet d'avoir un test positif :T+ (pour une maladie) sachant que le sujet est vraiment atteint de la maladie : P(T+/M). Elle est en général notée Se. Elle représente la probabilité conditionnelle de détection d'un test, sachant que le sujet est malade. Plus la sensibilité est élevée plus le test est efficace.

La Spécificité est la probabilité pour un sujet d'avoir un test négatif :T- (pour une maladie) sachant que le sujet n'est pas atteint de la maladie = P(T-/$\overline{M}$). Elle est en général notée Spé. Elle permet d'estimer le taux de fausses alarmes associé au test. Plus la spécificité est élevée plus le taux de fausse alarme du test est faible : Spécificité = 1 – (Probabilité de fausse alarme)
Elle représente la probabilité conditionnelle de ne pas détecter une maladie (à l'aide d'un test diagnostique), sachant que le sujet n'est vraiment pas malade.
Plus la spécificité est élevée plus le test est efficace.

La Valeur prédictive positive est la probabilité conditionnelle pour un sujet d'être réellement atteint d'une maladie sachant que le sujet a eu un test positif pour cette maladie = P(M/T+). Elle est en général notée **VPP**. Il est donc préférable d'avoir une VPP élevée.

La Valeur prédictive négative est la probabilité conditionnelle pour un sujet de ne pas être atteint d'une maladie sachant que le sujet a eu un test négatif pour cette maladie = P($\overline{M}$/T-). Elle est en général notée **VPN**. Il est donc préférable d'avoir une VPN élevée.

Estimation de la sensibilité, de la spécificité, de la VPP et de la VPN :

Se =VP/(FN+VP) Spé=VN/(FP+VN)

Vrais positifs (VP)	Faux positifs (FP)	**Test positif P(T+)**	$VPP \approx VP/(VP + FP)$
Faux négatifs (FN)	Vrais négatifs (VN)	**Test négatif P(T-)**	$VPN \approx VN/(VN + FN)$
Atteints par la maladie P(M)	**Non atteints par la maladie**		

Où :
VP = nombre de vrais positifs = P(T+ n M)
FP = nombre de faux positifs = P(T+ n $\overline{M}$)
FN = nombre de faux négatifs = P(T- n M)
VN = nombre de vrais négatifs = = P(T- n $\overline{M}$)

- Pour calculer la Spécificité ou la Sensibilité :

Se ≈ VP/(VP + FN) = P(T+ n M)/P(M) = P(T+/M)
Spé ≈ VN/(VN + FP) = P(T- n $\overline{M}$)/P($\overline{M}$) = P(T-/$\overline{M}$)

Exemple :
Nous disposons de deux tests T1 et T2 pour diagnostiquer une maladie. Le test T1 a une sensibilité de 90%, le test T2 a une sensibilité de 80%. Nous décidons d'utiliser les deux tests T1 et T2 successivement sur les sujets à analyser. Nous décidons que nous considérerons que le test est globalement positif uniquement lorsque les tests T1 et T2 seront tous les deux positifs. Quelle est la sensibilité du test global (T) sachant que les tests T1 et T2 sont indépendants l'un de l'autres ?

Test T1 : P(T1+/M) = Se (T1) = 0,9
Test T2 : P(T2+/M) = Se (T2) = 0,8
Test global : P(T+/M) = Se (Test global) = P(T1+ n T2+ / M)
Puisque T1 et T2 sont indépendants : P(T+/M) = P(T1+/M) * P(T2+/M) = 0,9*0,8 = 0,72
donc Se (global) = 0,72

Le test T1 a une spécificité de 80%, le test T2 a une spécificité de 70%. Nous décidons d'utiliser les deux tests T1 et T2 successivement sur les sujets à analyser. Nous décidons que nous considérerons que le test est globalement négatif lorsque l'un des tests T1 **ou** T2 sera négatif (il suffit que l'un des deux tests soit négatif). Quelle est la spécificité du test global (T) sachant que T1 et T2 sont indépendants l'un de l'autres ?

Test T1 : P(T1-/$\overline{M}$) = Spé (T1) = 0,8 et P(T1+/$\overline{M}$) = 1 - P(T1-/$\overline{M}$) = 0,2
Test T2 : P(T2-/$\overline{M}$) = Spé (T2) = 0,7 et P(T2+/$\overline{M}$) = 1 - P(T2-/$\overline{M}$) = 0,3

Test global : P(T-/$\overline{M}$) = Spé (Test global) = P(T1- n T2+ /$\overline{M}$) + P(T1+ n T2- /$\overline{M}$) + P(T1- n T2- /$\overline{M}$)
Puisque T1 et T2 sont indépendants P(T1- n T2+ /$\overline{M}$) + P(T1+ n T2- /$\overline{M}$) + P(T1- n T2- /$\overline{M}$) = P(T1-/$\overline{M}$) * P(T2+/$\overline{M}$) + P(T1+/$\overline{M}$) * P(T2- /$\overline{M}$) + P(T1-/$\overline{M}$) * P(T2- /$\overline{M}$) = (1-0,8)*(0,7) + 0,8*(1-0,7) + (1-0,8) * (1-0,7) = 0,2*0,7 + 0,8 * 0,3 + 0,3 * 0,2 = 0,44
Donc Spé (global) = 0,44

- Il existe deux façons différentes de calculer la VPP et la VPN

1ère façon de calculer : Si la Prévalence dans la population est équivalente à la proportion de malades (P(M)) dans l'échantillon (échantillon représentatif):
VPP ≈ VP/(VP + FP) = P(T+ n M)/P(T+) = P(M/T+)
VPN = VN/(VN + FN) = P(T- n $\overline{M}$)/P(T-) = P(M/$\overline{T}$-)

2ème façon de calculer : Si la Prévalence dans la population n'est pas équivalente à la proportion de malades dans l'échantillon (échantillon non représentatif) :

$$VPP = \frac{[Se * P(M)]}{[Se * P(M) + (1 - Spé) * (1 - P(M))]}$$

$$VPN = \frac{[Spé * (1 - P(M))]}{[(1-Se)*P(M) + Spé * (1 - P(M))]}$$

La 2ème façon de calculer est exacte dans tous les cas, alors que la 1ère façon de calculer décrite un peu plus haut, n'est exacte que dans le cas où l'échantillon est représentatif (Prévalence dans la population équivalente à la proportion de malades dans l'échantillon).
Dans de nombreux cas, surtout lors de l'étude de maladies rares, la proportion de malades dans l'échantillon est plus élevée que dans la population générale, pour pouvoir mieux étudier cette maladie. Il devient donc important d'utiliser la 2nde façon d'estimer la VPP et la VPN plutôt que la première.

La prévalence (proportion de malades dans la population) n'a pas d'influence ni sur la sensibilité, ni sur la spécificité, mais peut avoir un effet sur l'estimation de la VPP ou de la VPN. Pour éviter que des erreurs dans l'estimation de la proportion de malades à partir de l'échantillon (de taille souvent limitée), ne viennent biaiser l'estimation de la VPP ou de la VPN, on utilise la 2nde méthode de calcul. La 2nde méthode de calcul est basée sur l'estimation de la prévalence dans la population toute entière (et non plus à partir d'un échantillon de taille souvent réduite), une mauvaise estimation de la prévalence ne risque donc plus d'influencer la VPP ou la VPN.

Exemple :
Sur un échantillon de 200 personnes, nous avons inclu 100 personnes affectées par une maladie rare que nous désirons étudier. Un test a été utilisé pour diagnostiquer la maladie, le test est positif pour 120 sujets parmi les 200 personnes de l'échantillon. 90 sujets qui étaient réellement malades ont été testés positifs. Calculer La sensibilité ainsi que la spécificité du test utilisé :

	Malades	Non Malades	
Test Positif	90	Faux Positifs ?	120
Test Négatif	Faux Négatifs ?	Vrais Négatifs ?	80
	100	100	200

Faux Négatifs = 120 – 90 = 30Faux Positifs = 100 – 90 = 10
Vrais Négatifs = 100 – 30 = 80 – 10 = 70
Se = 90/100 = VP / Malades = 0,9
Spé = 70/100 = VN / Non Malades = 0,7

Calculer la VPP et la VPN du test utilisé :
Nous ne connaissons pas la prévalence et nous ne sommes pas sûrs que la proportion de malades dans l'échantillon soit réellement représentative.