Probabilités et Biostatistique

1 - Probabilités et probabilités conditionnelles
Evaluation d'un test diagnostique

PAES Faculté de Médecine P. et M. Curie
V. Morice

Pourquoi la biostatistique : la variabilité entre individus

Un exemple : comment comparer 2 traitements A et B pour la même affection

- Critère de comparaison
- Choisir les patients A et ceux B
  $\Rightarrow$ **échantillons** A et B
- Pouvoir généraliser à la **population** (*taille* d'échantillon, *comparabilité* , *représentativité*)

Variabilité

- Variabilité = métrologique + biologique
- Variabilité biologique =
  inter-individuelle + intra-individuelle
- Grandeurs mesurées = **aléatoires**
  $\Rightarrow$ les variations ne sont pas maîtrisées

Probabilités et statistique

- **Probabilités** concernent
  - *Populations*, modèles, théorie
  - On ne peut y faire des mesures
- **Statistique** concerne
  - *Échantillons*, monde réel, pratique
  - On fait des mesures sur des individus

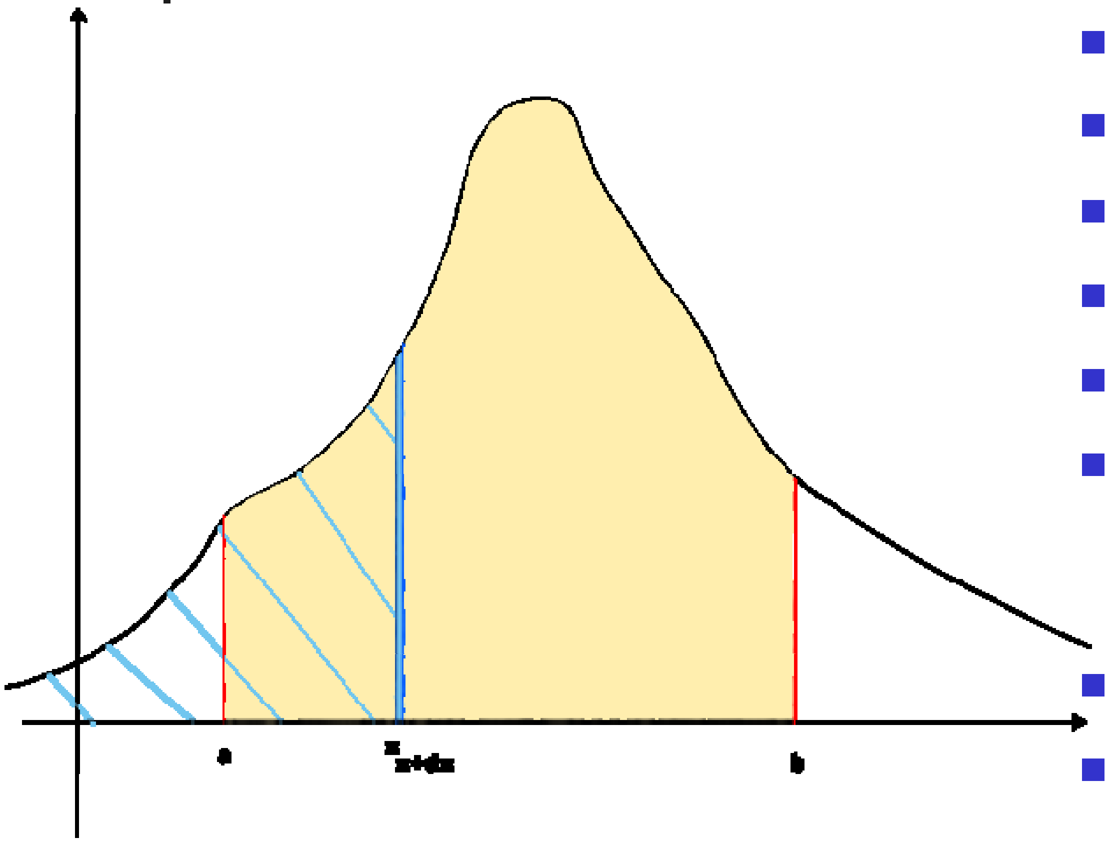
Liens entre probabilités et statistique

- Statistique descriptive
  - Mesures sur un échantillon
  - Résumer/représenter les mesures (moyenne, histogramme)
- Statistique inférentielle, estimation
  - Généraliser les résultats à la population (espérance mathématique, loi de probabilité)
  - $\Rightarrow$ **définir un modèle**
- Tests d'hypothèses
  - Contrôler la validité d'un modèle
  - En utilisant des mesures sur échantillons
  - En prenant en compte les **fluctuations d'échantillonnage**

Rappels sur les ensembles

- Ensemble **fini** (nombre fini d'éléments)
- Ensemble **infini dénombrable** (les éléments peuvent être numérotés ; ex. $\mathbb{N}$)
- Ensemble **infini non dénombrable** (les éléments ne peuvent pas être numérotés ; ex. $\mathbb{R}$)
- $\mathbf{A \cap B}$: intersection $\Leftrightarrow$ **A et B**
  - $A \cap B = \varnothing \Leftrightarrow$ A et B sont **disjoints**
- $\mathbf{A \cup B}$: réunion $\Leftrightarrow$ **A ou B**
- $\mathbf{\complement A}$ ou $\mathbf{\overline{A}}$: complémentaire ou négation $\Leftrightarrow$ **non A**
- $\mathbf{A \times B}$: produit (ex. $A=\{p,f\}$, $B=\{1,2,3,4,5,6\}$ $A \times B=\{(p,1),(f,1),(p,2),\ldots,(f,6)\}$)

Rappels sur les intégrales



- $\int_a^b f(x)dx$ = surface jaune
- $f(x)dx$ = surface bleue
- $\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx$
- $\int_a^b \lambda f(x)dx = \lambda \int_a^b f(x)dx$
- $\int_a^b f(x)dx = \int_a^b f(t)dt$
- $\int_{-\infty}^x f(t)dt = F(x)$
  = **primitive** de $f(x)$
  = surface hachurée
- $\int_a^b f(x)dx = F(b) - F(a)$
- $f(x) = \dfrac{dF(x)}{dx}$

Équations aux dimensions

- Deux grandeurs physiques comparables possèdent la même **dimension**
- 7 dimensions de base :
  - longueur **L**, masse **M**, temps **T**, intensité électrique **I**, température **Θ**, intensité lumineuse **J**, quantité de matière **N**
- La dimension de toute grandeur physique s'exprime en fonction des dimensions de base par une équation aux dimensions :
  - $v = x/t \Rightarrow [v] = LT^{-1}$
  - $w = Fl = m\gamma l \Rightarrow [w] = ML^2T^{-2}$

Systèmes d'unités

- Deux grandeurs de même dimension doivent avoir la même unité pour être comparables
- Unités les plus utilisées définies par le **Système International**
- Unités SI des 7 dimensions de base
  - Mètre (m) pour une longueur L
  - Kilogramme (kg) pour une masse M
  - Seconde (s) pour un temps T
  - Ampère (A) pour une intensité électrique I
  - Degré Kelvin (K) pour une température Θ
  - Candela (cd) pour une intensité lumineuse J
  - Mole (mol) pour une quantité de matière N
- Unités SI dérivées : pascal, joule, newton, volt, …

Erreurs de mesure

- La mesure d'une grandeur X est entachée d'une erreur ΔX
  - Imprécision de l'outil de mesure
  - Biais introduit par l'outil de mesure
  - Variabilité intra-individuelle
- Si G est fonction de variables X, Y, Z mesurables avec des erreurs ΔX, ΔY, ΔZ
  - $G = f(X, Y, Z)$
  - $\Delta G = \left|\frac{\partial f}{\partial X}\right| \Delta X + \left|\frac{\partial f}{\partial Y}\right| \Delta Y + \left|\frac{\partial f}{\partial Z}\right| \Delta Z$
  - Si une seule variable : $G = f(X)$, alors $\Delta G = \left|\frac{df}{dX}\right| \Delta X$

    Ce qui revient à confondre la dérivée et $\frac{\Delta G}{\Delta X}$

Expérience aléatoire, ensemble fondamental

- Expérience aléatoire
  - Ex : lancer de dé, glycémie de 100 personnes
  - Réalisation $\Rightarrow$ mesures $\Rightarrow$ statistique
  - Étude des résultats possibles $\Rightarrow$ probabilités
- Ensemble fondamental
  - $E=\{r_1, r_2, \ldots, r_n\}$: liste des résultats possibles
  - E peut être fini ou infini, dénombrable ou non

Événements

- Sous-ensemble de résultats possibles
  - Si $E=\{1,2,3,4,5,6\}$, l'événement *résultat pair* est $\{2,4,6\}$
  - L'événement se produit si le résultat de l'expérience fait partie du sous-ensemble
- Cas particuliers
  - $\{r_i\}$ = événement **élémentaire**
  - $\varnothing$ = ensemble vide = événement **impossible**
  - E = événement **certain**
  - Événements A et B **incompatibles** ou **exclusifs** $\Leftrightarrow$ sous-ensembles A et B disjoints $\Leftrightarrow A \cap B = \varnothing$

Règles de combinaison d'événements

- Si A et B sont 2 événements, on veut
  - $A \cap B$ est un événement
  - $A \cup B$ est un événement
  - $\overline{A}$ est un événement
- Si E est **fini** ou **infini dénombrable**, tout sous-ensemble de E est un événement
- Si E est **infini non dénombrable** ($\mathbb{R}$), un événement est un **intervalle** ou une combinaison d'intervalles

Probabilité d’un événement

- La théorie des probabilités ne permet pas de calculer toutes les probabilités
- Elle permet le calcul pour les combinaisons d’événements de probabilités connues
- Définition utilisée : **probabilité = limite de fréquence**
- Autres définitions possibles (jeux, probabilités subjectives, …)

Règles (axiomes) du calcul des probabilités

1. $Pr(A) \geq 0$
2. $Pr(E) = 1$
3. Si $A \cap B = \varnothing$, $Pr(A \cup B) = Pr(A) + Pr(B)$

Conséquences

- $Pr(A) \leq 1$
- $Pr(\varnothing) = 0$

  Car $E \cap \varnothing = \varnothing \Rightarrow Pr(E \cup \varnothing) = Pr(E) = Pr(E) + Pr(\varnothing)$

  NB : si $Pr(A) = 0$, A n'est pas nécessairement $\varnothing$
- $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$

Probabilités à définir sur un ensemble fondamental fini

- On doit se donner les probabilités de tout événement élémentaire $\{r_i\}$:
  $Pr(\{r_i\})=Pr(r_i)$, pour tout r_i de E
  - $Pr(r_i) \geq 0$
  - $\sum_{i=1,n} Pr(r_i)=1$ (n = nombre d'événements élémentaires)
- Si $A=\{r_1,r_4,r_5\}$, $Pr(A)= Pr(r_1)+ Pr(r_4)+ Pr(r_5)$
- Ensemble équiprobable : les événements élémentaires ont tous la même probabilité $1/n$
  - Si A possède k éléments, sa probabilité est k/n (nombre de cas favorables sur nombre de cas total)

Définition des probabilités sur un ensemble infini non dénombrable ($\mathbb{R}$)

- Il faut définir les probabilités de tout intervalle
- On utilise une fonction qui dépend des bornes de l'intervalle
  - La **fonction de répartition** permet un calcul par simple soustraction
  - La **densité de probabilité** nécessite un calcul d'intégrale

Probabilités conditionnelles : introduction

- Expérience considérée sur une population $\mathcal{P}$
- Événement A de probabilité $Pr(A)$
- Que devient $Pr(A)$ si on se restreint à une sous-population de $\mathcal{P}$
  - A = taille $\in$ [170 ; 175]
    Sous-population = les hommes
  - A = présence d'une maladie M
    Sous-population = les individus présentant un signe S

Probabilités conditionnelles : notations

- B = événement conditionnant,
  qui définit la sous-population
  - B=être un homme ; B=présenter le signe S
- L'ensemble fondamental doit parler de B
  - Ensemble produit
  - Ex : {(M,S), (M,$\overline{\text{S}}$),($\overline{\text{M}}$,S),($\overline{\text{M}}$, $\overline{\text{S}}$)}
- ***Pr(A/B)*** = Probabilité de A pour les individus présentant B
  = Probabilité de A sachant que B s'est produit
  = Probabilité de A parmi les B
  = Probabilité de A si B
  = Probabilité de A sachant que B
- Confusion fréquente entre $Pr(A/B)$ et $Pr(A \cap B)$

Probabilités conditionnelles : formule de calcul

- $\boldsymbol{Pr(A/B)=\dfrac{Pr(A\cap B)}{Pr(B)}}$
- $Pr(B)$ ne doit pas être nul
- $Pr(A/B)$ est une véritable probabilité
  - $(A \cap B) \subset B \Rightarrow Pr(A \cap B) \leq Pr(B) \Rightarrow Pr(A/B) \leq 1$
  - Si $A_1 \cap A_2 = \varnothing$ (au moins chez les B) $\Rightarrow$ $Pr((A_1 \cup A_2)/B) = Pr(A_1/B) + Pr(A_2/B)$

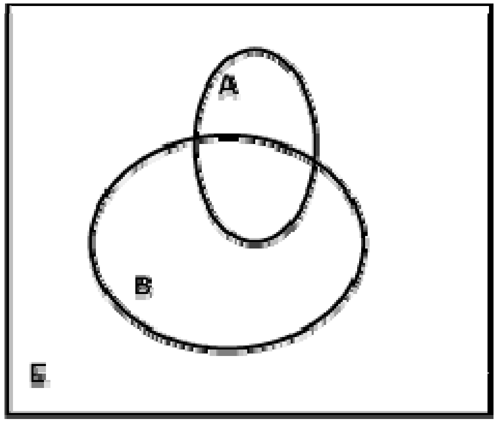
Probabilités conditionnelles : interprétation 1

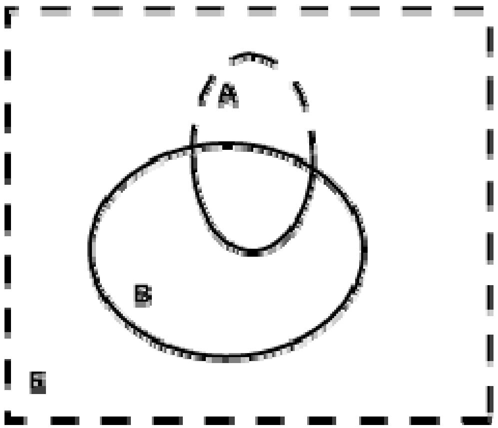
- Interprétation **fréquentielle**
  - Expérience répétée n fois
  - Fréquence de $A = n_A/n$
  - Fréquence de $B = n_B/n$
  - Fréquence de $A \cap B = n_{AB}/n$
  - Fréquence de A parmi les B =
    nb cas favorables sur nb cas total =

$$n_{AB}/n_B = \frac{n_{AB}/n}{n_B/n}$$

Probabilités conditionnelles : interprétation 2

- Interprétation **graphique**





- Représentation avec surfaces proportionnelles aux probabilités
- $Pr(A) = |A|/|E|$
- Les A parmi les B sont représentés par la surface de A incluse dans B
- $Pr(A/B) = |A \cap B|/|B| = \dfrac{|A \cap B|/|E|}{|B|/|E|}$

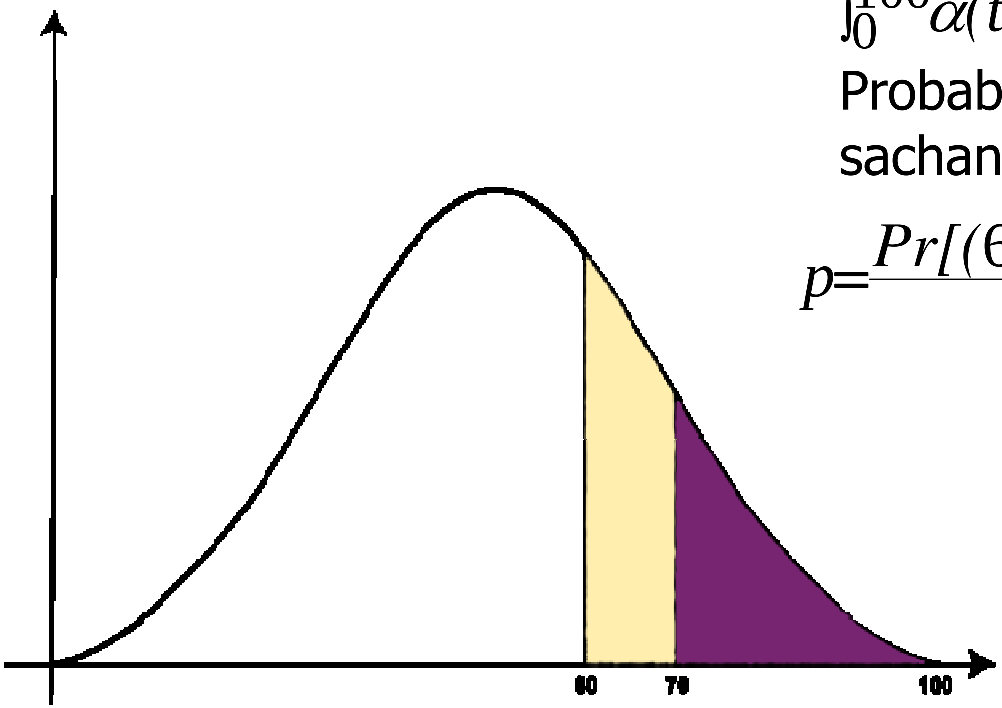
Probabilités conditionnelles : exemple 1 (E fini)

- On lance 2 dés. La somme des 2 résultats est 6. Probabilité qu'un des résultats soit 2 ?
- On a 36 résultats possibles, équiprobables
- A = un résultat est 2 = {(2,1)(2,2)(2,3)(2,4)(2,5)(2,6) (1,2)(3,2)(4,2)(5,2)(6,2)}
- B = somme vaut 6 = {(1,5)(2,4)(3,3)(4,2)(5,1)}
- $A \cap B$ = {(2,4)(4,2)}
- $Pr(A \cap B)$ = 2/36 = 5,6%
- $Pr(A/B)$ = 2/5 = 40% ($= \frac{2/36}{5/36}$)

Probabilités conditionnelles : exemple 2 ($E = \mathbb{R}$)

Durée de vie t : $Pr(t_1 \le t \le t_2) = \int_{t1}^{t2} \alpha(t)dt$

Avec $\alpha(t) = At^2(100-t)^2$ si $0 \le t \le 100$. Sinon $\alpha(t) = 0$



$$\int_0^{100} \alpha(t)dt = 1 \Rightarrow A = 3 \times 10^{-9}$$

Probabilité p de décès entre 60 et 70 sachant qu'on a déjà vécu 60 ?

$$p = \frac{Pr[(60 \le t \le 70) \cap (t \ge 60)]}{Pr(t \ge 60)} = \frac{Pr(60 \le t \le 70)}{Pr(t \ge 60)}$$

$$Pr(60 \le t \le 70) = \int_{60}^{70} \alpha(t)dt$$

$$Pr(t \ge 60) = \int_{60}^{100} \alpha(t)dt$$

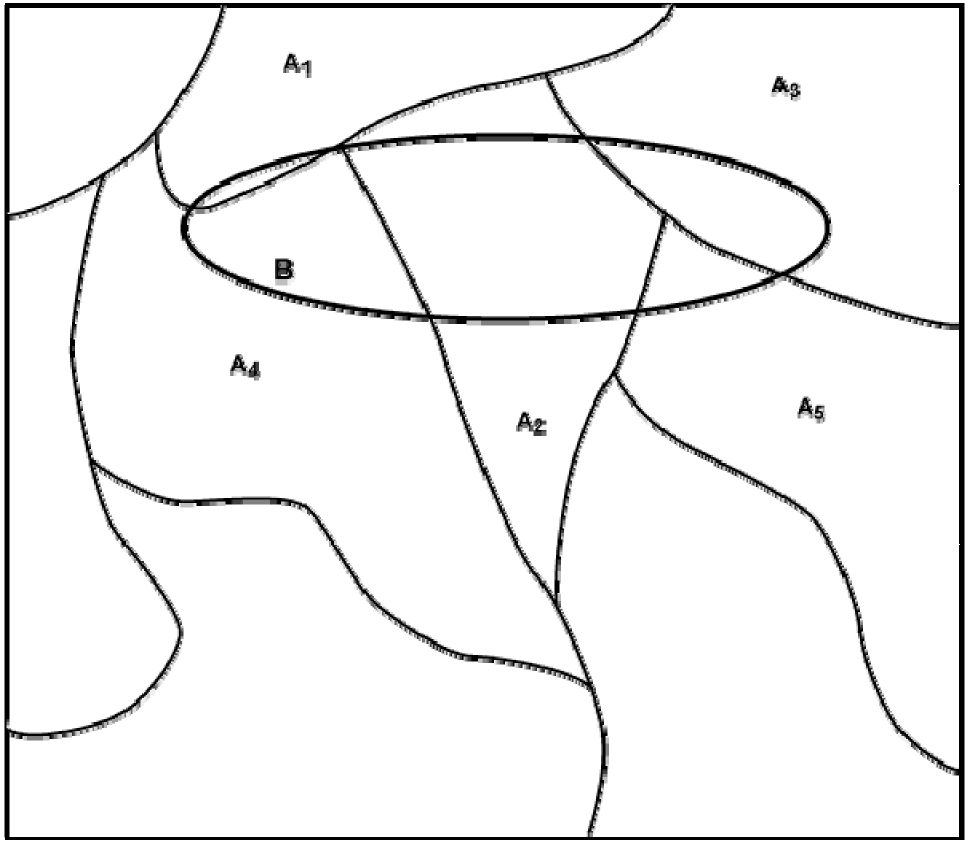
$$p = 0,486$$

Théorème de la multiplication

- Rappel : $Pr(A/B) = \dfrac{Pr(A \cap B)}{Pr(B)}$
- $\mathbf{Pr(A \cap B) = Pr(A/B)Pr(B)}$
- 2ème rappel : $Pr(B/A) = \dfrac{Pr(A \cap B)}{Pr(A)}$
- Autre forme du théorème :
  $Pr(A \cap B) = Pr(B/A)Pr(A)$

Probabilités totales

$Pr(B) = Pr(B/A_1)\,Pr(A_1) + Pr(B/A_2)\,Pr(A_2) + \ldots + Pr(B/A_n)\,Pr(A_n)$



À condition que les A_i forment une partition de E

(les A_i sont exclusifs et $\cup A_i = E$)

$B = B \cap E = B \cap (A_1 \cup A_2 \ldots \cup A_n) =$
$(B \cap A_1) \cup (B \cap A_2) \ldots \cup (B \cap A_n)$

Les $(B \cap A_i)$ sont exclusifs car les A_i le sont. D'où :

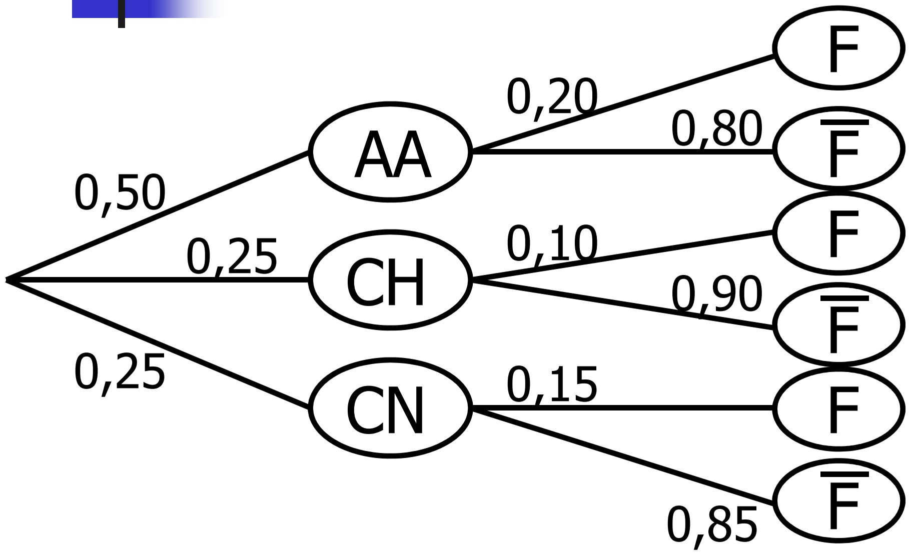
$Pr(B) = Pr(B \cap A_1) + Pr(B \cap A_2) \ldots + Pr(B \cap A_n)$

Théorème de la multiplication :

$Pr(B \cap A_i) = Pr(B/A_i)\,Pr(A_i)$

D'où le résultat annoncé

Probabilités totales : exemple



- Douleur aiguë de l'abdomen
- 3 pathologies
  - AA (50% des cas)
  - CH (25% des cas)
  - CN (25% des cas)
- 20% des AA ont de la fièvre
- 10% des CH et 15% des CN

- Probabilité de fièvre en cas de douleur aiguë de l'abdomen
- $Pr(\text{F}) = Pr(\text{F/AA})Pr(\text{AA}) + Pr(\text{F/CH})Pr(\text{CH}) + Pr(\text{F/CN})Pr(\text{CN})$
  $= 0{,}162$

Formule de Bayes

- Théorème de la multiplication :

$$Pr(A \cap B) = Pr(A/B)Pr(B) = Pr(B/A)Pr(A)$$

- $$Pr(A/B) = \frac{Pr(B/A)Pr(A)}{Pr(B)}$$

- Interprétation :
  - B est une conséquence, A une cause
  - La formule permet de remonter aux causes

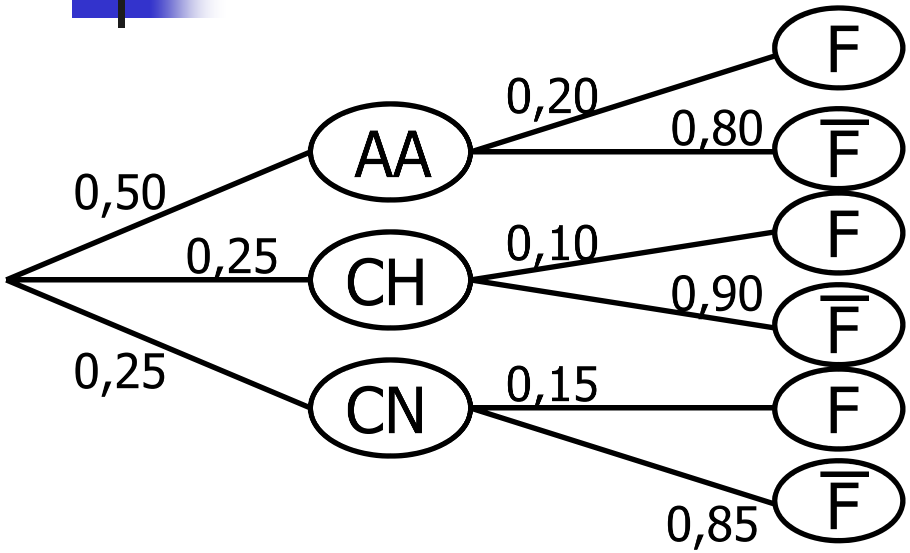
Théorème de Bayes

- Soit n événements A_i formant une partition de E, et un autre événement B (souvent n=2 et $A_2=\overline{A_1}$)
- Pour chaque A_i : $Pr(A_i/B)=\dfrac{Pr(B/A_i)Pr(A_i)}{Pr(B)}$
- Utilisons le théorème des probabilités totales :

$$Pr(A_i/B)=\frac{Pr(B/A_i)Pr(A_i)}{Pr(B/A_1)Pr(A_1)+Pr(B/A_2)Pr(A_2)+\ldots+Pr(B/A_n)Pr(A_n)}$$

- Pour calculer une probabilité conditionnelle, utiliser :
  - La définition
  - Ou Bayes (formule ou théorème)

Bayes : exemple



- Douleur aiguë de l'abdomen
- Un patient présente de la fièvre. Probabilité de chacune des causes
- Quel est le diagnostic le plus probable ?
- Rappel : $Pr(\text{F}) = 0,162$

- $Pr(\text{AA/F}) = \dfrac{Pr(\text{F/AA})Pr(\text{AA})}{Pr(\text{F})} = \dfrac{0,2 \times 0,5}{0,162} = 0,62$

- $Pr(\text{CH/F}) = 0,15$. $Pr(\text{CN/F}) = 0,23$

Indépendance entre deux événements

- A est indépendant de B si la réalisation ou non de B n'influe pas sur celle de A
- $\boldsymbol{Pr(A) = Pr(A/B)}$
- De $Pr(A/B) = \dfrac{Pr(A \cap B)}{Pr(B)}$ on tire

$$\boldsymbol{Pr(A \cap B) = Pr(A)\, Pr(B)}$$

- Autres formes : $Pr(A) = Pr(A/\overline{B})$, $Pr(B) = Pr(B/A)$

Indépendance et incompatibilité

- Ne pas confondre événements indépendants et événements exclusifs (incompatibles)
- Si A et B sont exclusifs
  - La réalisation de B influe sur celle de A : elle l'empêche
  - $Pr(A/B) = Pr(B/A) = 0$
  - $Pr(A \cap B) = 0$ $[\neq Pr(A)Pr(B)]$
- Si A et B sont indépendants
  - La réalisation de B n'influe pas sur celle de A
  - $Pr(A \cap B) = Pr(A)Pr(B)$

Indépendance entre deux événements : exemple

- Trois jets consécutifs d'une pièce
  - A = le premier jet donne face
  - B = le deuxième jet donne face
  - C = deux jets consécutifs donnent face
- Indépendance entre A et B, A et C, B et C ?
- E contient 8 éléments équiprobables {FFF,FFP,FPF,FPP,PFF,PFP,PPF,PPP}
- A={FFF,FFP,FPF,FPP}. Donc $Pr(A) = 1/2$
- B={FFF,FFP,PFF,PFP} $\Rightarrow Pr(B) = 1/2$. C={FFF,FFP,PFF} $\Rightarrow Pr(C) = 3/8$
- $A \cap B$={FFF,FFP} $\Rightarrow Pr(A \cap B) = 1/4 = Pr(A)Pr(B) = 1/2\times 1/2$
- $A \cap C$={FFF,FFP} $\Rightarrow Pr(A \cap C) = 1/4 \neq Pr(A)Pr(C) = 1/2\times 3/8 = 3/16$
- $B \cap C$={FFF,FFP,PFF} $\Rightarrow Pr(B \cap C) = 3/8 \neq Pr(B)Pr(C) = 1/2\times 3/8 = 3/16$

Intérêt diagnostique des informations médicales

On se place dans la situation suivante :

- On considère une maladie qui peut être présente (M) ou absente ($\overline{M}$)
  - On dit aussi que le diagnostic de M est vrai (D) ou faux ($\overline{D}$)
- On considère un examen dont le résultat est un signe qui peut être présent (S) ou absent ($\overline{S}$)
  - Idée : si S on conclut M, si $\overline{S}$ on conclut $\overline{M}$
  - Si le signe est le résultat d'un examen en tout ou rien, la présence du signe désigne la caractéristique pathologique
  - Si l'examen fournit une valeur numérique, on définit un seuil.
    D'un côté du seuil, les valeurs sont dites normales.
    De l'autre, elles sont dites pathologiques, et S est présent

Sensibilité et spécificité

- **Sensibilité** Se d'un test diagnostique
  - Le test est d'autant plus sensible que les **sujets atteints** de M présentent **plus souvent S**
  - $\mathbf{Se} = Pr(\mathbf{S} / \mathbf{M})$
- **Spécificité** Sp d'un test diagnostique
  - Le test est d'autant plus spécifique que les **sujets non atteints** de M présentent **moins souvent S**
  - $\mathbf{Sp} = Pr(\overline{\mathbf{S}} / \overline{\mathbf{M}})$

Dépistage et confirmation de la maladie M

- Un test diagnostique parfait aurait une sensibilité et une spécificité de 1
- Un test avec une bonne sensibilité sera positif chez presque tous les malades
  - $\Rightarrow$ utilisable pour un dépistage
- Un test avec une bonne spécificité sera négatif chez presque tous les non malades
  - $\Rightarrow$ utilisable pour une confirmation

Valeur prédictive positive

- La VPP est la probabilité d'être **atteint** de M si on **présente le signe** S : **VPP = Pr(M / S)**
- Par Bayes : $\text{VPP} = Pr(\text{M/S}) = \dfrac{Pr(\text{S/M})Pr(\text{M})}{Pr(\text{S/M})Pr(\text{M}) + Pr(\text{S}/\overline{\text{M}})Pr(\overline{\text{M}})}$
- Donc $\text{VPP} = Pr(\text{M/S}) = \dfrac{\text{Se} \times Pr(\text{M})}{\text{Se} \times Pr(\text{M}) + (1 - \text{Sp})(1 - Pr(\text{M}))}$
- La VPP dépend de Pr(M), **prévalence** de la maladie

Valeur prédictive négative

- La VPN est la probabilité de ne **pas être atteint** de M si on ne présente **pas le signe** S :
  $\mathbf{VPN} = \boldsymbol{Pr}(\overline{\mathbf{M}}/\overline{\mathbf{S}})$
- Par Bayes : $\mathrm{VPN} = Pr(\overline{\mathrm{M}}/\overline{\mathrm{S}}) = \dfrac{Pr(\overline{\mathrm{S}}/\overline{\mathrm{M}})\,Pr(\overline{\mathrm{M}})}{Pr(\overline{\mathrm{S}}/\mathrm{M})Pr(\mathrm{M}) + Pr(\overline{\mathrm{S}}/\overline{\mathrm{M}})\,Pr(\overline{\mathrm{M}})}$
- Donc $\mathrm{VPN} = Pr(\overline{\mathrm{M}}/\overline{\mathrm{S}}) = \dfrac{\mathrm{Sp} \times (1 - Pr(\mathrm{M}))}{(1 - \mathrm{Se}) \times Pr(\mathrm{M}) + \mathrm{Sp} \times (1 - Pr(\mathrm{M}))}$
- La VPN dépend de la prévalence de la maladie

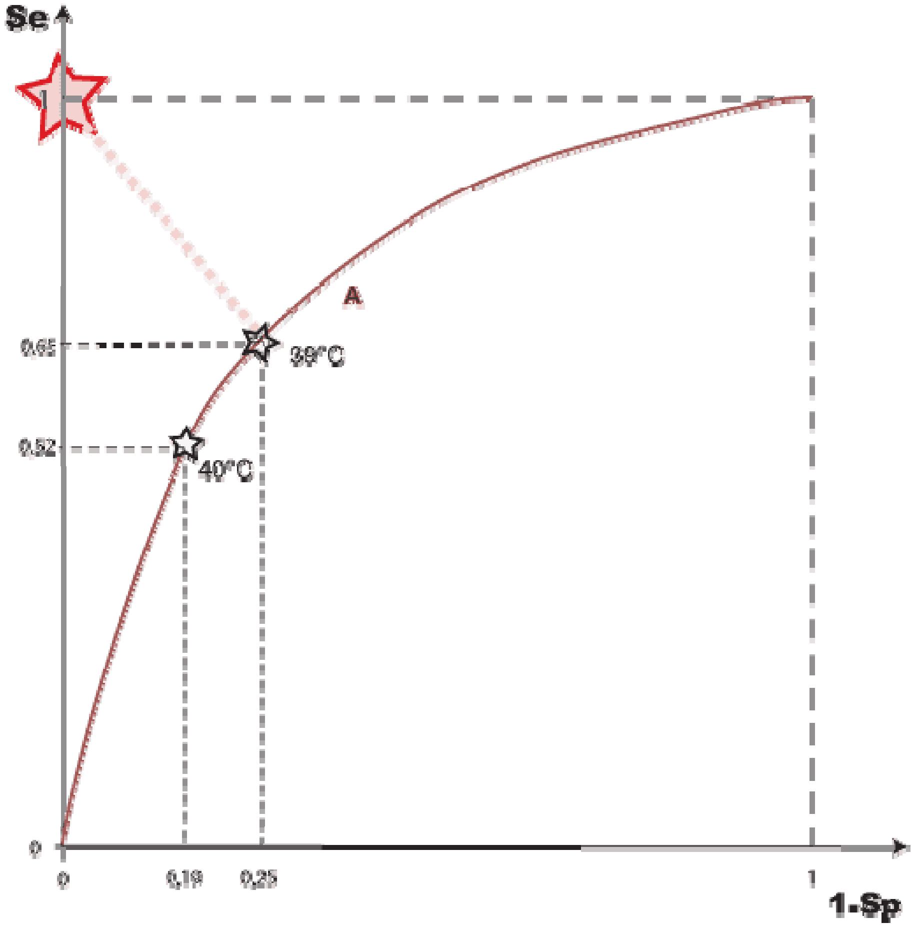
Sensibilité et spécificité vs valeurs prédictives

- Les valeurs prédictives semblent plus naturelles pour juger de l'intérêt d'un examen
  Si on observe S, on connaît la probabilité de M (VPP)
- Mais elles dépendent de la prévalence de M
  Deux centres ont des recrutements différents, et des prévalences différentes. Les valeurs prédictives ne sont pas comparables.
- Ce n'est pas le cas du couple sensibilité/spécificité

Critère continu : influence du seuil

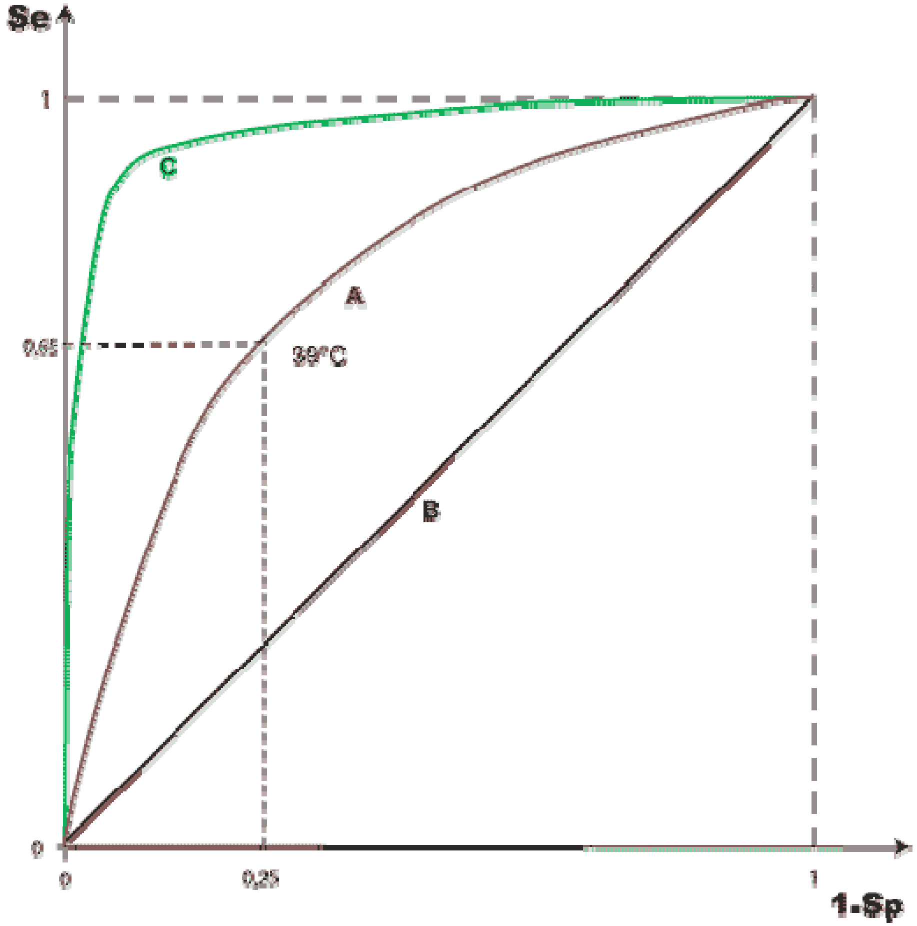
- Maladie = grippe.
  - Signe présent = température $\theta > 39°C$
  - Signe présent = température $\theta > 40°C$
- Se = Pr (S/M) = Pr (θ>seuil/grippe)
  - Décroît lorsque le seuil augmente
- Sp = $Pr(\overline{S}/\overline{M})$ = Pr (θ<seuil/pas grippe)
  - Croît avec le seuil
- **Se et Sp varient en sens inverse**

Critère continu : courbes ROC



- Aide à choisir le seuil
- Courbe ROC : Se = f(1-Sp)
  selon le seuil
- Courbe ROC θ pour grippe
- Meilleur compromis
  Si coûts d'erreur identiques, choisir le point le plus proche du coin supérieur gauche

Critère continu : courbes ROC



- A = courbe θ pour grippe
- Courbe B : examen inutile
  S et M indépendants $Pr(S/M) = Pr(S/\overline{M})$
- Courbe C : bon critère diagnostique

Sensibilité, spécificité, VPP, VPN : estimation (1)

	M	$\overline{M}$
S	VP	FP
$\overline{S}$	FN	VN

- **Un seul échantillon**
- On compte les VP, FN, etc
- Se ≈ VP/(VP+FN)
- Sp ≈ VN/(VN+FP)
- VPP ≈ VP/(VP+FP)
- VPN ≈ VN/(VN+FN)

Sensibilité, spécificité, VPP, VPN : estimation (2)

	M	$\overline{M}$
S	VP	FP
$\overline{S}$	FN	VN

- **Deux échantillons** (malades et non malades)
- La proportion malades/non malades n'est plus respectée
- Se ≈ VP/(VP+FN)
- Sp ≈ VN/(VN+FP)
- VPP et VPN se calculent
  - par Bayes
  - en utilisant la prévalence de M

$$VPP = \frac{Se \times Pr(M)}{Se \times Pr(M) + (1-Sp)(1-Pr(M))}$$

$$VPN = \frac{Sp \times (1-Pr(M))}{(1-Se) \times Pr(M) + Sp \times (1-Pr(M))}$$