

ELEMENTS DE CALCUL DE PROBABILITES

Sensibilité et spécificité d'un test de diagnostic

Dans une phase d'évaluation, un test est appliqué à un groupe d'individus "malades" et à un groupe d'individus "non malades". La présence de la maladie est établie à l'aide d'un *test de référence* (*gold standard*) dont le résultat est considéré comme sûr. Pour chaque individu, on s'intéresse donc aux caractères suivants:

M = avoir la "maladie",

\bar{M} = ne pas avoir la "maladie",

T = avoir un résultat positif au test,

\bar{T} = avoir un résultat négatif au test.

On détermine les *fréquences absolues* (comptages) des quatre résultats possibles:

	M	\bar{M}
T	n_{TM}	$n_{T\bar{M}}$
\bar{T}	$n_{\bar{T}M}$	$n_{\bar{T}\bar{M}}$
Définitions: Total	n_M	$n_{\bar{M}}$

Sensibilité = $\frac{n_{TM}}{n_M}$ = proportion de "+" parmi les malades,

Spécificité = $\frac{n_{\bar{T}\bar{M}}}{n_{\bar{M}}}$ = proportion de "-" parmi les sains.

Exemple

Le test a été administré à 1000 personnes avec M
et à 1000 personnes sans M :

	M	\bar{M}
T	950	10
\bar{T}	50	990
Total	1000	1000

Donc:

$$\text{Sensibilité} = \frac{950}{1000} = 95\%, \quad \text{Spécificité} = \frac{990}{1000} = 99\%.$$

Remarque

La “précision” des valeurs obtenues (“estimations”) dépend du nombre d’individus testés. Cet aspect n’est pas traité ici.

Problème

Supposons que la sensibilité et la spécificité d'un certain test soient:

$$\text{Sensibilité} = 95\%$$

$$\text{Spécificité} = 99\%.$$

Le médecin applique ce test à un patient et obtient un résultat positif.

Quelle est la probabilité que le patient soit réellement malade ?

Pour résoudre ce problème une information supplémentaire est nécessaire: la *fréquence (relative)* de la maladie M dans la population ou *prévalence*.

Supposons que

$$\text{Prévalence} = \frac{1}{10\,000}.$$

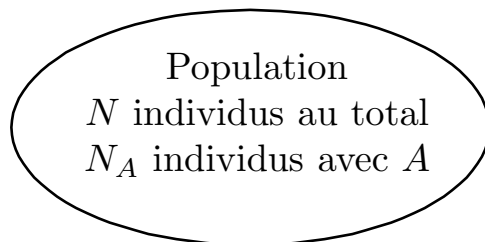
La prévalence de M dans la population est la probabilité *a priori* (avant connaissance du résultat du test) que le patient soit malade.

Il conviendra d'utiliser les concepts fondamentaux et le formalisme du calcul des probabilités.

Concepts de calcul des probabilités

Définitions

Considérons une population de taille N et soit N_A le nombre d'individus avec le caractère A . Supposons le tirage au sort d'un individu.



La *probabilité* de tirer un individu avec A est:

$$P(A) = \frac{N_A}{N}.$$

On dit aussi que $P(A)$ est la probabilité de l'*événement* A . Dans notre définition elle coïncide avec la proportion d'individus avec A .

Supposons maintenant que les individus aient un deuxième caractère B , et indiquons par N_{AB} le nombre d'individus avec les deux caractères A et B simultanément.

La *probabilité conjointe* de A et B est:

$$P(A \cap B) = \frac{N_{AB}}{N}.$$

La *probabilité conditionnelle* de B donné A est:

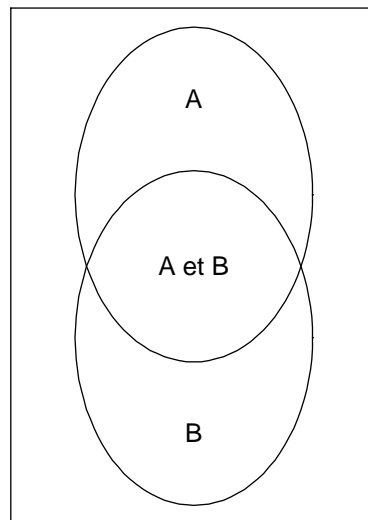
$$P(B|A) = \frac{N_{AB}}{N_A}.$$

Elle est la proportion d'individus avec B (et A) dans la sous-population d'individus avec A .

On dit que A et B sont *indépendants* si:

$$P(B|A) = P(B).$$

Population



Propriétés mathématiques élémentaires des probabilités

- $0 \leq P(A) \leq 1$ pour tout événement A .
- Si \bar{A} signifie “ne pas avoir A ”, alors, pour tout événement A ,

$$P(\bar{A}) = 1 - P(A).$$

- Si A et B sont des *événements incompatibles*, c'est-à-dire, tels que $A \cap B = \emptyset$ (*événement impossible*), alors:

$$P(A \cup B) = P(A) + P(B).$$

- En général

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

- Pour tout A et tout B :

$$P(B|A) = \frac{P(A \cap B)}{P(A)},$$
$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

- Si A et B sont indépendants:

$$P(B|A) = P(B),$$
$$P(A|B) = P(A),$$
$$P(A \cap B) = P(A) \cdot P(B).$$

- *Formule de Bayes:*

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

- *Formule de la probabilité totale:*

$$P(A) = P(A \cap B) + P(A \cap \bar{B})$$
$$= P(A|B)P(B) + P(A|\bar{B})P(\bar{B}).$$

Solution du problème à l'aide de la formule de Bayes

On veut déterminer

$P(M|T)$ = Probabilité que le patient soit malade
“donné le résultat positif du test”
(sachant que le résultat du test est positif).

On sait que:

$$\begin{aligned}P(T|M) &= 95\% = \text{sensibilité,} \\P(\bar{T}|\bar{M}) &= 99\% = \text{spécificité,} \\P(M) &= 1/10\,000 = \text{prévalence,}\end{aligned}$$

et donc

$$P(T|\bar{M}) = 1\% \quad P(\bar{M}) = 9\,999/10\,000.$$

Selon les formules de Bayes et de la probabilité totale:

$$P(M|T) = \frac{P(T|M)P(M)}{P(T)} = \frac{P(T|M)P(M)}{P(T|M)P(M) + P(T|\bar{M})P(\bar{M})}.$$

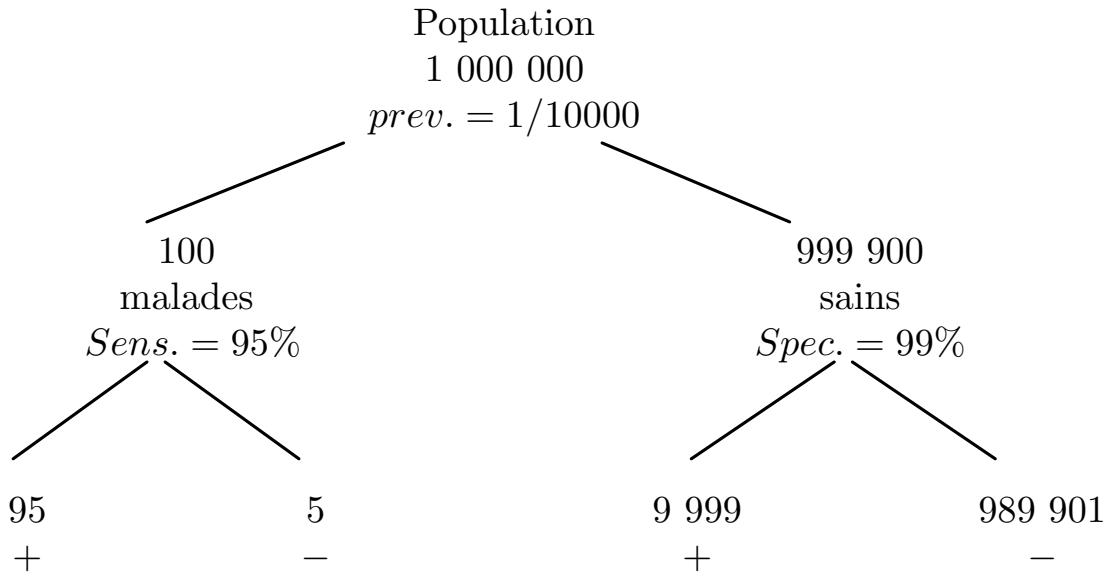
On obtient:

$$P(M|T) = \frac{0.95 \times 0.0001}{0.95 \times 0.0001 + 0.01 \times 0.9999} = 0.0094.$$

$P(M|T)$ est la probabilité *a posteriori* (après connaissance du résultat du test) que le patient soit malade.

Solution intuitive

La prévalence de 1/10'000 nous permet d'affirmer que dans une population hypothétique de 1'000'000 d'individus, on peut s'attendre à 100 malades et 999'900 sains. Le test dépiste 95 cas positifs et 5 cas négatifs parmi les malades, car sa sensibilité est de 95%. Le test trouve aussi 9'999 résultats positifs et 989'901 résultats négatifs dans la partie saine de la population.



Cas positifs et négatifs attendus dans une population de 1 000 000

Ce schéma peut aussi être représenté dans un tableau de fréquences attendues:

Fréquences attendues

	malades	sains	Total
positifs	95	9 999	10 094
négatifs	5	989 901	989 906
Total	100	999 900	1 000 000

En conclusion, la proportion de malades parmi les cas positifs est de 95/10'094, ce qui indique que les chances qu'un individu positif au test soit réellement malade sont seulement de 0.0094 ($\approx 1\%$). Assurez-vous qu'on trouve les mêmes proportions si la taille de la population est changée, par exemple 4'000'000.

Terminologie

$$P(T|M) = \textit{sensibilité du test},$$

$$P(\bar{T}|\bar{M}) = \textit{spécificité du test},$$

$$P(M|T) = \textit{valeur prédictive positive du test},$$

$$P(\bar{M}|\bar{T}) = \textit{valeur prédictive négative du test},$$

$$P(T|\bar{M}) = \textit{taux de faux positifs} = 1 - \textit{spécificité},$$

$$P(\bar{T}|M) = \textit{taux de faux négatifs} = 1 - \textit{sensibilité}.$$

Attention: pour certains auteurs:

$$\textit{taux de faux positifs} = P(\bar{M}|T),$$

$$\textit{taux de faux négatifs} = P(M|\bar{T}).$$

Vérifiez donc toujours la définition utilisée !

Le rôle de la prévalence

Il est souvent difficile de connaître $P(M)$ avec précision. Il convient alors d'examiner le test pour différentes valeurs de $P(M)$.

Par exemple, si $P(T|M) = 0.95$ et $P(\bar{T}|\bar{M}) = 0.99$, on obtient:

$P(M)$	$P(\bar{M} T)$	$P(M \bar{T})$
1/1 000 000	0.9999	0.00000
1/100 000	0.9991	0.00000
1/10 000	0.9906	0.00001
1/1000	0.9132	0.00005
1/500	0.8401	0.00010
1/200	0.6769	0.00025
1/100	0.5103	0.00051

Le taux $P(M|\bar{T})$ est faible: dans le pire des cas ($P(M) = 1\%$), 5 malades sur 10 000 échappent au test. Par contre le taux $P(\bar{M}|T)$ est élevé ($> 50\%$): sur 100 individus positifs plus de 50 sont sains. La décision de maintenir un tel test dépendra de l'importance de la maladie, des conséquences du test, des coûts des examens complémentaires et de l'éventuel traitement, des chances de succès du traitement, etc.

Il est parfois possible de réduire les taux d'erreur en combinant deux (ou plusieurs) tests.

Combinaison de deux ou plusieurs tests

Notation:

M = le patient est malade	\bar{M} = le patient n'est pas malade
T_1 = le premier test est positif	\bar{T}_1 = le premier test est négatif
T_2 = le deuxième test est positif	\bar{T}_2 = le deuxième test est négatif.

Condition: Les résultats des deux tests sont “indépendants”
et cela chez les malades et chez les sains.

En pratique: les tests sont menés de façon indépendante.

Informations disponibles:

$P(M)$	Prévalence	par exemple 10%
$P(T_1 M)$	Sensibilité du premier test	par exemple 75%
$P(\bar{T}_1 \bar{M})$	Spécificité du premier test	par exemple 80%
$P(T_2 M)$	Sensibilité du deuxième test	par exemple 90%
$P(\bar{T}_2 \bar{M})$	Spécificité du deuxième test	par exemple 95%

Valeur prédictive positive du premier test

$$\begin{aligned} P(M|T_1) &= \frac{P(T_1|M)P(M)}{P(T_1|M)P(M) + P(T_1|\bar{M})P(\bar{M})} \\ &= \frac{0.750 \times 0.100}{0.750 \times 0.100 + 0.200 \times 0.900} = 0.294. \end{aligned}$$

Valeur prédictive positive de la combinaison

Après le premier test, $P(M|T_1)$ devient une nouvelle probabilité a priori (avant le deuxième test). Donc, après les deux tests,

$$\begin{aligned} P(M|T_2) &= \frac{P(T_2|M)P(M|T_1)}{P(T_2|M)P(M|T_1) + P(T_2|\bar{M})P(\bar{M}|T_1)} \\ &= \frac{0.900 \times 0.294}{0.900 \times 0.294 + 0.05 \times 0.706} = 0.882. \end{aligned}$$

On procède de façon similaire avec un troisième, un quatrième test, etc. jusqu'au moment où il est raisonnable de prendre une *décision médicale*.

Evaluation basée sur un seul échantillon

Dans certaines études d'évaluation, on ne considère pas deux groupes séparés (malades et non malades) de tailles fixées (n_M et $n_{\bar{M}}$): un seul "échantillon" de taille $n_{..}$ est étudié; ses éléments sont classés dans les quatre cases du tableaux:

	M	\bar{M}	Total
T	n_{TM}	$n_{T\bar{M}}$	n_T
\bar{T}	$n_{\bar{T}M}$	$n_{\bar{T}\bar{M}}$	$n_{\bar{T}}$
Total	n_M	$n_{\bar{M}}$	$n_{..}$

On obtient

$$\text{Sensibilité} = P(T|M) = \frac{P(M \cap T)}{P(M)} \approx \frac{n_{TM}}{n_M},$$

$$\text{Spécificité} = P(\bar{T}|\bar{M}) = \frac{P(\bar{M} \cap \bar{T})}{P(\bar{M})} \approx \frac{n_{\bar{T}\bar{M}}}{n_{\bar{M}}},$$

$$\text{Valeur préd. pos.} = P(M|T) = \frac{P(M \cap T)}{P(T)} \approx \frac{n_{TM}}{n_T},$$

$$\text{Valeur préd. neg.} = P(\bar{M}|\bar{T}) = \frac{P(\bar{M} \cap \bar{T})}{P(\bar{T})} \approx \frac{n_{\bar{T}\bar{M}}}{n_{\bar{T}}}.$$

Problèmes de génétique classique

Supposons que les individus d'une population aient les caractères R et r (allèles d'un même locus).

Les individus adultes peuvent être de trois types: RR , Rr , rr .

Nous considérerons certains types de mariage et de fils :

		père		mère	
Mariages:	A_1 :	RR	\sim	RR	
	A_2 :	RR	\sim	Rr	
	A_3 :	Rr	\sim	RR	Fils: B_1 : RR
	A_4 :	Rr	\sim	Rr	B_2 : Rr
	A_5 :	Rr	\sim	rr	B_3 : rr

Hypothèse A. La transmission des caractères suit le modèle de Mendel:

Mariage	filis	Probabilité
A_1	$\left\{ \begin{array}{l} RR \\ RR \end{array} \right. \longrightarrow RR : B_1$	1
A_2	$\left\{ \begin{array}{l} RR \\ Rr \end{array} \right. \begin{array}{l} \longrightarrow RR : B_1 \\ \longrightarrow Rr : B_2 \end{array}$	$\begin{array}{l} 1/2 \\ 1/2 \end{array}$
A_3	$\left\{ \begin{array}{l} Rr \\ RR \end{array} \right. \begin{array}{l} \longrightarrow RR : B_1 \\ \longrightarrow Rr : B_2 \end{array}$	$\begin{array}{l} 1/2 \\ 1/2 \end{array}$
A_4	$\left\{ \begin{array}{l} Rr \\ Rr \end{array} \right. \begin{array}{l} \longrightarrow RR : B_1 \\ \longrightarrow Rr : B_2 \\ \longrightarrow rr : B_3 \end{array}$	$\begin{array}{l} 1/4 \\ 1/2 \\ 1/4 \end{array}$
A_5	$\left\{ \begin{array}{l} Rr \\ rr \end{array} \right. \begin{array}{l} \longrightarrow rr : B_3 \\ \longrightarrow Rr : B_2 \end{array}$	$\begin{array}{l} 1/2 \\ 1/2 \end{array}$

Hypothèse B. Les couples se forment au hasard et donc les génotypes du père et de la mère sont indépendants.

En utilisant les abréviations p =père, m =mère, on a, par exemple,

$$P(A_1) = P(p = RR \cap m = RR) = P(p = RR) \cdot P(m = RR)$$

Problème 1. Pierre est sain, son frère souffre de mucoviscidose, leurs parents sont sains. Quel est le risque pour Pierre d'avoir un enfant atteint alors que son épouse a une soeur atteinte de mucoviscidose et des parents sains.

Solution. Soit r le gène de la mucoviscidose.

Abbréviations:

\mathcal{P} = Pierre,

\mathcal{E} = épouse de Pierre

\mathcal{G} = enfant de \mathcal{P} et de \mathcal{E} .

On sait que:

Le frère de \mathcal{P} est atteint \Rightarrow le mariage des parents de \mathcal{P} est de type A_4 .

\mathcal{P} est sain $\Rightarrow \mathcal{P}$ n'est pas rr .

Donc

$$\begin{aligned} P(\mathcal{P} \text{ est } Rr \mid \mathcal{P} \text{ non } rr) &= P(\mathcal{P} \text{ est } Rr \cap \mathcal{P} \text{ non } rr) / P(\mathcal{P} \text{ non } rr) \\ &= (1/2) / (1/2 + 1/4) = 2/3, \end{aligned}$$

$$P(\mathcal{P} \text{ est } RR \mid \mathcal{P} \text{ non } rr) = 1/3.$$

Analoguement, pour son épouse,

$$P(\mathcal{E} \text{ est } Rr \mid \mathcal{E} \text{ non } rr) = 2/3,$$

$$P(\mathcal{E} \text{ est } RR \mid \mathcal{E} \text{ non } rr) = 1/3.$$

Grâce aux hypothèses A et B on obtient

$$P(\mathcal{G} \text{ est } rr)$$

$$= P(\mathcal{G} \text{ est } rr \mid \text{le mariage de } \mathcal{P} \text{ et } \mathcal{E} \text{ est } A_4) P(\text{le mariage de } \mathcal{P} \text{ et } \mathcal{E} \text{ est } A_4)$$

$$= (1/4) P(p = Rr) P(m = Rr) = (1/4)(2/3)(2/3) = 1/9.$$

Hypothèse C. Les fréquences alléliques dans la première génération se distribuent selon le modèle de Hardy-Weinberg.

Supposons que les fréquences de R et r soient p et $q = 1 - p$ respectivement. Alors la fréquence du génotype homozygote RR dans la première génération est p^2 ; la fréquence du génotype homozygote rr est q^2 et la fréquence du génotype hétérozygote Rr est $2pq$.

Problème 2. Quelle est la probabilité pour un individu sain d'avoir un enfant avec une maladie autosomique récessive dont la fréquence allélique est de 20% sachant qu'un de ses parents est malade.

Solution. Voir TP2