éduSCOL
Informer et accompagner les professionnels de l'éducation

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

MINISTÈRE DE L'ÉDUCATION NATIONALE ET DE LA JEUNESSE

VOIE GÉNÉRALE

2DE | 1RE | TLE

Enseignement scientifique

ENSEIGNEMENT COMMUN

THÈME 3
SOUS-THÈME 3-5 : L'INFÉRENCE BAYÉSIENNE

Mots-clés

Probabilités des causes, diagnostic, faux positifs, vrais négatifs, formule de Bayes, détection de spams.

Références au programme

Savoirs

L'inférence bayésienne est une méthode de calcul des probabilités des causes à partir des probabilités de leurs effets. Elle est utilisée en apprentissage automatique pour modéliser des relations au sein de systèmes complexes, notamment en vue de prononcer un diagnostic (médical, industriel, détection de spam…). Cela permet de détecter une anomalie à partir d'un test imparfait.

Savoir-faire

À partir de données, par exemple issues d'un diagnostic médical fondé sur un test, produire un tableau de contingence afin de calculer des fréquences de faux positifs, faux négatifs, vrais positifs, vrais négatifs. En déduire le nombre de personnes malades suivant leur résultat au test.

Notions mathématiques travaillées

- Proportions, pourcentages,fréquences
- Probabilité a priori, probabilité aposteriori
- Tableau de contingence
- Probabilités conditionnelles, formule de Bayes

Histoire, enjeux, débats

L'inférence bayésienne fait référence au révérend Thomas Bayes, mathématicien et pasteur britannique né à Londres aux environs de l'année 1702 et mort en 1761. Ses découvertes en probabilités ont été résumées dans son *Essay Towards Solving a Problem in the Doctrine of Chances* (Essai sur la manière de résoudre un problème dans la théorie des risques) publié à titre posthume en 1763 par un de ses amis, Richard Price, dans les comptes rendus de l'académie royale de Londres. On lui doit notamment le théorème de Bayes, très utilisé dans tout ce qui relève du classement automatique (diagnostic médical, filtrage de spams).

Inférence bayésienne et diagnostic médical

La probabilité $P(M)$ d'être atteint d'une maladie peut être interprétée le pourcentage de chances d'être malade avant de prendre en compte des observations (par exemple le résultat d'un test de dépistage). On l'appelle la prévalence. C'est une **probabilité a priori**.

La probabilité $P_{T+}(M)$ d'être malade sachant qu'on réagit positivement au test et la probabilité $P_{T-}(\bar{M})$de ne pas être malade sachant qu'on réagit négativement au test de dépistage peuvent être interprétées comme le pourcentage de chances d'être malade après la prise en compte du résultat du test. On les appelle les **probabilités a posteriori**.

Ces probabilités a posteriori sont déduites d'études cliniques menées au préalable sur des personnes dont on sait si elles sont atteintes de la maladie ou non. On note $P_{\mathrm{M}}(\mathrm{T}^{+})$ la probabilité que le test soit positif lorsque la personne est malade, et $P_{\bar{M}}(\mathrm{T}^{+})$celle qu'il soit positif lorsque la personne n'est pas malade. **Ce sont les probabilités de l'effet conditionnellement aux causes**.

La formule de Bayes $P_{T^+}(M) = \frac{P_M(T^+)\times P(M)}{P_M(T^+)\times P(M)+P_{\overline{M}}(T^+)\times P(\bar{M})}$ permet de calculer la probabilité a posteriori de la cause (être malade) à partir des probabilités de l'effet observé (les chances de test positif lorsqu'on est malade et lorsqu'on ne l'est pas), en prenant en compte la probabilité a priori de la cause. En d'autres termes, elle permet de réviser le pourcentage de chances a priori d'être malade en fonction des observations.

La formule de Bayes pourra être démontrée ou admise selon la connaissance par les élèves des probabilités conditionnelles.

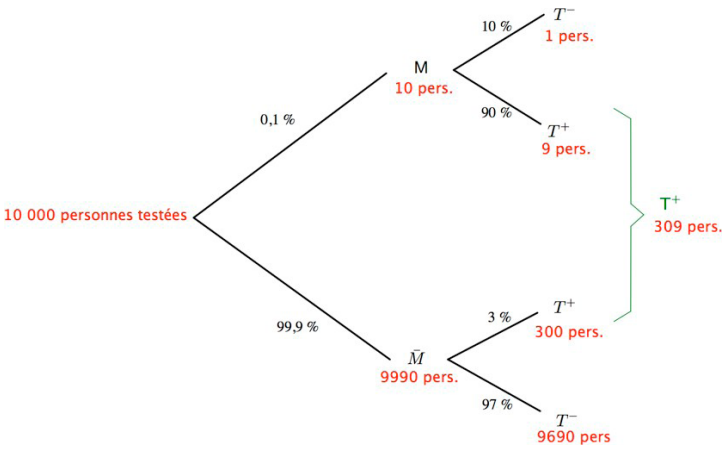
Un exemple introductif (d'après un article de Science étonnante)

Une personne vient de passer un test de dépistage d'une maladie rare. On sait qu'elle ne touche que 0,1 % de la population. Le médecin lui annonce que le résultat du test est positif. La personne demande au médecin si le test est fiable. Sa réponse est sans appel :

« Si vous êtes malade, le test est positif dans 90 % des cas et si vous n'êtes pas malade, il est négatif dans 97 % des cas ».

Problème posé : quelle est la probabilité que cette personne soit effectivement malade ?

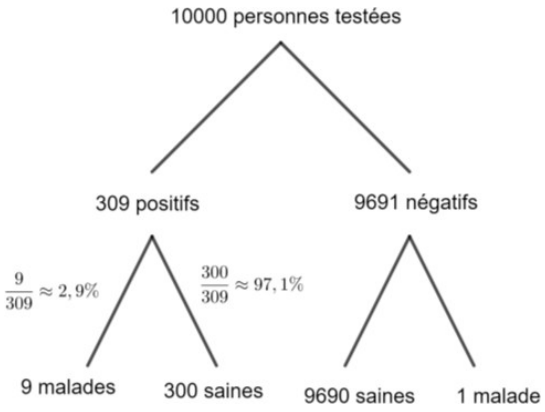
Pour simplifier, on part d'une population de référence de 10 000 personnes ayant passé le test. On peut représenter la situation par un arbre



ou par un tableau (appelé tableau de contingence)

	Test positif	Test négatif	Total
Malades	9	1	10
Non malades	300	9 690	9 990
Total	309	9 691	10 000

Puisque la maladie touche 0,1% de la population, il y a 10 malades parmi ces 10 000 personnes. Comme parmi ces malades, 90 % réagissent positivement au test, il y en a 9 qui réagissent positivement au test. On considère maintenant les personnes saines : ils sont 9990. Puisque dans 97 % des cas le test donne un résultat négatif chez une personne saine, il y a 9690 (valeur entière arrondie) tests négatifs, et donc 300 tests positifs chez ces 9 990 personnes saines. Le bilan de cette analyse est représenté sur le schéma ci-contre.



Sur les 309 personnes qui sont testées positivement, 9 seulement sont réellement malades (ce sont les vrais positifs, notés VP) et 300 sont saines (ce sont les faux négatifs, notés FN).

La personne considérée, dont le test est positif a donc $\frac{9}{309} \approx 2,9\ \%$ de risque d'être réellement malade, et 97,1 % de chance d'être un faux positif, et donc d'être sain.

Pourquoi ce résultat est-il contre-intuitif ?

Face à une telle situation, on est interpelé par les données (90 % des malades réagissent positivement au test, 97 % des non malades réagissent négativement au test).
Une réaction positive au test pourrait lancer penser qu'il y a un risque important d'être effectivement malade, ce qui est démenti par le calcul précédent.
Nous allons démontrer que la probabilité d'être malade pour un individu réagissant positivement au test dépend aussi du caractère plus ou moins rare de la maladie.

Caractéristiques d'un test

Pour pouvoir utiliser un test, on a besoin de déterminer ses caractéristiques.
Cette détermination se fait lors d'une phase de calibrage sur échantillon : le test est appliqué sur un échantillon de n des personnes dont on sait qu'il contient a personnes malades et $b = n - a$ personnes saines. Il faut que l'échantillon de calibrage soit représentatif de la population totale pour que les valeurs caractéristiques du test, calculées à partir de l'échantillon, puissent servir à calculer des probabilités portant sur la population totale. Ainsi, la probabilité $p = P(M)$ pour qu'une personne de la population totale soit malade est estimée par la proportion $\frac{a}{n}$ de personnes malades dans l'échantillon de calibrage.

On observe les réactions au test de ces différentes personnes que l'on classe en fonction de leur état de santé (M = malades, $\overline{M}$ non malades) et de leur résultat au test étudié (T^+s'ils réagissent positivement au test, T^-s'ils réagissent négativement au test).

Les vrais positifs (dont le nombre est noté vp) sont les sujets de l'échantillon qui sont malades et qui réagissent positivement au test ($M \cap T^+$).

Les faux positifs (dont le nombre est noté fp) sont les sujets de l'échantillon qui ne sont pas malades et qui réagissentpositivementautest($\overline{M} \cap T^+$).

Les vrais négatifs (dont le nombre est noté vn) sont les sujets de l'échantillon qui ne sont pas malades et qui ne régissent pas au test ($\overline{M} \cap T^-$).

Les faux négatifs (dont le nombre est noté fn) sont les sujets de l'échantillon qui sont malades et qui ne régissent pas au test ($M \cap T^-$).

Le tableau à deux entrées qui rassemble ces données est appelé tableau de contingence relatif au test de calibrage.

	Test positif ($T+$)	Test négatif ($T-$)	Total
Malades (M)	vp	fn	$vp + fn = a$
Nonmalades $\overline{M}$	fp	vn	$fp + vn = b$
Total	$vp + fp$	$fn + vn$	n

Les résultats ainsi obtenus permettent de déterminer deux caractéristiques du test : sa sensibilité et sa spécificité.

La sensibilité du test, notée S_e, est la probabilité $P_M(T^+)$ qu'une personne malade réagisse positivement au test. Elle est estimée par la proportion $\frac{vp}{vp+fn}$ de vrais positifs parmi les sujets malades de l'échantillon de calibrage.

La spécificité du test, notée S_p, est la probabilité $P_{\overline{M}}(T^-)$ qu'une personne non malade réagissené gativement au test. Elle est estimée par la proportion $\frac{vn}{fp+vn}$ de vrais négatifs parmi les sujets non malades de l'échantillon de calibrage.

La qualité des estimations de S_e et S_p à partir des résultats du test de calibrage dépend de la représentativité de l'échantillon.

Valeurs prédictives d'un test

La **valeur prédictive positive** du test dans une population donnée (**qui n'est plus l'échantillon de calibrage**), notée VPP, est la probabilité qu'un individu de cette population qui réagit positivement au test soit effectivement malade.

Cette probabilité pourrait théoriquement être estimée par la proportion de malades parmi tous les individus de la population qui réagissent positivement au test. Mais, comme on ne peut pas effectuer le test sur la totalité de la population, on n'a pas un accès direct à cette valeur.

La formule de Bayes permet de calculer la valeur prédictive positive à partir des caractéristiques du test et de la prévalence de la maladie.

Démonstration de la formule de Bayes

La formule de Bayes pourra être démontrée à partir de résultats sur les probabilités conditionnelles ou justifiée à partir d'un raisonnement sur les proportions :

Pour des élèves ayant étudié les probabilités conditionnelles en spécialité mathématique de première

La valeur prédictive positive est interprétée comme une probabilité conditionnelle :

$$VPP = P_{T^+}(M) = \frac{P(T^+ \cap M)}{P(T^+)}$$

De le même manière, $P_{\mathrm{M}}(T^+) = \frac{P(T^+ \cap M)}{P(\mathrm{M})}$.

Cela donne pour le calcul du numérateur : $P(T^+ \cap M) = P(M) \times P_{M}(T^+)$

Sous réserve de représentativité du test, on suppose que la probabilité $P_{M}(T^+)$ qu'une personne malade réagisse positivement au test est la même dans la population totale et dans l'échantillon. C'est la sensibilité S_{e} du test.

Ainsi $P(T^+ \cap M) = p \times S_{\mathrm{e}}$.

Pour le dénominateur $P(T^+)$ un calcul similaire permet d'abord de montrer que $P(T^+ \cap \overline{M}) = P(\overline{M}) \times P_{\overline{M}}(T^+) = (1-p) \times (1-S_p)$,

puisque $P(\overline{M}) = 1-P(M) = 1-p$ et $P_{\overline{M}}(T^+) = 1-P_{\overline{M}}(T^-) = 1-S_p$.

Les événements $(T^+ \cap M)$ et $(T^+ \cap \overline{M})$ étant incompatibles et de réunion T^+,

$P(T^+) = P(T^+ \cap M) + P(T^+ \cap \overline{M}) = p \times S_e + (1-p) \times (1-S_p)$.

D'où $VPP = P_{\mathrm{T}}(M) = \frac{p \times S_e}{p \times S_e + (1-p) \times (1-S_p)}$

De manière analogue, on définit la **valeur prédictive négative** du test dans une population donnée, notée VPN, comme la probabilité qu'un individu de cette population réagissant négativement au test soit sain.

Des calculs similaires permettent de montrer que : $VPP = P_{T-}(\overline{M}) = \frac{(1-p) \times S_p}{p \times (1-S_e) + (1-p) \times S_p}$

Pour des élèves n'ayant pas une connaissance préalable des probabilités conditionnelles
L'égalité

$$P(T^+ \cap M) = P(M) \times P_M(T^+)$$

peut être interprétée comme une proportion de proportion : la proportion, dans la population totale, de personnes à la fois malades et réagissant positivement au test est égale au produit de la proportion de personnes malades dans la population totale (la prévalence) par la proportion de personnes réagissant positivement au test parmi la sous-population des personnes malades.

On justifie de même l'égalité $P(T^+ \cap M) = P(T^+) \times P_{T^+}(\mathrm{M})$, qui permet d'écrire

$$\mathrm{VPP} = \frac{P(T^+ \cap M)}{P(T^+)} = \frac{P(M) \times P_M(T^+)}{P(T^+)}$$

Il reste à calculer $P(T^+)$, approchée par la proportion de personnes réagissant positivement au test dans la population totale.

Parmi les personnes réagissant positivement au test, il y a des personnes malades et des personnes saines. La proportion de personnes réagissant positivement au test est la somme de la proportion de personnes réagissant positivement au test en étant malades et de la proportion de personnes réagissant positivement au test en étant non malades.

La proportion de personnes à la fois malades et réagissant positivement au test vient d'être exprimée comme une proportion de proportion et on fait de même pour calculer la proportion de personnes à la fois non malades et réagissant positivement au test. On raisonne de même pour la VPN.

On peut donc ainsi justifier les égalités :

$$VPP = P_{T^+}(M) = \frac{p \times S_e}{p \times S_e + (1-p) \times (1-S_p)} \qquad VPN = P_{T^-}(\overline{M}) = \frac{(1-p) \times S_p}{p \times (1-S_e) + (1-p) \times S_p}$$

L'appliquette GeoGebra « Probabilités conditionnelles : une visualisation» permet de représenter des probabilités conditionnelles à l'aide d'un arbre et de les interpréter en termes d'aires.

L'appliquette GeoGebra « inférence bayésienne et dépistage avec un arbre» permet d'illustrer l'influence de la sensibilité, de la spécificité et de la prévalence sur les VPP et VPN.

L'ensemble des probabilités mentionnées ci-dessus peuvent être récapitulées en réécrivant le tableau précédent non plus en termes de nombre de personnes au sein d'un échantillon de n individus, mais de proportions au sein de la population totale.

	Test positif (T^+)	Test négatif (T^-)	Total
Malades (M)	$P(T^+ \cap M) = P(M) \times P_M(T^+) = p \times S_e$	$P(T^- \cap M) = P(M) \times P_M(T^-) = p \times (1-S_e)$	$P(M) = p$
Non malades ($\overline{M}$)	$P(T^+ \cap \overline{M}) = P(\overline{M}) \times P_M(T^+) = (1-p) \times (1-S_p)$	$P(T^- \cap \overline{M}) = P(\overline{M}) \times P_M(T^-) = (1-p) \times S_p$	$P(\overline{M}) = 1-p$
Total	$P(T^+) = p \times S_e + (1-p) \times (1-S_p)$	$P(T^-) = p \times (1-S_e) + (1-p) \times S_p$	1

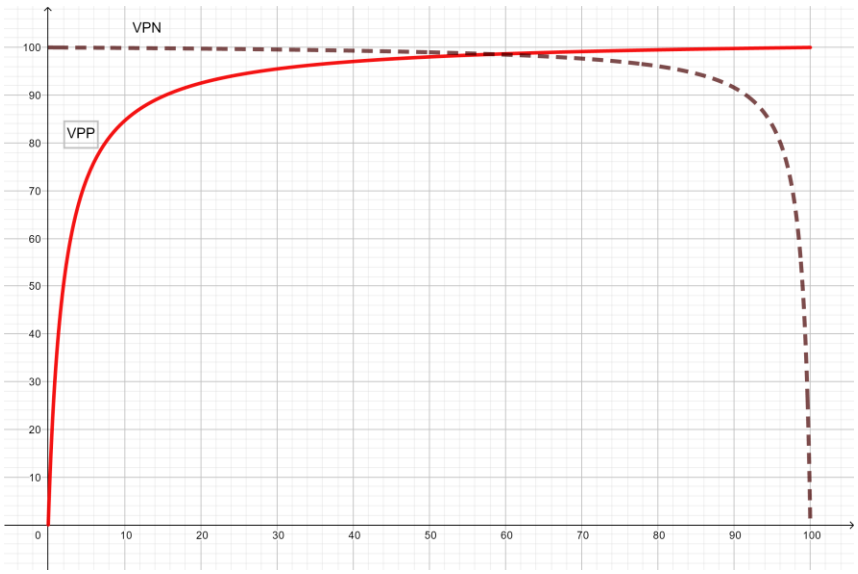
Évolution des valeurs prédictives positives et négatives en fonction de la prévalence

Le tableau suivant donne, pour un test de caractéristiques $S_e = 0{,}99$ et $S_p = 0{,}98$ quelques valeurs de VPP et de VPN en fonction de la prévalence exprimée en pourcentage.

Prévalence	VPP	VPN
0,1 %	< 5 %	99,99 %
1 %	33 %	99,98 %
5 %	72 %	99,95 %
10 %	84 %	99,89 %
30 %	95 %	99,56 %

Comme on le voit sur le tableau, quand la prévalence augmente, la VPP augmente et la VPN diminue. Si la prévalence de la maladie M est de 0,1 % dans la population générale, de 1 % dans la clientèle d'un généraliste et 5 % dans celle d'un spécialiste de la maladie, le tableau indique que l'utilisation systématique du test a une bonne valeur diagnostique pour le spécialiste (VPP = 72 %), que cette valeur est déjà moindre pour le généraliste (VPP = 33 %) et qu'elle est trop basse pour être utilisable pour un dépistage de masse (VPP < 5%).

Les courbes ci-dessous, pour les valeurs $S_e = 0{,}99$ et $S_p = 0{,}95$, représentent l'évolution de la VPN et la VPP, exprimées en pourcentage, en fonction de la prévalence, elle aussi exprimée en pourcentage.



On constate qu'en cas de maladie rare (faibles valeurs de p), la VPN est élevée alors que la VPP est faible. Cela signifie qu'un patient ayant réagi positivement au test a cependant une probabilité faible d'être atteint par la maladie. Des examens complémentaires doivent être envisagés. En revanche, il y a de fortes chances qu'un patient ayant réagi négativement au test ne soit pas malade. Il est à noter que, même en cas d'épidémie, la prévalence (proportion d'individus malades dans une population) atteint rarement des valeurs supérieures à 40 %.

L'appliquette GeoGebra « Valeur prédictive en fonction de la prévalence» permet de visualiser différentes courbes de valeurs prédictives en fonction des caractéristiques du test.

Inférence bayésienne et détection de spams

Un des premiers programmes de filtrage bayésien du courrier électronique était le programme iFile de Jason Rennie, publié en 1996.

Le principe, analogue à celui du diagnostic médical, repose sur le fait que les mots du dictionnaire ont des probabilités différentes d'apparaître dans les spams et dans les courriers légitimes.

Le filtre de détection des spams ne connaît pas à l'avance les probabilités d'apparition de ces mots, c'est pourquoi il lui faut une phase d'apprentissage pour les évaluer. Cette phase d'apprentissage est analogue à la phase de calibrage du test médical étudié ci-dessus.

L'apprentissage se fait à partir de l'observation du comportement des utilisateurs, qui doivent indiquer manuellement si un message est un spam ou non. Pour chaque mot de chaque message « appris », le filtre ajustera les probabilités de rencontrer ce mot dans un spam ou dans un courrier légitime et le stockera dans sa base de données.

On note $P_S(M)$ la probabilité qu'un spam contienne le mot M et $P_{\bar{S}}(M)$ la probabilité qu'un courrier légitime contienne le mot M. Ces deux probabilités sont estimées au cours de la phase d'apprentissage, tout comme la probabilité $P(S)$ qu'un message quelconque soit un spam (analogue à la prévalence $P(M)$ dans le test médical).

Une fois ces valeurs déterminées, la formule de Bayes permet de calculer la probabilité qu'un message donné soit un spam sachant qu'il contient le mot M selon la formule.

$$P_M(S) = \frac{P(M \cap S)}{P(M)} = \frac{P_S(M) \times P(S)}{P_S(M) \times P(S) + P_{\bar{S}}(M) \times (1 - P(S))}$$

Cette probabilité est comparée à un seuil ; si elle est supérieure au seuil, le filtre classera ce message dans les spams.

Dans la réalité, on travaille non pas sur un seul mot M, mais sur un stock de mots, en faisant l'hypothèse naïve que les mots présents dans un message sont indépendants les uns des autres. Cela est faux dans les langages naturels, où par exemple la probabilité de trouver un adjectif est influencée par celle de trouver un nom. De plus, cette technique de filtrage, connue sous le nom de filtrage bayésien naïf, ne tient pas compte du sens des mots, alors qu'il a une incidence sur la présence simultanée de certains mots à l'intérieur du message. Par exemple, la présence du mot « anniversaire » n'est pas indépendante de celle du mot « joyeux ».

Propositions d'activités

Activité 1

Parmi les femmes de 40 ans ayant effectué une mammographie, 1 % a un cancer du sein.
À la suite de mammographies sur échantillon, on a établi que :

- pour 82 % des femmes ayant un cancer du sein, la mammographie détecte une anomalie;
- pour 9 % des femmes n'ayant pas de cancer du sein, la mammographie détecte une anomalie.

On suppose que 10 000 de 40 ans ont effectué une mammographie.

1. Préciser les caractéristiques (sensibilité, spécificité) d'unemammographie.
2. Compléter le tableau ci-dessous :

	Anomalie détectée	Pas d'anomalie détectée	Total
Malades			
Non malades			
Total			10 000

3. Une femme de 40 ans a subi une mammographie qui a détecté une anomalie. Quelle est la probabilité qu'elle soit atteinte d'un cancer du sein?
4. Calculer les valeurs prédictives positive et négative d'une mammographie chez les femmes de 40 ans.

L'appliquette GeoGebra « Sensibilité, spécificité, valeurs prédictives » permet de s'exercer sur d'autres activités de ce type.

Activité 2

Montrer que si $S_e+S_p=1$, alors le test est inutile, dans le sens où $P_{T+}(M)=P(M)$ et $P_{T-}(\overline{M})=P(\overline{M})$. Les probabilités a posteriori sont égales aux probabilités a priori et le test est inutile.

Activité 3 : dépistage du VIH

L'infection par le virus de l'immunodéficience humaine reste un problème de santé publique à l'échelle mondiale.

1. Aujourd'hui, il existe des tests rapides pour l'infection à VIH appelés « Test Rapide d'Orientation Diagnostique » ou TROD. Ces tests ont l'avantage de pouvoir être réalisés à partir d'un échantillon de salive ou à partir d'une goutte de sang prélevée au bout du doigt. Pour comparer les caractéristiques de ces deux tests (salivaire et sanguin), on a réalisé les tests TROD sur 10 000 personnes dont on sait qu'elles sont infectées par le VIH et sur 100 000 personnes noninfectées.

Les caractéristiques des tests salivaire et sanguin sont les suivantes :

	Personnes infectées par le VIH	Personnes non infectées par le VIH
Test salivaire positif	9803	260
Test sanguin positif	9968	90

Calculer la spécificité et la sensibilité de chacun de ces tests.

2. Influence de la prévalence sur les valeurs prédictives destests.
   a. En 2017, la population mondiale exposée était estimée à 6 milliards et parmi elle, le nombre de personnes infectées par le VIH à 37millions. Calculer la valeur prédictive positive de chacun des deux tests pour la population mondiale exposée.
   b. En 2017, la population française exposée était estimée à 50 millions et le nombre de personnes infectées par le VIH à 150000. Calculer la valeur prédictive positive de chacun des deux tests pour la population française exposée.
   c. En 2017, la population sud-africaine exposée était estimée à 35 millions et le nombre de personnes infectées par le VIH à 7millions. Calculer la valeur prédictive positive de chacun des deux tests pour la population sud-africaine exposée.

Activité 4 : aide au diagnostic

Trois maladies virales peuvent être transmises par les moustiques : dengue, chikungunya et zika. Elles provoquent des symptômes qui peuvent être assez proches. Il peut être difficile de les différencier directement. Ici on s'intéresse à la mise en place d'une aide statistique au diagnostic. Pour cela, on va s'appuyer sur des données obtenues chez des personnes dont le diagnostic a pu être certifié par des examens biologiques. Pour simplifier, on supposera que ces caractères apparaissent indépendamment chez les personnes infectées.

Symptômes	Dengue	Chikungunya	Zika
Fièvre	95 %	75 %	75 %
Courbatures	75 %	95 %	50 %
Douleur oculaire	50 %	25 %	50 %
Déficit globules blancs	50 %	50 %	25 %
Hémorragie	25 %	5 %	5 %

À partir de ces données, on veut déterminer les probabilités de chaque maladie selon les symptômes présentés et dans des conditions différentes.

1. On suppose qu'une personne malade revient d'un pays dans lequel aucune de ces maladies n'est épidémique. On considère donc a priori que les trois maladies sont équiprobables. Quelles sont les probabilités de chaque maladie si cette personne présente à la fois de la fièvre, pas de courbatures et des douleurs oculaires ? Quel est le diagnostic le plus probable dans ce cas?
2. On suppose qu'une personne malade revient d'un pays dans lequel sévit une épidémie de Zika. A priori, y a 80 % de chances qu'elle ait été infectée par Zika et 10 % par chacune des deux autres maladies. Quelles sont les probabilités a posteriori de chaque maladie si cette personne présente à la fois de la fièvre, pas de courbatures et des douleurs oculaires ? Quel est le diagnostic le plus probable dans ce cas ?