Une maladie (exemple : cancer) est présente dans une population dans la proportion d'une personne malade sur 10 000, soit 0,01 %.

Un patient vient de passer un test pour le dépistage de cette maladie.
Le médecin le convoque pour lui annoncer le résultat : mauvaise nouvelle, il est positif.
Il lui indique alors que ce test est plutôt fiable :

« Si vous avez cette maladie, le test sera positif dans 99 % des cas.
Si vous ne l'avez pas, il sera négatif dans 99,8 % des cas ».

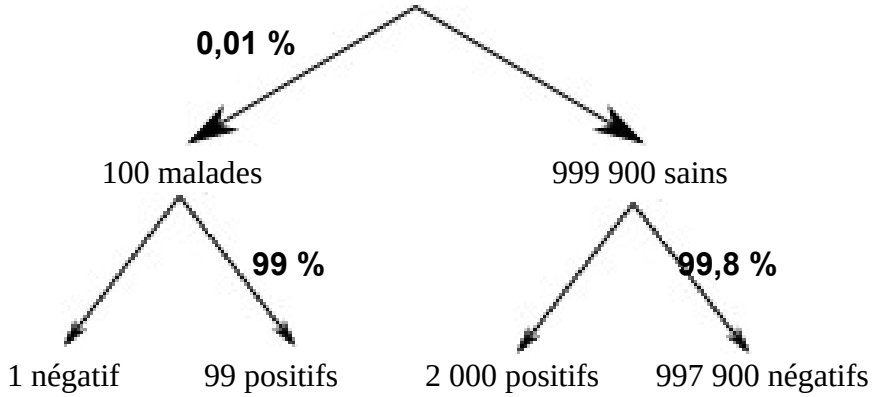
A votre avis, puisque le test est positif, quelle est la probabilité que le patient ait la maladie ?

☐ 90 % ? ☐ 80 % ☐ 70 % ☐ 60 % ☐ moins de 60 % ☐ moins de 30 %

Si vous avez répondu autre chose que « moins de 30 % », c'est que vous avez été trompé par ce biais cognitif bien connu, appelé « **oubli de la fréquence de base** » (aussi connue sous le nom de négligence de la taille de l'échantillon). Autrement dit, vous avez oublié de considérer la fréquence de base de l'occurrence de l'événement dont on cherche la probabilité… Le plus souvent, cela conduit à surestimer cette probabilité.

Les exemples les plus typiques de cette surestimation sont, en médecine, les surdiagnostics concernant le dépistage de certains cancers (seins, prostate, mais aussi poumons et thyroïde), l'asthme ou encore les troubles du déficit de l'attention.

Prenons un exemple : 1 000 000 personnes testées



Avec 1 000 000 de personnes testées, il y a 100 malades et 999 900 non malades puisque 0,01 % de la population est malade.
D'après les affirmations du médecin sur la fiabilité du test, on a alors :

- parmi les 100 malades, **99 auront un test positif** ;
- parmi les 999 900 non malades, **2 000 auront un test positif**
  (puisque $0{,}2\% \times 999\,900 \approx 2\,000$).

Il y a donc 2 099 tests positifs, parmi lesquels 99 correspondent à des personnes malades.

$\frac{99}{99+2000} \approx 0{,}047$ donc :

avec un test positif, la probabilité que le patient ait la maladie est d'environ 4,7 %.

Autrement dit, il y a **95,3 % de faux positifs** : 95,3 % des tests positifs désignent des personnes saines !

De même, avec un test négatif, la probabilité que le patient soit sain est :

$$\frac{997900}{997901} \approx 99{,}9998998\ \%.$$

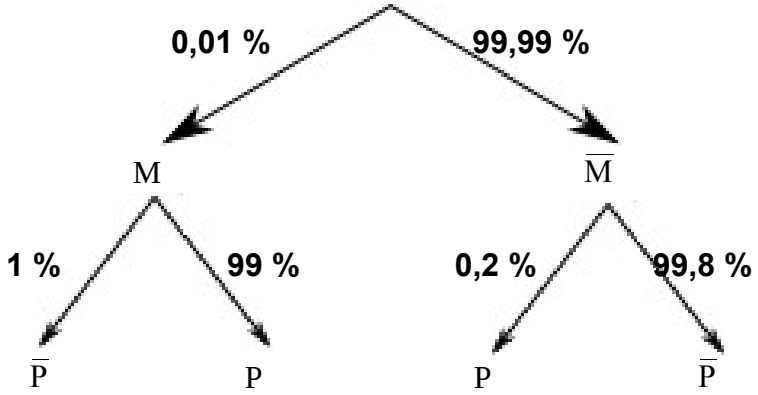
Autrement dit, il y a **0,0001 % de faux négatifs.**

Conclusion :

Pratiquement tous les malades présentent un test positif
…
mais pratiquement tous les tests positifs désignent des personnes saines !

SOLUTION AVEC LES PROBABILITÉS CONDITIONNELLES

M : « la personne est malade » P : « le test est positif »

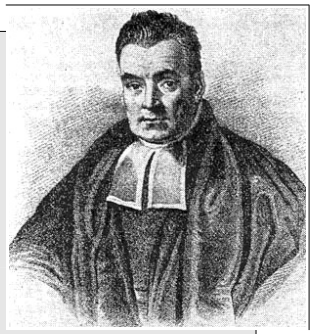


$$p_P(M)=\frac{p_M(P)\times p(M)}{p(P)} \quad \leftarrow \text{c'est ce qu'on appelle la formule de Bayes}$$
$$=\frac{p_M(P)\times p(M)}{p_M(P)\times p(M)+p_{\overline{M}}(P)\times p(\overline{M})}$$
$$=\frac{99\%\times 0{,}01\%}{99\%\times 0{,}01\%+0{,}2\%\times 99{,}99\%}$$
$$=\frac{0{,}000099}{0{,}0020988}$$
$$\approx 0{,}0471698$$

FORMULE DE BAYES

$$\text{proba. cause sachant } \textit{conséquence}=\frac{\text{proba. } \textit{conséquence} \text{ sachant cause}\times\text{proba. cause}}{\text{proba. } \textit{conséquence}}$$

Ce théorème est aussi appelé "formule de probabilité des causes" : elle permet en effet de **calculer la probabilité d'une cause sachant celle de sa (ses) conséquence(s)**.



Thomas Bayes (1702-1761) est un mathématicien britannique et pasteur de l'Église presbytérienne.

À sa mort, Bayes laisse à son ami Richard Price ses articles non terminés.
Il prendra l'initiative de publier l'article de Bayes et de l'envoyer à la Royal Society deux ans plus tard.

Il est probable que Richard Price ait lui-même contribué de manière significative à la rédaction de l'article final et qu'il soit ainsi avec Thomas Bayes l'auteur du théorème connu sous le nom de théorème de Bayes.

Probabilité que : → Sachant que : ↓	Test positif	Test négatif	Malade	Non malade (saine)
Test positif	-		4,7 %	95,3 % (faux positifs)
Test négatif			0,0001 % (faux négatifs)	99,9998998 %
Malade	99 %	1 %	-	
Non malade (saine)	0,2 %	99,8 %		

Vocabulaire

• **Sensibilité** : probabilité que le test soit positif sachant que la personne est malade.
• **Spécificité** : probabilité que le test soit négatif sachant que la personne est saine.
En statistique, la sensibilité (ou sélectivité) d'un test mesure sa capacité à donner un résultat positif lorsqu'une hypothèse est vérifiée. Elle s'oppose à la spécificité, qui mesure la capacité d'un test à donner un résultat négatif lorsque l'hypothèse n'est pas vérifiée.

• **Valeur prédictive positive (VPP)** : probabilité que la personne soit malade sachant que le test est positif.
• **Valeur prédictive négative (VPN)** : probabilité que la personne soit saine sachant que le test est négatif.
En notant VP les vrais positifs, VN les vrais négatifs, FP les faux positifs et VP les vrais positifs, on arrive facilement aux formules suivantes :

$$\text{VPP} = \frac{\text{VP}}{\text{VP}+\text{FP}} \text{ et } \text{VPN} = \frac{\text{VN}}{\text{VN}+\text{FN}}.$$

La probabilité qu'une personne soit malade dans la population est appelée **prévalence de la maladie** (dans mon exemple, c'est donc 0,01 %).

Dans mon exemple, je donnais donc la prévalence de la maladie, la sensibilité et la spécificité.
Je demandais alors la valeur prédictive positive (VPP).

Selon le message que je souhaite faire passer concernant les liens entre ce test et la maladie qu'il diagnostique, je peux facilement choisir le pourcentage approprié…
Comme dirait A. Levenstein, les statistiques, c'est comme le bikini : ce qu'elles révèlent est suggestif mais ce qu'elles dissimulent est essentiel !

Mais alors, puisque la probabilité qu'une personne soit malade sachant que son test est positif est très faible (4.7 %), voilà que ce test nous paraît un peu "inutile"… Non ?

Pas tant que ça, car cette probabilité (en vert) est liée à la probabilité qu'un patient soit sain sachant que son test est négatif (en rouge/rose). Et mieux vaut que cette dernière soit très proche de 100 % : il vaut mieux inquiéter quelqu'un à tort que de lui dire que tout va bien alors que ce n'est pas le cas… En médecine comme ailleurs, **on mesure les risques et on essaie de les équilibrer**.

D'autre part, en faisant ce test à une population, il sera positif pour environ 0.21 % des personnes*. Certes, beaucoup de ces gens seront en réalité non malades, mais il suffira pour cela de faire des tests complémentaires plus long et plus onéreux…

En faisant ce test, on a évité de faire faire les tests complémentaires à toute la population : **on a ainsi divisé la population de départ à examiner par plus de 476** !

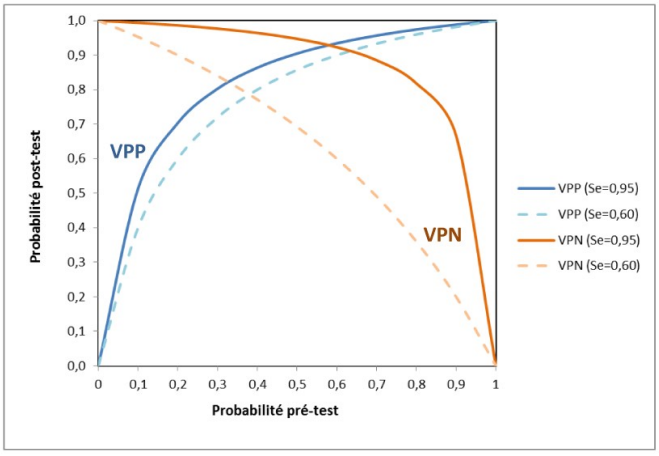
* calcul effectué : 99 %×0.01 %+0.2 %×99.99 %=0.20988 %

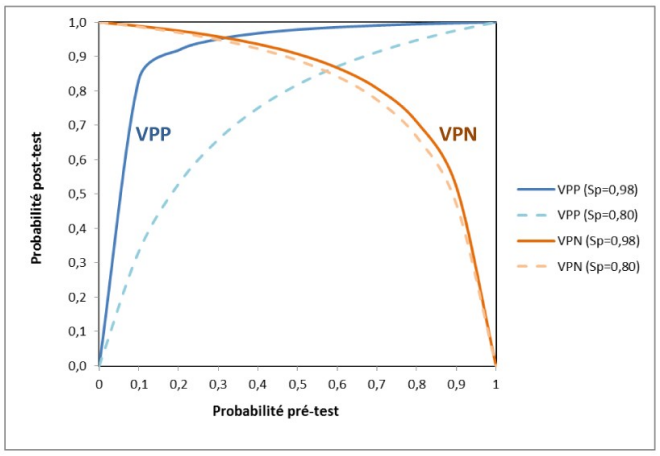
> Imaginons ce test sur une population de 40 000 000 de personnes : il sera positif pour 83 952 personnes, dont 4 000 seulement seront vraiment malades.
> Il faudra donc effectuer les tests complémentaires sur ces 83 952 personnes au lieu des 40 millions.
> Par ailleurs, n'oublions pas que 40 personnes auront un test négatif tout en étant malades… :(

COMPLÉMENTS

La VPP et la VPN dépendent de la prévalence de la maladie, de la sensibilité et de la spécificité.
On peut montrer que :

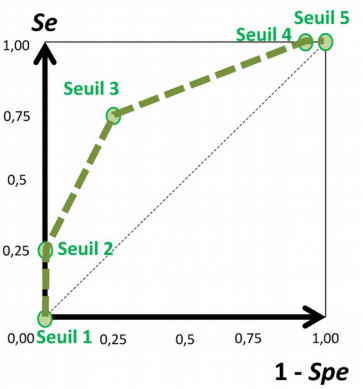
- classiquement, sensibilité et spécificité varient en sens opposé ;
- si la prévalence augmente, alors la VPP augmente et la VPN diminue ;
- l'influence d'un gain de sensibilité est plus marqué sur la VPN que sur la VPP ;
- l'influence d'un gain de spécificité est plus marqué sur la VPP que sur la VPN.



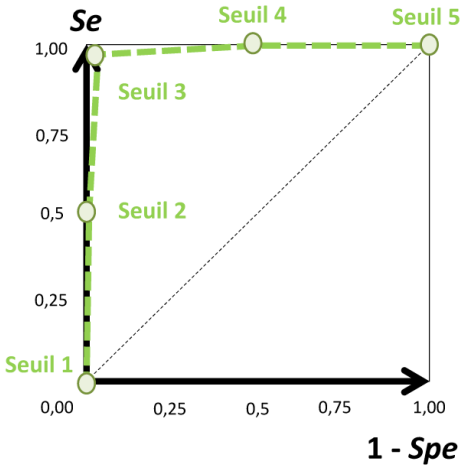


Lorsque le diagnostic se fonde sur la valeur d'une variable quantitative, il faut choisir un seuil au-delà duquel on considère le diagnostic comme positif. Pour représenter les différentes valeurs de sensibilité et spécificité associées aux différents seuils possibles, on construit une **courbe ROC** (*Receiver Operating Curve* ou *Receiver Operating Characteristic*, pour « caractéristique de fonctionnement du récepteur »).

Seuil	Se	Sp	1 - Sp
Seuil 1	0%	100%	0%
Seuil 2	25%	100%	0%
Seuil 3	75%	75%	25%
Seuil 4	100%	2%	98%
Seuil 5	100%	0%	100%



Voici l'exemple d'un test diagnostique qui permet de distinguer les malades des non malades pratiquement sans erreur : la courbe ROC se rapproche du coin supérieur gauche.



CELA SERT-IL DANS D'AUTRES DOMAINES ?

Bien sûr !

• Par exemple, le raisonnement bayésien est aussi utilisé pour le **filtrage des spams**.

L'hypothèse initiale H est par exemple « tel message est un spam », puis l'algorithme réalise un certain nombre d'observations concernant le contenu du message (son expéditeur, les mots employés, la présence de liens, etc.)

A chacune de ces observations, grâce au théorème de Bayes, l'algorithme met à jour son estimation de la probabilité que le message soit un spam : il détermine la probabilité d'une cause sachant les observations faites. Une fois toutes les observations effectuées, en fonction de la valeur de la probabilité a posteriori, il peut décider de classer ou non le message comme spam.

• On l'utilise pour l'auto-apprentissage machine en **intelligence artificielle : analyse d'images, cassage de codes, reconnaissance visuelle ou de la parole, deep learning, etc**.

• En criminalistique, c'est très souvent utilisé. Vous pouvez télécharger ce document pour en savoir un peu plus : https://www.lpc-expert.fr/Articles/article_Bayes_criminalistique.pdf

On y voit un exemple marquant où la probabilité qu'un suspect soit la source d'une trace ADN sachant que le test ADN est positif est très faible… Contre-intuitif !

• En **physique des particules**, on utilise le théorème de Bayes pour évaluer la probabilité d'existence d'une particule. En effet, ils produisent des particules ayant une durée de vie trop courte pour être observable : s'il n'est donc pas possible de voir directement ces particules, il est en revanche possible d'observer ce qui reste après leur désintégration.

Malheureusement, plusieurs particules peuvent avoir les même produits de désintégration. En observant ces produits de désintégration, c'est-à-dire un événement se produisant avec une probabilité donnée, les physiciens cherchent donc à mesurer la probabilité d'avoir produit une particule donnée en fonction des produits de dés intégration qu'ils observent.
La difficulté qu'ils rencontrent, qui est d'ailleurs souvent le principal obstacle à une utilisation efficace du théorème de Bayes, est qu'il n'est pas facile de déterminer une valeur acceptable pour la probabilité de chacune des causes possibles. Autrement dit, on est conduit à faire des hypothèses qui peuvent être sujettes à caution. Elles sont d'ailleurs l'objet d'une polémique, car elles ne s'appuient pas toujours sur des arguments physiques.

Source : Tangente HS n°17 (Nicolas Delerue)

Une application étonnante : la contrebande d'ivoire

Gilles Guillot, de l'Université technique du Danemark, décrit une application originale : les statistiques bayésiennes sont utilisées pour **identifier l'origine des ivoires d'Afrique saisis par la douane aux aéroports**.

L'ADN prélevé sur les ivoires est comparé à celui d'éléphants dont l'origine géographique est bien identifiée ; la formule de Bayes utilise ces informations pour calculer la probabilité que l'échantillon provienne d'une certaine latitude et longitude, et pour identifier ainsi son origine probable.

A l'échelle du continent africain, la moitié des échantillons peuvent ainsi être localisés avec une erreur inférieure à 500 km.

QUAND UTILISER LES STATISTIQUES BAYÉSIENNES ?

Les deux approches se complètent, la statistique classique étant en général préférable lorsque les informations sont abondantes et d'un faible coût de collecte.

Ainsi, un sondage d'opinion ne coûte que quelques euros et un test en fin de chaîne de fabrication que quelques centimes : les statistiques classiques conviennent alors parfaitement.

Lorsqu'il est question de s'informer en effectuant un forage pétrolier, le coût des mesures devient tel que les méthodes bayésiennes, qui les minimisent, sont préférables.

En cas de profusion de données, les résultats sont asymptotiquement les mêmes dans chaque méthode, la bayésienne étant simplement plus coûteuse en calcul.

En revanche, la méthode bayésienne permet de traiter des cas où la statistique ne disposerait pas suffisamment de données pour qu'on puisse en appliquer les théorèmes.