Revenons sur le texte d'Étienne Klein que je vous ai invité à lire.

«Faisons une expérience de pensée. Imaginons que dans une population donnée apparaisse une nouvelle maladie, qui affecte une personne sur mille. Les symptômes de cette pathologie n'étant ni visibles ni ressentis, nul ne sait dire qui est malade et qui ne l'est pas. Mais les chercheurs s'activent et finissent par mettre au point un test de dépistage dont l'efficacité est de 95 %.

Cela signifie que sur cent personnes positives à ce test, en moyenne quatre-vingt-quinze sont effectivement malades et cinq sont ce qu'on appelle des «faux positifs» (c'est-à-dire sont positifs au test sans être malades).¹

Soit maintenant une personne qui se révèle positive au test : quelle est la probabilité qu'elle soit malade ?

Si vous réalisez un sondage dans votre entourage, vous constaterez que la proportion de ceux qui répondent «95 %» à cette question est très élevée.

Or, la bonne réponse est... seulement de 2 % !

Autrement dit, une personne positive au test a quatre-vingt dix-huit chances sur cent de ne pas être malade !

Ce résultat violemment contre-intuitif s'obtient à l'issue d'un raisonnement qui est pourtant simple.»
(fin de la citation d'Étienne Klein)

C'est un exercice classique du domaine des probabilités et de la formule de Bayes sur les probabilités conditionnelles.

Le mot probabilités conditionnelles ne me plaît pas trop. Appelons le "probabilités réduites à un sous-ensemble de l'univers".

Elles correspondent à un redécoupage de l'univers.

Notons Ω notre univers. ici, la population.

Un événement est une partie de cet univers : par exemple les personnes affectées A . On a alors l'événement contraire $\bar{A}$ les personnes non affectées.

La probabilité de $\bar{A}$ est évidemment $1 - p(A)$.

événement	A	$\bar{A}$
probabilité	$\frac{1}{1000}$	$\frac{999}{1000}$

On a ensuite un autre événement : test positif qu'on va noter T .

On ne connaît pas la probabilité de T ni celle de $\bar{T}$ sur tout l'univers Ω .

On peut donc découper Ω en quatre parties :

	A	$\bar{A}$
T	affectés positifs au test	non affectés positifs au test
$\bar{T}$	affectés négatifs au test	non affectés négatifs au test

On ne connaît a priori la probabilité d'aucun de ces événements. On va leur donner des noms :

	A	$\bar{A}$
T	α	β
$\bar{T}$	γ	δ

avec $\alpha + \beta + \gamma + \delta = 1$.

On peut alors reconstituer la probabilité de chacun des événements fondamentaux, en sommant en

colonnes ou en lignes

	A	$\bar{A}$
T	α	β
$\bar{T}$	γ	δ

$p(T) = \alpha + \beta$
 $p(\bar{T}) = \gamma + \delta$
 $p(A) = \alpha + \gamma$ $p(\bar{A}) = \beta + \delta$

On retrouve au passage $p(\bar{A}) = 1 - p(A)$. On sait aussi $\alpha + \gamma = \frac{1}{1000}$: "une nouvelle maladie, qui affecte une personne sur mille".

1. on reformulera plus loin cette hypothèse

	A	$\bar{A}$	
T	α	β	$p(T) = \alpha + \beta$

On découpe maintenant le sous-univers (la condition) "test positif"

	γ	δ	
$\bar{T}$	γ	δ	$p(\bar{T}) = \gamma + \delta$
	$p(A) = \alpha + \gamma$	$p(\bar{A}) = \beta + \delta$	

Et on se retrace à cet univers (de probabilité non nulle) :

	A	$\bar{A}$
T	α	β

 $p(T) = \alpha + \beta$
 et on lit des probabilités dites conditionnelles car il faut se ramener à "masse totale=1" :

	A	$\bar{A}$	
T	α	β	$p(T) = \alpha + \beta$
	$P_T(A) = \frac{\alpha}{\alpha + \beta}$	$P_T(\bar{A}) = \frac{\beta}{\alpha + \beta}$	

On retrouve au passage $P_T(\bar{A}) = 1 - P_T(A)$. On les note aussi $P_T(A) = P(A|T)$ (et on lit probabilité de A sachant T, ou probabilité de A sur le sous-univers T).

La formule explicite est $P_T(A) = \frac{P(A \cap T)}{P(T)}$.

On peut aussi découper dans l'autre sens par exemple sur les individus sains :

	A	$\bar{A}$			
T	α	β	$p(T) = \alpha + \beta$	puis	probabilités conditionnelles
$\bar{T}$	γ	δ	$p(\bar{T}) = \gamma + \delta$		
	$p(A) = \alpha + \gamma$	$p(\bar{A}) = \beta + \delta$			
					$p_{\bar{A}}(T) = \frac{\beta}{\beta + \delta}$
					$P_{\bar{A}}(\bar{T}) = \frac{\delta}{\beta + \delta}$
					$P_{\bar{A}}(\bar{A}) = 1$

On lit cette fois la probabilité d'erreur de dépistage : probabilité d'être déclaré positif alors qu'on n'est pas atteint = $\frac{\beta}{\beta + \delta}$.

On nous rappelle une information : *sur cent personnes positives à ce test, en moyenne quatre-vingt-quinze sont effectivement malades et cinq sont ce qu'on appelle des «faux positifs».*
 Non, celle là ne me plait pas dans la narration : une personne saine n'a que cinq pour cent de risque d'être déclarée atteinte.

On la traduit : $P_{\bar{A}}(T) = \frac{95}{100}$. On a donc une nouvelle équation $\frac{\beta}{\beta + \delta} = \frac{5}{100}$.

On a assez d'informations pour compléter le tableau ?

Non, on n'a que trois équations pour quatre inconnues : $\frac{\beta}{\beta + \delta} = \frac{5}{100}$, $\beta + \delta = \frac{1}{1000}$ et $\alpha + \beta + \gamma + \delta = 1$.

Il nous faut une hypothèse de plus qui est implicite dans le texte d'Étienne Klein, sur l'erreur de seconde espèce.

On distingue dans ce cadre effectivement deux types d'erreur de la part du test :

type 1 : vous êtes affectés, mais le test ne le détecte pas

type 2 : vous n'êtes pas affecté mais le test se plante et vous dit "positif"

Ici, Étienne Klein donne l'erreur de type 2 (5 pour cent d'erreurs de ce type).

On va estimer que l'erreur de type 1 est impossible. Nul ne passe au travers du test. On a donc $\bar{T} \cap A = \emptyset$ et $\gamma = 0$ (certains se disent peut être "enfin, je l'avais dit dès le début moi").

Cette fois, on a assez d'équations pour compléter le tableau :

	A	$\bar{A}$
T	1	999
	$\frac{1}{1000}$	$\frac{20000}{18981}$
$\bar{T}$	0	$\frac{20000}{20000}$

On revient cette fois à l'univers des personnes positives au test :

	A	$\bar{A}$	
T	1	999	$\frac{1\ 019}{20\ 000}$
	$\frac{1}{1\ 000}$	$\frac{999}{20\ 000}$	

On calcule les probabilités conditionnelles :

	A	$\bar{A}$	
T	1	999	$\frac{1\ 019}{20\ 000}$
	$\frac{1}{1\ 000}$	$\frac{999}{20\ 000}$	

$$P_T(A) = \frac{20}{1\ 019} \quad P_T(\bar{A}) = \frac{999}{1\ 019}$$

On interprète : sachant que le teste est déclaré positif, la probabilité que vous soyez effectivement affecté est de $P_T(A) = \frac{20}{1\ 019}$ soit bel et bien le 2 pour cent de l'article.

La formule qu'on pouvait même utiliser est directement celle des Bayes :
on écrit $P_T(A) = \frac{P(A \cap T)}{P(T)}$ et $P_{\bar{A}}(T) = \frac{P(A \cap T)}{P(A)}$, on élimine $P(A \cap T)$: $P_T(A) = P_{\bar{A}}(T) \cdot \frac{P(A)}{P(T)}$
Et même $P_T(A) = P_{\bar{A}}(T) \cdot \frac{P(A)}{P_{\bar{A}}(T) \cdot P(\bar{A}) + P_{\bar{A}}(T) \cdot P(A)}$ en ayant écrit $P(T) = P(T \cap \bar{A}) + P(T \cap A) = P_{\bar{A}}(T) \cdot P(\bar{A}) + P_{\bar{A}}(T) \cdot P(A)$.

Bilan : il faut effectivement se méfier des assertions trop rapides.

Même de la part d'Étienne Klein.
La phrase *Cela signifie que sur cent personnes positives à ce test, en moyenne quatre-vingt-quinze sont effectivement malades et cinq sont ce qu'on appelle des «faux positifs»* racontait des choses sur $P_T(\bar{A})$. Elle voulait en fait raconter des choses sur $P_{\bar{A}}(T)$, c'est à dire "si une personne n'est pas malade, le test est positif dans 5% des cas".
Le texte a été trop vite publié.

Je remercie Gallien de m'avoir mis la puce à l'oreille. En effet, les deux phrases
« *Cela signifie que sur cent personnes positives à ce test, en moyenne quatre-vingt-quinze sont effectivement malades et cinq sont ce qu'on appelle des «faux positifs»* »
et « *une personne qui se révèle positive au test : quelle est la probabilité qu'elle soit malade ?* »
expriment un peu trop la même chose, non ?

On félicitera quand même Étienne Klein pour son grand travail de vulgarisateur scientifique, pour son œuvre physique énorme, et pour ses anagrammes ludiques :
le saviez vous : "CHAUVE SOURIS" est anagramme de "SOUCHE À VIRUS".