Les statistiques : bien les comprendre pour mieux décider

Avner Bar-Hen

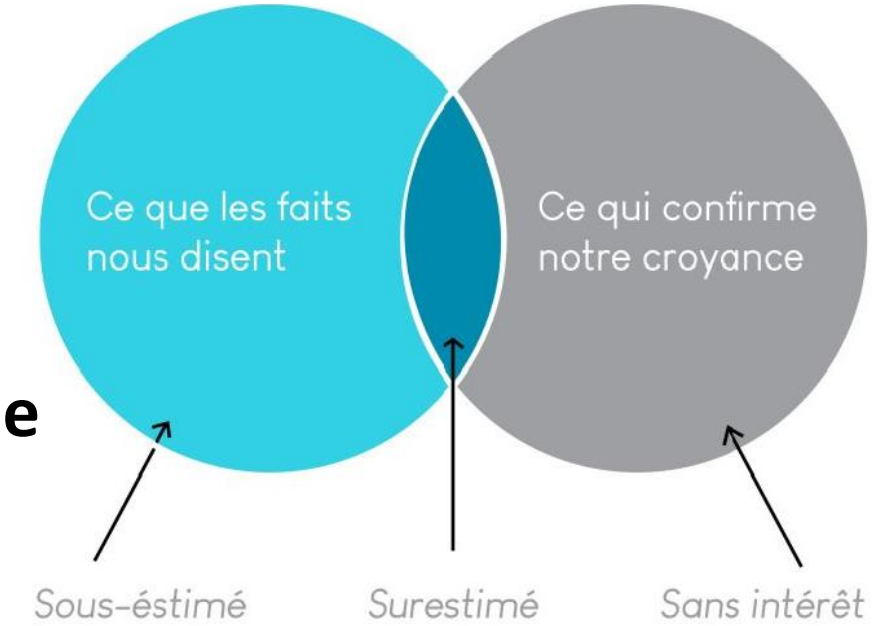
Professeur du Cnam, chaire «Statistique et données massives».

©Domo

- Dans de nombreux domaines, les données s’accumulent en masse.

- **Analyser les données pour les valoriser**

Biais de confirmation



Le **biais de confirmation**, également dénommé **biais de confirmation** d'hypothèse, est le biais cognitif qui consiste à privilégier les informations confirmant ses idées préconçues ou ses hypothèses et/ou à accorder moins de poids aux hypothèses et informations jouant en défaveur de ses conceptions (réticence à changer d'avis).

Exemple

Jours	Pluie	Pas de pluie
Arthrite	14	6
Pas d'arthrite	7	2

©Wikipedia

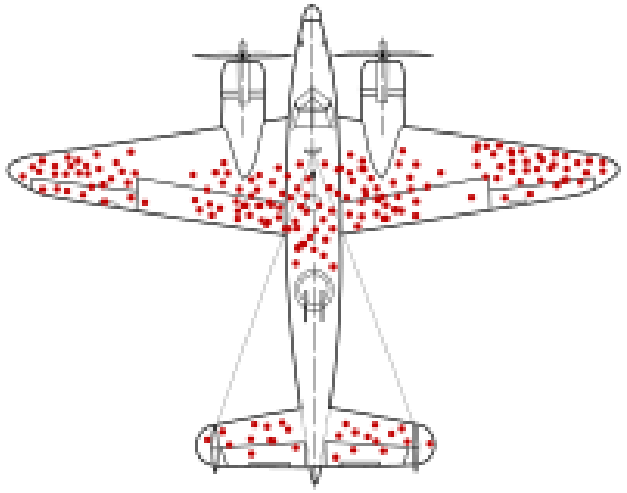
Durant les jours de pluie, 66 % (14/21) de personnes souffrent de l'arthrite

MAIS

75 % (6/8) de personnes souffrent de cette maladie durant les jours où il n'y a pas de pluie.

Biais du survivant

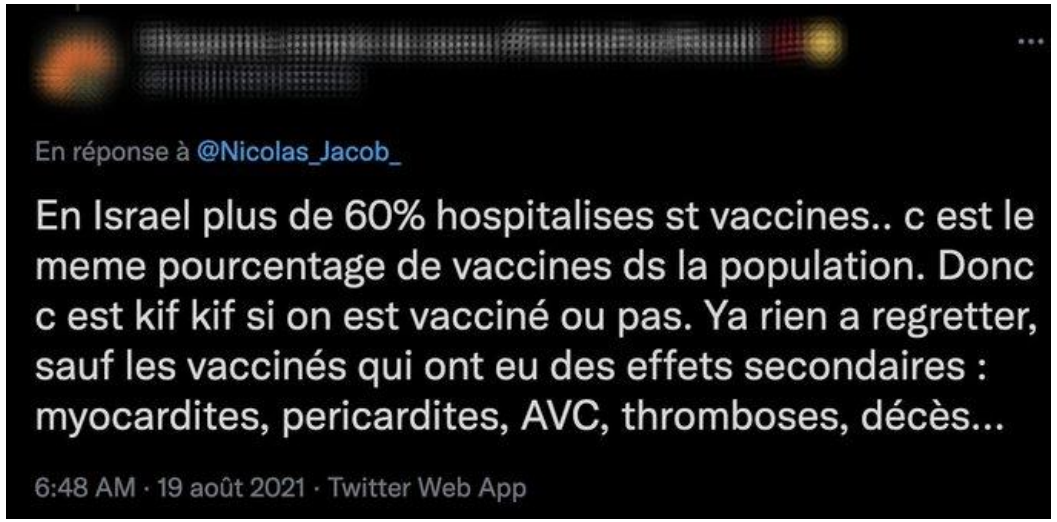
- Le **biais du survivant** est une forme de biais de sélection consistant à surévaluer les chances de succès d'une initiative en concentrant l'attention sur les sujets ayant réussi mais qui sont des exceptions statistiques (des « survivants ») plutôt que des cas représentatifs.



Les endroits endommagés des avions revenus du front montrent les endroits où ils peuvent être endommagés et espérer revenir à la base. Les avions endommagés aux autres endroits ne reviennent pas.

Biais du survivant

Paradoxe de Simpson



AU 15 AOÛT 2021

Population	Vaccinés 2 doses	Hospitalisations pour cas sévère	dont vaccinés 2 doses
9 053 000	5 634 634	515	301
100%	62%	100%	58,4%

Source : Israeli government data dashboard

301/515 ≈ 0,584

Paradoxe de Simpson

- Données brutes, en valeur absolue, sans tenir compte des effectifs respectifs entre vax et non vax
- Normalisons pour 100 000 vax et 100 000 non vax

AU 15 AOÛT 2021

Population		Cas sévères		Efficacité contre les cas sévères
Non Vaccinés	Vaccinés 2 doses	Non Vaccinés	Vaccinés 2 doses	
3 418 366 soit 37,7%	5 634 634 soit 62,3%	214 soit 6,26 pour 100 000	301 soit 5,34 pour 100 000	14,7%

Source : Israeli government data dashboard

Efficacité : (1-(5,34/6,26))x100=14,7% (critère OMS : 50% min.)

Paradoxe de Simpson

Un facteur confondant : un élément qui vient influer non seulement sur les conséquences que nous étudions (ici la fréquence des cas sévères) mais aussi sur la cause que l'on teste (ici le statut vaccinal)

La vaccination n'est pas ouverte aux < 12 ans

AU 15 AOÛT 2021

Population >12 ans		Cas sévères		Efficacité contre les cas sévères
Non Vaccinés	Vaccinés 2 doses	Non Vaccinés	Vaccinés 2 doses	
1 302 912 soit 18,2%	5 634 634 soit 78,7%	214 soit 16,4 pour 100 000	301 soit 5,3 pour 100 000	67,5%

Source : Israeli government data dashboard

(1-(5,3/16,4)x100=67,5% d'efficacité.

Paradoxe de Simpson

L'âge influe sur le statut vaccinal **ET** sur la probabilité de faire une forme grave.

AU 15 AOÛT 2021

Âge	Population		Cas sévères		Efficacité contre les cas sévères
	Non Vaccinés	Vaccinés 2 doses	Non Vaccinés	Vaccinés 2 doses	
Tous âges	3 418 366 soit 37,7%	5 634 634 soit 62,3%	214 soit 6,26 pour 100 000	301 soit 5,34 pour 100 000	14,7%
>12ans	1 302 912 soit 18,2%	5 634 634 soit 78,7%	214 soit 16,4 pour 100 000	301 soit 5,3 pour 100 000	67,5%
12-50ans	1 116 834 soit 24,2%	3,501 118 soit 75,8%	43 soit 3,9 pour 100 000	11 soit 0,3 pour 100 000	91,8%
50 et plus	186 078 soit 8%	2 133 516 soit 92%	171 soit 91,9 pour 100 000	290 soit 13,6 pour 100 000	85,2%

Source : Israeli government data dashboard

Paradoxe de Simpson

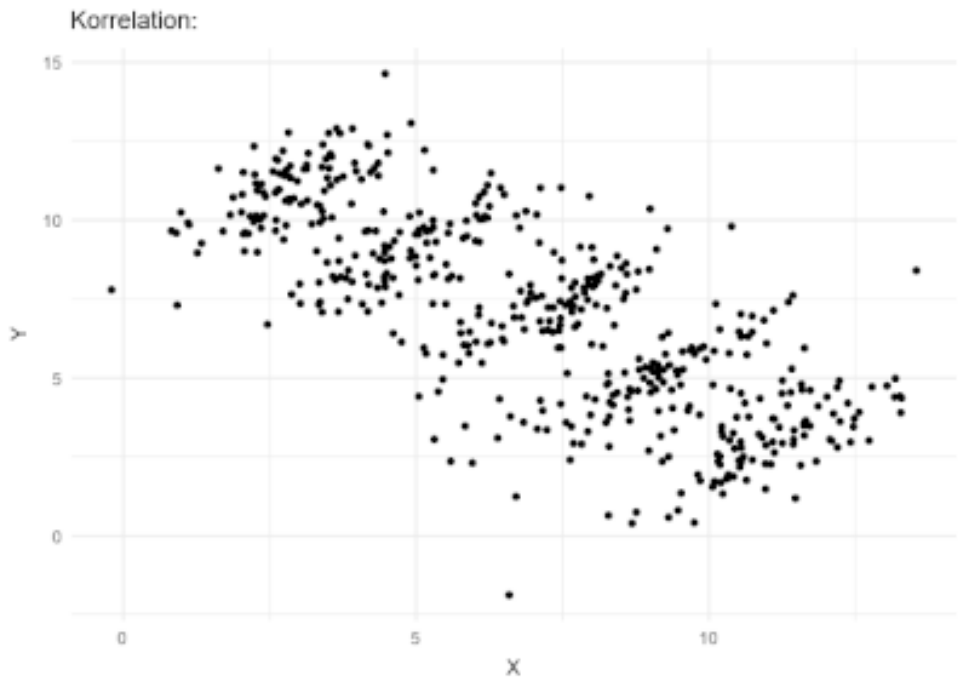
AU 15 AOÛT 2021

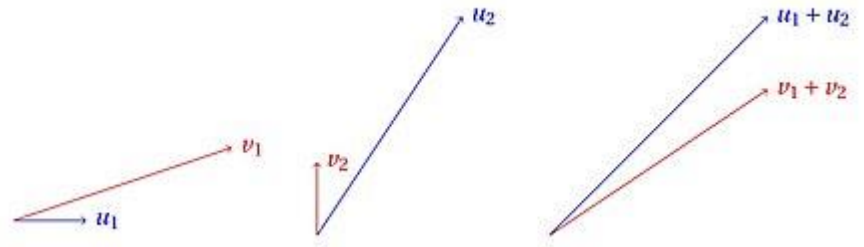
Âge	Cas sévères / 100 000		Efficacité contre les cas sévères
	Non Vaccinés	Vaccinés 2 doses	
12-15	0,30	0	100%
16-19	1,60	0	100%
20-29	1,50	0	100%
30-39	6,20	0,20	96,8%
40-49	16,50	1,00	93,9%
50-59	40,20	2,90	92,8%
60-69	76,60	8,70	88,7%
70-79	190,10	19,80	89,6%
80-89	252,30	47,90	81,1%
90+	510,9	38,60	92,4%

Source : Israeli government data dashboard

Paradoxe de Simpson

Une tendance ou un résultat présent lorsque les données sont en groupes
qui s'inverse ou disparaît lorsque les données sont combinées.





la pente de u_1 est strictement inférieure à celle de v_1
la pente de u_2 est strictement inférieure à celle de v_2

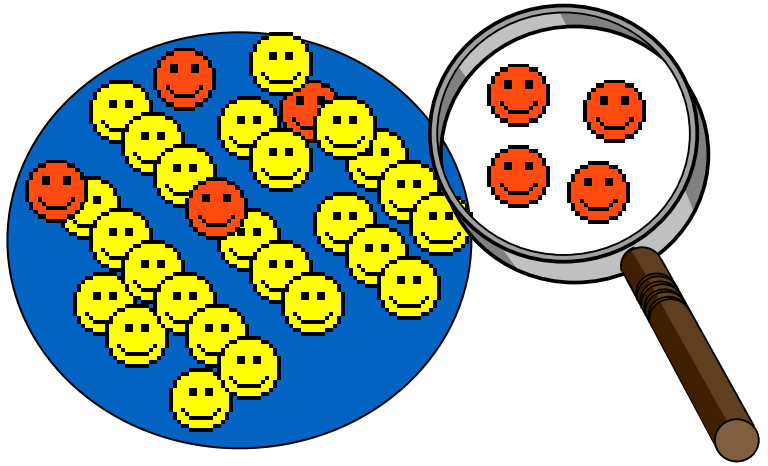
mais la pente de $u_1 + u_2$ est strictement supérieure à celle de $v_1 + v_2$

Enquête en population

Quelle est la prévalence de malades dans la population ?

Enquête exhaustive

- *causes de décès*
- *registre*

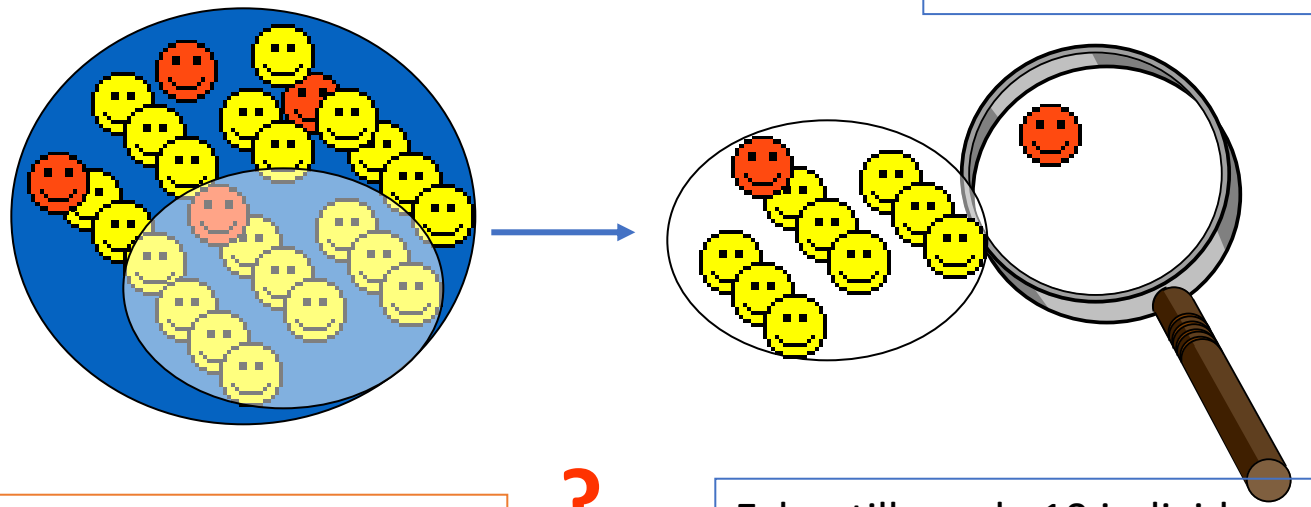


Population de 25 individus
4 individus malades
Prévalence dans la population = 4/ 25

Enquête en population

Quelle est la prévalence de malades dans la population ?

Enquête sur un échantillon de la population



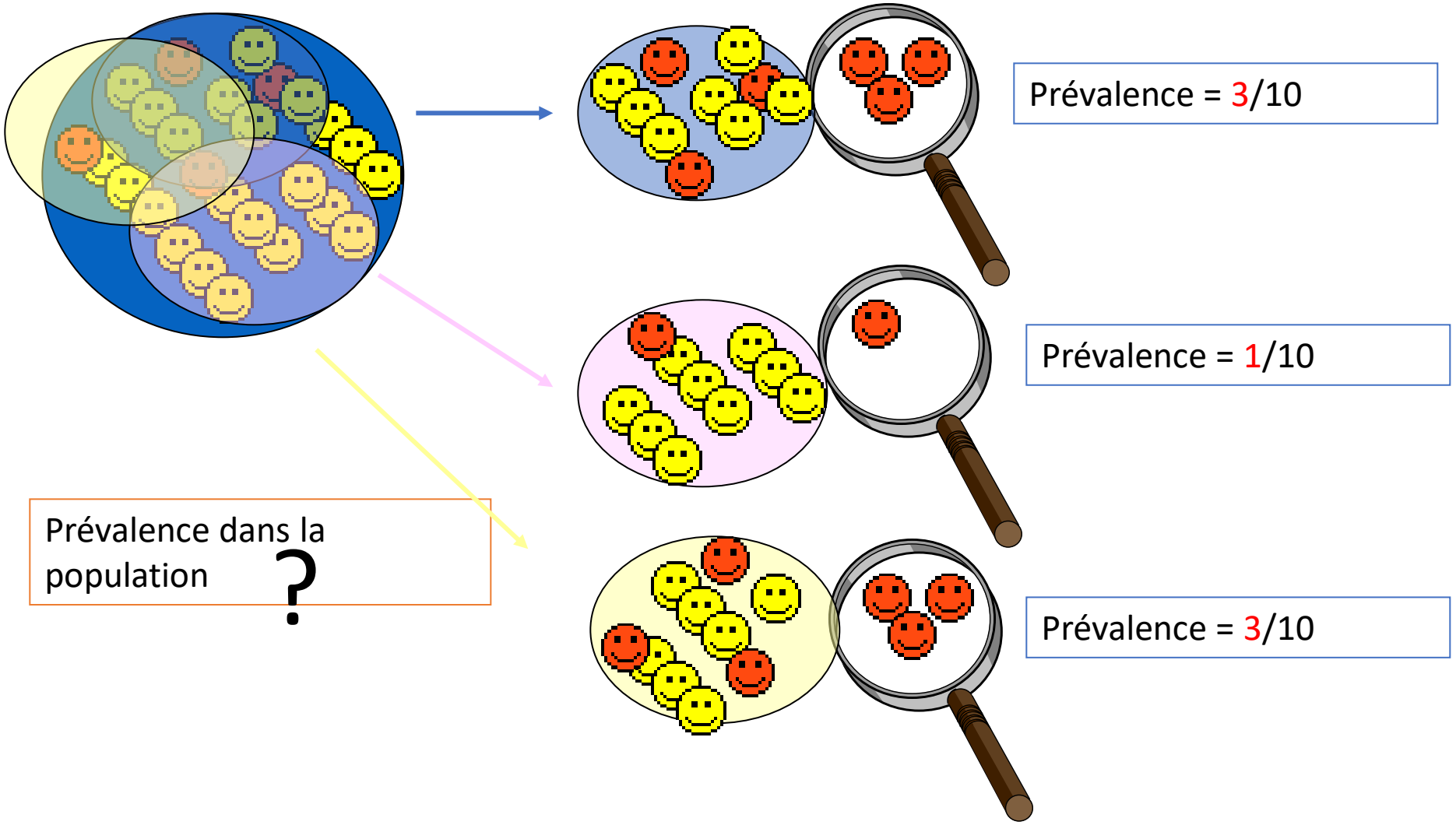
Prévalence dans la population=1/10

?

Echantillons de 10 individus
1 individu malade
Prévalence dans l'échantillon = 1/10

Fluctuation d ’échantillonnage

Quelle est la prévalence de malades dans la population ?



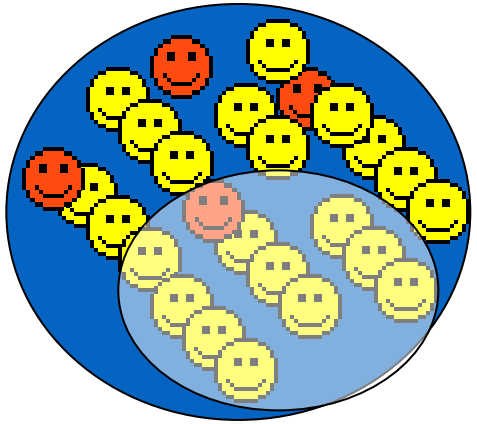
Prévalence = 3/10

Prévalence = 1/10

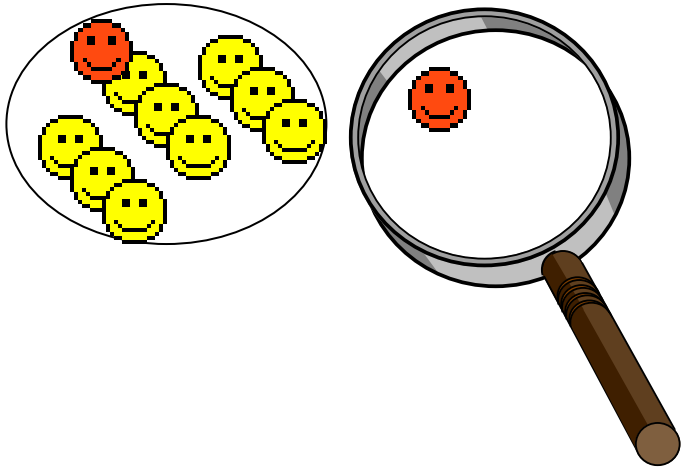
Prévalence = 3/10

Prévalence dans la population ?

Comment définir la prévalence dans la population?



Hasard



Estimation statistique

Prévalence dans la population =

Intervalle de valeurs très probables

Echantillons de 10 individus
1 individu malade
Prévalence dans l'échantillon = 1/10

Voir exposé d'Aurélie

Constituer un échantillon représentatif

Faire intervenir le hasard pour sa constitution

Avoir un *nombre suffisant* d'individus dans l'échantillon pour pouvoir faire de l'inférence statistique

Voir exposé d'Aurélie

Qui est malade ?

Prévalence 10%
Sensibilité 90%
Spécificité 95%

	Malades	Non malades	
Test +	90 (a)	45 (b)	135
Test -	10 (c)	855 (d)	865
	100 (a+c)	900 (b+d)	1000 (N)

- Prévalence est illustrée par le rapport a+c/N.
- Sensibilité = P (T+/M+) = a/a+c = 90%
- Spécificité = P(T- /M-) = d/ d +b = 95%
- Valeur Prédictive Positive = a/a+b
- Valeur prédictive négative = d/c+d

Lecture verticale du tableau

Lecture horizontale du tableau

NOTIONS DE PROBABILITE

$$P(T+) = \frac{n(T+)}{n(T)}$$

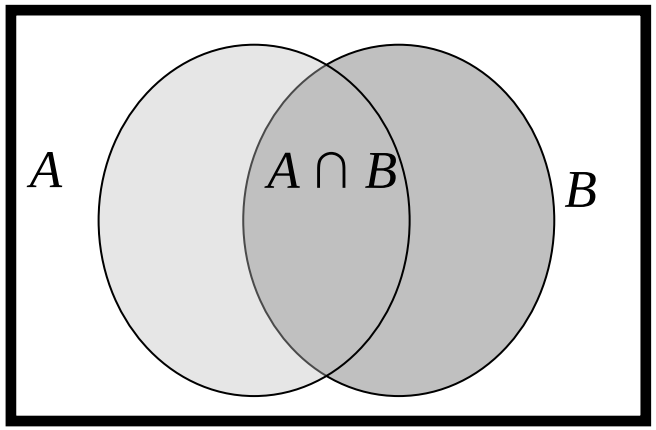
C’est ce qu’on appelle la probabilité fréquentielle.

Si T est grand la probabilité fréquentielle tend à se rapprocher de la probabilité théorique

$$\text{Probabilité FRÉQUENTIELLE} = \frac{\text{Nombre de fois qu’un résultat s’est produit}}{\text{Nombre d’expériences réalisées}}$$

Probabilité conditionnelle

Supposons que nous nous intéressions au calcul de la probabilité de l'événement A et qu'on nous ait dit que l'événement B s'est produit. Alors la probabilité conditionnelle de **A sachant B** est définie comme étant :



$$P\left[A\middle|B\right]=\frac{P\left[A\cap B\right]}{P\left[B\right]} \quad \text{Si} \quad P\left[B\right]\neq 0$$

$$P\left[A\cap B\right]=P\left[A\right]P\left[B\right]$$ **Si** A et B indépendants

Règle de Bayes

$$P[B|A] = \frac{P[B \cap A]}{P[A]}$$

$$P\left[A|B\right] = \frac{P\left[A \cap B\right]}{P\left[B\right]}$$

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

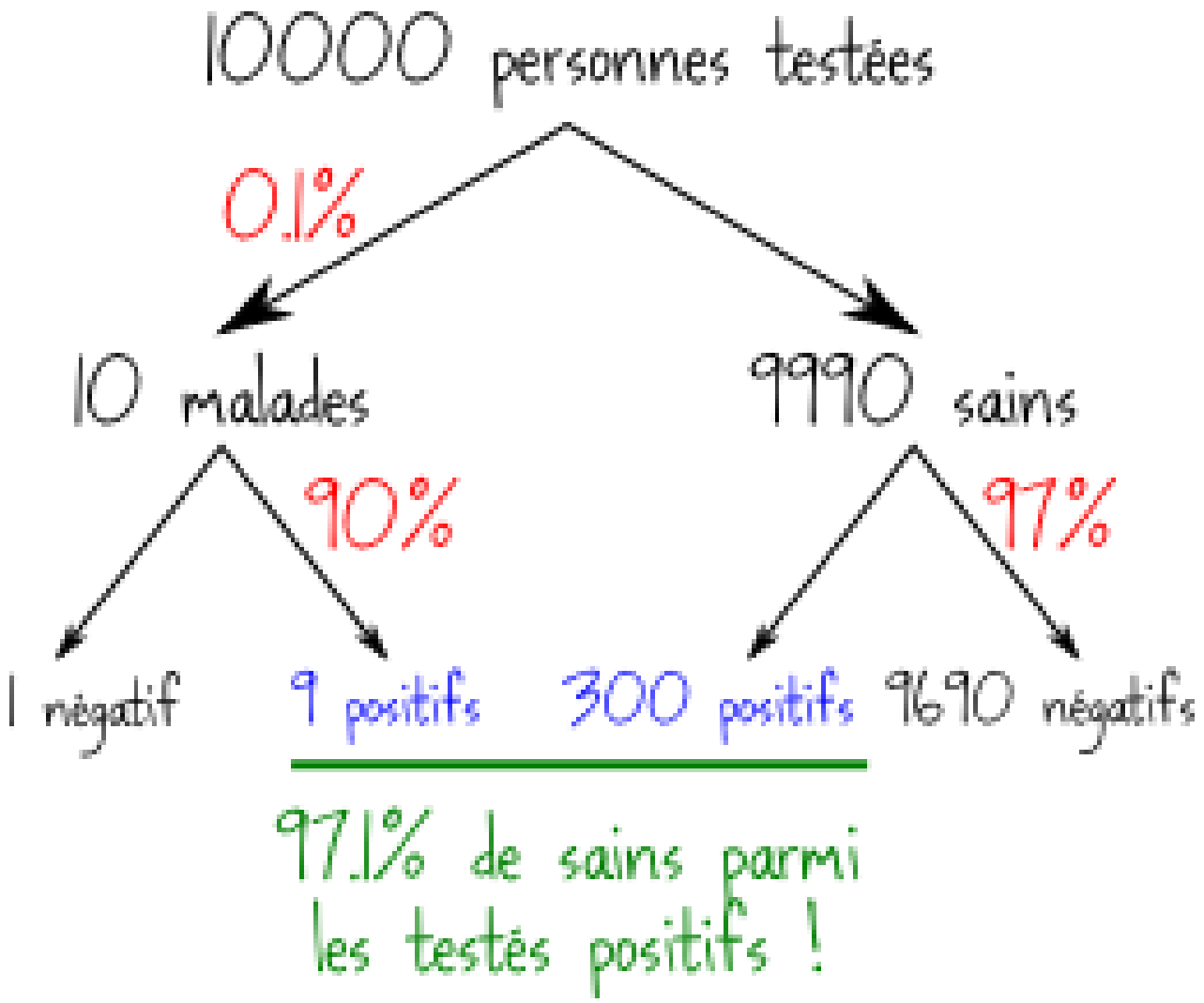
1702-1761

Vous venez de passer un test pour le dépistage d'une maladie. Le médecin vous convoque et vous annonce que le résultat est positif. Ce type de maladie ne touche que 0.1% de la population.

Patient, *fébrile* **-** Le test est-il fiable?

Médecin. - Si vous avez cette maladie, le test sera positif dans 90% des cas ; alors que si vous ne l'avez pas, il sera négatif dans 97% des cas.

- Quelle est la probabilité que vous ayez cette maladie ?


10000 personnes testées
0.1%
10 malades
9990 sains
90%
97%
1 négatif
9 positifs
300 positifs
9690 négatifs
97.1% de sains parmi
les testés positifs !

Test	Malade - M	$\overline{M}$	Total
Positif - P	9	300	309
$\overline{P}$	1	9 690	9 691
Total	10	9 990	10 000

Fréquence en colonne :

Test	Malade - M	$\overline{M}$	Total
Positif - P	90 %	3 %	3 %
$\overline{P}$	10 %	97 %	97 %
Total	100 %	100 %	100 %

Fréquence en ligne :

Test	Malade - M	$\overline{M}$	Total
Positif - P	2,9 %	97,1 %	100 %
$\overline{P}$	0,01 %	99,99 %	100 %
Total	0,1 %	99,9 %	100 %

La condition des sentiments

- Sur les 309 personnes testées positives, 9 sont malades et 300 sont saines (faux positifs).
- Si vous êtes positif, vous n'avez que 2,9% (=9/309) de risque d'être malade et 97,1% (=300/309) de chance d'être un faux positif.
- Pourquoi ce résultat est-il contre-intuitif?

La confusion des sentiments ?

- Soit l'événement M: « *être malade* »
- Soit l'événement P: « *avoir un résultat positif au test* »

➢ Confusion entre « *la probabilité d'être malade sachant que le test est positif* » et « *la probabilité d'être testé positif sachant que l'on est malade* ».

La confusion des sentiments ?

- Si vous êtes testé positif et que vous vous demandez si vous avez cette maladie, vous cherchez la probabilité suivante
  - P(M|T+) : « la probabilité d'être malade sachant que le test est positif »

- Si le médecin vous dit que si vous avez cette maladie, le test sera positif dans 90% des cas, vous cherchez cette probabilité
  - P(T+|M) : « la probabilité d'être testé positif sachant que l'on est malade »

Comment passer de $P(M|T+)$ à $P(T+|M)$?

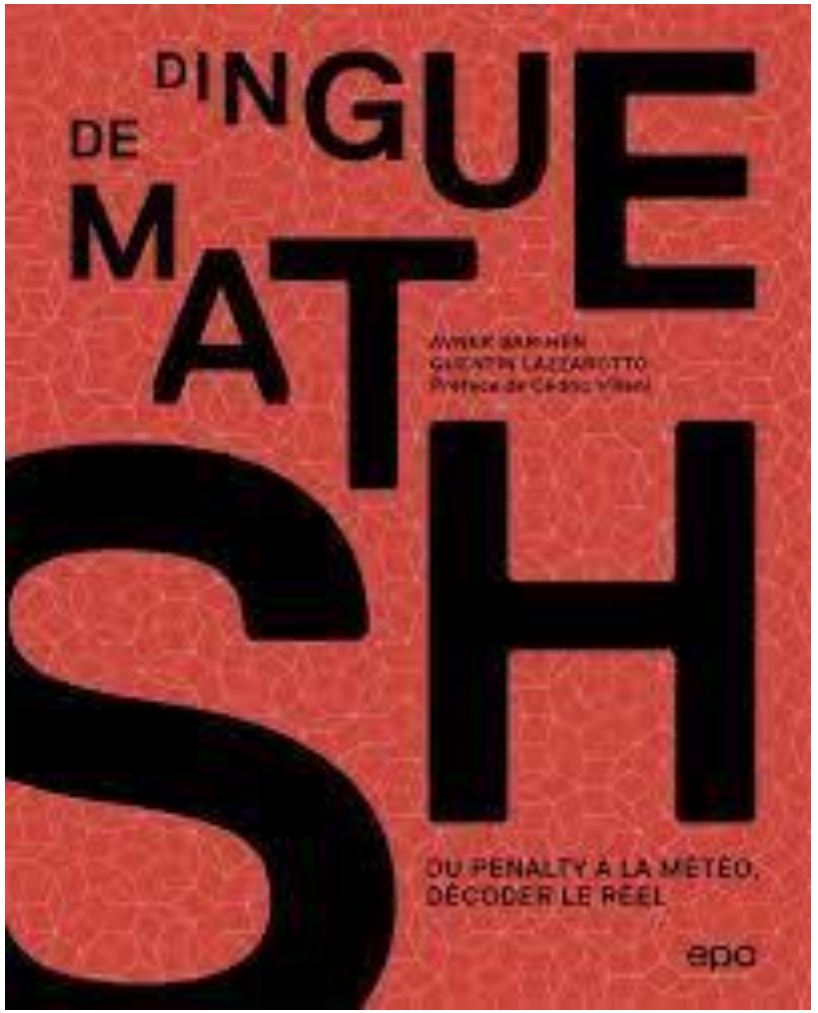
- $P(M|T+) = \frac{P(T+|M)*p(M)}{p(T+)}$
- $P(M|T+) = \frac{0{,}90*0{,}001}{0{,}0309}$
- $P(M|T+) = 0{,}029 = 2{,}9\%$

- La probabilité d'être malade sachant que le test est positif
  - $P(M|T+)$= 2,9 %
- La probabilité d'être testé positif sachant que l'on est malade
  - $P(T+|M)$ = 90%

Conclusions

- Les statistiques n’empêchent pas de réfléchir
- Un nombre n’a d’intérêt que si il est replacé dans un contexte
- C’est facile de mentir avec des statistiques MAIS c’est encore plus simple de mentir sans statistiques
- Ne pas confondre critique et complotisme

Dingue de maths
Du pénalty à la météo, décoder le réel
Avner BAR-HEN & Quentin LAZZAROTTO



Merci de votre attention