F - Courbes ROC et optimisation des tests

Comme on l'a vu, les tests sont essentiellement construits à partir de la loi d'une statistique de test sous H_0, et l'on utilise au choix la p-valeur ou un quantile de cette loi sous H_0 : jamais H_1 n'est vraiment considérée, à part pour latéraliser un test.
(On a vu aussi que la p-valeur est un objet « conflictuel », bref on peut émettre des réserves sur l'ensemble de la procédure des tests !)

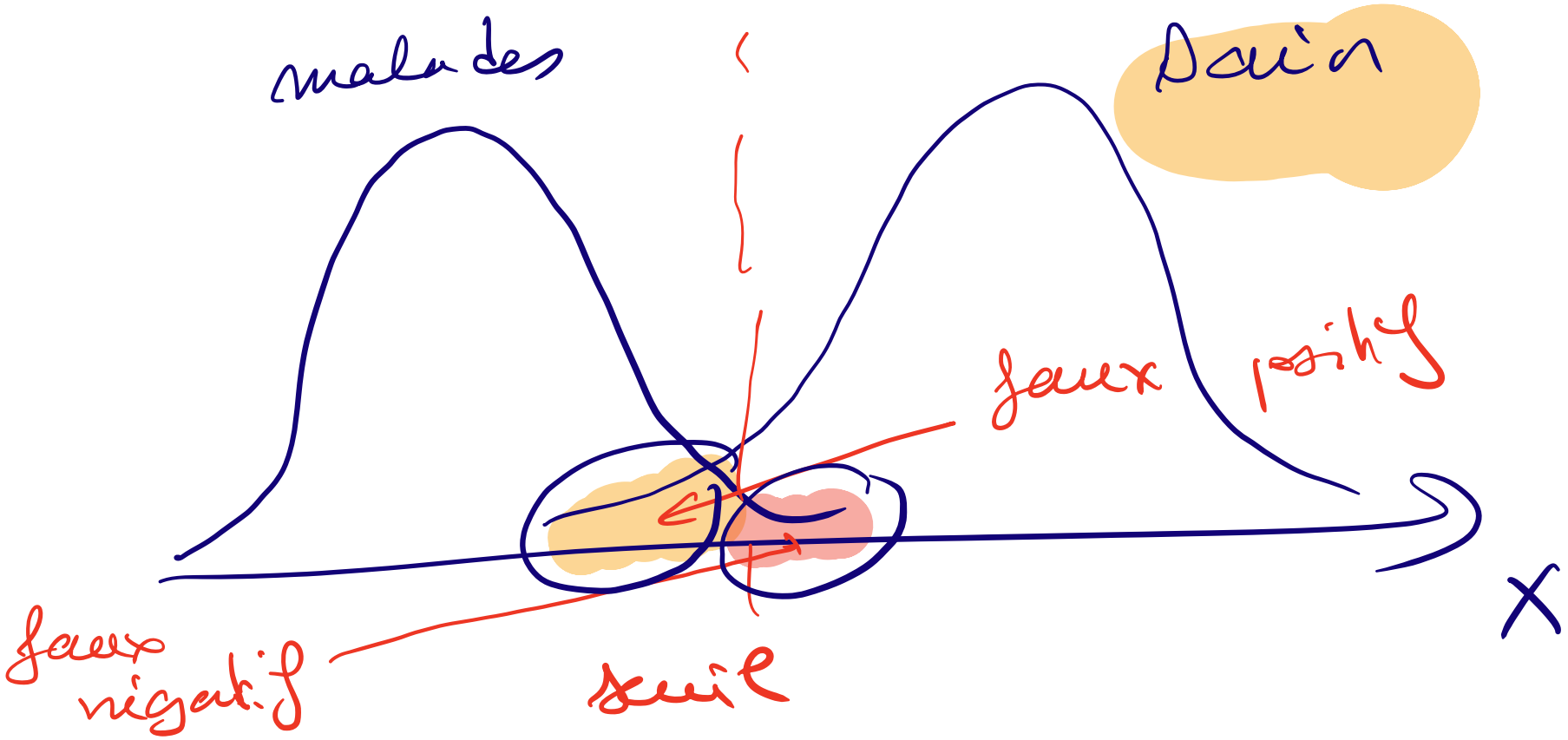
F - Courbes ROC et optimisation des tests

Comme on l'a vu, les tests sont essentiellement construits à partir de la loi d'une statistique de test sous H_0, et l'on utilise au choix la p-valeur ou un quantile de cette loi sous H_0 : jamais H_1 n'est vraiment considérée, à part pour latéraliser un test.
(On a vu aussi que la p-valeur est un objet « conflictuel », bref on peut émettre des réserves sur l'ensemble de la procédure des tests !)

La courbe ROC est une façon de choisir intelligemment un seuil de test **individuel** : dans le cas où, naturellement, la question est unilatérale, au sens où l'on suppose *a priori* que les deux échantillons ont des valeurs de la mesure observée qui se séparent naturellement, par exemple prenons le cas de personnes porteuses ou pas d'une maladie, et dont on mesure une quantité de `substance_X` dans le sang qui est *plutôt* :

- ➜ *petite* pour le premier échantillon (les personnes non porteuses par exemple) ;
- ➜ *grande* pour le second échantillon (les personnes porteuses).

La construction des tests, par exemple dans un cadre gaussien, permet d'affirmer que les deux populations de personnes malades et de personnes saines ont bien des moyennes différentes, mais que faire si on voit arriver une personne avec sa mesure, comment va-t-on dire si elle est malade ?



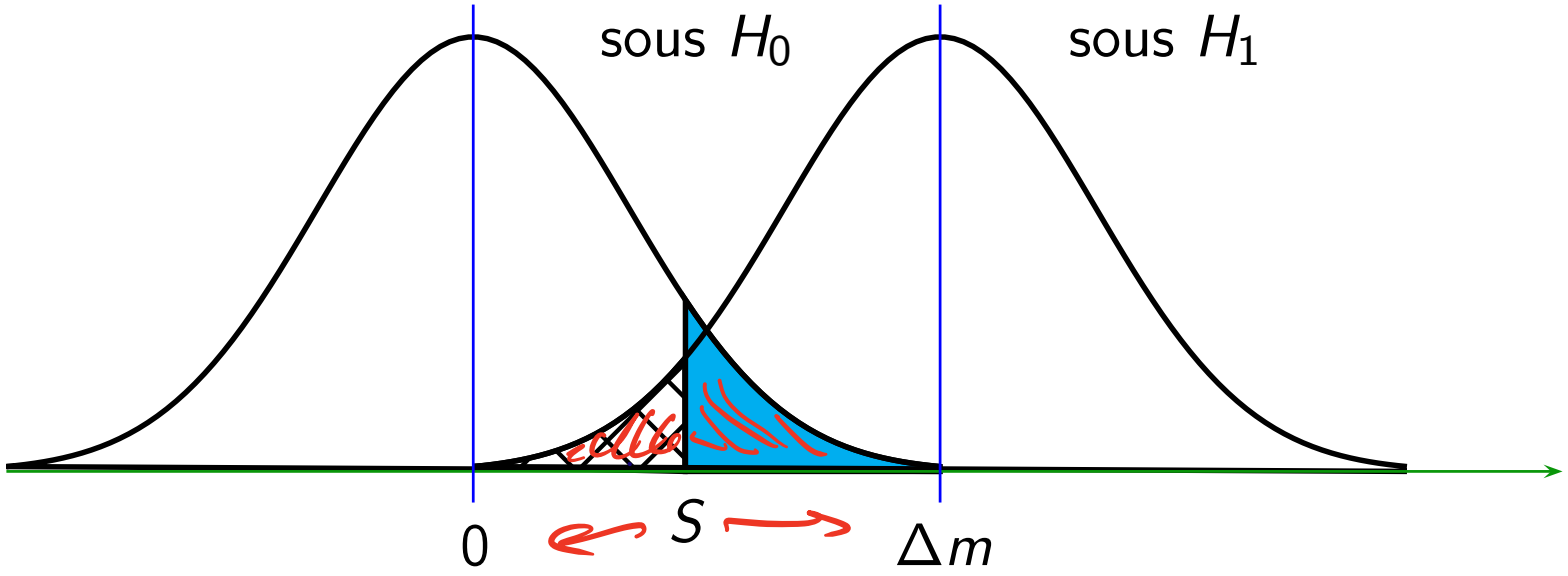
La construction des tests, par exemple dans un cadre gaussien, permet d'affirmer que les deux populations de personnes malades et de personnes saines ont bien des moyennes différentes, mais que faire si on voit arriver une personne avec sa mesure, comment va-t-on dire si elle est malade ?

Alors il est naturel de se proposer de fixer un seuil pour déclarer qu'au delà d'une certaine valeur, la personne sera déclarée porteuse de la maladie : si l'on note x la quantité, et S le seuil, alors en faisant cela on fera des erreurs :

- ➜ il y aura des **faux positifs** ;
- ➜ et des **faux négatifs**.

→ minimisation globale de ces deux types d'erreurs.

Regardons les deux erreurs que l'on va faire si l'on effectue notre test avec une décision au point indiqué (à gauche on ne rejette pas l'hypothèse H_0 (« être sain »), et à droite on la rejette) :



on voit bien que si l'on décale le seuil S de décision à droite on va réduire l'erreur de première espèce et augmenter celle de deuxième espèce, dans la pratique cela va créer/supprimer des « faux positifs » et des « faux négatifs »

ces cas sont résumés dans un tableau, lorsque l'on effectue le test sur un nombre n de cas, on obtiendra

	H_0	H_1
Rejet	*faux positif* $\mathbf{P}_{H_0}$ (rejet)	*vrai positif* $\mathbf{P}_{H_1}$ (rejet)
Non rejet	*vrai négatif* $\mathbf{P}_{H_0}$ (non rejet)	*faux négatif* $\mathbf{P}_{H_1}$ (non rejet)

On introduit alors :

La **spécificité** : proportion de vrais négatifs chez les non-malades ;

La **sensibilité** : la proportion de vrais positifs parmi les malades.

On voit donc que la spécificité est associée à l'erreur de première espèce, en ce sens que si l'on avait un grand nombre de personnes testées, on aurait

$$\textbf{spécificité} = \textbf{Sp} \simeq 1 - \textbf{erreur de première espèce},$$

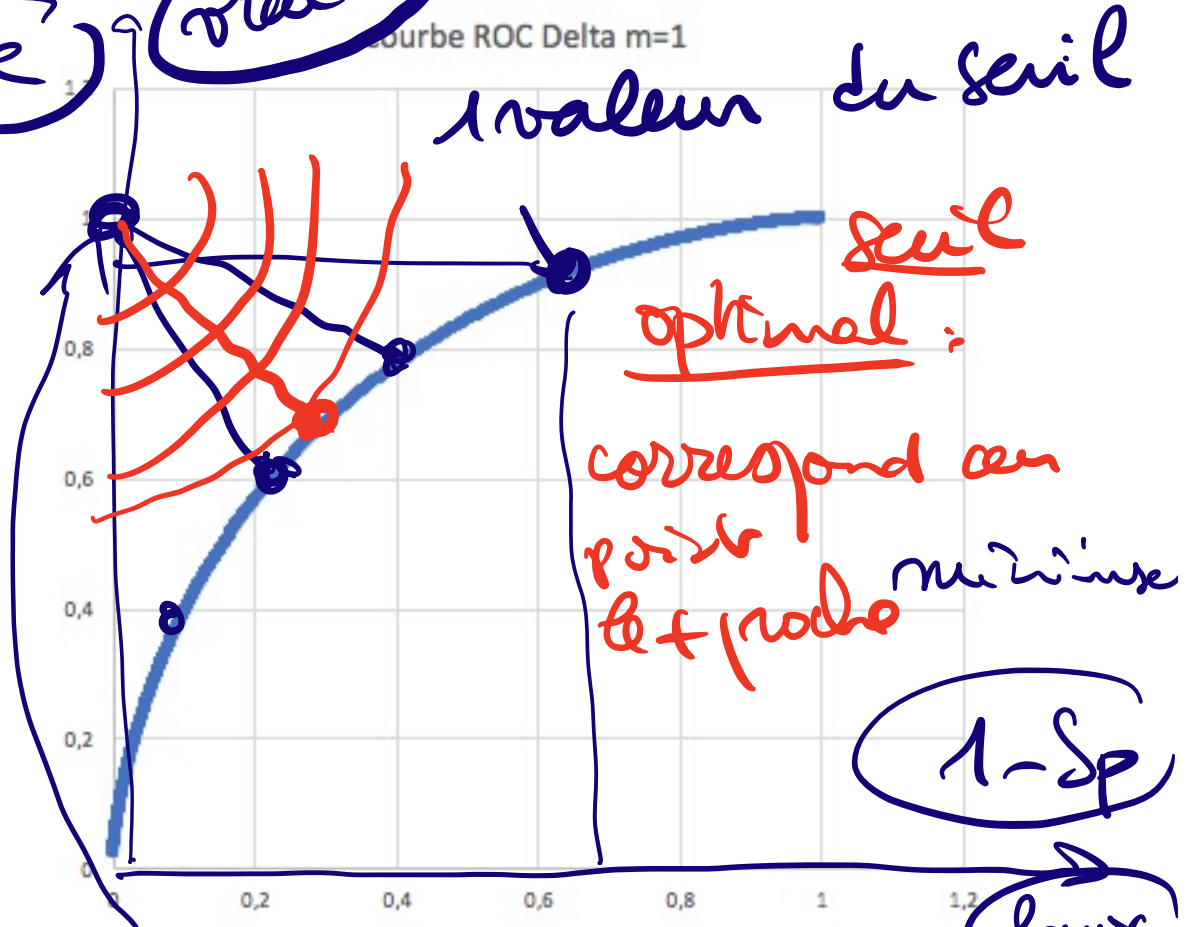
et la sensibilité est aussi reliée à l'erreur de seconde espèce :

$$\textbf{sensibilité} = \textbf{Se} \simeq 1 - \textbf{erreur de deuxième espèce}.$$

On souhaite donc maximiser ces deux quantités. . . la procédure utilisée jusqu'à présent a été de fixer un seuil sur l'erreur de première espèce, mais maintenant on souhaite une optimisation plus globale sur les deux types d'errreur.

On souhaite donc maximiser ces deux quantités... la procédure utilisée jusqu'à présent a été de fixer un seuil sur l'erreur de première espèce, mais maintenant on souhaite une optimisation plus globale sur les deux types d'erreur.

Pour proposer une solution à cette question, on peut proposer la courbe ROC (Receiver Operating Characteristic). Pour cela on trace la courbe composée des points de coordonnées $(1 - \mathbf{Sp}(S), \mathbf{Se}(S))$.

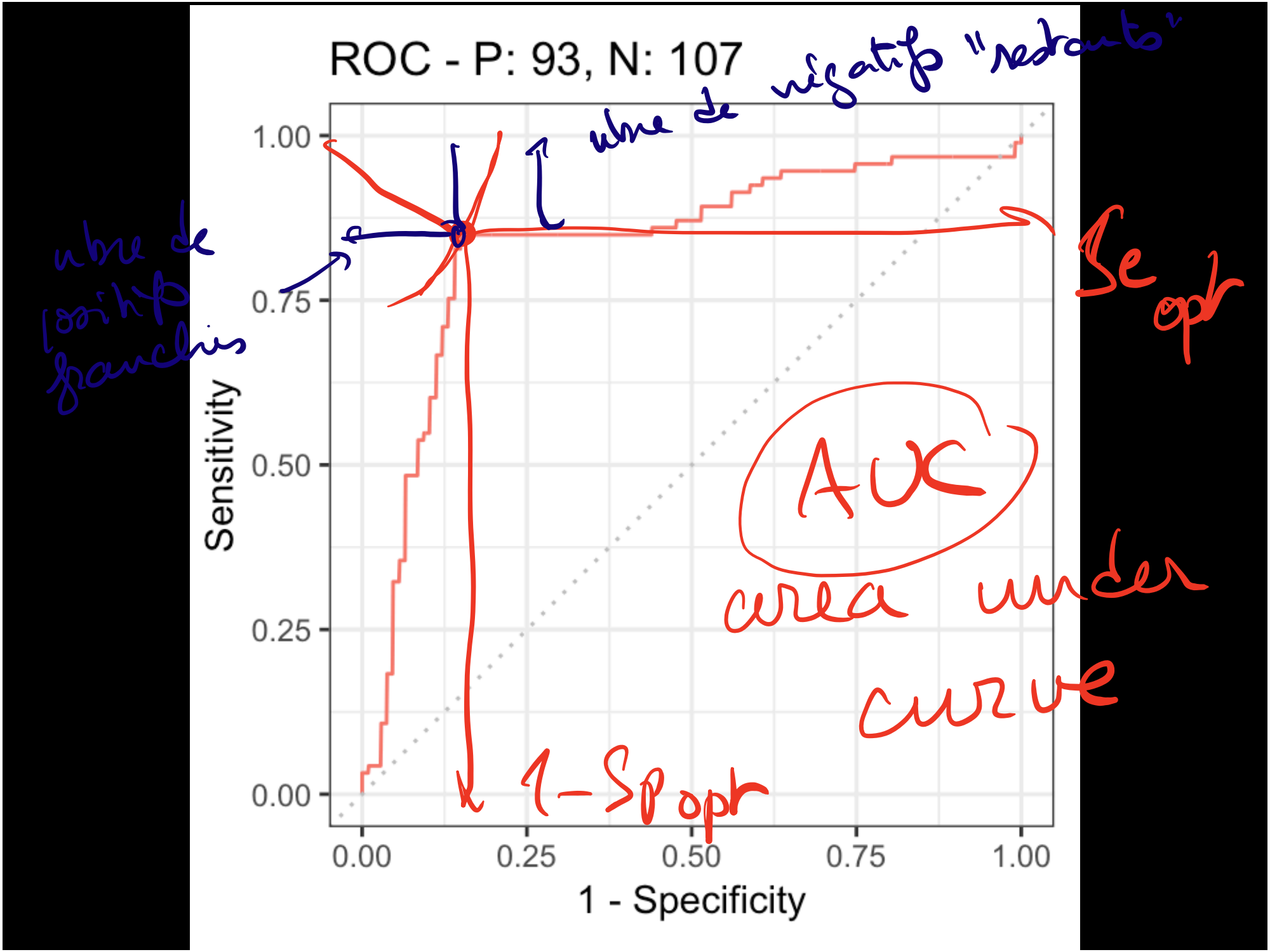


On souhaite maximiser la spécificité et la sensibilité, une première approche serait par exemple de prendre le seuil S correspondant au point de la courbe ROC le plus proche du point « parfait » $(0, 1)$.

En fait cette question est facile (c'est-à-dire uniquement calculatoire) si l'on connait effectivement les lois sous l'hypothèse H_0 et H_1 de notre statistique de test, mais en général ce n'est pas le cas, et on est plutôt confronté à une situation du genre :

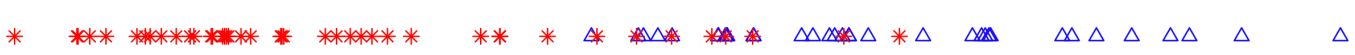
- on connait parfaitement une population avec des personnes malades et des personnes saines (situation validée par des indicateurs fiables à 100%), respectivement n_1 individus sains et n_2 individus malades ;
- on mesure des quantités numériques $x_1^s, \ldots, x_{n_1}^s$ et $x_1^m, \ldots, x_{n_2}^m$ sur ces personnes ;
- on établit un tableau de test en comptant les faux positifs etc, pour différentes valeurs du seuil S ;
- et on établit une approximation de la courbe ROC en positionnant les points approchant $1 - \mathbf{Sp}(S)$ et $\mathbf{Se}(S)$ par les taux de faux positifs et vrais négatifs respectivement.

On obtient ainsi une courbe ROC « en escaliers » sur les deux échantillons qui permet de proposer aussi un seuil optimal.


ROC - P: 93, N: 107
Sensitivity
1 - Specificity
0.00
0.25
0.50
0.75
1.00
nbre de négatifs "restants"
nbre de positifs manqués
Se opt
1 - Sp opt
AUC
area under curve

Comment tracer efficacement cette courbe

Imaginons qu'on a notre nuage de points de la quantité x coloré selon que les individus sont malades ou pas :



et l'on va faire partir S de $-\infty$ et le faire aller à $+\infty$:

Comment tracer efficacement cette courbe

Imaginons qu'on a notre nuage de points de la quantité x coloré selon que les individus sont malades ou pas :

et l'on va faire partir S de $-\infty$ et le faire aller à $+\infty$:

➜ au début, tout le monde est considéré comme malade : $\mathbf{Sp}(-\infty) = 0\%$ et $\mathbf{Se}(-\infty) = 100\%$: le point de la courbe ROC est donc le point $(1, 1)$;

Comment tracer efficacement cette courbe

Imaginons qu'on a notre nuage de points de la quantité x coloré selon que les individus sont malades ou pas :

et l'on va faire partir S de $-\infty$ et le faire aller à $+\infty$:

- ➜ au début, tout le monde est considéré comme malade : $\mathbf{Sp}(-\infty) = 0\%$ et $\mathbf{Se}(-\infty) = 100\%$: le point de la courbe ROC est donc le point $(1, 1)$;
- ➜ si S « franchit » une personne *saine*, la spécificité augmente de $1/n_1$ et la sensibilité ne bouge pas : cela correspond à tracer un segment de longueur $1/n_1$ vers la gauche ;

Comment tracer efficacement cette courbe

Imaginons qu'on a notre nuage de points de la quantité x coloré selon que les individus sont malades ou pas :

et l'on va faire partir S de $-\infty$ et le faire aller à $+\infty$:

- ➜ au début, tout le monde est considéré comme malade : $\mathbf{Sp}(-\infty) = 0\%$ et $\mathbf{Se}(-\infty) = 100\%$: le point de la courbe ROC est donc le point $(1, 1)$;
- ➜ si S « franchit » une personne *saine*, la spécificité augmente de $1/n_1$ et la sensibilité ne bouge pas : cela correspond à tracer un segment de longueur $1/n_1$ vers la gauche ;
- ➜ si S « franchit » une personne *malade*, la spécificité ne bouge pas et la sensibilité diminue de $1/n_1$: cela correspond à tracer un segment de longueur $1/n_2$ vers le bas ;

Comment tracer efficacement cette courbe

Imaginons qu'on a notre nuage de points de la quantité x coloré selon que les individus sont malades ou pas :

et l'on va faire partir S de $-\infty$ et le faire aller à $+\infty$:

- ➜ au début, tout le monde est considéré comme malade : $\mathbf{Sp}(-\infty) = 0\%$ et $\mathbf{Se}(-\infty) = 100\%$: le point de la courbe ROC est donc le point $(1, 1)$;
- ➜ si S « franchit » une personne *saine*, la spécificité augmente de $1/n_1$ et la sensibilité ne bouge pas : cela correspond à tracer un segment de longueur $1/n_1$ vers la gauche ;
- ➜ si S « franchit » une personne *malade*, la spécificité ne bouge pas et la sensibilité diminue de $1/n_1$: cela correspond à tracer un segment de longueur $1/n_2$ vers le bas ;
- ➜ à la fin, tout le monde est considéré comme sain : $\mathbf{Sp}(+\infty) = 100\%$ et $\mathbf{Se}(+\infty) = 0\%$: le point de la courbe ROC est donc le point $(0, 0)$.

Un package parmi d'autres

Le package `PRROC` permet de le faire simplement, en ayant deux vecteurs de valeurs $x_1 =$ `sains` et $x_2 =$ `malades`, si le test est de classer à gauche du seuil les personnes saines et à droite les personnes malades, alors on procède ainsi :

```
> courbe<-roc.curve(malades,sains,curve=TRUE)
> plot(courbe)
```