

Du génome au protéome ...

## Méthodologie d'évaluation



Irène Buvat  
U678 INSERM, Paris

[buvat@imed.jussieu.fr](mailto:buvat@imed.jussieu.fr)

25 février 2005

L'évaluation est un travail difficile, mais pourtant indispensable, auquel on est confronté dans toutes les disciplines scientifiques. L'objectif de cet exposé est :

- d'une part, de vous présenter, un cadre méthodologique simple permettant de bien formuler un problème d'évaluation et,
- d'autre part, de vous donner quelques éclairages sur les outils qui existent pour mener à bien un travail d'évaluation de façon objective et rigoureuse.

## Objectif de l'évaluation

Déterminer si le biomarqueur X  
est un indicateur pertinent concernant la présence d'une pathologie  
chez un groupe de sujets Y

Quoi évaluer ?

Dans quel but ?

Sur quelles données ?

Typiquement, dans votre contexte, il est nécessaire de faire appel à des méthodes d'évaluation pour répondre à des questions du type :

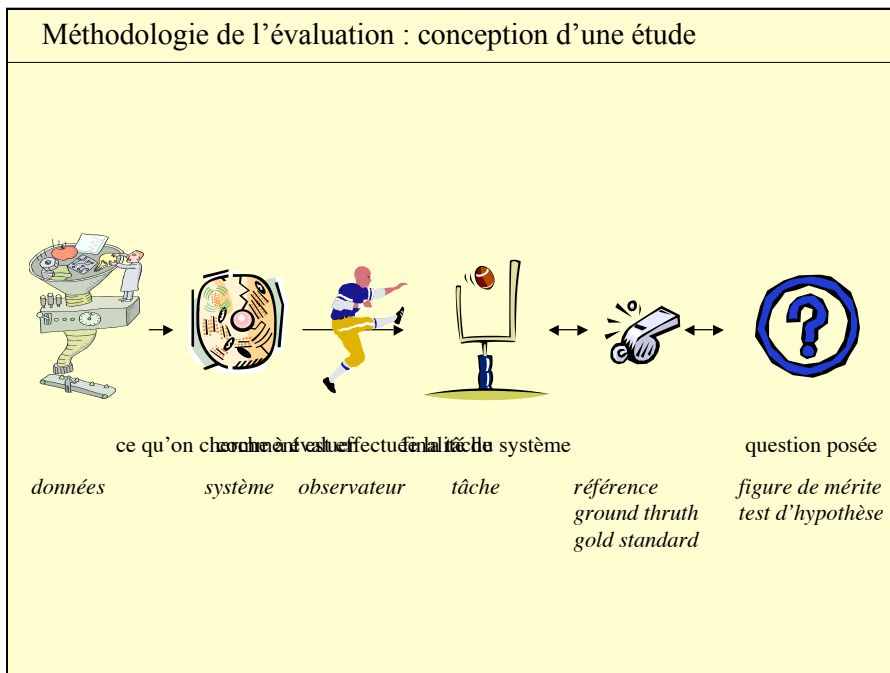
« tel biomarqueur est-il un indicateur pertinent concernant la probabilité de présence d'une pathologie chez tel groupe de sujets ? ».

Cette seule formulation révèle déjà 3 éléments qui doivent être clairement défini dans un travail d'évaluation :

- ce qu'on cherche à évaluer
- la question à laquelle on cherche à répondre
- et le contexte dans lequel on souhaite répondre à cette question.

Ces éléments peuvent être clairement précisés en formalisant relativement simplement le problème d'évaluation

## Méthodologie de l'évaluation : conception d'une étude



Il faut donc commencer par définir ce qu'on veut évaluer. Typiquement, il peut s'agir d'un biomarqueur, d'une combinaison de biomarqueurs, ou encore la façon de mesurer le taux d'un biomarqueur par exemple. On peut désigner ce qu'on cherche à évaluer sous le nom générique du système.

La deuxième notion qu'il est indispensable de préciser est la finalité du système, qu'on appelle la tâche. On peut distinguer 2 types de tâches : d'abord, les tâches de classification pour lesquelles la réponse est l'appartenance à une catégorie, parmi un petit nombre de catégories, par exemple bénin-malin. D'autre part, les tâches d'estimation qui visent à estimer un paramètre sur une échelle continue, comme un taux de biomarqueur. Il est fondamental de savoir définir le type de tâche au centre du problème d'évaluation, car les méthodes d'évaluation à mettre en œuvre vont totalement dépendre de la tâche.

Une 3ème notion à identifier est comment la tâche est effectuée : l'entité qui va effectuer la tâche est appelée, de façon générique, l'observateur.

Il faut aussi préciser sur quelles données va être réalisé le travail d'évaluation. De ces données dépendra la portée des conclusions qui pourront être tirées.

Une autre notion fondamentale en évaluation est celle de la référence, encore appelée « gold standard » ou « ground truth ». Ce qu'on appelle ainsi, c'est la vraie valeur du paramètre que l'on cherche à évaluer par exemple, ou la catégorie réelle dans laquelle l'échantillon devrait être classé dans les tâches de classification.

Enfin, l'objectif de tout travail d'évaluation est de répondre à une question. Cette question peut appeler une réponse chiffrée, ou figure de mérite, ou se formuler comme un test d'hypothèse, qui appelle une réponse binaire : l'hypothèse est acceptée ou rejetée. En fait, les tests d'hypothèses sous-tendent systématiquement une figure de mérite. C'est donc davantage l'usage que l'on fait de la figure de mérite qui distingue ces deux cas. Soit on l'interprète tel quel, soit on l'utilise pour tester une hypothèse appelant une réponse binaire.

Conception d'une étude : choix de l'outil d'évaluation

Quel type de tâche ?



Classification  
ou  
Estimation

La composante principale qui permet d'orienter le choix de la méthode d'évaluation à adopter est la tâche : tâche de classification ou tâche d'estimation.

Nous allons maintenant voir les approches adaptées pour ces différents types de tâches.

## Outils pour les tâches de classification

sensibilité - spécificité - exactitude - valeurs prédictives -  
rapports de vraisemblance

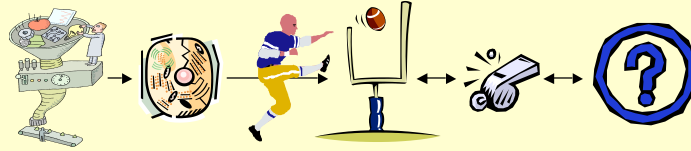


approche ROC



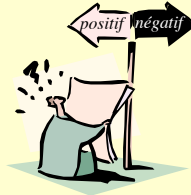
Pour les tâches de classification, une approche souvent utilisée consiste à faire des calculs de sensibilité et de spécificité de détection, ou d'index similaires. Je vais expliquer dans quel cas ces approches sont appropriées, et aussi comment on peut, dans certains cas, aller au delà de cette approche, en utilisant la méthodologie ROC.

## Tâches de classification : contexte général



Données divisibles en 2 (ou +) catégories *e.g., avec et sans anomalie*  
(positif ou négatif)  
*anomalie A versus anomalie B*

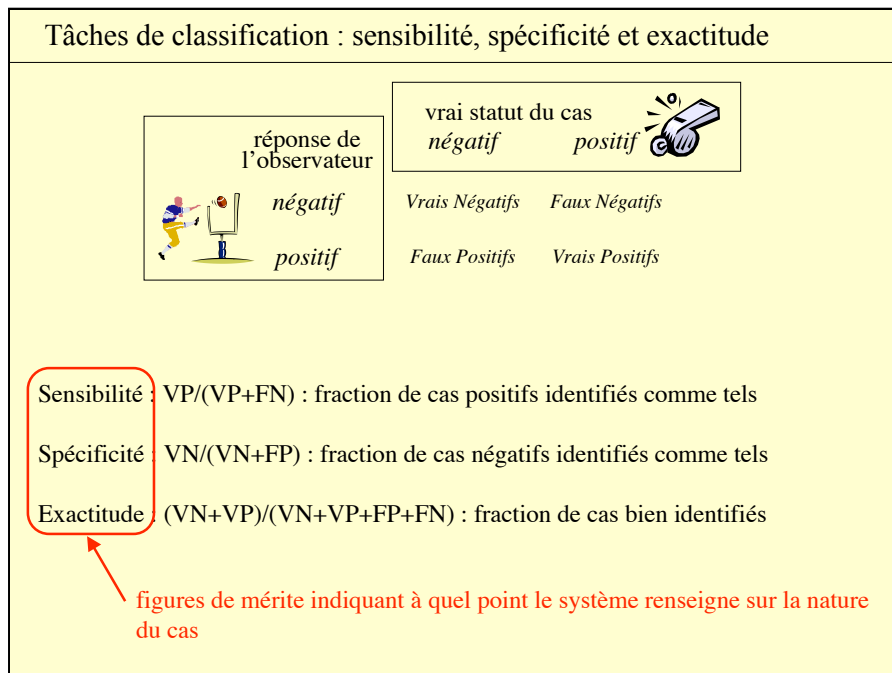
Tâche : déterminer la catégorie à laquelle appartient chaque élément du jeu de données



Tout d'abord, précisons le contexte des tâches de classification, dans le cas de 2 classes pour simplifier.

Les données peuvent être divisées en 2 catégories : par exemple, certains éléments comportent une anomalie, d'autres pas.

La tâche assignée à l'observateur est donc de déterminer, pour chaque élément, à laquelle des 2 catégories il appartient, c'est-à-dire s'il contient ou non l'anomalie. C'est donc une décision binaire.



L'analyse des réponses des observateurs va se faire en triant les réponses négatives et positives en fonction de la vraie réponse, c'est à dire la référence. Ceci conduit à séparer 4 cas de figures :

- les cas positifs et identifiés comme tels : on parle de vrais positifs.
- les cas négatifs et identifiés comme tels : ce sont les vrais négatifs.
- les cas positifs classés comme négatifs : ce sont les faux négatifs.
- enfin, les cas négatifs classés comme positifs : ce sont les faux positifs.

Les effectifs dans ces 4 catégories permettent de calculer des index caractérisant la justesse de la classification, et en particulier :

- la sensibilité, définie comme le pourcentage de cas positifs correctement identifiés.
- et la spécificité, définie comme le pourcentage de cas négatifs correctement identifiés.

On utilise parfois aussi l'exactitude, c'est-à-dire le pourcentage de cas, positifs ou négatifs, correctement identifiés.

A priori, ces grandeurs, qui sont des figures de mérite, pourraient être suffisantes pour évaluer une méthode de classification en 2 catégories. En fait, elles présentent très vite des limites, comme l'illustrent les exemples suivants.

## Pourquoi cette approche est insuffisante ? Deux exemples



Comment évaluer le compromis sensibilité/spécificité ?

*e.g. :* biomarqueur A : sensibilité=80% spécificité=65%

biomarqueur B : sensibilité=90% spécificité=50%

**Quel est le meilleur ?**

Insuffisance de l'exactitude pour répondre :

*biomarqueur A : sensibilité=70%, spécificité=90%, exactitude=87%*

*biomarqueur B : sensibilité=40%, spécificité=96%, exactitude=87%*

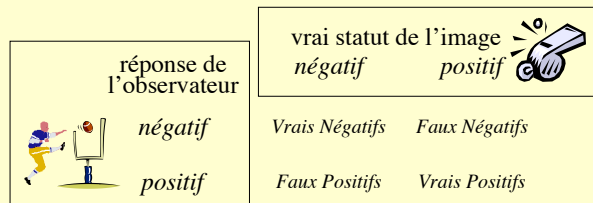
**Une même exactitude peut correspondre à des compromis sensibilité/spécificité très différents**

La première limite est que pour évaluer les performances de la méthode de classification, il faut généralement prendre en compte à la fois sa sensibilité et sa spécificité. Or, on arrive souvent à des situations du type de celle-ci. Quelle est alors dans ce cas le meilleur des 2 biomarqueurs ?

On pourrait penser que l'exactitude apporte une réponse. Cependant, cet exemple montre que ce n'est pas le cas. En effet, une même exactitude peut correspondre à des compromis sensibilité/spécificité bien différents, et la considération de l'exactitude sans considérer les couples sensibilité et spécificité peut conduire à des conclusions erronées.



## Tâches de classification : valeurs prédictives



Valeur prédictive positive :  $VP/(VP+FP)$  : fraction de cas identifiés comme positifs qui sont effectivement positifs  
Valeur prédictive négative :  $VN/(VN+FN)$  : fraction de cas identifiés comme négatifs qui sont effectivement négatifs

Dépendent non seulement de la justesse du système, mais aussi de la prévalence

Mesurent la valeur « clinique » du système

Une autre façon de mesurer les performances d'un système est de déterminer la qualité des prédictions issues du système.

Ceci est généralement réalisé en calculant deux figures de mérite :

- la valeur prédictive positive, qui représente la fraction de cas réellement positifs parmi tous les cas identifiés comme positifs par le test.
- de façon similaire, la valeur prédictive négative, qui représente la fraction de cas réellement négatifs parmi tous les cas identifiés comme négatifs par le test.

Contrairement aux sensibilité et spécificité, les valeurs prédictives dépendent de la prévalence de la pathologie. Elles mesurent donc essentiellement l'utilité diagnostique d'un système, mais pas sa justesse intrinsèque.

Dans la plupart des cas, on cite en fait les valeurs prédictives en plus des valeurs de sensibilité et spécificité.

## Tâches de classification : rapports de vraisemblance

Rapport de vraisemblance positif (RV+) : VP/FP :

probabilité d'une observation positive chez les cas réellement positifs  
par rapport à la probabilité d'une observation positive chez les cas  
réellement négatifs

Rapport de vraisemblance négatif (RV-) : FN/VN :

probabilité d'une observation négative chez les cas réellement positifs  
par rapport à la probabilité d'une observation négative chez les cas  
réellement négatifs

Système idéal :  $RV+ = +\infty$  et  $RV- = 0$

Système non informatif :  $RV+ = 1$  et  $RV- = 1$



Mesurent le gain informatif apporté par le système sur la  
probabilité de présence de la pathologie (prédiction)

risque d'avoir la pathologie après le test =  
risque d'avoir la pathologie avant le test  $\times$  RV+

... mais ne dépendent pas de la prévalence

Une dernière façon de mesurer les performances d'un système est d'évaluer l'information apportée par le système concernant la nature du cas, par rapport à l'information dont on dispose avant d'utiliser le système.

Les figures de mérite permettant de faire cela sont des rapports de vraisemblance :

- le rapport de vraisemblance positif représente la probabilité d'observer un résultat positif chez les cas positifs divisé par la probabilité d'observer un résultat positif chez les cas négatifs. Idéalement, ce rapport de vraisemblance doit être infini.

- de façon similaire, le rapport de vraisemblance négatif représente la probabilité d'observer un résultat négatif chez les cas positifs par rapport à la probabilité d'observer un résultat négatif chez un cas négatif. Idéalement, ce rapport de vraisemblance doit être zéro.

Ces figures de mérite mesurent l'information apportée par le système, sans dépendre de la prévalence de la maladie.

## Domaines d'applications



Si le système produit un résultat binaire, les 3 approches présentées précédemment sont les seules possibles avec :



- 1) la possibilité de tester si les différences entre figures de mérite sont significatives
- 2) la possibilité d'étudier l'impact de certains facteurs externes sur les résultats (modélisation par régression)
- 3) la possibilité de combiner les résultats de plusieurs tests binaires et de caractériser la pertinence de la combinaison (e.g., régression logistique)

... mais souvent, l'affectation binaire est faite à partir d'une observable continue donnée par le système

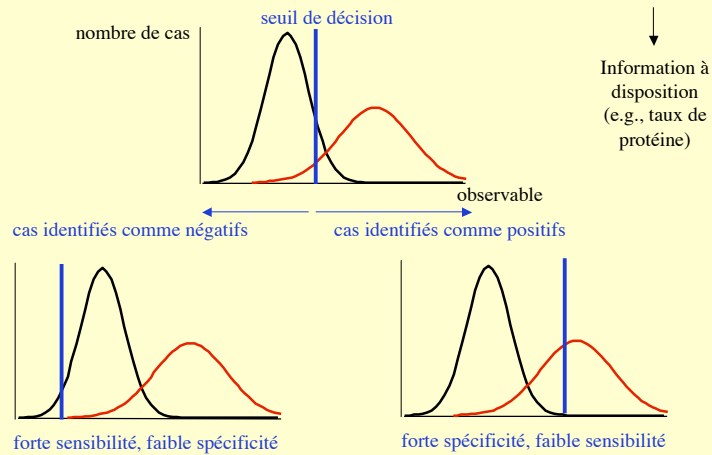
Si le système produit une observable binaire, les 3 approches que je viens de présenter sont les seules possibles.

Il existe des procédures permettant de comparer statistiquement des systèmes à partir de ces figures de mérite, et également de déterminer l'impact de certains facteurs sur les valeurs de ces figures de mérite par des approches de régression. Avec ces approches, on peut aussi déterminer la combinaison des résultats de plusieurs tests binaires qui conduit à la meilleure classification, en utilisant différents types de modèles, comme la régression logistique.

Cependant, souvent, l'affectation à une classe parmi deux se fait par l'observateur, à partir d'une observable continue délivrée par le système. Dans ce cas, il existe des approches plus performantes pour caractériser les performances du système.

## Modèle des observables pour les tâches de classification

Modèle : 2 populations qui se recouvrent partiellement au sens de l'observable



Le résultat de l'évaluation dépend du seuil de décision

Pour décrire ces approches, il faut considérer un modèle des observables. Dans le cas de deux catégories, les observables peuvent être représentées comme étant issues de 2 populations, qui se recouvrent partiellement au sens de l'observable. L'observable correspond à l'information à disposition, par exemple un taux de protéine.

L'observateur, pour donner une réponse binaire à la question qui lui est posée, se fixe ce qu'on appelle un seuil de décision. Au delà de ce seuil de décision, il qualifie le cas de positif, et en dessous, il le considère comme négatif.

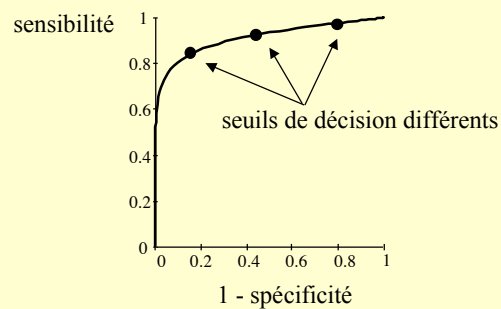
Plus ce seuil est bas, plus la sensibilité va être élevée, mais plus la spécificité sera faible. Et inversement.

Ce modèle met clairement en évidence le fait que les résultats, en termes de sensibilité et de spécificité, et des autres indices précédemment définis, dépendent totalement du seuil de décision.

## Solution : l'approche ROC (Receiver Operating Characteristics)



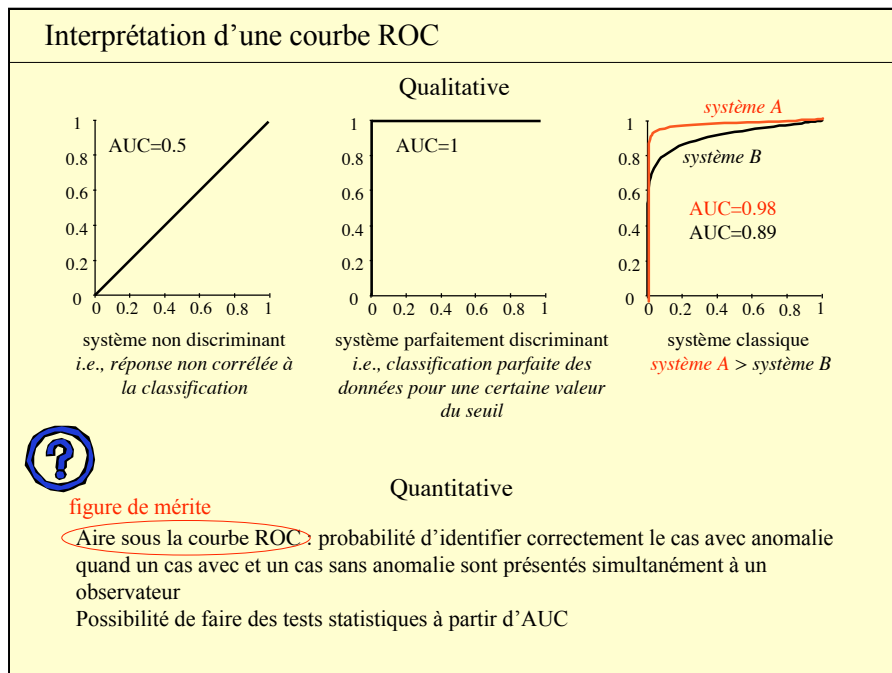
Caractérisation des performances de détection indépendante d'un seuil de décision



L'évaluation par les approches ROC apportent une solution à ce problème de seuil de décision puisque tout leur intérêt est de permettre l'évaluation d'un système indépendamment du choix d'un seuil de décision.

Plus précisément, l'approche ROC consiste à représenter la valeur de la sensibilité en fonction de (1-spécificité) pour toutes les valeurs de seuil possibles, et à joindre ces points par une courbe. Chaque point de la courbe représente le compromis sensibilité/spécificité correspondant à un seuil de décision spécifique.

La courbe ROC résume l'ensemble des compromis sensibilité/spécificité pour les différentes valeurs de seuil de décision.



L'interprétation d'une courbe ROC peut se faire de manière qualitative, en considérant simplement la forme de la courbe.

Une courbe confondue avec la diagonale correspond à un système non discriminant, c'est à dire que la réponse donnée par l'observateur n'est nullement corrélée à la présence ou à l'absence d'une anomalie.

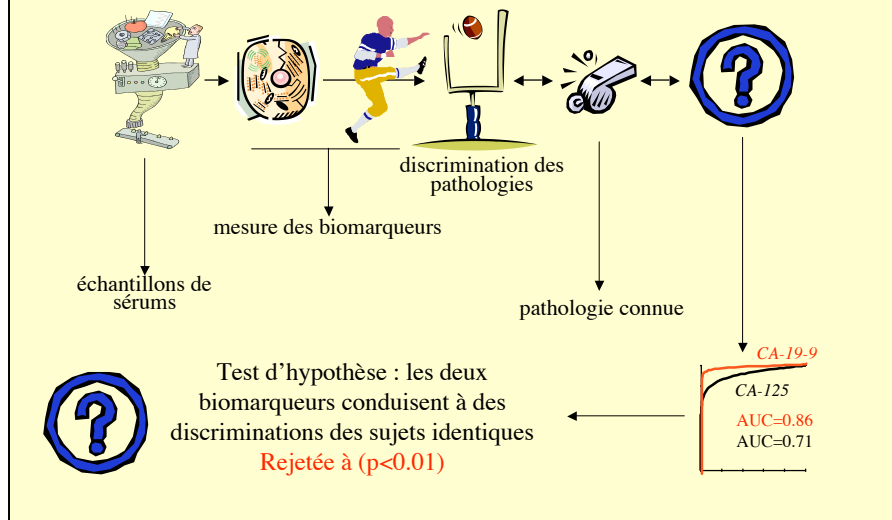
Une courbe de cette forme correspond à un système parfaitement discriminant, pour lequel il existe un seuil de décision séparant parfaitement les deux catégories.

Dans la pratique, on rencontre des courbes de forme intermédiaire. Par exemple ici, la courbe correspondant au système A révèle des performances de détection meilleure que celle correspondant au système B.

Lorsqu'on veut réaliser un test d'hypothèse, ou classer les méthodes par performances croissantes, la figure de mérite associée le plus classiquement aux courbes ROC est l'aire sous la courbe ROC, qui représente la probabilité d'identifier correctement le cas avec anomalie lorsqu'un cas avec anomalie et un cas sans sont présentés simultanément à l'observateur.

## Application de la méthodologie ROC : exemple

Déterminer lequel parmi deux biomarqueurs (CA-19-9 et CA-125) distingue le mieux les sujets atteints de cancer du pancréas de ceux atteints de pancréatite



Cette méthodologie ROC s'applique à de nombreux problèmes d'évaluation. Elle permet par exemple de comparer la capacité de deux biomarqueurs à discriminer cancer du pancréas et pancréatite.

Dans ce cas, la tâche consisterait à identifier la pathologie parmi deux. Le système à évaluer est la mesure, par un observateur, du taux d'un biomarqueur. On va donc calculer une courbe ROC pour chaque système, correspondant à chaque biomarqueur.

La superposition des deux courbes ROC donne déjà des indications sur la potentielle supériorité d'une méthode par rapport à une autre.

Les aires sous les courbes ROC peuvent être calculées comme figure de mérite et permettre de conclure, par un test statistique approprié, à la supériorité effective de l'un ou l'autre des biomarqueurs.

## Potentialités de l'approche ROC



- Exploitation de l'appariement si les mêmes cas sont traitées par deux méthodes à comparer (augmentation de la puissance statistique des tests)
- Données continues (valeur d'une observable) ou discrètes (scores attribués par des biologistes)
- Modèle paramétrique (2 lois normales) ou non paramétrique de l'observable
- Anomalies multiples par échantillon : extensions FROC et AFROC
- Mesures multiples du même paramètre par échantillon : approche Dorfman-Berbaum-Metz (1992)
- Absence de gold standard : travaux de Henkelman (1990) et Beiden (2000)
- ROC à 3 catégories
- Caractérisation ROC des performances de la combinaison de plusieurs paramètres (régression logistique)

L'approche ROC peut se décliner sous de multiples facettes :

Pour augmenter la puissance des analyses ROC, on peut tirer parti de l'appariement de données, par exemple lorsque chaque échantillon donne lieu à deux mesures à comparer entre elles.

L'approche ROC est adaptée que les données soient discrètes (scores attribués par des observateurs à des échantillons) ou continues (index issu d'un algorithme).

La méthodologie ROC repose soit sur un modèle de loi binormale, soit sur aucun modèle paramétrique. Elle est donc toujours applicable pour évaluer des tâches de classification en deux catégories à partir d'une variable continue ou prenant plus de 2 valeurs discrètes.

Des variantes de l'approche ROC traitent des cas où l'on s'intéresse à la détection de plusieurs anomalies potentiellement présentes dans l'échantillon. Ce sont les méthodes FROC et AFROC.

L'approche ROC permet de gérer des résultats obtenus au moyen d'observateurs multiples.

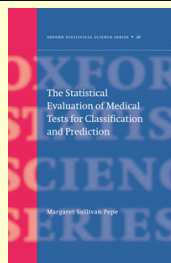
Une méthode a été proposée pour mettre en œuvre une analyse ROC en l'absence de gold standard.

L'approche est également applicable lorsqu'on s'intéresse à une classification en 3 catégories.

Enfin, il est possible de caractériser, par l'approche ROC, les performances de la combinaison de plusieurs paramètres, par des modèles de régression logistique.



## En pratique...



Nombreux programmes disponibles en ligne :

<http://www.bio.ri.ccf.org/Research/ROC/>

[http://xray.bsd.uchicago.edu/krl/roc\\_soft.htm](http://xray.bsd.uchicago.edu/krl/roc_soft.htm)

<http://www.mips.ws/>

Bibliographie :

<http://www.guillemet.org/irene/equipe4/ressources.html>

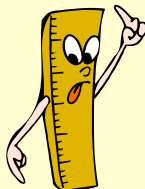
*The statistical evaluation of medical tests for classification and prediction, MS Pepe, Oxford University Press*

En pratique, il existe différentes ressources permettant de mettre en œuvre l'approche ROC sans avoir à reprogrammer quoi que ce soit. Je vous invite en particulier à visiter ces sites Web où vous trouverez les liens vers les pages où des programmes peuvent être téléchargés.

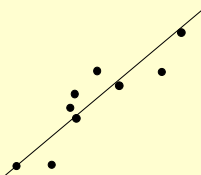
Enfin, des références bibliographiques figurent sur cette page Web, pour vous permettre d'en savoir plus sur toutes les approches ROC. Je recommande également cet ouvrage dans lequel de nombreux exemples sont relatifs à des biomarqueurs.

## Outils pour les tâches d'estimation

Biais et variabilité

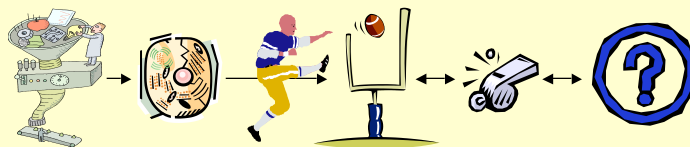


Corrélation



Passons maintenant aux outils nécessaires à l'évaluation de tâches d'estimation.

## Tâches d'estimation : contexte général



Données quelconques

Tâche : extraire une valeur à partir d'un échantillon



Il s'agit cette fois d'extraire une valeur à partir de données quelconques.

## Tâches d'estimation : biais



Requièrent la connaissance de la vraie valeur du paramètre  
Applicables seulement à des données parfaitement caractérisées



Figure de mérite : biais  $\pm$  écart-type  
Tests d'hypothèse possibles

Plusieurs mesures du biais possibles :

$$\% \text{ erreur} = 1/N[\sum_{\text{observations}_i} (p_{i\_estimé} - p_i) / p_i]$$

$$\% \text{ erreur absolue} = 1/N[\sum_{\text{observations}_i} |p_{i\_estimé} - p_i| / p_i]$$

$$\text{erreur quadratique moyenne} = 1/N[\sum_{\text{observations}_i} (p_{i\_estimé} - p_i)^2 / p_i^2]$$

à choisir en fonction du contexte

L'approche la plus intuitive va consister à calculer l'erreur affectant le paramètre estimée, c'est-à-dire le biais.

Une limite à cette approche est qu'elle nécessite bien sur de connaître la vraie valeur du paramètre, ce qui est souvent difficile dans un contexte clinique.

Si on est capable d'estimer un biais, on peut bien sur l'utiliser directement comme figure de mérite, et on peut également faire des tests d'hypothèses, par exemple pour déterminer si les biais obtenus pour deux méthodes sont identiques.

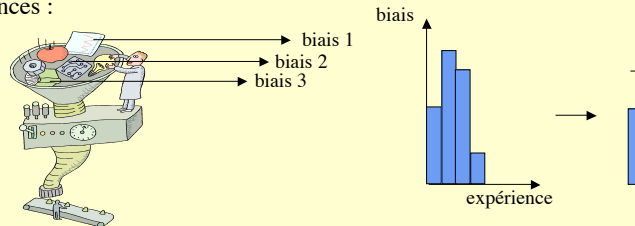
En fonction du contexte, plusieurs index sont possibles pour caractériser le biais, comme le pourcentage d'erreur, le pourcentage d'erreur absolue, ou l'erreur quadratique moyenne. La mesure la plus pertinente dépend du contexte.

## Tâches d'estimation : variabilité



Tout résultat en terme de biais doit être accompagné d'une estimation de la variabilité du biais pour être interprétable

Attention, la variabilité doit être calculée à partir d'un grand nombre d'expériences :



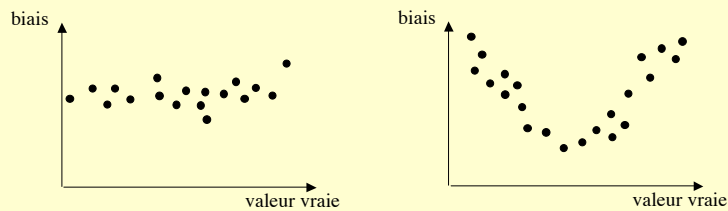
Si loi normale réaliste : écart-type des mesures

Sinon : bootstrap (*Efron et Tibshirani, An introduction to the bootstrap, Chapman and Hall*)

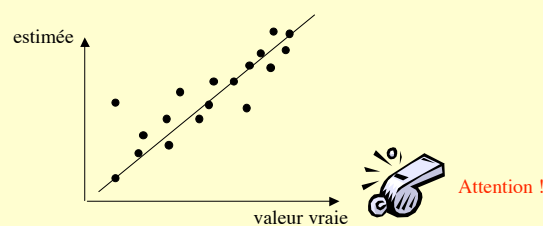
Il est indispensable d'assortir le biais d'une mesure de la variabilité du biais. Cette mesure de variabilité ne peut se faire qu'en répétant un grand nombre de fois l'expérience.

A partir d'expériences multiples, si les mesures suivent a priori une loi normale, la variabilité peut être caractérisée par l'écart type des mesures. Si l'hypothèse de la loi normale n'est pas justifiée, il est préférable d'utiliser une approche de bootstrap non paramétrique pour estimer la variabilité des mesures.

## Insuffisance des mesures de biais et variabilité



Evaluation plus complète : la régression linéaire ?



Le biais et la variabilité sont des index très synthétiques, mais ils peuvent masquer une dépendance du biais à la vraie valeur du paramètre qu'il peut être important de connaître.

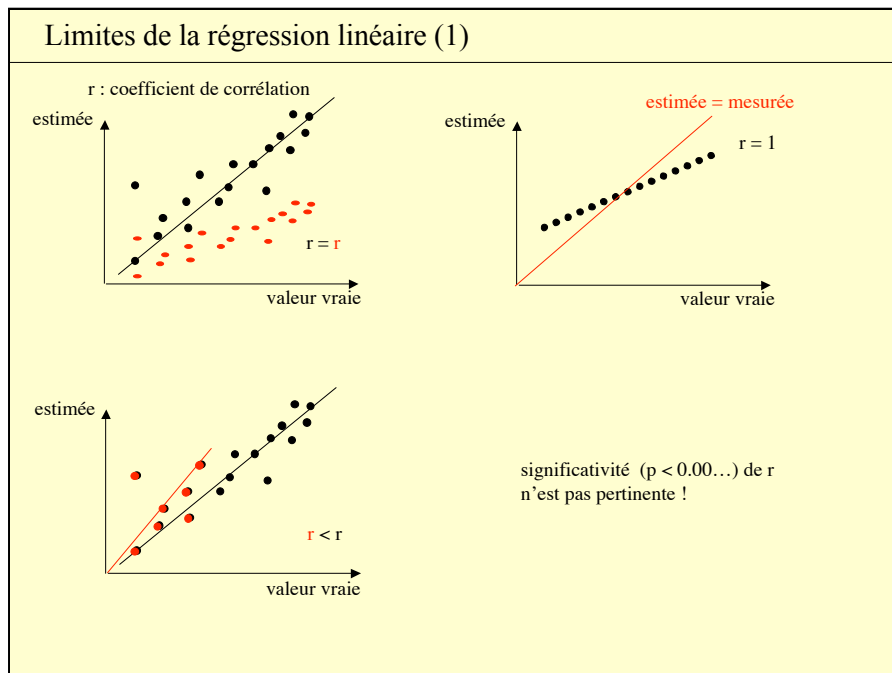
Schématiquement, biais et variabilité peuvent suffire lorsque le biais dépend peu de la valeur du paramètre. En revanche, lorsque le biais dépend sensiblement de la valeur du paramètre, l'évaluation de la méthode nécessite de faire appel à une approche d'évaluation plus complète.

On réalise alors souvent une étude de la régression linéaire entre la valeur estimée et la valeur vraie, pour une large étendue des valeurs vraies potentiellement observables en pratique.

La régression linéaire permet de déterminer :

- si il existe une relation linéaire entre la valeur estimée et la vraie valeur du paramètre. L'existence de cette relation linéaire peut parfois suffire à une interprétation des données.
- si le biais dépend de la vraie valeur du paramètre.

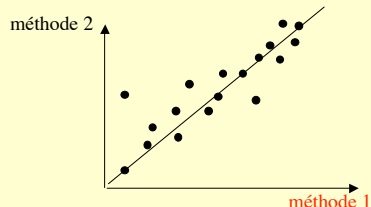
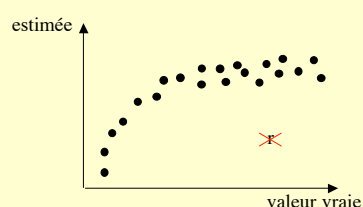
Attention, la régression linéaire suppose que la référence est connue.



La régression linéaire a cependant ces limites.

- Tout d'abord, on caractérise souvent le résultat d'une régression linéaire par un coefficient de corrélation. On conclut souvent que la méthode est d'autant plus fiable que ce coefficient est élevé. Mais il faut bien avoir conscience que ce coefficient représente le degré de corrélation entre l'estimée et la référence, mais pas l'accord entre ces deux quantités. Autrement dit, on peut parfaitement avoir un  $r = 1$  avec une mesure biaisée.
- La valeur du coefficient de corrélation est insensible au changement d'échelle, alors que le biais l'est bien évidemment.
- La valeur du coefficient de corrélation dépend directement de l'étendue des valeurs de références utilisées pour le calculer. Il est d'autant plus élevé que cette étendue est grande.
- Enfin, certains auteurs fondent leurs conclusions sur le degré de significativité de  $r$ . Il est évident que l'estimée est quasiment toujours significativement corrélée au paramètre que l'on cherche à évaluer. La significativité apporte donc très peu d'informations sur la méthode la moins biaisée...

## Limites de la régression linéaire (2)



\* étude de la corrélation entre les deux méthodes, indépendamment du biais !  
\* peu sensible



L'analyse de la corrélation linéaire entre l'estimée et la valeur vraie est utile pour une interprétation correcte du biais moyen et de sa variabilité.

La caractérisation des performances d'une méthode ou la comparaison des performances de deux méthodes via le coefficient de corrélation est hasardeuse...

La régression linéaire conduit également à des résultats anormalement pessimistes si la relation entre l'estimée et la valeur vraie n'est pas une relation linéaire. Il est donc indispensable de toujours visualiser cette relation pour juger de la pertinence de la valeur du coefficient de corrélation.

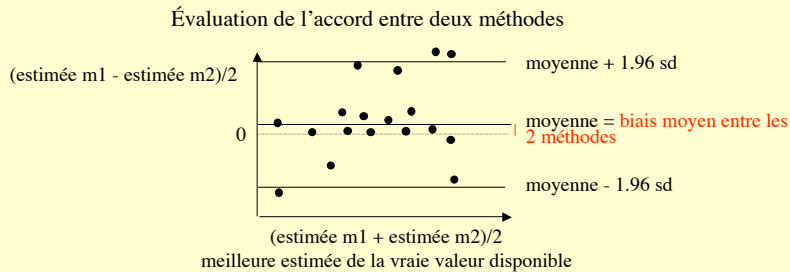
Enfin, souvent, on utilise la régression linéaire pour juger de l'accord entre deux méthodes, lorsque aucune d'elles ne peut faire office de gold standard. En fait, nous allons voir que cette approche est peu sensible, et que lorsqu'on ne dispose pas de gold standard, il existe des approches plus performantes pour mesurer l'accord entre deux méthodes, voire pour estimer la moins biaisée.

En conclusion, l'analyse de la corrélation linéaire entre l'estimée et la valeur vraie est utile pour une interprétation correcte du biais moyen et de sa variabilité.

En revanche, la caractérisation des performances d'une méthode et la comparaison des performances de deux méthodes via le coefficient de corrélation sont hasardeuses.



## Absence de gold standard : approche de Bland-Altman (1)



Réf : Bland and Altman. *Lancet*, 307-10, 1986.



La plupart des différences sont comprises dans l'intervalle [moyenne - 2 sd ; moyenne + 2 sd]

L'étendue de cet intervalle doit permettre de conclure à l'interchangeabilité des méthodes ou non, **mais PAS au fait que l'une est moins biaisée que l'autre !**

En l'absence de gold standard, on cherche souvent à caractériser la cohérence des résultats produits par la nouvelle méthode par rapport aux résultats obtenus par la méthode faisant office de standard faute de mieux.

Pour étudier cette cohérence, une approche plus puissante qu'une régression linéaire est l'approche proposée par Bland et Altman. Il s'agit d'une approche simple à mettre en œuvre : il suffit de représenter, pour chaque cas, la différence moyenne entre les estimées issues des 2 méthodes en fonction de la moyenne des deux estimées. En l'absence d'informations supplémentaires, cette moyenne représente en effet la meilleure estimée de la valeur vraie du paramètre.

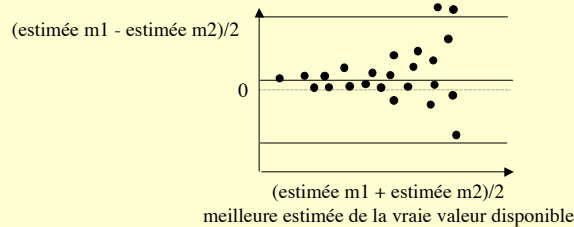
Sur un tel graphe, la moyenne des différences correspond au biais moyen entre les deux méthodes. Si on fait l'hypothèse que les différences suivent une loi normale, 95% des différences seront comprises entre cette valeur moyenne - 1.96 x écart-type des différences, et la valeur moyenne + 1.96 x écart-type des différences. C'est la raison pour laquelle on représente également ces deux lignes sur le graphe.

A quelles questions ce type de graphe permet-il de répondre ?

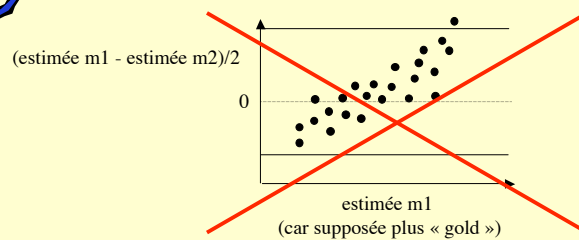
En fonction de l'étendue de cet intervalle, il permet de déterminer si les deux méthodes sont interchangeables (par exemple en se fixant la valeur de différence maximale tolérable pour que les méthodes puissent être utilisées l'une à la place de l'autre).

Cependant, cette approche ne permet en aucun cas de conclure à la supériorité d'une méthode par rapport à une autre, au sens d'une estimation moins biaisée.

## Absence de gold standard : approche de Bland-Altman (2)



Permet de détecter des différences systématiques entre les méthodes



Réf : Bland and Altman. *Lancet*, 346: 1085-7, 1995.

La représentation de Bland-Altman permet en outre de mettre en évidence des différences entre les méthodes qui dépendraient de la valeur du paramètre à estimer. Par exemple ici, les méthodes produisent des résultats similaires pour de faibles valeurs du paramètres, puis divergents quand la valeur du paramètre augmente.

Des utilisations abusives de la méthode de Bland-Altman ont fait l'objet d'une mise au point par les auteurs.

Parfois, on a tendance à avoir plus confiance dans une méthode que dans une autre, et la représentation de Bland-Altman a été utilisée en représentant la différence des estimées produits par les 2 méthodes en fonction de la valeur estimée par une seule méthode. Bland et Altman montrent que cette représentation est incorrecte, car elle conduit toujours à l'observation d'une tendance, qui n'est pas réelle.

**Régression sans gold standard : détermination de la meilleure méthode**

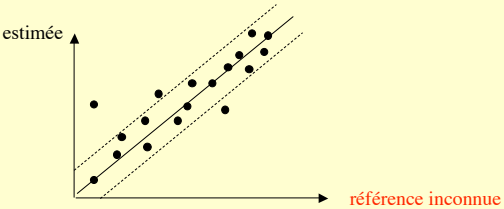
Hypothèses :


- \* évaluation de la justesse des estimées résultant de la méthode m
- \*  $p_{mi} = a_m p_i + b_m + \varepsilon_{mi}$  avec i indiquant le cas
- \*  $p_i$  inconnus
- \*  $\varepsilon_m$  suit une loi normale centrée (écart-type  $\sigma_m$ )
- \*  $p_i$  suit une loi de probabilité de forme connue (sans que les paramètres  $r$  de cette loi soient eux même connus, e.g., loi normale)

Méthode :

Détermination des paramètres du modèle qui maximisent la vraisemblance des observables :  $a_m, b_m, \sigma_m, r$

Figure de mérite :  $\sigma_m / a_m$  (le plus faible possible)



 Permet de déterminer la méthode la plus fiable quantitativement en l'absence de gold standard

Au delà de l'analyse de Bland-Altman, il existe en fait une méthode qui permet d'évaluer la justesse d'une méthode d'estimation en l'absence de gold standard.

Evidemment, il n'y a pas de miracles, donc l'absence de gold standard doit être suppléée par un modèle sur le lien entre les observables et le gold standard inconnu.

Le modèle le plus simple consiste à supposer que l'observable est une fonction linéaire de la valeur vraie du paramètre,  $p_i$ , qui est elle même inconnue, à une erreur près, qui est supposée gaussienne centrée, mais de variance inconnue.

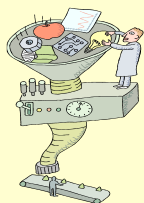
On doit en outre supposée connue le type de loi des  $p_i$ , sans que les paramètres de la loi soient connus. Par exemple, on peut supposer que les vraies valeurs du paramètres suivent une loi normale, de moyenne et d'écart-type inconnus.

La méthode consiste alors à déterminer les paramètres du modèle qui maximisent la vraisemblance des observables. Par exemple, en utilisant un algorithme de type EM, on obtient des estimées des paramètres du modèle. Cette estimation revient à faire une régression linéaire par rapport à une variable inconnue. La meilleure méthode sera celle qui conduira à l'erreur la plus faible, c'est à dire au rapport  $\sigma_m/a_m$  le plus faible.

Si on effectue la procédure pour plusieurs méthodes, on peut donc déterminer quelle méthode est a priori la plus fiable.

## Régression sans gold standard : détermination de la meilleure méthode

Caractéristiques de l'approche :



- \* 2 méthodes sont suffisantes
- \* généralisable à une dépendance non linéaire :  
$$p_{mi} = f(\mathbf{p}_i, v_m) + \varepsilon_{mi}$$
 avec  $i$  indiquant le cas
- \* robuste même lorsque l'hypothèse sur la distribution des  $\mathbf{p}_i$  est approximative
- \* 25 cas au moins sont nécessaires

Réf: Kupinski et al. *Academic Radiology*, 9: 290-297, 2002  
Hoppin et al. *IEEE Trans Med Imaging*, 21: 441-449, 2002

Pour appliquer la méthode avec suffisamment de robustesse, 2 estimées du gold standard par deux méthodes différentes sont nécessaires. On peut également avoir davantage d'estimées, mais le gain en justesse est alors modéré.

Je vous ai présenté le modèle sous-jacent à la méthode lorsqu'il existe une relation linéaire entre estimée et valeur supposée du gold standard. En fait, la méthode peut tout à fait s'appliquer lorsque cette relation n'est pas linéaire. C'est essentiellement le temps de calcul qui augmente dans ce cas.

Les auteurs ont montré que l'approche reste robuste même lorsque l'hypothèse sur la distribution statistique des  $\mathbf{p}_i$  est approximative.

Enfin, concernant le nombre de cas nécessaire, des premiers résultats suggèrent que 25 cas au moins sont nécessaires pour aboutir à une classification correcte des méthodes entre elles.

Conclusions

Deux types de travaux d'évaluation :

- tâche de classification
- tâche d'estimation

Pour chaque type, méthodes d'évaluation rigoureuses

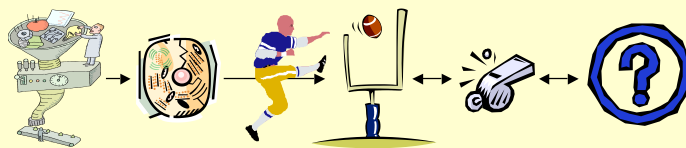
- approches type ROC et dérivées
- calcul de biais et variabilité après observation de la dépendance de l'erreur à la valeur vraie

Même en l'absence de gold standard, une évaluation objective rigoureuse est possible

En conclusion, les messages à retenir sont les suivants :

- il existe deux types de travaux d'évaluation, ceux à mettre en œuvre pour les tâches de classification, et ceux concernant les tâches d'estimation.
- Pour chacune de ces tâches, il existe des méthodes rigoureuses d'évaluation objective. Pour les tâches de classification, les méthodes les plus sophistiquées relèvent de l'analyse ROC. Pour les tâches d'estimation, il s'agit des calculs de biais et de variabilité, après inspection de la relation entre l'erreur et la valeur vraie du paramètre.
- Enfin, la plupart de ces méthodes connaissent des extensions applicables en l'absence de gold standard. Par conséquent, même en l'absence de gold standard, une évaluation objective et rigoureuse est possible.

Copie de la présentation



<http://www.guillemet.org/irene/equipe4/conferences.html>