

## Evaluation en imagerie médicale

Irène Buvat

IMIV

UMR 1023 Inserm/CEA/Université Paris Sud

ERL 9218 CNRC

Orsay

[irene.buvat@u-psud.fr](mailto:irene.buvat@u-psud.fr)

<http://www.guillemet.org/irene>

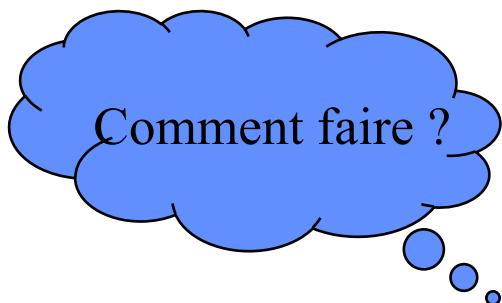
Novembre 2018

# Objectifs pédagogiques

- Comprendre le contexte de l'évaluation en imagerie médicale



- Savoir concevoir une étude d'évaluation



- Comprendre la distinction entre les 2 problématiques d'évaluation auxquelles on peut être confronté
- Connaître les méthodes d'évaluation appropriées pour évaluer des méthodes de classification
- Connaître les méthodes d'évaluation appropriées pour évaluer des méthodes d'estimation



# Plan du cours

---

- L'évaluation en imagerie médicale
- Notions de base
  - population, échantillon et inférence statistique
  - distribution et caractérisation d'une population
  - intervalles de confiance
  - hypothèses H<sub>0</sub>, H<sub>1</sub>, tests unilatéral et bilatéral
  - valeur de p ( $p < 0,05$ ), niveau de significativité
  - degrés de liberté
  - tests d'hypothèse
- Evaluation pour des tâches de classification
  - Définitions classiques
  - Insuffisance des indices classiques
  - Approche ROC
- Evaluation pour des tâches d'estimation
  - Définitions classiques
  - Régression linéaire
  - Approche de Bland Altman
  - Evaluation sans gold standard
- Annexes et compléments

# Contexte de l'évaluation

---

Déterminer si la méthode X  
conduit à des informations pertinentes  
concernant (la présence d') une pathologie  
chez un groupe de sujets Y

Quoi évaluer ?

Dans quel but ?

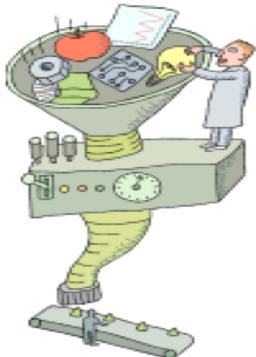
Sur quelles données ?

# Exemples

- Y a-t-il une corrélation entre la valeur d'un indice extrait d'une image et la présence de l'anomalie ?
- Le volume segmenté délimite t-il bien le volume réel de la structure ?
- L'indice 1 est-il meilleur ou moins “bon” que l'indice 2 pour différencier deux pathologies ?
- L'image filtrée par le filtre A contient-elle une information mieux corrélée au diagnostic que celle filtrée par le filtre B ?



# Conception d'une étude



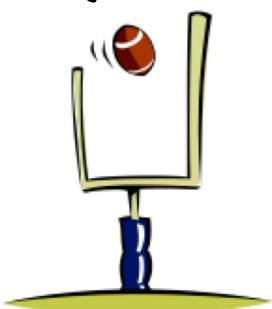
*données*



*ce qu'on cherche à évaluer  
système*



*comment est effectuée la tâche  
observateur*



*finalité du système  
tâche*

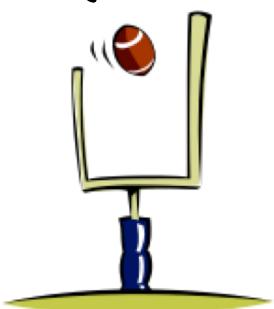
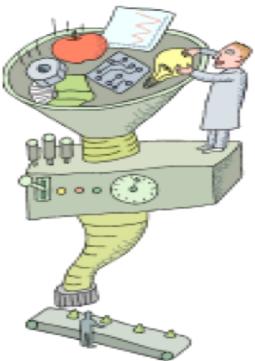


*référence  
ground truth  
gold standard*



*question posée  
figure de mérite  
test d'hypothèse*

# Choix de l'outil d'évaluation



Quel type de tâche ?

Classification  
ou  
estimation ?

## Effectuer des tests statistiques

- Procédure permettant de déterminer si une hypothèse est acceptée ou rejetée, de façon statistiquement significative
- Utilisé quasiment systématiquement pour présenter des résultats de façon objective
- Repose sur un certain nombre de notions :
  - population, échantillon et inférence statistique
  - distribution d'une population
  - moyenne, écart-type
  - intervalles de confiance
  - hypothèse H0
  - tests unilatéral et bilatéral
  - valeur de p ( $p < 0,05$ )
  - erreurs de type I et II,  $\alpha$  et  $\beta$



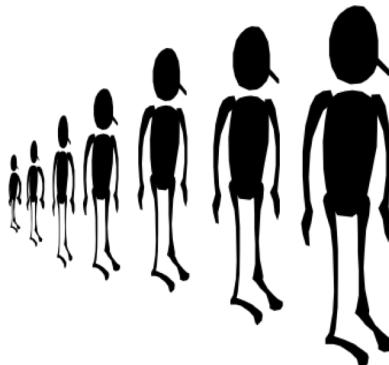
# Notions de base

---

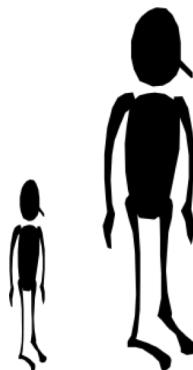


# Population, échantillons et inférence statistique

- Population : grand ensemble de valeurs, d'observations, d'objets, trop grand en pratique pour pouvoir faire l'objet de mesures exactes pour les caractériser



- Echantillon : petite partie d'une population qu'il est possible d'observer et sur laquelle il est possible de faire des mesures

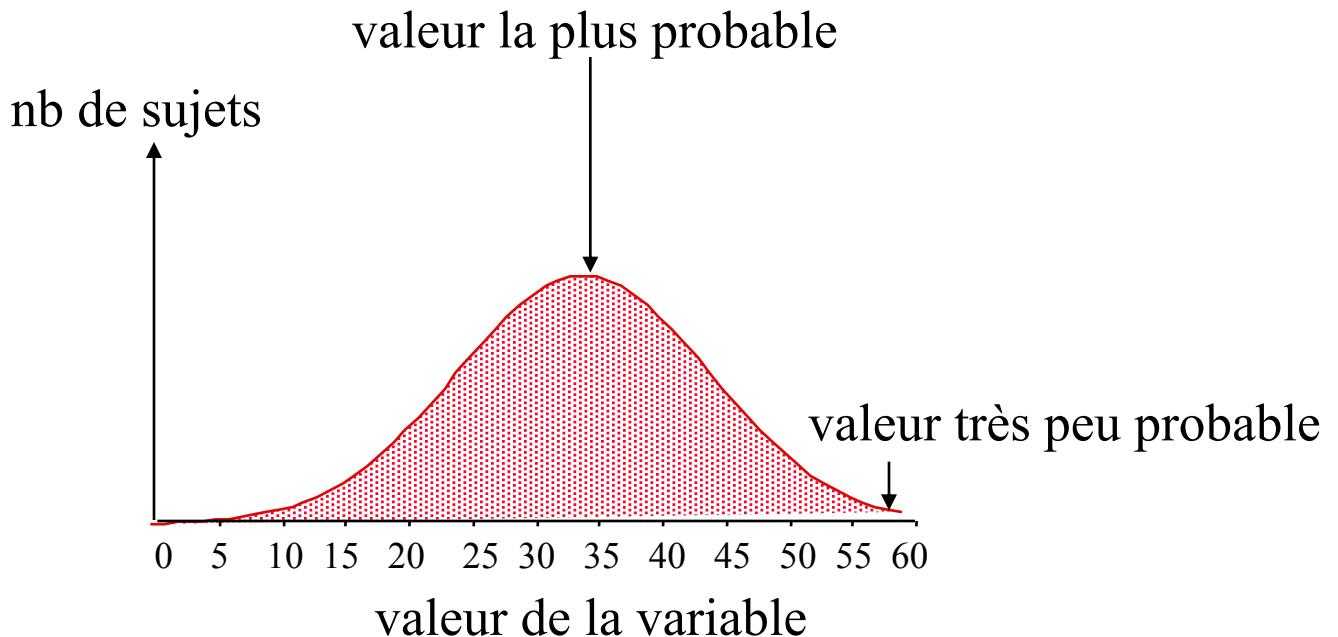


- Inférence statistique : tirer des enseignements sur la population à partir des observations faites sur l'échantillon
- Exemple : estimer la moyenne d'une population à partir de la moyenne d'un échantillon



# Distribution d'une population

- Diagramme représentant la fréquence des différentes valeurs prises par la population



- Il existe différentes formes de distributions, la plus fréquente étant la distribution normale



# Caractérisation d'une population : moyenne, écart-type

7, 6, 3, 11, 5, 7, 7, 9, 6, 5

- Moyenne : somme de toutes les observations  $x_i$  divisée par le nombre d'observation  $N$

$$\begin{aligned}\text{moyenne} &= x_{\text{moyen}} = \sum_i x_i / N \\ &= (7 + 6 + 3 + 11 + 5 + 7 + 7 + 9 + 6 + 5) / 11 \\ &= 66 / 10 = 6,6\end{aligned}$$

Caractérise la « position » de la distribution

- Ecart-type (racine carré de la variance) caractérise la dispersion (l'étalement) de la distribution

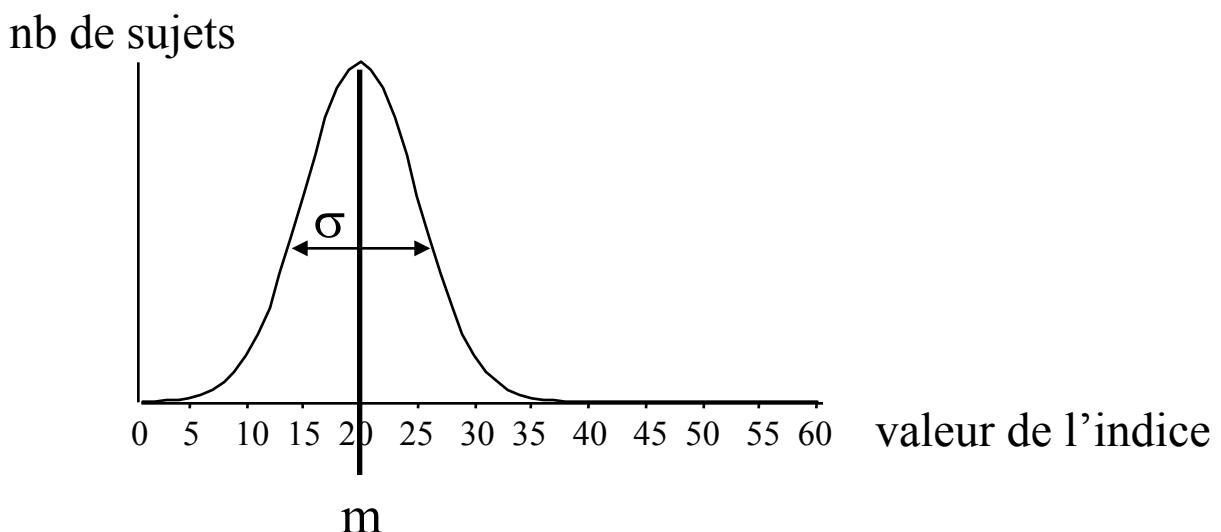
$$\begin{aligned}\text{Écart-type} &= \sigma = \{\sum_i (x_i - x_{\text{moyen}})^2 / (N-1)\}^{1/2} \\ &= \{[\sum_i x_i^2 - (\sum_i x_i)^2 / N] / (N-1)\}^{1/2} \\ &= 2,2\end{aligned}$$

- Coefficient de variation =  $100 * \text{écart-type}/\text{moyenne}$



# La loi normale (ou gaussienne)

- Loi symétrique, en forme de cloche, décrivant bien la distribution statistique de nombreux indices « naturels » (tension artérielle, taille, etc).
- Entièrement caractérisée par 2 paramètres : moyenne  $m$  et écart-type  $\sigma$



$$y = \frac{1}{\sqrt{2 \pi \sigma^2}} \exp \frac{-(x - m)^2}{2 \sigma^2}$$



# Calcul d'écart-types

- Ecart-type de la moyenne d'un échantillon de taille N (à ne pas confondre avec l'écart-type de l'échantillon)

$$ET = \text{écart-type de l'échantillon} / N^{1/2}$$

- Ecart-type de la différence entre les 2 moyennes  $m_1$  et  $m_2$  d'échantillons de tailles  $N_1$  et  $N_2$  et d'écart-type  $\sigma_1$  et  $\sigma_2$

$$ET = [(\sigma_1^2/N_1 + \sigma_2^2/N_2)]^{1/2}$$



En pratique, une valeur résultat devrait toujours être accompagnée de son écart-type, qui donne une indication sur la confiance qu'on peut accorder à la valeur

# Intervalles de confiance

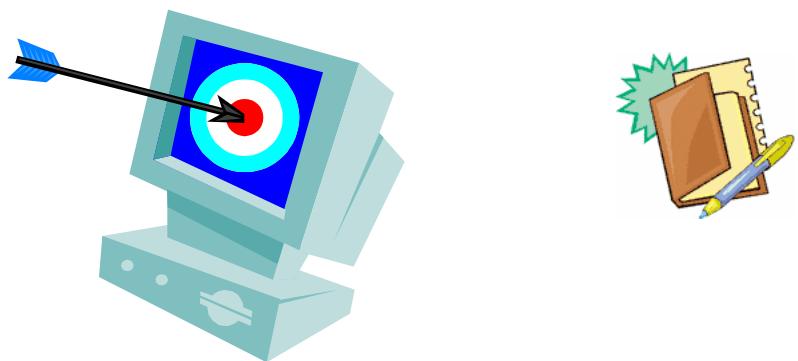
- Intervalle dans lequel la probabilité P pour que la valeur du paramètre k recherché se trouve est de X%
- Pour une population suivant une loi normale, cet intervalle est donné par

$$[k_{\text{moyen}} - 1,96 \text{ ET} ; k_{\text{moyen}} + 1,96 \text{ ET}]$$

où  $k_{\text{moyen}}$  est la moyenne sur l'échantillon observé  
ET représente l'écart-type de  $k_{\text{moyen}}$   
et P = 95% (N>30)

$$[k_{\text{moyen}} - 1,64 \text{ ET} ; k_{\text{moyen}} + 1,64 \text{ ET}] \text{ pour } P = 90\%$$

$$[k_{\text{moyen}} - 2,576 \text{ ET} ; k_{\text{moyen}} + 2,576 \text{ ET}] \text{ pour } P = 99\%$$



# Comment répondre objectivement à une question ?



Effectuer des tests statistiques



# Notion d'hypothèse H0

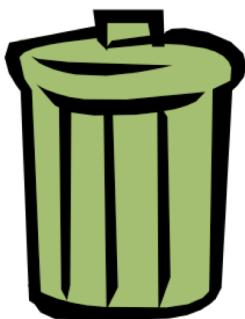
- H0 : hypothèse nulle correspondant à l'hypothèse que l'on souhaite tester

e.g. :

H0 : A = B : il n'y a pas de différence entre les valeurs de fraction d'éjection mesurées à partir des images de type A et celles mesurées à partir des images de type B

H0 : les observations proviennent d'une distribution de Poisson

L'hypothèse nulle est généralement formulée de façon à être rejetée (sauf dans le cas d'ajustement de courbes)



Pourquoi ?

Car il y a en général plusieurs hypothèses alternatives possibles, et qu'il est plus facile de formuler l'hypothèse à rejeter que toutes les alternatives possibles



# Notion d'hypothèse alternative

- H1 : hypothèse à opposer à H0

e.g. :

H0 : A = B : il n'y a pas de différence entre les valeurs de fraction d'éjection mesurées à partir des images de type A et celles mesurées à partir des images de type B

H1 : A  $\neq$  B : les valeurs de fraction d'éjection mesurées à partir des images de type A sont différentes de celles mesurées à partir des images de type B



- 2 types d'hypothèses alternatives :

- bilatérale (2-tailed)

H1 : A  $\neq$  B : les valeurs sont différentes

- unilatérale (1-tailed)

H1 : A > B

ou

H1 : A < B



# Valeurs de p

- $p < 0,05$

Valeur de p associée à un test = probabilité d'observer la valeur obtenue si l'hypothèse H<sub>0</sub> est vraie

Exemple :

- les patients ayant reçu la thérapie X (groupe I) ont eu moins d'accidents cardiaques que les patients sous placebo (groupe II) ( $p < 0,05$ )

Signifie que la survenue d'accidents cardiaques telles qu'elle a été observée dans les 2 groupes s'observerait dans moins de 5% des cas (moins de 1 cas sur 20) si la thérapie n'avait pas d'effet.

Autrement dit, la probabilité d'observer ce qu'on observe en termes de survenue d'accident cardiaques dans les 2 groupes « par hasard » (sans relation avec la thérapie) est de moins d'1 chance sur 20.



# Significativité statistique, niveau de significativité

- Fixer un niveau de significativité  $\alpha$ , c'est choisir avec quelle probabilité on accepte de rejeter  $H_0$  alors que  $H_0$  est vraie
- Rejeter  $H_0$  alors que  $H_0$  est vraie = erreur de type I, habituellement notée  $\alpha$  ( $\alpha = 0,05 = 5\%$ ) : observer une différence alors qu'il n'y en a pas.
- Erreur de type II = accepter  $H_0$  alors qu'elle est fausse, habituellement notée  $\beta$  : ne pas observer de différence alors qu'il y en a une.
- Puissance du test = probabilité de rejeter  $H_0$  à juste titre =  $1 - \beta$
- $\alpha$  et  $\beta$  ne sont pas indépendants : si  $\alpha$  augmente,  $\beta$  diminue et réciproquement



# Degrés de liberté

- Nombre d'observations variant indépendamment après que certaines contraintes aient été placées sur ces observations
- Le plus souvent,  $\text{ddl} = N-1$
- Exemple : choisir dans une boîte de  $N$  chocolats

Pour chaque chocolat, on a le choix, jusqu'à ce qu'il ne reste plus qu'un chocolat dans la boîte, où là, on n'a plus le choix. On a donc  $N-1$  choix.



# Généralités sur les tests d'hypothèse

- La procédure implique le calcul d'une variable, qu'on appelle la statistique de test
- Le test peut être paramétrique (faisant des hypothèses sur la distribution de la population impliquée dans le test) ou non paramétrique
- Il existe un grand ensemble de tests statistiques, chacun étant approprié pour tester certains types d'hypothèse H0
- e.g.,
  - Pour un appareil de comptage :  
H0 : il n'y a pas de différence entre les nombres de coups prédits par une théorie et ceux mesurés  
Test approprié : test du Chi-2
  - Pour l'étude de l'effet d'une substance :  
H0 : la tension artérielle mesurée après traitement est identique à celle mesurée avant traitement  
Test approprié : test t apparié

Nécessité d'avoir un bon ouvrage de référence pour faire les bons choix

# Principe d'un test d'hypothèse statistique

- Formuler clairement le problème à étudier
- Formuler l'hypothèse  $H_0$  correspondante
- Déterminer le test d'hypothèse approprié pour tester  $H_0$  :  
e.g. :  
test t de Student, test z, test du Chi-2
- Choisir un degré de significativité  $\alpha$ , i.e., la probabilité d'erreur de type I : probabilité de rejeter  $H_0$  alors que  $H_0$  est vraie (erreur de type I, habituellement,  $\alpha = 0,05 = 5\%$ )
- Calculer la statistique de test  $s$   
e.g. : valeur du t de Student, valeur de z, valeur du Chi-2
- Consulter la table appropriée indiquant la probabilité p d'observer la valeur de la statistique  $s$  si  $H_0$  est vraie
- Si  $p < \alpha$ , rejeter  $H_0$ , sinon, accepter  $H_0$



# Exemple : le test t de Student

---

- Test paramétrique qui suppose que :
  - la distribution statistique des populations desquelles sont issus les échantillons observés suivent des lois normales
  - les échantillons sont indépendants
  - les populations normales ont la même variance
  - les variables caractérisant les populations sont continues
- Test robuste, c'est-à-dire restant valide pour de faibles écarts aux conditions ci-dessus
- Test approprié pour de petits échantillons ( $N < 30$ )
- Statistique de test :

$$t = \frac{\text{moyenne1} - \text{moyenne2}}{\text{écart-type}_{(\text{moyenne1} - \text{moyenne2})}}$$

## Exemple : test t appliqué à un échantillon

---

- On dispose d'un échantillon de N mesures, de moyenne  $m$  et d'écart-type  $\sigma$
- On veut savoir si la moyenne  $M$  de la population dont est issu l'échantillon est significativement différente de la moyenne connue  $\mu$  d'une population donnée
- Statistique de test :

$$t = \frac{m - \mu}{\sigma/\sqrt{N}}$$

- Comparer la valeur de  $t$  à la valeur dans la table en considérant  $N-1$  ddl

# Table des valeurs de t

**TABLE A-2**

Critical values

n	p=0,1	p=0,05	p=0,025	p=0,01	p=0,005 (unilatéral)	ddl
	p=0,2	p=0,1	p=0,05	p=0,02	p=0,01 (bilatéral)	
2	3.078	6.314	12.706	31.821	63.657	1
3	1.886	2.920	4.303	6.965	9.925	2
4	1.638	2.353	3.182	4.541	5.841	3
5	1.533	2.132	2.776	3.747	4.604	4
6	1.476	2.015	2.571	3.365	4.032	5
7	1.440	1.943	2.447	3.143	3.707	6
8	1.415	1.895	2.365	2.998	3.499	7
9	1.397	1.860	2.306	2.896	3.355	8
10	1.383	1.833	2.262	2.821	3.250	9
11	1.372	1.812	2.228	2.764	3.169	10
12	1.363	1.796	2.201	2.718	3.106	11
13	1.356	1.782	2.179	2.681	3.055	12
14	1.350	1.771	2.160	2.650	3.012	13
15	1.345	1.761	2.145	2.624	2.977	14
16	1.341	1.753	2.131	2.602	2.947	15
17	1.337	1.746	2.120	2.583	2.921	16
18	1.333	1.740	2.110	2.567	2.898	17
19	1.330	1.734	2.101	2.552	2.878	18
20	1.328	1.729	2.093	2.539	2.861	19
21	1.325	1.725	2.086	2.528	2.845	20
22	1.323	1.721	2.080	2.518	2.831	21
23	1.321	1.717	2.074	2.508	2.819	22
24	1.319	1.714	2.069	2.500	2.807	23
25	1.318	1.711	2.064	2.492	2.797	24
26	1.316	1.708	2.060	2.485	2.787	25
27	1.315	1.706	2.056	2.479	2.779	26
28	1.314	1.703	2.052	2.473	2.771	27
29	1.313	1.701	2.048	2.467	2.763	28
30	1.311	1.699	2.045	2.462	2.756	29
inf.	1.282	1.645	1.960	2.326	2.576	inf.

- Exemple :

Echantillon de 20 sujets dont la tension artérielle diastolique vaut en moyenne  $m = 98$  ( $\sigma=6,2$ ).

$H_0$  : cet échantillon est issu d'une population dont la tension artérielle diastolique moyenne est  $\mu=93$

$H_1$  : cet échantillon est issu d'une population dont la tension artérielle diastolique moyenne est supérieure à 93

$$t = (98-93)/(6,2/20^{1/2}) = 3,61$$

$$p<0,005$$

# Exemple : test z entre 2 échantillons indépendants

- On dispose de 2 échantillons de N<sub>1</sub> et N<sub>2</sub> mesures chacun (**N<sub>1</sub>>30 et N<sub>2</sub>>30**), de moyennes m<sub>1</sub> et m<sub>2</sub> et d'écart-type σ<sub>1</sub> et σ<sub>2</sub>
- Les distributions des populations dont sont issus les 2 échantillons sont approximativement normales
- H<sub>0</sub> : il n'y a pas de différence entre la valeur du paramètre entre les 2 populations dont sont issus les 2 échantillons
- Statistique de test :

$$z = \frac{m_1 - m_2}{\text{écart-type}_{(m1-m2)}}$$

- écart-type<sub>(m1-m2)</sub> =  $[(\sigma_1^2/N_1 + \sigma_2^2 /N_2)]^{1/2}$
- Comparer à la valeur de z critique dans la table

# Table des valeurs de probabilités pour une loi z

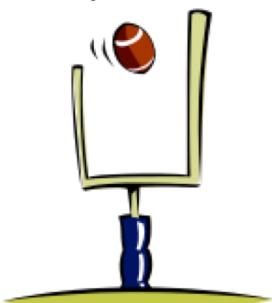
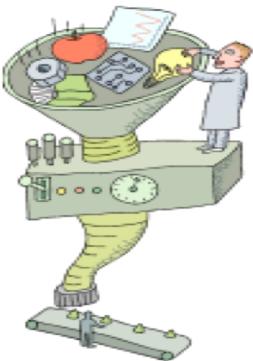
**TABLE A-1**

Probability table for the normal distribution

$z$	$\alpha$														
0.00	.5000	0.35	.3632	0.70	.2420	1.05	.1469	1.40	.0808	1.75	.0401	2.10	.0179	2.45	.0071
0.01	.4960	0.36	.3594	0.71	.2389	1.06	.1446	1.41	.0793	1.76	.0392	2.11	.0174	2.46	.0069
0.02	.4920	0.37	.3557	0.72	.2358	1.07	.1423	1.42	.0778	1.77	.0384	2.12	.0170	2.47	.0068
0.03	.4880	0.38	.3520	0.73	.2327	1.08	.1401	1.43	.0764	1.78	.0375	2.13	.0166	2.48	.0066
0.04	.4840	0.39	.3483	0.74	.2296	1.09	.1379	1.44	.0749	1.79	.0367	2.14	.0162	2.49	.0064
0.05	.4801	0.40	.3446	0.75	.2266	1.10	.1357	1.45	.0735	1.80	.0359	2.15	.0158	2.50	.0062
0.06	.4761	0.41	.3409	0.76	.2236	1.11	.1335	1.46	.0721	1.81	.0351	2.16	.0154	2.51	.0060
0.07	.4721	0.42	.3372	0.77	.2206	1.12	.1314	1.47	.0708	1.82	.0344	2.17	.0150	2.52	.0059
0.08	.4681	0.43	.3336	0.78	.2177	1.13	.1292	1.48	.0694	1.83	.0336	2.18	.0146	2.53	.0057
0.09	.4641	0.44	.3300	0.79	.2148	1.14	.1271	1.49	.0681	1.84	.0329	2.19	.0143	2.54	.0055
0.10	.4602	0.45	.3264	0.80	.2119	1.15	.1251	1.50	.0668	1.85	.0322	2.20	.0139	2.55	.0054
0.11	.4562	0.46	.3228	0.81	.2090	1.16	.1230	1.51	.0655	1.86	.0314	2.21	.0136	2.56	.0052
0.12	.4522	0.47	.3192	0.82	.2061	1.17	.1210	1.52	.0643	1.87	.0307	2.22	.0132	2.57	.0051
0.13	.4483	0.48	.3156	0.83	.2033	1.18	.1190	1.53	.0630	1.88	.0301	2.23	.0129	2.58	.0049
0.14	.4443	0.49	.3121	0.84	.2005	1.19	.1170	1.54	.0618	1.89	.0294	2.24	.0125	2.59	.0048
0.15	.4404	0.50	.3085	0.85	.1977	1.20	.1151	1.55	.0606	1.90	.0287	2.25	.0122	2.60	.0047
0.16	.4364	0.51	.3050	0.86	.1949	1.21	.1131	1.56	.0594	1.91	.0281	2.26	.0119	2.61	.0045
0.17	.4325	0.52	.3015	0.87	.1922	1.22	.1112	1.57	.0582	1.92	.0274	2.27	.0116	2.62	.0044
0.18	.4286	0.53	.2981	0.88	.1894	1.23	.1093	1.58	.0571	1.93	.0268	2.28	.0113	2.63	.0043
0.19	.4247	0.54	.2946	0.89	.1867	1.24	.1075	1.59	.0559	1.94	.0262	2.29	.0110	2.64	.0041
0.20	.4207	0.55	.2912	0.90	.1841	1.25	.1056	1.60	.0548	1.95	.0256	2.30	.0107	2.65	.0040
0.21	.4168	0.56	.2877	0.91	.1814	1.26	.1038	1.61	.0537	1.96	.0250	2.31	.0104	2.66	.0039
0.22	.4129	0.57	.2843	0.92	.1788	1.27	.1020	1.62	.0526	1.97	.0244	2.32	.0102	2.67	.0038
0.23	.4090	0.58	.2810	0.93	.1762	1.28	.1003	1.63	.0516	1.98	.0239	2.33	.0099	2.68	.0037
0.24	.4052	0.59	.2776	0.94	.1736	1.29	.0985	1.64	.0505	1.99	.0233	2.34	.0096	2.69	.0036
0.25	.4013	0.60	.2743	0.95	.1711	1.30	.0963	1.65	.0495	2.00	.0228	2.35	.0094	2.70	.0035
0.26	.3974	0.61	.2709	0.96	.1685	1.31	.0951	1.66	.0485	2.01	.0222	2.36	.0091	2.71	.0034
0.27	.3936	0.62	.2676	0.97	.1660	1.32	.0940	1.67	.0475	2.02	.0217	2.37	.0089	2.72	.0033
0.28	.3897	0.63	.2643	0.98	.1635	1.33	.0938	1.68	.0465	2.03	.0212	2.38	.0087	2.73	.0032
0.29	.3859	0.64	.2611	0.99	.1611	1.34	.0930	1.69	.0455	2.04	.0207	2.39	.0084	2.74	.0031
0.30	.3821	0.65	.2578	1.00	.1587	1.35	.0885	1.70	.0446	2.05	.0202	2.40	.0082	2.75	.0030
0.31	.3783	0.66	.2546	1.01	.1562	1.36	.0869	1.71	.0436	2.06	.0197	2.41	.0080	2.76	.0029
0.32	.3745	0.67	.2514	1.02	.1539	1.37	.0853	1.72	.0427	2.07	.0192	2.42	.0078	2.77	.0028
0.33	.3707	0.68	.2483	1.03	.1515	1.38	.0838	1.73	.0418	2.08	.0188	2.43	.0075	2.78	.0027
0.34	.3669	0.69	.2451	1.04	.1492	1.39	.0823	1.74	.0409	2.09	.0183	2.44	.0073	2.79	.0026

- Valeurs adaptées aux tests unilatéral
- Pour test bilatéral, multiplier la valeur de probabilité par 2

# Choix de l'outil d'évaluation



Quel type de tâche ?

Classification  
ou  
estimation ?

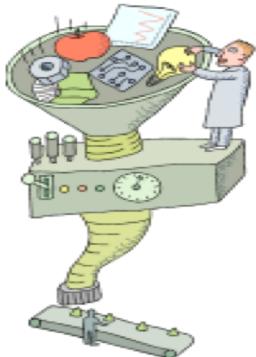
# Outils pour les tâches de classification

---

Sensibilité, spécificité, exactitude,  
valeurs prédictives, rapports de  
vraisemblance, approches ROC

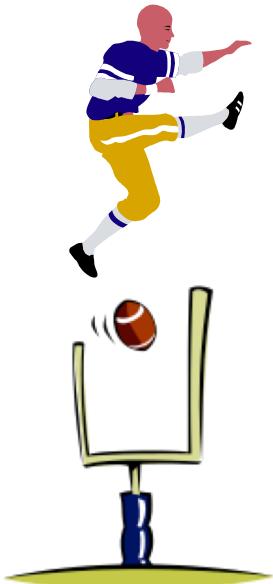


# Tâches de classification : contexte général

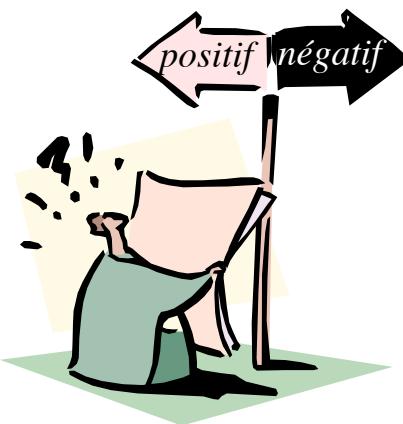
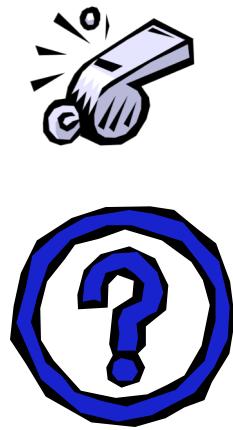


Données divisibles en 2 (ou +) catégories

*e.g., avec et sans anomalie  
(positif ou négatif)  
anomalie A versus anomalie B*



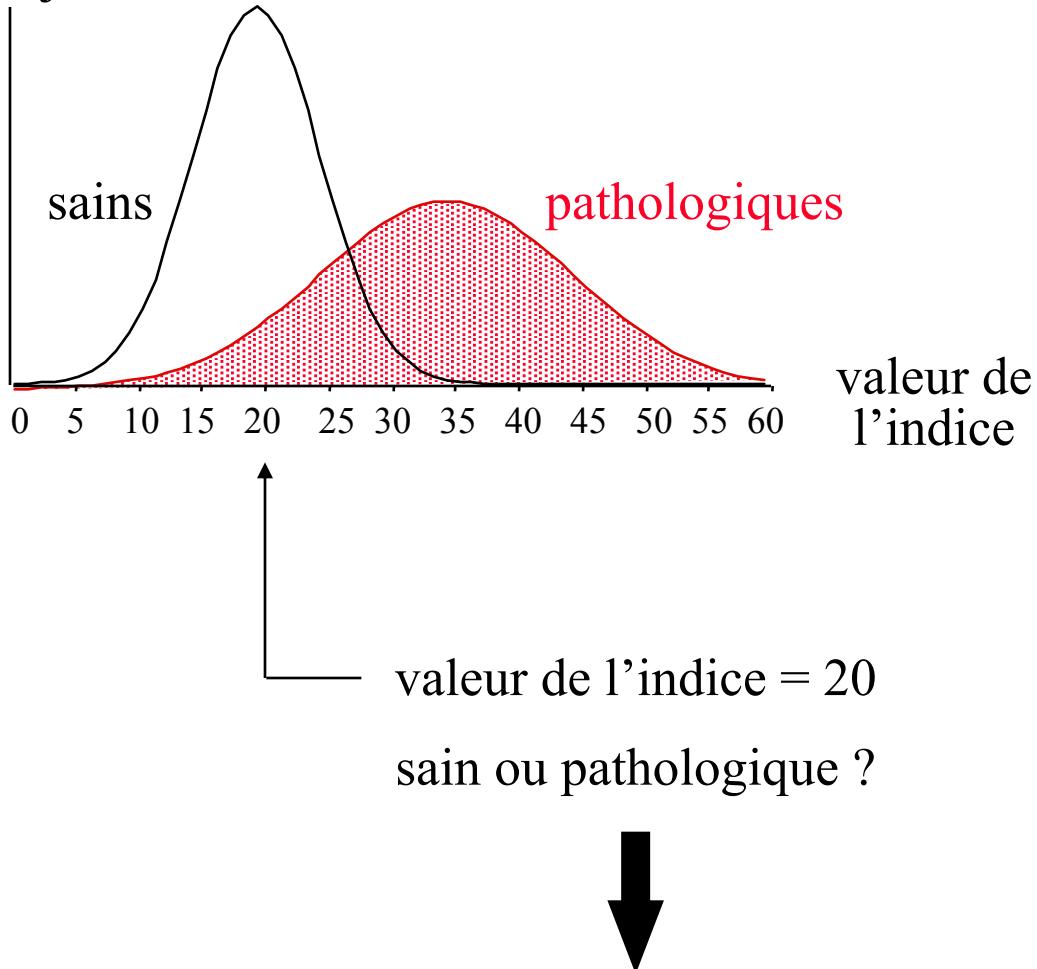
Tâche : déterminer la catégorie à laquelle appartient chaque élément du jeu de données



# Situation concrète

- Recouvrement des valeurs de l'indice pour les populations avec et sans l'anomalie

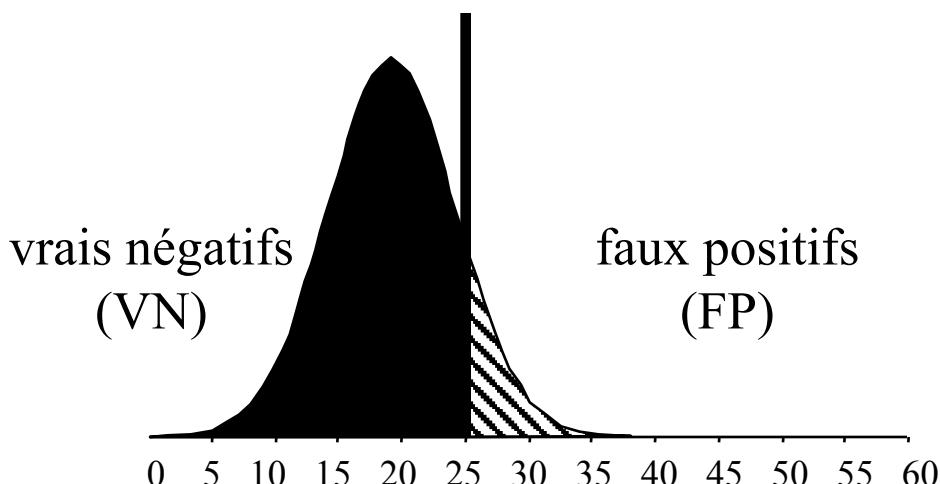
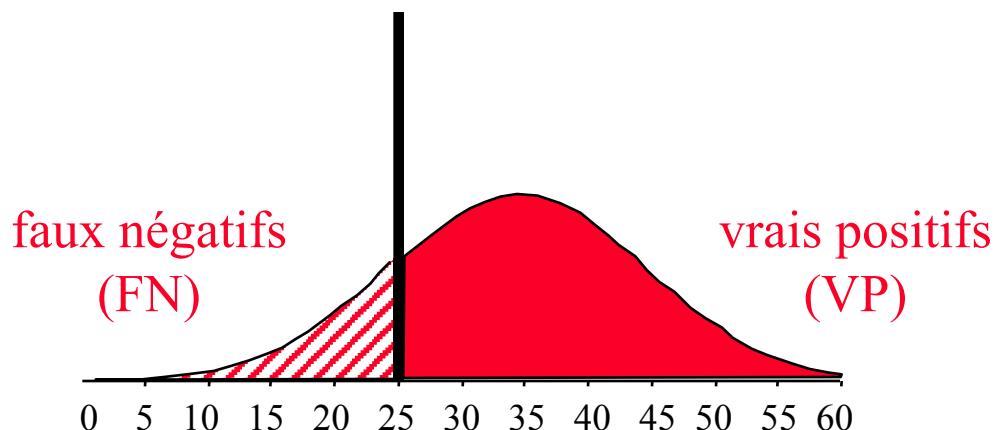
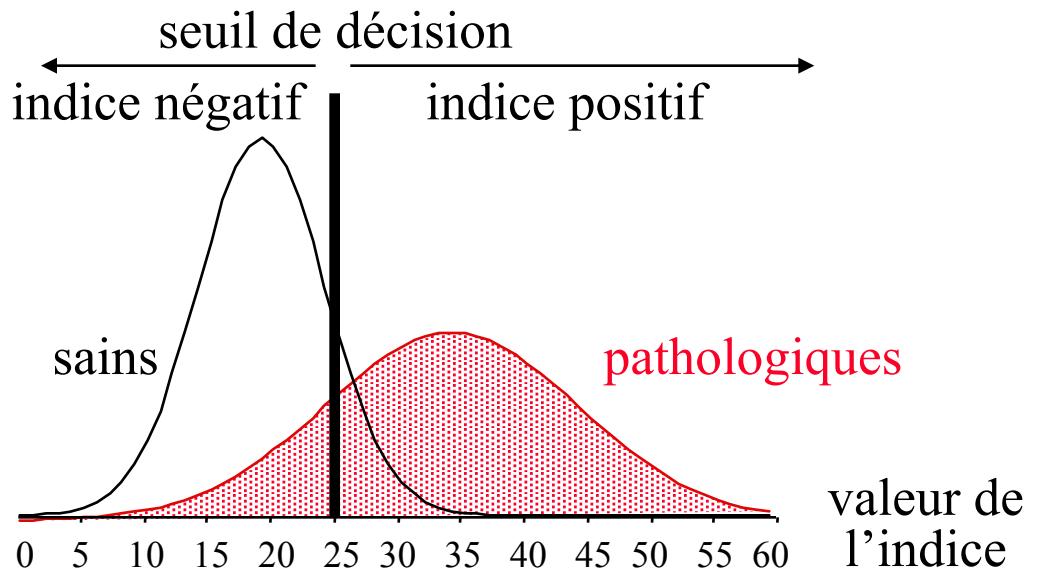
nb de sujets



nécessité de choisir un seuil de décision

→ les performances diagnostiques de l'indice dépendent du seuil de décision

# Seuil de décision et classification des sujets



# Définitions pour l'évaluation diagnostique



INDICE  
négatif  
positif

		PATHOLOGIE absente	présente
INDICE négatif	vrais négatifs VN	faux négatifs FN	
	faux positifs FP	vrais positifs VP	



- Fraction de vrais positifs  $FVP = \frac{VP}{VP + FN} = \text{sensibilité}$
- Fraction de vrais négatifs  $FVN = \frac{VN}{VN + FP} = \text{spécificité}$
- Fraction de faux positifs  $FFP = \frac{FP}{FP + VN} = 1 - FVN$
- Fraction de faux négatifs  $FFN = \frac{FN}{FN + VP} = 1 - FVP$
- Valeur prédictive positive  $VPP = \frac{VP}{VP + FP}$
- Valeur prédictive négative  $VPN = \frac{VN}{VN + FN}$

# Exemple

Population de 1200 sujets

		PATHOLOGIE	
		absente	présente
INDICE négatif	absente	vrais négatifs 900	faux négatifs 60
	présente	faux positifs 100	vrais positifs 140

- sensibilité ?

- spécificité ?

- FFP ?

- FFN ?

- VPP ?

- VPN ?

- Prévalence ?

# Solution

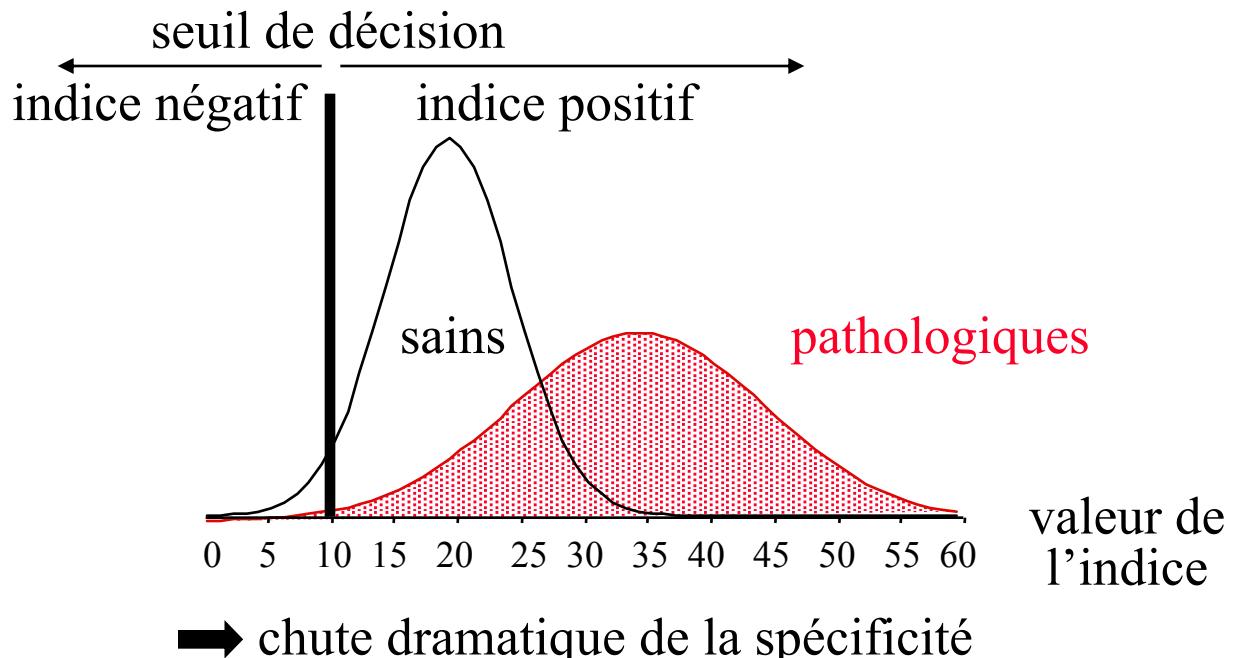
Population de 1200 sujets

		PATHOLOGIE	
		absente	présente
INDICE négatif	absente	vrais négatifs 900	faux négatifs 60
	présente	faux positifs 100	vrais positifs 140

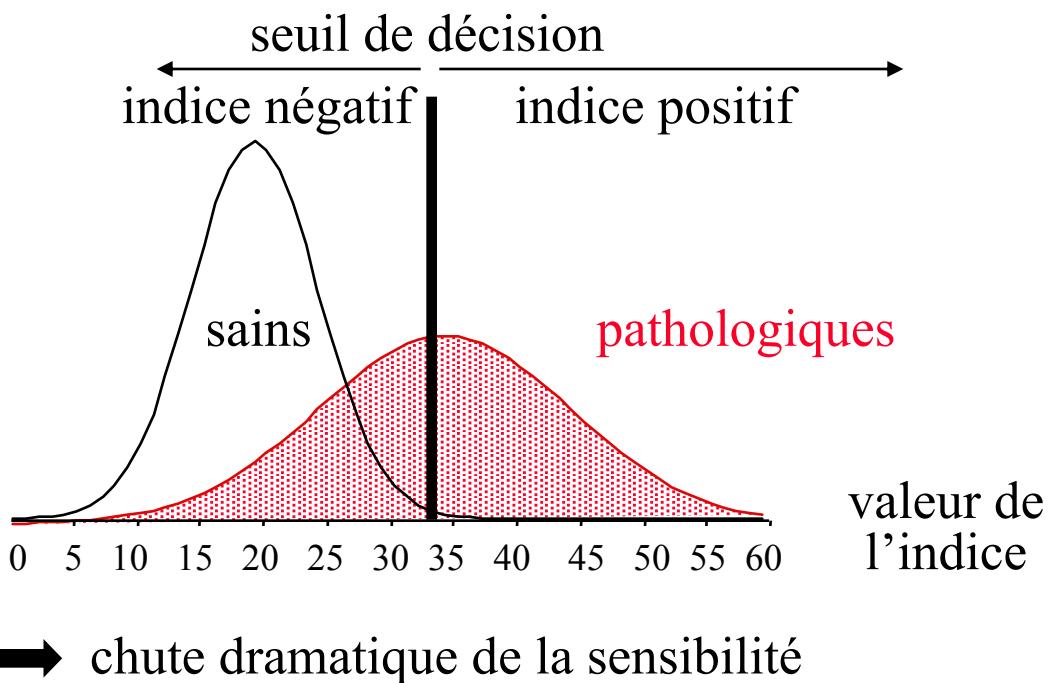
- sensibilité =  $FVP = 140 / 200 = 0.70 = 70\%$
- spécificité =  $FVN = 900 / 1000 = 0.90 = 90\%$
- $FFP = 100 / 1000 = 0.1 = 10\%$
- $FFN = 60 / 200 = 0.3 = 30\%$
- $VPP = 140 / 240 = 0.58 = 58\%$
- $VPN = 900 / 960 = 0.94 = 94\%$
- Prévalence =  $200 / 1200 = 0.17 = 17\%$

# Sensibilité, spécificité et seuil de décision

- Comment choisir le seuil de décision pour maximiser la sensibilité d'un indice ?

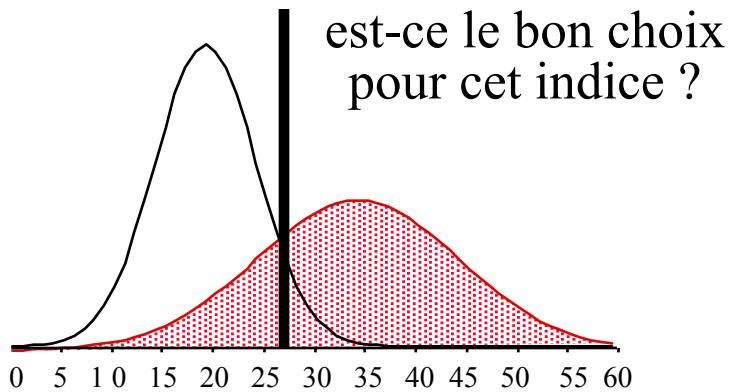


- Comment choisir le seuil de décision pour maximiser la spécificité d'un indice ?



# Insuffisance des sensibilités et spécificités

- Valeurs dépendant du choix du seuil de décision



- Comment comparer deux indices A et B ?

Le seuil de décision optimal pour un indice n'est pas nécessairement optimal pour l'autre

- Comment comparer deux indices A et B ?

indice A : sensibilité = 80% spécificité = 65%  
indice B : sensibilité = 90% spécificité = 50%

Quel est le meilleur ?

- Contrôle du seuil de décision

indice = valeur numérique  
→ aisément contrôlable

indice = diagnostic posé à partir d'une image  
→ seuil subjectif, difficilement contrôlable

# Indice composite ?

- Combiner sensibilité et spécificité en un nouveau critère ?

$$\text{exactitude} = \frac{VP + VN}{VP + VN + FP + FN}$$

		PATHOLOGIE	
		absente	présente
INDICE A	négatif	vrais négatifs 900	faux négatifs <b>60</b>
	positif	faux positifs <b>100</b>	vrais positifs 140
INDICE B	négatif	vrais négatifs 960	faux négatifs <b>120</b>
	positif	faux positifs <b>40</b>	vrais positifs 80

sensibilité = 70%  
spécificité = 90%  
exactitude = 87%

sensibilité = 40%  
spécificité = 96%  
exactitude = 87%

→ même exactitude pour des performances différentes

- peu affectée par les valeurs de sensibilité dans certains cas (faible ou forte prévalence) :

e.g., si prévalence = 5% et tous les cas diagnostiqués négatifs

→ exactitude = 95% !

# Valeurs prédictives

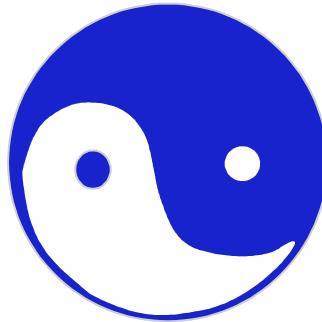
---

- 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

# Applications



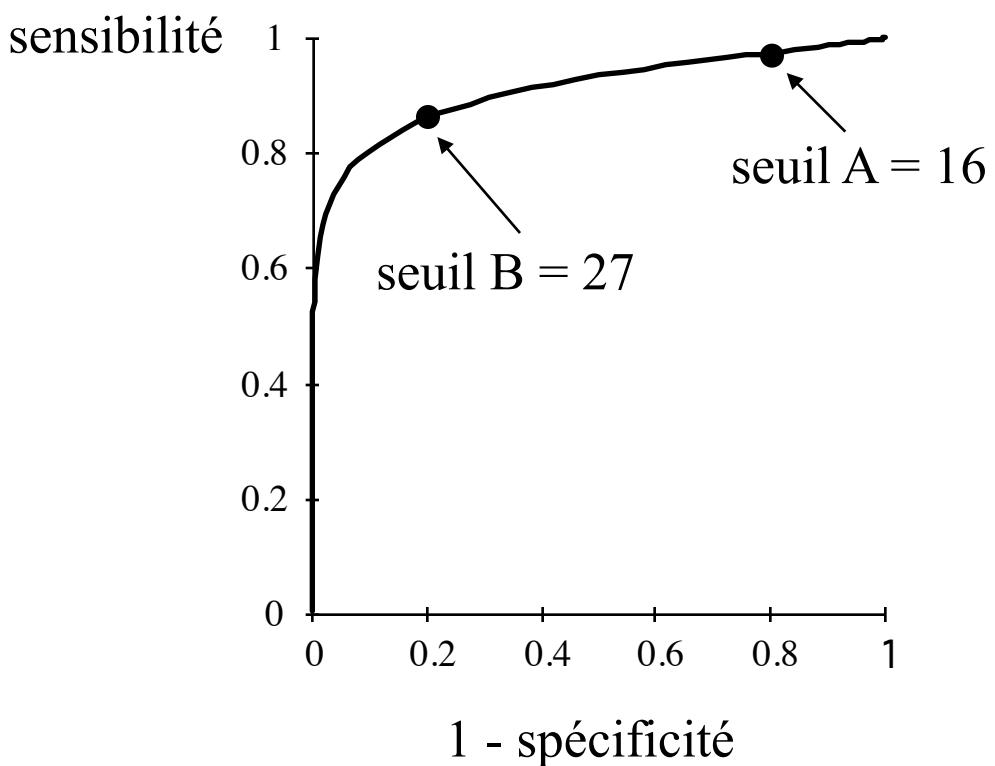
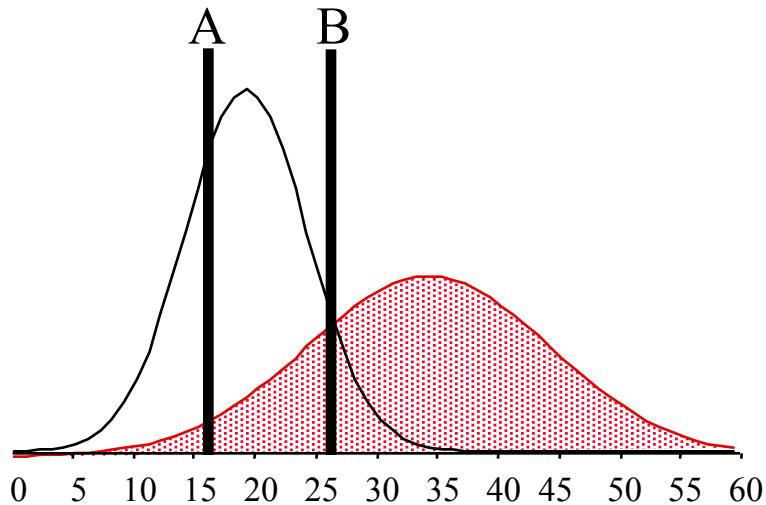
Si le système produit un résultat binaire, les 3 approches présentées précédemment sont les seules possibles avec :

- 
- 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

# Quelle solution ? Courbes ROC

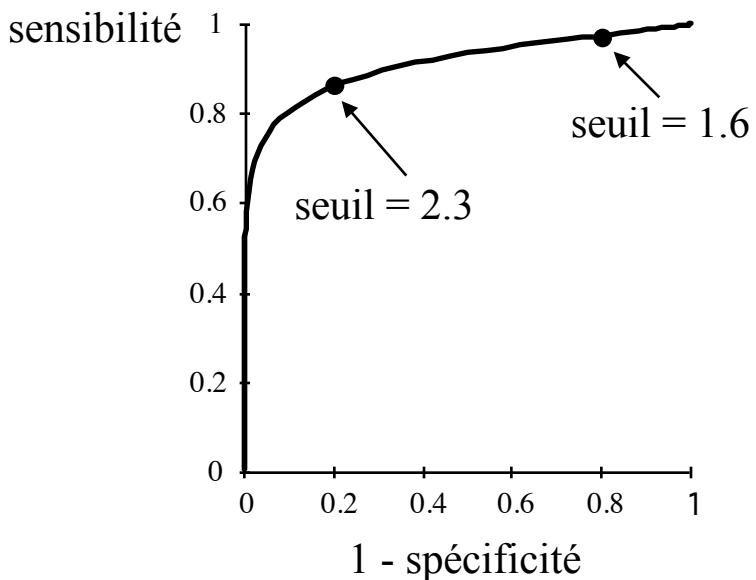
- ROC : Receiver Operating Characteristics
- Prend en compte la variabilité des performances de classification en fonction de la valeur du seuil de décision



# Construction d'une courbe ROC (1)

- Indice correspondant à des valeurs numériques

individu	1	2	3	4	5	....	N
nature de l'individu (S ou P)	S	P	S	S	P	....	P
valeur de l'indice	1.8	2.3	2.1	1.7	2.1		1.9



➡ variation quasi-continue du seuil de décision



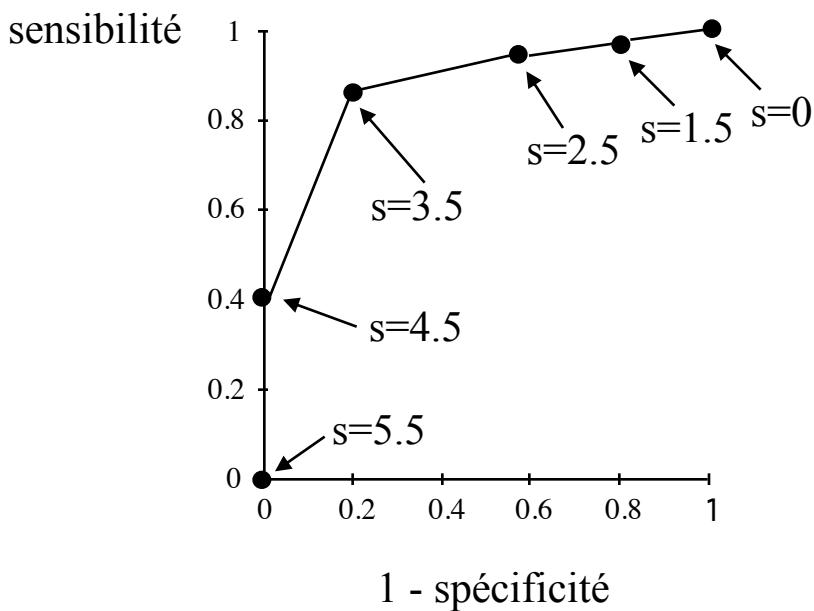
# Construction d'une courbe ROC (2)

- Indice correspondant à une interprétation subjective

image	1	2	3	4	5	....	N
nature de l'image (S ou P)	S	P	S	S	P	....	P
score attribué par un observateur	3	4	2	1	3		2

échelle de scores :

- 1. lésion certainement absente ..... → seuil 0
- 2. lésion probablement absente ..... → seuil 1.5
- 3. lésion possiblement présente ..... → seuil 2.5
- 4. lésion probablement présente ..... → seuil 3.5
- 5. lésion certainement présente ..... → seuil 4.5
- ..... → seuil 5.5

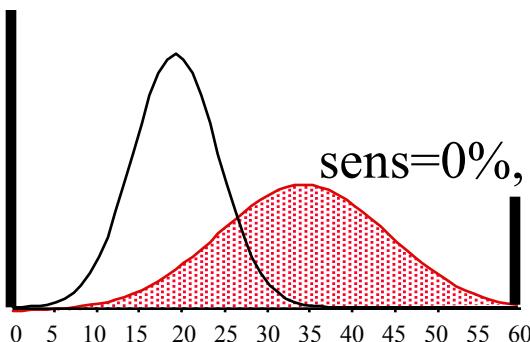


→ variation discrète du seuil de décision

# Caractéristiques d'une courbe ROC

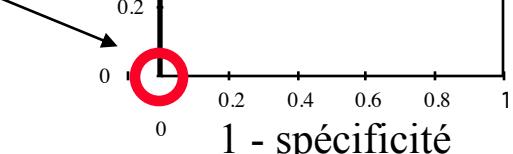
- Passe par les points  $(0,0)$  et  $(1,1)$

sens=100%, spéc=0%



sensibilité

sens=0%, spéc=100%

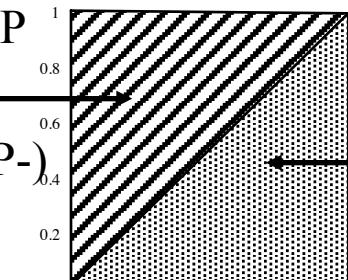


- Courbe ROC au dessus de la diagonale

sensibilité = FVP

FVP > FFP

Prob(T+|P+) > Prob (T+|P-)

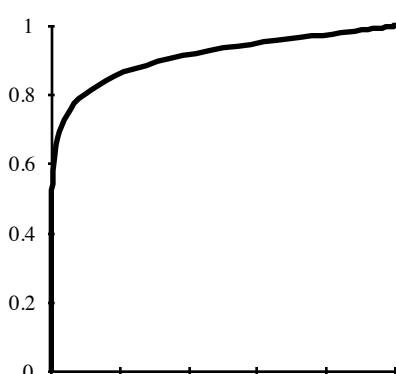


FVP < FFP  
Prob(T+|P+) < Prob (T+|P-)

1 - spécificité = FFP

- Pente décroissante de  $(0,0)$  à  $(1,1)$

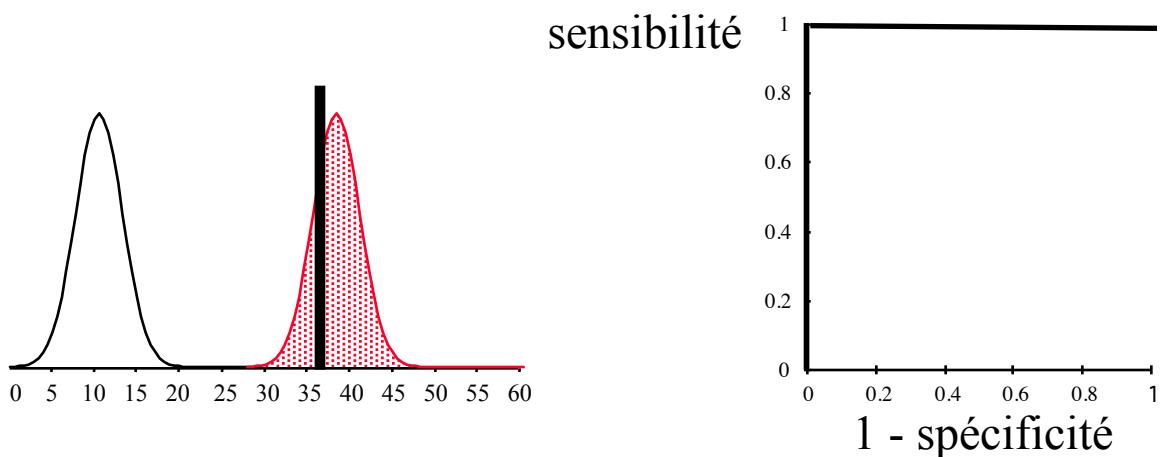
sensibilité



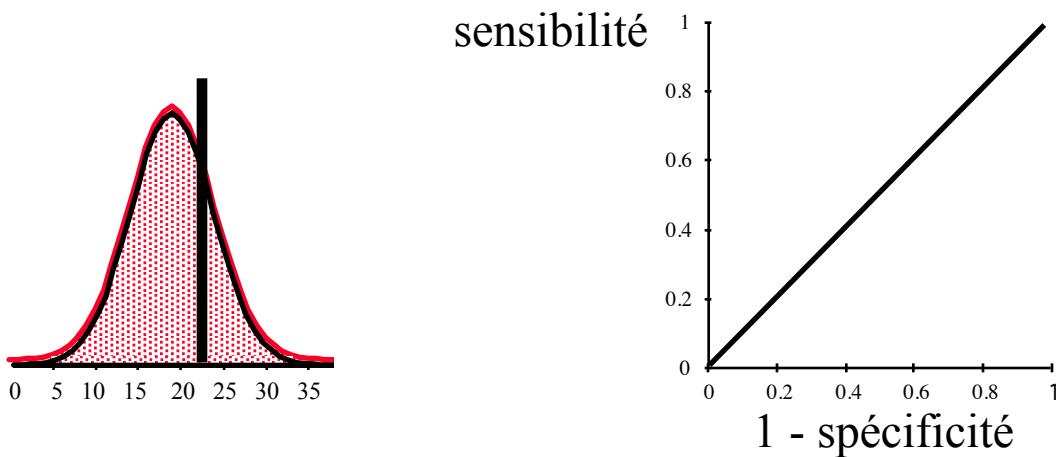
1 - spécificité

# Interprétation d'une courbe ROC

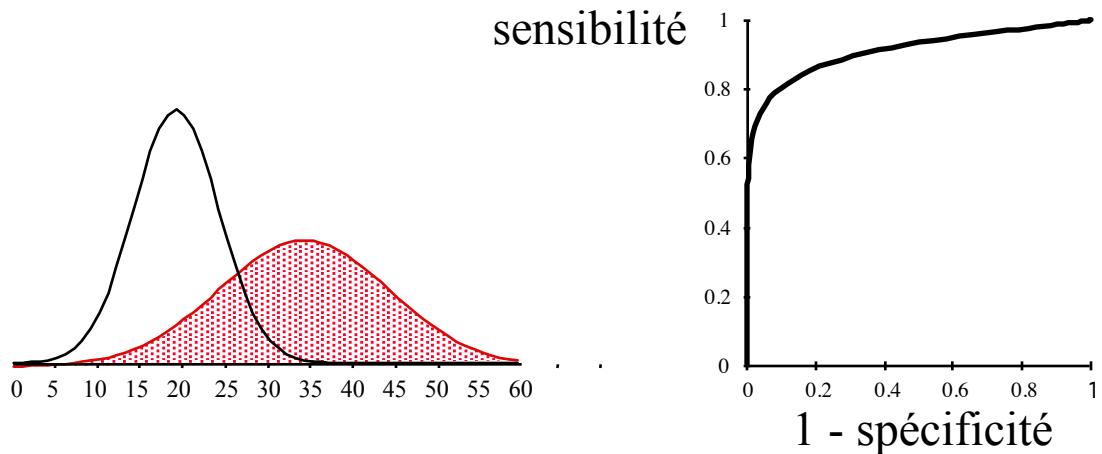
- Indice idéal



- Indice non informatif

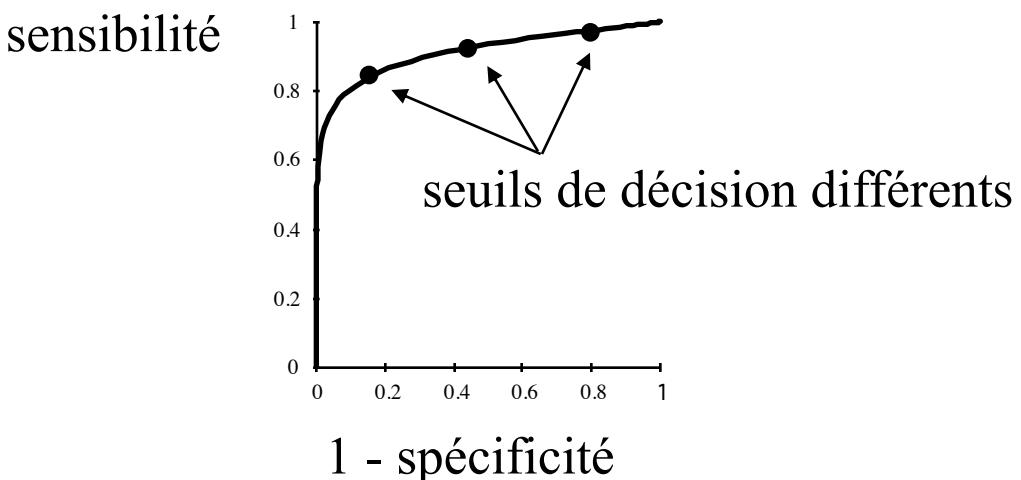


- Indice courant



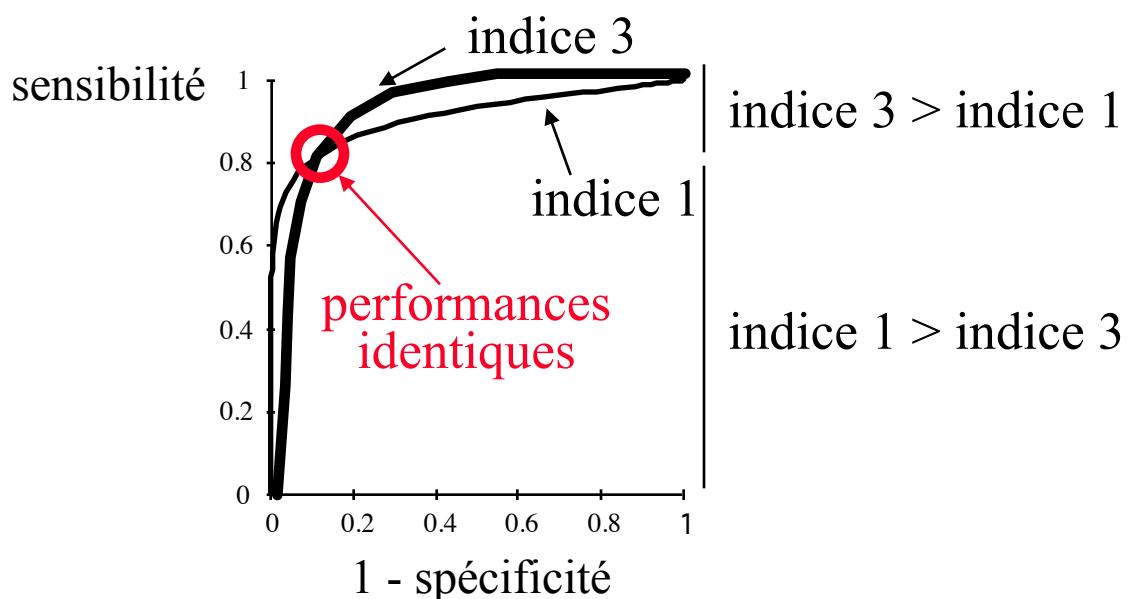
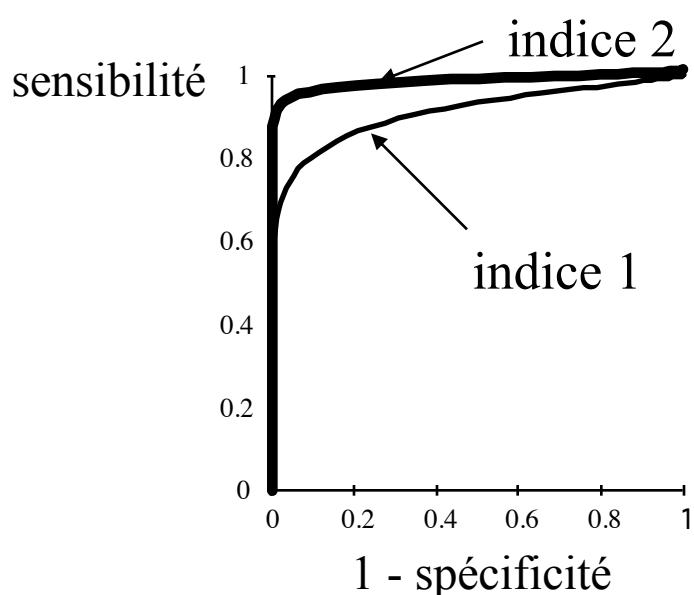
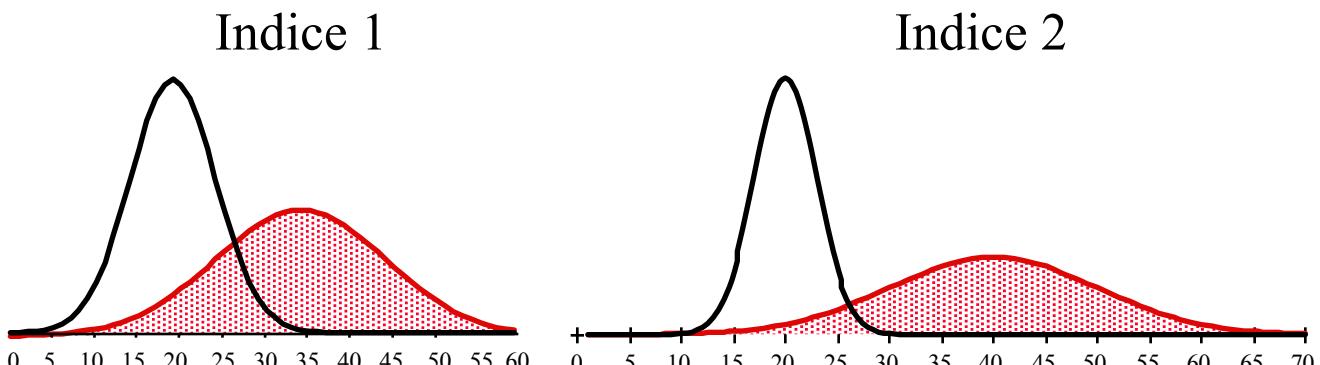
# Propriétés d'une courbe ROC

- Ne dépend pas de la valeur du seuil de décision



- Ne dépend pas de la prévalence
- Pour un indice quantitatif : caractérise les performances de l'indice
- Pour un diagnostic posé à partir de l'observation d'une image : caractérise la combinaison (image ; observateur)  
→ intérêt des observateurs multiples

# Comparaison subjective de deux courbes ROC

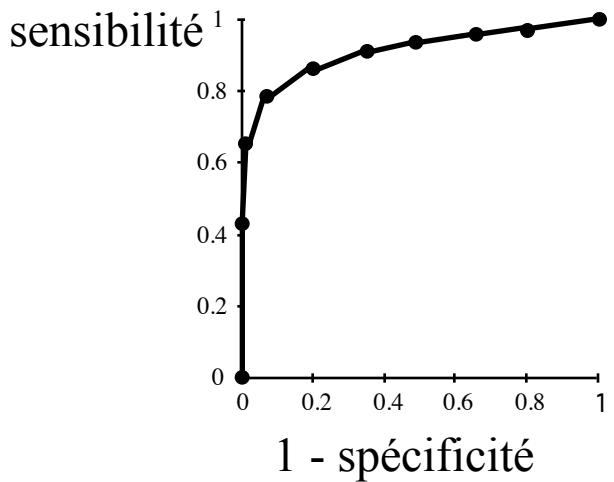
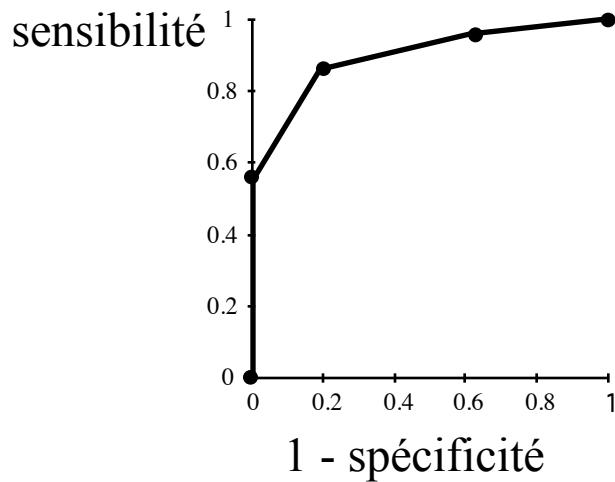


# Tracé d'une courbe ROC : approche non paramétrique

- Approche ne reposant pas sur un modèle statistique sous-jacent

ligne brisée reliant les points de mesure

→ aucun modèle sous-jacent



→ courbe d'autant plus lisse que le nombre de points de mesure est élevé

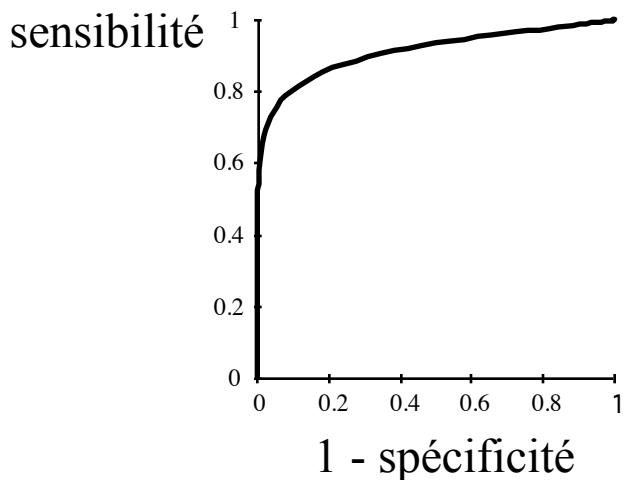
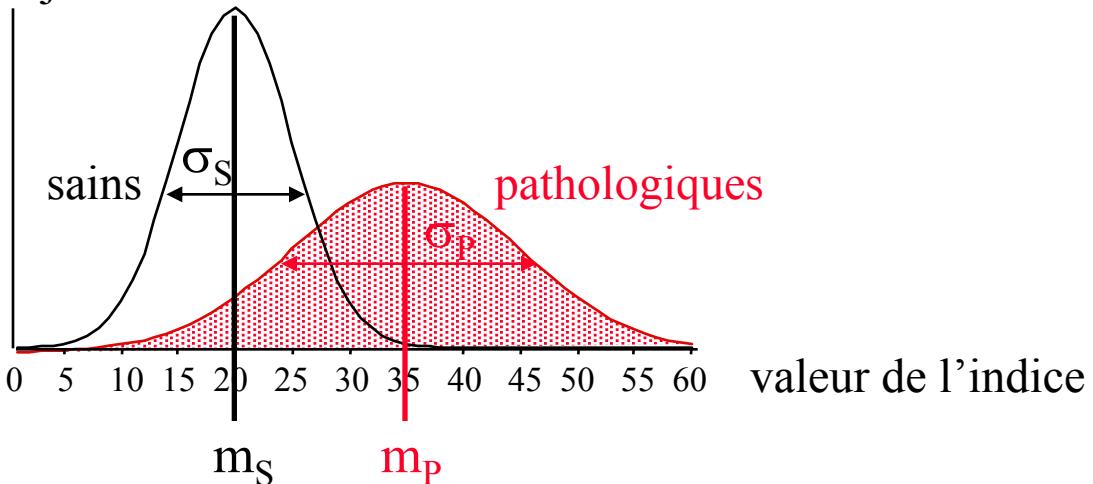
→ courbe caractérisée par l'ensemble des triplets (seuil, 1-spécificité, sensibilité)

# Tracé d'une courbe : approche paramétrique

- Modèle binormal sous-jacent

Hypothèse : chaque population suit une loi normale

nb de sujets



$$A = \frac{m_P - m_S}{\sigma_S}$$

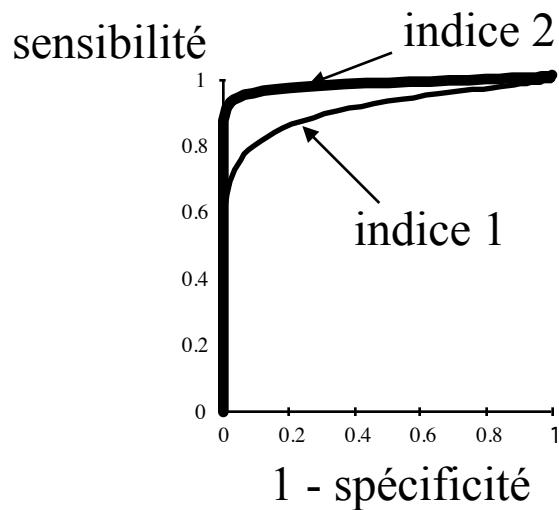


$$B = \frac{\sigma_S}{\sigma_P}$$

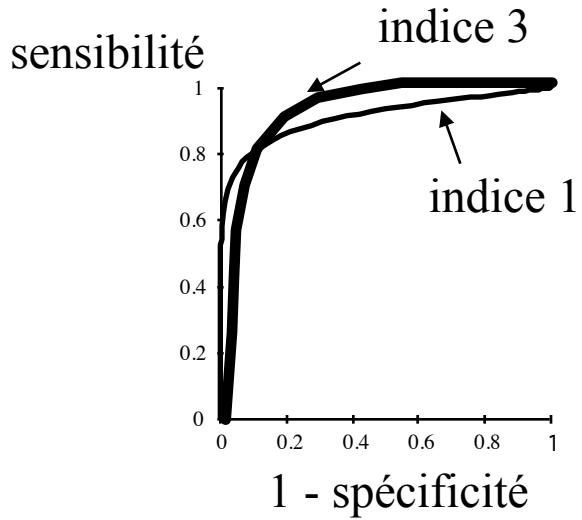
→ détermination des paramètres A et B à partir des triplets (seuil, 1- spécificité, sensibilité) par ajustement maximisant la vraisemblance (i.e., la probabilité que les observations soient expliquées par le modèle)  
e.g., programme ROCKIT

# Comparaison de 2 courbes ROC

- Comparaison subjective (visuelle)



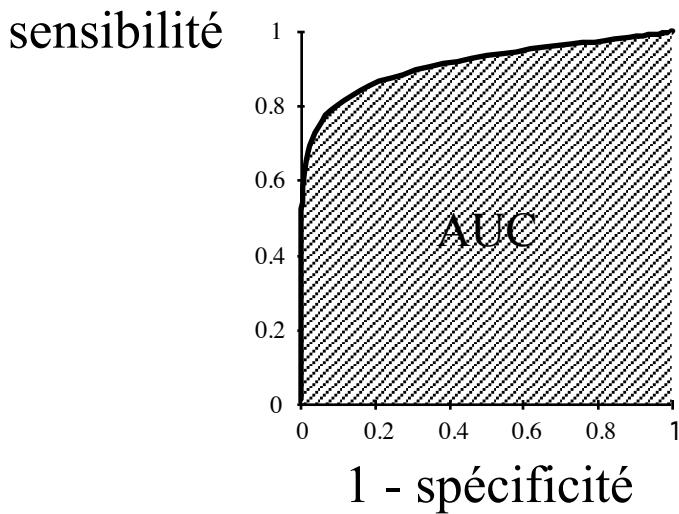
indice 2 > indice 1



indice 3 > indice 1  
ou indice 3  $\leq$  indice

- Comparaison objective : réalisation d'un test d'hypothèse, le plus souvent à partir du calcul de l'aire sous la courbe ROC (AUC)

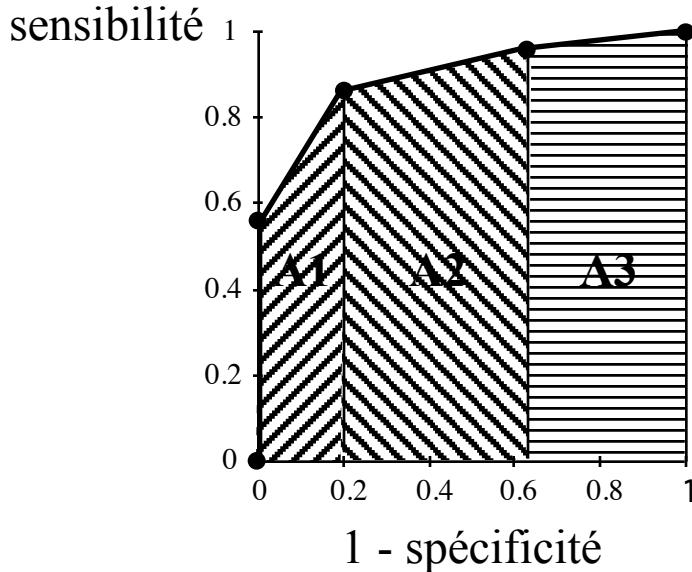
# Aire sous la courbe ROC



AUC = probabilité d'identifier correctement l'image avec anomalie quand une image avec et une image sans anomalie sont présentées simultanément à un observateur

# Estimation de AUC : approche non paramétrique

- Approche trapézoïdale



$$W = \widehat{AUC} = A1 + A2 + A3$$

- $W = \text{statistique de Wilcoxon} = \frac{1}{n_{P+} \cdot n_{P-}} \sum_1^{n_{P+}} \sum_1^{n_{P-}} S(t_{P+}, t_{P-})$
- avec  $S(t_{P+}, t_{P-}) = \begin{cases} 1 & \text{si } t_{P+} > t_{P-} \\ 0.5 & \text{si } t_{P+} = t_{P-} \\ 0 & \text{si } t_{P+} < t_{P-} \end{cases}$

- $W = AUC$  si l'échelle de cotation présente un grain infini

$$\bullet SE(W) = \sqrt{\frac{W(1-W) + (n_{P+} - 1)(Q_1 - W^2) + (n_{P-} - 1)(Q_2 - W^2)}{n_{P+} \cdot n_{P-}}}$$

avec  $Q_1 = \text{Prob}(t_{P+} > t_{P-} \text{ et } t'_{P+} > t_{P-})$   
 $Q_2 = \text{Prob}(t_{P+} > t_{P-} \text{ et } t_{P+} > t'_{P-})$

# Estimation de AUC : approche paramétrique

- Modèle binormal :

AUC = Az estimé à partir de l'ajustement des paramètres du modèle par maximisation de la vraisemblance

estimation conjointe de SE(Az)

Condition 1: rating

Total number of actually-negative cases = 58.  
Total number of actually-positive cases = 51.

Data effectively collected in 5 categories.

Category 5 represents the strongest evidence of positivity.  
(e.g., that the disease is present)

Response Data:

Category	1	2	3	4	5
Actually-Negative Cases	33	6	6	11	2
Actually-Positive Cases	3	2	2	11	33

[...]

=====

Final Estimates of the Binormal ROC Parameters

=====

Binormal Parameters and Area Under the Estimated ROC :

a	=	1.6568
b	=	.7130
Area (Az)	=	.9113

Estimated Standard Errors of these Values:

Std. Err. (a)	=	.3121
Std. Err. (b)	=	.2162
Std. Err. (Az)	=	.0296

# Approche non paramétrique vs paramétrique

Scores t attribués	1	2	3	4	5
P-	33	6	6	11	2
P+	3	2	2	11	33

- Approche non paramétrique

$$W = 0.893$$

$$SE(W) = 0.032$$

- Approche paramétrique

=====

Final Estimates of the Binormal ROC Parameters

=====

Area Under the Estimated ROC :

$$Area (Az) = .9113$$

Estimated Standard Error :

$$Std. Err. (Az) = .0296$$

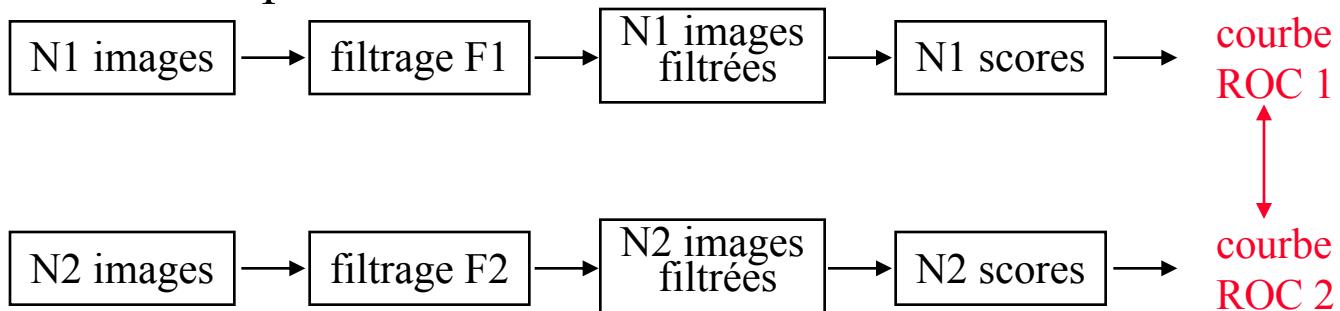
- $W < Az$  car échelle de cotation discrète
- $SE(W) > SE(Az)$ , i.e.  $SE(W)$  conservatif

# Comparaison de 2 courbes ROC

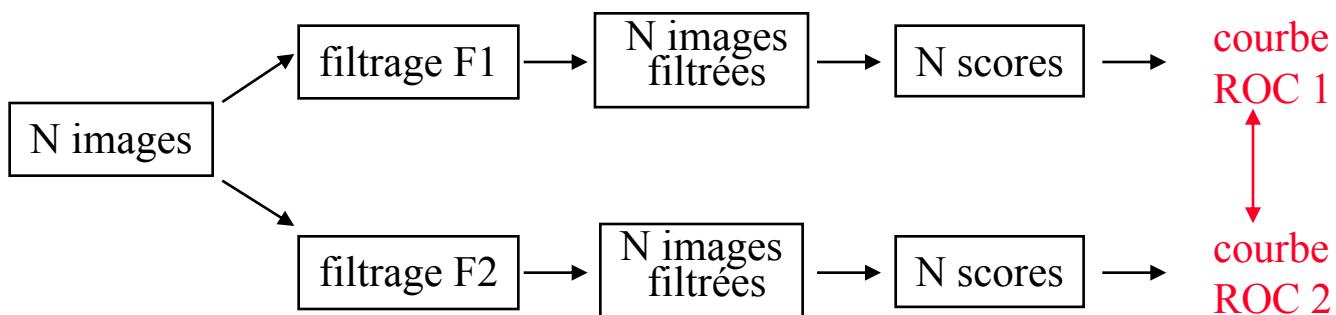
		<u>APPROCHE</u>
<u>COURBES ROC</u>	non paramétrique	paramétrique
indépendantes	test z entre $W_1$ et $W_2$	test z entre $Az_1$ et $Az_2$
appariées	test z intégrant la corrélation entre $W_1$ et $W_2$	test z apparié entre $Az_1$ et $Az_2$

## Comparaison de deux méthodes de filtrage F1 et F2

### Indépendance



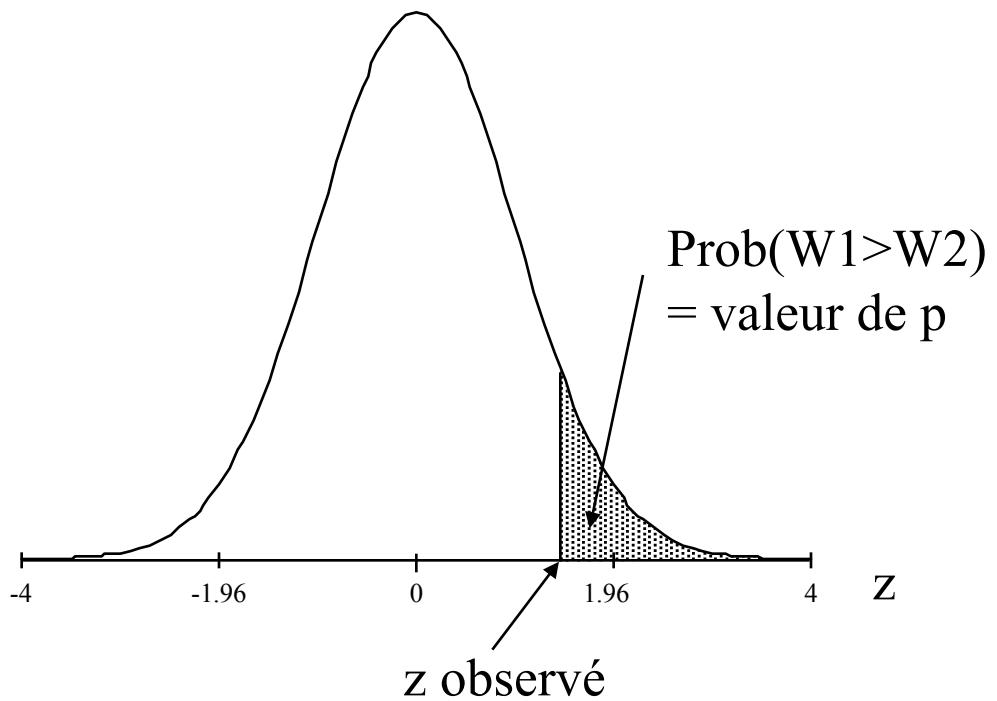
### Appariement



# Cas indépendant - non paramétrique

$$z = \frac{W_1 - W_2}{\sqrt{SE(W_1)^2 + SE(W_2)^2}}$$

puis référence à une table de loi normale :



p correspond à la probabilité que l'on ait observé les données que l'on a (au travers de la valeur de z calculée) si H<sub>0</sub> est vraie

# Cas apparié - non paramétrique

$$z = \frac{W_1 - W_2}{\sqrt{SE(W_1)^2 + SE(W_2)^2 - 2 r \ SE(W_1)SE(W_2)}}$$

avec  $r$  = corrélation entre  $W_1$  et  $W_2$

- $r = f(r_+, r_-)$  avec
    - $r_+$  : corrélation des scores correspondant aux cas P+
    - $r_-$  : corrélation des scores correspondant aux cas P-
  - $r_+$  et  $r_-$  calculés par :
    - coefficient de corrélation de Pearson
      - si les indices  $t_{P+}$  et  $t_{P-}$  sont des valeurs quantitatives variant continument
    - coefficient tau de Kendall
      - si les indices  $t_{P+}$  et  $t_{P-}$  sont des scores
  - $r$  donné par la table (annexe 1) :
- |                 |       |     | $(W_1+W_2)/2$ |
|-----------------|-------|-----|---------------|
| $(r_+ + r_-)/2$ | 0.700 | ... | 0.975         |
| 0.02            |       |     |               |
| ⋮               |       |     |               |
| 0.90            |       |     |               |

# Cas indépendant - paramétrique

$$z = \frac{Az1 - Az2}{\sqrt{SE(Az1)^2 + SE(Az2)^2}}$$

Hypothèses :

- Sujets choisis aléatoirement parmi la population d'intérêt
- Sujets assignés aléatoirement à l'estimation d'un seul des deux indices t1 ou t2
- Indices 1 et 2 interprétés indépendamment
- Lois binormales sous-jacentes pour chaque indice
  - Maximisation d'une fonction de vraisemblance
  - Estimation de Az1, Az2, SE(Az1) et SE(Az2)
  - Test z

# Cas apparié - paramétrique

---

$$z = \frac{Az1 - Az2}{\sqrt{SE(Az1)^2 + SE(Az2)^2 - 2 \operatorname{Covar}(Az1, Az2)}}$$

Hypothèses :

- Sujets choisis aléatoirement parmi la population d'intérêt
- Indices t1 et t2 calculés systématiquement pour tous les sujets
- Indices 1 et 2 interprétés indépendamment
- lois binormales sous-jacentes pour chaque indice
  - Maximisation d'une fonction de vraisemblance
  - Estimation de Az1, Az2, SE(Az1), SE(Az2), Covar(Az1, Az2)
  - Test z intégrant la corrélation entre les deux ensembles de scores

# Impact de la technique de comparaison

## Sujets sains

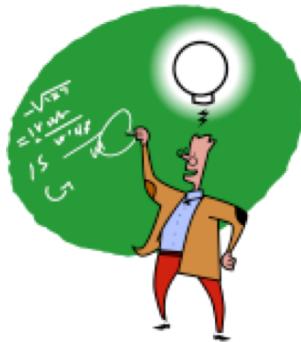
Scores t avec la modalité 1	Scores t avec la modalité 2					
	1	2	3	4	5	6
1	9	3	-	-	-	-
2	17	9	2	-	-	-
3	3	4	1	-	-	-
4	1	2	2	1	-	-
5	1	1	-	2	-	-
6	-	-	-	-	-	-

## Sujets pathologiques

Scores t avec la modalité 1	Scores t avec la modalité 2					
	1	2	3	4	5	6
1	-	-	1	-	-	-
2	1	-	2	-	-	-
3	1	1	1	3	-	-
4	1	1	1	9	1	-
5	-	-	-	7	10	5
6	-	-	-	-	4	5

Test	AUC1	AUC2	r	z
NP, Ind	.883±.033	.932±.026	0	1.181
NP, App	.883±.033	.932±.026	.531	1.550
P, Ind	.895±.030	.941±.025	0	1.165
P, App	.897±.030	.940±.027	.537	1.571

# Potentialités de l'approche ROC



- Peut traiter des cas partiellement appariés
- Possibilité de comparer plus de 2 courbes ROC (plus de deux indices, et/ou plus de 2 observateurs, et/ou plus d'une lecture par observateur)
- Mesures multiples du même paramètre par échantillon : approche Dorfman-Berbaum-Metz (1992)
- Absence de gold standard : travaux de Henkelman (1990) et Beiden (2000)
- Caractérisation ROC des performances de la combinaison de plusieurs paramètres (régression logistique)
- Variantes des courbes ROC pour, par exemple, prendre en compte la localisation des lésions

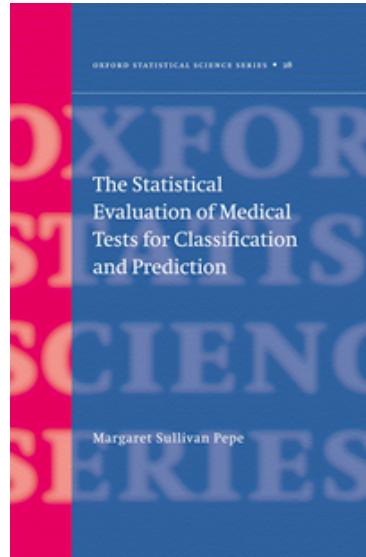
# Considérations pratiques

---

- Nombre de cas à inclure : il existe des méthodes pour les calculer sous certaines conditions
- Sélection des cas : se placer dans des conditions de difficultés intermédiaires
- Ordre de lecture : à soigner sinon biais
- Gold standard imparfait : il existe des méthodes pour prendre en compte un gold standard imparfait
- Echelle de cotation : 5 niveaux suffisent
- Paramétrique ou non paramétrique : non paramétrique si suffisant, paramétrique sinon
- Lectures multiples : améliore la robustesse

Cf annexes

# Pour en savoir plus



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)

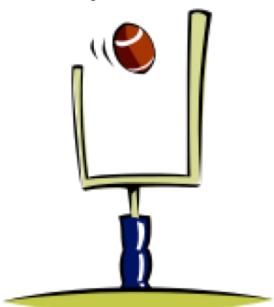
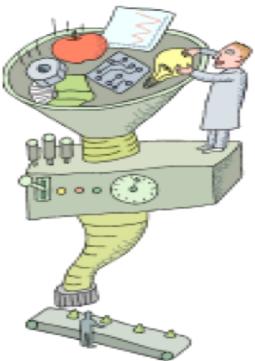
<http://www.mips.ws/>

Bibliographie :

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

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

# Choix de l'outil d'évaluation

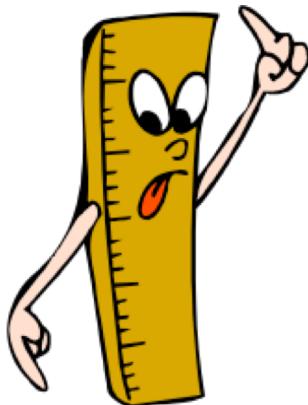


Quel type de tâche ?

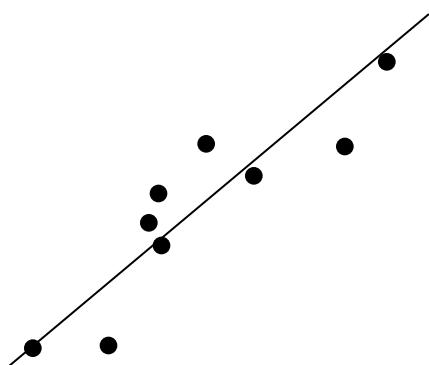
Classification  
ou  
estimation ?

# Outils pour les tâches d'estimation

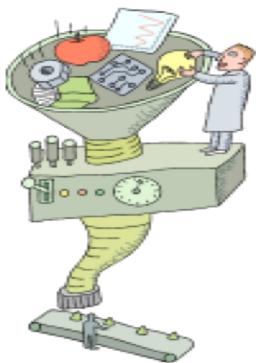
## Biais et variabilité



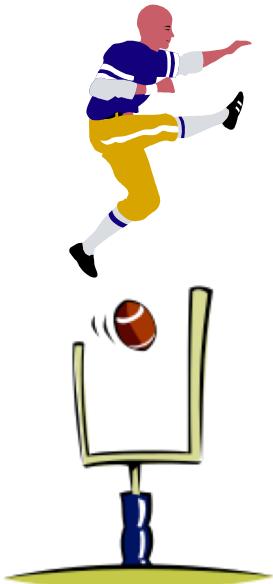
## Corrélation



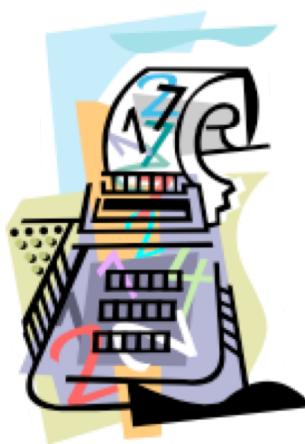
# Tâches d'estimation : contexte général



Données quelconques



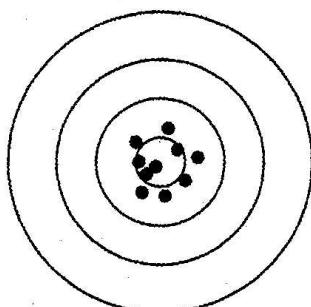
Tâche : extraire une valeur à partir d'un échantillon



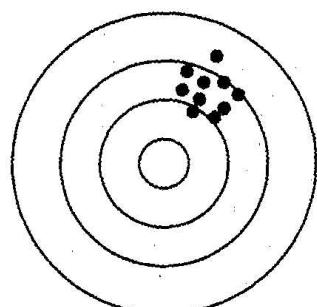
# Evaluation de la qualité d'une mesure

- Contexte : on cherche à caractériser les performances d'une méthode pour estimer un paramètre x
- Ces performances doivent être décrites par 2 grandeurs : le biais d'estimation, et la variabilité
- Biais : caractérise la distance moyenne entre la valeur estimée du paramètre et sa vraie valeur
- Variabilité : caractérise la dispersion des valeurs estimées pour une même valeur vraie

faible variabilité

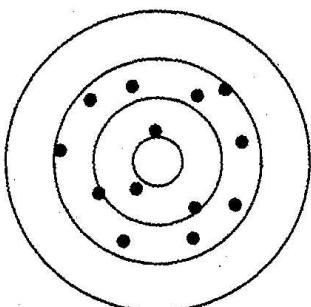


(a)

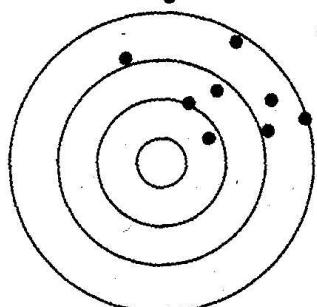


(b)

forte variabilité



(c)



(d)

non biaisé



biaisé

# 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 :

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

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

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

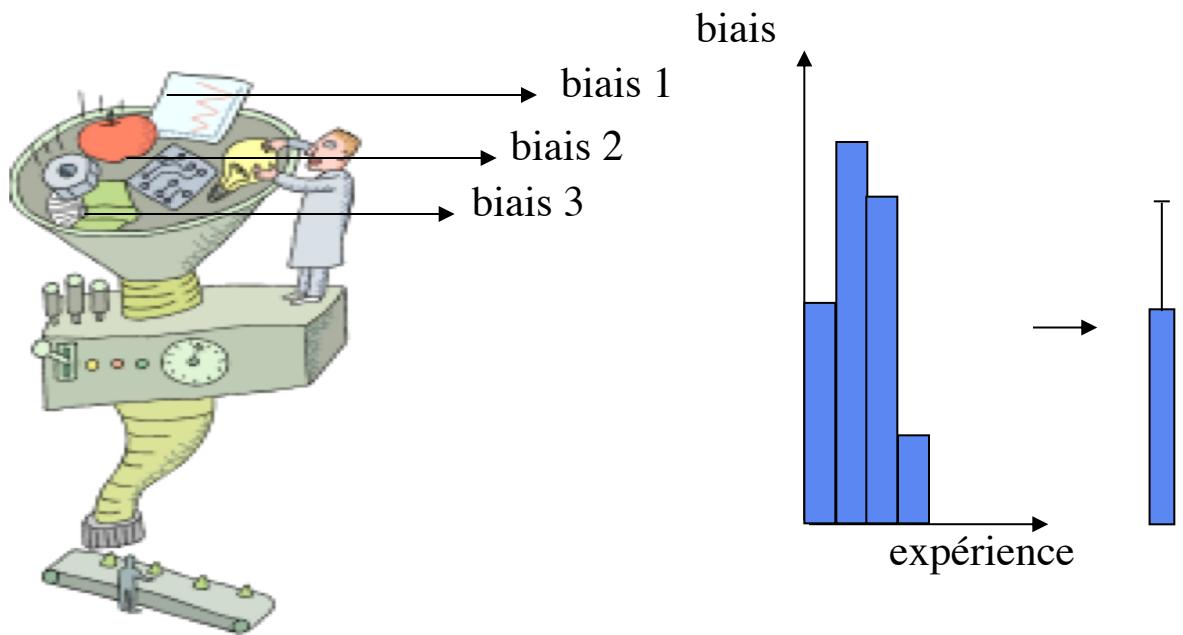
à choisir en fonction du contexte

# Tâches d'estimation : variabilité



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

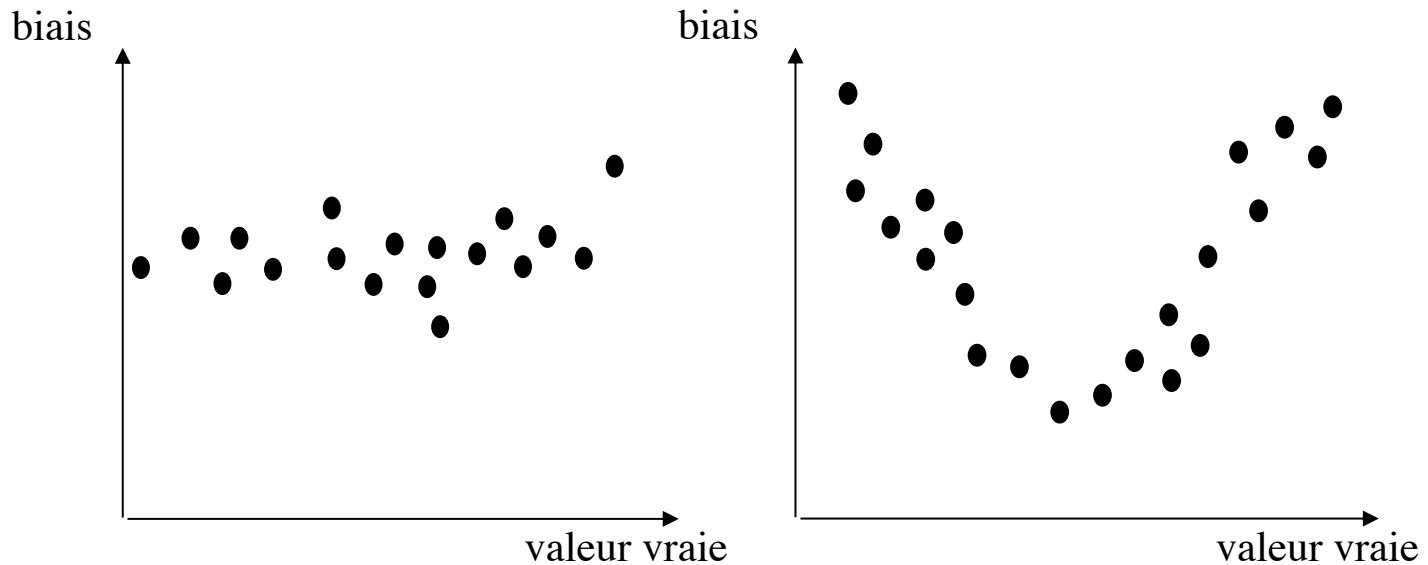
Attention, la variabilité doit être calculée à partir d'un grand nombre d'expériences :



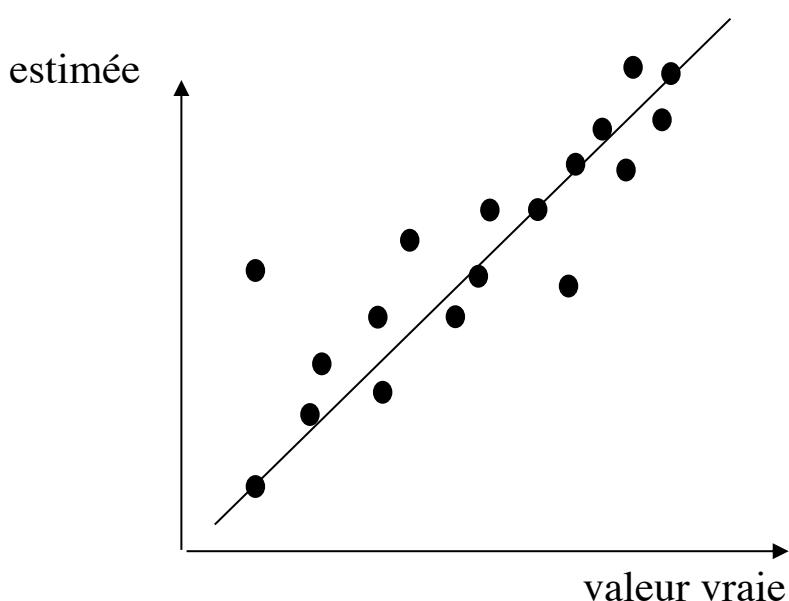
Si loi normale réaliste : écart-type des mesures

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

# Insuffisance des mesures de biais et variabilité



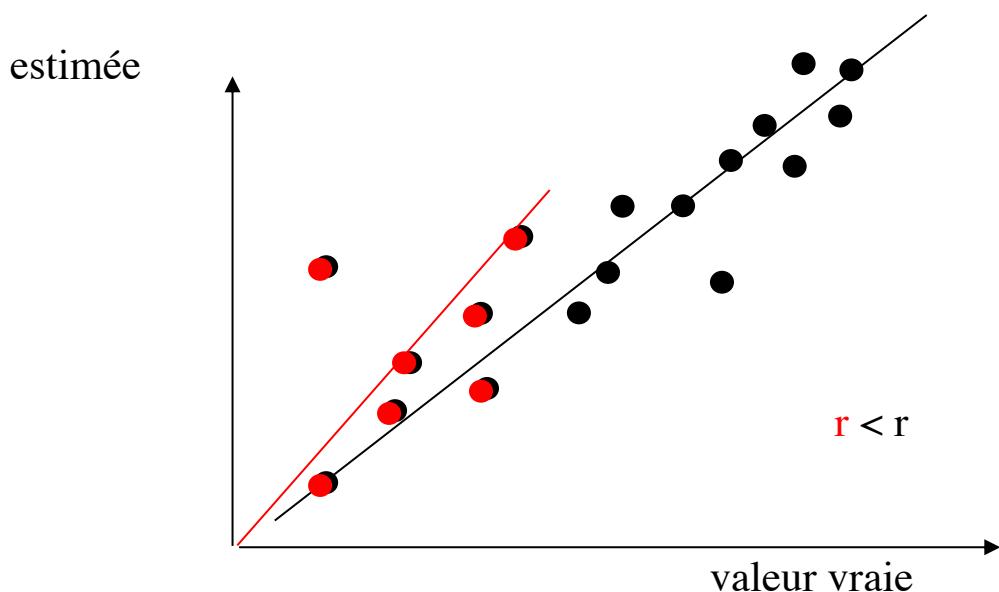
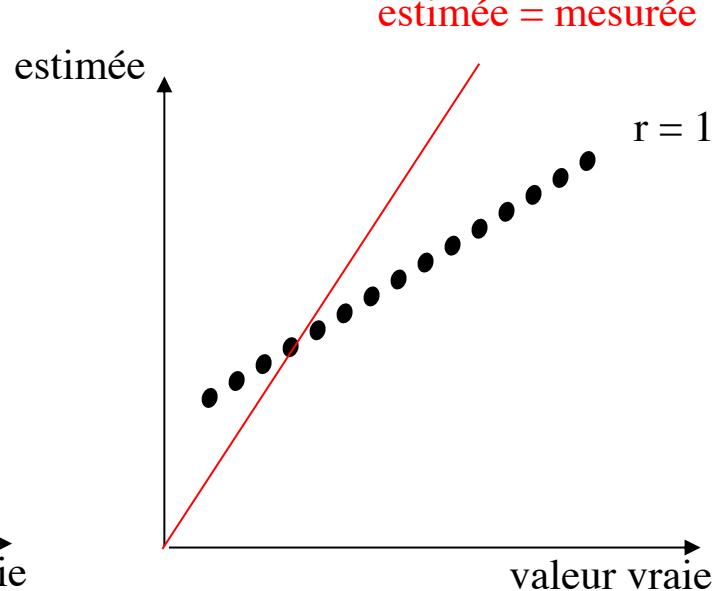
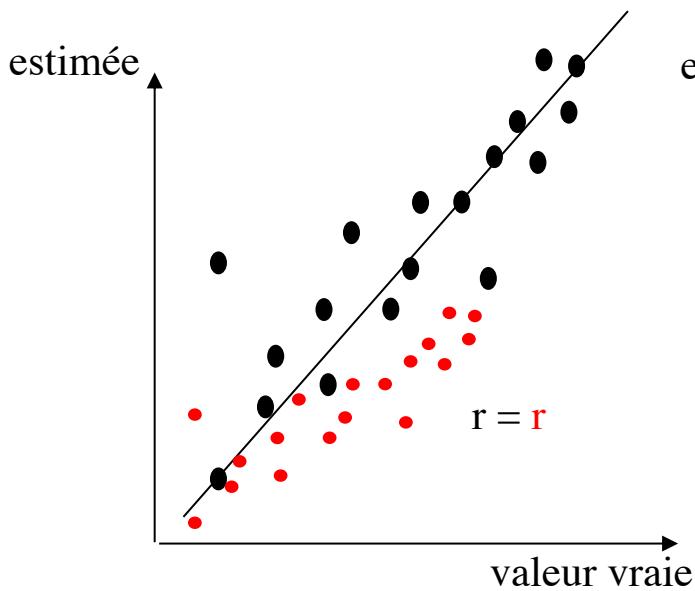
Evaluation plus complète :  
la régression linéaire ?



Attention !

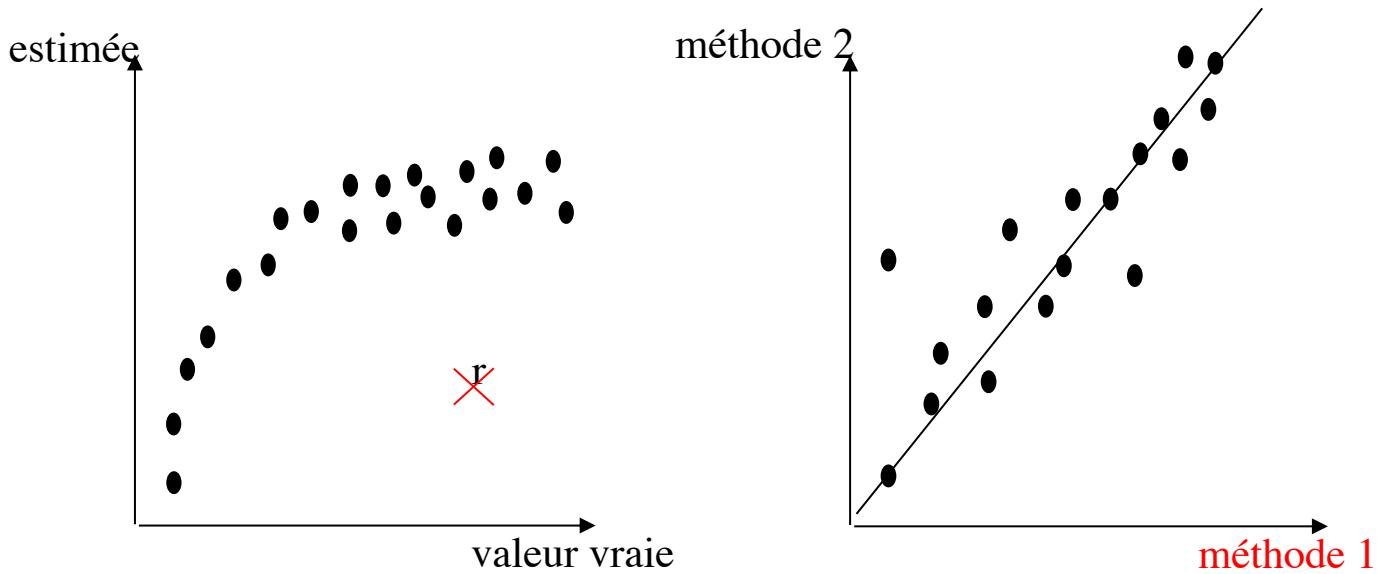
# Limites de la régression linéaire (1)

$r$  : coefficient de corrélation



significativité ( $p < 0.00\dots$ ) de  $r$  n'est pas pertinente !

# Limites de la régression linéaire (2)



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



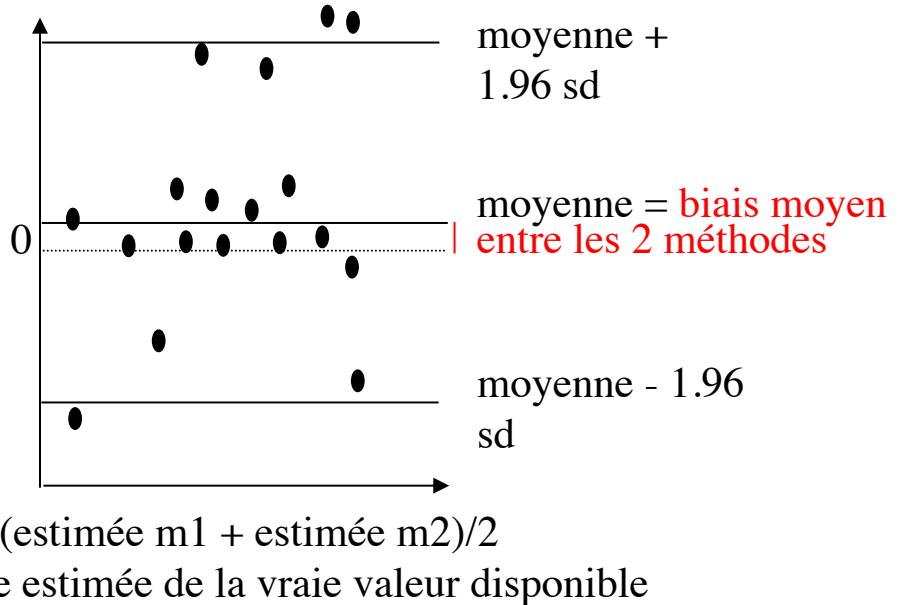
L'analyse de la corrélation linéaire entre estimée et valeur vraie est utile pour une interprétation correcte du biais moyen et de sa variabilité.

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

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

Évaluation de l'accord entre deux méthodes

(estimée m1 - estimée m2)



$(\text{estimée m1} + \text{estimée m2})/2$   
meilleure estimée de la vraie valeur disponible

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

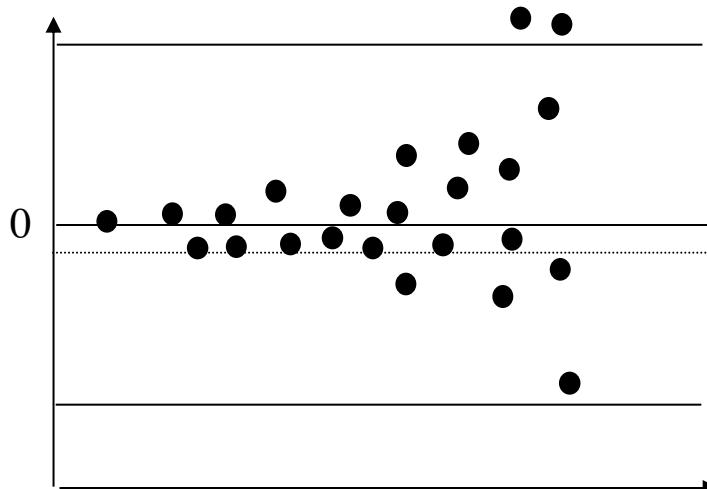


95% des différences sont comprises dans l'intervalle [moyenne - 1.96 sd ; moyenne + 1.96 sd]

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 !

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

estimée m1 - estimée m2



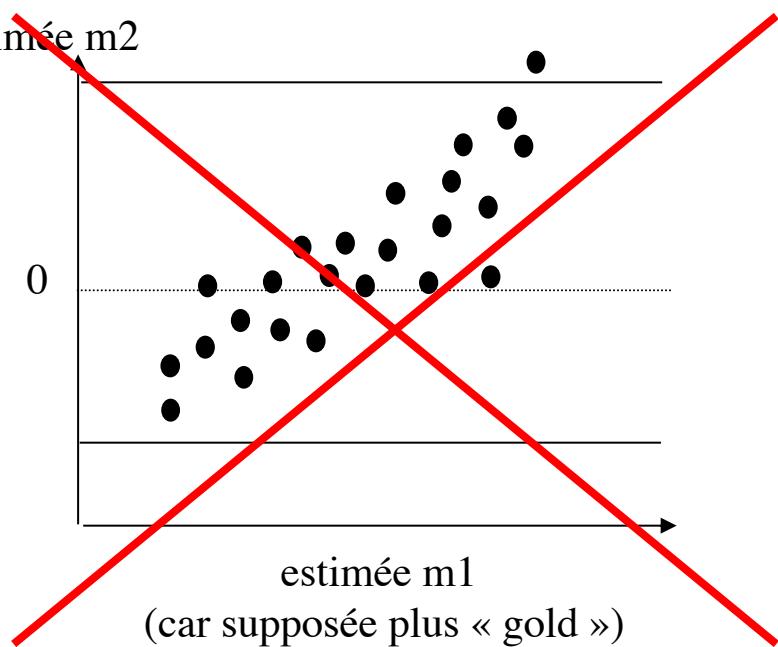
(estimée m1 + estimée m2)/2

meilleure estimée de la vraie valeur disponible



Permet de détecter des différences systématiques entre les méthodes

estimée m1 - estimée m2



estimée m1

(car supposée plus « gold »)

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

# Régression sans gold standard

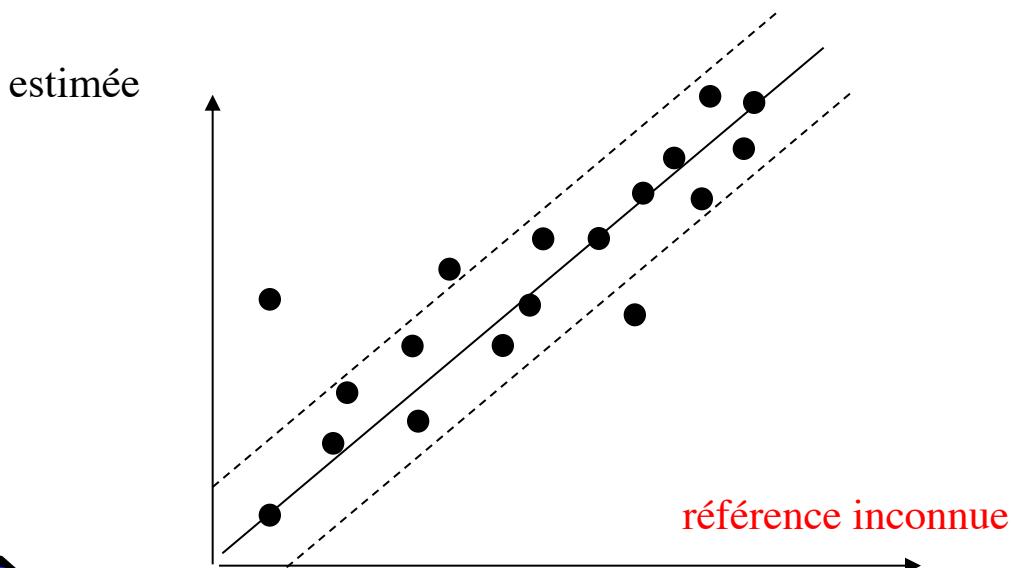
Hypothèses :

- \* 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 :  $a_m, b_m, \sigma_m, r$

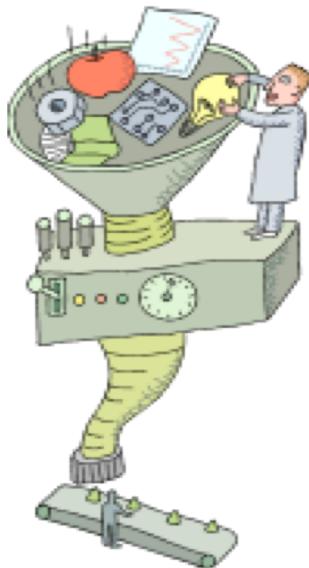
Figure de mérite :  $\sigma_m / a_m$  (le plus faible possible)



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

# Régression sans gold standard

Caractéristiques de l'approche :



- \* 2 méthodes sont suffisantes
- \* robuste même lorsque l'hypothèse sur la distribution des  $p_i$  est approximative
- \* 25 cas au moins sont nécessaires

Généralisation de l'approche :

- \* généralisation à une dépendance quadratique
- \* nouvelle figure de mérite Q plus performante
- \* intervalle de confiance bootstrap autour de Q

Réf: Kupinski et al. Academic Radiology, 9: 290-297, 2002

Hopkin et al. IEEE Trans Med Imaging, 21: 441-449, 2002

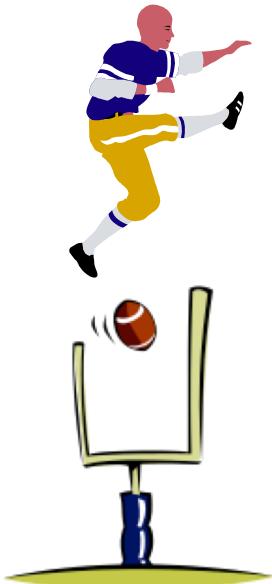
Leenberg et al, IEEE Trans Med Imaging, 31: 1651-1660, 2012

# Conclusion



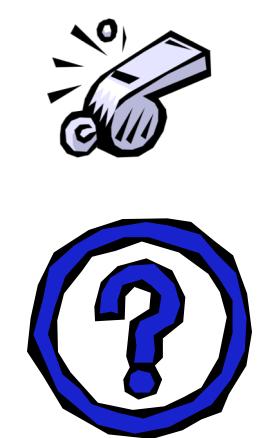
Deux types de travaux d'évaluation :

- tâche de classification
- tâche d'estimation



Pour chaque type, méthodes d'évaluation rigoureuse :

- approches type ROC
- biais, variabilité



Même en l'absence de gold standard (cas le plus fréquent en évaluation clinique), possibilité d'effectuer une évaluation rigoureuse

# Annexe 1 : table des coefficients r entre W1 et W2

Corrélation moyenne	AUC moyenne												
	.700	.725	.750	.775	.800	.825	.850	.875	.900	.925	.950	.975	
.02	.02	.02	.02	.02	.02	.02	.02	.01	.01	.01	.01	.01	.01
.04	.04	.04	.03	.03	.03	.03	.03	.03	.03	.02	.02	.02	.02
.06	.05	.05	.05	.05	.05	.05	.05	.04	.04	.04	.03	.03	.02
.08	.07	.07	.07	.07	.07	.06	.06	.06	.06	.05	.04	.04	.03
.10	.09	.09	.09	.09	.08	.08	.08	.07	.07	.06	.06	.06	.04
.12	.11	.11	.11	.10	.10	.10	.09	.09	.08	.08	.07	.07	.05
.14	.13	.12	.12	.12	.12	.11	.11	.11	.10	.09	.08	.08	.06
.16	.14	.14	.14	.14	.13	.13	.13	.12	.11	.11	.09	.09	.07
.18	.16	.16	.16	.16	.15	.15	.14	.14	.13	.12	.11	.09	.09
.20	.18	.18	.18	.17	.17	.17	.16	.15	.15	.14	.12	.10	
.22	.20	.20	.19	.19	.19	.18	.18	.17	.16	.15	.14	.11	
.24	.22	.22	.21	.21	.21	.20	.19	.19	.18	.17	.15	.12	
.26	.24	.23	.23	.23	.22	.22	.21	.20	.19	.18	.16	.13	
.28	.26	.25	.25	.25	.24	.24	.23	.22	.21	.20	.18	.15	
.30	.27	.27	.27	.26	.26	.25	.25	.24	.23	.21	.19	.16	
.32	.29	.29	.29	.28	.28	.27	.26	.26	.24	.23	.21	.18	
.34	.31	.31	.31	.30	.30	.29	.28	.27	.26	.25	.23	.19	
.36	.33	.33	.32	.32	.31	.31	.30	.29	.28	.26	.24	.21	
.38	.35	.35	.34	.34	.33	.33	.32	.31	.30	.28	.26	.22	
.40	.37	.37	.36	.36	.35	.35	.34	.33	.32	.30	.28	.24	
.42	.39	.39	.38	.38	.37	.36	.36	.35	.33	.32	.29	.25	
.44	.41	.40	.40	.40	.39	.38	.38	.37	.35	.34	.31	.27	
.46	.43	.42	.42	.42	.41	.40	.39	.38	.37	.35	.33	.29	
.48	.45	.44	.44	.43	.43	.42	.41	.40	.39	.37	.35	.30	
.50	.47	.46	.46	.45	.45	.44	.43	.42	.41	.39	.37	.32	
.52	.49	.48	.48	.47	.47	.46	.45	.44	.43	.41	.39	.34	
.54	.51	.50	.50	.49	.49	.48	.47	.46	.45	.43	.41	.36	
.56	.53	.52	.52	.51	.51	.50	.49	.48	.47	.45	.43	.38	
.58	.55	.54	.54	.53	.53	.52	.51	.50	.49	.47	.45	.40	
.60	.57	.56	.56	.55	.55	.54	.53	.52	.51	.49	.47	.42	
.62	.59	.58	.58	.57	.57	.56	.55	.54	.53	.51	.49	.45	
.64	.61	.60	.60	.59	.59	.58	.58	.57	.55	.54	.51	.47	
.66	.63	.62	.62	.62	.61	.60	.60	.59	.57	.56	.53	.49	
.68	.65	.64	.64	.64	.63	.62	.62	.61	.60	.58	.56	.51	
.70	.67	.66	.66	.66	.65	.65	.64	.63	.62	.60	.58	.54	
.72	.69	.69	.68	.68	.67	.67	.66	.65	.64	.63	.60	.56	
.74	.71	.71	.70	.70	.69	.69	.68	.67	.66	.65	.63	.59	
.76	.73	.73	.72	.72	.72	.71	.71	.70	.69	.67	.65	.61	
.78	.75	.75	.75	.74	.74	.73	.73	.72	.71	.70	.68	.64	
.80	.77	.77	.77	.76	.76	.76	.75	.74	.73	.72	.70	.67	
.82	.79	.79	.79	.79	.78	.78	.77	.77	.76	.75	.73	.70	
.84	.82	.81	.81	.81	.81	.80	.80	.79	.78	.77	.76	.73	
.86	.84	.84	.83	.83	.83	.82	.82	.81	.81	.80	.78	.75	
.88	.86	.86	.86	.85	.85	.85	.84	.84	.83	.82	.81	.79	
.90	.88	.88	.88	.88	.87	.87	.87	.86	.86	.85	.84	.82	

## Annexe 2 : références bibliographiques

- Berbaum KS, Dorfman DD, Franken EA. Measuring observer performance by ROC analysis: indications and complications. *Invest Radiol*, 1989; 24: 228-233
- DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated Receiver Operating Characteristic curves: a nonparametric approach. *Biometrics*, 1988; 44: 837-845
- de Vries DJ, King MA, Soares EJ, Tsui BMW, Metz CE. Effects of scatter subtraction on detection and quantitation in hepatic SPECT. *J Nucl Med*, 1999; 40: 1011-1023
- Gifford HC, King MA, de Vries DJ, Soares EJ. Channelized Hotelling and human observer correlation for lesion detection in hepatic SPECT imaging. *J Nucl Med*, 2000; 41: 514-521
- Gilland DR, Tsui BMW, Metz CE, Jaszczak RJ, Perry JR. An evaluation of maximum likelihood-expectation maximization reconstruction for SPECT by ROC analysis. *J Nucl Med*, 1992; 33: 451-457
- Gur D, King JL, Rockette HE, Britton CA, Thaete FL, Hoy RJ. Practical issues of experimental ROC analysis: selection of controls. *Invest Radiol*, 1990; 25: 583-586
- Hajian-Tilaki KO, Hanley JA, Joseph L, Collet JP. A comparison of parametric and nonparametric approaches to ROC analysis of quantitative diagnostic tests. *Med Decis Making*, 1997; 17: 94-102
- Hanley JA. The robustness of the “binormal” assumptions used in fitting ROC curves. *Med Decis Making*, 1988; 8: 197-203
- Hanley JA, McNeil BJ. The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. *Radiology*, 1982; 143: 29-36
- Hanley JA, McNeil BJ. A method of comparing the areas under Receiver Operating Characteristic curves derived from the same cases. *Radiology*, 1983; 148: 839-843
- Henkelman RM, Kay I, Bronskill MJ. Receiver Operating Characteristic (ROC) analysis without truth. *Med Decis Making*, 1990; 10: 24-29
- LaCroix KJ, Tsui BMW, Frey EC, Jaszczak RJ. Receiver Operating Characteristic evaluation of iterative reconstruction with attenuation correction in  $^{99m}\text{Tc}$ -sestamibi myocardial SPECT images. *J Nucl Med*, 2000; 41: 502-513
- Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. *Invest Radiol*, 1989; 24: 234-245
- Metz CE, Herman BA, Roe CA. Statistical comparison of two ROC-curve estimates obtained from partially paired datasets. *Med Decis Making*, 1998; 18: 110-121
- Metz CE, Shen JH. Gains in accuracy from replicated readings of diagnostic images: prediction and assessment in terms of ROC analysis. *Med Decis Making*, 1992; 12: 60-75
- Obuchowski NA. Computing sample size for Receiver Operating Characteristic studies. *Invest Radiol*, 1994; 29: 238-243
- Rockette HE, Gur D, cooperstein LA, Obuchowski NA, King JL, Fuhrman CR, Tabor EK, Metz CE. Effect of two rating formats in multi-disease ROC study of chest images. *Invest Radiol*, 1990; 25: 225-229
- Swensson RG. Unified measurement of observer performance in detecting and localizing target objects on images. *Med Phys*, 1996; 23: 1709-1725
- Thibodeau LA. Evaluating diagnostic tests. *Biometrics*, 1981; 37: 801-804
- logiciel rockit : site [www-radiology.uchicago.edu/sections/roc](http://www-radiology.uchicago.edu/sections/roc)

# Comparaison courbes ROC : appariement partiel

Sujets 1 à n1	indice t1 seulement
Sujets (n1+1) à n2	indice t2 seulement
Sujets (n2+1) à N	indices t1 et t2

Alternatives :

- considérer uniquement les sujets (n2+1) à N
  - gaspillage et perte de puissance
- considérer tous les sujets mais ignorer l'appariement
  - perte de puissance
- utiliser un test gérant les données partiellement appariées
  - test paramétrique seulement

Hypothèses :

- Echantillons appariés et non appariés issus de la même population
- Echantillons appariés et non appariés indépendants
- Indices 1 et 2 interprétés indépendamment
- lois binormales sous-jacentes pour chaque indice
  - maximisation d'une fonction de vraisemblance
    - estimation de Az1, Az2, SE(Az1), SE(Az2), covariance
    - test z intégrant la corrélation entre les données corrélées

# Comparaison de plus de 2 courbes ROC

Situations possibles :

Nb d'indices à comparer	Nb d'observateurs	Nb de lectures par observateur
2	1	>1
2	>1	1
2	>1	>1
>2	1	1
>2	>1	1
>2	1	>1
>2	>1	>1

→ A chaque situation, une procédure de test !

→ Pas de consensus sur les procédures les plus appropriées

## 2 indices, plusieurs observateurs, 1 lecture (1)

Nb d'indices à comparer	Nb d'observateurs	Nb de lectures par observateur
2	$K > 1$	1

- Pour chaque observateur  $k$  :  
calculer  $Az1_k$  et  $Az2_k$  (ou  $W1_k$  et  $W2_k$ )
  - Pour chaque observateur  $k$  :  
calculer  $\Delta_k = Az1_k - Az2_k$  (ou  $\Delta_k = W1_k - W2_k$ )
  - Réaliser un test t de Student testant :  
 $H_0 : \text{moyenne}(\Delta_k) = 0$
- Beaucoup de lecteurs nécessaires pour atteindre une puissance statistique correcte (test t)
- La conclusion est valable pour l'échantillon étudié seulement, donc s'assurer de la représentativité de l'échantillon : attention aux petits échantillons !
- Une courbe ROC moyenne pour chaque méthode peut être obtenue en moyennant (sur les observateurs) les valeurs  $A1_k$  et  $B1_k$  estimées par le modèle binormal :
- $$A1 = \sum_k A1_k / K, \quad B1 = \sum_k B1_k / K,$$
- $$A2 = \sum_k A2_k / K, \quad B2 = \sum_k B2_k / K,$$

Réf : Metz, 1989, exemple dans La Croix et al, 2000

## 2 indices, plusieurs observateurs, 1 lecture (2)

Nb d'indices à comparer	Nb d'observateurs	Nb de lectures par observateur
2	$K > 1$	1

- Pour chaque observateur  $k$  :  
→ calculer  $Az1_k$  et  $Az2_k$  (ou  $W1_k$  et  $W2_k$ )
  - Pour chaque observateur  $k$  :  
tester  $H_0 : Az1_k = Az2_k$  (ou  $H_0 : W1_k - W2_k$ )  
→ valeur de  $p$  pour chaque observateur  $k$
  - Conclure :
    - Avec  $K$  observateurs, la différence moyenne entre  $Az1$  et  $Az2$  vaut ...
    - L'indice 1 s'est avéré significativement meilleur que l'indice 2 avec  $p < 0.05$  pour ... observateurs sur  $K$
- Pas de conclusion générale

Réf : Metz, 1989, exemple dans La Croix et al, 2000

# M indices, 1 observateur, 1 lecture

Nb d'indices à comparer	Nb d'observateurs	Nb de lectures par observateur
M	1	1

- Utilisation de la statistique de Wilcoxon pour calculer W1, W2, ... WM
  - Calcul analytique de la matrice variance-covariance (M x M) S du vecteur  $\mathbf{W} = (W_1, \dots, W_M)$
  - Expression d'une matrice L de contraste linéaire correspondant à l'hypothèse à tester  
exemples :
    - indice 1 meilleur que la moyenne des indices 2 et 3 ?  
 $L = (1 \quad -0.5 \quad -0.5)$
    - indice 1 meilleur que indice 2 ou meilleur que indice 3 ?  
 $L = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix}$
  - Calcul d'une statistique du Chi-2 = f(L, S) et comparaison à la valeur critique
- approche non paramétrique, test apparié

Réf et exemple : De Long et al, 1988

# M indices, K observateurs, 1 lecture (1)

Nb d'indices à comparer	Nb d'observateurs	Nb de lectures par observateur
M	K	1

- pour chaque observateur k :  
calculer les  $Azm_k$  (ou  $Wm_k$ )
- pour chaque observateur k :  
tester  $H_0 : Azm_k = Azm'_k$  (ou  $H_0 : Wm_k - Wm'_k = 0$ ) pour les couples de méthodes m et m' d'intérêt  
→ valeur de p pour chaque observateur k et chaque couple (m, m')

OU

- pour chaque observateur k :  
calculer les  $Azm_k$  (ou  $Wm_k$ )
- pour chaque observateur k :  
calculer  $\Delta_{mm'k} = Azm_k - Azm'_k$  (ou  $\Delta_k = Wm_k - Wm'_k$ )
- réaliser des tests t de Student testant :  
 $H_0 : \text{moyenne}(\Delta_{mm'k}) = 0$

Réf et exemple : Gilland et al, 1992

## M indices, K observateurs, 1 lecture (2)

Nb d'indices à comparer	Nb d'observateurs	Nb de lectures par observateur
M	K	1

- pour chaque observateur k :  
calculer les  $Azm_k$  (ou  $Wm_k$ )
- analyse de variance à 2 critères de classification (2-way ANOVA) pour tester  $H_0$  :  
tous les  $Azm$  sont équivalentes ( $H_1$  : les indices diffèrent entre eux)
- si  $H_0$  est rejeté :  
test de Scheffé de comparaisons multiples appariées

Réf et exemples : de Vries et al, 1999, Gifford et al, 2000

# 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  $\times RV+$

... et ne dépendent pas de la prévalence

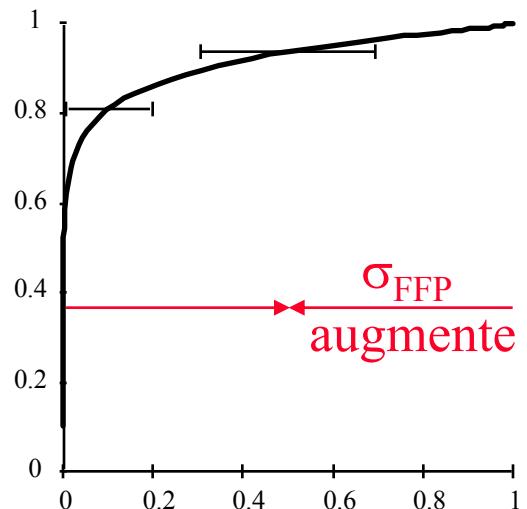
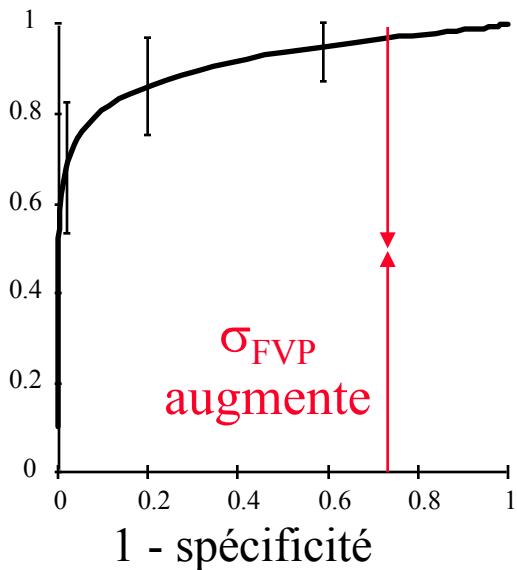


# Fiabilité d'une courbe ROC

$$\sigma_{FVP} = \sqrt{\frac{FVP(1 - FVP)}{n_{P+} - 1}}$$

$$\sigma_{FFP} = \sqrt{\frac{FFP(1 - FFP)}{n_{P-} - 1}}$$

sensibilité



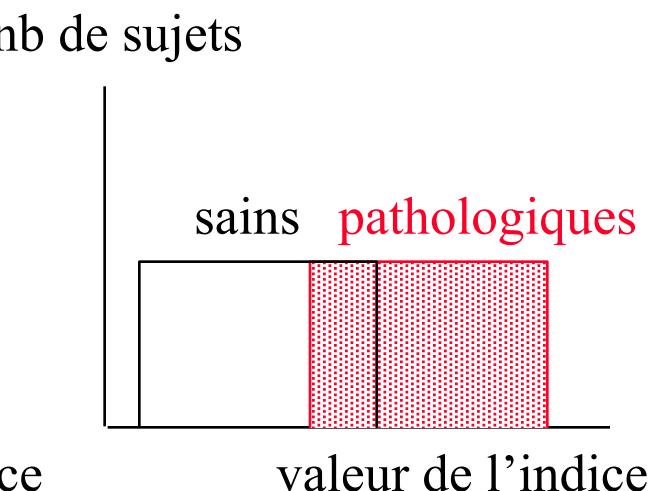
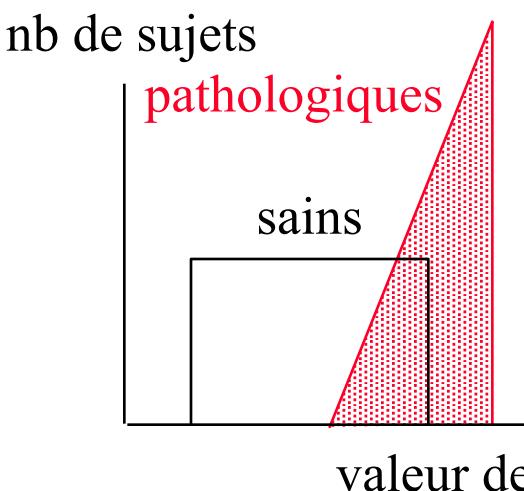
- $\sigma_{FVP}$  augmente quand on se rapproche de  $FVP = 0.5$
- $\sigma_{FFP}$  augmente quand on se rapproche de  $FFP = 0.5$
- $\square \sigma_{FVP}$  varie inversement avec le nombre de sujets pathologiques ( $n_{P+}$ )
- $\square \sigma_{FFP}$  varie inversement avec le nombre de sujets sains ( $n_{P-}$ )



choisir plutôt  $n_{P+} \sim n_{P-}$

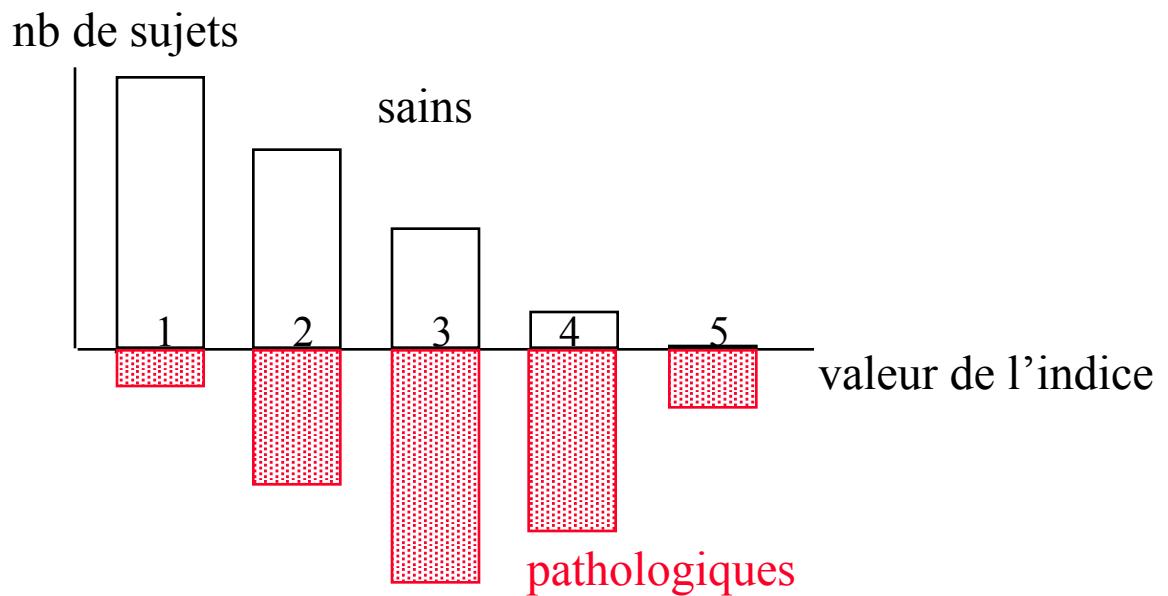
# Justification du modèle binormal (1)

- Modèle bien adapté lors de jugements subjectifs par un observateur exprimés sur une échelle comportant un petit nombre de catégories, même pour :
  - distributions binomiales
  - distributions de Poisson
  - distributions du Chi-2
  - distributions gamma
- Problèmes possibles avec :
  - distributions puissance :  $FVP = FFP^k$
  - distributions triangulaire ou rectangulaire

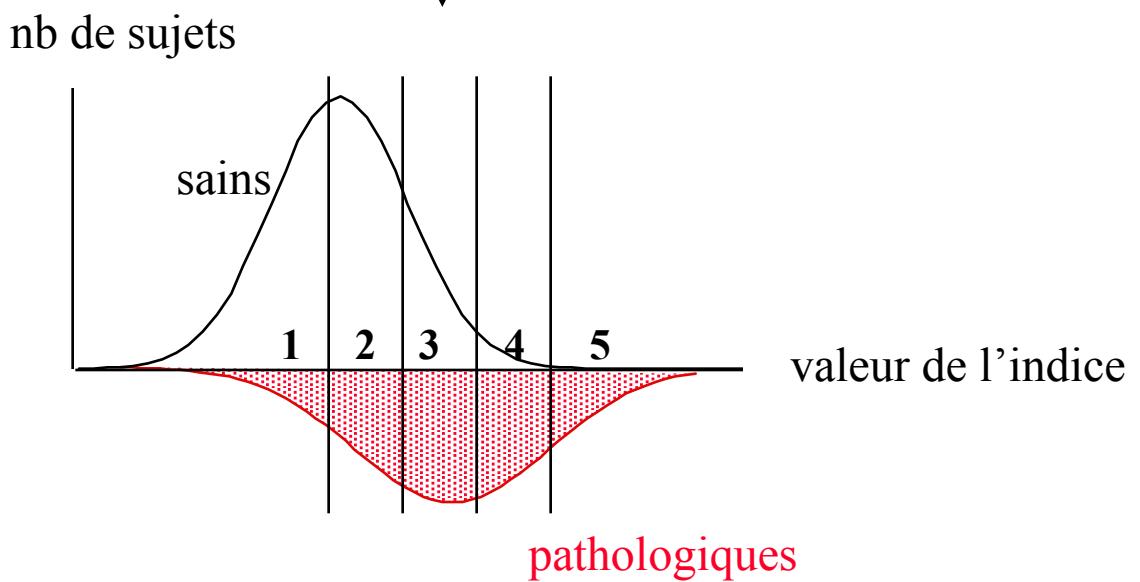


# Justification du modèle binormal (2)

- Faible nombre de catégories dans l'échelle des scores



ajustement par deux lois normales



# Considérations pratiques ROC : nb de cas à inclure

Combien de cas nécessaire pour mettre en évidence une différence significative entre AUC1 et AUC2 ?

## Hypothèses

- Connaissance approximative de AUC1 et AUC2
  - Optionnellement, connaissance approximative du rapport des écarts types  $\sigma_S$  et  $\sigma_P$  entre les deux lois normales sous-jacentes
- si indice continu :  
SE(AUC) donnée par Hanley et McNeil, 1992
- si indice = score sur échelle discrète :  
SE(Az) donnée par Obuchowski, 1994



déduction de  $n_{P+}$  et  $n_P$   
pour un niveau de signification  $\alpha$  donné

# Considérations pratiques ROC : sélection des cas

- Le plus souvent : comparaison des performances des deux indices
  - pour maximiser la puissance des tests, choisir des cas de difficulté intermédiaire :  
 $AUC \sim 0,75 - 0,80$
  - conduire éventuellement une étude pilote pour identifier les cas de difficulté intermédiaire



Si les cas sont « trop faciles » : les 2 méthodes vont toujours conduire à la bonne classification

Si les cas sont « trop difficiles », les 2 méthodes vont aussi toujours conduire à une classification érronée ou hasardeuse

# Considérations pratiques ROC : ordre de lecture

- Image indice 1 présentée systématiquement avant image indice 2

→ biais potentiel

- Deux alternatives

→ 1 avant 2 pour la moitié des observateurs  
2 avant 1 pour l'autre moitié

→ pour chaque observateur :  
- 1 avant 2 pour la moitié des images  
- 2 avant 1 pour l'autre moitié  
- éloignement maximal des images

correspondant aux mêmes cas vues par les 2 indices

- ordre différent pour chaque observateur

Exemple : 2 indices 1 et 2, 4 observateurs

→ diviser les images en deux sous-groupes A et B de même effectif

→ assigner un ordre pour chaque observateur :  
(A1, B2, A2, B1)  
(A2, B1, A1, B2)  
(B1, A2, B2, A1)  
(B2, A1, B1, A2)

# Considérations pratiques ROC : gold standard

- Gold standard :
  - la bonne réponse
  - référence
  - ground truth
  - vérité terrain



# Considérations pratiques ROC : gold std imparfait

- Deux cas
  - erreur gold standard non corrélée à l'erreur indice
  - erreur gold standard corrélée à l'erreur indice
- Non corrélation des erreurs :
  - courbe ROC comprimée vers la diagonale
  - AUC sous-estimée
  - sensibilité et spécificité systématiquement sous-estimées
- Corrélation des erreurs :
  - courbe ROC optimiste
  - AUC surestimée
  - sensibilité et spécificité systématiquement surestimées
- Que faire ?
  - utiliser une définition très stricte pour estimer la sensibilité
  - utiliser une définition moins rigoureuse pour estimer la spécificité
  - considérer des populations différentes pour estimer les sensibilité (forte prévalence) et spécificité (faible prévalence)
  - possible correction des biais a posteriori si on sait si les erreurs gold standard - indice sont corrélées

# Considérations pratiques ROC : échelle de cotation

- Nombre de catégories ?
  - ➡ précision accrue lorsqu'on passe de 5 à 10
  - ➡ au delà de 10, plus de gain substantiel
- Format de cotation si plusieurs (A) types d'anomalies ?
  - format général : 1 score par image
    - absence ou présence d'une anomalie (avec degré de confiance associé) quelle qu'elle soit
    - exemple : anomalie 1 ou 2 ou 3 est :
      1. très certainement absente
      2. probablement absente
      3. possiblement absente
      4. probablement présente
      5. très certainement présente
  - format spécifique :
    - pour chaque type d'anomalie, absence ou présence cotée sur l'échelle de 1 à 5, d'où A scores par image
    - score le plus élevé retenu, d'où 1 score par image
- ➡ pas de différence significative entre les 2 formats si on s'intéresse seulement à la présence d'une anomalie, et pas à sa nature.
- ➡ si on limite l'analyse ROC aux images présentant un type d'anomalie, influence du format plus ambiguë

Réf : Rockette et al, 1990

# Considérations pratiques ROC : paramétrique ou non ?

## REGLES

- indice résultant de jugements d'observateurs et exprimé sur une échelle de cotation discrète
  - modèle binormal bien adapté
- indice continu
  - approche non paramétrique plus générale si les distributions sous-jacentes sont inconnues
- approche non paramétrique moins puissante que l'approche paramétrique

## EN PRATIQUE

- données effectivement binormales
    - Az et SE pratiquement non biaisées que l'approche soit paramétrique ou non
  - données non binormales
    - estimées très fiables des AUC par les 2 approches (biais < 1% en NP, <2.5% en P)
    - tendance à surestimer SE avec les deux approches
- CHOISIR LE PLUS PRATIQUE !

Réf : Hajian-Tilaki et al, 1997

# Considérations pratiques ROC : lectures multiples

## REGLES POUR AMELIORER L'ESTIMATION ROC

- Si les observateurs présentent des performances voisines, plusieurs observateurs préférables à plusieurs lectures par un même observateur
- Réduction des variabilités d'autant plus importantes que les résultats issus des différents observateurs ou observations sont peu corrélés.
- Combinaison des scores puis estimation ROC

Exemple :

obs 1 :	scores	1	2	3	4	5
obs 2 :	scores	1	2	3	4	5
obs “1 et 2” :	scores	2	3	4	5	6    7    8    9    10

analyse ROC à partir des scores combinés

Gains à espérer :

1 obs :  $Az = 0.836$

obs “1 et 2” :  $Az = 0.860$

Réf : Metz et Shen, 1992

# Exemple : test t entre 2 échantillons indépendants

- On dispose de 2 échantillons de N1 et N2 mesures chacun, de moyennes  $m_1$  et  $m_2$  et d'écart-type  $\sigma_1$  et  $\sigma_2$
- $H_0$  : il n'y a pas de différence entre la valeur du paramètre entre les 2 populations dont sont issus les 2 échantillons
- Statistique de test :

$$t = \frac{m_1 - m_2}{\text{écart-type}_{(m1-m2)}}$$

- écart-type<sub>(m1-m2)</sub> =  $[s^2(1/N_1 + 1/N_2)]^{1/2}$

où  $s^2 = [(N_1-1) \sigma_1^2 + (N_2-1) \sigma_2^2] / (N_1+N_2-2)$

et  $ddl = N_1 + N_2 - 2$

Exemple :

Echantillon 1 : 2, 4, 3 et 5

Echantillon 2 : 3, 3, 1 et 2

$H_1$  : il y a une différence entre la valeur du paramètre entre les 2 populations

$$t = 1,56$$

$$ddl = 6$$

$0,1 < p < 0,2$  : on ne peut pas rejeter  $H_0$

# Équivalence avec les probabilités conditionnelles

P = statut réel de la pathologie

T = résultat issu de l'indice

quand

- Fraction de vrais positifs FVP =  $\frac{VP}{VP + FN} = \text{Prob}(T+ | P+)$
- Fraction de vrais négatifs FVN =  $\frac{VN}{VN + FP} = \text{Prob}(T- | P-)$
- Fraction de faux positifs FFP =  $\frac{FP}{FP + VN} = \text{Prob}(T+ | P-)$
- Fraction de faux négatifs FFN =  $\frac{FN}{FN + VP} = \text{Prob}(T- | P+)$
- Valeur prédictive positive VPP =  $\frac{VP}{VP + FP} = \text{Prob}(P+ | T+)$
- Valeur prédictive négative VPN =  $\frac{VN}{VN + FN} = \text{Prob}(P- | T-)$
- Prévalence de la pathologie Prév =  $\text{Prob}(P+)$



# Variantes de l'approche ROC

- LROC : Localized Response Operating Characteristic
  - 1 anomalie par image
  - détecter ET localiser l'anomalie dans une image
  - ordonnée : FVP (bien détectés et localisés),  
abscisse : FFP (pour les images sans anomalie)
- FROC : Free Response Operating Characteristic
  - plusieurs anomalies possibles par image
  - ordonnée : FVP (bien détectés et localisés)  
abscisse : FFP (nb moyen de FP par image)
  - potentiellement plus puissant que ROC ou LROC pour quantifier des performances de détection
- AFROC : Alternative FROC
  - plusieurs anomalies possibles par image
  - ordonnée : FVP (bien détectés et localisés) = ordonnée de FROC  
abscisse : fraction d'images contenant au moins un FP = abscisse de ROC

Limite majeure :

Modèles complexes pour ajuster les observations  
→ pas de test statistique adapté à la comparaison de deux courbes LROC ou de deux courbes FROC

Réf : Swensson RG, 1996