MS BGD
MDI 720 : Statistiques

François Portier et Joseph Salmon
http://josephsalmon.eu
Télécom Paristech, Institut Mines-Télécom

Sommaire

Sommaire

Tests d'hypothèses pour le "Pile ou face"

- On veut tester une hypothèse sur le paramètre θ.
- On l'appelle hypothèse nulle $\mathcal{H}_0$
  Exemple : 'la pièce est non biaisée" : $\mathcal{H}_0 = \{p = 0.5\}$.
  Exemple : 'la pièce est peu biaisée', $\mathcal{H}_0 = \{0.45 \leqslant p \leqslant 0.55\}$
- L'hypothèse alternative $\mathcal{H}_1$ est (souvent) le contraire de $\mathcal{H}_0$.
  Exemple : $\mathcal{H}_1 = \{p \neq 0.5\}$
  Exemple : $\mathcal{H}_1 = \{p \notin [0.45, 0.55]\}$
- « Faire un test » : déterminer si les données permettent de rejeter l'hypothèse $\mathcal{H}_0$. On cherche une région R pour laquelle si $(y_1, \dots, y_n) \in R$ on rejette l'hypothèse $\mathcal{H}_0$. R est la région de **rejet**.

Rejet ou acceptation ?

Présomption d'innocence en faveur de $\mathcal{H}_0$

Même si $\mathcal{H}_0$ n'est pas rejetée par le test, on ne peut en général pas conclure que $\mathcal{H}_0$ est vraie !

Rejeter $\mathcal{H}_1$ est souvent impossible car $\mathcal{H}_1$ est trop générale.
e.g., $\{p \in [0, 0.5[\cup]0.5, 1]\}$ ne peut pas être rejetée !

- $\mathcal{H}_0$ s'écrit sous la forme $\{\theta \in \Theta_0\}$, avec $\Theta_0 \subset \Theta$
- $\mathcal{H}_1$ s'écrit sous la forme $\{\theta \in \Theta_1\}$, avec $\Theta_1 \subset \Theta$

Rem: $\{\theta \in \Theta_0\}$ et $\{\theta \in \Theta_1\}$ sont disjoints.

Risques de première et de seconde espèce

	$\mathcal{H}_0$	$\mathcal{H}_1$
Non rejet de $\mathcal{H}_0$	Juste	Faux (acceptation à tort)
Rejet de $\mathcal{H}_0$	Faux (rejet à tort)	Juste

- Risque de 1^{re} espèce : probabilité de rejeter à tord
$$\alpha = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta((y_1, \ldots, y_n) \in R)$$

- Risque de 2^{nde} espèce : probabilité d'accepter à tord
$$\sup_{\theta \in \Theta_1} \mathbb{P}_\theta\left((y_1, \ldots, y_n) \notin R\right)$$

Niveau/Puissance

Niveau du test

$1 - \alpha =$ probabilité d'« accepter » à raison (si $\mathcal{H}_0$ est valide)

Puissance du test

$1 - \beta =$ probabilité de rejeter $\mathcal{H}_0$ à raison (si $\mathcal{H}_1$ est valide)

En général, lorsqu'on parle de « test à 95% » on parle d'un test de niveau $1 - \alpha \geqslant 95\%$.

Statistique de test et région de rejet

Objectif classique : construire un test de niveau $1-\alpha$

- On cherche une fonction des données $T_n(y_1,\ldots,y_n)$ dont on connaît la loi si $\mathcal{H}_0$ est vraie : T_n est appelée *statistique de test*.
- On définit une région de rejet ou région critique de niveau α, une région R telle que, sous $\mathcal{H}_0$,
$$\mathbb{P}(T_n(y_1,\ldots,y_n)\in R)\leqslant\alpha$$
- Règle de rejet de $\mathcal{H}_0$: on rejette si $\boxed{T_n(y_1,\ldots,y_n)\in R}$

Exemple gaussien : nullité de la moyenne

- Modèle : $\Theta = \mathbb{R}$, $\mathbb{P}_\theta = \mathcal{N}(\theta, 1)$.
- Hypothèse nulle : $\mathcal{H}_0 : \{\theta = 0\}$
- Sous $\mathcal{H}_0$, $\quad T_n(y_1, \ldots, y_n) = \frac{1}{\sqrt{n}} \sum_i y_i \sim \mathcal{N}(0, 1)$
- Région critique pour T_n ? Quantiles gaussiens : sous H_0,
$$\mathbb{P}(T_n \in [-1.96, 1.96]) = 0.95$$

  On prend $R = [-1.96, 1.96]^C =]-\infty, -1.96[\cup]1.96, +\infty[$.
- <u>**Exemple numérique**</u> : si $T_n = 1.5$, on ne rejette **PAS** $\mathcal{H}_0$ au niveau 95%

Sommaire

Tester la nullité des coefficients (I)

Rappel : prenons $X \in \mathbb{R}^{n \times p}$, alors $\hat{\sigma}^2 = \|\mathbf{y} - X\hat{\boldsymbol{\theta}}\|_2^2/(n - \mathrm{rg}(X))$ est un estimateur sans biais de la variance. Ainsi

$$\text{Si } \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \,\mathrm{Id}_n), \text{ alors } \quad \boxed{T_j = \frac{\hat{\theta}_j - \theta_j^*}{\hat{\sigma}\sqrt{(X^\top X)^{-1}_{j,j}}} \sim \mathcal{T}_{n-\mathrm{rg}(X)}}$$

où $\mathcal{T}_{n-\mathrm{rg}(X)}$ est une loi dite de Student (de degré $n - \mathrm{rg}(X)$).

Sa densité, ses quantiles, etc... peuvent être calculés numériquement.

Tester la nullité des coefficients (I)

$H_0 : \theta_j^* = 0$ ce qui revient à prendre $\Theta_0 = \{\boldsymbol{\theta} \in \mathbb{R}^p : \theta_j = 0\}$.
Sous H_0 on connaît donc la distribution de $\hat{\theta}_j$:

t-statiques : $$T_j := \frac{\hat{\theta}_j}{\hat{\sigma}\sqrt{(X^\top X)^{-1}_{j,j}}} \sim \mathcal{T}_{n-\mathrm{rg}(X)}$$

Ainsi en choisissant comme région de rejet $[-t_{1-\alpha/2}, t_{1-\alpha/2}]^c$ (en notant $t_{1-\alpha/2}$ un quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{T}_{n-\mathrm{rg}(X)}$), on peut former le test (de Student) :

$$\mathbb{1}_{\{|T_j|>t_{1-\alpha/2}\}}$$

c'est-à-dire que l'on rejette H_0 au niveau α, si $|T_j| > t_{1-\alpha/2}$

cf. Tsybakov (2006) pour plus de détails

Lien IC et Test

Rappel (modèle gaussien) :

$$IC_\alpha := \left[\hat{\theta}_j - t_{1-\alpha/2}\hat{\sigma}\sqrt{(X^\top X)^{-1}_{j,j}}, \hat{\theta}_j + t_{1-\alpha/2}\hat{\sigma}\sqrt{(X^\top X)^{-1}_{j,j}}\right]$$

est un IC de niveau α pour θ_j^*. Dire que "$0 \in IC_\alpha$" signifie que

$$|\hat{\theta}_j| \leqslant t_{1-\alpha/2}\hat{\sigma}\sqrt{(X^\top X)^{-1}_{j,j}} \quad \Leftrightarrow \quad \frac{|\hat{\theta}_j|}{\hat{\sigma}\sqrt{(X^\top X)^{-1}_{j,j}}} \leqslant t_{1-\alpha/2}$$

Cela est donc équivalent à accepter l'hypothèse $\theta_j^* = 0$ au niveau α. Le α le plus petit telle que $0 \in IC_\alpha$ est appelé la p-**value**.

Rem: On sait que si l'on prend α très proche de zéro un IC_α va recouvrir l'espace entier, on peut donc trouver (par continuité) un α qui assure l'égalité dans les équations ci-dessus.

Sommaire

Base de données "diabetes"

	age	sex	bmi	bp	Serum measurements						Resp
patient	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	y
1	59	2	32.1	101	157	93	38	4	4.9	87	151
2	48	1	21.6	87	183	103	70	3	3.9	69	75
. . .	. . .										. . .
. . .	. . .										. . .
441	36	1	30.0	95	201	125	42	5	5.1	85	220
442	36	1	19.6	71	250	133	97	3	4.6	92	57

$n = 442$ patients diabétiques, $p = 10$ variables "baseline" body mass index, bmi), average blood pressure (bp), etc. ont été mesurées. Objectif : prédire la progression de la maladie un an après les mesures"baseline" [EHJT04]

- Chacunes des variables de la base de `sklearn` a été standardisée préalablement
- On applique une version peu couteuse de la méthode "forward variable selection" (voir par exemple [Zha09])

Base de données "diabetes"

- On définit le vecteur des covariables avec intercept $\tilde{X} = (\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_{10})$.

Etape 0

- pour chacune des variables $\tilde{X}_k$, $k = 1, \ldots, 11$, on considère le modèle
$$\mathbf{y} \simeq \beta_k \mathbf{x}_k$$
- on test si son coefficient de régression est nulle, *i.e.*,
$$H_0 : \beta_k = 0$$
via la statistique $\hat{\beta}_k / \hat{s}_k$ avec $\hat{s}_k$ l'écart type estimé.
- on compare toutes les p-valeurs, on garde celle ayant la plus petite. On sauvegarde les résidus dans $\mathbf{r}_0$.

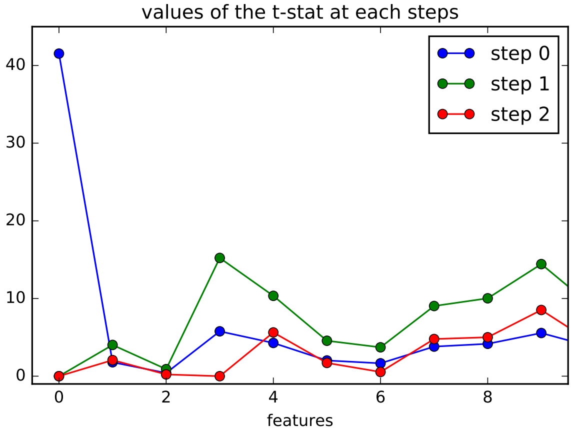
Base de données "diabetes"

Etape ℓ

On a sélectionné ℓ variable(s) : $\tilde{X}^{(\ell)} \in \mathbb{R}^{\ell}$. Les autres sont noté $\tilde{X}^{(-\ell)} \in \mathbb{R}^{p-\ell}$. On dispose du vecteur des résidus $\mathbf{r}_{\ell-1}$ calculé à l'étape précédente.

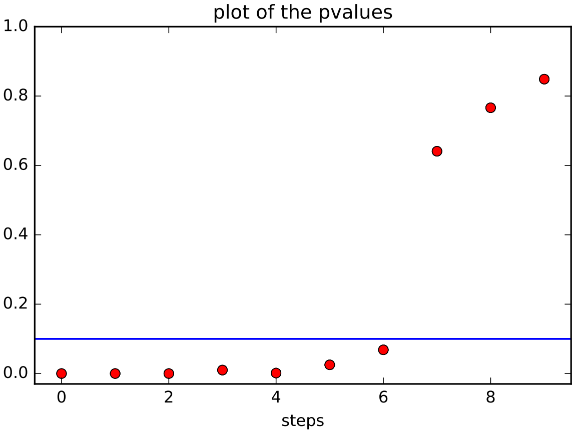
- pour chacune des variables $\mathbf{x}_k$, dans $\tilde{X}^{(-\ell)}$, on considère le modèle
$$\epsilon_{\ell-1} \simeq \beta_k \mathbf{x}_k$$
- on test si son coefficient de régression est nulle, *i.e.*,
$$H_0 : \beta_k = 0$$
via la statistique $\hat{\beta}_k / \hat{s}_k$ avec $\hat{s}_k$ l'écart type estimé.
- on compare toutes les p-valeurs, on garde celle ayant la plus petite. On sauvegarde les résidus dans $\mathbf{r}_\ell$.

Valeurs de la statistique de test à chaque étape



- la statistique d'une variable sélectionnée est mise à 0 aux étapes suivantes
- L'intercept est la première variable sélectionnée, ensuite x_3...

Valeurs de la statistique de test à chaque étape



- variables sélectionnées lors d'un test de niveau $.1$:
  [0, 3 ,9 ,5 ,4 ,2 ,7]

Sommaire

Contexte médical

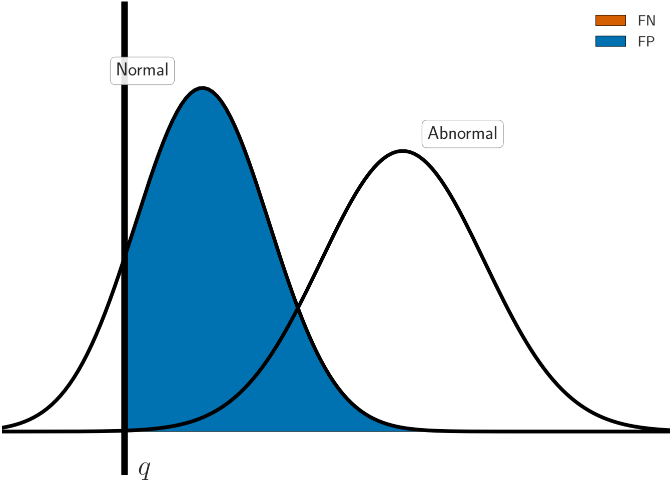
- Un groupe de patients $i = 1, \dots, n$ est suivi pour un dépistage.
- Pour chaque individu, le test se base sur une variable aléatoire $X_i \in \mathbb{R}$ et un seuil $q \in \mathbb{R}$

$$\begin{cases} \text{Si } X_i > q & \text{le test est } \textbf{positif} \\ \text{Sinon} & \text{le test est } \textbf{négatif} \end{cases}$$

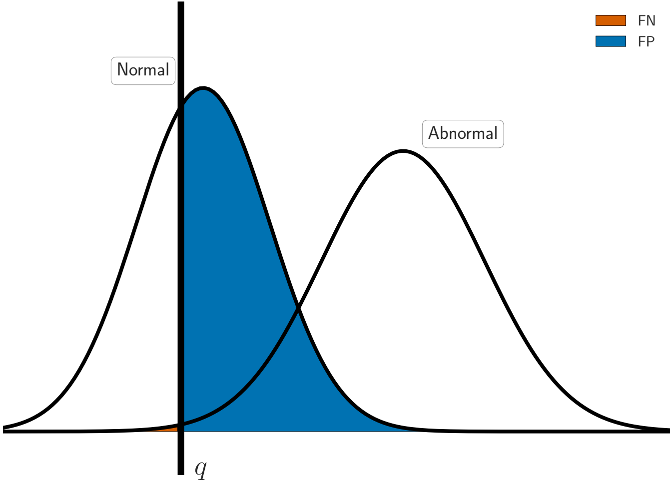
Ensemble des configurations possibles

	Normal H_0	Atteint H_1
négatif	vrai négatif	faux négatif (FN)
positif	faux positif (FP)	vrai positif

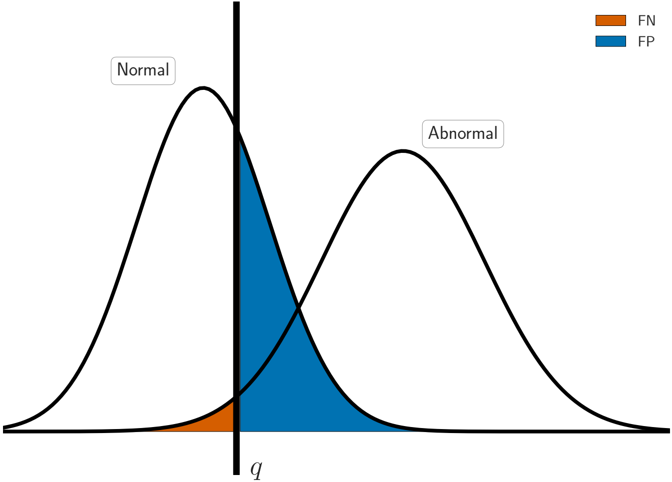
Faux positif vs faux négatif



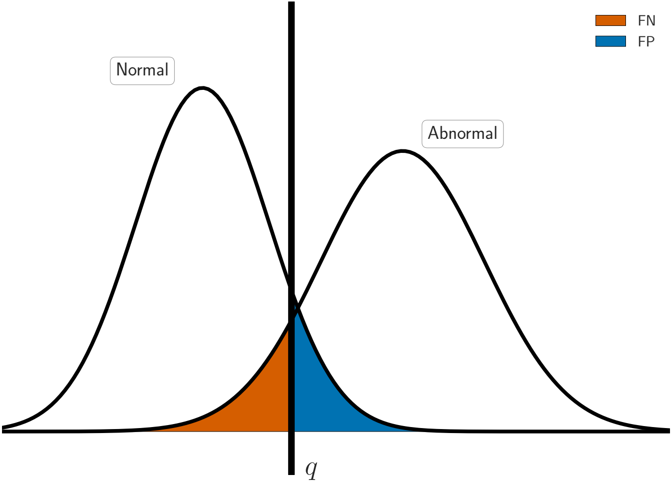
Faux positif vs faux négatif



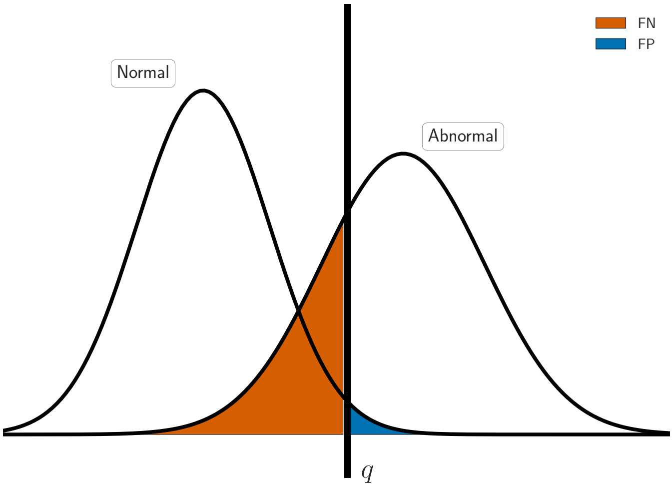
Faux positif vs faux négatif



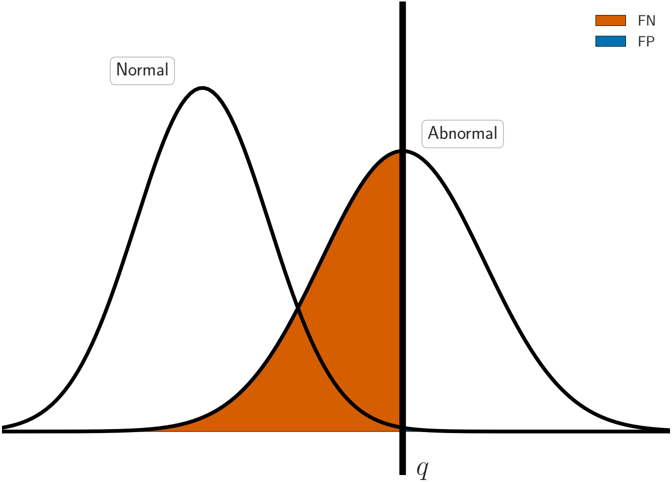
Faux positif vs faux négatif



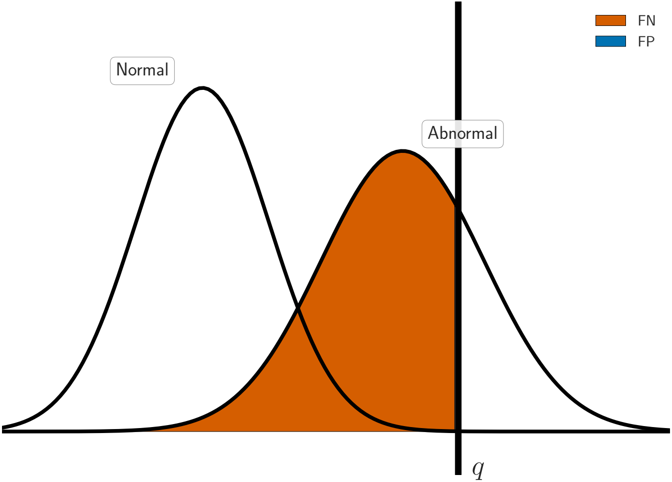
Faux positif vs faux négatif



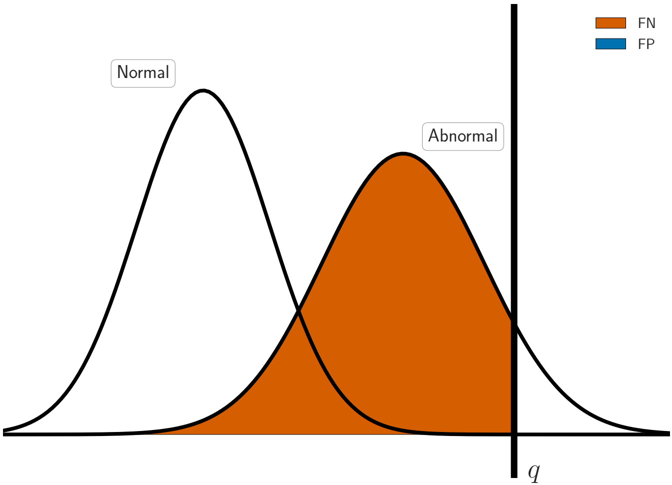
Faux positif vs faux négatif



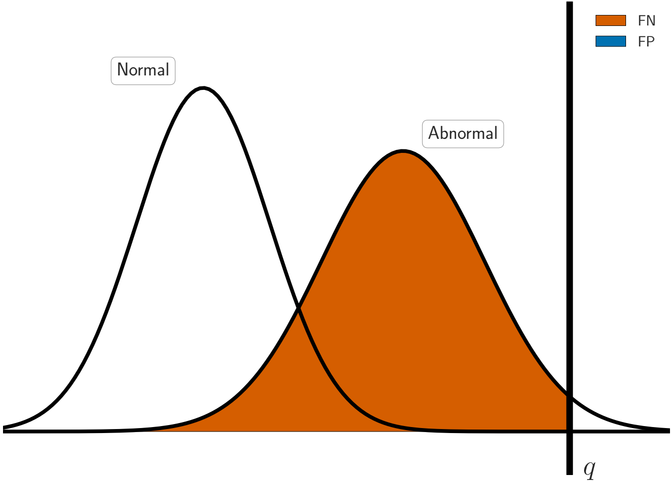
Faux positif vs faux négatif



Faux positif vs faux négatif



Faux positif vs faux négatif



Sensibilité - Spécificité

- On suppose que les individus normaux ont la même fonction de répartition F
- On suppose que les individus atteints ont la même fonction de répartition G

Définition

- Sensibilité : $\mathrm{Se}(q) = 1 - G(q)$ $\quad$ $(1-$ risque de 2^{nde} espèce$)$
- Spécificité : $\mathrm{Sp}(q) = F(q)$ $\quad$ $(1-$ risque de 1^{re} espèce$)$

Courbe ROC

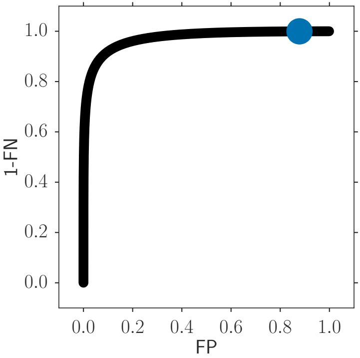
Définition

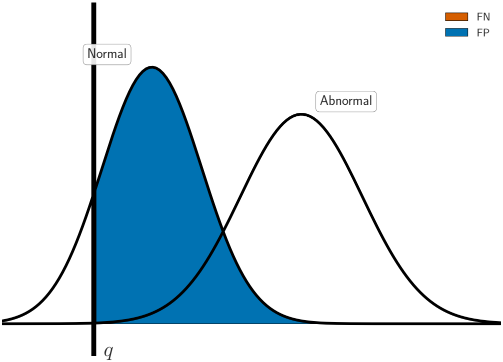
La courbe ROC est la courbe décrite par $(1 - \mathrm{Sp}(q), \mathrm{Se}(q))$, quand $q \in \mathbb{R}$. C'est donc la fonction $[0,1] \to [0,1]$

$$\mathrm{ROC}(t) = 1 - G(F^{-}(1-t))$$

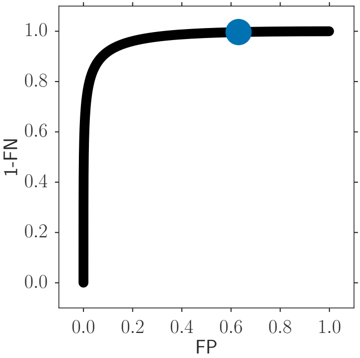
où $F^{-}(1-t) = \inf\{x \in \mathbb{R} \,:\, F(x) \geqslant 1-t\}$.

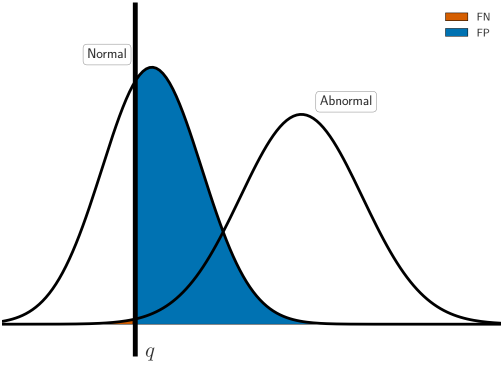
ROC curve



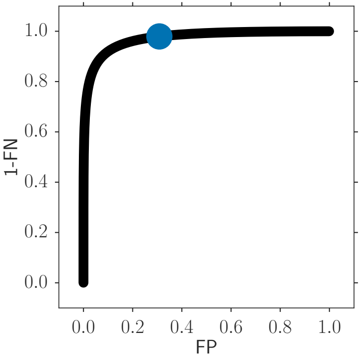


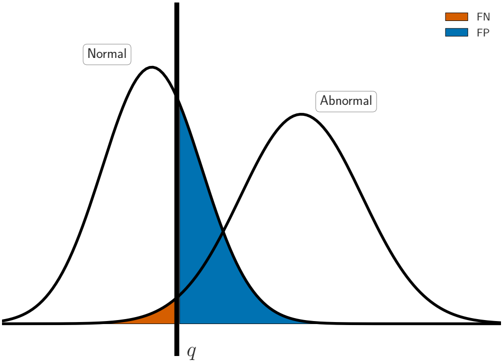
ROC curve



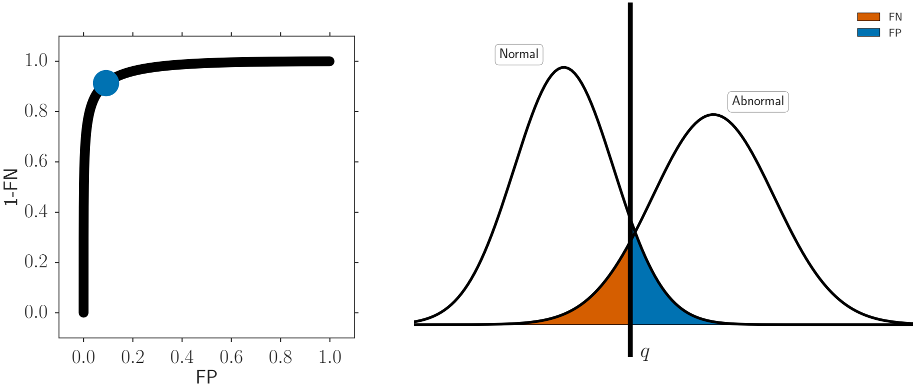


ROC curve

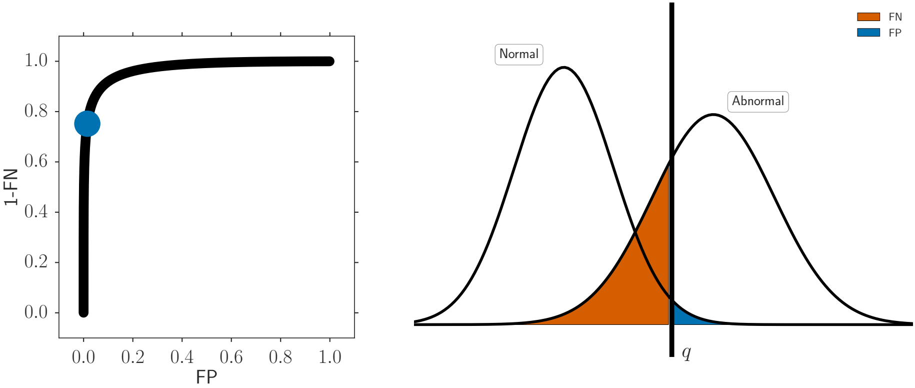




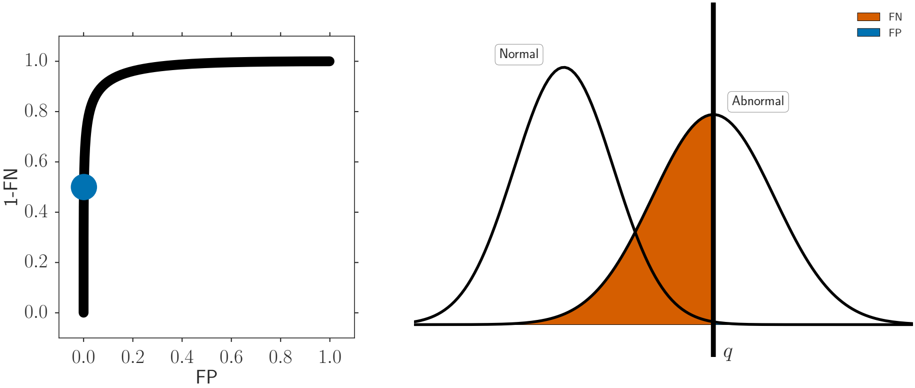
ROC curve



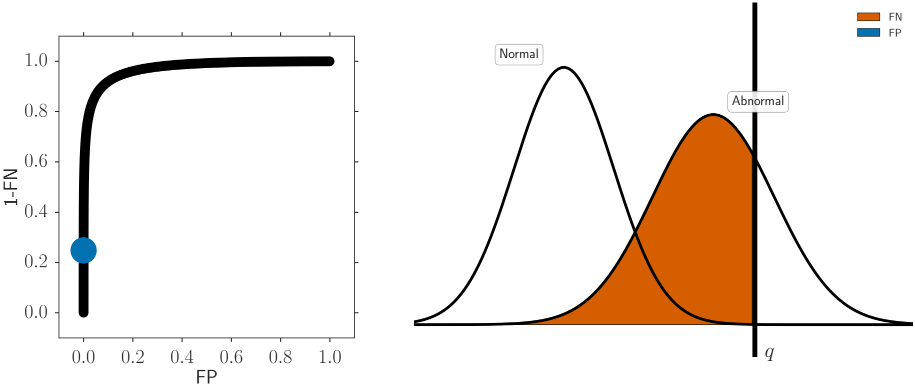
ROC curve



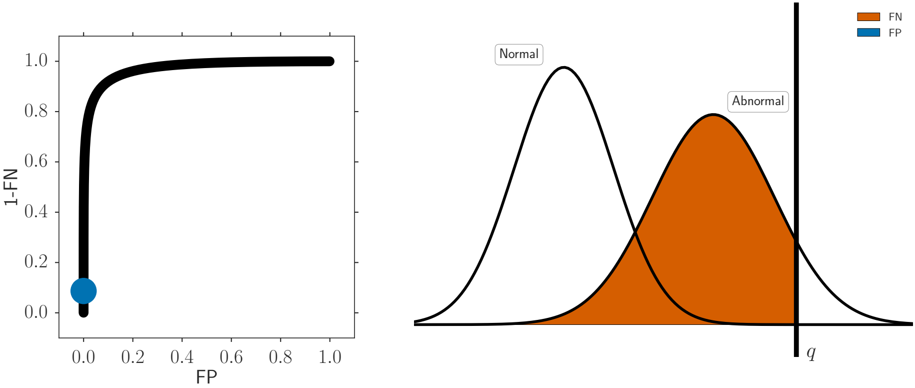
ROC curve



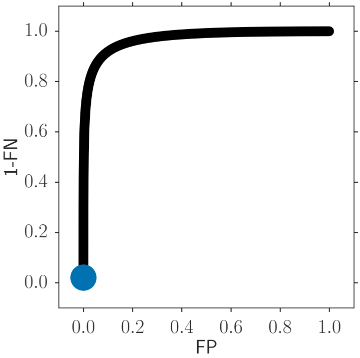
ROC curve

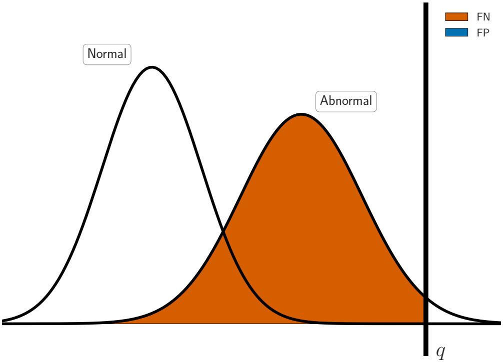


ROC curve



ROC curve

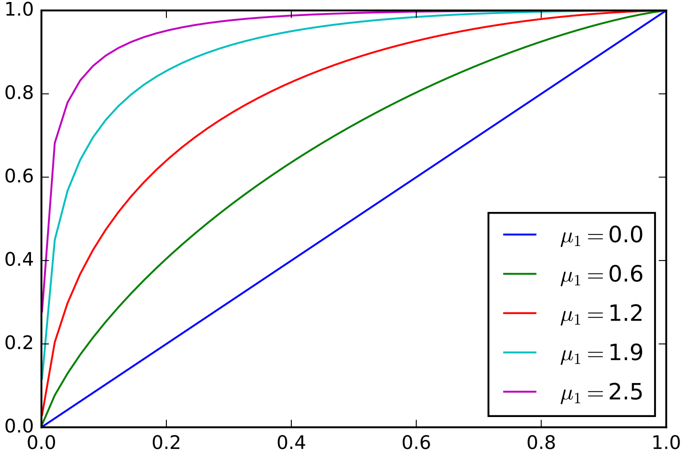




Sommaire

La courbe ROC dans le cas bi-normal

- F et G sont des gaussiennes de paramètres μ_0, σ_0 et μ_1, σ_1, respectivement.
- On spécifie $\mu_0 = 0$, $\sigma_0 = \sigma_1 = 1$, on fait varier μ_1



Estimation–application

Estimation de la courbe ROC

- Maximum de vraisemblance
- Non-paramétrique
- Bayésien avec variable d'état latente
- Estimation de l'aire sous la courbe ROC

Application

- Pour comparer différents tests statistiques
- Pour comparer différents algorithmes d'apprentissage supervisé
- Pour comparer des méthodes de sélection de support du Lasso

(: *ROC=Receiver Operating Characteristic*)

Références I

- B. Efron, T. Hastie, I. M. Johnstone, and R. Tibshirani.
  Least angle regression.
  Ann. Statist., 32(2) :407–499, 2004.
  With discussion, and a rejoinder by the authors.

- A. B. Tsybakov.
  Statistique appliquée, 2006.
  http://josephsalmon.eu/enseignement/ENSAE/StatAppli_tsybakov.pdf.

- Tong Zhang.
  Adaptive forward-backward greedy algorithm for sparse learning with linear models.
  In *Advances in Neural Information Processing Systems*, pages 1921–1928, 2009.