MS BGD
MDI 720 : Statistiques

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

Sommaire

Sommaire

Intervalle de confiance

- Contexte : on a une estimation $\hat{g}(y_1, \ldots, y_n)$ d'une grandeur $g(\theta)$. On veut un intervalle $\hat{I}$ autour de $\hat{g}$ qui contient g avec une grande probabilité.
- On construit $\hat{I} = [A, B]$ à partir des observations $(y_1, \ldots, y_n)$: l'intervalle est une variable aléatoire

$$\mathbb{P}(\hat{I} \text{ contient } g) = \mathbb{P}(A \leqslant g \text{ et } B \geqslant g) = 95\%$$

Intervalle de confiance de niveau α

Intervalle de confiance

Un intervalle de confiance de niveau α pour la grandeur $g = g(\theta)$ est une fonction de l'échantillon

$$\hat{I} : (y_1, \ldots, y_n) \mapsto \hat{I} = [A(y_1, \ldots, y_n), B(y_1, \ldots, y_n)]$$

telle que, quelle que soit le paramètre $\theta \in \Theta$,

$$\mathbb{P}\left[g(\theta) \in \hat{I}(y_1, \ldots, y_n)\right] \geqslant 1 - \alpha \qquad \text{lorsque } y_i \sim \mathbb{P}_\theta$$

Rem: des choix classiques sont $\alpha = 5\%, 1\%, 0.1\%$, etc.
Rem: Dans la suite on notera IC pour Intervalle de Confiance

Exemple : sondage

- Sondage d'une élection à deux candidats : A et B. Le choix du i-ème sondé suit une loi de Bernoulli de paramètre p, $y_i = 1$ s'il vote A, 0 sinon. Ainsi,

$$\Theta = [0, 1] \text{ et } \theta = p.$$

- But : estimer $g(\theta) = p$.
- échantillon de taille n : un estimtateur raisonnable est alors

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} y_i = \overline{y}_n$$

intervalle de confiance pour p ?

Sondage : intervalle de confiance

- Chercher un intervalle $\hat{I} = [\hat{p} - \delta, \hat{p} + \delta]$ tel que $\mathbb{P}(p \in \hat{I}) \geqslant 0.95 \Leftrightarrow$ chercher δ tel que $\mathbb{P}\left[\,|\hat{p} - p|\ > \delta\right] \leqslant 0.05$
- Ingrédient : inégalité de **Tchebyschev** (si $\mathbb{E}(X^2) < +\infty$)

$$\boxed{\forall \delta > 0, \quad \mathbb{P}(|X - \mathbb{E}(X)| > \delta) \leqslant \frac{\mathrm{Var}(X)}{\delta^2}}$$

Pour $X = \hat{p} = \frac{1}{n}\sum_{i=1}^n Y_i$ on a $\mathbb{E}(\hat{p}) = p$ et $\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n}$:

$$\forall p \in (0,1), \forall \delta > 0, \quad \mathbb{P}(|\hat{p} - p| > \delta) \leqslant \frac{p(1-p)}{n\delta^2} \leqslant \frac{1}{4n\delta^2}$$

<u>**Application numérique :**</u> pour un IC à 95%, choisir δ tel que $\frac{1}{4n\delta^2} = 0.05$, *i.e.*, $\delta = (0.2n)^{-1/2}$. Si $n = 1000$, $\hat{p} = 55\%$:

$$\delta = 0.07 \; ; \quad \hat{I} = [0.48, 0.62]$$

Sommaire

Théorème central limite

- $y_1, y_2, \dots,$ des variables aléatoires *i.i.d.* de carré intégrable.
- μ et σ leur espérance et écart-type théoriques.

Théorème central limite (TCL)

La loi de la moyenne empirique renormalisée $\sqrt{n}\left(\frac{\bar{y}_n - \mu}{\sigma}\right)$ converge vers une loi normale centrée réduite $\mathcal{N}(0,1)$

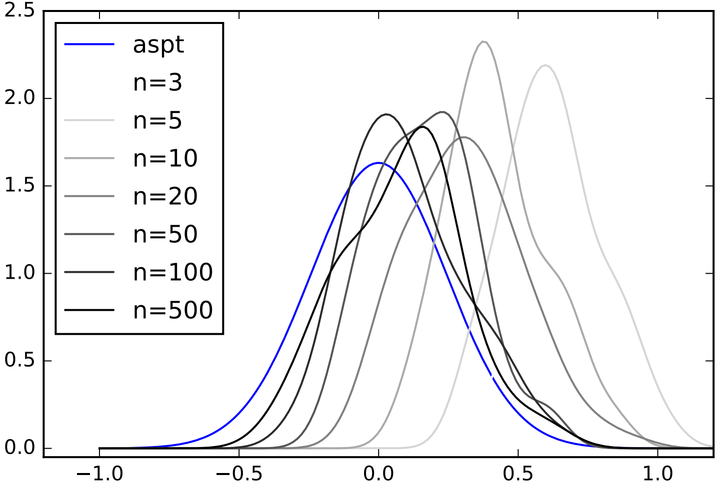
- σ is unknown

Lemme de Slutsky

La loi de la moyenne empirique "studentizée" $\sqrt{n}\left(\frac{\bar{y}_n - \mu}{\hat{\sigma}}\right)$ converge vers une loi normale centrée réduite $\mathcal{N}(0,1)$ quand $\hat{\sigma} \to \sigma$

Reformulation : $\bar{y}_n \simeq \mathcal{N}(\mu, \hat{\sigma}^2/n)$

Illustration



Intervalles de confiance asymptotiques

- Exemple du sondage : $y_i \in \{0, 1\}$, $n = 1000$,

$$\hat{p} = n^{-1} \sum_{i=1}^{n} y_i = 0.55$$

- On suppose que n est suffisamment grand pour que

$$\sqrt{n} \left(\frac{\hat{p} - p}{\hat{\sigma}} \right) \sim \mathcal{N}(0, 1)$$

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (y_i - \hat{p})^2 = \hat{p} - \hat{p}^2$$

- On connaît les quantiles de la loi normale (numériquement)
- D'après le TCL, et l'approximation des quantiles gaussiens

$$\mathbb{P} \left[-1.96 < \sqrt{n} \, \frac{0.55 - p}{\hat{\sigma}} < 1.96 \right] \approx 0.95$$

nouvel intervalle de confiance : $\hat{I} = [0.52, 0.58]$: meilleur ! (**plus optimiste**)

En Python

```
import numpy as np
from scipy.stats import norm
```

```
x = np.random.binomial(1,.5,n)
pchap = np.mean (x)
sig = np.sqrt(pchap * (1 - pchap))
q = norm.ppf(.975)
borneinf = pchap - sig * q / np.sqrt(n)
bornesup = pchap - sig * (1-q) /np.sqrt(n)
print('IC=[' + str(borneinf) +
      ',' + str(bornesup) + ']')
```

Sommaire

IC pour les moindres carrés (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 :

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

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.

IC pour les moindres carrés (II)

Sous l'hypothèse gaussienne, comme

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

et en notant $t_{1-\alpha/2}$ un quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{T}_{n-\mathrm{rg}(X)}$, alors l'intervalle de confiance suivant est de niveau α

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

pour la quantité θ_j^*.
Rem: $\mathbb{P}(|T_j| < t_{1-\alpha/2}) = 1-\alpha$ car la loi de Student est symétrique

Limites des IC précédent

Dans la partie précédente, l'intégralité des raisonnement repose sur le modèle gaussien.
Attention : si le modèle est (trop) faux alors les IC obtenus ne seront pas forcément pertinents.

Une alternative possible est le *bootstrap*, qui est une méthode non-paramétrique reposant sur le ré-échantillonnage, bien fondée (théoriquement) pour des statistiques régulières telle que la moyenne, les quantiles, etc., mais pas pour le max ou le min.
Pour aller plus loin : Efron et Tibshirani (1994)

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 $(y_1, \ldots, y_n) \in R$ 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 $\Theta_1 \subset \Theta$

<u>Rem</u>: $\{\theta \in \Theta_0\}$ 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 mauvaise détection

$$\alpha = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta((y_1, \ldots, y_n) \in R)$$

- Risque de 2^{nde} espèce : probabilité de fausse alarme

$$\sup_{\theta \in \Theta_1} \mathbb{P}_\theta\left((y_1, \ldots, y_n) \notin R\right)$$

<u>Rem</u>: Pour le vocabulaire, prendre l'exemple H_0 un missile arrive vs. H_1 il n'y pas de missile, et ainsi comprendre le risque de fausse alarme

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 $R = [-1.96, 1.96]^C =]-\infty, -1.96[\cup]1.96, +\infty[$.

- **Exemple numérique** : 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

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

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 $\Theta_0 = \{\theta \in \mathbb{R}^p : \theta_j = 0\}$.
Sous H_0 on connaît donc la distribution de $\hat{\theta}_j$:

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

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

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

Contexte médical

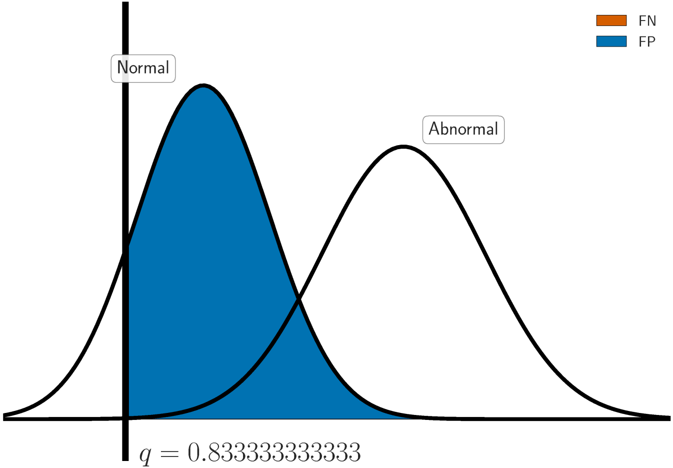
- Un groupe de patients $i = 1, \ldots, n$ est suivi pour un dépistage.
- Pour chaque individu, le test se base sur une variable aléatoire $X_i \in \mathbb{R}$ et un seuil $q \in \mathbb{R}$

dès lors que	$X_i > q$	le test est **positif**
sinon		le test est **négatif**

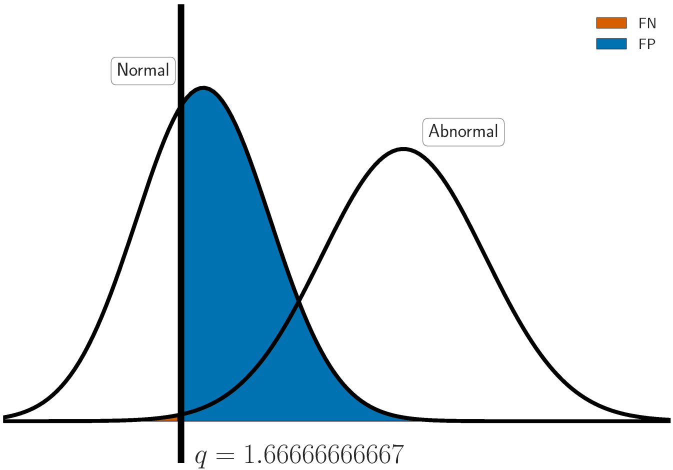
Ensemble des configurations possibles

	Normal H_0	Atteint H_1
négatif	vrai négatif	faux négatif
positif	faux positif	vrai positif

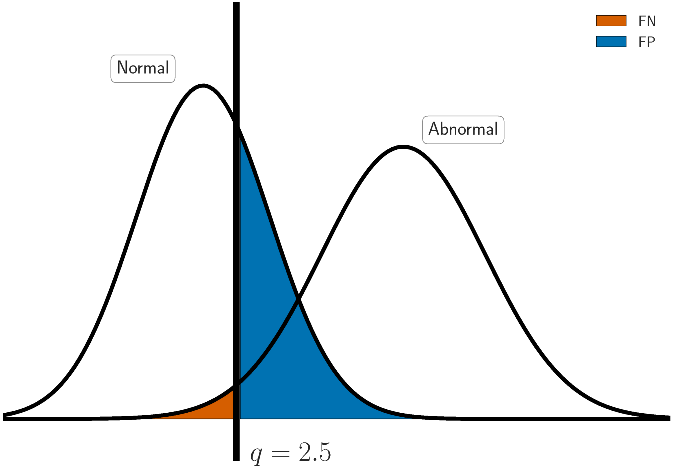
Faux positif vs faux négatif



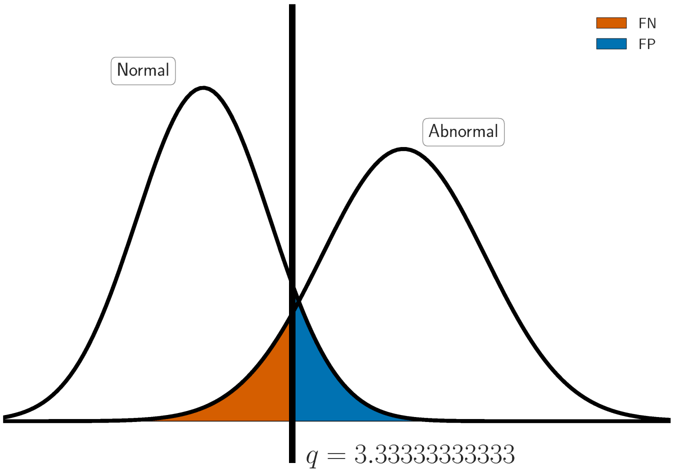
Faux positif vs faux négatif



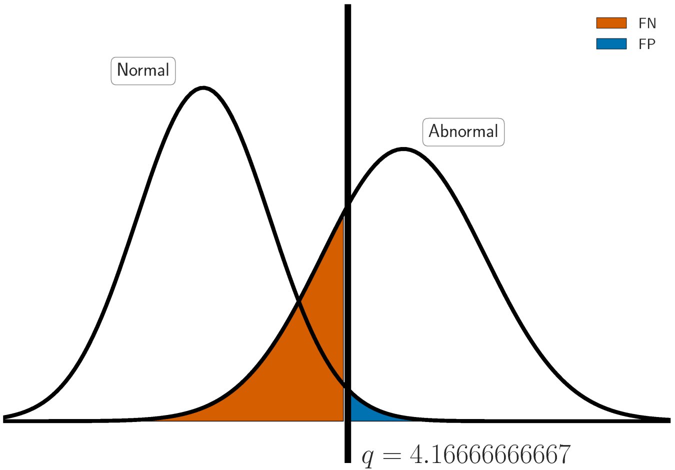
Faux positif vs faux négatif



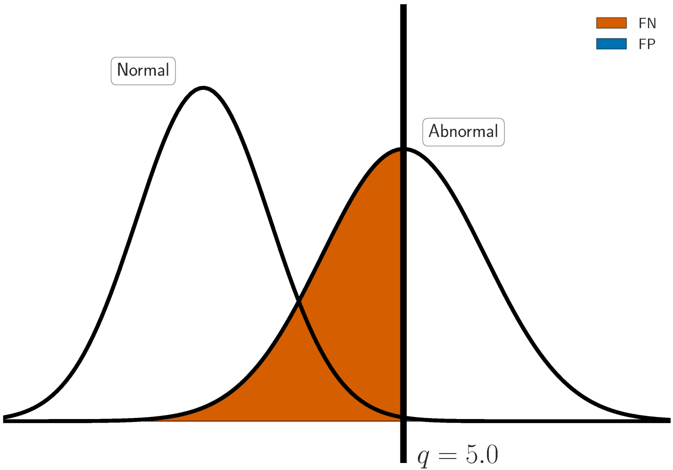
Faux positif vs faux négatif



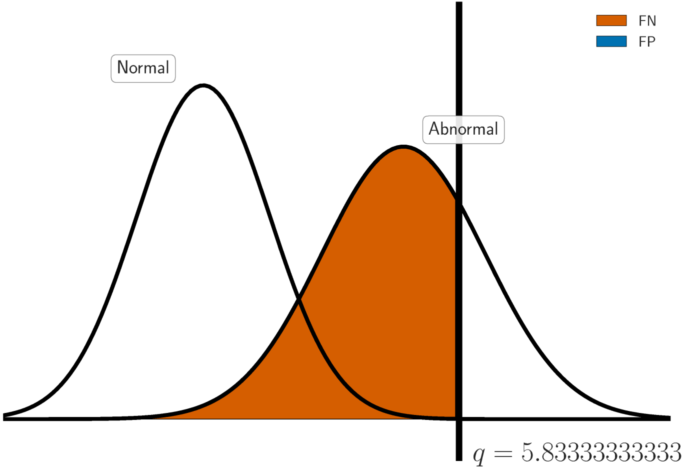
Faux positif vs faux négatif



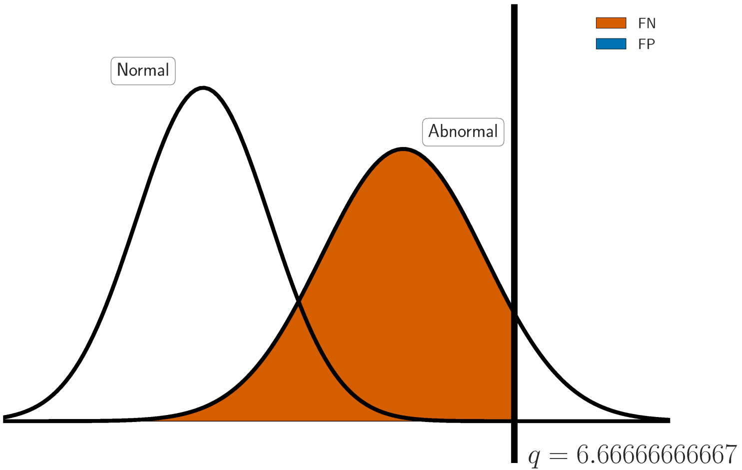
Faux positif vs faux négatif



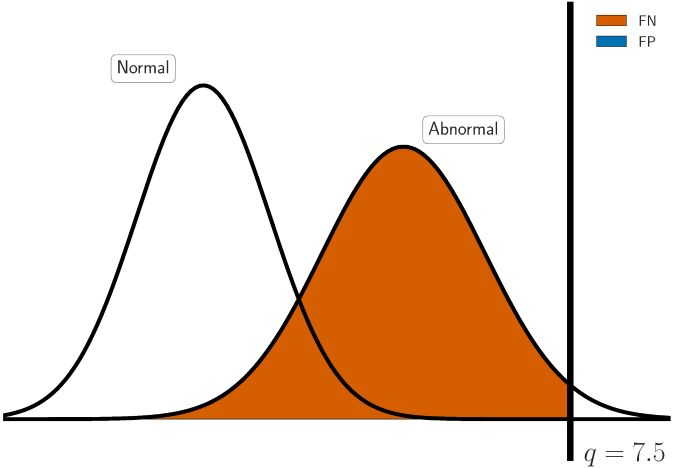
Faux positif vs faux négatif



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

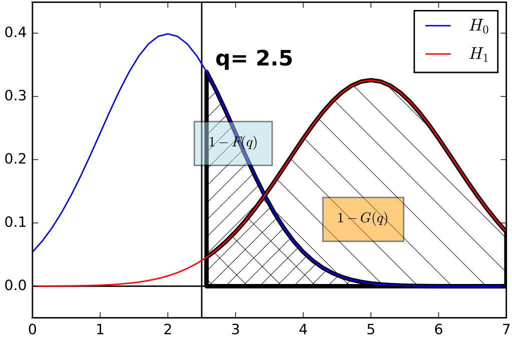
Courbe ROC

Définition

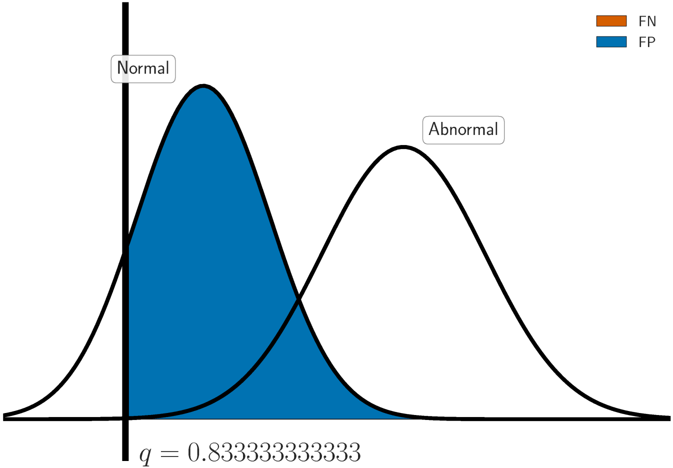
La courbe ROC est la courbe décrit par $(1 - \mathrm{Sp}(q), \mathrm{Se}(q))$, quand $q \in \mathbb{R}$. C'est donc la fonction $[0,1] \to [0,1]$

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

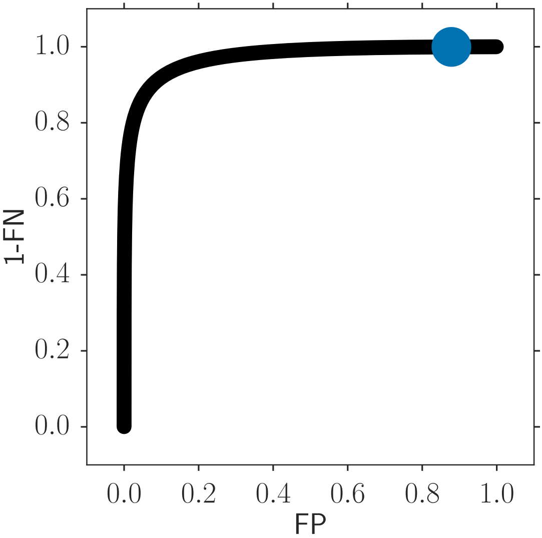
où $F^{-}(1-t) = \inf\{x \in \mathbb{R} \,:\, F(x) \geqslant 1-t\}$.



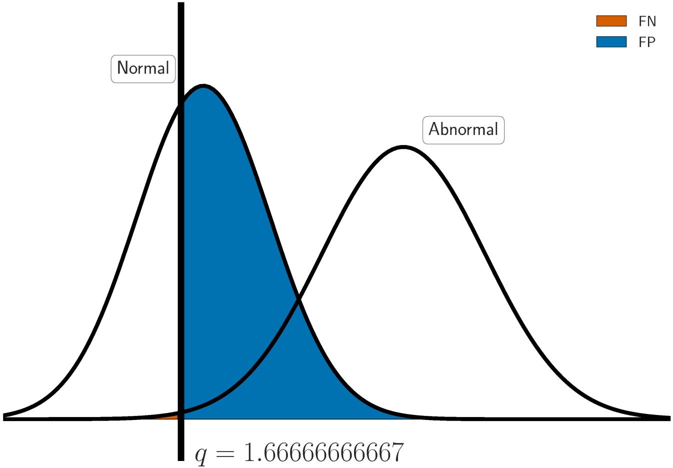
Courbe ROC



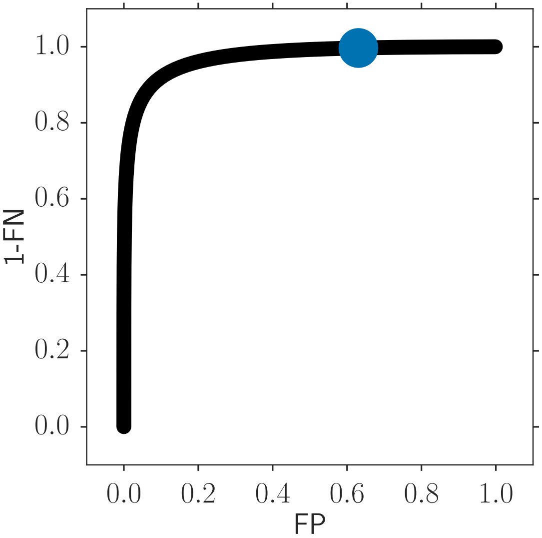
Courbe ROC



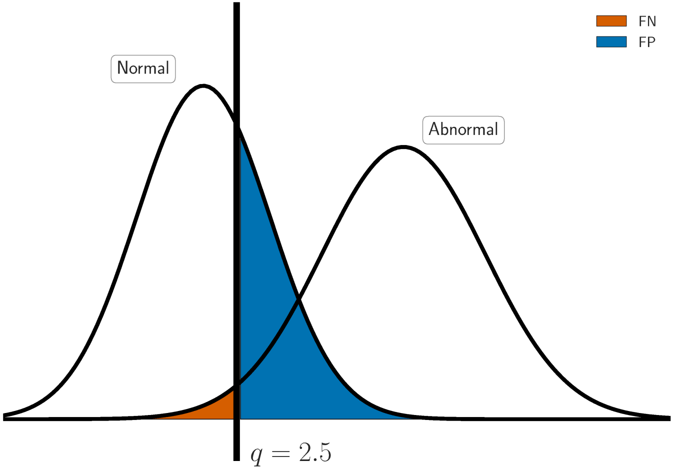
Courbe ROC



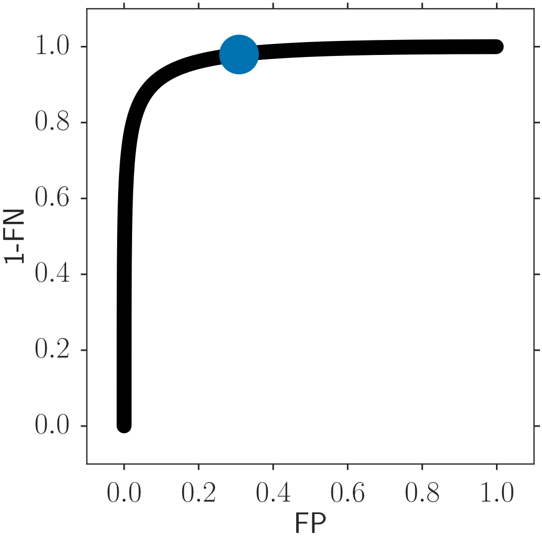
Courbe ROC



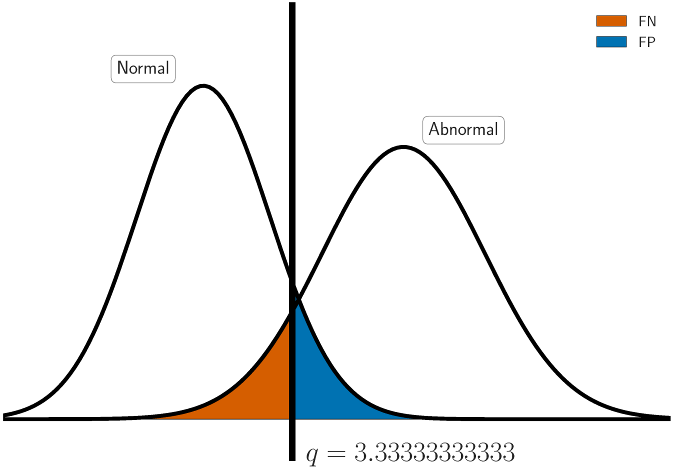
Courbe ROC



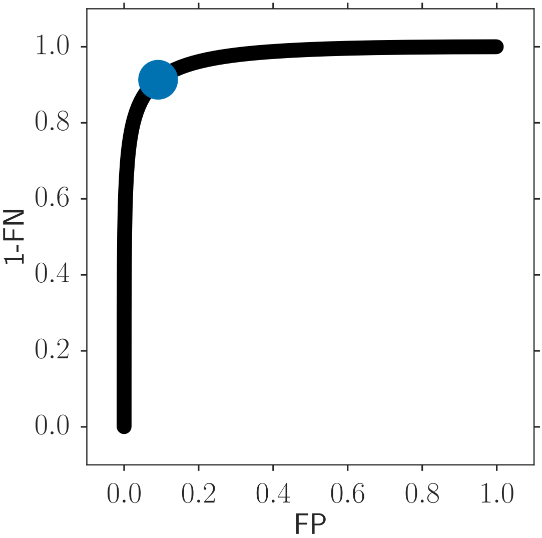
Courbe ROC



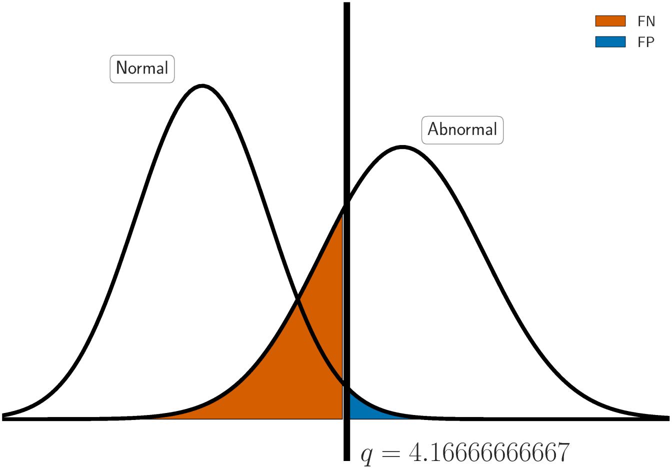
Courbe ROC



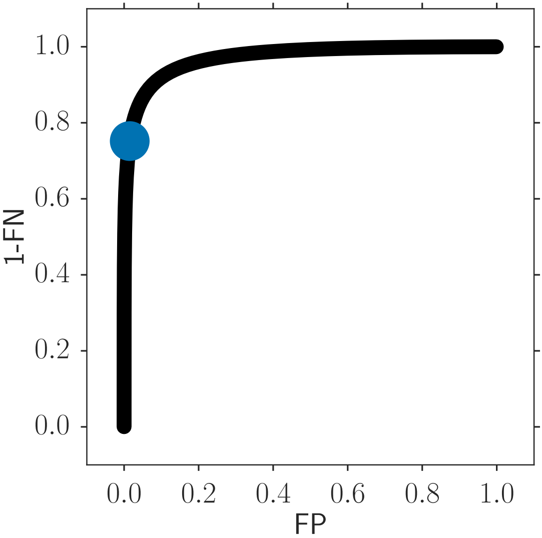
Courbe ROC



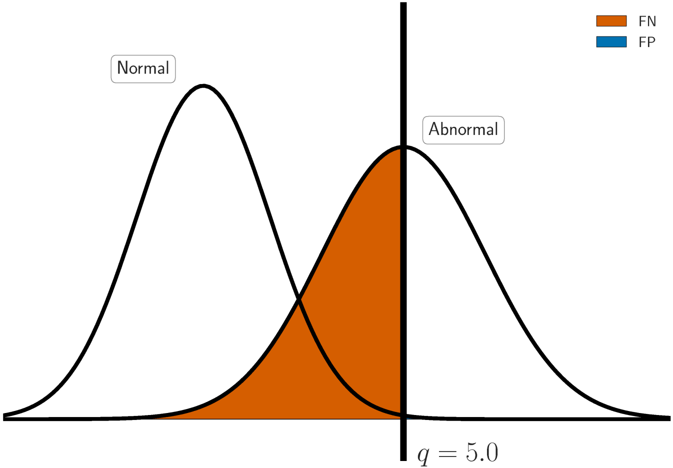
Courbe ROC



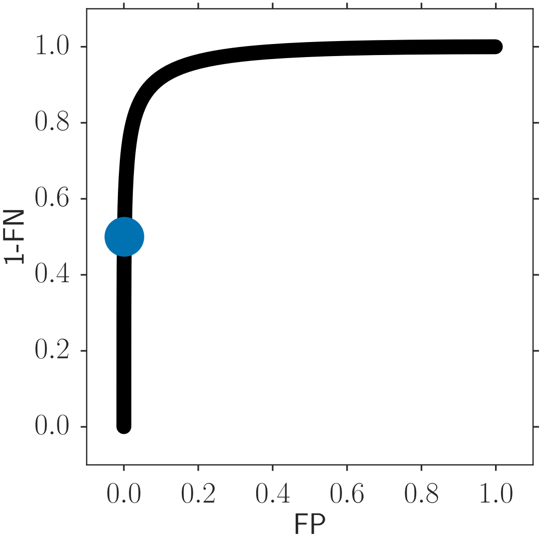
Courbe ROC



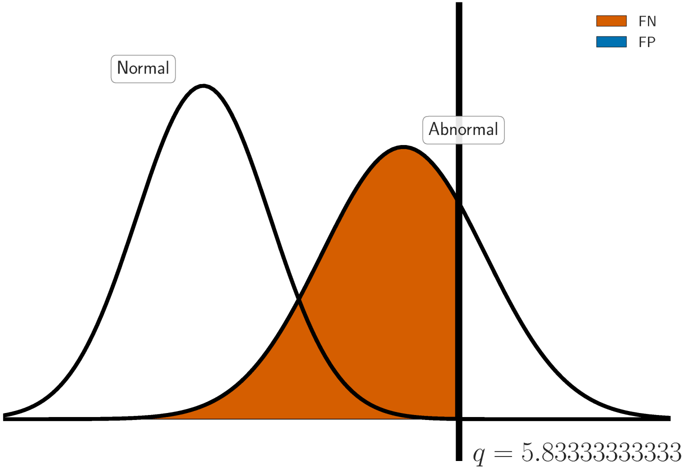
Courbe ROC



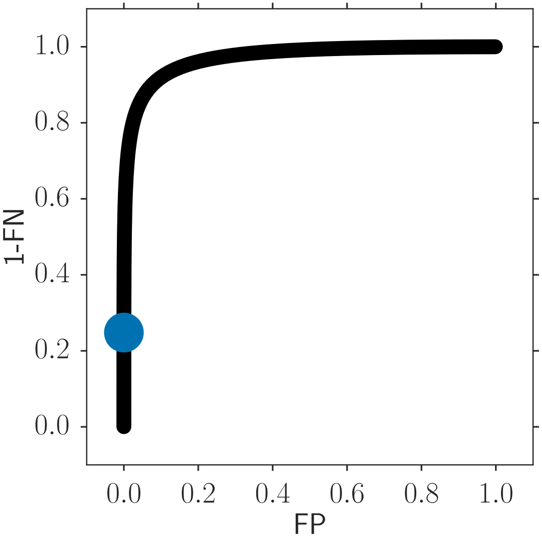
Courbe ROC



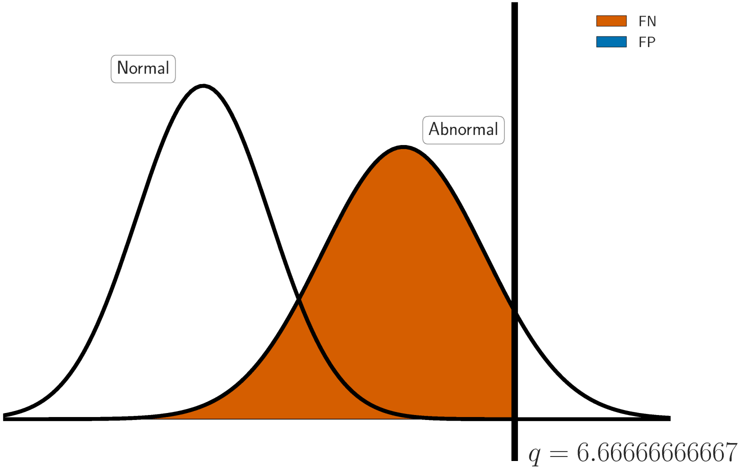
Courbe ROC



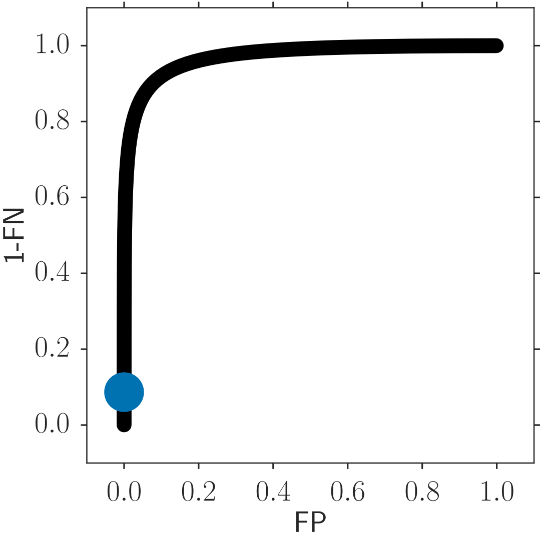
Courbe ROC



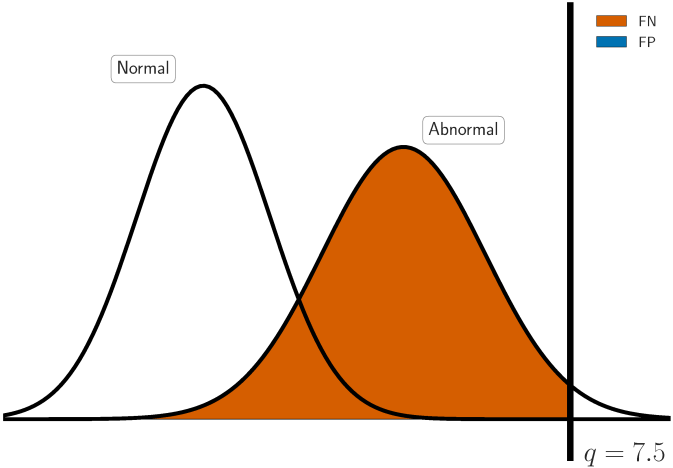
Courbe ROC



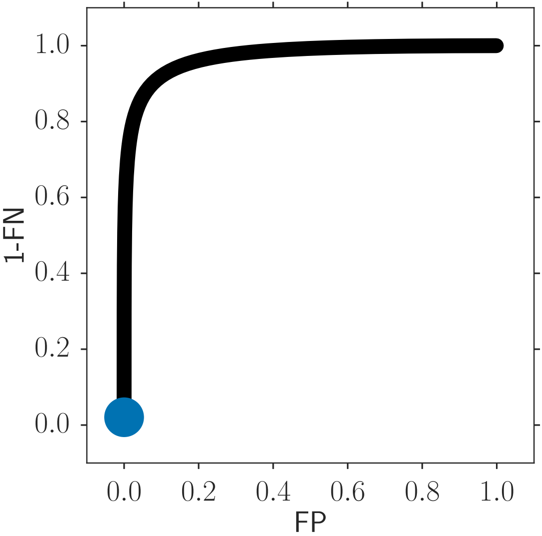
Courbe ROC



Courbe ROC



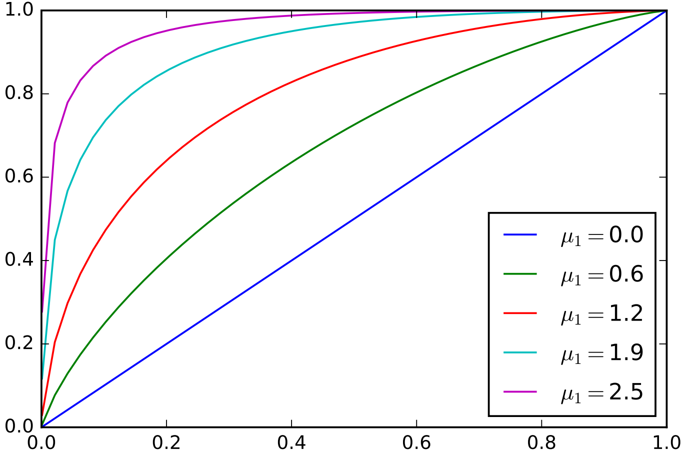
Courbe ROC



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.

nb : ROC = Receiver Operating Characteristic

Références I

- B. Efron and R. Tibshirani.
  An introduction to the bootstrap.
  CRC press, 1994.