MS BGD
MDI 720 : Statistiques

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

Plan

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 $\hat{I} = [\hat{A}, \hat{B}]$ à partir des observations $(y_1, \ldots, y_n)$: l'intervalle est une variable aléatoire

$$\mathbb{P}(\hat{I} \text{ contient } g) = \mathbb{P}(\hat{A} \leqslant g \text{ et } \hat{B} \geqslant g) = 1 - \alpha$$

Rem: souvent $1 - \alpha = 95\%$, car on veut que cette probabilité soit grande

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

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

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

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

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 : Albert et Bertrand. Le choix du i^{e} sondée suit une loi de Bernoulli de paramètre p

$$y_i = \begin{cases} 1, & \text{si le} i^{\text{e}} \text{ individu vote Albert} \\ 0, & \text{si le} i^{\text{e}} \text{ individu vote Bertrand} \end{cases}$$

Ainsi,

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

- But : estimer $g(\theta) = p$
- un estimateur raisonnable pour un échantillon de taille n :

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

Question : 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 $\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n}$:

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

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

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

Sommaire

Théorème central limite

- $y_1, \ldots, y_n$: 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 re-normalisée $\sqrt{n}\left(\frac{\bar{y}_n - \mu}{\sigma}\right)$ converge vers une loi normale centrée réduite $\mathcal{N}(0,1)$

- Reformulation : la moyenne empirique se comporte approximativement comme une loi normale $\mathcal{N}(\mu, \sigma^2/n)$

Intervalles de confiance asymptotiques

- Exemple du sondage : $\hat{p} = 0.55$, $n = 1000$
- On suppose que n est suffisamment grand pour que

$$\sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^{n} y_n - p}{\sqrt{p(1-p)}} \sim \mathcal{N}(0,1) \qquad \underline{\text{Rappel}} : p(1-p) = \mathrm{Var}(Y)$$

- D'après le TCL, et l'approximation des quantiles gaussiens

$$\mathbb{P}\left[-1.96 < \sqrt{n}\,\frac{0.55 - p}{\sqrt{p(1-p)}} < 1.96\right] \approx 0.95$$

On résout en p (équations de degré deux) :

$$\mathbb{P}\left[0.52 < p < 0.58\right] = 0.95$$

nouvel intervalle de confiance : $\hat{I} = [0.52, 0.58]$: meilleur !

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 :

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

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

La densité de la loi est connue explicitement, pour tout degré T :

$$\boxed{f_T(t) = \frac{1}{\sqrt{k\pi}} \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})} \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}}}$$

où Γ est la fonction Gamma d'Euler (qui extrapole les factoriels)

IC pour les moindres carrés (II)

Sous l'hypothèse gaussienne : $T_j = \dfrac{\hat{\theta}_j - \theta_j^\star}{\hat{\sigma}\sqrt{(X^\top X)^{-1}_{j,j}}} \sim \mathcal{T}_{n-\mathrm{rg}(X)}$

en notant $t_{1-\alpha/2}$ un quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{T}_{n-\mathrm{rg}(X)}$, l'IC suivant est de niveau α pour la quantité $\theta_j^\star$

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

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

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 complémentaire d' $\mathcal{H}_0$.
  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 $\mathcal{H}_0$.

  La région R est appelée 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 pas en général 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 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

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

Rem: pour le vocabulaire, prendre l'exemple H_0 : "un missile arrive" vs. H_1 : "il n'y pas de missile" (d'où le nom **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)

Rem: 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 sous $\mathcal{H}_0$: 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

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

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

Sa densité est connue explicitement pour un degré T :

$$\boxed{f_T(t) = \frac{1}{\sqrt{k\pi}} \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})} \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}}}$$

Tester la nullité des coefficients (I)

$H_0 : \theta_j^\star = 0$ 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 $\theta^\star_j$. Dire que "$0 \in IC_\alpha$" signifie que

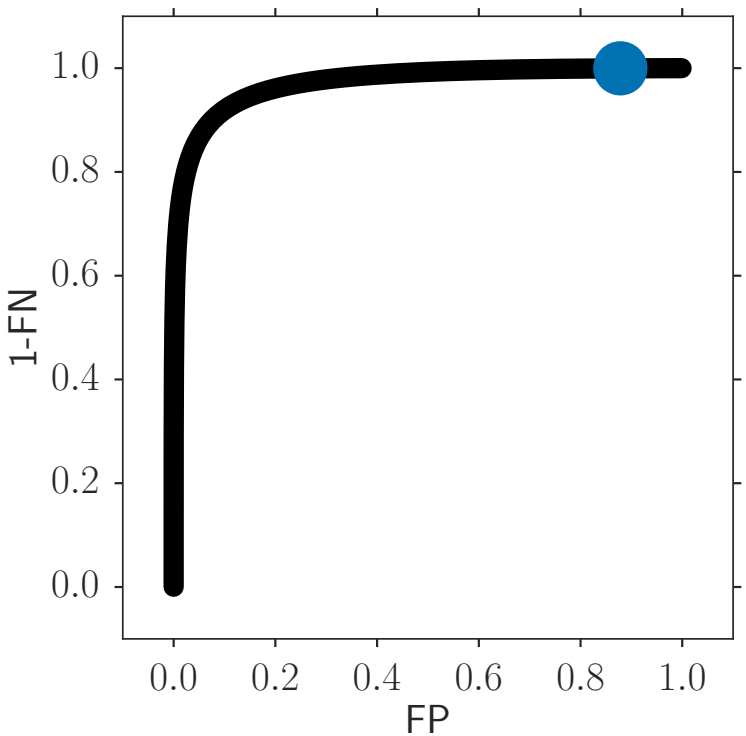
$$|\hat{\theta}_j| \leqslant t_{1-\alpha/2}\hat{\sigma}\sqrt{(X^\top X)^{-1}_{j,j}} \quad \Leftrightarrow \quad \frac{|\hat{\theta}_j|}{\hat{\sigma}\sqrt{(X^\top X)^{-1}_{j,j}}} \leqslant t_{1-\alpha/2}$$

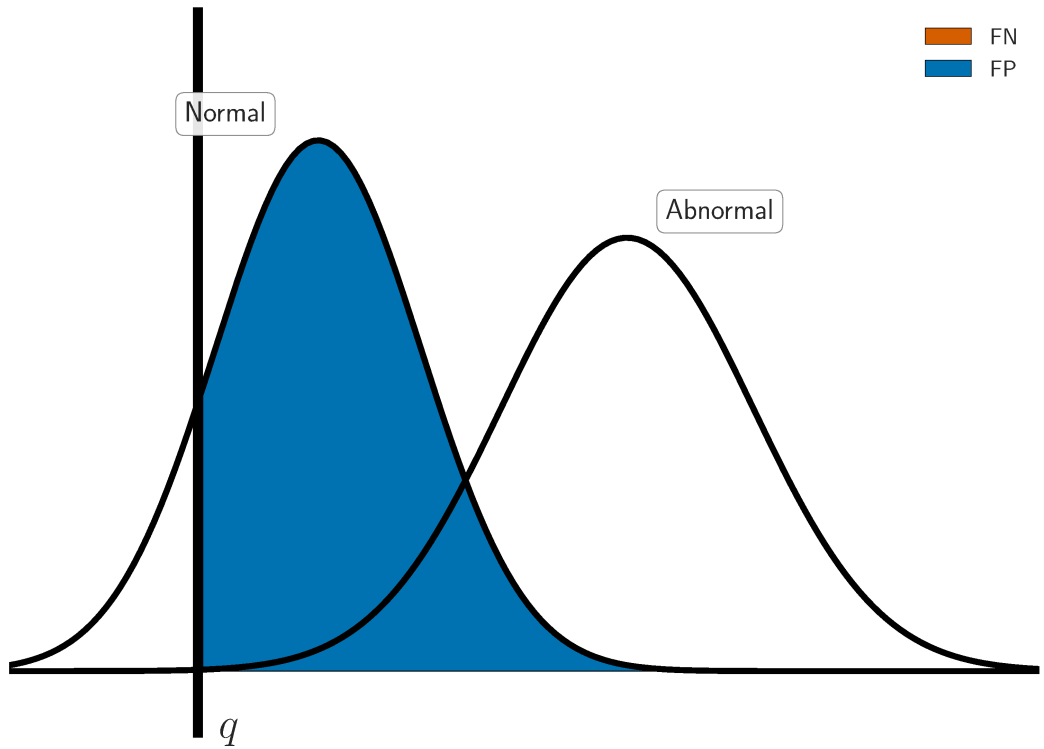
Cela est équivalent à accepter l'hypothèse $\theta^\star_j = 0$ au niveau α
Le α le plus petit tel que $0 \in IC_\alpha$ est appelé la **p-value**.

Rem: si α est proche de zéro un IC_α recouvre l'espace entier ; on peut trouver (par continuité) un α assurant l'égalité ci-dessus

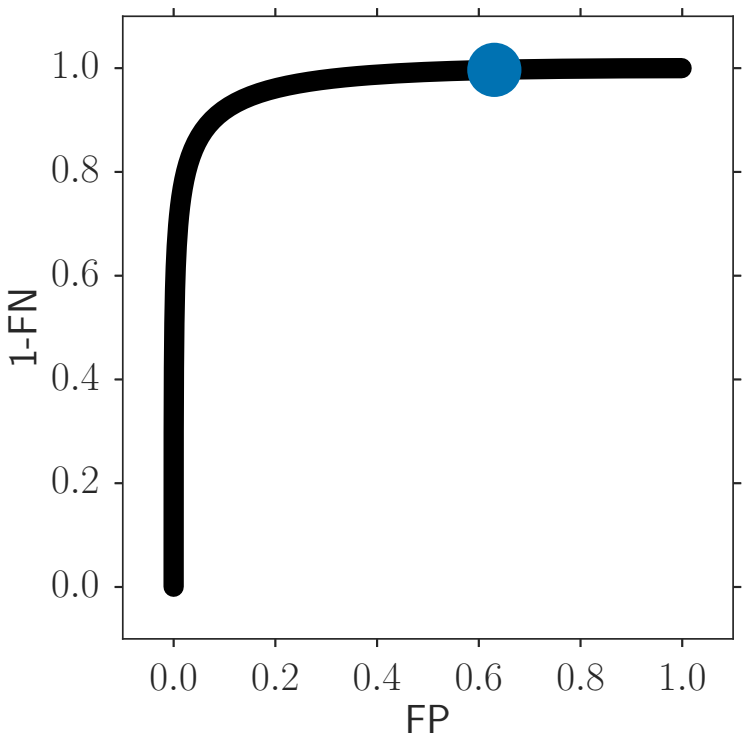
Sommaire

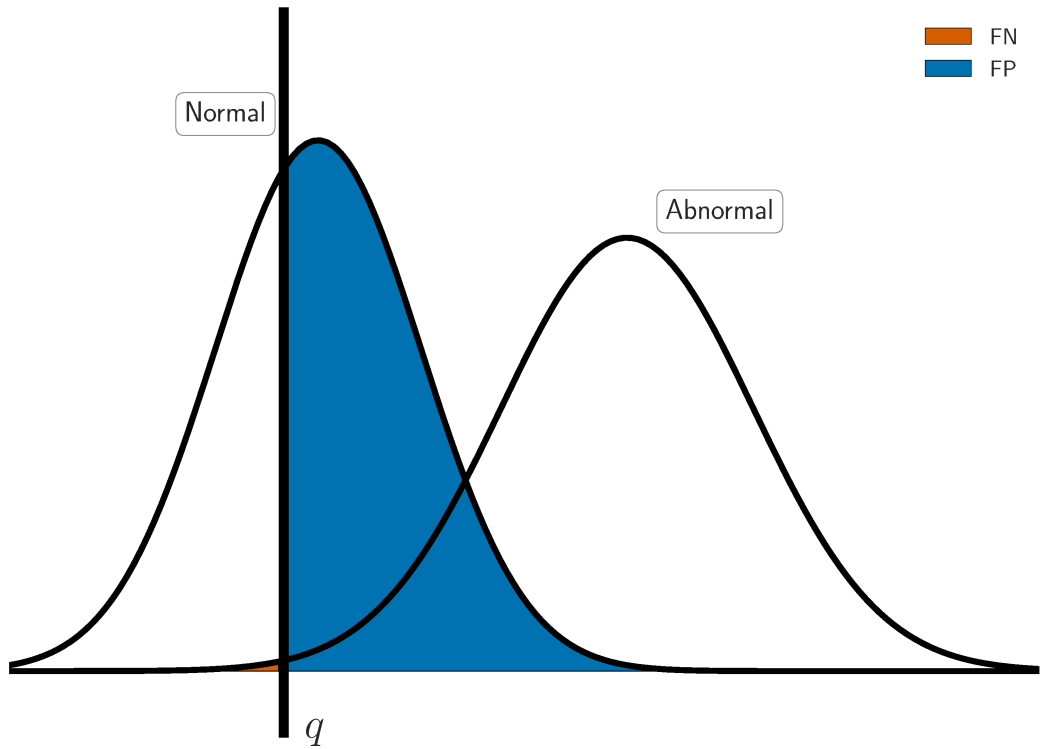
ROC curve



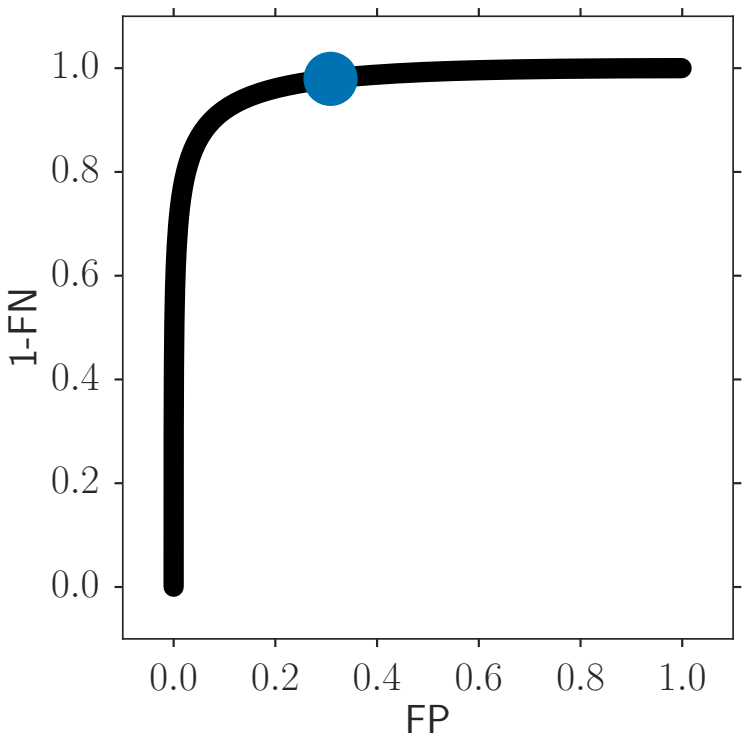


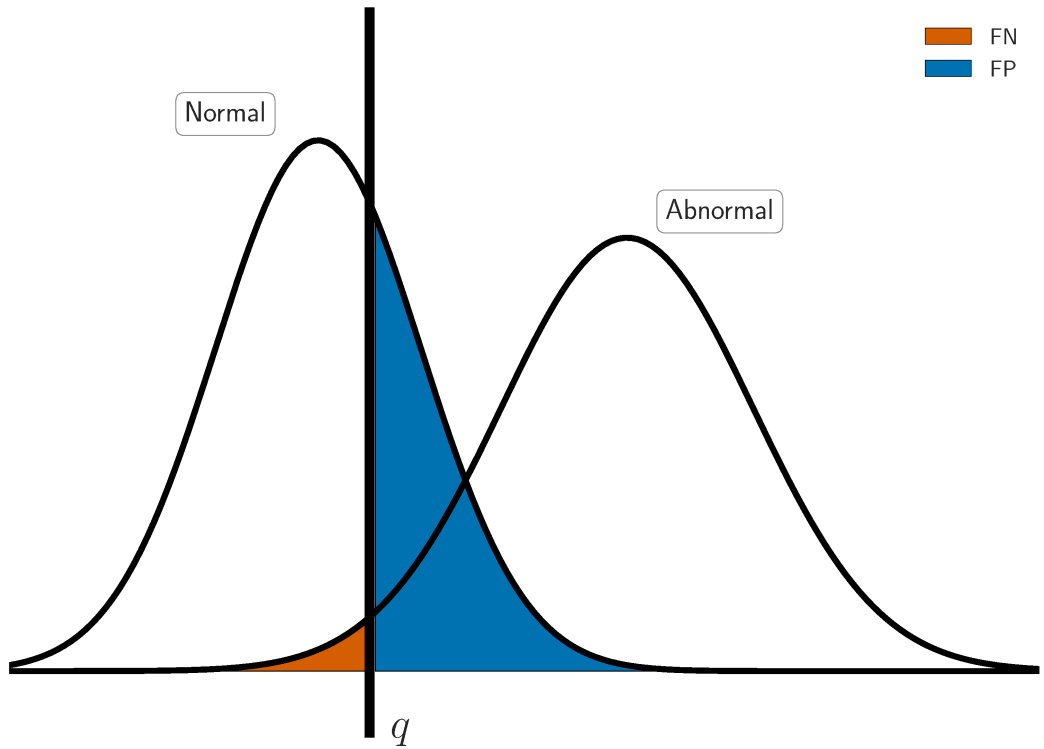
ROC curve



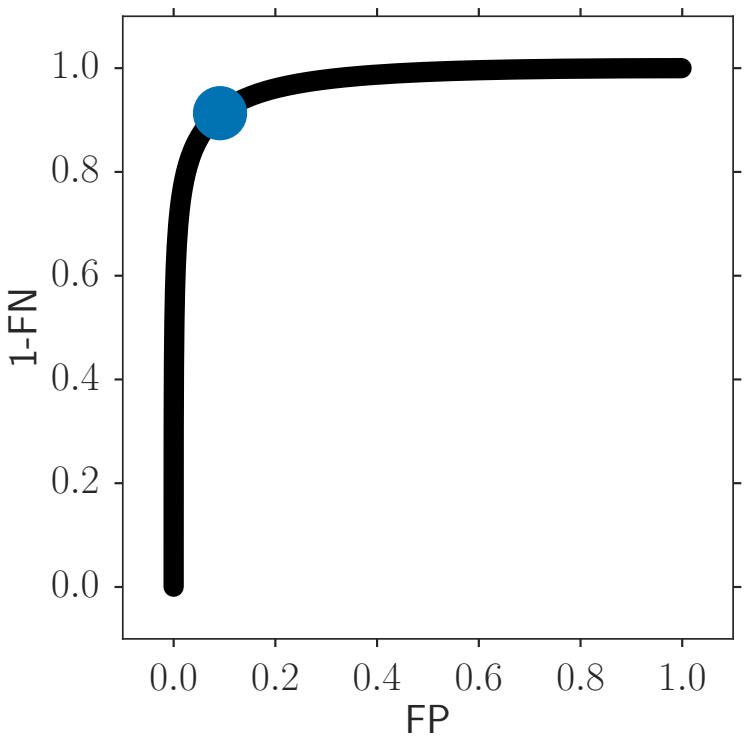


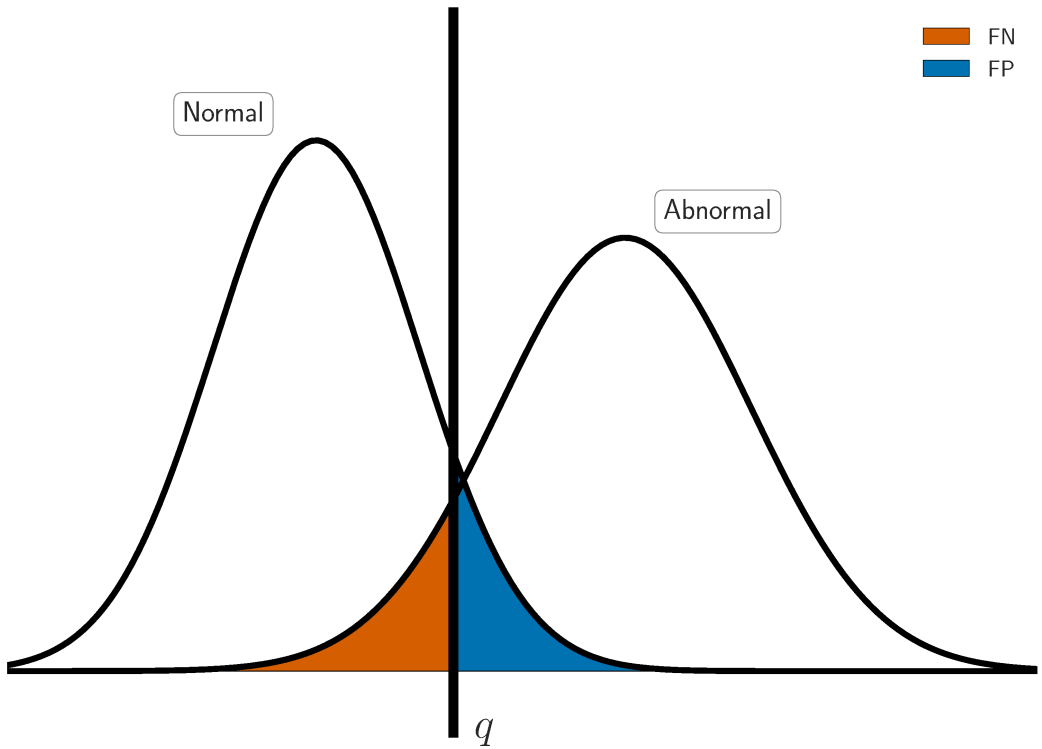
ROC curve



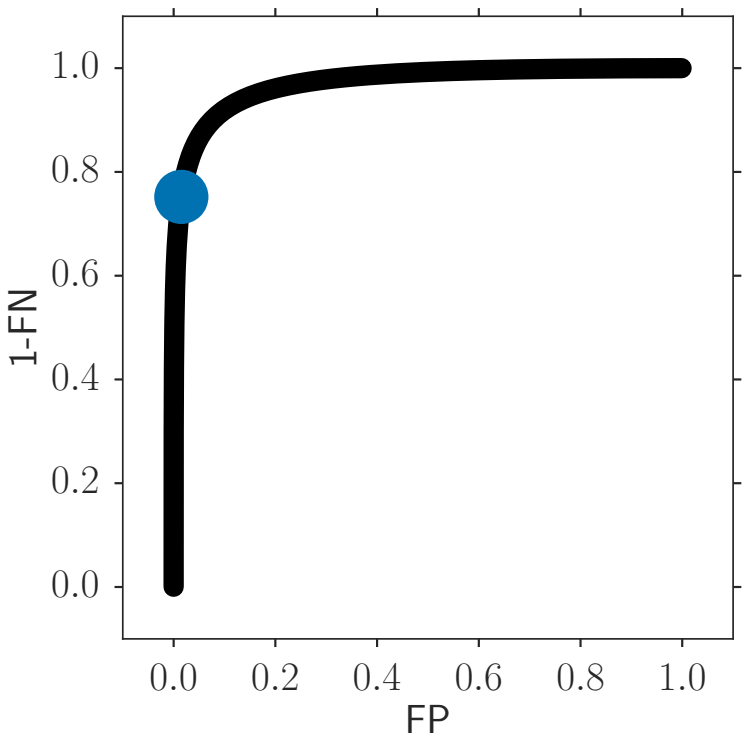


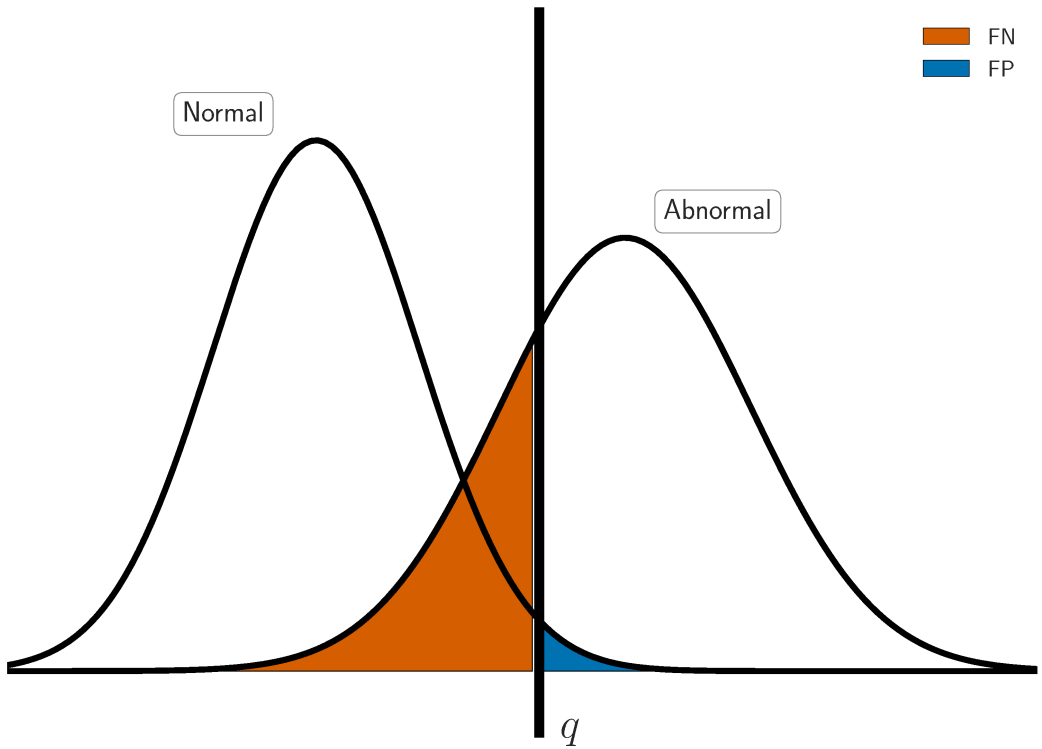
ROC curve



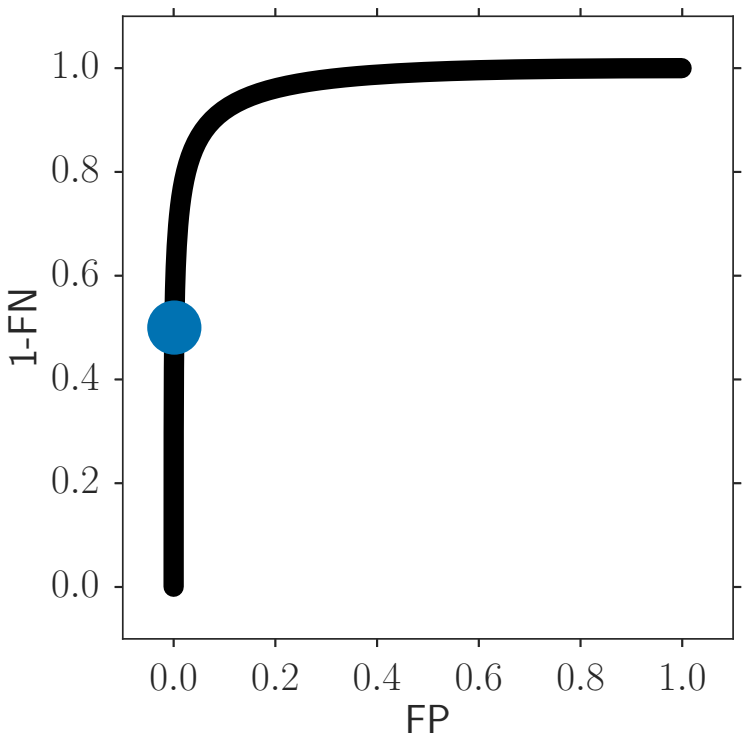


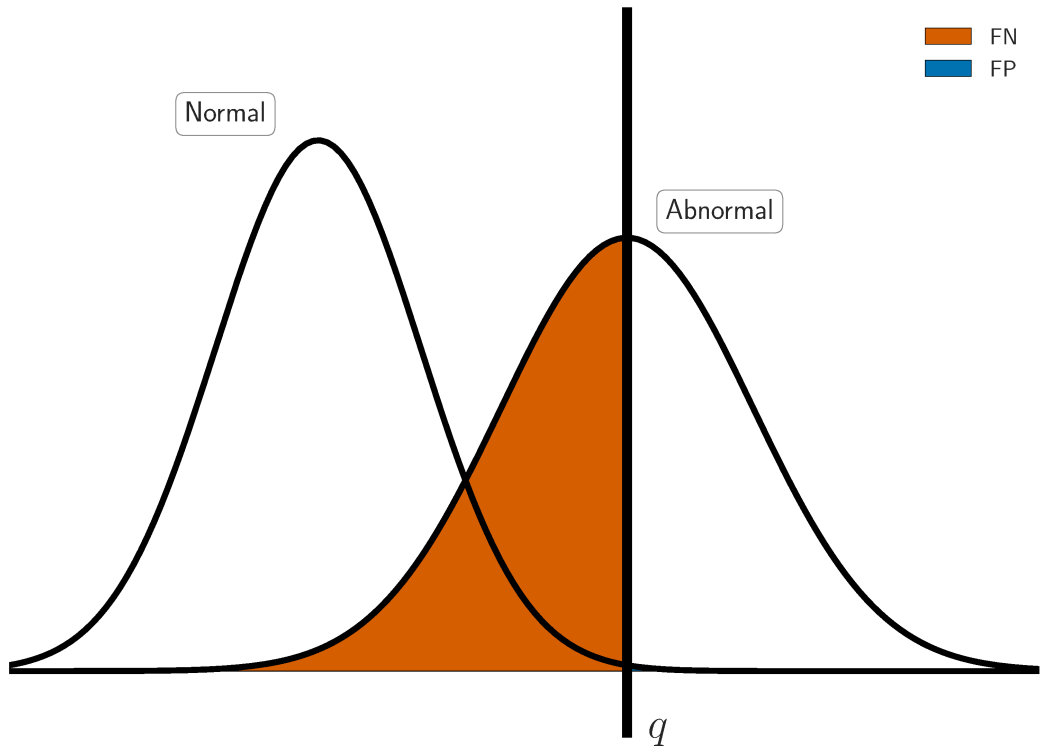
ROC curve



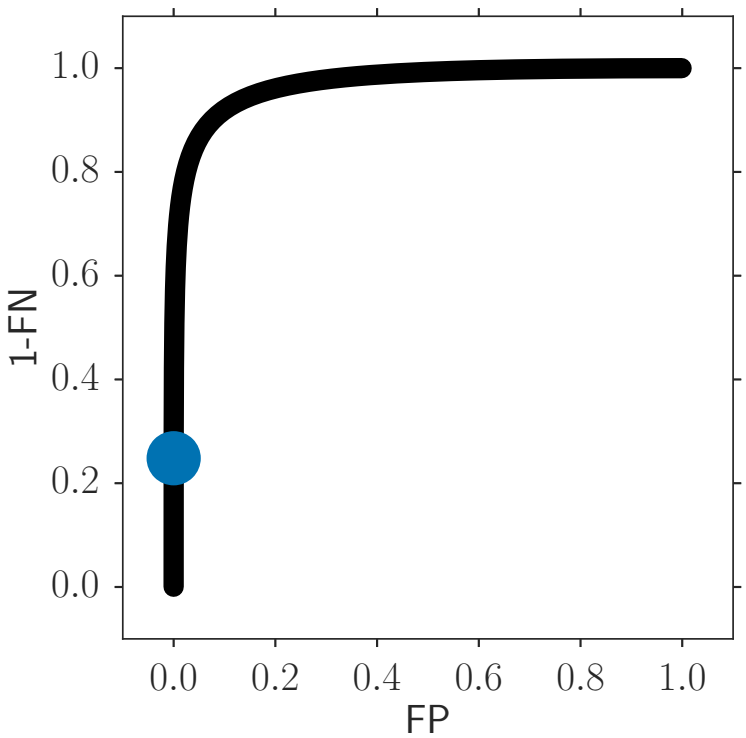


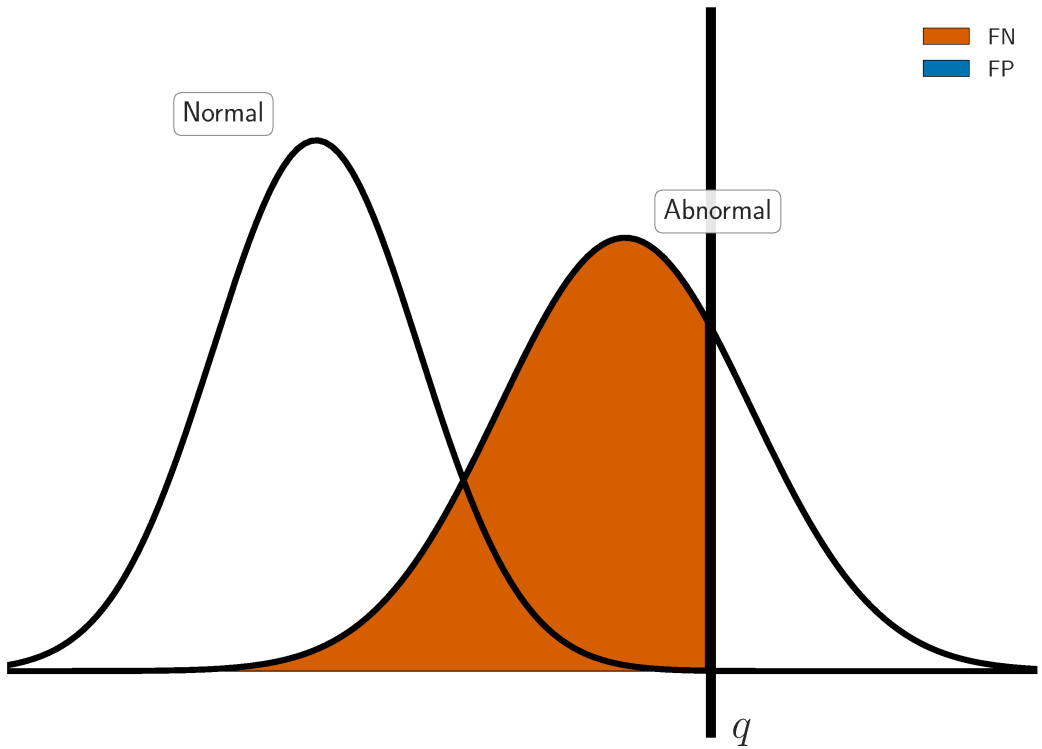
ROC curve



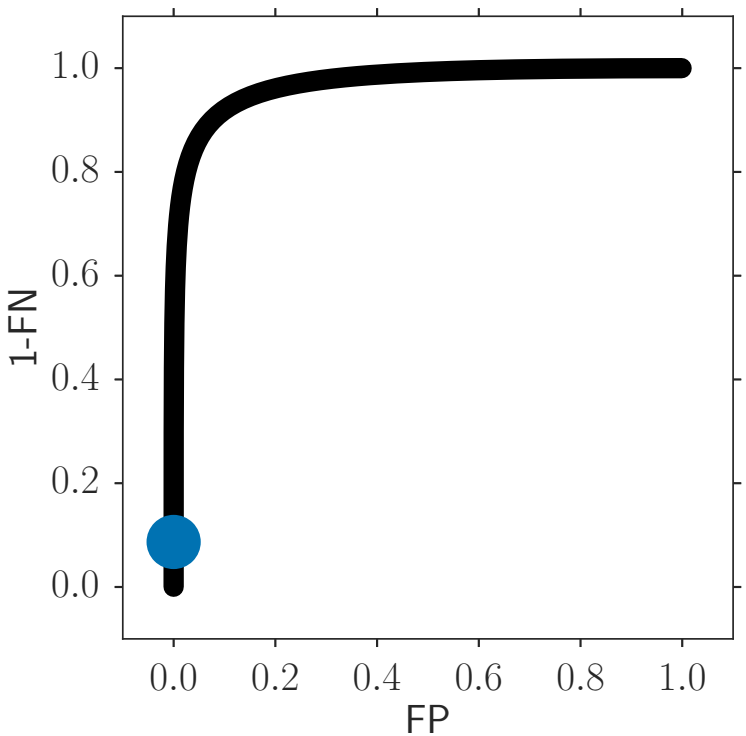


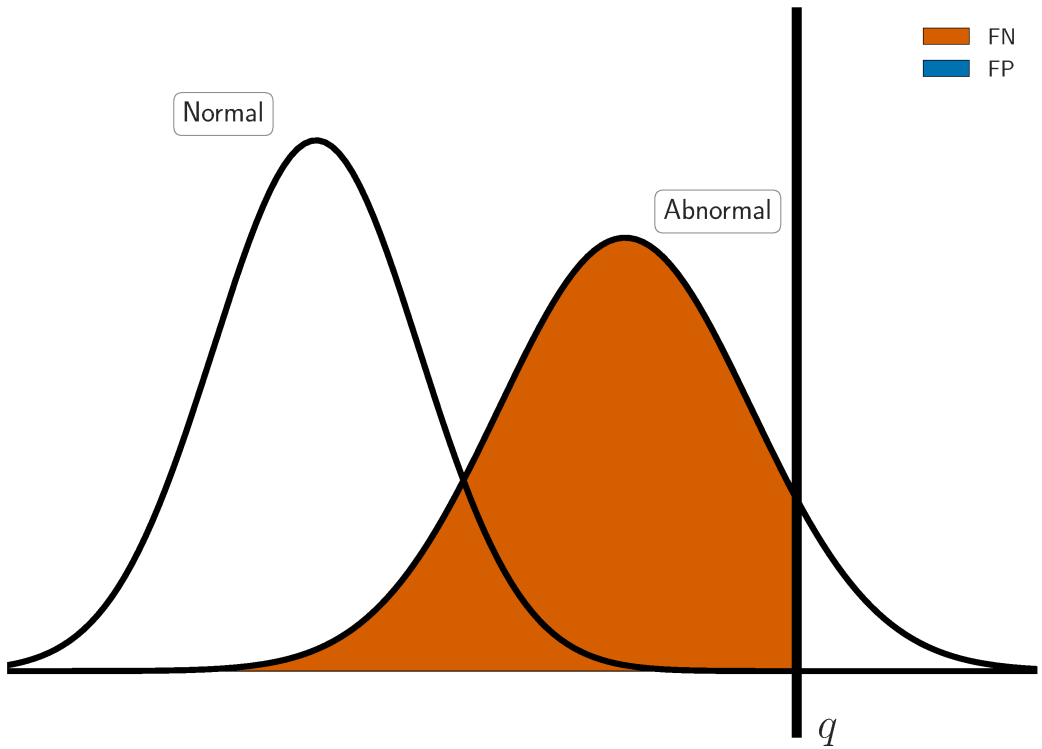
ROC curve



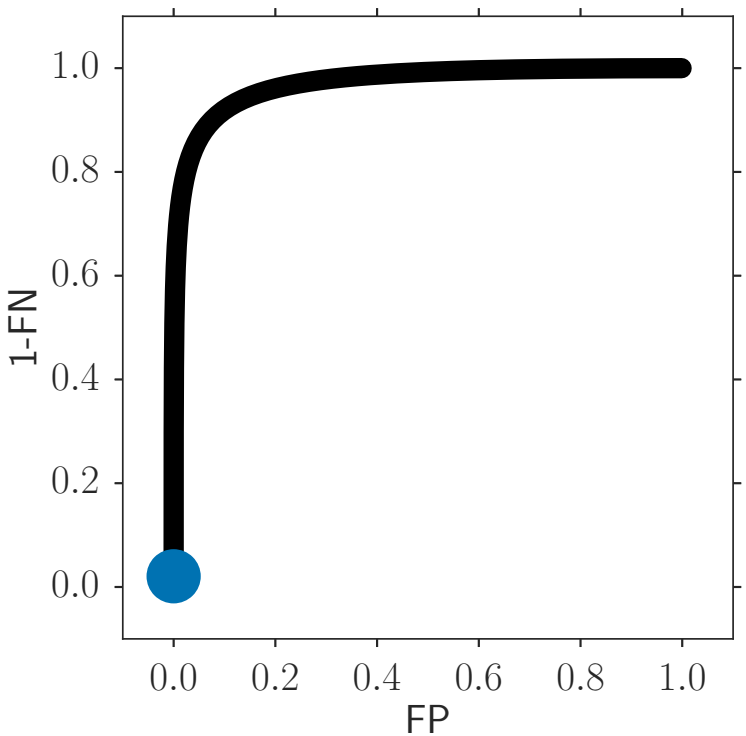


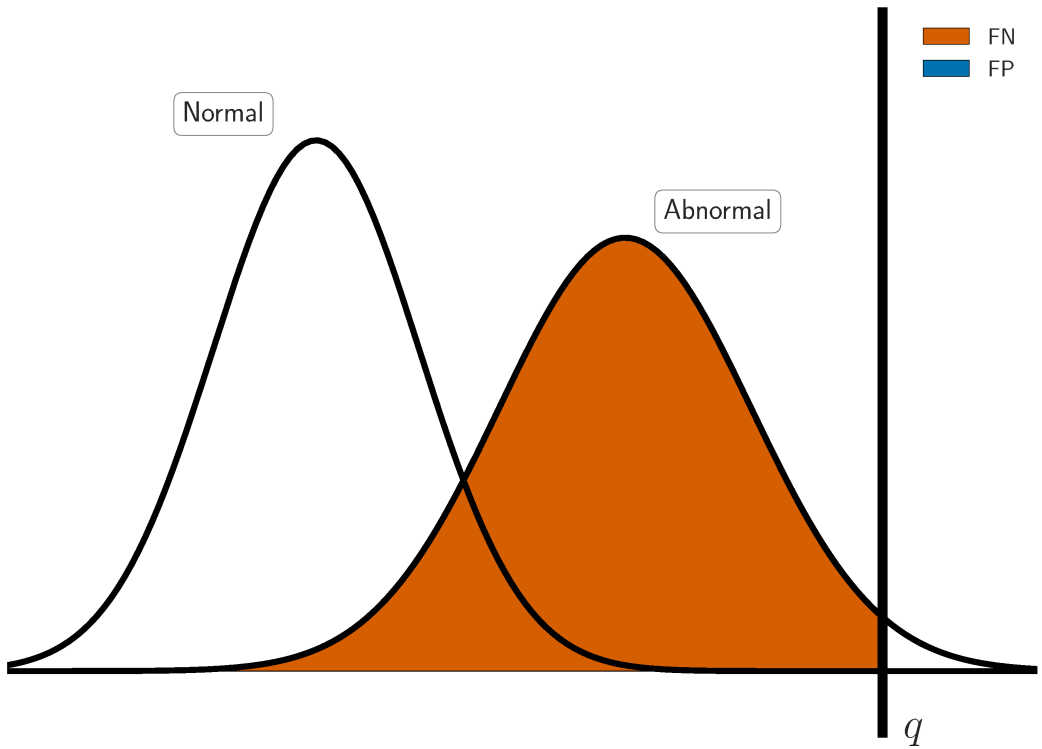
ROC curve





ROC curve





Références I