

SD-TSIA204  
Statistical hypothesis testing (for linear model)

Ekhine Irurozki  
Telecom Paris

December 2021

# Outline

## 1. Statistical hypothesis testing

- Definition

- The  $p$ -value

- Tests for linear regression

## 2. Illustration: forward variable selection

- Data set “diabetes”

## 3. ROC Curve

- Presentation

- Examples

## 1. Statistical hypothesis testing

### Definition

The  $p$ -value

Tests for linear regression

## 2. Illustration: forward variable selection

## 3. ROC Curve

# General principle

## Context

- We observe  $X_1, \dots, X_n$  from a common distribution  $\mathcal{P}$
- We are interested in  $\theta \in \Theta$ , a parameter of  $\mathcal{P}$

## Goal

To decide whether an assumption on  $\theta$  is likely (or not)

$$\mathcal{H}_0 = \{\theta \in \Theta_0\}$$

against some alternative

$$\mathcal{H}_1 = \{\theta \in \Theta_1\}$$

Call  $\mathcal{H}_0$  the null hypothesis,  $\mathcal{H}_1$  : the alternative

# General principle

## Means

Determine a **test statistic**  $T(X_1, \dots, X_n)$  and a region  $R$  such that if

$$T(X_1, \dots, X_n) \in R \Rightarrow \text{we reject } \mathcal{H}_0$$

In other words the observed data discriminates between  $H_0$  and  $H_1$

# Hypothesis testing for “heads or tails”

When flipping a coin the model is a Bernoulli distribution with parameter  $p$ ,  $\mathcal{B}(p)$ .

**Is the coin fair?**

$$\mathcal{H}_0 = \{p = 0.5\} \quad \text{against} \quad \mathcal{H}_1 = \{p \neq 0.5\}$$

**Is the coin possibly unfair?**

$$\mathcal{H}_0 = \{0.45 \leq p \leq 0.55\} \quad \text{against} \quad \mathcal{H}_1 = \{p \notin [0.45, 0.55]\}$$

What happens when  $n$  goes to  $\infty$ ?

## Do we reject or do we accept ?

In most practical situations,  $\mathcal{H}_0$  is simple, i.e.,

$$\Theta_0 = \{\theta_0\}$$

and  $\Theta_1 = \Theta \setminus \Theta_0$  is large

( $\mathcal{H}_0$  is often an hypothesis on which we care particularly, e.g., something acknowledged to be true, easy to formulate)

### We only reject $\mathcal{H}_0$

If  $\mathcal{H}_0$  is not rejected we cannot conclude  $\mathcal{H}_0$  is true because  $\mathcal{H}_1$  is too general

*e.g.*  $\{p \in [0, 0.5 \cup 0.5, 1]\}$  can not be rejected!

## 2 types of error

	$\mathcal{H}_0$	$\mathcal{H}_1$
$\mathcal{H}_0$ is not rejected	Correct (True positive)	Wrong (False negative)
$\mathcal{H}_0$ is rejected	Wrong (False positive)	Correct (True negative)

- **Type I:** probability of a wrong reject

$$\mathbb{P}(T(X_1, \dots, X_n) \in R \mid \mathcal{H}_0)$$

- **Type II:** probability of wrong non-reject

$$\mathbb{P}(T(X_1, \dots, X_n) \notin R \mid \mathcal{H}_1)$$



# Significance level and power

## Significance level $\alpha$ if

$$\limsup_{n \rightarrow +\infty} \mathbb{P}(T(X_1, \dots, X_n) \in R \mid \mathcal{H}_0) \leq \alpha$$

(We speak of 95%-test when  $\alpha$  is 0.05%)

## Consistency

A test statistics (given by  $T(X_1, \dots, X_n)$  and a region  $R$ ) is said to be  $\alpha$ -consistent if the **significant level** is  $\alpha$  and if the **power** goes to one, i.e.,

$$\limsup_{n \rightarrow +\infty} \mathbb{P}(T(X_1, \dots, X_n) \in R \mid \mathcal{H}_0) \leq \alpha$$

$$\lim_{n \rightarrow \infty} \mathbb{P}(T(X_1, \dots, X_n) \in R \mid \mathcal{H}_1) = 1$$

# Test statistic and reject region

Goal: to build a  $\alpha$ -consistent test

- (1) Define the test statistic  $T(X_1, \dots, X_n)$  and the level  $\alpha$  you wish
- (2) Do some maths to determine a reject region  $R$  that achieves a significance level  $\alpha$
- (3) Prove the consistency
- (4) Rule decision: reject whenever  $T_n(X_1, \dots, X_n) \in R$

# Famous tests

- Test of the equality of the mean for 1 sample
- Test of the equality of the means between 2 samples
- Chi-square test for the variance
- Chi-square test of independence
- Regression coefficient non-effects test

## Example: Gaussian mean

- Model:  $\Theta = \mathbb{R}$ ,  $\mathbb{P}_\theta = \mathcal{N}(\theta, 1)$
- Observe  $(X_1, \dots, X_n)$  i.i.d. from this model
- Null hypothesis:  $\mathcal{H}_0 : \{\theta = 0\}$
- Under  $\mathcal{H}_0$ ,  $T_n(X_1, \dots, X_n) = \frac{1}{\sqrt{n}} \sum_i X_i \sim \mathcal{N}(0, 1)$
- Critical region for  $T_n$  ? Gaussian quantile:

$$\mathbb{P}(T_n \in [-1.96, 1.96] \mid \mathcal{H}_0) = 0.95$$

- Take  $R = ]-\infty, -1.96[ \cup ]1.96, +\infty[$ .
- Numerical example: If  $T_n = 1.5$ , we do **not** reject  $\mathcal{H}_0$  at level 95%

## 1. Statistical hypothesis testing

Definition

The  $p$ -value

Tests for linear regression

## 2. Illustration: forward variable selection

## 3. ROC Curve

## Usage of the $p$ -value

- The decision to accept or reject  $\mathcal{H}_0$  is subject to the chosen significance level  $\alpha$ .
- To avoid making this choice in advance, in particular in software, the notion of the  $p$ -value is used to represent the result of a test.
- **The  $p$ -value is the probability that, under  $\mathcal{H}_0$ , the test statistic  $T_n$  takes a value at least as extreme as its observed value.**
- Relation to the critical region:
  - If the test is one-sided with  $R = \{t \mid t > c\}$  then for the observed  $T_n$  the  $p$ -value is  $\mathbb{P}(T > t_0 \mid \mathcal{H}_0)$ .
  - If the test is one-sided with  $R = \{t \mid t < c\}$  then for the observed  $T_n$  the  $p$ -value is  $\mathbb{P}(T < T_n \mid \mathcal{H}_0)$ .
  - If the test is two-sided with  $R = \{t \mid t \in ]-\infty; c_1) \cup (c_2; +\infty[ \}$  then for the observed  $T_n$  the  $p$ -value is  $2\mathbb{P}(T < T_n \mid H_0)$  if  $T_n$  is smaller than the median, and  $2\mathbb{P}(T > T_n \mid H_0)$  if  $T_n$  is larger than the median.

## Usage of the $p$ -value: example

- Model:  $\Theta = \mathbb{R}$ ,  $\mathbb{P}_\theta = \mathcal{N}(\theta, 1)$
- Observe  $(X_1, \dots, X_n)$  i.i.d. from this model
- Null hypothesis:  $\mathcal{H}_0 : \{\theta \leq 5\}$
- Under  $\mathcal{H}_0$ ,  $T_n(X_1, \dots, X_n) = \frac{\bar{X}_n - 5}{\frac{1}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$

The test decision:

- Reject  $\mathcal{H}_0$  if  $\bar{X}_n > 5 + z_{1-\alpha} \frac{1}{\sqrt{n}}$ .

Using the  $p$ -value:

- Assume  $n = 10$  and  $\bar{X}_n = 5.75$ .
- The  $p$ -value equals  $\mathbb{P}(\bar{X} > 5.75)$  with  $\bar{X} \sim \mathcal{N}(5, \frac{1}{10})$ ,  
*i.e.*  $\mathbb{P}(Z > 2.3717)$  with  $Z \sim \mathcal{N}(0, 1)$ , which equals 0.0089.
- This indicates directly that one should reject at a level 0.05 and even 0.01.
- If the test would be two sided, *i.e.* with  $\mathcal{H}_0 : \{\theta = 5\}$ , the  $p$ -value for  $\bar{X}_n = 5.75$  would be  $0.0089 \times 2 = 0.0178$  implying **reject** at a level 0.05 but **not** 0.01.

# Cons of hypothesis testing

- No limits on the sample size
- What if we don't reject?



## 1. Statistical hypothesis testing

Definition

The  $p$ -value

Tests for linear regression

## 2. Illustration: forward variable selection

## 3. ROC Curve

## Test of no-effect : Gaussian case

### Gaussian Model

$$y_i = \theta_0^* + \sum_{k=1}^p \theta_k^* x_{i,k} + \varepsilon_i$$

$$\mathbf{x}_i^\top = (1, x_{i,1}, \dots, x_{i,p}) \in \mathbb{R}^{p+1} \text{ (deterministic)}$$

$$\varepsilon_i \stackrel{i.i.d}{\sim} \mathcal{N}(0, \sigma^2), \text{ for } i = 1, \dots, n$$

### Theorem

Let  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top \in \mathbb{R}^{n \times (p+1)}$  of full rank, and  $\hat{\sigma}^2 = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}}\|_2^2 / (n - (p+1))$ , then

$$\hat{T}_j = \frac{\hat{\theta}_j - \theta_j^*}{\hat{\sigma} \sqrt{(\mathbf{X}^\top \mathbf{X})_{jj}^{-1}}} \sim \mathcal{T}_{n-(p+1)}$$

where  $\mathcal{T}_{n-p}$  is a Student law (with  $n - (p+1)$  degrees of freedom)

## Test of no-effect : Gaussian case

### Null hypothesis

Aim is to test

$$\mathcal{H}_0 : \theta_j^* = 0$$

equivalently,  $\Theta_0 = \{\theta \in \mathbb{R}^p : \theta_j = 0\}$

Under  $\mathcal{H}_0$ , we know the value of  $\hat{T}_j$  :

$$T_j := \frac{\hat{\theta}_j}{\hat{\sigma} \sqrt{(X^\top X)^{-1}_{jj}}} \sim \mathcal{T}_{n-(p+1)}$$

Choosing  $R = [-t_{1-\alpha/2}, t_{1-\alpha/2}]^c$  with  $t_{1-\alpha/2}$  the  $1 - \alpha/2$ -quantile of  $\mathcal{T}_{n-(p+1)}$ , we decide to reject  $\mathcal{H}_0$  whenever

$$|\hat{T}_j| > t_{1-\alpha/2}$$

## Test of no-effect : Random-design case

### Random design Model

$$y_i = \theta_0^* + \sum_{k=1}^p \theta_k^* x_{i,k} + \varepsilon_i$$

$$\mathbf{x}_i^\top = (1, x_{i,1}, \dots, x_{i,p}) \in \mathbb{R}^{p+1}$$

$$(\varepsilon_i, \mathbf{x}_i) \stackrel{i.i.d}{\sim} (\varepsilon, \mathbf{x}), \text{ for } i = 1, \dots, n$$

$$\mathbb{E}(\varepsilon|\mathbf{x}) = 0, \text{ Var}(\varepsilon|\mathbf{x}) = \sigma^2$$

### Theorem

If  $\text{var}(\mathbf{x})$  has full rank, then

$$\hat{T}_j = \frac{\hat{\theta}_j - \theta_j^*}{\hat{\sigma} \sqrt{(\mathbf{X}^\top \mathbf{X})_{jj}^{-1}}} \xrightarrow{d} \mathcal{N}(0, 1)$$

## Test of no-effect : Random-design case

### Null hypothesis

Aim is to test

$$\mathcal{H}_0 : \theta_j^* = 0$$

equivalently,  $\Theta_0 = \{\theta \in \mathbb{R}^p : \theta_j = 0\}$

Under  $\mathcal{H}_0$ , we know the value of  $\hat{T}_j$  :

$$T_j := \frac{\hat{\theta}_j}{\hat{\sigma} \sqrt{(X^\top X)^{-1}_{j,j}}} \xrightarrow{d} \mathcal{N}(0, 1)$$

Choosing  $R = [-z_{1-\alpha/2}, z_{1-\alpha/2}]^c$  with  $z_{1-\alpha/2}$  the  $1 - \alpha/2$ -quantile of  $\mathcal{N}(0, 1)$ , we decide to reject  $\mathcal{H}_0$  whenever

$$|\hat{T}_j| > z_{1-\alpha/2}$$

## Link between IC and test

Reminder (Gaussian model):

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

is a CI at level  $\alpha$  for  $\theta_j^*$ . Stating “ $0 \in IC_\alpha$ ” means

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

It is equivalent to accepting the hypothesis  $\theta_j^* = 0$  at level  $\alpha$ . The smallest  $\alpha$  such that  $0 \in IC_\alpha$  is called the **p-value**.

Rem: Taking  $\alpha$  close to zero  $IC_\alpha$  covers the full space, hence one can find (by continuity) an  $\alpha$  achieving equality in the aforementioned equations.

# Outline

## 1. Statistical hypothesis testing

- Definition

- The  $p$ -value

- Tests for linear regression

## 2. Illustration: forward variable selection

- Data set “diabetes”

## 3. ROC Curve

- Presentation

- Examples

1. Statistical hypothesis testing
2. Illustration: forward variable selection  
Data set “diabetes”
3. ROC Curve



## “Diabetes” data set

patient	age	sex	bmi	bp	Serum measurements						Resp
	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 having diabetes,  $p = 10$  variables “baseline” body mass index (bmi), average blood pressure (bp), *etc.*... have been measured.

**Goal:** predict disease progression one year in advance after the “baseline” measurement [EHJT04].

- Each variable of the data set from *sklearn* has been previously standardized.
- We apply an “expensive” version of the **forward variable selection** method (see, *e.g.*, [Zha09])

## “Diabetes” data set

- We define a vector of covariates with intercept  $\tilde{X} = (\mathbb{1}, x_1, \dots, x_{10})$ .

### Step 0

- for each variable  $\tilde{X}_k$ ,  $k = 1, \dots, 11$ , we consider the model

$$y \simeq \beta_k x_k$$

- we test whether its regression coefficient equals zero, *i.e.*

$$H_0 : \beta_k = 0$$

using the statistic  $\frac{\hat{\beta}_k}{\hat{s}_k}$  with  $\hat{s}_k$  being the estimated standard deviation.

- we compare all of the  $p$ -values, and keep the one possessing the smallest  $p$ -value. We save the residuals in the vector  $r_0$ .

## “Diabetes” data set

### Step $\ell$

We have selected  $\ell$  variable(s) :  $\tilde{X}^{(\ell)} \in \mathbb{R}^\ell$ . Those not selected are noted  $\tilde{X}^{(-\ell)} \in \mathbb{R}^{p-\ell}$ . We possess the vector of residuals  $\mathbf{r}_{\ell-1}$  calculated on the previous step.

- for each variable  $\mathbf{x}_k$  in  $\tilde{X}^{(-\ell)}$ , we consider the model

$$\mathbf{r}_{\ell-1} \simeq \beta_k \mathbf{x}_k$$

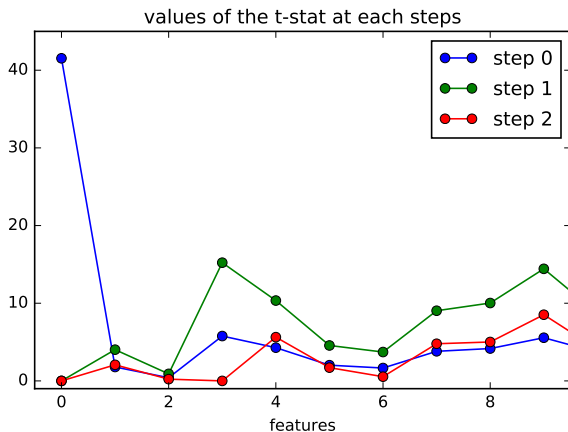
- we test if its regression coefficient equal zero, *i.e.*

$$H_0 : \beta_k = 0$$

using the test statistic  $\frac{\hat{\beta}_k}{\hat{s}_k}$  with  $\hat{s}_k$  being the estimated standard deviation.

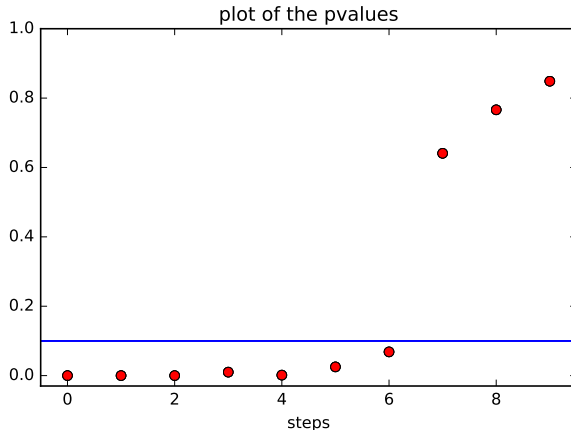
- we compare all of the  $p$ -values, and keep the one possessing the smallest  $p$ -value. We save the residuals in the vector  $\mathbf{r}_\ell$ .

## Values of the test statistics at each step



- The test statistic of the selected variable is 0 on the following steps.
- The intercept is the first selected variable, then  $x_3$ , *etc.*...

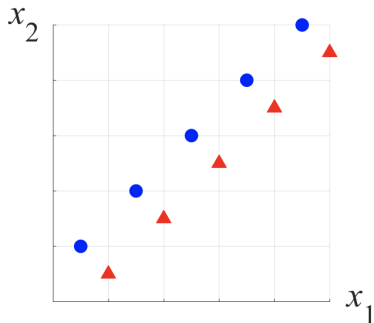
## Values of the test statistics at each step



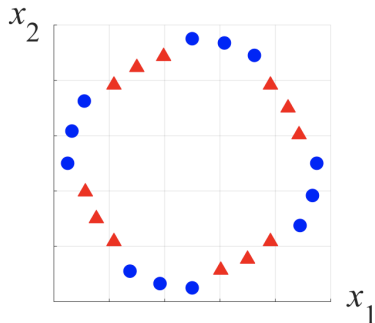
- Sequence of the selected variables with the test size 0.1 :

[ 0, 3, ,9 ,5 ,4 ,2 ,7 ]

Other heuristics: rank the coefficients by coefficient, p-values, backwards selection, ...



$$\begin{aligned} \text{(a)} \quad \hat{e}^{LOO}(\text{LDC}) &= 0 \\ \hat{e}^{LOO}(\text{1NN}) &= 1 \end{aligned}$$



$$\begin{aligned} \text{(b)} \quad \hat{e}^{LOO}(\text{LDC}) &= 1 \\ \hat{e}^{LOO}(\text{1NN}) &= 0 \end{aligned}$$

[KMAGR20]

# Outline

## 1. Statistical hypothesis testing

- Definition

- The  $p$ -value

- Tests for linear regression

## 2. Illustration: forward variable selection

- Data set “diabetes”

## 3. ROC Curve

- Presentation

- Examples

1. Statistical hypothesis testing
2. Illustration: forward variable selection
3. ROC Curve
  - Presentation
  - Examples



## Medical context

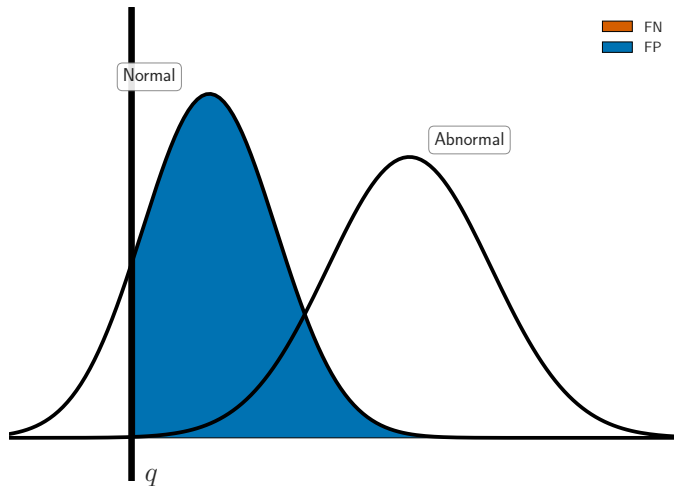
- A group of patients  $i = 1, \dots, n$  is followed for disease screening.
- For each individual, the test relies on a random variable  $X_i \in \mathbb{R}$  and a threshold  $q \in \mathbb{R}$

as soon as  $X_i > q$  the test is **positive**  
o.w. the test is **negative**

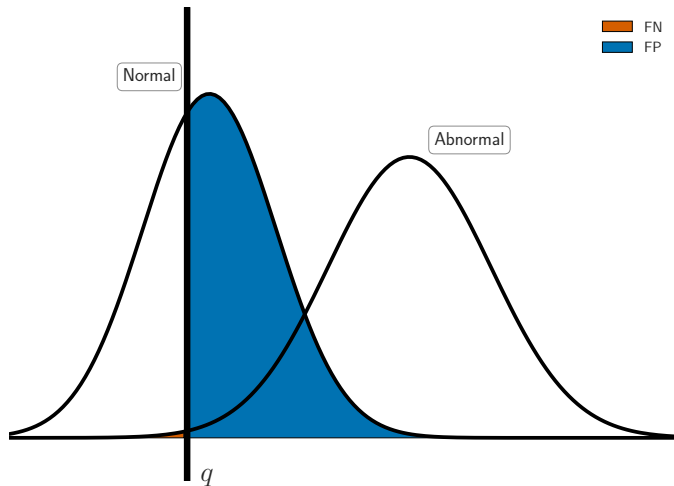
## Set of possible configurations

	Normal $H_0$	Sick $H_1$
negative	true negative	false negative
positive	false positive	true positive

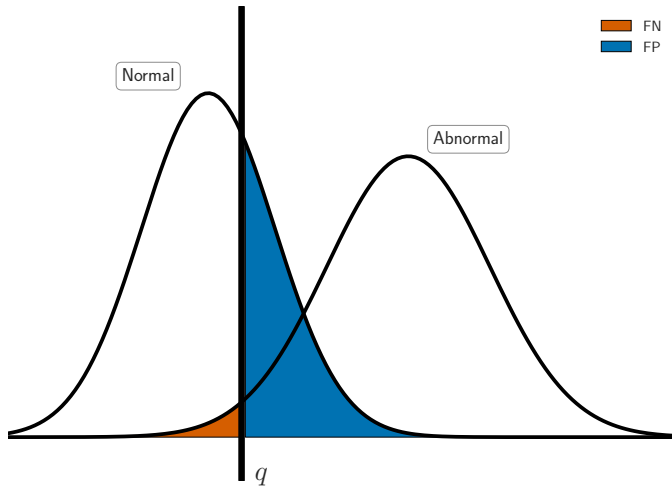
## False positive vs. false negative



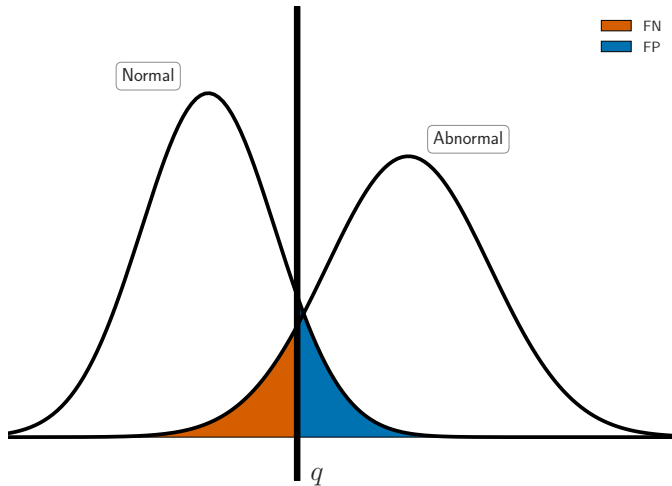
## False positive vs. false negative



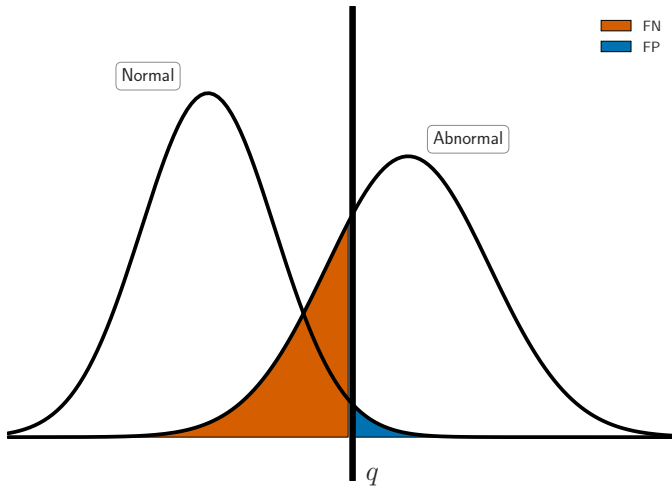
## False positive vs. false negative



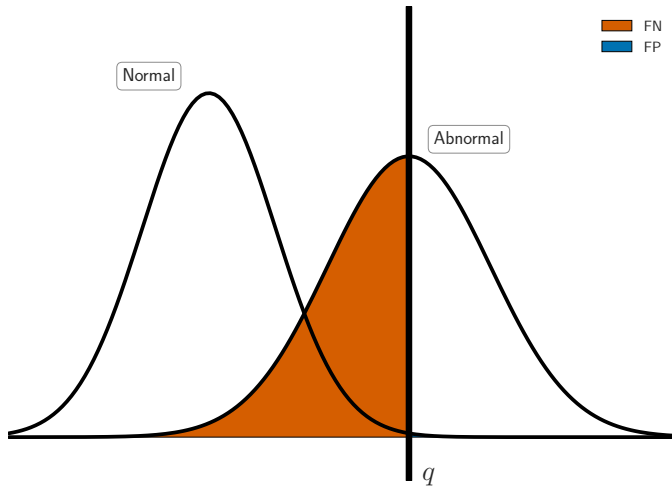
## False positive vs. false negative



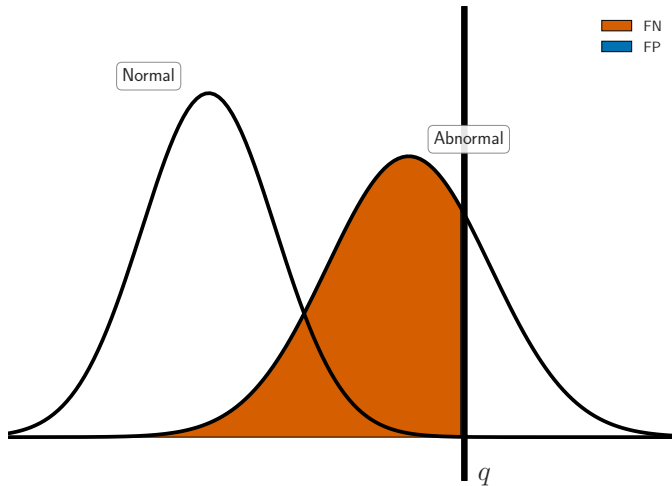
## False positive vs. false negative



## False positive vs. false negative

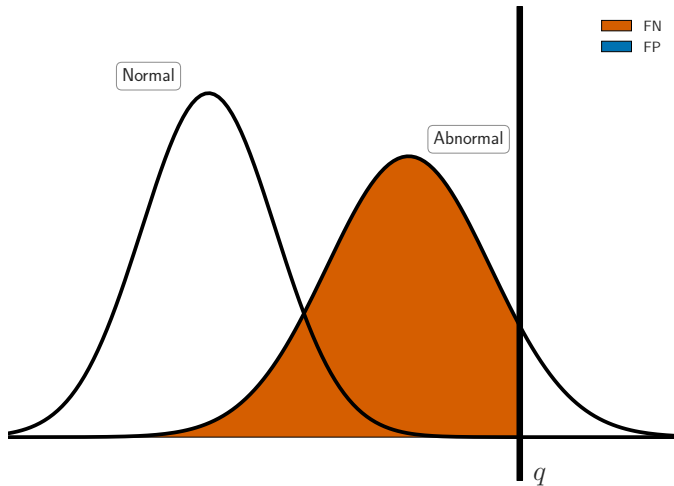


## False positive vs. false negative

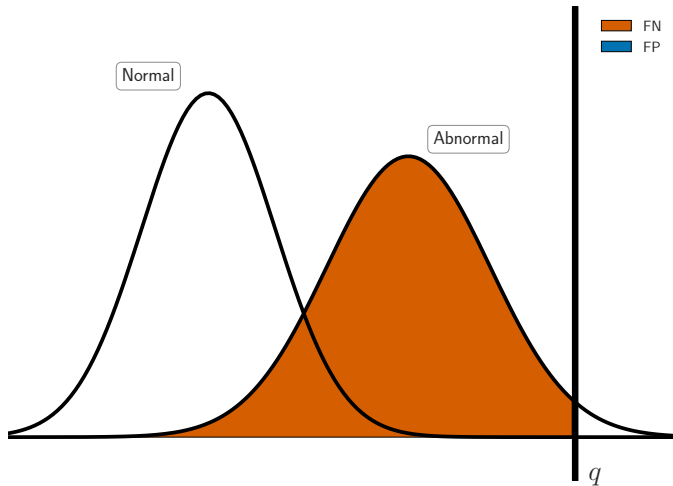




## False positive vs. false negative



## False positive vs. false negative



# Sensitivity - Specificity

- Assumption: Normal individuals have the same c.d.f.  $F$
- Assumption: Sick individual have the same c.d.f  $G$

## Definition

- Sensitivity :  $Se(q) = 1 - G(q)$  (1- type 2nd error)
- Specificity :  $Sp(q) = F(q)$  (1- type 1st error)

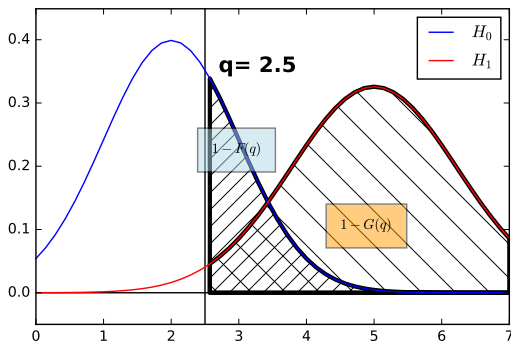
# ROC curve

## Definition

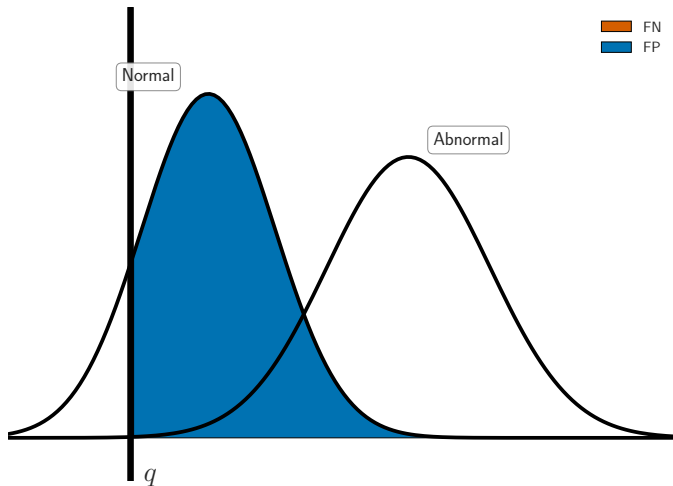
The ROC curve is the curve described by  $(1 - \text{Sp}(q), \text{Se}(q))$ , when  $q \in \mathbb{R}$ . Hence, it is the function  $[0, 1] \rightarrow [0, 1]$

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

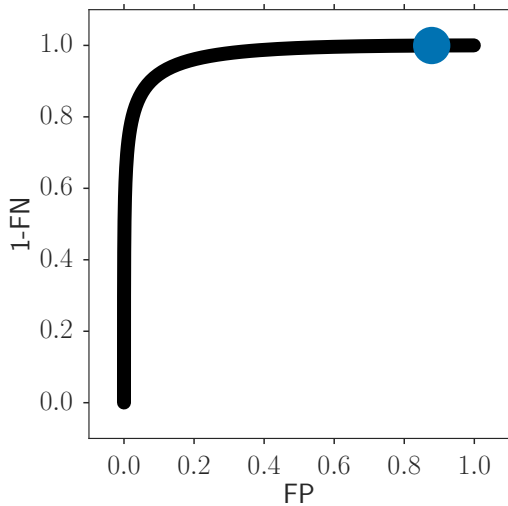
where  $F^{-1}(1 - t) = \inf\{x \in \mathbb{R} : F(x) \geq 1 - t\}$ .



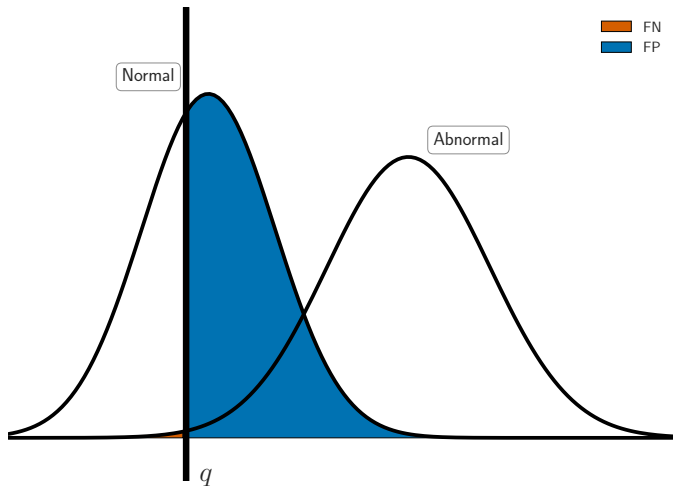
# ROC Curve



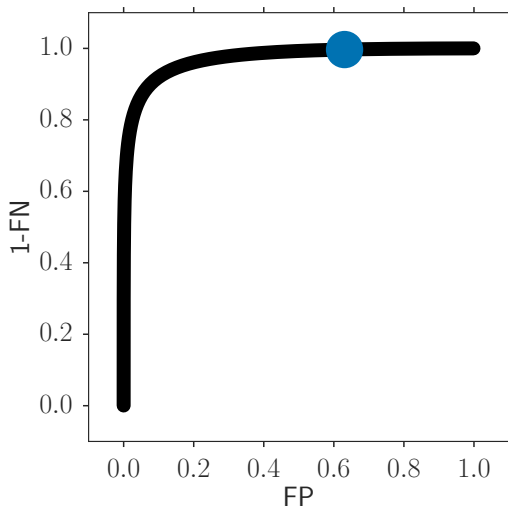
## ROC Curve



# ROC Curve

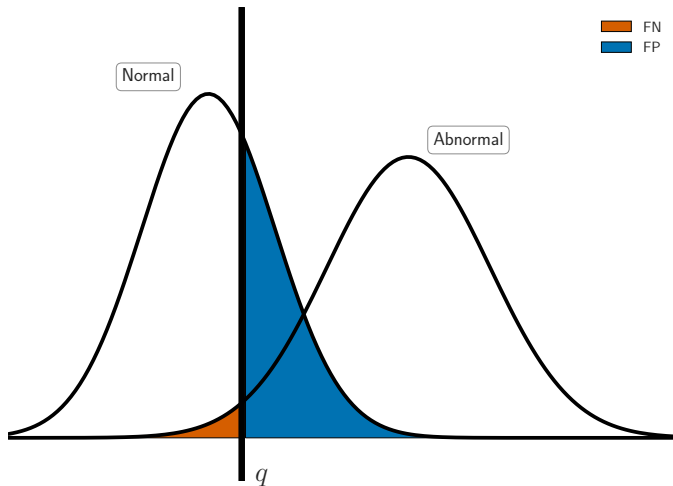


## ROC Curve

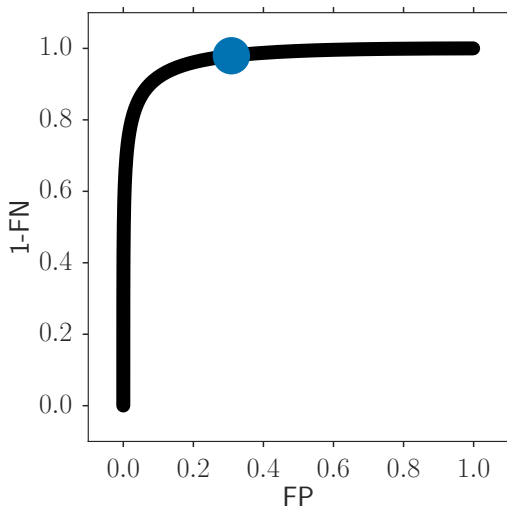




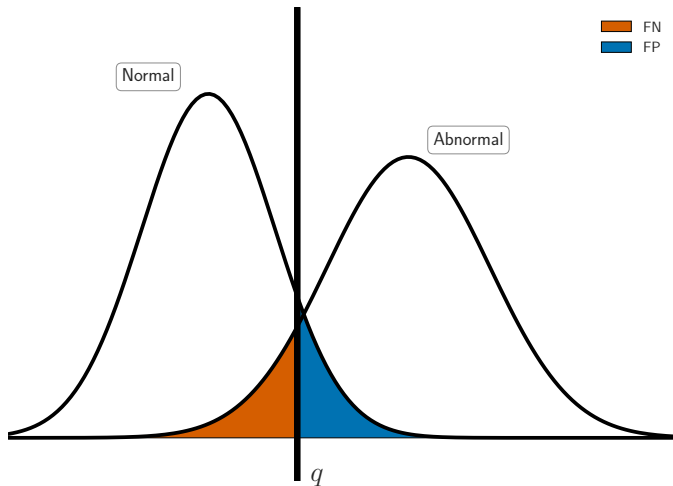
# ROC Curve



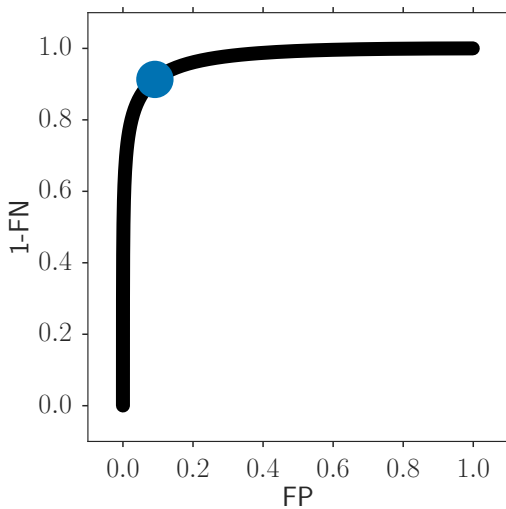
## ROC Curve



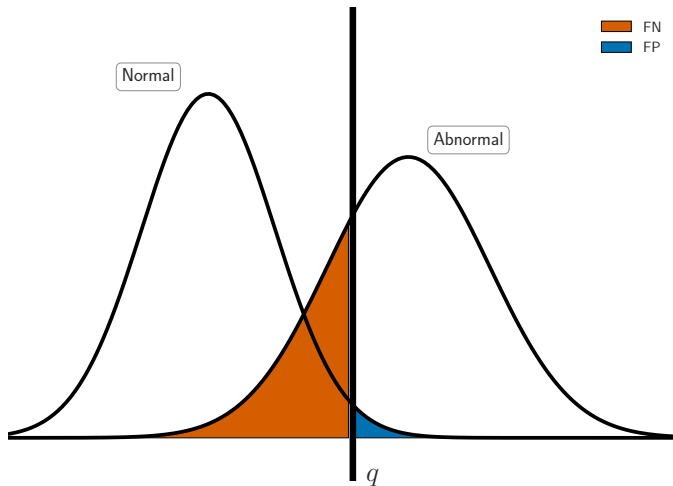
# ROC Curve



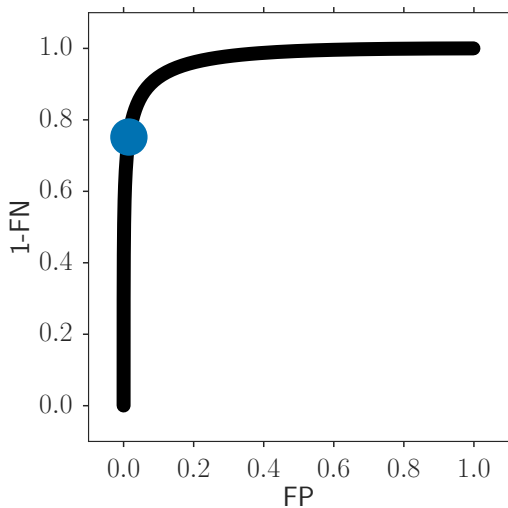
## ROC Curve



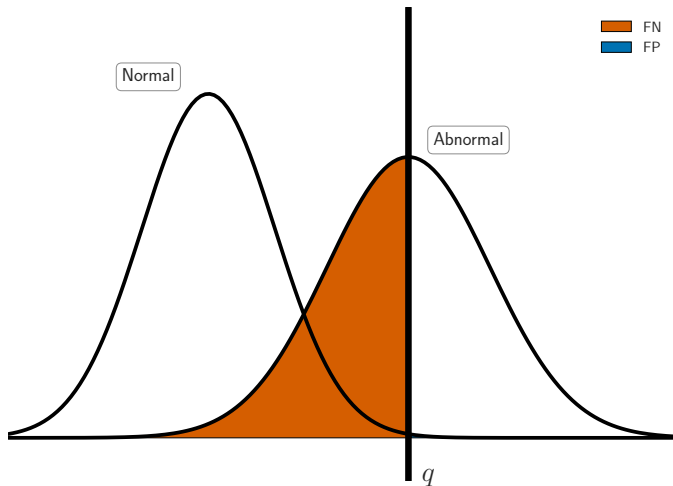
# ROC Curve



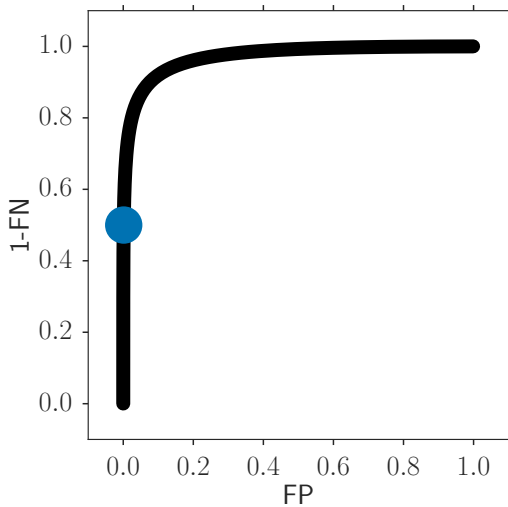
## ROC Curve



# ROC Curve

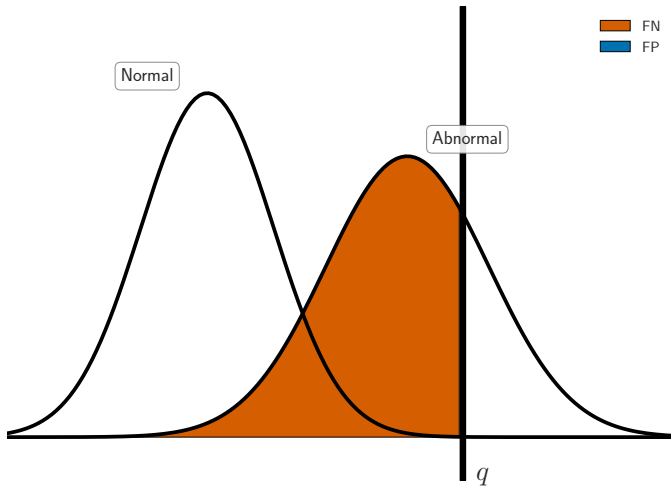


## ROC Curve

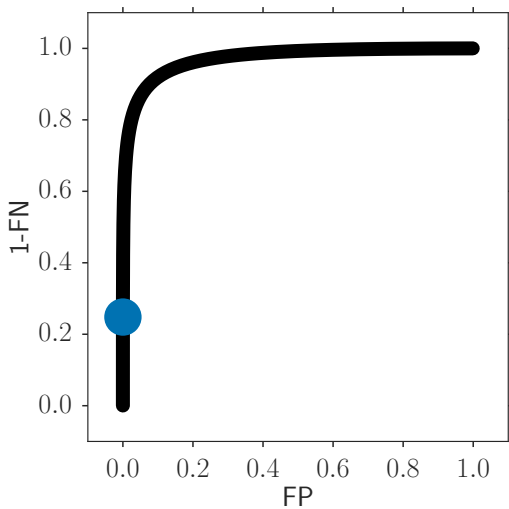




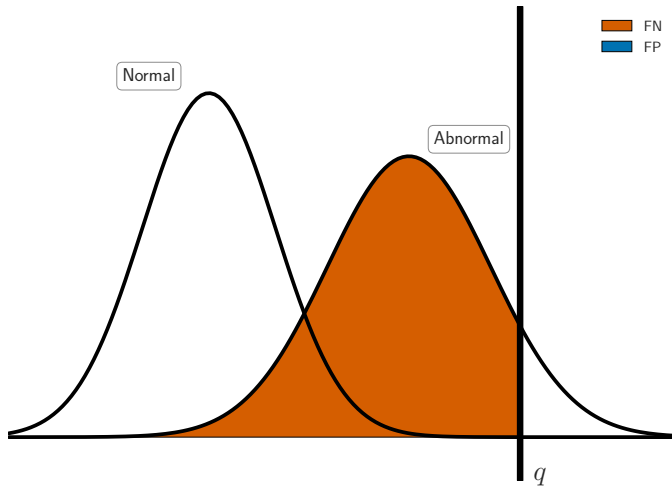
# ROC Curve



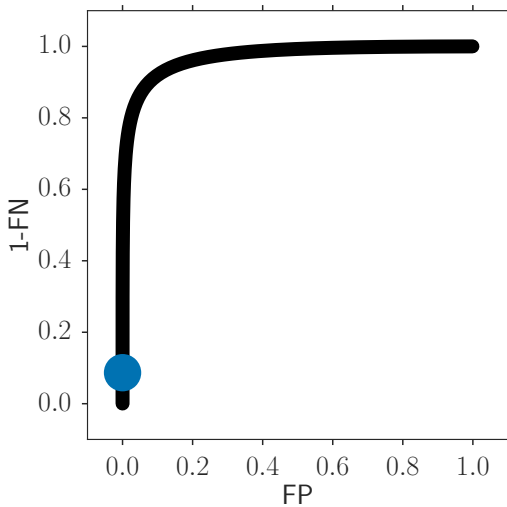
## ROC Curve



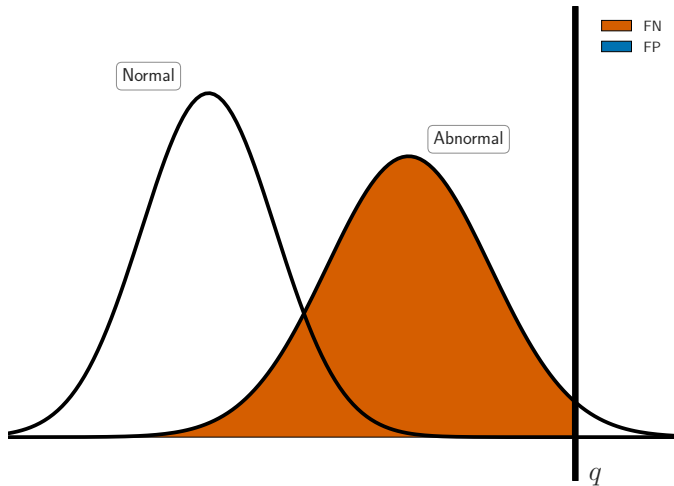
# ROC Curve



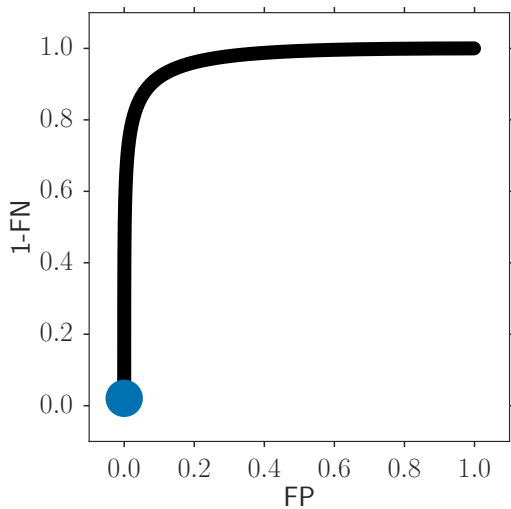
## ROC Curve



# ROC Curve



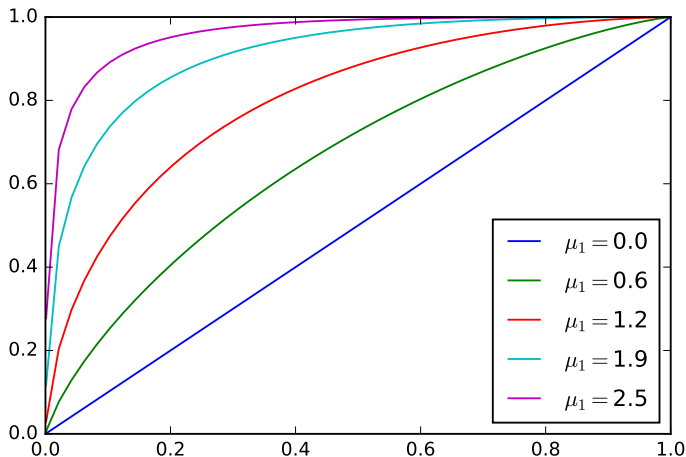
## ROC Curve



1. Statistical hypothesis testing
2. Illustration: forward variable selection
3. ROC Curve
  - Presentation
  - Examples

## ROC curves for bi-normal case

- $F$  and  $G$  are Gaussian with parameter  $\mu_0, \sigma_0$  and  $\mu_1, \sigma_1$ , respectively.
- Here  $\mu_0 = 0$ ,  $\sigma_0 = \sigma_1 = 1$ , and  $\mu_1$  varies





# Estimation–application

## ROC curve estimation

- Maximum likelihood
- Non-parametric
- Bayesian with latent variables
- Estimation of the area under the ROC curve (AUC)

## Application

- To compare different statistic tests
- To compare different (supervised) learning algorithm
- To compare variable selection methods (*e.g.* Lasso, OMP, etc.)

nb: ROC = Receiver Operating Characteristic

# References I

- [EHJT04] 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.
- [KMAGR20] Ludmila I Kuncheva, Clare E Matthews, Álgvar Arnaiz-González, and Juan J Rodríguez. Feature selection from high-dimensional data with very low sample size: A cautionary tale. *arXiv preprint arXiv:2008.12025*, 2020.
- [Zha09] Tong Zhang. Adaptive forward-backward greedy algorithm for sparse learning with linear models. In *Advances in Neural Information Processing Systems*, pages 1921–1928, 2009.
- Some of these slides have been prepared by Anne Sabourin and Josef Salmon, the authors express their gratitude for this.