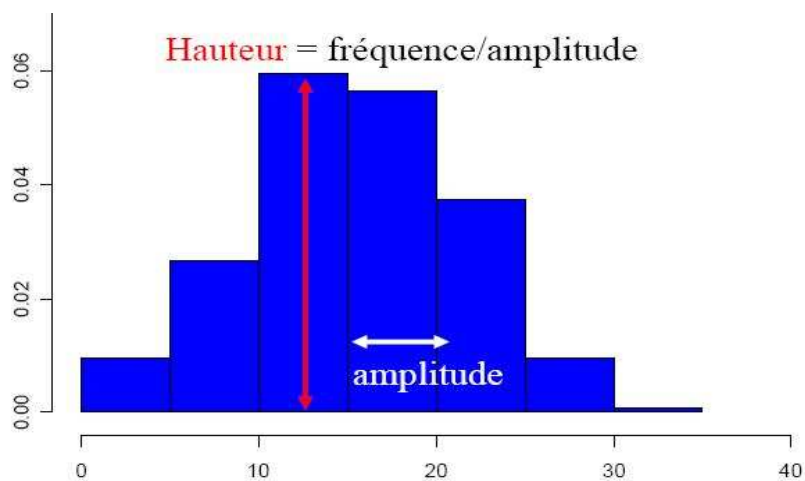


Cours N°3 : Variables aléatoires continues et tests diagnostiques.

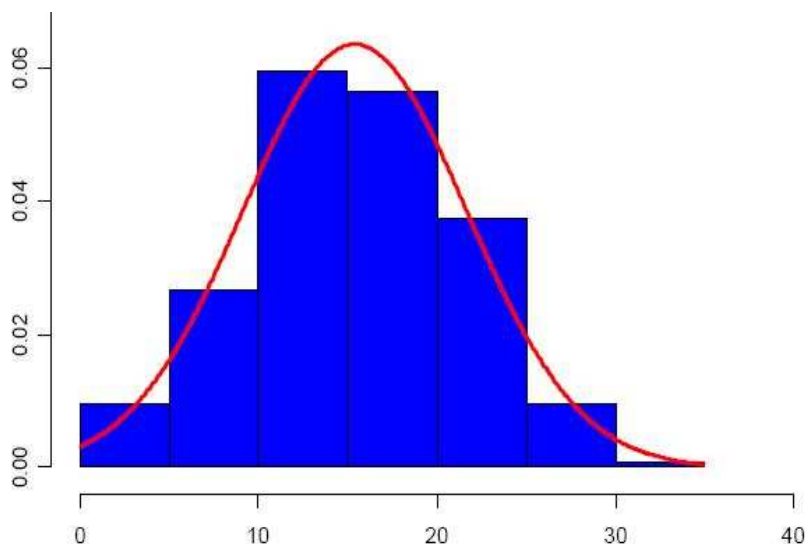
I. Variables aléatoires continues

On a vu que pour une **variable aléatoire discrète**, l'univers est **dénombrable** et on peut utiliser un **diagramme en bâtons**, en déterminant $P(X = x)$.

Pour une **variable aléatoire continue**, l'univers n'est **pas dénombrable** puisqu'il est dans l'ensemble des réels \mathbb{R} . On a **TOUJOURS** $P(X = x) = 0$. On ne sait définir qu'une **probabilité d'un intervalle** $P(X \in [a; b])$. On utilise alors un **histogramme**, où la **superficie** d'un rectangle (**hauteur x amplitude**) donne la **fréquence**, et où la superficie totale, donc la somme des fréquences, fait 1.



On « lisse » ensuite l'histogramme par une fonction appelée **densité de probabilité**.



Attention : l'amplitude des classes n'est pas nécessairement la même pour toutes les classes !

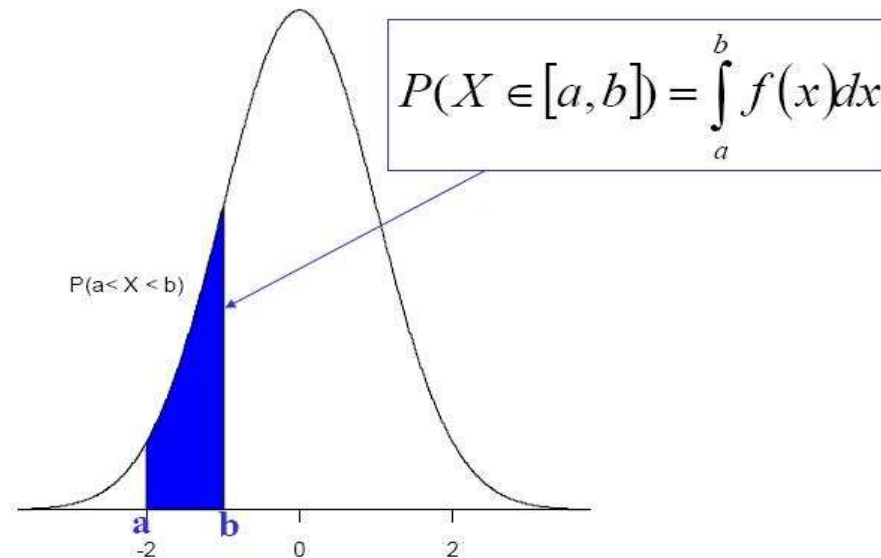
A. Densité de probabilité

Une densité de probabilité est définie par **2 conditions** :

- $f(x) \geq 0$ [$f(x)$ n'est jamais négative !]
- $\int_{x \in E} f(x) dx = 1$

On peut écrire l'approximation $P(X \in [x; x + dx]) \approx f(x) dx$

On peut également **définir la probabilité d'un intervalle [a ; b]** [en fait, c'est la seule chose qu'on peut faire concrètement)].



Prenons par exemple une loi uniforme sur [0 ; 1]. Il est logique que, pour que l'aire sous la courbe soit de 1, il faut que l'ordonnée soit de 1. On peut retrouver ce résultat par la densité de probabilité.

$$\int_0^1 f(x) dx = 1$$

On pose k l'ordonnée que l'on cherche pour que l'aire sous la courbe soit de 1, et on a $f(x) = k$ puisque la loi est uniforme. On a donc :

$$\int_0^1 f(x) dx = \int_0^1 k dx = [kx]_0^1 = k = 1$$

Remarque : le passage du discret au continu transforme les sommes Σ en intégrales \int , et les p_i en $f(x) dx$.

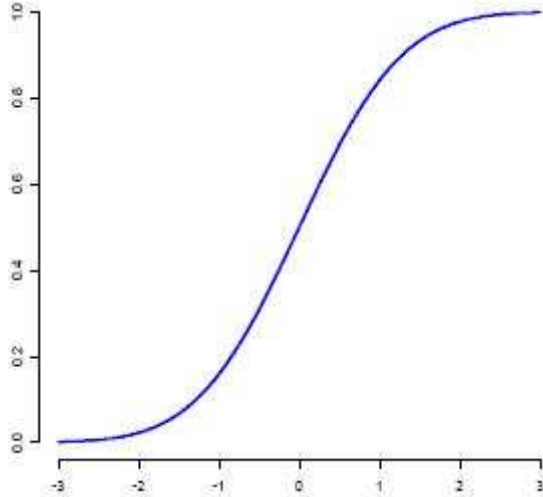
B. Fonction de répartition F(x)

1. Définition

La **fonction de répartition F(x)**, pour les **variables aléatoires continues**, est donnée par la relation :

$$P(X) \in [-\infty ; t] = \int_{-\infty}^t f(x) dx = P(X \leq t) = F(t)$$

Cette fonction de répartition est **monotone** et **croissante de 0 à 1**, et **continue**.



Il faut savoir également **la fonction de répartition F(X) est la primitive de la densité de probabilité f(x)**. Ainsi, on peut écrire :

$$F'(x) = f(x)$$

Si on considère par exemple la loi uniforme sur [0 ; 1].

L'univers E est compris entre 0 et 1 (E = [0 ; 1]), et f(x) = 1.

On peut donc retrouver la fonction de répartition F(x) en faisant une **intégrale** :

$$F(x) = \int_0^x f(t) dt = \int_0^x 1 dt = [t]_0^x = x$$

Ainsi, la fonction de répartition F(x) est une fonction linéaire de type y = x. Elle est bien **monotone** et **croissante de 0 à 1**.

Plus généralement, on peut définir directement une **densité de probabilité f(x)** et une **fonction de répartition F(x)** d'une **loi uniforme sur [a ; b]** :

$$f(x) = k = \frac{1}{b-a}$$

$$F(x) = \frac{x-a}{b-a}$$

2. Intérêts de la fonction de répartition

La fonction de répartition est valable aussi bien pour les variables continues que pour les variables discrètes. Elle permet de **calculer des probabilités d'intervalle** (et plus particulièrement l'intervalle $]-\infty ; x]$), ce qui est surtout utile pour les variables aléatoires continues.

C. Probabilité d'un intervalle [a ; b]

Le calcul d'une probabilité d'un intervalle $[a ; b]$ se calcule de la façon suivante :

$$P(X \in [a ; b]) = \int_a^b f(x) dx$$

$$P(X \in [-\infty ; t]) = \int_{-\infty}^t f(x) dx = P(X \leq t) = F(t)$$

$$\boxed{P(X \in [a ; b]) = F(b) - F(a)}$$

Percentiles de X

On appelle $p^{\text{ième}}$ percentile de X, le réel L_p tel que :

$$\boxed{F(L_p) = P(X \leq L_p) = p}$$

Exemples : la médiane est le $50^{\text{ème}}$ percentile, le premier quartile Q1 est le $25^{\text{ème}}$ percentile, et le troisième quartile Q3 est le $75^{\text{ème}}$ percentile.

D. Espérance de X

Pour une variable aléatoire continue, l'**espérance de X** est donnée par la relation :

$$\boxed{E(X) = \int_{-\infty}^{+\infty} x f(x) dx}$$

Il existe également quelques **propriétés** se rapportant à l'espérance d'une variable aléatoire continue, on a :

- $E(aX + b) = aE(X) + b$
- $E(\sum X) = \sum E(X)$

Par exemple, pour la loi uniforme sur $[0,1]$, où $f(x) = 1$ et où $F(x) = x$, on a :

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx = \int_0^1 x dx = \left[\frac{x^2}{2} \right]_0^1 = \frac{1}{2}$$

[Il ne faut pas la confondre avec l'espérance de la variable aléatoire discrète donnée par la relation $E(X) = \sum_{x \in E} x \times P(X = x)$]

E. Variance de X

La **variance** d'une variable aléatoire continue est donnée par les mêmes relations que pour les variables aléatoires discrètes. On a :

$$\boxed{Var(X) = E[(X - E[X])^2]}$$

$$\boxed{Var(X) = E(X^2) - [E(X)]^2}$$

Mais avec
$$E(X^2) = \int_{-\infty}^{+\infty} x^2 f(x) dx$$

Il existe aussi des **propriétés** concernant la variance d'une variable aléatoire continue :

- $Var(aX) = a^2 Var(X)$
- $Var(\sum X) = \sum Var(X)$ si les variables aléatoires **X** sont **indépendantes** entre elles.

Par exemple, pour une loi uniforme sur $[0 ; 1]$, où $f(x) = 1$ et $F(x) = x$, on a :

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx = \int_0^1 x dx = \left[\frac{x^2}{2} \right]_0^1 = \frac{1}{2}$$

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 f(x) dx = \int_0^1 x^2 dx = \left[\frac{x^3}{3} \right]_0^1 = \frac{1}{3}$$

Ainsi, $Var(X) = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12}$

II. « Catalogue » de lois continues

A. Variable aléatoire continue uniforme sur [a ; b]

La **variable aléatoire continue uniforme sur [a ; b]**, notée aussi $X \sim U[a ; b]$, est définie par $f(x) = k$, avec :

- $k \geq 0$, $\int_a^b k dx = k \int_a^b dx = k(b-a) = 1$ d'où $f(x) = k = \frac{1}{b-a}$

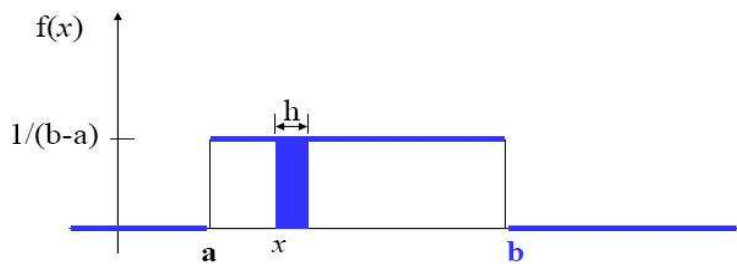
- $F(x) = \frac{x-a}{b-a}$

- $E(X) = \frac{(a+b)}{2}$; C'est également la médiane.

- $Var(X) = \frac{(b-a)^2}{12}$

[Pour démontrer ça, vous pouvez vous amuser à tout intégrer...Sinon, admettez-le.]

On peut calculer la **surface du rectangle** ci-dessous **de base h** et de hauteur $\frac{1}{(b-a)}$



Pour calculer cette surface, on utilise la relation suivante :

$$\int_x^{x+h} k dt = k \int_x^{x+h} dt = \frac{1}{(b-a)}(x+h-x) = \frac{h}{(b-a)} = P(x \in [x; x+h])$$

La loi uniforme **modélise les tirages des nombres « au hasard »**. En effet, on peut donner facilement les probabilités de tomber dans un intervalle donné.

B. Variable aléatoire continue exponentielle

La variable aléatoire continue exponentielle est définie sur l'univers \mathbf{R}^+ par :

- λ le **paramètre de la variable aléatoire continue exponentielle**, correspondant également à l'ordonnée à l'origine.
- La **densité de probabilité** $f(x) = \lambda e^{-\lambda x}$
- La **fonction de répartition** $F(X) = 1 - e^{-\lambda x}$
- L'**espérance** $E(X) = \frac{1}{\lambda}$
- La **variance** $Var(X) = \frac{1}{\lambda^2}$
- La **médiane** telle que $F(Me) = \frac{1}{2}$ et donc $Me = \frac{\ln(2)}{\lambda}$

D'après la fonction de répartition, on en déduit aussi que :

$$P(X > a) = P(X \geq a) = 1 - F(a) = e^{-\lambda a}$$

$$P(X < a) = P(X \leq a) = F(a) = 1 - e^{-\lambda a}$$

La loi exponentielle est dite « **sans mémoire** », on ne tient pas compte des faits antérieurs :

$$P(X > a+b | X > a) = \frac{P(X > a+b \cap X > a)}{P(X > a)} = \frac{P(X > a+b | X > a)}{P(X > a)} = \frac{e^{-\lambda(a+b)}}{e^{-\lambda a}} = e^{-\lambda(a+b-a)} = P(X > b)$$

La loi exponentielle modélise les durées de vie lorsque le vieillissement n'intervient pas.

[Pour les calculs d'exponentielle, voir le second tableau page 9, dans la deuxième colonne]

III. Applications aux tests diagnostiques

Le plus souvent, pour établir un diagnostic, c'est-à-dire distinguer un sujet malade d'un non malade, on utilise des **grandeurs continues** (pression artérielle, taux de glucose sanguin, ...). La 1^{ère} étape est de **définir un seuil** au-dessus duquel les résultats seront considérés comme **positifs**, et en dessous duquel les résultats seront considérés comme **négatifs** (ou le contraire...).

A. Seuil diagnostique d'un test

A partir de la valeur de X, on peut tenter de « discriminer » les sujets malades et non malades. Par exemple, le test sera « positif » si $X \leq \text{seuil}$, et « négatif » si $X > \text{seuil}$.

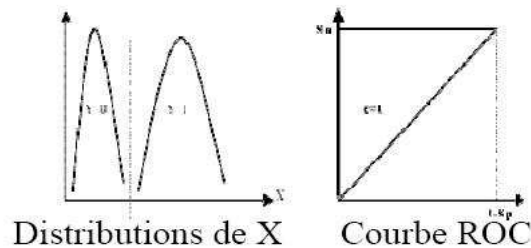
Pour quantifier sa valeur diagnostique, on calcule **pour un seuil** la **sensibilité et spécificité**.

A chaque seuil correspond des valeurs de sensibilité (% « vrais positifs ») et de spécificité (% « vrais négatifs ») ou de son complémentaire (% « faux positifs »). On peut alors tracer ces résultats sur un plan.

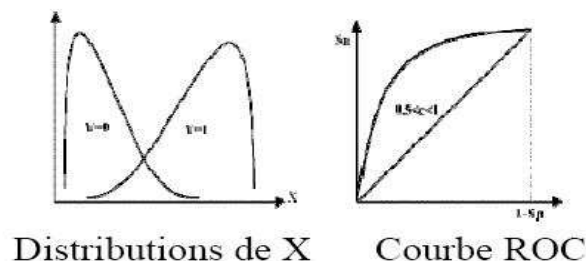
Par tradition, on trace **en abscisse 1-Spécificité (FP)** et **en ordonnée la Sensibilité (VP)**. La courbe obtenue est dite « **Courbe ROC** » [*promesses tenues*].

B. Courbe ROC

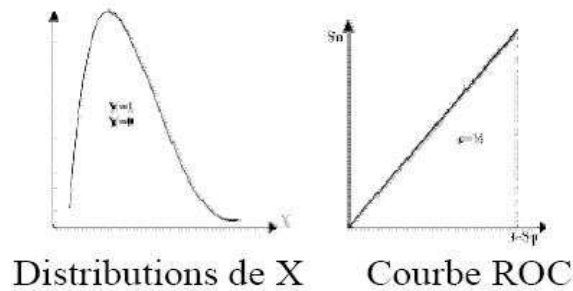
En situation **discriminante totale** (les distributions de X chez les malades et les non malades sont totalement distinctes [*du quasi jamais vu !*]), **la surface sous la courbe ROC est égale à 1**.



Sinon, si **les deux distributions de X des malades et non malades se chevauchent partiellement** [*cas le plus fréquent*], **la surface sous la courbe ROC est comprise entre 0,5 et 1**.



Si les deux distributions de X des malades et non malades sont superposées [ce qui est totalement inutile, ce qui explique qu'on change alors de signe pour le seuil.], la surface sous la courbe ROC est égale à 0,5.



La courbe ROC montre bien le **compromis entre sensibilité et spécificité** (quand l'une augmente, l'autre diminue). **Plus la courbe est proche de l'angle supérieur gauche, plus le marqueur a de bonnes valeurs diagnostiques.** Plus la courbe se rapproche de la diagonale, moins sa valeur diagnostique est bonne.

L'aire sous la courbe ROC est une mesure de la performance diagnostique du marqueur.

Ce document, ainsi que tous les cours P1, sont disponibles gratuitement sur <http://coursplbichat-larib.weebly.com>