CM 3
Méthodes de Simulation Informatique

Amaya Nogales Gómez
amaya.nogales-gomez@univ-cotedazur.fr

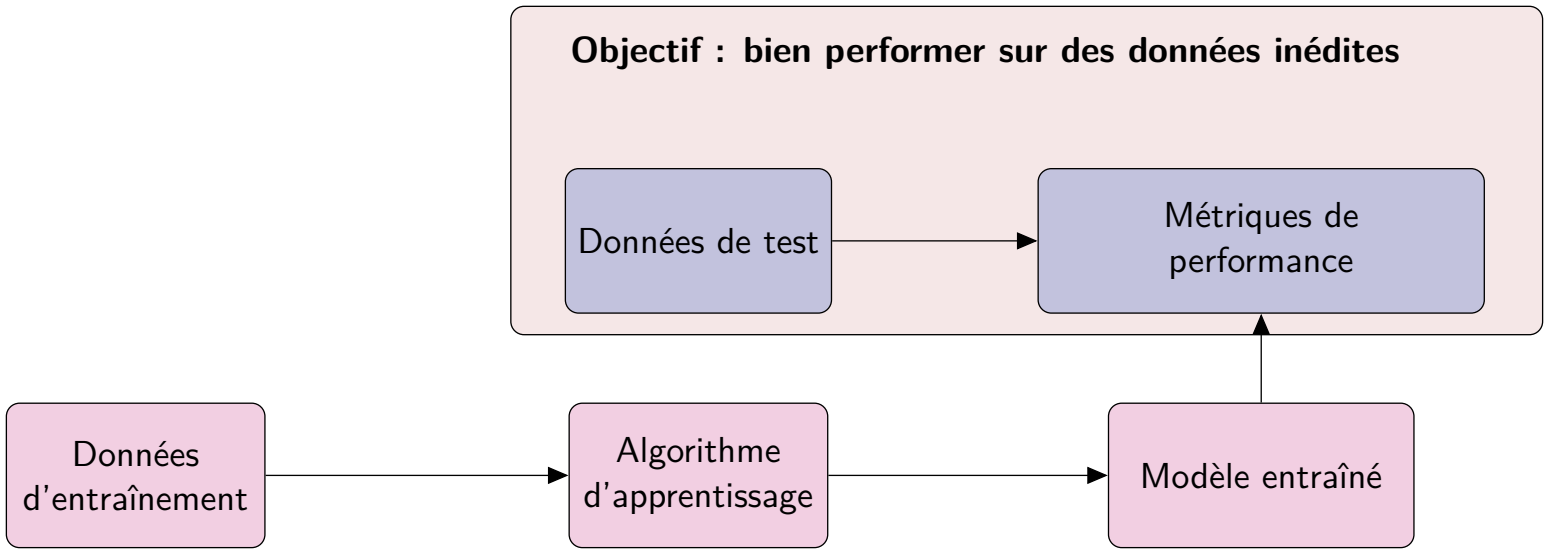
Licence 3 Informatique
Université Côte d'Azur

25 février 2022

Plan du cours

1. Introduction
   - Préliminaires
   - Python: numpy, pandas
2. Base de données
   - Generation des données synthétiques
   - Base de données reels
3. Analyse descriptive
4. Techniques d'apprentissage supervisée
   - Support Vector Machines
   - Régression logistique
5. Contrôle de connaissances
6. Techniques de validation
7. Elements de la méthodologie scientifique
8. LaTeX
   - Écriture de textes scientifiques
   - Beamer: présentations et posters scientifiques

Approche d'apprentissage automatique de base



Notation

- Base de données (X, Y)
- X: variables
- $Y \in \{-1, +1\}$ étiquette, classe du groupe
- $\hat{Y} = f(X)$: prédiction

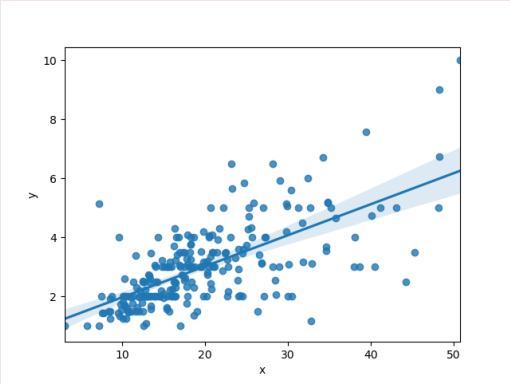
Abus de notation

- variable $\equiv$ attribut $\equiv$ caractéristique
- objet $\equiv$ individu $\equiv$ observation

Apprentissage supervisé: régression et classification

Régression

Sortie : fonction continue.

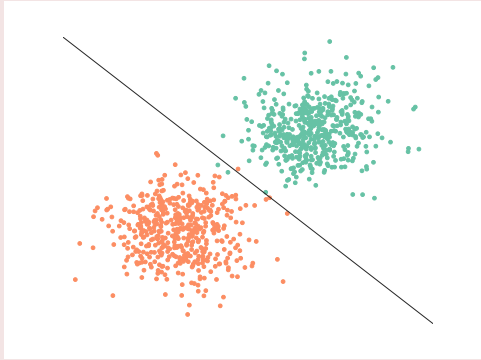


Exemples:

- Prévision (météo)
- Taille des animaux
- Actions

Classification

Sortie: règle de séparation.

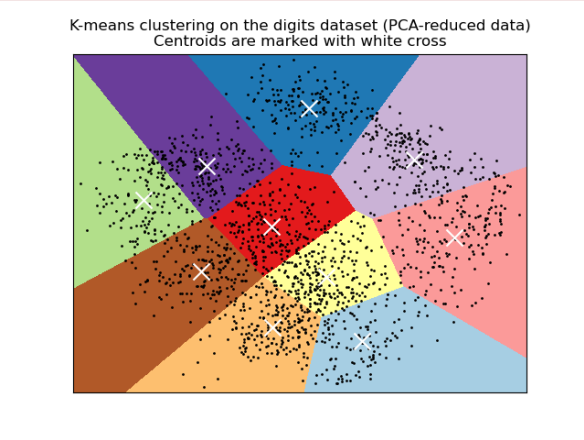


Examples:

- Rembourser un prêt
- Acceptation d'entrée à l'université
- Classement des images

Apprentissage non supervisé

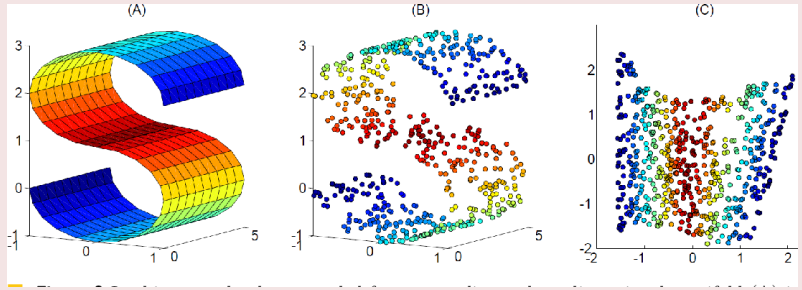
Clustering



Exemples:

- Systèmes de recommandation
- Analyse des réseaux sociaux

Réduction de la dimensionnalité



Exemples:

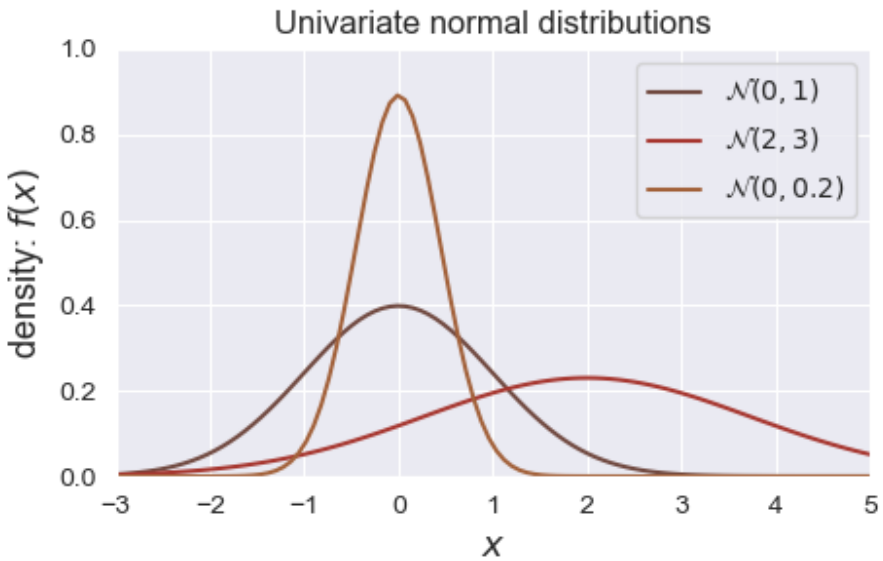
- Visualisation de données
- Stockage de données
- Complexité de calcul

Distribution gaussienne

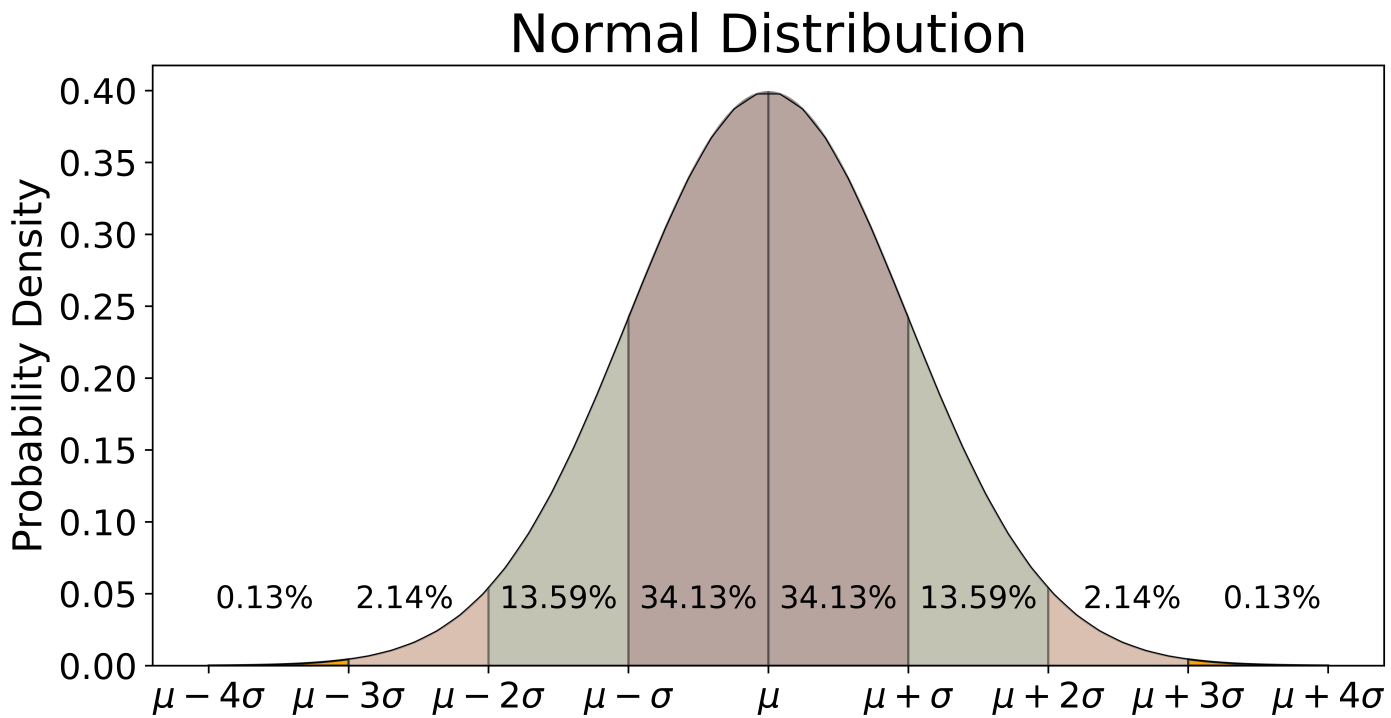
Une distribution normale (ou gaussienne) est un type de distribution de probabilité continue pour une variable aléatoire à valeur réelle. La forme générale de sa fonction de densité de probabilité est

$$X \sim N(\mu, \sigma^2)$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



$$P(\mu - n\sigma \leq X \leq \mu + n\sigma) = F(\mu + n\sigma) - F(\mu - n\sigma)$$
$$F(x) = \int_{-\infty}^{x} f(x)dx$$



Définition des statistiques

- A l'origine, l'objectif principal de la statistique était de recueillir des données démographiques, sociologiques et économiques.

Définition des statistiques

- A l'origine, l'objectif principal de la statistique était de recueillir des données démographiques, sociologiques et économiques.
- L'activité statistique est **ancienne**, et résulte de l'intérêt des **gouvernements** pour connaître les ressources de ceux qui pourraient apporter.

Définition des statistiques

- A l'origine, l'objectif principal de la statistique était de recueillir des données démographiques, sociologiques et économiques.
- L'activité statistique est **ancienne**, et résulte de l'intérêt des **gouvernements** pour connaître les ressources de ceux qui pourraient apporter.
- Le mot statistique a été utilisé pour la première fois dans **Allemagne** vers le milieu du **17e siècle**, et fait référence à la compilation de données et de documents útiles pour la **administration d'état**.

Définition des statistiques

- A l'origine, l'objectif principal de la statistique était de recueillir des données démographiques, sociologiques et économiques.
- L'activité statistique est **ancienne**, et résulte de l'intérêt des **gouvernements** pour connaître les ressources de ceux qui pourraient apporter.
- Le mot statistique a été utilisé pour la première fois dans **Allemagne** vers le milieu du **17e siècle**, et fait référence à la compilation de données et de documents útiles pour la **administration d'état**.
- Avant le **Empire romain**, ils avaient déjà été recensés ou compté les richesses, les soldats, les navires, les loyers public, les habitants, etc.

Définition des statistiques

- A l'origine, l'objectif principal de la statistique était de recueillir des données démographiques, sociologiques et économiques.
- L'activité statistique est **ancienne**, et résulte de l'intérêt des **gouvernements** pour connaître les ressources de ceux qui pourraient apporter.
- Le mot statistique a été utilisé pour la première fois dans **Allemagne** vers le milieu du **17e siècle**, et fait référence à la compilation de données et de documents útiles pour la **administration d'état**.
- Avant le **Empire romain**, ils avaient déjà été recensés ou compté les richesses, les soldats, les navires, les loyers public, les habitants, etc.
- La racine "**status**" (état des choses) justifie donc le mot Statistique.

Définition des statistiques

En raison du grand progrès de la science, les objectifs initiaux ont été élargis et sont inclus dans la définition suivante:

Définition des statistiques

La statistique est la **science** qui traite de la théorie et de l'application de méthodes appropriées pour **collecter, représenter, résumer des données, les analyser et en tirer des conclusions**.

Deux parties fondamentales se distinguent dans la définition de la statistique:

- Collecte et analyse des données pour donner une description des caractéristiques étudiées sur un ensemble d'individus, tirer des conclusions sur sa structure et ses relations existant avec d'autres groupes auxquels il est comparé.

  L'ensemble de ces techniques est appelé **Statistiques Descriptives**
- Faire des déductions sur les caractéristiques de la population pour de l'étude d'un sous-ensemble de la population ou de l'échantillon.

  C'est le but de **Statistiques inductives ou inférentielles**, qui, sur la base du calcul des probabilités, déduit, induit ou estime des lois générales sur le comportement de la population.

Concepts généraux

- **Population:** Ensemble d'éléments sur lesquels porte l'enquête et dont les données sont extraites.

Concepts généraux

- **Population:** Ensemble d'éléments sur lesquels porte l'enquête et dont les données sont extraites.
- **Unité Statistique ou Individu:** Chacun des éléments qui composent la population.

  C'est une entité observable qui n'a pas besoin d'être une personne, ce peut être un objet ou même quelque chose d'abstrait.

Concepts généraux

- **Population:** Ensemble d'éléments sur lesquels porte l'enquête et dont les données sont extraites.
- **Unité Statistique ou Individu:** Chacun des éléments qui composent la population.

  C'est une entité observable qui n'a pas besoin d'être une personne, ce peut être un objet ou même quelque chose d'abstrait.
- **Recensement:** Examen de tous les individus qui composent la population.

Concepts généraux

- **Population:** Ensemble d'éléments sur lesquels porte l'enquête et dont les données sont extraites.
- **Unité Statistique ou Individu:** Chacun des éléments qui composent la population.

  C'est une entité observable qui n'a pas besoin d'être une personne, ce peut être un objet ou même quelque chose d'abstrait.
- **Recensement:** Examen de tous les individus qui composent la population.
- **Échantillon:** Sous-ensemble d'éléments de la population.

Concepts généraux

- **Population:** Ensemble d'éléments sur lesquels porte l'enquête et dont les données sont extraites.
- **Unité Statistique ou Individu:** Chacun des éléments qui composent la population.

  C'est une entité observable qui n'a pas besoin d'être une personne, ce peut être un objet ou même quelque chose d'abstrait.
- **Recensement:** Examen de tous les individus qui composent la population.
- **Échantillon:** Sous-ensemble d'éléments de la population.
- **Caractère: Qualité ou propriété observable chez un individu.**

Concepts généraux

- **Population:** Ensemble d'éléments sur lesquels porte l'enquête et dont les données sont extraites.
- **Unité Statistique ou Individu:** Chacun des éléments qui composent la population.

  C'est une entité observable qui n'a pas besoin d'être une personne, ce peut être un objet ou même quelque chose d'abstrait.
- **Recensement:** Examen de tous les individus qui composent la population.
- **Échantillon:** Sous-ensemble d'éléments de la population.
- **Caractère: Qualité ou propriété observable chez un individu.**
- **Modalité**: Différentes situations ou variantes possibles du caractère.

Exemple

- **Population:** Ensemble d'ordinateurs des laboratoires Petit Valrose.
- **Caractères:** Système d'exploitation, vitesse du processeur, mémoire RAM (Gb.)
- **Modalités:**
  - **Système d'exploitation:** { Windows XP, Linux, Windows 7, ... }.
  - **Vitesse du processeur:** { 1.4, 1.7, 1.5, 2.4, ... }.
  - **Mémoire RAM (Gb.):** { 1, 2, 4, 8, ... }

Types de caractères

Types de caractères

- **Quantitatif:** Ils peuvent être mesurés ou quantifiés.
- **Qualitatif:** Ils ne peuvent pas être mesurés.

Types de caractères

Types de caractères

- **Quantitatif:** Ils peuvent être mesurés ou quantifiés.
- **Qualitatif:** Ils ne peuvent pas être mesurés.

Caractères quantitatifs

- Elles sont appelées **variables**.
- A chaque modalité est attribué un nombre réel nommé **valeur**.
- Ils peuvent être de deux types :
  - **Variables discrètes:** entre deux valeurs consécutives, la variable ne peut pas prendre une autre valeur.
    Exemple: Nombre de cœurs de processeur, nombre de ports USB.
  - **Variables continues:** entre deux valeurs, la variable peut prendre des valeurs infinies. *Exemple: Temps de démarrage, température du processeur.*

Caractères qualitatifs

- Ils sont appelés **attributs** ou variables categorielles.
- Ils peuvent être de deux types :
  - **Nominaux:** étant donné deux modalités ou plus, nous ne pouvons que vérifier qu'ils soient différents ou non.

    Exemple: Système d'exploitation: Windows XP, Ubuntu, Redhat, Mac OS, Debian, Windows 7
  - **Ordinaux:** étant donné deux modalités ou plus, non seulement pouvons-nous vérifier s'ils sont différents ou non, mais aussi un ordre peut être établi.

    Exemple: type de processeur Intel Core: i3, i5, i7

Analyse statistique descriptive

La première étape d'une analyse statistique d'un échantillon de données: obtenir des **tables** ou d'autres sorties de visualisation permettant de **résumer** et **ordonner** les données, aidant à son analyse postérieure.

- Considérons un échantillon composé de n individus, pour lequel on observera la variable X, ayant n données : $x_1, x_2, ..., x_n$.
- Soit $x_1, \ldots, x_k$ les k différentes **valeurs** observées.

Fréquence absolue

La fréquence (ou fréquence absolue) d'un événement x_i est le nombre n_i de fois où l'observation s'est produite dans une expérience.

$$\sum_{i=1}^{k} n_i = n$$

fréquence relative

La fréquence relative de x_i, notée f_i, est la proportion d'occurrences observées pour cet événement, c'est-à -dire,

$$f_i = \frac{n_i}{n}, \qquad 1 \leq i \leq k$$

$$\sum_{i=1}^{k} f_i = 1$$

Fréquences absolues, relatives et cumulées

Les concepts suivants n'ont de sens que pour les **variables** et **attributs ordinaux**.

Définition (fréquence cumulée)

La **fréquence cumulée (absolue)** de x_i, qui on notera N_i, c'est le nombre d'observations de valeur inférieur ou égal à x_i,

$$N_i = \sum_{j=1}^{i} n_j, \qquad 1 \leq i \leq k.$$

Il faut que: $N_1 = n_1, \quad N_k = n, \quad n_i = N_i - N_{i-1}, \ 2 \leq i \leq k$

Fréquences absolues, relatives et cumulées

Définition (fréquence relative cumulée)

La **fréquence relative cumulée** de x_i, que nous noterons F_i, est la proportion d'observations avec une valeur inférieure ou pareil que x_i,

$$F_i = \frac{N_i}{n} = \sum_{j=1}^{i} f_j, \qquad 1 \leq i \leq k.$$

Il faut que: $F_1 = f_1, \quad F_k = 1, \quad f_i = F_i - F_{i-1}, \ 2 \leq i \leq k$

La fréquence relative cumulée exprimée en % est appelée **pourcentage cumulé**.

Distribution des fréquences

Une distribution de fréquence est un tableau (*tableau de fréquences*) ou un graphe (diagramme à barres ou histogramme) qui affiche la fréquence des événements dans un échantillon. Chaque entrée du tableau contient la fréquence des occurrences de valeurs dans un groupe ou un intervalle particulier.

x_i	n_i	N_i	f_i	F_i
x_1	n_1	N_1	f_1	F_1
x_2	n_2	N_2	f_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	N_k	f_k	F_k
	n		1	

Tableau des fréquences

Si X est une variable qui prend plusieurs valeurs différentes, alors il est habituel de les regrouper en **intervalles**.

$(L_{i-1}, L_i]$	n_i	N_i	f_i	F_i	x_i	a_i	h_i
$(L_0, L_1]$	n_1	N_1	f_1	F_1	x_1	a_1	h_1
$(L_1, L_2]$	n_2	N_2	f_2	F_2	x_2	a_2	h_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$(L_{k-1}, L_k]$	n_k	N_k	f_k	F_k	x_k	a_k	h_k
	n		1				

- $x_i = (L_i + L_{i-1})/2$ est le centre de chaque intervalle, nommé **marque de classe** et représentant l'intervalle.
- $a_i = L_i - L_{i-1}$ est le **amplitude** de l'intervalle.
- $h_i = n_i/a_i$ est la **densité de fréquence**.

Exemple

Le score COMPAS de récidive calculé pour les 15 suspects un jour donné était:

$$4,\ 3,\ 7,\ 5,\ 6,\ 4,\ 5,\ 4,\ 5,\ 6,\ 7,\ 7,\ 3,\ 4,\ 5$$

Exemple

Le score COMPAS de récidive calculé pour les 15 suspects un jour donné était:

$$4,\ 3,\ 7,\ 5,\ 6,\ 4,\ 5,\ 4,\ 5,\ 6,\ 7,\ 7,\ 3,\ 4,\ 5$$

Le tableau des fréquences pour cet échantillon est:

x_i	n_i	N_i	f_i	F_i
3	2	2	0.133	0.133
4	4	6	0.266	0.4
5	4	10	0.266	0.666
6	2	12	0.133	0.8
7	3	15	0.2	1
	15		1	

Exemple

Les longueurs sont mesurées en millimètres d'un certain composant de un système, obtenant les résultats suivants

$$0.2,\ 0.6,\ 1.1,\ 1.7,\ 1.9,\ 3.7,\ 3.8,\ 4.2,\ 4.5,\ 4.8,\ 5.3,$$
$$5.7,\ 6.2,\ 6.7,\ 7.5,\ 8.1,\ 8.5,\ 8.7,\ 9.2,\ 9.5$$

Exemple

Les longueurs sont mesurées en millimètres d'un certain composant de un système, obtenant les résultats suivants

$$0.2,\ 0.6,\ 1.1,\ 1.7,\ 1.9,\ 3.7,\ 3.8,\ 4.2,\ 4.5,\ 4.8,\ 5.3,$$
$$5.7,\ 6.2,\ 6.7,\ 7.5,\ 8.1,\ 8.5,\ 8.7,\ 9.2,\ 9.5$$

- Très peu de valeurs sont répétées (presque toutes les fréquences sont à 1), donc le tableau des fréquences qui serait obtenu serait trop long.
- Il est conseillé de regrouper les données en intervalles afin de résumer et de comprendre les informations contenues dans les données.
- Nous considérerons des intervalles de la forme $(a, b]$ lors de la division.

Exemple

Les longueurs sont mesurées en millimètres d'un certain composant de un système, obtenant les résultats suivants

$$0.2,\ 0.6,\ 1.1,\ 1.7,\ 1.9,\ 3.7,\ 3.8,\ 4.2,\ 4.5,\ 4.8,\ 5.3,$$
$$5.7,\ 6.2,\ 6.7,\ 7.5,\ 8.1,\ 8.5,\ 8.7,\ 9.2,\ 9.5$$

Exemple

Les longueurs sont mesurées en millimètres d'un certain composant de un système, obtenant les résultats suivants

$$0.2,\ 0.6,\ 1.1,\ 1.7,\ 1.9,\ 3.7,\ 3.8,\ 4.2,\ 4.5,\ 4.8,\ 5.3,$$
$$5.7,\ 6.2,\ 6.7,\ 7.5,\ 8.1,\ 8.5,\ 8.7,\ 9.2,\ 9.5$$

- Le **critère de division** n'est pas objectif. Principes généraux:
  - Il ne devrait pas y avoir trop peu d'intervalles, car trop d'informations seront perdues.
  - Il ne doit pas y avoir beaucoup d'intervalles, car le tableau résultant n'aura pas une taille adéquate.
  - Les intervalles doivent couvrir toutes les valeurs possibles et ne pas se chevaucher.

Exemple

$$0.2,\ 0.6,\ 1.1,\ 1.7,\ 1.9,\ 3.7,\ 3.8,\ 4.2,\ 4.5,\ 4.8,\ 5.3,$$
$$5.7,\ 6.2,\ 6.7,\ 7.5,\ 8.1,\ 8.5,\ 8.7,\ 9.2,\ 9.5$$

Exemple

$$0.2,\ 0.6,\ 1.1,\ 1.7,\ 1.9,\ 3.7,\ 3.8,\ 4.2,\ 4.5,\ 4.8,\ 5.3,$$
$$5.7,\ 6.2,\ 6.7,\ 7.5,\ 8.1,\ 8.5,\ 8.7,\ 9.2,\ 9.5$$

Nous avons considéré la division $(0,1], (1,3], (3,5], (5,6], (6,8], (8,10]$. Le tableau des fréquences serait :

$(L_{i-1}, L_i]$	n_i	N_i	f_i	F_i	x_i	a_i	h_i
(0,1]	2	2	0.1	0.1	0.5	1	2
(1,3]	3	5	0.15	0.25	2	2	1.5
(3,5]	5	10	0.25	0.5	4	2	2.5
(5,6]	2	12	0.1	0.6	5.5	1	2
(6,8]	3	15	0.15	0.75	7	2	1.5
(8,10]	5	20	0.25	1	9	2	2.5
	20		1				

Remarque: Tous les intervalles ne doivent pas nécessairement avoir la même amplitude.

Méthodes graphiques

Représentations graphiques

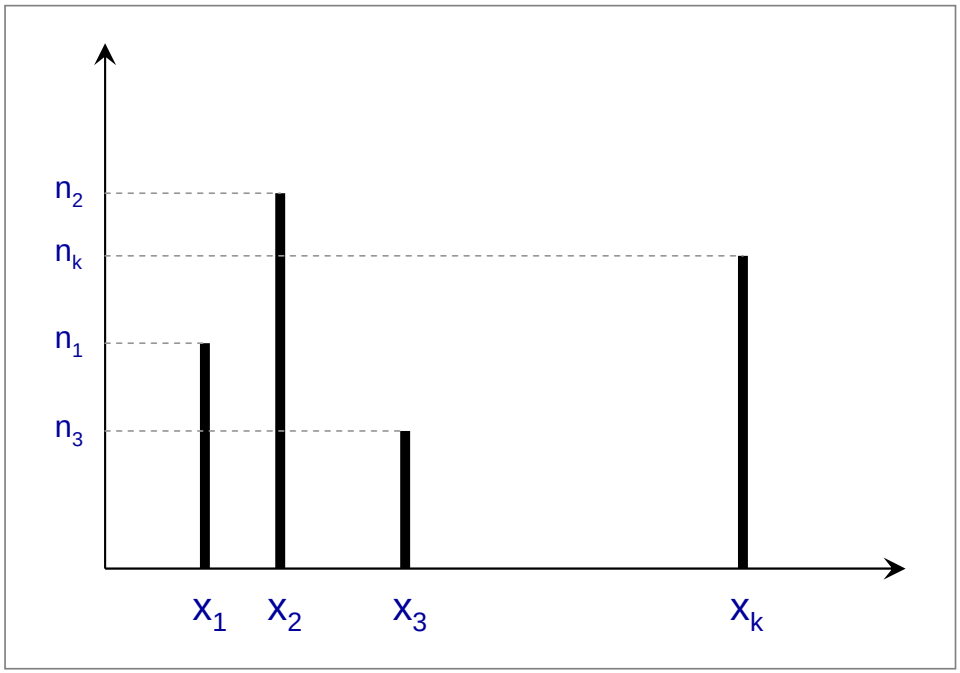
- Ils permettent de montrer de manière claire et concise les caractéristiques des données.
- Ils sont un élément auxiliaire d'analyse.
- Pour eux-mêmes, ils ne sont pas utiles pour effectuer une étude rigoureuse des informations contenues dans les données.

Représentations Graphiques

- Variable
  - Discrète
    - Diagramme à barres
    - Polygone de fréquence
    - Courbe cumulative
  - Continue
    - Histogramme
    - Polygone de fréquence
    - Courbe cumulative
- Attribut
  - Diagramme à barres
  - Diagramme circulaire
  - Pictogramme

Diagramme à barres

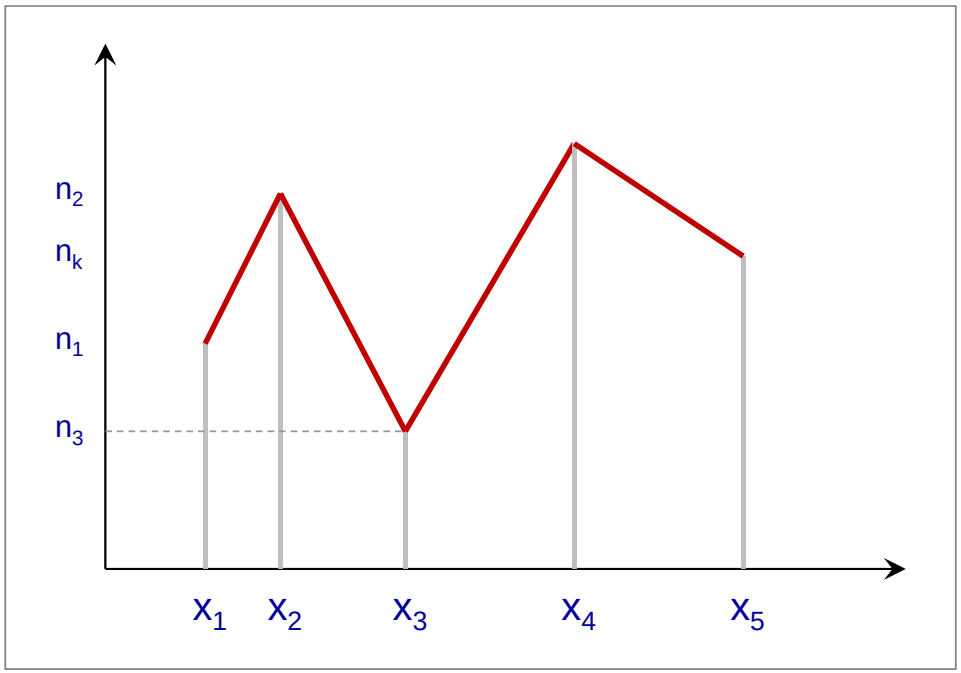
- Les valeurs observées de la variable sont représentées sur l'axe des abscisses.
- Pour chaque valeur observée, un segment de hauteur égale à sa fréquence absolue est relevé.



Les fréquences relatives peuvent être représentées au lieu d'être absolues. Dans ce cas, la hauteur de chaque segment serait f_i, au lieu de n_i.

Polygone de fréquences

- C'est une autre façon de représenter les mêmes informations que le graphique à barres.
- Il est construit en joignant les extrémités supérieures des segments du graphique à barres.
- Il peut également être utilisé pour représenter les fréquences relatives au lieu d'absolu.



Courbe cumulative

- Il permet de représenter les fréquences cumulées.
- C'est une fonction échelon qui est égale à 0 depuis $-\infty$ jusqu'à la plus petite valeur observée.
- Entre la première et la deuxième plus petite valeur, la fonction est égale à N_1.
- Entre la deuxième plus petite valeur et la troisième plus petite, la fonction est égale à N_2.
- A partir de la plus grande valeur, la fonction est constante et égale au nombre d'observations n.
- La fonction tracée est discontinue à chaque valeur observée, continue vers la droite.
- Aussi les fréquences cumulées peuvent être représentées relatif, au lieu d'absolu, auquel cas à partir de la plus grande, la fonction est constante et égale à 1.

Histogramme

- Sur l'axe des abscisses se trouvent les intervalles dans lesquels les données ont été regroupées.
- Sur chaque intervalle est tracé un rectangle dont la aire est égale à la fréquence absolue, n_i.

$$\text{Aire} = \text{largeur} \times \text{longueur} \quad \Rightarrow \quad \text{longueur} = \frac{n_i}{a_i} = h_i$$

- Des fréquences relatives au lieu d'absolues peuvent également être représentées. Dans un tel cas, l'aire des rectangles serait f_i et sa hauteur:

$$h_i = \frac{f_i}{a_i}$$

- Si tous les intervalles ont la même largeur, n_i (ou f_i) peut être prise comme longueur des rectangles.

Polygone de fréquences

- C'est une manière alternative de représenter les mêmes informations que l'histogramme.
- Pour le construire, joignez les centres des faces supérieures des rectangles.
- Par conséquent, il est construit en joignant les points de coordonnées (x_i, h_i).
- Il peut également être utilisé pour représenter des fréquences relatives au lieu de fréquences absolues.

Courbe cumulative

- Comme dans le cas discret, il permet de représenter les fréquences cumulées.
- C'est une ligne brisée qui vaut 0 de $-\infty$ jusqu'à L_0.
- Dans le premier intervalle, $(L_0, L_1]$, est un segment joignant les points $(L_0, 0)$ y (L_1, N_1)
- Dans le second intervalle, $(L_1, L_2]$, est un segment qui joint les points (L_1, N_1) y (L_2, N_2).
- ...
- De extrémité droite du dernier intervalle à $+\infty$ est toujours n.
- La fonction dessinée est continue.
- Vous pouvez également tracer les fréquences cumulées relatif, au lieu d'absolu, auquel cas à partir de la extrémité droite du dernier intervalle, la fonction est constante et égale à 1.

Exemple: variable discrète

Considérons la variable X="score de risque COMPAS du défendeur".

x_i	n_i	N_i	f_i	F_i
3	2	2	0.133	0.133
4	4	6	0.266	0.4
5	4	10	0.266	0.666
6	2	12	0.133	0.8
7	3	15	0.2	1
	15		1	

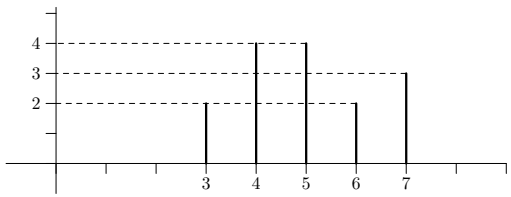
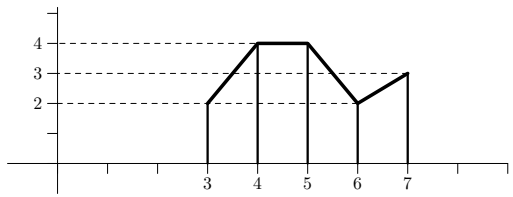
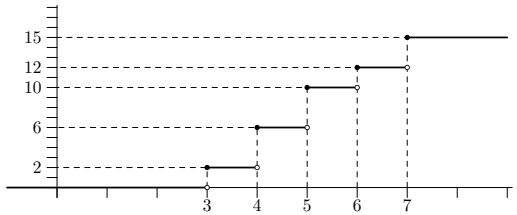


Diagramme à barres



Polygone de fréquences

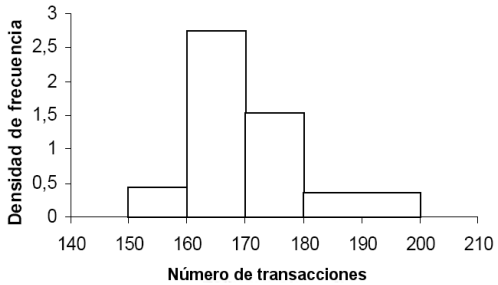


Courbe cumulative

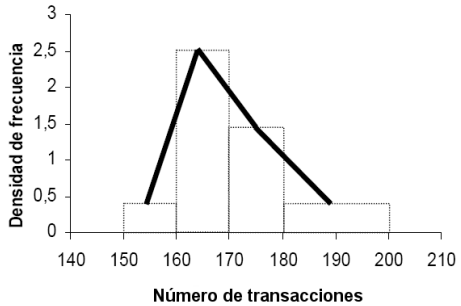
Exemple: variable continue

Considérons la variable X="Taille en cm", observé chez $n = 50$ suspects.

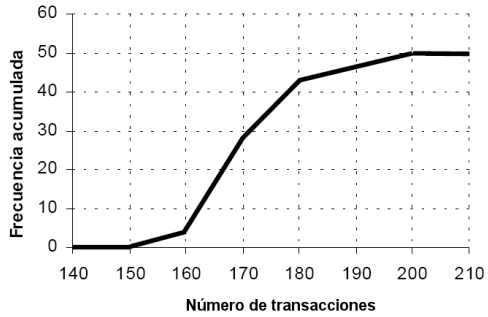
$(L_{i-1}, L_i]$	n_i	a_i	h_i	N_i
$(150, 160]$	4	10	0.4	4
$(160, 170]$	25	10	2.5	29
$(170, 180]$	14	10	1.4	43
$(180, 200]$	7	20	0.35	50



Histogramme



Polygone de fréquences



Courbe cumulative

Diagramme à barres

- Présente des variables catégorielles avec des barres rectangulaires avec des hauteurs ou des longueurs proportionnelles aux fréquences.
- Les barres peuvent être tracées verticalement ou horizontalement.

EXEMPLE

La répartition des 12 utilisateurs selon le système d'exploitation utilisé est la suivante :

Système d'exploitation	n_i
Linux (L)	5
Windows (W)	3
Mac OS (M)	4

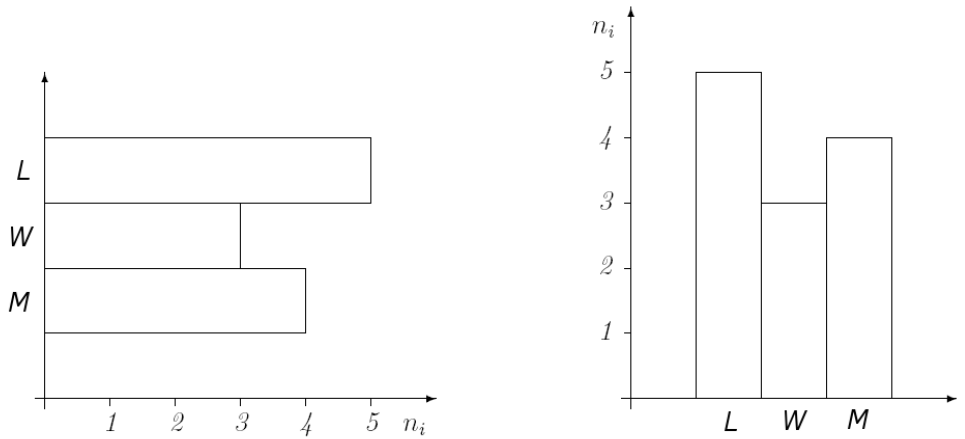
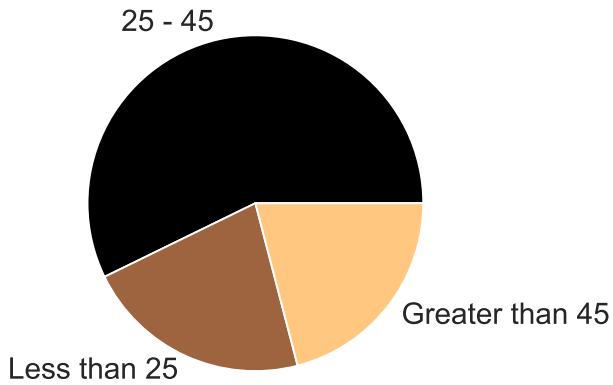


Diagramme circulaire

A l'intérieur d'un cercle, chaque catégorie est affectée à un secteur proportionnel à sa fréquence.

Âge du défenseur	n_i	f_i	$f_i \times 360$
Moins de 25	1347	0.22	79^o
Entre 25 et 45	3532	0.57	206^o
Plus de 45	1293	0.21	75^o
	6172	1	360^o



Pictogramme

Dans ce type de graphiques, on utilise des figures liés à la phénomène étudié, de sorte que sa taille ou nombre indique la fréquence associée à chaque modalité.

EXEMPLE

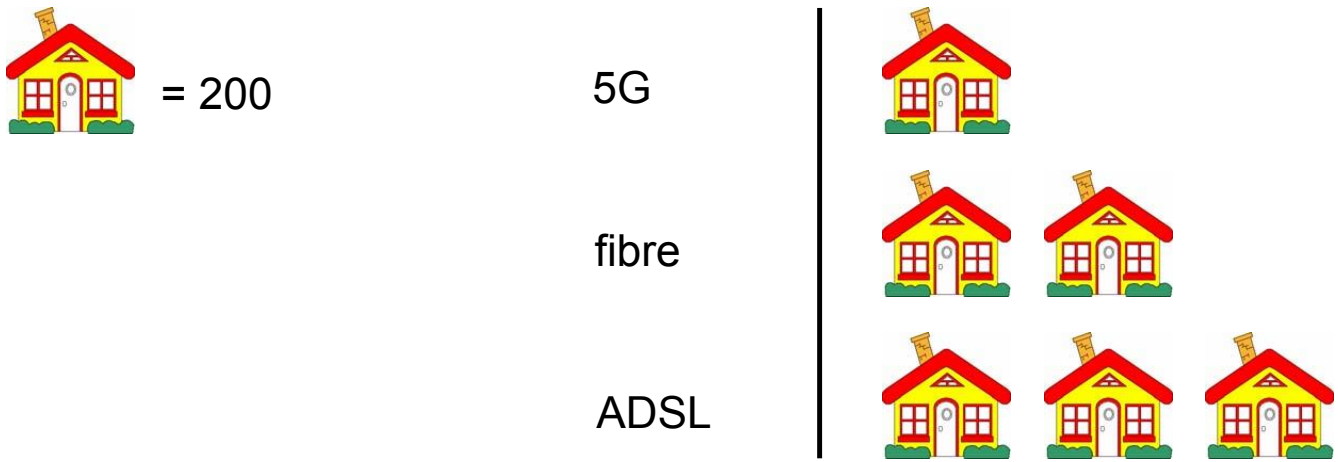
Le tableau suivant présente les moyens d'accès à Internet utilisés dans 1200 foyers d'une certaine localité :

moyens d'accès	n_i
5G	200
fibre	400
ADSL	600
	1200

Pictogramme

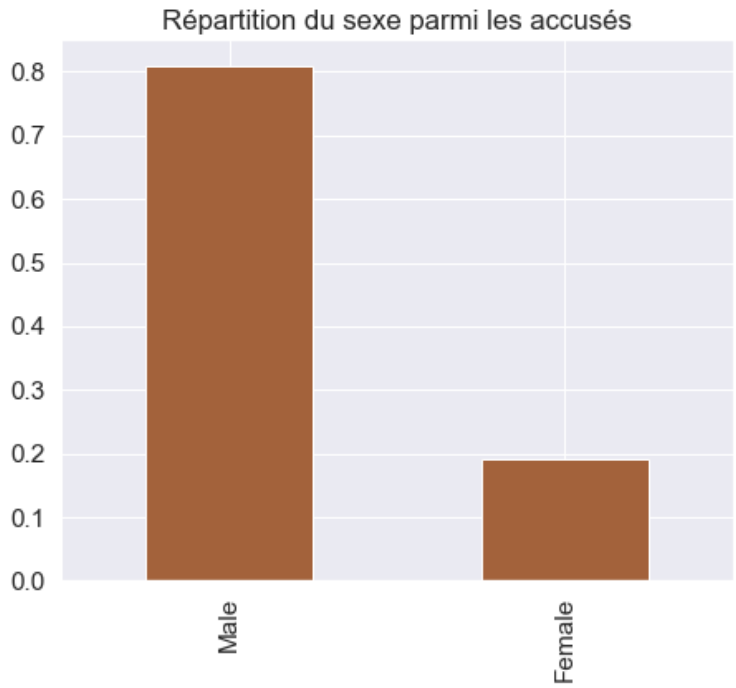
Pictogramme avec des figures de taille proportionnelle aux fréquences:

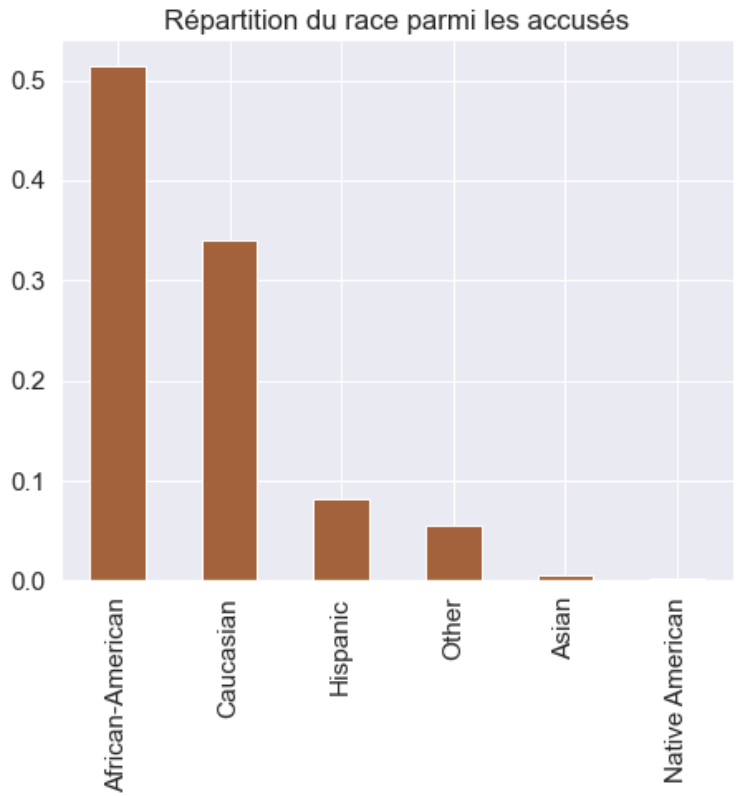
Pictogramme avec nombre de figures proportionnel aux fréquences:



Analyse de la base de données COMPAS

- Nous allons maintenant étudier l'algorithme COMPAS sur les données collectées par ProPublica.
- Nous analyserons les scores COMPAS pour le "Risque de récidive". Une analyse équivalente pourrait être faite pour le "Risque de récidive avec violence".
- Nous analysons: la répartition par sexe, les races, la répartition des scores déciles COMPAS pour différents groupes, l'analyse d'équité.





Près de 80% des accusés sont classés comme hommes, tandis que les accusés blancs et noirs représentent environ 85% de la population totale des accusés.

```
fig, axes = plt.subplots(1,2, figsize=(16,6))

(
    df.sex.value_counts(normalize=True)
    .plot(kind='bar', title='Répartition du sexe parmi les accusés', ax=axes[0],color='#A3623B')
)

(
    df.race.value_counts(normalize=True)
    .plot(kind='bar', title='Répartition du race parmi les accusés', ax=axes[1],color='#A3623B')
);


plt.savefig("Hist_sex_race2.pdf")
```

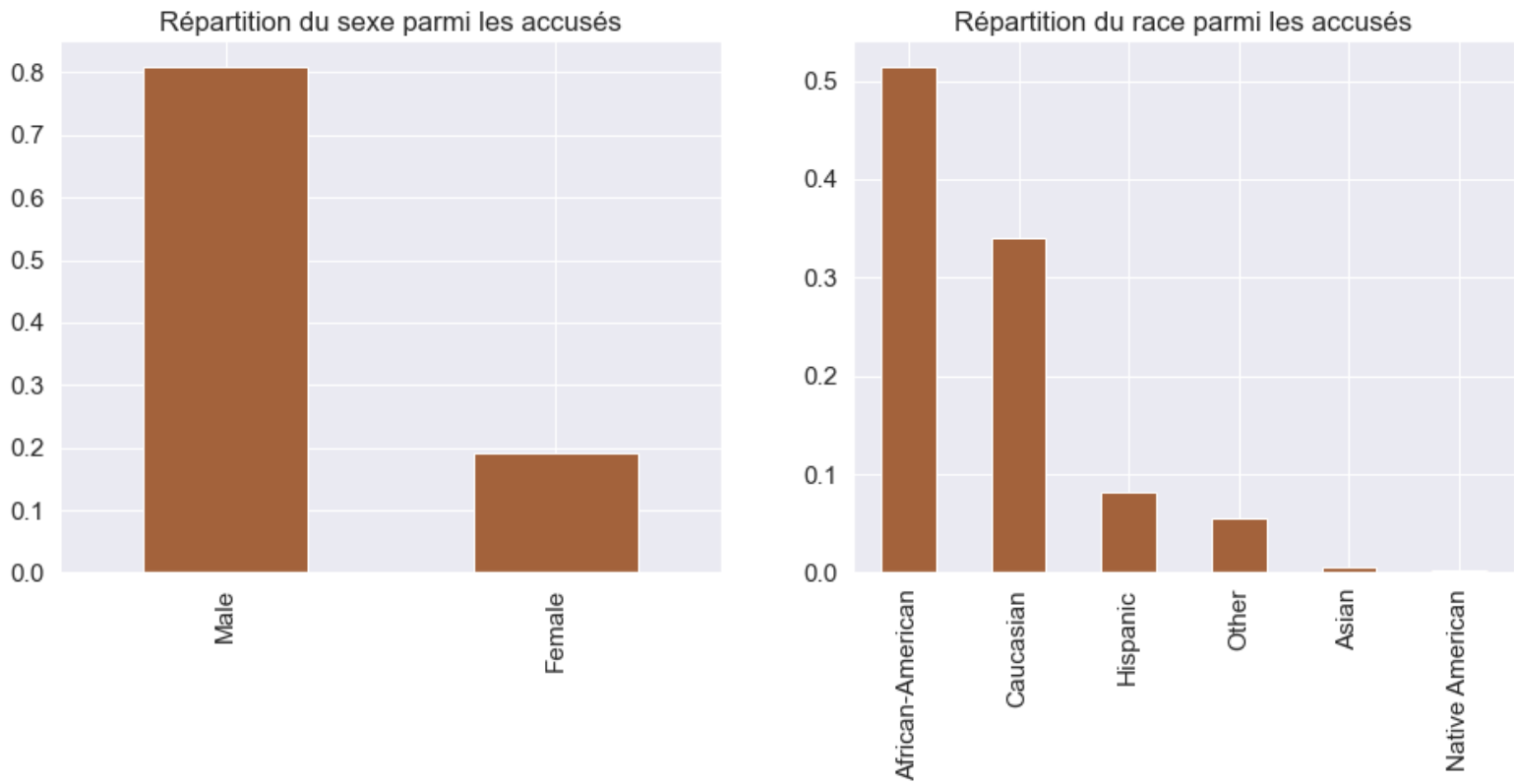
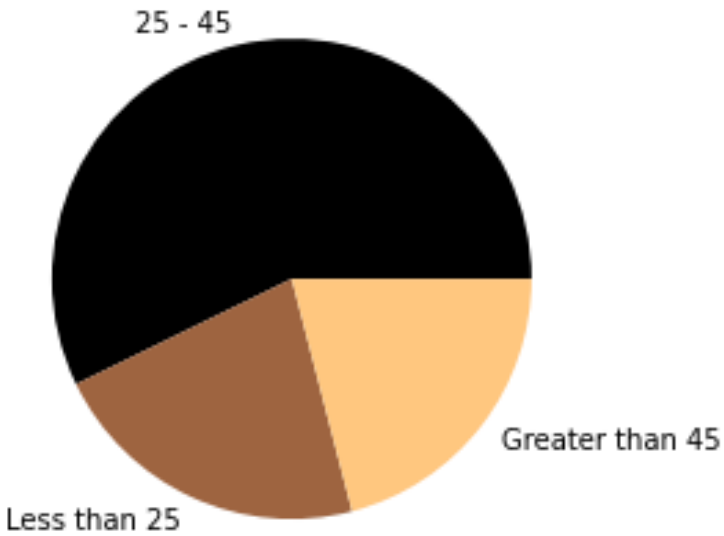


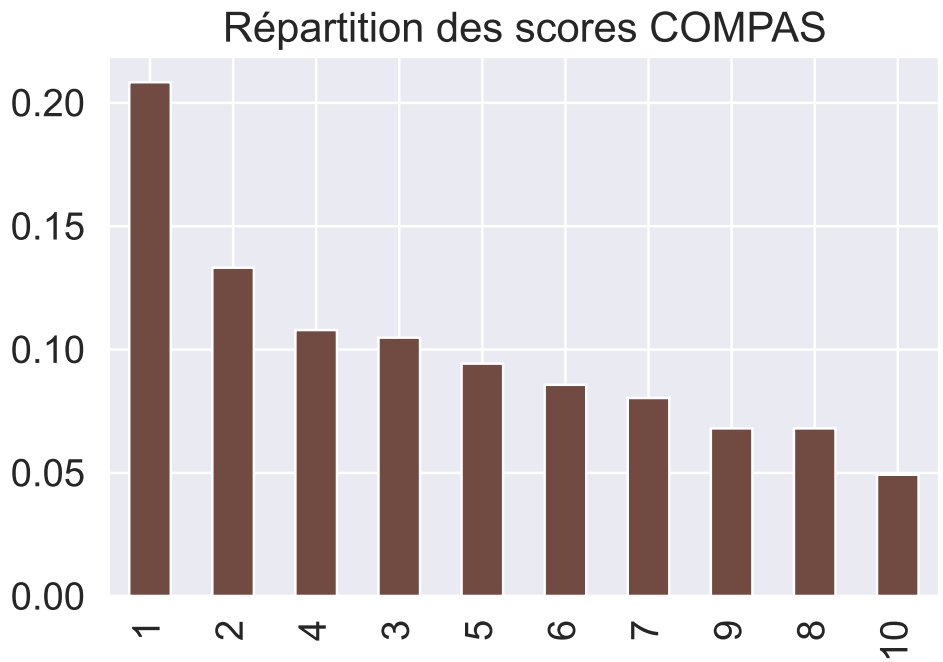
Diagramme circulaire

```
ax = df.age_cat.value_counts(normalize=False).plot(kind='pie',cmap='copper')
ax.set_ylabel('')
#df.plot.pie(y='age_cat', figsize=(5, 5))
plt.savefig("sect.pdf")
df.age_cat.value_counts(normalize=True)
```

```
25 - 45              0.572262
Less than 25         0.218244
Greater than 45      0.209494
Name: age_cat, dtype: float64
```



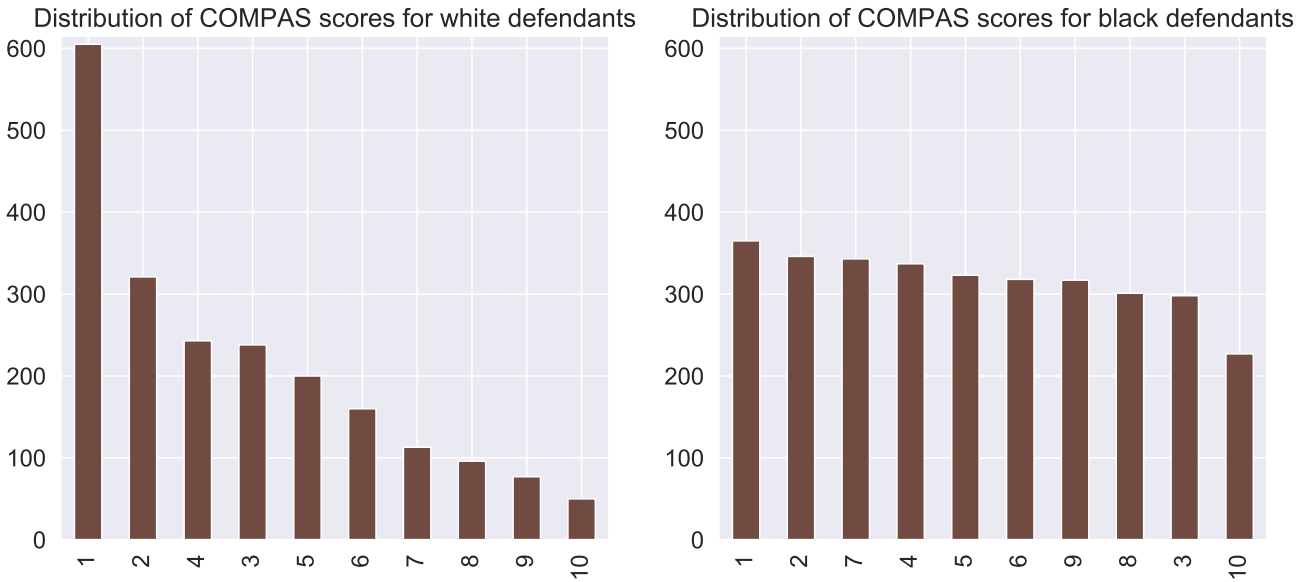
- Nous analysons les scores COMPAS pour le "Risque de récidive".
- Nous traçons la distribution des scores déciles COMPAS.
- Nous traçons la distribution de ces scores pour l'ensemble des 6 172 prévenus qui n'avaient pas été arrêtés pour une nouvelle infraction ou qui avaient récidivé dans les deux ans.



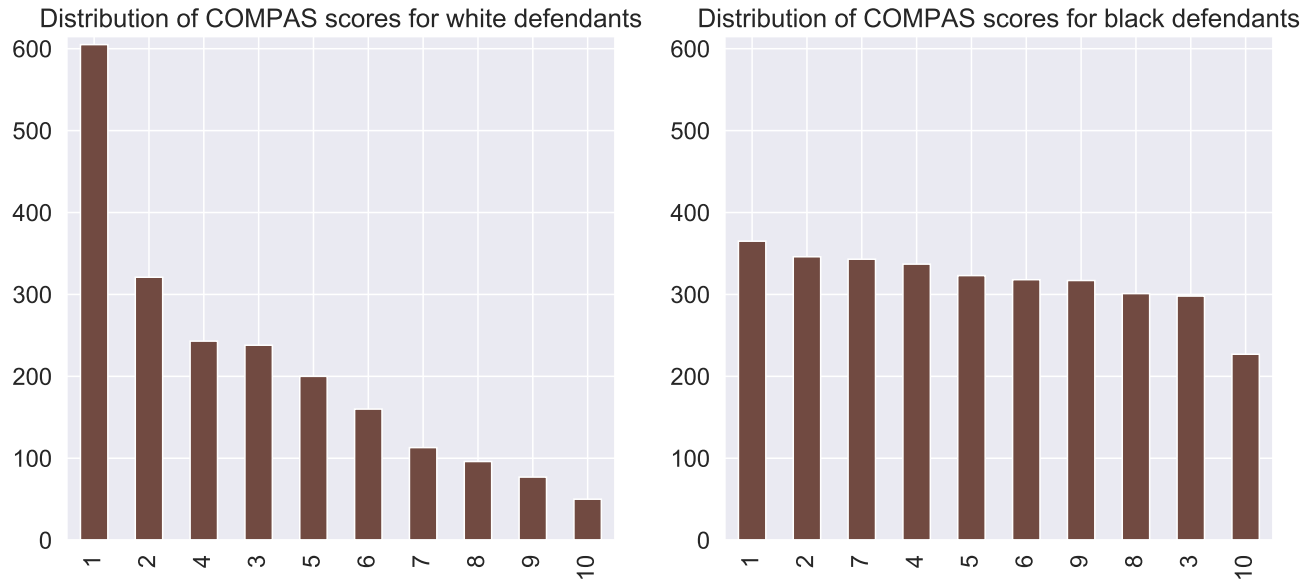

```
ax=df.decile_score.value_counts(normalize=True).plot(kind='bar',
                                                      title='Répartition des scores COMPAS', color='#714A41')

plt.savefig("Hist_score_all.pdf")
```

Comparaison du score COMPAS en fonction de la race

- Il existe une différence qualitative dans les distributions entre les accusés noirs et blancs.
- Les scores des accusés blancs étaient biaisés vers les catégories à faible risque.
- Les scores des accusés noirs étaient répartis de manière égale entre les scores.
- Ces observations ne prouvent aucun biais démographique.

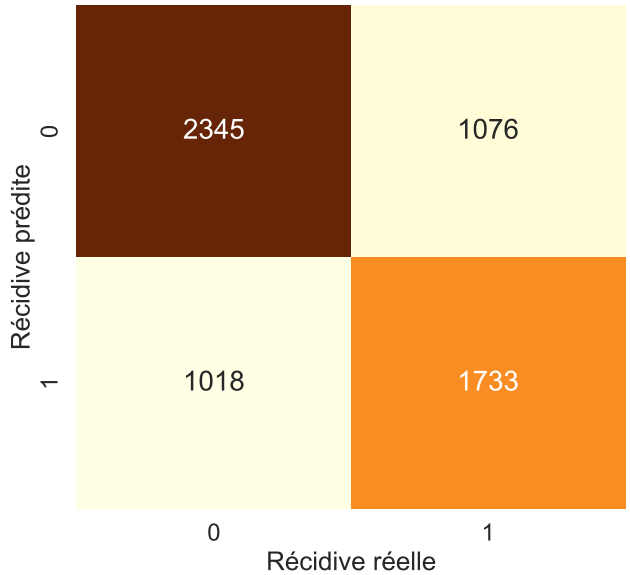


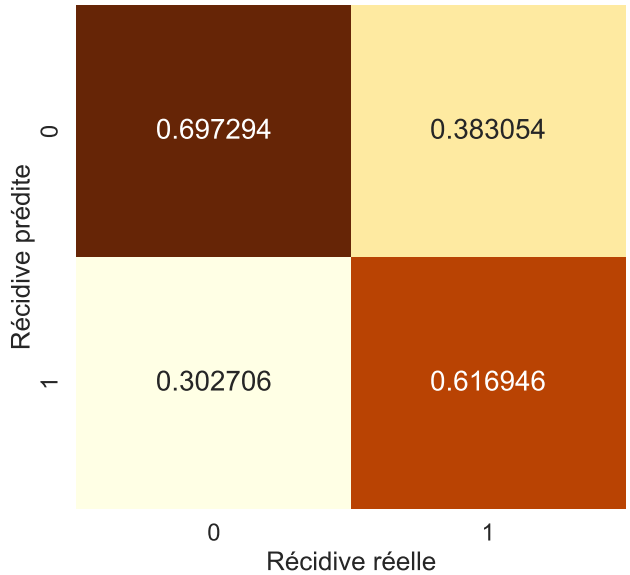
```
#Nous calculons la fréquence absolue maximale des decile_scores pour être caucasien et afro-américain
#afin de définir la limite de l'axe y
max_y=max(df.decile_score.loc[df.race.isin(['Caucasian'])].value_counts().max(),
          df.decile_score.loc[df.race.isin(['African-American'])].value_counts().max())



fig, axes = plt.subplots(1,2, figsize=(14,6),subplot_kw={'ylim': (0,max_y+10)})
#axes.set_ylim(0,600)

(
    df.decile_score.loc[df.race.isin(['Caucasian'])].value_counts()
    .plot(kind='bar', title='Distribution of COMPAS scores for white defendants', ax=axes[0], color='#714A41')
)

(
    df.decile_score.loc[df.race.isin(['African-American'])].value_counts()
    .plot(kind='bar', title='Distribution of COMPAS scores for black defendants', ax=axes[1], color='#714A41')
);


#df.decile_score.loc[df.race.isin(['Caucasian'])].value_counts()
plt.savefig("Hist_score_BW.pdf")
```



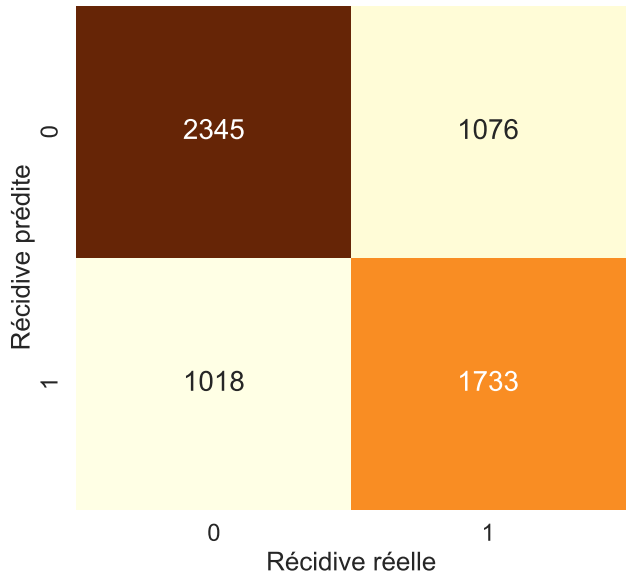
- L'algorithme COMPAS, sur l'ensemble de données, est relativement équilibré.
- La classe positive (récidive réels) représente 46% de la base de données, coincidant légèrement avec les récidive prédits (45%).
- 34% de la population ont fait l'objet d'une décision incorrecte, à peu près équilibrée entre les faux positifs et les faux négatifs.

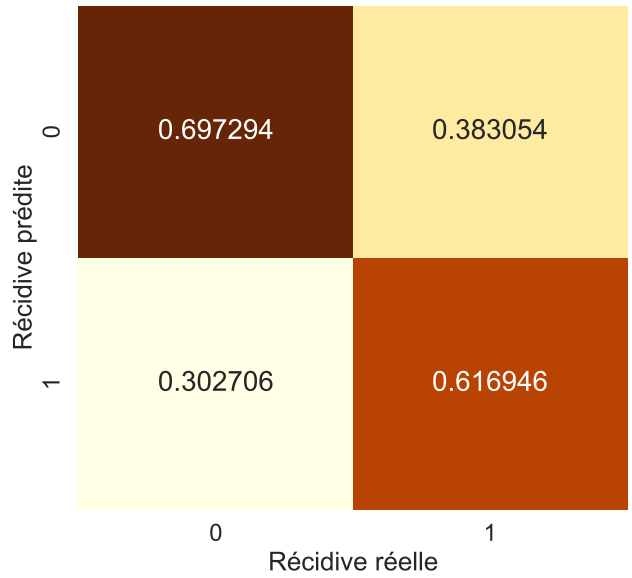
	Acc(%)	TFN	TFP
All	66.0	0.38	0.30





```
#Attention : il faut normaliser par colonne pour obtenir le tableau TFP
cm=pd.crosstab(df['COMPAS_Decision'], df['two_year_recid'],
               rownames=['Récidive prédite'],colnames=['Récidive réelle'])
cm1=pd.crosstab(df['COMPAS_Decision'], df['two_year_recid'],
                rownames=['Récidive prédite'],colnames=['Récidive réelle'],normalize='columns')

fig, axes = plt.subplots(1,2, figsize=(14,6))
(
sns.heatmap(cm,annot=True,fmt="d",cbar=False, ax=axes[0],cmap='YlOrBr', annot_kws={"size":18})
)

(
sns.heatmap(cm1,annot=True,fmt="f",cbar=False,ax=axes[1],cmap='YlOrBr', annot_kws={"size":18})
);
sns.set(font_scale=1.4)
#plt.savefig("FreqTable_all.pdf");
```

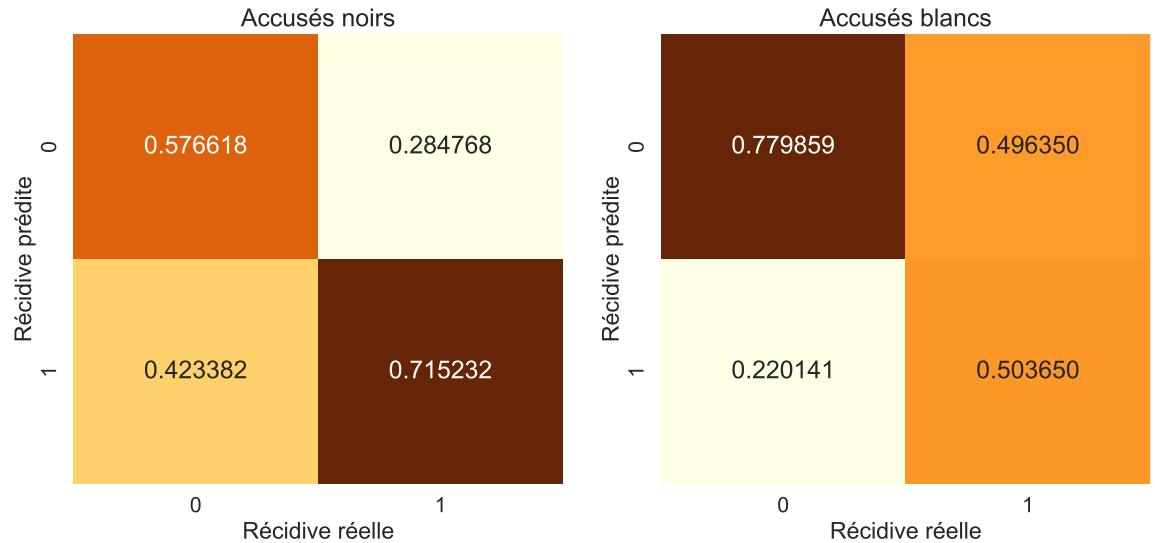




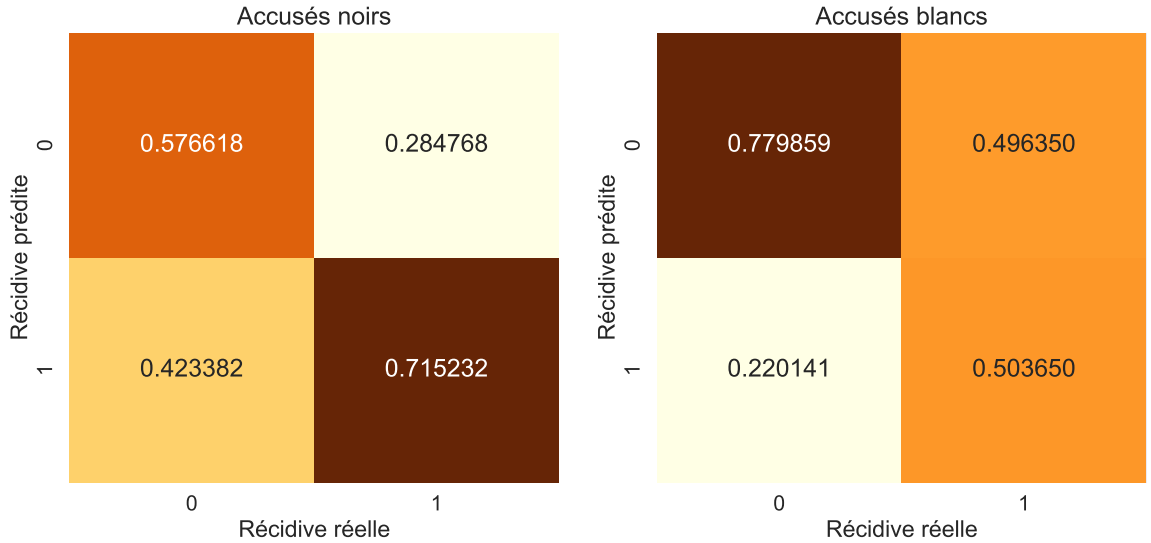
Si nous regardons les populations noires et blanches séparément:

- Une proportion plus grande d'accusés noirs souffrent d'une prédiction incorrecte de "récidive" que d'accusés blancs.
- Une proportion plus grande d'accusés blancs souffrent d'une prédiction incorrecte de "non-récidive" que d'accusés noirs.

	Acc(%)	TFN	TFP
Tous	66.0	0.38	0.30
Noir	64.9	0.28	0.42
Blancs	67.2	0.49	0.22



```
FT_black=pd.crosstab(b_recid['COMPAS_Decision'], b_recid['two_year_recid'],
                     rownames=['Récidive prédite'],colnames=['Récidive réelle'],normalize='columns')
FT_white=pd.crosstab(w_recid['COMPAS_Decision'], w_recid['two_year_recid'],
                     rownames=['Récidive prédite'],colnames=['Récidive réelle'],normalize='columns')

fig, axes = plt.subplots(1,2, figsize=(14,6))
axes[0].set_title('Accusés noirs',fontsize = 18)
axes[1].set_title('Accusés blancs',fontsize = 18)

(
sns.heatmap(FT_black,annot=True,fmt="f",cbar=False, ax=axes[0],cmap='YlOrBr', annot_kws={"size":18})
)

(
sns.heatmap(FT_white,annot=True,fmt="f",cbar=False,ax=axes[1],cmap='YlOrBr', annot_kws={"size":18})
);
sns.set(font_scale=1.4)
```



Quelques lectures intéressantes

- https://www.propublica.org/
- https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
- https://github.com/propublica/compas-analysis/
- https://www.foia.gov/
- https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html