Introduction aux principes du traitement de l'information

Benoit Gaüzère, Stéphane Canu
benoit.gauzere@insa-rouen.fr

INSA Rouen Normandie - ITI

19 janvier 2026

Présentation de l'EC

Enseignants

- Benoit Gaüzère (moi)
  - CM et TD
  - benoit.gauzere@insa-rouen.fr
- Gilles Gasso : 2 TD
- Florian Yger : 1 TD

Organisation

- 14 semaines CM + TD
- Toutes les ressources : Cours Moodle
- Une question ? : Forum Moodle

Présentation de l'EC

Notes

- **I.S.** Examen médian sur machine : 30 %
- **D.S.** Examen final sur machine : 40 %
- **Projet** Jeu de données à analyser : 30 %

Conseils

- Du travail régulier
- Ne laissez rien incompris
- N'ayez pas peur ni des maths, ni de Python
- Posez des questions !

Introduction aux bases de la Science des Données

- **Traiter et structurer des données**
  - Données tabulaires, textuelles, temporelles
  - Données bruitées, incomplètes, hétérogènes
- **Décrire et comprendre les données**
  - Résumer un jeu de données
  - Visualiser pour détecter structures et anomalies
  - Identifier des relations entre variables
- **Modéliser et prédire**
  - Expliquer une variable par d'autres
  - Comparer des groupes
  - Réduire la dimension et synthétiser l'information
- **Décider en présence d l'incertitude**
  - Séparer signal et bruit
  - Quantifier l'incertitude
  - Évaluer la fiabilité d'une conclusion

M8 = Statistiques + Informatique

Les outils de M8

Python

- Utilisation en script
- Nombreuses bibliothèques
- Numpy, Matplotlib, ...
- Implémentation des concepts vus en M8

Jupyter notebook

- Interface pour implémenter les outils en python
- Permet de prototyper
- Rédaction de rapports

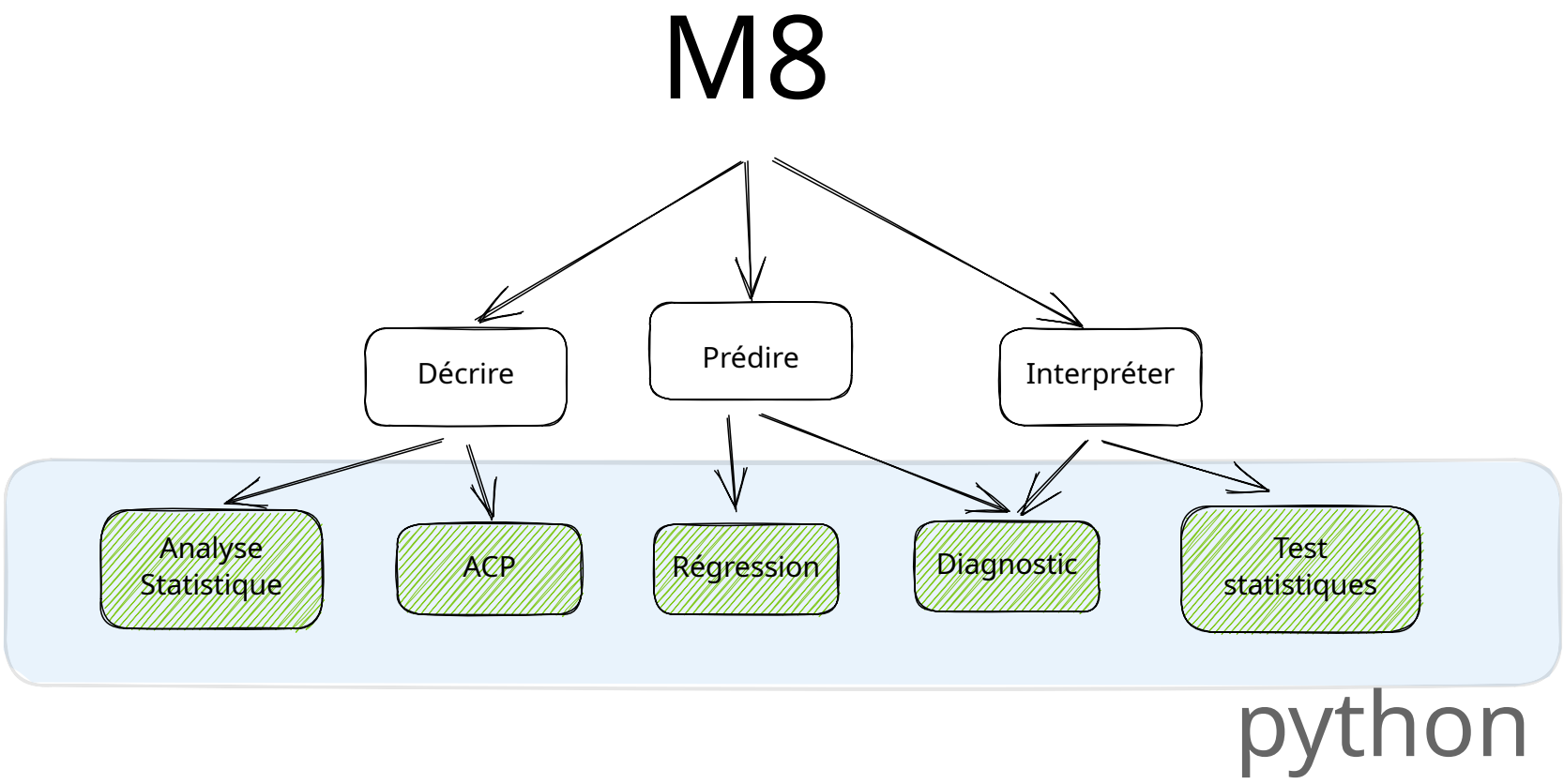
Les questions d'une étude de *data science*

- Descriptive :
  - → quels sont les principaux groupes ?
  - → nettoyer les données (variables/individus) ?
- Expérimentale : établir un lien de corrélation
  - → émission de CO_2 et la température
- Comparative/prédictive : y a-t-il une différence entre deux groupes ?
  - → ce médicament est-il efficace ?

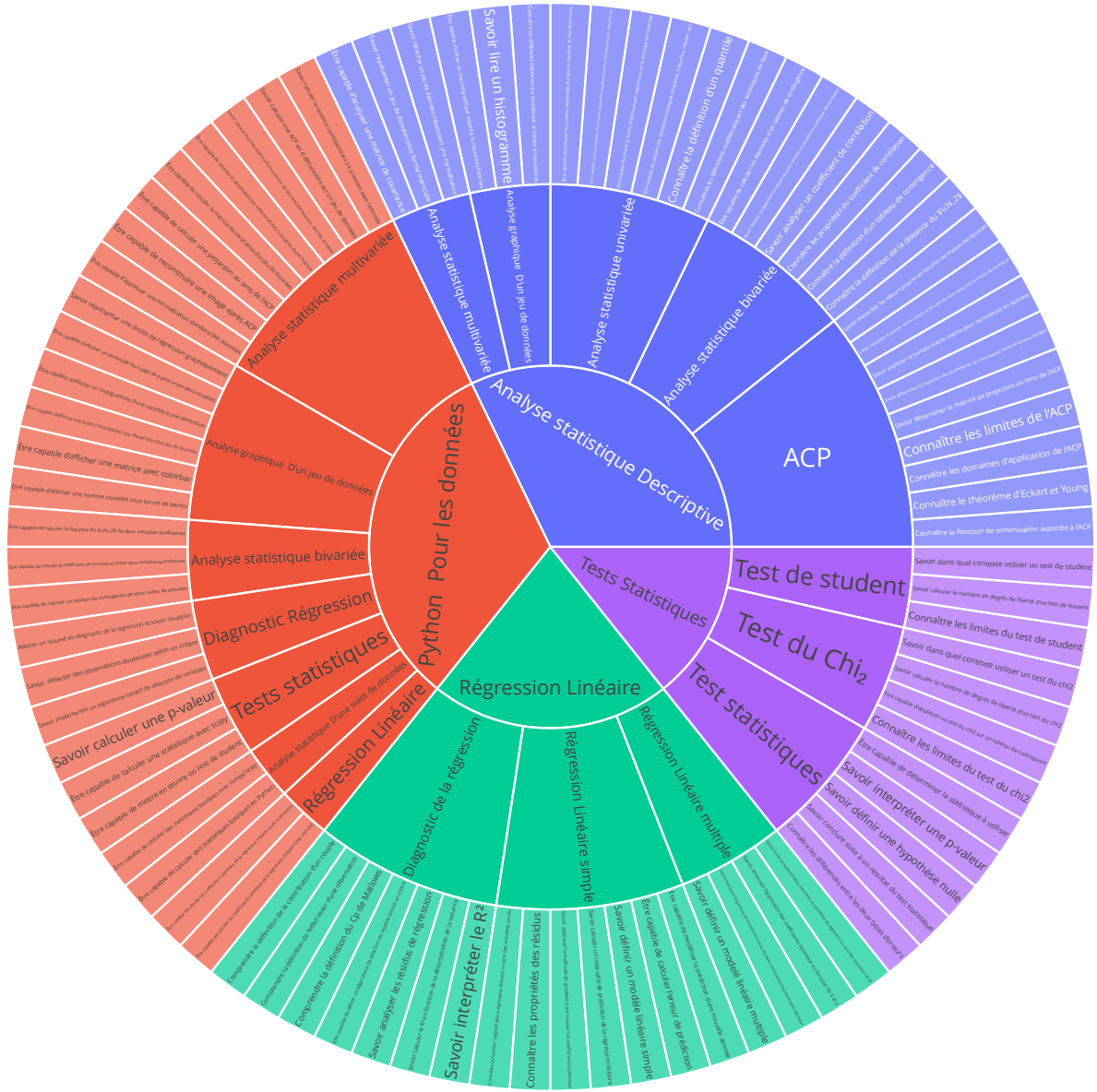
Les données permettent de répondre à une question

1. Poser la question
2. Récupérer les données / les nettoyer (pré traitement)
3. Modéliser poser un modèle et des hypothèses (statistiques)
4. Interpréter les données (inférence statistique et vérif. des hypothèses)

Plan de l'EC


M8
Décrire
Prédire
Interpréter
Analyse Statistique
ACP
Régression
Diagnostic
Test statistiques
python

Compétences M8



Où se posent les questions éthiques ?

- **Avant l'analyse**
  - Quelles données sont collectées ?
  - Qui est inclus / exclu du jeu de données ?
  - Comment les données ont-elles été obtenues ?
- **Pendant l'analyse**
  - Quels choix de modèles et de critères ?
  - Quels biais sont introduits par la méthode ?
  - Que mesure-t-on réellement ?
- **Après l'analyse**
  - Comment les résultats sont-ils interprétés ?
  - À quelles décisions mènent-ils ?
  - Pour qui, et avec quelles conséquences ?

Éthique des données

- **Confidentialité et anonymisation**
  - Données personnelles, sensibles, indirectement identifiantes
  - Anonymisation $\neq$ impossibilité de ré-identification
- **Méthode de collecte**
  - Données observées vs données provoquées
  - Consentement, contexte de production des données
- **Biais de données**
  - Échantillon non représentatif
  - Variables manquantes ou proxies
  - Biais historiques reproduits par les modèles

Point clé

Les modèles apprennent ce que les données contiennent, pas ce qui est juste ou souhaitable.

Éthique des méthodes et des usages

- **Éthique de la méthode**
  - Hypothèses du modèle explicites ou implicites
  - Choix des critères d'optimisation
  - Limites de validité des résultats
- **Éthique de l'usage**
  - Quelle question est réellement posée ?
  - Décision automatisée ou aide à la décision ?
  - Qui porte la responsabilité finale ?

Message clé

Un résultat statistique n'est jamais une décision, mais un élément parmi d'autres.

Cas concret : biais algorithmiques dans la justice pénale

- **Contexte**
  - Algorithme *COMPAS* utilisé aux États-Unis
  - Objectif : prédire le risque de récidive
- **Problème observé**
  - Taux de faux positifs plus élevé pour certaines populations
  - Même score $\Rightarrow$ décisions différentes selon les groupes
- **Lien avec M8**
  - Biais dans les données d'apprentissage
  - Choix du critère d'optimisation
  - Corrélation $\neq$ causalité

Références

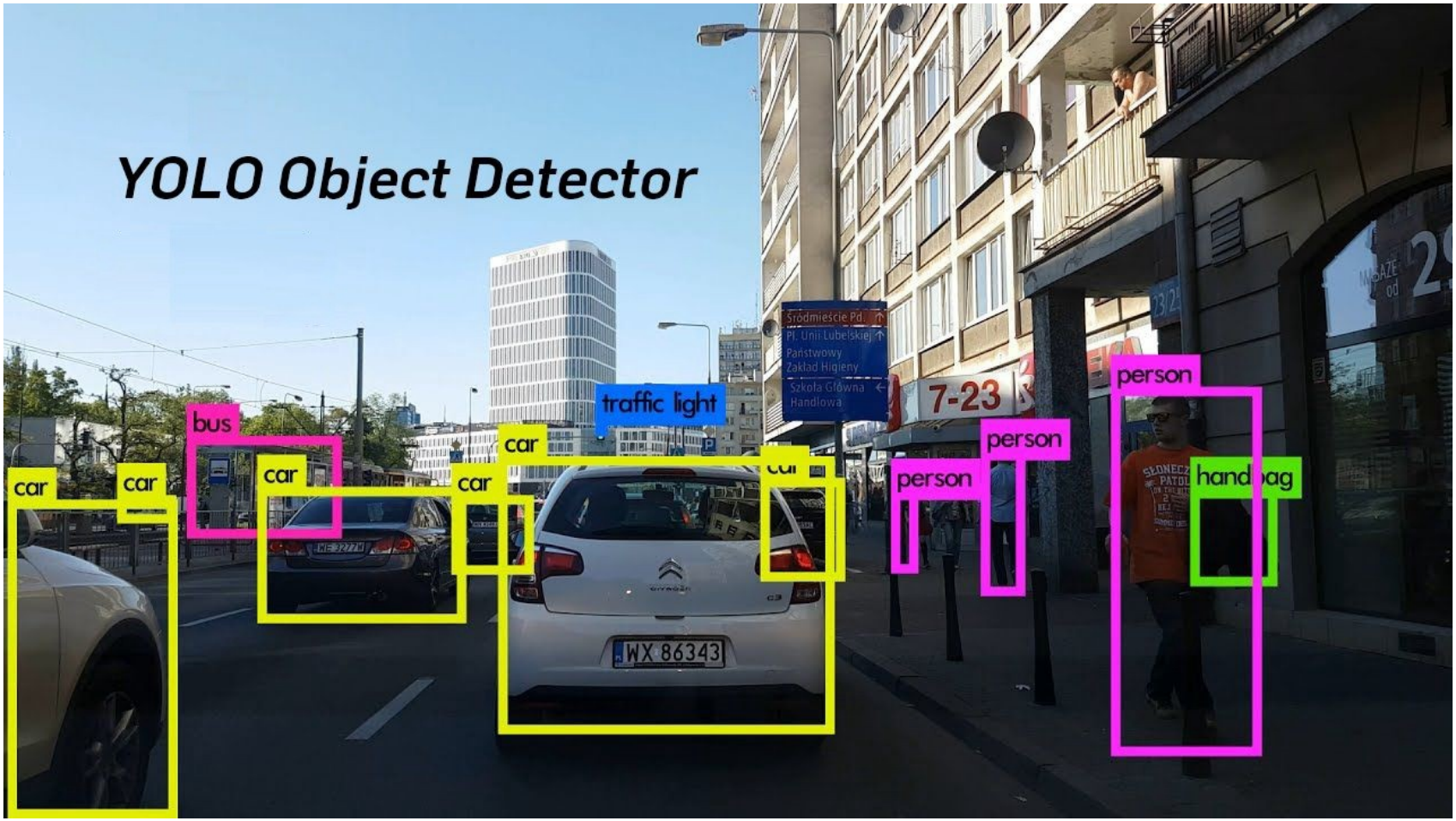
- Article sur COMPAS
- Article du Monde sur COMPAS

Message clé

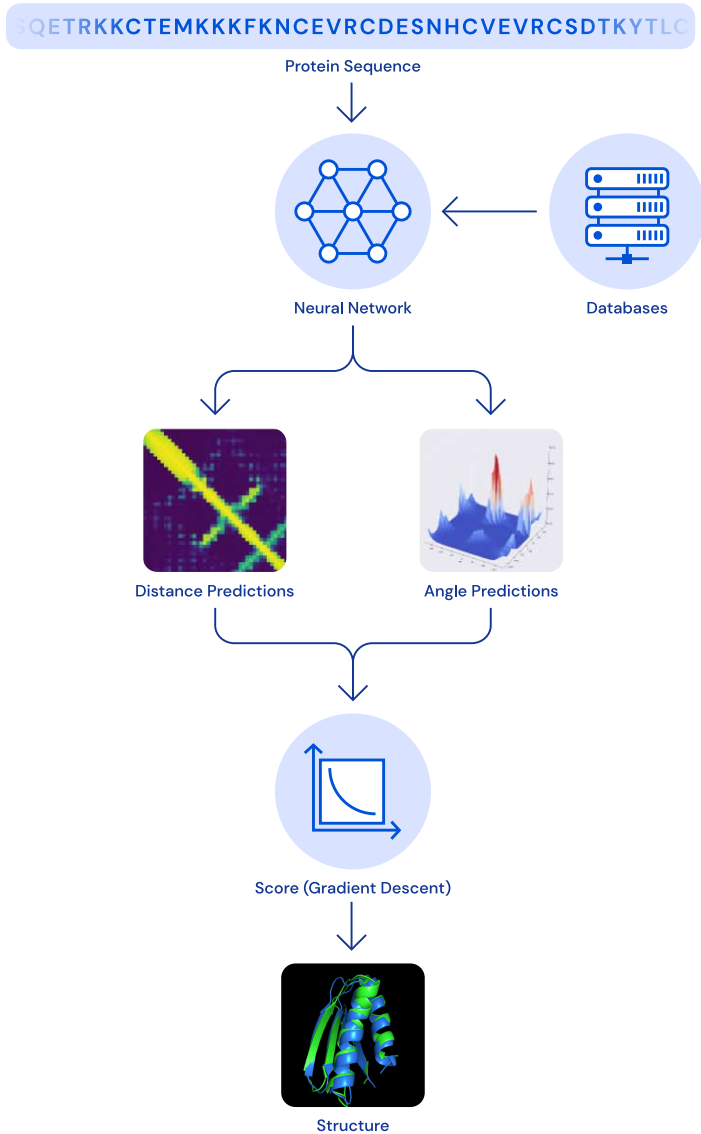
- Les données ne sont jamais neutres
- Les modèles optimisent un critère, pas la justice ni la vérité
- Une bonne corrélation ne garantit pas une bonne décision

La responsabilité reste humaine

Cas concrets d'utilisation de la science des données



https://blog.paperspace.com/how-to-implement-a-yolo-object-detector-in-pytorch/



https://deepmind.com/blog/article/AlphaFold-Using-AI-for-scientific-discovery

GPT-4

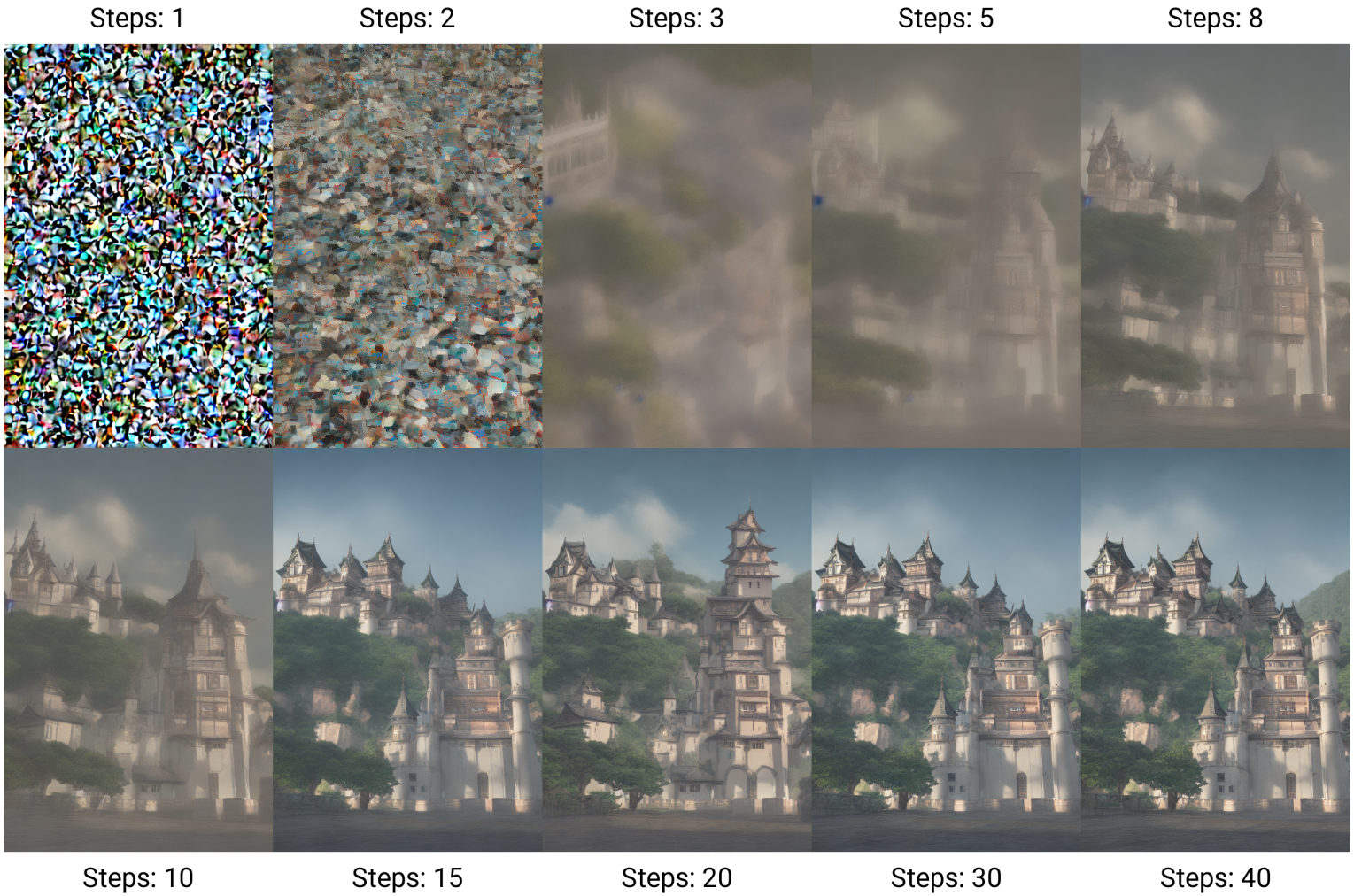
You
User
Donne moi un plan de cours sur 14 semaines pour l'introduction aux statistiques en deuxième année d'école d'ingénieur en une dizaines de lignes

ChatGPT
Bien sûr, voici un exemple de plan de cours pour une introduction aux statistiques sur 14 semaines, destiné à des étudiants en deuxième année d'école d'ingénieur :

1. Semaine 1 : Introduction aux statistiques - Concepts de base, types de données, et importance des statistiques dans l'ingénierie.
2. Semaine 2 : Méthodes de collecte de données - Techniques d'échantillonnage, études observationnelles vs expérimentales.
3. Semaine 3-4 : Probabilités - Règles de base, variables aléatoires, distributions de probabilité (binomiale, normale).
4. Semaine 5-6 : Statistique descriptive - Mesures de tendance centrale, de dispersion, graphiques et visualisation des données.
5. Semaine 7 : Tests d'hypothèses (partie 1) - Concepts de base, erreur de type I et II, test de significativité.
6. Semaine 8 : Tests d'hypothèses (partie 2) - Tests pour moyennes et proportions, puissance des tests.
7. Semaine 9 : Intervalles de confiance - Construction et interprétation pour moyennes et proportions.
8. Semaine 10 : Analyse de variance (ANOVA) - Comparaison de plusieurs moyennes, hypothèses et applications.
9. Semaine 11 : Régression linéaire - Modélisation, estimation des paramètres, tests et prédictions.
10. Semaine 12 : Régression multiple et corrélation - Extension de la régression linéaire, analyse de la corrélation.
11. Semaine 13 : Techniques non paramétriques - Tests de signe, test de rangs, autres méthodes.
12. Semaine 14 : Récapitulatif du cours et préparation pour l'examen final - Revue des concepts clés, séance de questions-réponses.

Chaque semaine pourrait inclure des lectures assignées, des devoirs pratiques, des

Modèles de diffusion



Plan

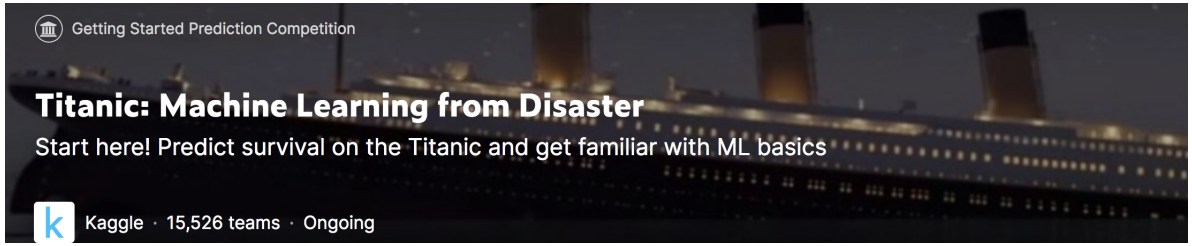
L'exemple des données bancaires

Nom de la variable	Exemple	Aléatoire	Discrète	Type
Matricule	242 335 AD	Non	-	-
nom	Peutu	Non	-	-
prénom	Stefen	Non	-	-
sexe	M	Oui	discrète	qualitative
age	38	Oui	continue	quantitative
Département	76	Oui	discrète	qualitative
PCS	Employé	Oui	discrète	qualitative
total des avoirs	11 240	Oui	continue	quantitative
max des entrées	1 570	Oui	continue	quantitative
taux d'endettement	31 %	Oui	continue	quantitative
nombre de visites	2	Oui	discrète	quantitative

Trois questions fondamentales

- Variable qualitative ou quantitative ?
- Discrète ou continue ?
- Variable aléatoire ou non (déterministe) ?

L'exemple du Titanic



- Survived : Outcome of survival (0 = No ; 1 = Yes)
- Pclass : Socio-economic class (1 = Upper ; 2 = Middle ; 3 = Lower)
- Name : Name of passenger
- Sex : Sex of the passenger
- Age : Age of the passenger (Some entries contain NaN)
- SibSp : Number of siblings and spouses of the passenger aboard
- Parch : Number of parents and children of the passenger aboard
- Ticket : Ticket number of the passenger
- Fare : Fare paid by the passenger
- Cabin Cabin number of the passenger (Some entries contain NaN)
- Embarked : Port of embarkation of the passenger (C = Cherbourg ; Q = Queenstown ; S = Southampton)

https://www.kaggle.com/c/titanic

Les trois distinctions I

Définition : Domaine d'une variable

Le **domaine d'une variable** est l'ensemble des valeurs que cette variable peut prendre.
Exemple : $\Omega_{sexe} = \{M, F\}$

Définition : Variable discrète

une **variable est discrète** si le cardinal de son domaine est dénombrable.
Exemples : sexe, département, PCS, nombre de visites...

Les trois distinctions II

Définition : Variable quantitative

une variable discrète est **quantitative** si l'ensemble de ses modalités est comparable[a]. Toutes les variables continues sont quantitatives.
Exemples : age, nombre de visites, ...

a. $\forall x, y$ deux modalités quelconques, soit $x < y$ soit $x \geq y$

Transformation des variables

- continue $\rightarrow$ continue (log)
- continue $\rightarrow$ discrète (quantification)
- discrète $\rightarrow$ continue (calcul d'une moyenne sur un age par exemple)
- continue $\rightarrow$ discrète (ordre,rang)

observations	2	-5,5	1	3,4
rang	3	1	2	4

Table – Exemple de transformation de rang.

Plan

Variables qualitatives

Variables qualitatives	**Variables quantitatives**
original ou copie (binaire)	date d'exécution
auteur : Vermeer	taille, poids, surface peinte
type de scène (multimodale)	prix

Variables qualitatives

Variable Qualitative

- Une variable qualitative peut être observée mais non mesurée
- $\neq$ quantitative
- Encodage par une énumération, numérique
- Aussi appelée variable catégorielle

Exemple

- Sexe
- Catégorie socio-professionnelle
- Code postal
- ...

Définition : Modalités

On appelle **modalités** l'ensemble des valeurs distinctes que peut prendre la variable observée. C'est le domaine d'une variable discrète. On note :

$$\Omega = \{m_1, m_2, ..., m_i, ..., m_r\}$$

avec $r = card(\Omega)$.

Exemples de modalités :

- oui / non (variable binaire)
- un ensemble de variétés de maïs (variable qualitative)
- un peu / moyen / beaucoup (qualitative ordinale)
- nombre des personnes à la caisse d'un supermarché (quantitative)
- nombre de mois de chômage entre deux emplois (quantitative)

Définition : Effectif

On appelle **effectif** d'une modalité m_i le nombre de fois n_i où cette modalité est apparue dans l'échantillon.

Définition : Fréquence

On appelle **fréquence** d'une modalité le rapport du nombre de fois où cette modalité est apparue dans un échantillon (n_i) divisé par la taille de l'échantillon (n). Pour la modalité m_i, la fréquence $f_i = \frac{n_i}{n}$.

Définition : Probabilité

La **probabilité** $\mathbb{P}(m_i)$ d'une modalité m_i est la limite de la fréquence observée lorsque la taille de l'échantillon tend vers l'infini [a].

---

a. Rigoureusement, il faudrait définir les probabilités à priori et considérer que l'on observe un échantillon d'une variable aléatoire tirée selon cette loi de manière indépendante et identiquement distribué (pour plus de détails voir par exemple `http://fr.wikipedia.org/wiki/Probabilité`).

Variables quantitatives discrètes

Exemples :

- Nombre de produits achetés par un client
- L'appréciation donnée par un spectateur à un film (1 à 5 étoiles)
- Intervalle (transformation d'une variable continue)
- Nombre des personnes à la caisse d'un supermarché (théoriquement infinie)

Fréquence cumulée

Définition : Effectif cumulé

On appelle **effectif cumulé** d'une modalité : $N_i = \sum_{j=1}^{j \leq i} n_j$

Définition : Fréquence cumulée

On appelle **fréquence cumulée** : $\widehat{F}_i = \sum_{j=1}^{i} f_j$

- On a $N_r = n$ et $\widehat{F}_r = 1$.
- Les $\widehat{F}_i$ définissent la loi cumulative empirique $\widehat{F}_i = \widehat{F}(m_i) = \hat{\mathbb{P}}(X \leq m_i)$
- On appelle $h_i = m_{i+1} - m_i$ la distance inter modalités.
- Ces quantités ont un sens lorsque la variable est quantitative. modalités ordonnées.

Variables quantitatives discrètes

Modalité (m_i)	effectifs n_i	N_i	$\widehat{f}_i = \frac{n_i}{n}$	$100 \times \widehat{F}_i$ (%)
24	1	1	$\frac{1}{38}$	2,63 %
26	2	3	$\frac{2}{38}$	7,89 %
29	3	6	$\frac{3}{38}$	15,79 %
31	2	8	$\frac{2}{38}$	21,05 %
33	4	12	$\frac{4}{38}$	31,58 %
35	3	15	$\frac{3}{38}$	39,47 %
37	3	18	$\frac{3}{38}$	47,37 %
38	1	19	$\frac{1}{38}$	50,5 %
41	6	25	$\frac{6}{38}$	69,79%
43	3	28	$\frac{3}{38}$	73,68 %
45	1	29	$\frac{1}{38}$	76,32 %
46	4	33	$\frac{4}{38}$	86,84 %
49	5	38	$\frac{5}{38}$	100 %
Total	38	-	1	-

Exemple

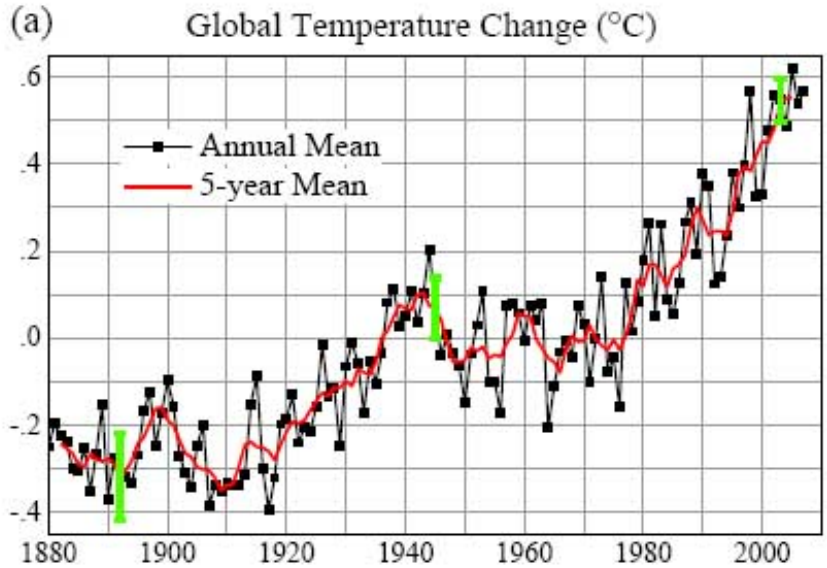
nom	**age**
louis	13
paul	35
joséphine	24
lucien	43
mike	56
pedro	27
...	...

$\Rightarrow$

m_i	n_i	f_i	N_i	F_i
13	381	$2,06$	381	2,05
14	576	$3,11$	957	5,17
15	441	$2,38$	1398	7,55
16	744	$4,02$	2142	11,57
17	553	$2,99$	2695	14,56
...	...	...	...	...
68	441	$2,38$	18509	100,00
Total	18509	100		

Dans le cas continu on peut toujours définir la fonction de répartition empirique $\widehat{F}_i = \widehat{F}(m_i) = \widehat{\mathbb{P}}(X \leq m_i)$

Variables quantitatives continues



Graphe des variations de températures annuelles du globe terrestre par rapport à la période de référence 1951-1980 (Air and ocean data from weather stations, ships and satellites) [1].

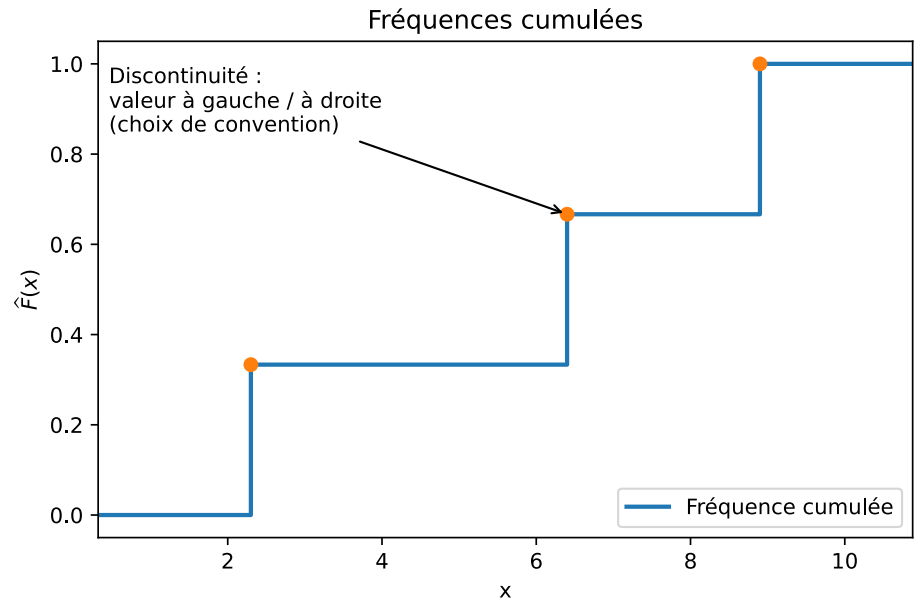
Exemples :

- Température ($\Omega = [-273.15, +\infty]$)
- Concentration, intensité, prix, poids, taille, distance, ...
- Rapport, proportion ($\Omega = [0, 1]$)

1. `http://www.nasa.gov/topics/earth/features/earth_temp.html`

Fonction de répartition empirique : le problème

x_i	f_i	F_i
2,3	$1/3$	$1/3$
6,4	$1/3$	$2/3$
8,9	$1/3$	1



- Échantillon fini ordonné
- $f_i = 1/n$
- $F_i = \sum_{j \leq i} f_j$

Difficulté

Fonction en escalier : la valeur de $\widehat{F}(x_i)$ dépend d'un choix de convention (gauche / droite).

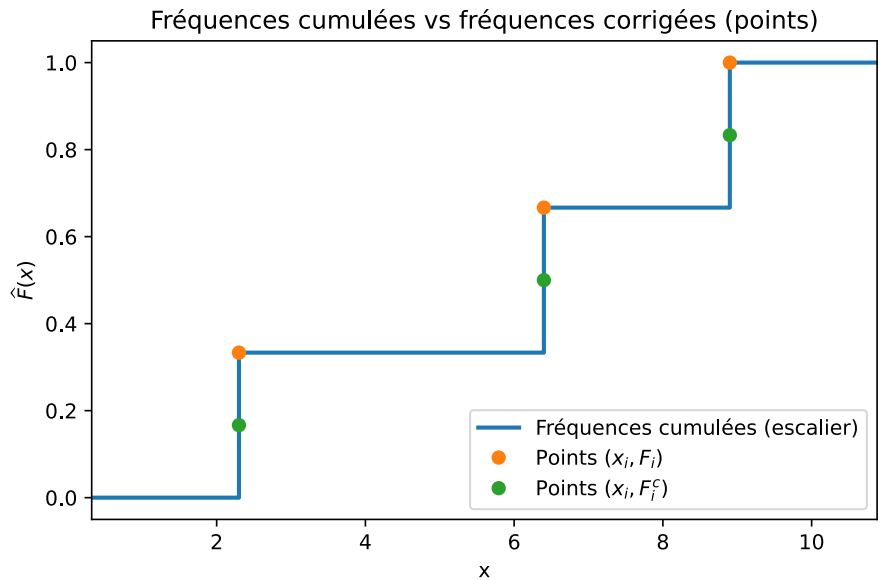
Lever l'ambiguïté : fréquences corrigées

Définition : Fréquences corrigées

$\widehat{F}_i^c = \widehat{F}_i - \frac{1}{2}(\widehat{F}_i - \widehat{F}_{i-1})$

$\rightarrow$ c'est le centre de l'intervalle $[\widehat{F}_{i-1}, \widehat{F}_i]$

x_i	f_i	F_i	F_i^c
2,3	1/3	1/3	1/6
6,4	1/3	2/3	1/2
8,9	1/3	1	5/6



Intuition

F_i^c place la masse de probabilité associée à x_i au centre de la marche (plutôt que sur un bord).

Fonction de répartition empirique

Définition : Fonction de répartition empirique

On définit $\widehat{F}_X$ comme l'interpolation linéaire entre les points (x_i, F_i^c) :

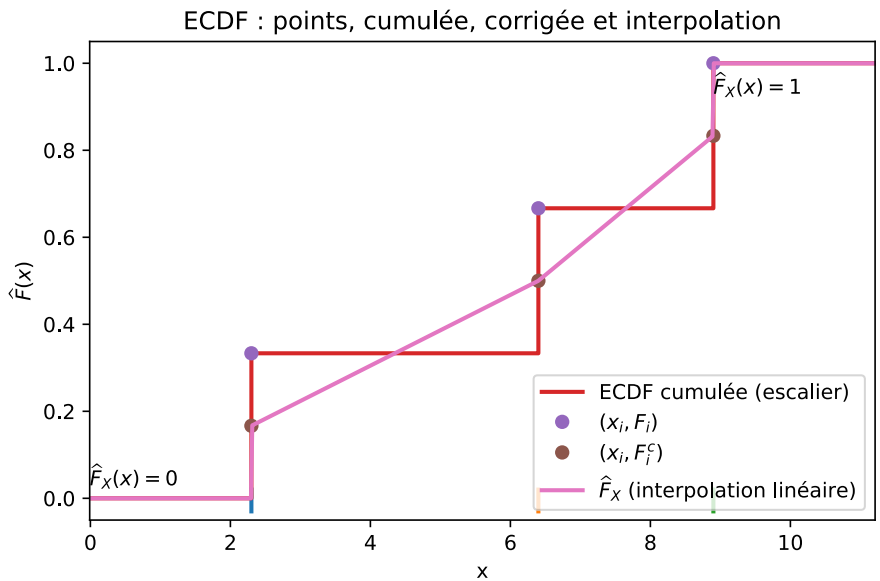
$$\widehat{F}_X(x) = \widehat{F}_{i-1}^c + \frac{\widehat{F}_i^c - \widehat{F}_{i-1}^c}{x_i - x_{i-1}}(x - x_{i-1}) \text{ si } x_{i-1} \leq x < x_i, i = 2, n$$

avec les conventions :

$$\widehat{F}_X(x) = 0 \text{ si } x < x_1 \qquad \text{et} \qquad \widehat{F}_X(x) = 1 \text{ si } x \geq x_n.$$

Fonction de répartition empirique : illustration

x_i	f_i	F_i	F_i^c
2,3	$1/3$	$1/3$	$1/6$
6,4	$1/3$	$2/3$	$1/2$
8,9	$1/3$	1	$5/6$



- Points originaux : (x_i)
- Points cumulés : (x_i, F_i)
- Points corrigés : (x_i, F_i^c)

Idée

La fonction empirique $\widehat{F}_X$ est l'interpolation linéaire entre les points (x_i, F_i^c) (avec $\widehat{F}_X(x) = 0$ avant x_1 et $\widehat{F}_X(x) = 1$ après x_n).

Fonctions de répartition : comparaison discret / continue I

modalité	**eff.**	$\widehat{f}$	**e. c.**	$\widehat{F}$
$]0,\ 18]$	1405	$7,59$	1405	7,59
$]18,\ 29]$	1875	$10,13$	3280	17,72
$]30,\ 39]$	1172	$6,33$	4452	24,05
$]40,\ 49]$	3047	$16,46$	7499	40,51
$]50,\ 59]$	4920	$26,58$	12419	67,09
$60 \leq$	4920	$32,91$	18510	100
total	18509	100		

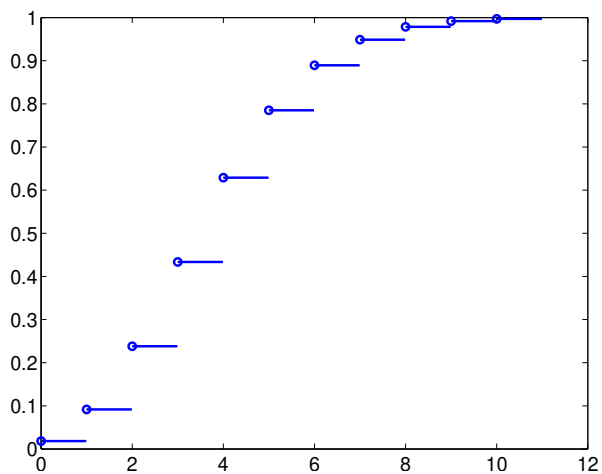
x_i	**eff.** n_i	N_i	$\widehat{f}_j = \frac{n_i}{n}$	$100 \times \widehat{F}_j$
24	1	1	$\frac{1}{6}$	16,67 %
26	1	2	$\frac{1}{6}$	33,33%
29	1	3	$\frac{1}{6}$	50 %
31	1	4	$\frac{1}{6}$	66,67%
33	1	5	$\frac{1}{6}$	83,33%
35	1	6	$\frac{1}{6}$	100 %
Total	6	-	1	-

Fonctions de répartition : comparaison discret / continue II

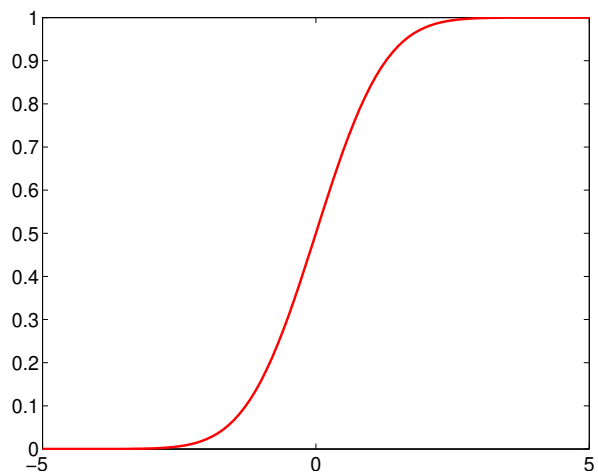
Définition : Fonction de répartition

La fonction de répartition de la variable aléatoire réelle X est la fonction F_X qui à tout réel x associe

$$F_X(x) = \mathbb{P}(X \leq x)$$



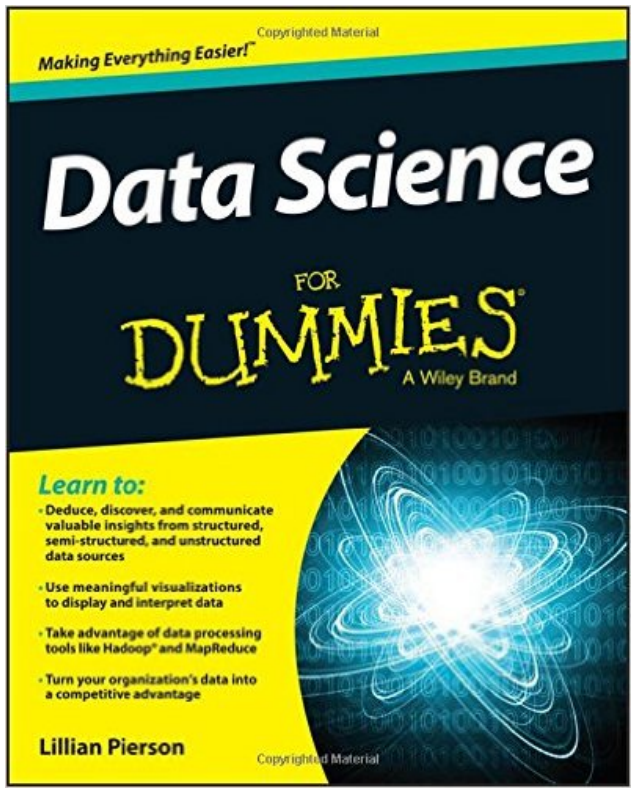
$F_X(x)$ discret



$F_X(x)$ continu

Il faut ordonner les observations !

Plan

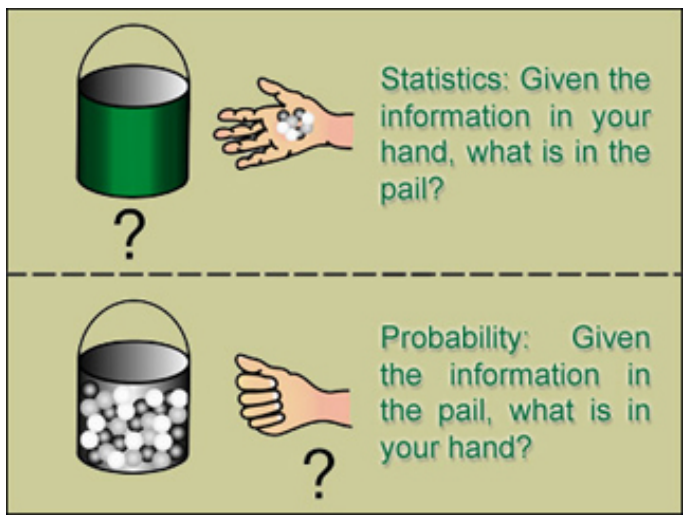


Conclusion

- Tableau de données :

  Ligne = une modalité (observation) / colonne = une variable

- Différents types de variables
  - Qualitatives : modalités / fréquence / probabilité
  - Quantitatives discrètes : effectif cumulé / fonction de répartition
  - Qualitative : plus de probabilités $\rightarrow$ une densité (continue)
- Objets empiriques / théoriques (parfois avec le même nom : probabilité)



References

Les livres de M8

- Statistics and Data Analysis from Elementary to Intermediate. Ajit C. Tamhane and Dorothy D. Dunlop, Prentice Hall, 2000.
- Statistical Inference. Casella, George, and Roger L. Berger. Belmont, CA : Duxbury Press, 1990.
- Statistique mathématique : applications commentées. Jean-Pierre Boulay, Ellipses, 2010