

Introduction aux principes du traitement de l'information

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

INSA Rouen Normandie - ITI

24 janvier 2025

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 !

Objectifs de M8

Introduction aux bases de la Science des Données

- ▶ Traiter des masses de données
- ▶ En extraire de l'information
 - ▶ le génome, la biologie et la médecine
 - ▶ l'astronomie et la physique
 - ▶ connaître ses clients : les tickets de caisse
 - ▶ Internet, par minute :
 - ▶ 480 000 Tweets
 - ▶ Instagram : 60 000 photos uploadées
 - ▶ 200 millions de mails
 - ▶ 4,2 millions de recherches Google
 - ▶ Youtube : 4,7 millions de vidéos vues
- ▶ Raisonner en présence d'incertitudes (de hasard)
 - ▶ Hasard vient de l'arabe al-zahr signifiant à l'origine "dés" mais aussi tous les domaines relevant de "la science de la Chance" (Averroès)

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

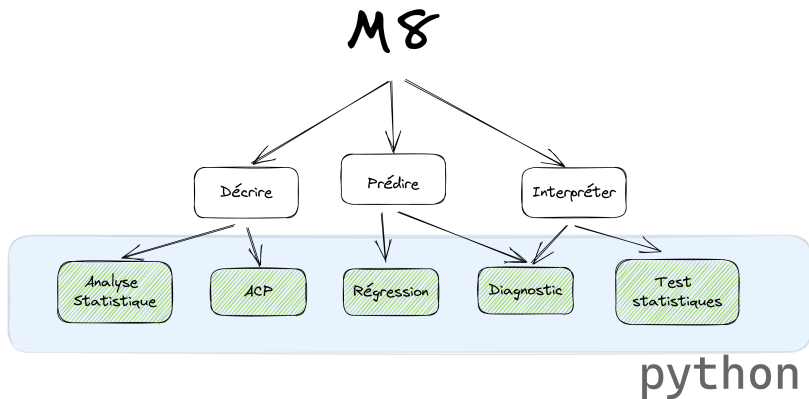
Les questions d'une étude de *data science*

- ▶ 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 ?

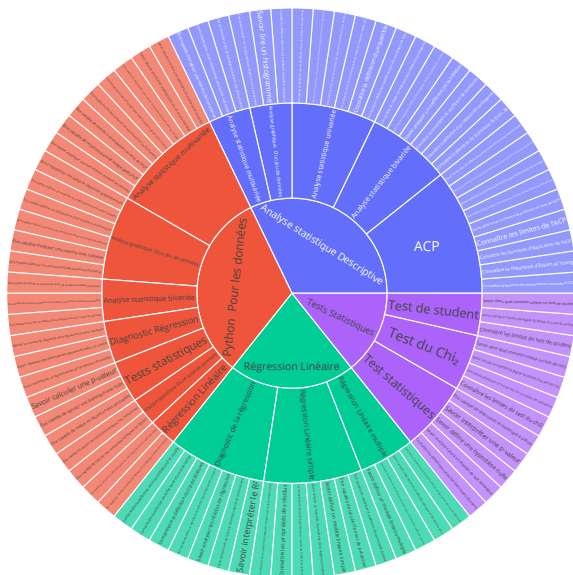
Les données permettent de répondre à une question

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

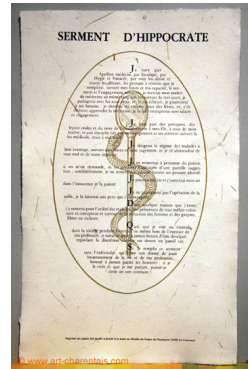


Compétences M8

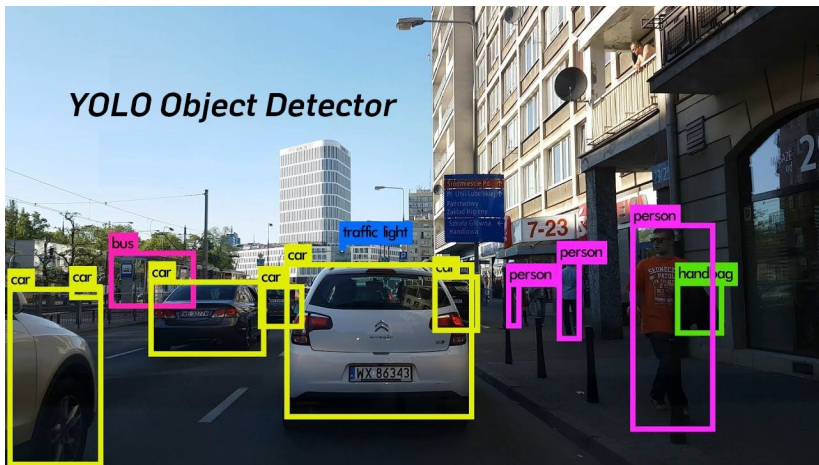


Questions éthiques

- ▶ L'éthique de la méthode
- ▶ L'éthique des données
 - ▶ confidentialité
 - ▶ méthode de collecte
 - ▶ biais
- ▶ L'éthique des usages : quelle est la question ?

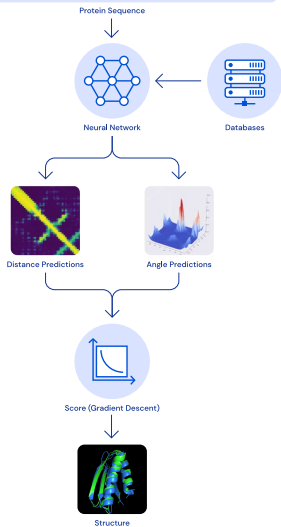


YOLO Object Detector



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

QETRRKKCTEMKKKFKNCEVRCDESNHCVEVRCSDTKYTLC



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



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 dizaine de lignes



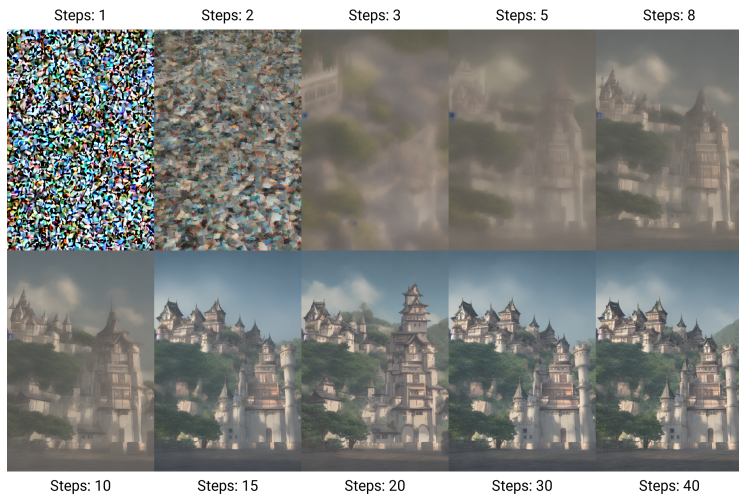
ChatGPT

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

Nature des variables statistiques

- L'exemple des données bancaires
- Les trois distinctions
- Transformation des variables

Description mono variable

- La démarche descriptive
 - Variable qualitative
 - Variables quantitatives discrètes
 - Variables quantitatives continues

Conclusion

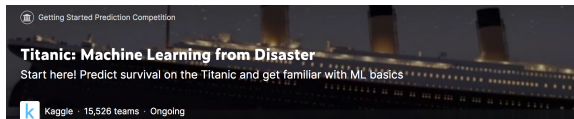
L'exemple des données bancaires

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)

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

Nature des variables statistiques

- L'exemple des données bancaires
- Les trois distinctions
- Transformation des variables

Description mono variable

- La démarche descriptive
 - Variable qualitative
 - Variables quantitatives discrètes
 - Variables quantitatives continues

Conclusion

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 :

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

avec $r = \text{card}(\Omega)$.

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** :
$$\hat{F}_i = \sum_{j, f_j \leq f_i} f_j$$

- ▶ On a $N_r = n$ et $\hat{F}_r = 1$.
- ▶ Les $\hat{F}_i$ définissent la loi cumulative empirique $\hat{F}_i = \hat{F}(m_i) = \hat{\mathbb{P}}(X \leq m_i)$
- ▶ 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

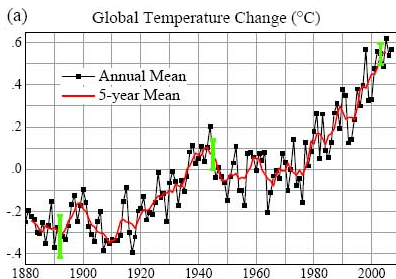
Modalité (m_i)	effectifs n_i	N_i	$\hat{f}_i = \frac{n_i}{n}$	$100 \times \hat{F}_i$ (%)
24	1	1	$\frac{1}{38}$	2,63 %
26	2	3	$\frac{2}{38}$	7,89 %
29	3	6	$\frac{3}{38}$	15,79 %
31	2	8	$\frac{2}{38}$	21,05 %
33	4	12	$\frac{4}{38}$	31,58 %
35	3	15	$\frac{3}{38}$	39,47 %
37	3	18	$\frac{3}{38}$	47,37 %
38	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		m_i	n_i	f_i	N_i	F_i
louis	13		13	381	2,06	381	2,05
paul	35		14	576	3,11	957	5,17
joséphine	24		15	441	2,38	1398	7,55
lucien	43	$\Rightarrow$	16	744	4,02	2142	11,57
mike	56		17	553	2,99	2695	14,56
pedro	27		...	...	...	...	...
...	...		68	441	2,38	18509	100,00
			Total	18509	100		

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

Variables quantitatives continues



Graphique 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

Définition : Fréquences corrigées

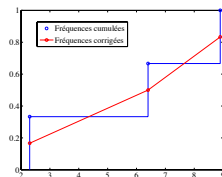
$$\widehat{F}_i^c = \widehat{F}_i - \frac{1}{2}(\widehat{F}_i - \widehat{F}_{i-1})$$

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

Exemple

► $n = 3$ observations : 6.4, 2.3 et 8.9

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



Fonction de répartition empirique

Définition : Fonction de répartition empirique

On appelle fonction de répartition empirique d'un échantillon pour la variable X la fonction $\hat{F}_X$ définie sur l'intervalle $[x_1, x_n]$ par :

$$\hat{F}_X(x) = \begin{cases} 0 & \text{si } x < x_1 \\ \hat{F}_{i-1}^c + \frac{\hat{F}_i^c - \hat{F}_{i-1}^c}{x_i - x_{i-1}}(x - x_{i-1}) & \text{si } x_{i-1} \leq x < x_i, \quad i = 2, n \\ 1 & \text{si } x \geq x_n \end{cases}$$

Exemples de fonctions de répartition empirique

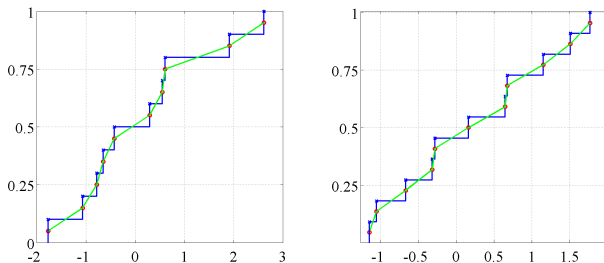


Figure – Fréquences cumulées (en bleue) et de fonction de répartition empirique (courbe verte) pour 11 observations (à droite).

Fonctions de répartition : comparaison discret / continue I

modalité	eff.	$\hat{f}$	e. c.	$\hat{F}$
]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		

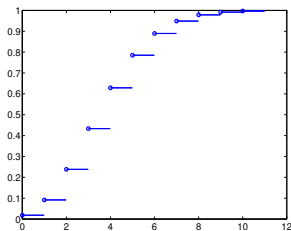
x_i	eff. n_i	N_i	$\hat{f}_j = \frac{n_i}{n}$	$100 \times \hat{F}_j$
24	1	1	$\frac{1}{6}$	16,67 %
26	1	2	$\frac{1}{6}$	33,33%
29	1	3	$\frac{1}{6}$	50 %
31	1	4	$\frac{1}{6}$	66,67%
33	1	5	$\frac{1}{6}$	83,33%
35	1	6	$\frac{1}{6}$	100 %
Total	6	-	1	-

Fonctions de répartition : comparaison discret / continu II

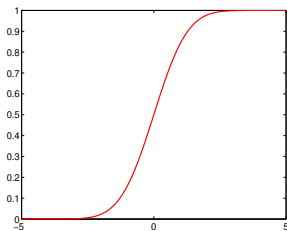
Définition : Fonction de répartition

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

Nature des variables statistiques

L'exemple des données bancaires

Les trois distinctions

Transformation des variables

Description mono variable

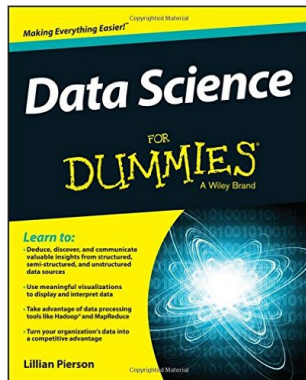
La démarche descriptive

Variable qualitative

Variables quantitatives discrètes

Variables quantitatives continues

Conclusion



Conclusion

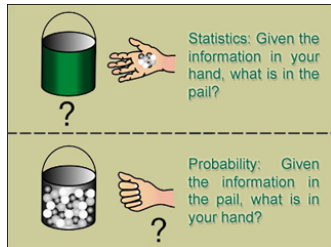
- ▶ **Tableau** de données :

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 → 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