Probabilités Discrètes

Cours n°1

EPITA 2025-2026

Introduction

Ce cours traite de probabilité, toutefois la première séance de ce cours portera sur l'étude de la statistique. Nous pouvons donc commencer par définir ce que sont les statistiques et les probabilités, et distinguer ces deux notions.

Lorsque l'on fait des **statistiques**, on commence par observer un phénomène, on relève différentes données le concernant, puis on étudie les données récoltées. On s'intéresse donc à un ou plusieurs évènements passés.
On appelle également statistique, ou **indicateur statistique**, une donnée issue des données observées et qui nous donne une information qualitative ou quantitave sur le phénomène que l'on étudie.

Exemple : Une moyenne de classe est une statistique qui représente le niveau global d'un groupe d'étudiant.

Lorsque l'on fait des **probabilités**, on va chercher à prévoir un phénomène, c'est-à-dire à envisager tous les évènements qui pourraient se produire dans une expérience et à déterminer si ces évènements ont de grandes chances ou non de se produire. Autrement dit on s'intéresse ici au futur.
Une **probabilité** est donc un nombre qui représente les chances qu'un évènement a de se produire dans un contexte prédéfini.

Exemple : Lorsqu'on en tire une pièce équilibrée à pile ou face, on peut prédire à l'avance que l'on a une chance sur deux de tomber sur pile. On dit que l'évènement "Tomber sur pile" a une probabilité égale à un demi.

1 Statistique

Comme on l'a vu en introduction, la statistique repose sur l'observation et l'analyse de données. Pour réaliser des observations et des analyses fiables, il existe des méthodes rigoureuses. Ces dernières permettent de définir des indicateurs statistiques et des outils de représentation de données, utilisables quelque soit l'expérience observée.

Définition : En statistique, un ensemble de données de même type est appelé **échantillon**.

Notation : Un échantillon s'écrit sous la forme {valeur_1, valeur_2,...}.

Exemple : L'ensemble $\{4, 11, 11, 13, 14, 14, 14, 19\}$ des notes obtenues par des étudiants à un contrôle est un échantillon.

Contre-exemple : L'ensemble {Zinédine, Zidane, 31 buts} n'est pas un échantillon : les données contenues dans cet ensemble ne sont pas de même type.

1.1 Indicateurs statistiques classiques

1.1.1 Quantiles

Définition : Les **quantiles** sont les valeurs qui divisent un échantillon en sous-ensembles de même taille.

Remarque : Pour effectuer cette division, on commence toujours par ordonner les valeurs de l'échantillon.

Exemple : Étudions l'échantillon $\{4, 2, 4, 5, 6, 4\}$. On souhaite le diviser en sous-ensemble de taille 2.
On commence par ordonner ses valeurs, par exemple par ordre croissant : $\{2, 4, 4, 4, 5, 6\}$.
On peut alors déterminer les quantiles de cet exemple : si on regarde les 2 premières de l'ensemble ordonné, la plus grande est 4, donc le premier quantile est égal à 4. On regarde ensuite les 2 valeurs suivantes de notre échantillon, on obtient que le second quantile est aussi égal à 4. Enfin, si on regarde de nouveau les 2 valeurs suivantes, on couvre tout notre échantillon, on peut donc s'arrêter.
On a ici deux quantiles, tout deux égaux à 4.

Remarque :
Si on divise notre échantillon en deux sous-ensembles, le quantile est appelé **médiane**.
Si on divise notre échantillon en deux sous-ensembles, les quantiles sont appelés **quartiles**.
De manière générale, si on divise notre échantillon en k sous-ensembles, les quantiles sont appelés **k-quantiles**.

Remarque : Un quantile est toujours une valeur présente dans l'échantillon, et si on étudie les k-quantiles, il y en a toujours $k - 1$.

1.1.2 Moyenne

Définition : Soit e un échantillon $\{v_1, ..., v_n\}$ de n nombres réels. On définit la **moyenne** m de e par la formule :

$$m = \frac{v_1 + ... + v_n}{n} \quad .$$

On dit que la moyenne est un indicateur statistique à tendance central : il représente le "centre" des valeurs de l'échantillon.

Remarque : La moyenne n'est pas forcément une valeur de l'échantillon. Par exemple la moyenne de 10 et 20 est égale à 15.

1.1.3 Variance et écart-type

Définition : Soient e un échantillon $\{v_1, ..., v_n\}$ de n nombres réels, et m la moyenne de e. On définit la **variance** v de e par la formule :

$$v = \frac{(v_1 - m)^2 + ... + (v_n - m)^2}{n} \quad .$$

La variance sert à représenter la "dispersion" des valeurs de l'échantillon, c'est-à-dire à savoir si les valeurs de l'échantillon sont éloignées ou proches de la moyenne de l'échantillon.

Définition : Soient e un échantillon de nombres réels, et v la variance de e. On définit l'**écart-type** σ de e par la formule :

$$\sigma = |\sqrt{v}| \quad .$$

Comme pour la variance, l'écart-type sert à représenter la "dispersion" des valeurs de l'échantillon.

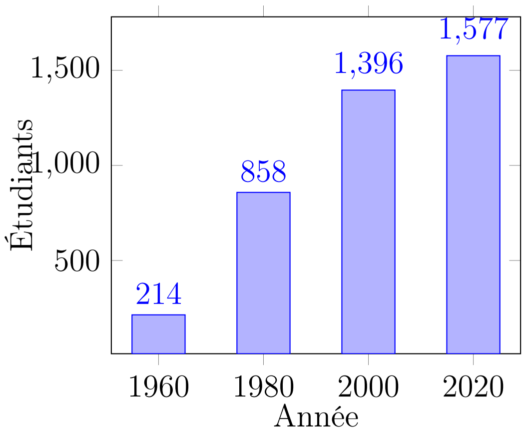
1.2 Représentations de données statistiques

1.2.1 Histogramme

L'histogramme, ou diagramme en bâton, permet de représenter un échantillon statistique par des rectangles, chacun correspondant à un sous-ensemble de l'échantillon. Plus le sous-ensemble contiendra de valeurs, plus l'aire du rectangle associé sera grande.

Exemple :

Nombre en milliers d'étudiants en université en France selon l'année.



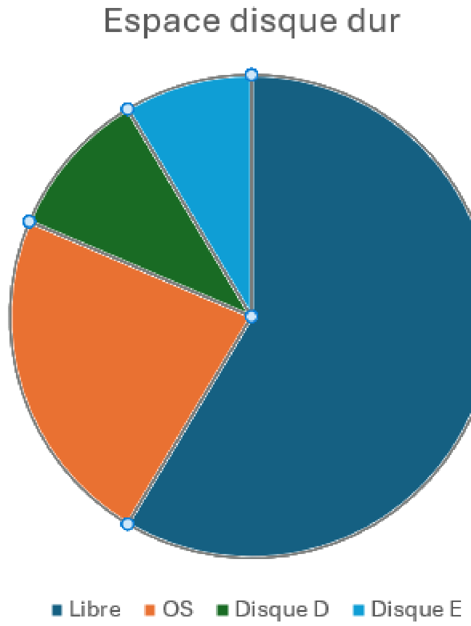
Dans un histogramme on retrouve :

- Le nombre de sous-ensembles par lequel on a divisé l'échantillon : c'est le nombre de rectangle.
- Le nombre de valeur compris dans chaque sous-ensemble : c'est la hauteur du rectangle correspondant, on la lit sur l'échelle à gauche.
- Le type des valeurs compris dans chaque sous-ensemble : on le lit dans la légende à gauche.
- La définition du sous-ensemble : on la lit sous le rectangle associé, et dans la légende en bas.

Remarque : Un histogramme peut également être représenter horizontalement.

1.2.2 Diagramme circulaire

Le diagramme circulaire permet de représenter un échantillon statistique en découpant un disque en plusieurs secteurs, chacun correspondant à un sous-ensemble de l'échantillon. Plus le sous-ensemble contiendra de valeurs, plus l'aire du secteur associé sera grande.



On peut observer dans l'exemple ci-dessus la répartition de la mémoire d'un disque dur.

Dans un diagramme circulaire on retrouve :

- Un disque découper en secteurs distincts.
- Une légende détaillant les informations contenues dans chaque secteurs.

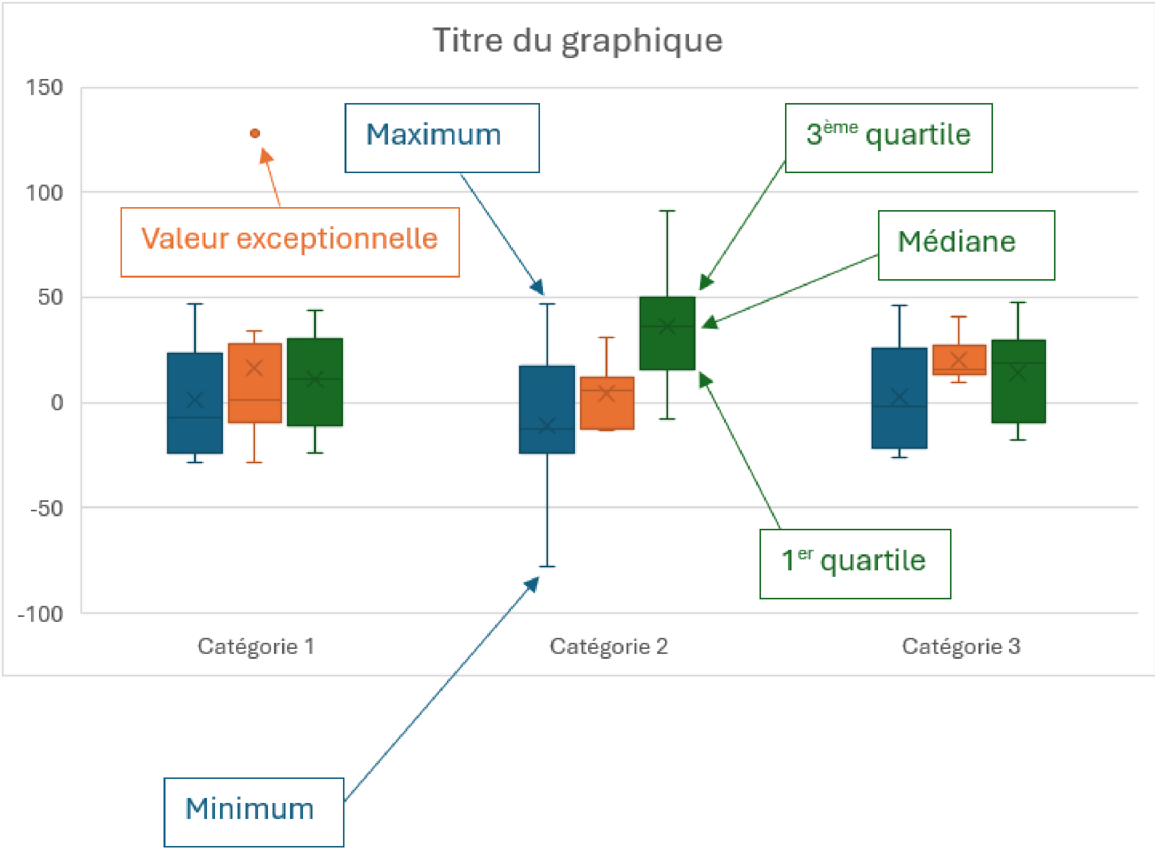
1.2.3 Boite à moustache

Définition : On appelle **valeurs exceptionelles**, ou **valeurs aberrantes**, d'un échantillon statistique les valeurs qui s'éloignent trop fortement des autres valeurs de l'échantillon.

La boite à moustache permet de représenter des valeurs représentatives d'un échantillon statistique :

- Son minimum et son maximum (en dehors des valeurs exceptionelles)
- Sa médiane
- Son premier et son troisème quartiles
- Ses valeurs exceptionelles (si il en a)

Elle est souvent utilisée pour comparer plusieurs échantillons statistiques issus d'une expérience qui se répète.



Dans le graphique ci-dessus on compare trois catégories d'échantillon statistiques, chacune de ces catégories contenant trois échantillons. Les valeurs des échantillons peuvent être lues à gauche.

2 Exercices

Exercice 1 :
Pour chacun des échantillons statistiques e suivants, donner son maximum, son minimum, ses quartiles, sa moyenne, sa variance et son écart-type. Dire également si certaines valeurs paraissent aberrantes.

1. $e = \{2, 1, 2, 3, 4, 8, 6, 1000\}$
2. $e = \left\{\frac{3}{5}, 1, \frac{5}{6}, \frac{5}{2}\right\}$
3. $e = \{32, 64, 512, 128, 64, 32, 32, 512, 512, 32, 64\}$
4. $e = \{32, 128, 128, 128, 128, 128, 128, 128, 128, 128, 128, 128, 128, 512\}$

Exercice 2 :
Soient e un échantillon statistique et v une valeur de e. On appelle **fréquence** de v le nombre de fois où v apparait dans e, divisé par le nombre de valeurs de e.
Pour chaque échantillon de l'exercice 1, donner la fréquence de chacune de ses valeurs.

Exercice 3 :
Dans une école il y a 150 ordinateurs. Tous les jours, chaque ordinateur envoie un message d'erreur à un serveur de l'école s'il a eu un problème. Sinon, il n'envoie rien.
On a relevé les erreurs reçues par le serveur pendant une semaine. On obtient les résultats suivants :

Jour	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi	Dimanche
Erreurs	22	35	15	149	28	23	17

Dessiner l'histogramme de l'échantillon des erreurs en excluant les potentielles valeurs exceptionnelles.

Exercice 4 :
Pour chaque échantillon de l'exercice 1, dessiner la boite à moustache correspondante.