

# Bilan Statistiques (et Probabilités)

## LP Rob&IA - 3

[R5.03 - Maths IT]

October 7, 2025

# Introduction

# Avant-Propos

- La **Statistique** est une méthode scientifique qui consiste à observer et à étudier une/plusieurs particularité(s) commune(s) chez un groupe de personnes ou de choses.
  - Les **Probabilités** quant à elles proposent des modèles théoriques permettant de structurer tous les phénomènes liés au hasard.
  - Nous rappelons le vocabulaire suivant :
    - *Population* : collection d'objets à étudier ayant des propriétés communes ;
    - *Échantillon* : partie étudiée de la population ;
    - *Variable* : propriété commune aux individus de la population que l'on souhaite étudier. Elle peut être :
      - *qualitative* : couleur, label par exemple
      - *quantitative* : par exemple taille, masse, pression
- Une variable quantitative peut être :
- *continue* : toute valeur d'un intervalle de  $\mathbb{R}$  possible
  - *discrète* : un nombre entier et fini de valeurs possibles

# Deux directions en Statistique

## ① Statistique **descriptive** :

But : décrire, résumer, représenter des données nombreuses ou suffisantes

- Data visualisation
- Détermination des paramètres de position, de dispersion, de relation
- Questions liées aux grands ou petits jeux de données

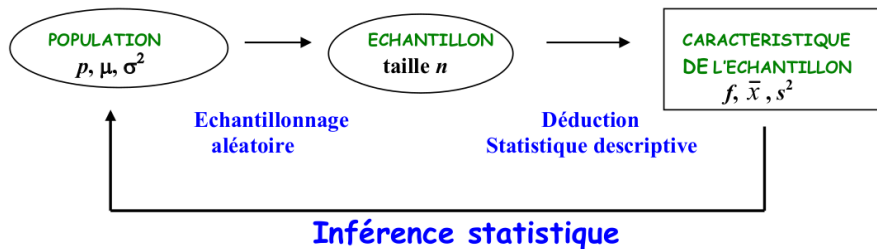
## ② Statistique **inférentielle** :

But : exploitation de données partielles d'une population (échantillon).

Les données sont des réalisations de variables aléatoires, qui suivent certaines lois de probabilité

- Intervalles de confiance
- Estimation de paramètres
- Tests d'hypothèse
- Modélisation comme droite de régression

# Schéma de synthèse



# Statistique Descriptive

# Nombre de variables d'une série statistique

Lorsque l'on observe une seule variable pour les individus de la population, on parle de statistique **univariée**, et de statistique **multivariée** lorsqu'on en observe au moins deux.

Exemple :

Vous vous rappelez peut-être du jeu de données des Iris de Fisher utilisé par M. Toffano en IA

- univarié : on garde uniquement la longueur des pétales pour prédire l'espèce Setosa, Versicolor ou Virginica
- multivarié : on prend longueur des sépales, largeur des sépales, longueur des pétales et largeur des pétales



# Cas de la statistique multivariée

Dans le cas multivarié (le plus courant et pertinent), l'étude est plus complexe et la représentation souvent impossible.

Donc on retrouve deux familles de méthodes :

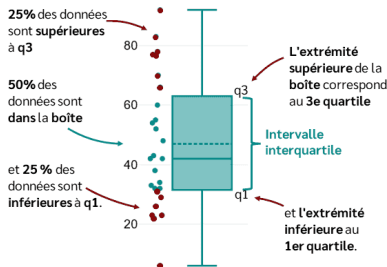
- Les **méthodes R** (factorielles) : recherche de réduction de la dimension comme avec l'Analyse en Composantes Principales par exemple.
- Les **méthodes Q** (classification) : réduire le nombre d'individus en formant des groupes homogènes (clusters)

En fait, ce sont les méthodes développées dans l'apprentissage non supervisé !

# Statistique descriptive à 1 variable (univariée)

Une série statistique est définie comme le vecteur colonne  $X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$

- 1 Sa moyenne  $\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$
- 2 Sa médiane  $m = x_{\frac{n+1}{2}}$   
(sépare en deux groupes de même effectif)
- 3 Ses quartiles (4 groupes), déciles (10 groupes), centiles (100 groupes)
- 4 Son étendue  $e = \max(x_i) - \min(x_i)$
- 5 Sa variance  $Var(X) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2$
- 6 Son écart-type est  $\sigma(X) = \sqrt{Var(X)}$



# Statistique descriptive à 2 variables (bivariée)

Une série statistique sera considérée comme la matrice  $X = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \\ x_{n1} & x_{n2} \end{pmatrix} = (X_1 \ X_2)$

- ❶ Sa moyenne  $\overline{X} = (\overline{X_1} \ \overline{X_2})$  (on parlera de point moyen)
- ❷ Sa variance  $Var(X) = (Var(X_1) \ Var(X_2))$
- ❸ Son écart-type est  $\sigma(X) = (\sigma(X_1) \ \sigma(X_2))$
- ❹ Sa covariance  $V(X) = Cov(X_1, X_2) = \left( \frac{1}{n} \sum_{i=1}^n x_{i1} x_{i2} \right) - \overline{X_1} \cdot \overline{X_2}$
- ❺ Son coefficient de corrélation  $R(X) = \rho(X_1, X_2) = \frac{Cov(X_1, X_2)}{\sigma(X_1)\sigma(X_2)}$

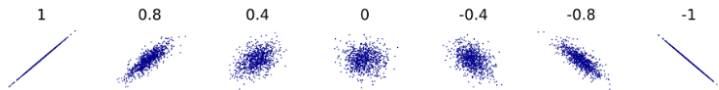
# Statistique descriptive à 2 variables (bivariée)

Le coefficient de corrélation  $\rho(X_1, X_2)$  est compris entre  $-1$  et  $1$  et mesure la relation entre les variables  $X_1$  et  $X_2$ .

Plus le coefficient est proche des valeurs extrêmes  $-1$  et  $1$ , plus la corrélation linéaire entre les variables est forte.

- Si  $\rho > 0$ , les valeurs prises par  $X_2$  ont tendance à croître quand les valeurs de  $X_1$  augmentent ;
- Si  $\rho < 0$ , les valeurs prises par  $X_2$  ont tendance à décroître quand les valeurs de  $X_1$  augmentent ;
- Si  $\rho = 0$ , les variables  $X_2$  et  $X_1$  sont indépendantes (linéairement !) ;

Exemples de coefficients de corrélation :



# Statistique descriptive à $p$ variables (multivariée)

Une série statistique sera considérée comme la matrice  $X = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ x_{21} & \dots & x_{2p} \\ \vdots & & \vdots \\ x_{n1} & & x_{np} \end{pmatrix}$

- 1 Sa moyenne  $\bar{X} = (\bar{X}_1 \ \bar{X}_2 \ \dots \ \bar{X}_p)$
- 2 Sa variance  $Var(X) = \frac{1}{n} \|X - \bar{X}\|^2$
- 3 Son écart-type est  $\sigma(X) = \frac{1}{\sqrt{n}} \|X - \bar{X}\|$
- 4 Sa covariance  $V(X) = Cov(X_i, X_j) = \left( \frac{1}{n} X_{i\cdot} X_{j\cdot} \right) - \bar{X}_i \bar{X}_j$  (Köning-Huygens)
- 5 Son coefficient de corrélation  $R(X) = \rho(X_i, X_j) = \frac{Cov(X_i, X_j)}{\sigma(X_i)\sigma(X_j)}$  (Bravais-Pearson)

# Statistique Inférentielle

# Synthèse Lois de Probabilités d'une variable aléatoire

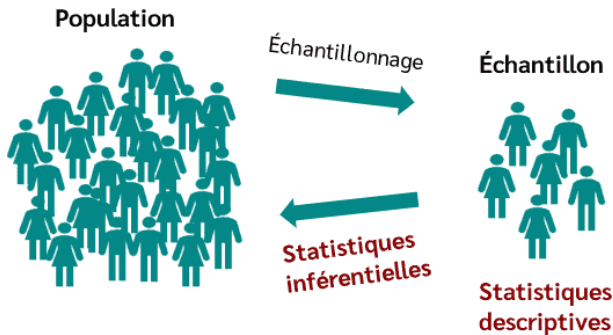
- On appelle variable aléatoire toute fonction  $X$  qui associe chaque éventualité de  $\Omega$  une valeur réelle.
- Une série statistique peut être considérée comme des valeurs prises par  $X = \{x_1, x_2, \dots, x_n\}$
- Du coup,  $X$  peut être discrète mais aussi continue ( $X = [a, b]$ )
- On utilise les notations courantes :
  - $p(X = k)$  la probabilité que  $X$  prenne la valeur  $k$
  - $E(X)$  l'espérance de  $X$  (tendance centrale)
  - $V(X)$  la variance de  $X$  (dispersion)
  - $\sigma(X)$  l'écart-type de  $X$  (dispersion dans la même unité que  $X$ )
- Une loi de probabilité structure la répartition des probabilités selon les valeurs prises par  $X$

→ Fiche Lois.

# Échantillonnage et estimation

L'échantillonnage et l'estimation sont deux problèmes inverses en fait :

- Dans l'échantillonnage, on connaît les paramètres de la population à étudier. On cherche dans quel intervalle on peut retrouver ces paramètres et avec quelle précision dans un échantillon de taille donnée.
- L'estimation ou statistique inférentielle est le problème inverse. On cherche à partir d'un échantillon à estimer les paramètres de la population à étudier.



# Tableau récapitulatif estimation

Nous avons vu l'an dernier les éléments suivants :

Paramètre de la population totale à estimer.	Valeur du paramètre dans l'échantillon de taille $n$	Estimation ponctuelle pour la population totale	Estimation par intervalle de confiance pour la population totale
Moyenne	$\bar{x}$	$\mu_0 = \bar{x}$	$\left[ \bar{x} - u_\alpha \frac{\sigma}{\sqrt{n}}; \bar{x} + u_\alpha \frac{\sigma}{\sqrt{n}} \right]$
Écart-type	$s$	$\hat{\sigma} = s \sqrt{\frac{n}{n-1}}$	
Fréquence	$f_e$	$\hat{p} = f_e$	$\left[ f_e - u_\alpha \sqrt{\frac{f_e(1-f_e)}{n-1}}; f_e + u_\alpha \sqrt{\frac{f_e(1-f_e)}{n-1}} \right]$

Dans la dernière colonne :

- $1 - \alpha$  correspond au niveau de confiance (souvent 95%)
- $u_\alpha$  le quantile d'ordre  $1 - \frac{\alpha}{2}$  (souvent associé à la loi  $N(0, 1)$ )

# Choix de la statistique de test et sa loi

- ❶ L'échantillon considéré (et donc le processus étudié) est-il considéré comme des variables gaussiennes ou non ? Pour synthèse :

$n \leq 30$ (petit)	$n > 30$ (grand)	$n > 100$ (très grand)
Loi de Student $t_{n-1}$ ou du Khi 2 $\chi_{n-1}^2$	$N(0, 1)$	$N(0, 1)$ ou pas ! (Bernoulli iid)

- ❷ Le paramètre étudié est une moyenne, une variance (ou écart-type) ou une proportion ?

Paramètre	Cas	Stat de test $T$	Loi associée
moyenne $\mu_0$	Gaussien	$T = \sqrt{n} \frac{\bar{x} - \mu_0}{s}$	Student $t_{n-1}$
moyenne $\mu_0$	non Gaussien ( $n$ grand)	$T = \sqrt{n} \frac{\bar{x} - \mu_0}{s}$	$N(0, 1)$
variance $\hat{\sigma}^2$	Gaussien	$T = (n-1) \frac{s^2}{\hat{\sigma}^2}$	Khi 2 $\chi_{n-1}^2$
proportion $\hat{p}$	non Gaussien ( $n$ grand)	$T = \sqrt{n} \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}}$	$N(0, 1)$

La logique est toujours : (estimateur – hypothèse) / (écart-type hypothèse)

# (Ré)-Introduction aux tests d'hypothèses

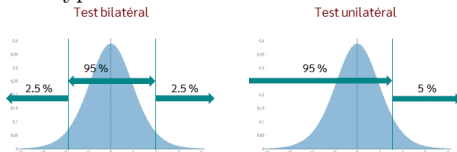
- On parle ici d'une méthodologie permettant de rejeter (ou pas) de manière rigoureuse et robuste une hypothèse statistique que l'on note  $H_0$  (hypothèse nulle).

Les données statistiques à disposition permettent-elles de réfuter  $H_0$  ?

- Il existe deux grands types de tests :
  - Les tests paramétriques : on fait une hypothèse paramétrique sur la loi des données sous  $H_0$  (loi normale, loi de Poisson...). Les hypothèses du test concernent alors les paramètres de cette loi.
  - Les tests non-paramétriques : ne nécessitant pas d'hypothèse sur la loi des données
- Un test d'hypothèse permet de répondre par exemple aux questions suivantes :
  - ① La taille moyenne des étudiants en LP est-elle de 1,67 mètre ?
  - ② L'écart type de leur taille est-il égal à 12,70 centimètres ?
  - ③ Les étudiants et les étudiantes en LP ont-ils la même taille ?
  - ④ La taille des étudiants en LP suit-elle une loi normale ?

# Construction d'un test d'hypothèse

## 1 Détermination du type de test : bilatéral ou unilatéral



- 2 Choix des hypothèses  $H_0$  (statu quo) et  $H_1$  (alternatif)
- 3 Choix de la distribution (normale, Student, ...)
- 4 Choix d'un seuil de signification  $\alpha$  : proba de rejeter  $H_0$  en sachant que  $H_0$  est vraie
- 5 **Approche classique** : On détermine la région de rejet à partir de  $\alpha$ , puis on regarde si la statistique tombe dedans.
- 6 **Approche p-value** : On calcule la  $p$ -value (probabilité de vraisemblance des données observées sous  $H_0$ ) et on la compare directement à  $\alpha$ .

La  $p$ -value nous donne une information plus riche car elle quantifie exactement "à quel point" nos données sont incompatibles avec  $H_0$ , pas seulement un "oui/non" de rejet.

# Synthèse

Diminuer le seuil  $\alpha$  du test a deux conséquences :

- On réduit l'erreur de première espèce (faux positif), c'est-à-dire rejeter  $H_0$  alors que  $H_0$  est vraie.
- On augmente la région d'acceptation de  $H_0$ .

On augmente ainsi un second risque : celui d'accepter  $H_0$  alors que  $H_0$  est fausse. C'est l'erreur de seconde espèce  $\beta$ .  $\beta$  est la probabilité d'accepter  $H_0$  alors que  $H_0$  est fausse (faux négatif).

Quand  $\alpha$  diminue,  $\beta$  augmente et inversement.  $\beta$  s'appelle la puissance du test. C'est la probabilité de rejeter  $H_0$  alors que  $H_0$  est fausse.

		Réalité	
		$H_0$ est vraie	$H_1$ est vraie
Conclusion du test	Rejeter $H_0$	Mauvaise décision Probabilité $\alpha$ Erreur de première espèce.	Bonne décision Probabilité $1 - \beta$ Puissance du test.
	Accepter $H_0$	Bonne décision Probabilité $1 - \alpha$ .	Mauvaise décision Probabilité $\beta$ Erreur de deuxième espèce.