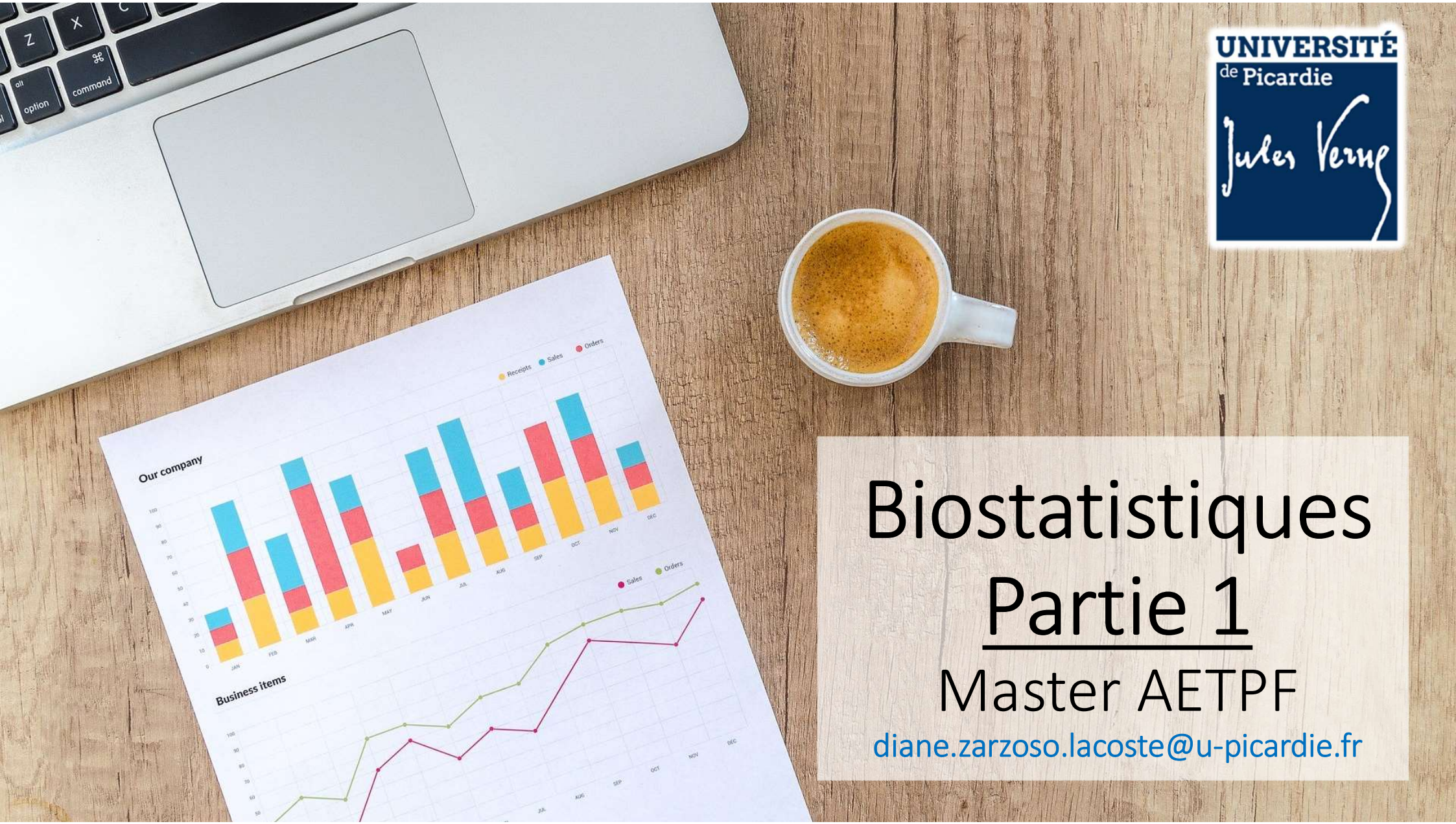


Biostatistiques
Partie 1
Master AETPF

diane.zarzoso.lacoste@u-picardie.fr

Modalités d'examen

Examen écrit de 2 h, le sujet sera constitué à minima de :

- Une partie QCM
- 1 ou 2 exercices d'interprétation (code et sorties R)

Une note de CC :

- S1: 2 questions type QCM sur chaque partie de cours → travail individuel
- *S2: préparation de fiches R sur fonctions utiles pour UE Diagnostic → par groupes de 3-4*

NB :

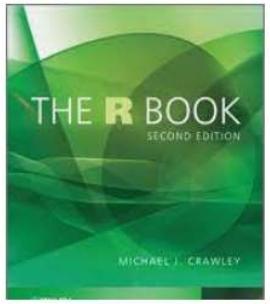
- Cours/fiches papier et calculatrice autorisés
- Ordinateurs, tablettes, téléphones interdits
- Savoir réaliser un test statistique manuellement
  (vidéos rappel/tuto + exercices corrigés sur MOODLE→ question QCM?)

Pour les séances de TP + CM/TD (partie 3 → fin)

→ **Venir avec votre ordinateur portable** (1 pour 2 si nécessaire –mais non recommandé-)

→ **R puis Rstudio doivent être installés** (dans cet ordre) sur votre ordi AVANT le TP1

Tuto installation (et documentation) R et Rstudio :
https://quanti.hypotheses.org/1813



Lien vers le R BOOK :
https://www.cs.upc.edu/~robert/teaching/estadistica/TheRBook.pdf

Plan général du cours

30 h de cours :

Partie 1. Démarche scientifique

Partie 2. Introduction aux statistiques

Partie 3. Les tests relevant de la statistique du χ^2

Partie 4. Les tests sur les différences de moyennes

Partie 5. Etude des relations entre variables

Objectifs :

- Savoir identifier le test approprié (fonction question scientifique et type données)
- Savoir réaliser le test et vérifier ses conditions d'application (manuellement et sous R)
- Savoir en interpréter les résultats

Plan du cours

Partie 1. Démarche scientifique

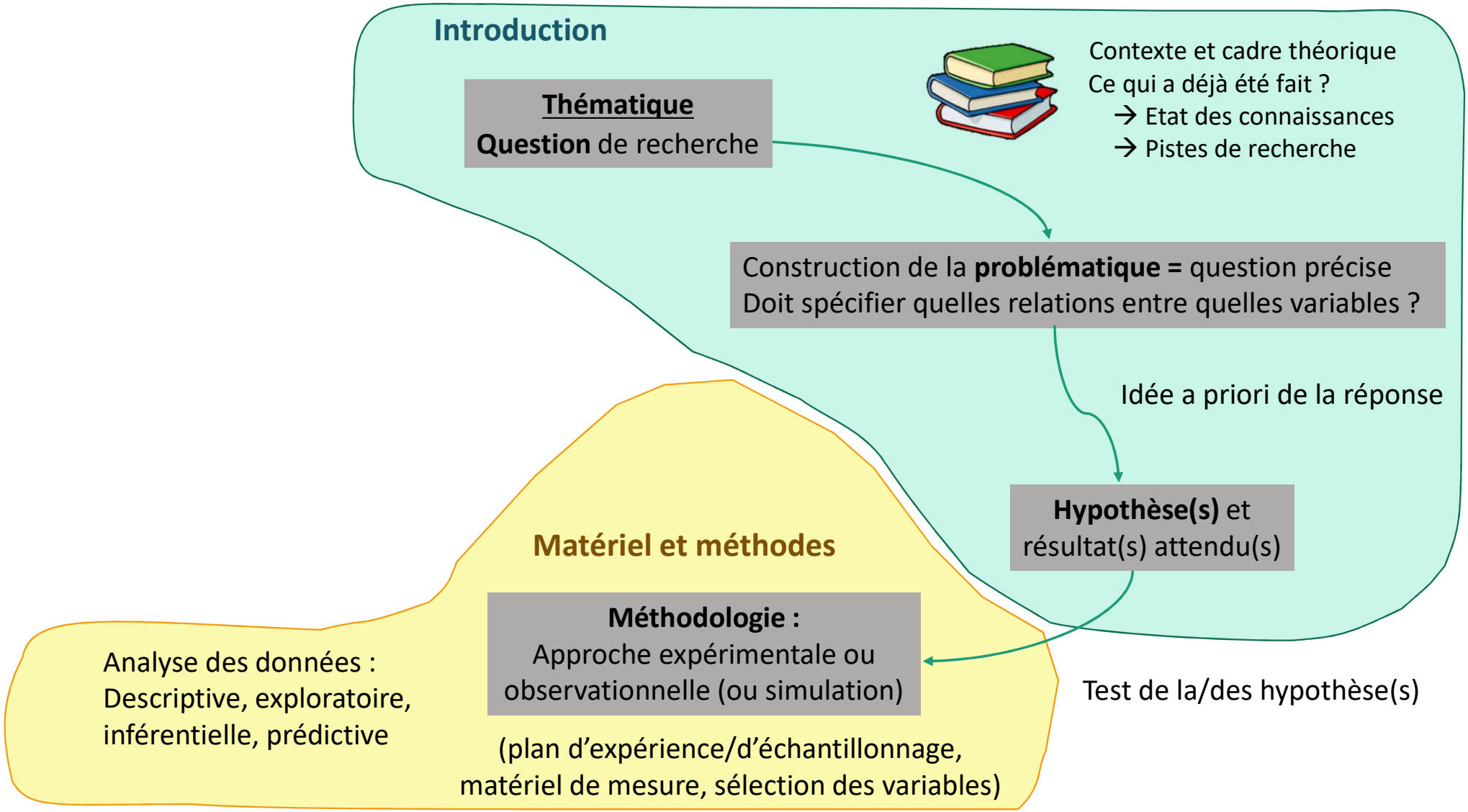
1. **Introduction : De la problématique à l'hypothèse de recherche**

2. Méthodologie de recherche : Protocole et planification

   - Les variables
   - Population, échantillon, individu statistique
   - Type d'investigation : Etude expérimentale vs observationnelle
   - Introduction aux plans expérimentaux
   - Introduction aux stratégies d'échantillonnage

3. Vers l'analyse statistique des données

Démarche scientifique : Approche hypothético-déductive



Démarche scientifique : Approche hypothético-déductive

Communication des résultats
(rapport, article, poster..)

Introduction

Thématique
Question de recherche

Contexte et cadre théorique
Ce qui a déjà été fait ?
→ Etat des connaissances
→ Pistes de recherche

Discussion / conclusions

Conclusion /
Généralisation ?

Construction de la **problématique** = question précise
Doit spécifier quelles relations entre quelles variables ?

Idée a priori de la réponse

Résultats

confrontation attendus *vs* observés
→ confirme ou infirme
hypothèse(s) de départ

Hypothèse(s) et
résultat(s) attendu(s)

Matériel et méthodes

Méthodologie :
Approche expérimentale ou
observationnelle (ou simulation)

Analyse des données :
Descriptive, exploratoire,
inférentielle, prédictive

(plan d'expérience/d'échantillonnage,
matériel de mesure, sélection des variables)

Test de la/des hypothèse(s)

Démarche scientifique : Approche hypothético-déductive

Etablir une méthodologie de recherche pour tester l'hypothèse théorique implique de nombreuses décisions:

On étudie l'effet de quoi sur quoi ?

- Quelles variables d'intérêt (mesurée ou observée, nature, rôle)

Pourquoi?

- Lien avec connaissances dispo (ampleur pbmtq, causes probables phéno)

Sur qui?

- Population cible, taille et nombre d'échantillons/groupes à comparer

Comment?

- Identification des analyses statistiques adaptées en amont
- Type d'investigation (observation vs expérimentation)
- Plan d'expérience / d'échantillonnage
- Instruments de mesure des variables (fiabilité, validité)

Où et Quand?

- Echelle d'investigation spatio-temporelle, organisation logistique

Définition des objectifs et de la problématique

- **Thématique :** Agriculture

- **Objectif étude :** Comparer les effets de ≠ engrais sur le développement des épis.

- **Problématique :** Est-ce que les effets de 3 engrais (A, B, C) sont les même sur les rendements en maïs?

- Peut être déclinée en plusieurs **sous-questions** plus précises → plan expérimental :
  - **Exemple 1** : Quels effets du type d'engrais (A, B, C) sur la rapidité de maturation des épis (temps)?
  - **Exemple 2** : Quels effets de la variété de maïs (traditionnelle vs OGM) **<u>et</u>** du type d'engrais (A, B, C) sur la qualité des épis (note de 0 à 100)?

… etc …

Trad.

OGM

Définir l'hypothèse de recherche

Réponse anticipée (affirmation provisoire) à la problématique qui demande à être scientifiquement vérifiée (soumise à l'épreuve des faits = recherche empirique).

Exemple 1 :

Quels effets des 3 **types d'engrais** (A, B, C) sur la **rapidité de maturation des épis maïs** (temps) ?

Nature des données ?

- **1 variable dépendante** (= à expliquer) **Quantitative**
- **1 variable indépendante** (=explicative) **Qualitative**

Hypothèse simple : Les effets des 3 **engrais** sur la **rapidité de maturation des épis** diffèrent

(ou : la rapidité de maturation des épis est supérieure avec l'engrais A → sens différence)

→ Expérimentation (conditions contrôlées) et ANOVA 1 à facteur?

Définir l'hypothèse de recherche

Exemple 2 :

Effet de **variété maïs** (trad. vs OGM) et **type d'engrais** (A, B, C) sur **qualité des épis** (note 0:100) ?

Nature des données ?

- **1 variable dépendante Quantitative**
- **2 variables indépendantes Qualitatives**

Hypothèses plus complexes: 3 effets peuvent être mesurés/testés

Effets principaux

- Effet de la **variété de maïs** sur la **qualité des épis**
- Effet du **type d'engrais** sur la **qualité des épis**

Effet d'interaction
variété * engrais

- Effet synergique ou croisé de **variété** et **type d'engrais** sur la **qualité des épis**

ECQ l'effet du type d'engrais sur la qualité de l'épis dépend de la variété utilisée ?

→ Expérimentation et ANOVA à 2 facteurs?

Plan du cours

Partie 1. Démarche scientifique et introduction aux statistiques

1. Introduction : De la problématique à l'hypothèse de recherche

2. **Méthodologie de recherche : Protocole et planification**

   - Les variables
   - Population, échantillon, individu statistique
   - Type d'investigation : Etude expérimentale vs observationnelle
   - Introduction aux plans expérimentaux
   - Introduction aux stratégies d'échantillonnage

3. Vers l'analyse statistique des données

Les variables : quelques définitions

Une variable c'est quoi ?

Caractéristique mesurable/observable sur les individus d'une population.

Sa valeur/modalité diffère d'un individu à l'autre (= variabilité inter-individuelle).

En statistiques, on travaille sur des variables aléatoires

Variable aléatoire : Toute variable pour laquelle il est impossible de prévoir a priori et avec certitude la valeur (ou modalité) qu'elle prendra chez un individu donné, même si toutes les conditions contrôlables de son observation sont fixées.

Sa valeur/modalité est donc le résultat (= la réalisation) d'une expérience aléatoire (= probabiliste).

Ex: Masse d'un épis de maïs, texture d'un sol, nombre de petit-pois dans une conserve, niveau d'appréciation d'un yaourt, …

Les variables : quelques définitions

Une variable aléatoire est :

- **Caractérisée par sa loi de probabilité** :

  1. L'ensemble des valeurs qu'elle peut prendre
  2. La probabilité d'observer chaque valeur dans la population (somme des probabilités = 1)

- **Représentée** par une **lettre majuscule** (X,Y,Z…) et ses **réalisations** par une **minuscule indicées** (x_i) → les données

- Sa **valeur** n'est cependant **pas totalement imprévisible** si l'on connait la nature de la variable et les conditions de sa réalisation → Parfois possible de prévoir une **moyenne**.

$$y_i = \bar{y} + \varepsilon_i$$

y_i : valeur observée sur l'individu i (réalisation)
$\bar{y}$: valeur moyenne attendue → terme déterminé
ε_i : écart à la valeur attendue → terme aléatoire

Ex: Masse d'un épis de maïs provenant d'un champ

Les variables : leur nature

VA qualitatives (facteurs)
(Valeurs catégorielles à ≥ 2 modalités/niveaux)

Nominales	**Ordinales**
Nommer/Qualifier	***Ordonner/Ranger***
Expriment une qualité non quantifiable sans gradation logique	**Expriment une qualité non quantifiable avec gradation logique**
Sexe, nom, génotype, couleur, variété …	*Stades vie, appréciation, fréquence évènement …*

Les opérations (somme, moyenne, multiplication…) n'ont pas de sens!

Les variables : leur nature

VA qualitatives (facteurs)

(Valeurs catégorielles à ≥ 2 modalités/niveaux)

Nominales

Nommer/Qualifier

Expriment une qualité non quantifiable sans gradation logique

Sexe, nom, génotype, couleur, variété …

Ordinales

Ordonner/Ranger

Expriment une qualité non quantifiable avec gradation logique

Stades vie, appréciation, fréquence évènement …

VA quantitatives

(valeurs numériques uniquement)

Discrètes

Compter

Prennent uniquement des valeurs entières

nb de descendants, de fleurs, de bactéries, d'espèces …

Continues

Mesurer

Peuvent prendre des valeurs décimales

Taille, concentration, masse, température …

Les opérations (somme, moyenne, multiplication…) n'ont pas de sens!

Les opérations ont un sens !!

Les variables : leur choix

En fonction de leur **rôle** :

- Variable **Dépendante** (VD = à expliquer, réponse, Y) que l'on **observe** et **mesure** sur les individus statistiques → celle dont on **veut expliquer les variations**.
- Variables **Indépendantes** (VI = explicatives, facteur, prédicteur, …, $X_{1 \text{ à } n}$) sont celles dont on suppose que les variations **causent/influencent** celles de la **VD Y** → Celles dont on veut tester/quantifier les effets (au moins 2 modalités).

➔ **En modélisation**, on cherche à quantifier: ***Si X_i varie de « tant », alors de combien varie Y ?***

Exemple :

VD

« Y »

Rapidité maturation épis

Dépend de :

VI

« X1 »

« X2 »

« Xn »

A B C

Type d'engrais

Température

Diversité microbienne sol

<u>En général :</u>

- *1 VI catégorielle est nommée « facteur »*
- *1 VI quantitative est nommée « covariable »*

Les variables : leur choix

Selon **mode de sélection** des valeurs/modalités des **VI** et **nature de leurs effets** sur **VD** :

- **VI fixes** : valeurs/**modalités** délibérément **choisies et fixées a priori** par le scientifique.

✓ Ces variables peuvent être manipulées/contrôlées pour créer conditions expérimentales

✓ Permet une étude/expérience facilement reproductible

✓ Les conclusions portent sur **ces modalités** particulières

(ex: classes de T°c ou de niveau d'irrigation choisies et fixées, marques engrais d'intérêt testées)

- **VI aléatoires** : valeurs/modalités **sélectionnées aléatoirement** parmi un ensemble de valeurs/modalités possibles (non contrôlées a priori).

✓ Ces variables sont des caractéristiques «naturelles» propres aux individus statistiques

✓ Elles sont recueillies sur le terrain ou à la suite d'une expérimentation

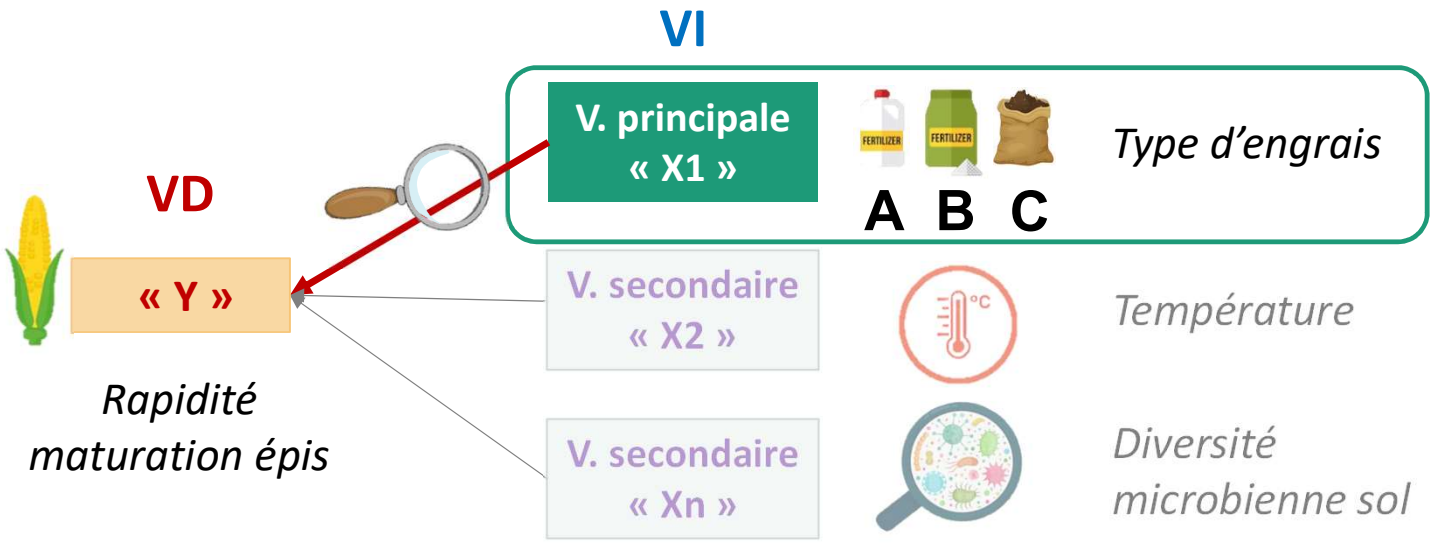
✓ S'intéresse à **l'effet global** de la variable/facteur, pas à des valeurs/modalités particulières

(ex: classes de T°c ou pluviométrie enregistrée sur le terrain, marques engrais quelconques parmi existantes)

Les variables : leur choix

En fonction des **objectifs de l'étude**, on distingue les :

- **Variables principales** : **VI qu'on choisi d'étudier** et sur lesquels portent la/les hypothèse(s)
- **Variables secondaires** : **VI** dont l'influence ne constitue **pas le centre d'intérêt** de l'étude

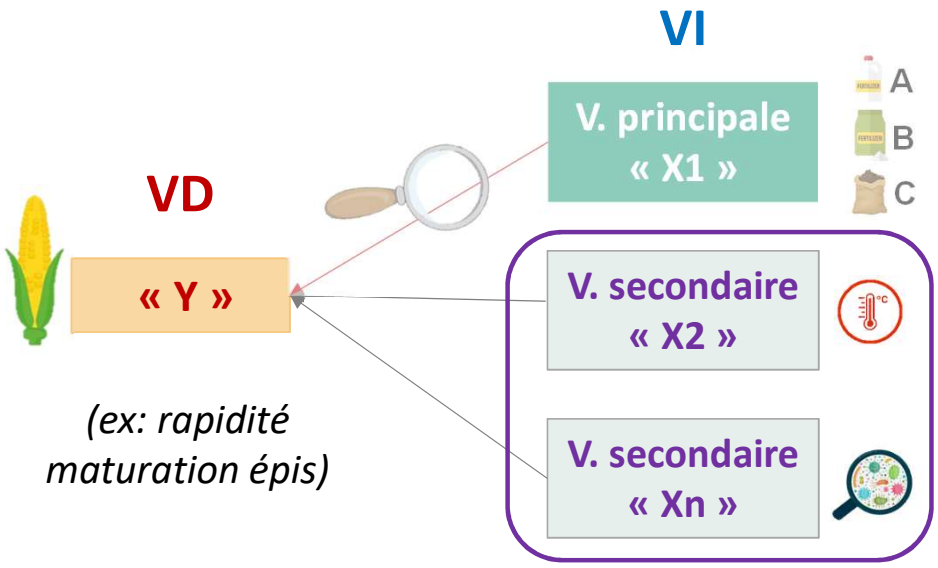


Les variables : leur choix

En fonction des **objectifs de l'étude**, on distingue les :

- **Variables principales** : **VI qu'on choisi d'étudier** et sur lesquels portent la/les hypothèse(s)
- **Variables secondaires** : **VI** dont l'influence ne constitue **pas le centre d'intérêt** de l'étude

↳ Doit **neutraliser leurs effets** afin d'isoler ceux des variables principales sur la **VD** :



✓**Variable contrôlée** : Effet neutralisé via stratégie de contrôle (variable/facteur maintenu constant, randomisé, contrebalancé) *(ex: irrigation, température…)*

✓**Variable parasite (facteur confondant, de confusion)** : non contrôlé → va influencer les variations de la VD sans que le scientifique ne s'en aperçoive ou puisse savoir/quantifier comment *(ex: diversité microbienne sol)*

Quelle est la nature des variables suivantes:

(Réponses possibles; A : Nominale, B : Ordinale, C : Discrète, D : Continue)

1. Le volume d'eau du lac ✓ **D**
2. Le nombre de poissons dans ce lac ✓ **C**
3. La profession ✓ **A**
4. Le niveau d'étude (diplôme) ✓ **B**
5. Le numéro de dossard dans une course ✓ **A**
6. Le temps mis pour compléter la course ✓ **D**

Quelle est le rôle des variables dans les phrases suivantes?

1. A Nöel, vous voulez estimer le temps de cuisson de votre dinde de 3.5 kg

A. Le temps de cuisson est la VI

B. La masse de la dinde est la VI ✓ **1. B**

2. Vous vous intéressez au nombre de jours nécessaires à la germination du blé

A. Le nombre de jours est la VI ✓ **2. A**

B. Le nombre de jours est la VD

Plan du cours

Partie 1. Démarche scientifique et introduction aux statistiques

1. Introduction : De la problématique à l'hypothèse de recherche

2. **Méthodologie de recherche : Protocole et planification**

   - Les variables
   - **Population, échantillon, individu statistique**
   - Type d'investigation : Etude expérimentale vs observationnelle
   - Introduction aux plans expérimentaux
   - Introduction aux stratégies d'échantillonnage

3. Vers l'analyse statistique des données

Rappel : population, échantillon, individu statistique

Population statistique : ensemble d'**individus statistiques** (objets, unités..) appartenant à 1 groupe pour lequel le scientifique étudie 1 ou + VA et souhaite généraliser des conclusions.

Ex : nb petit-pois dans boites conserve produites par 1 entreprise Française en déc 20

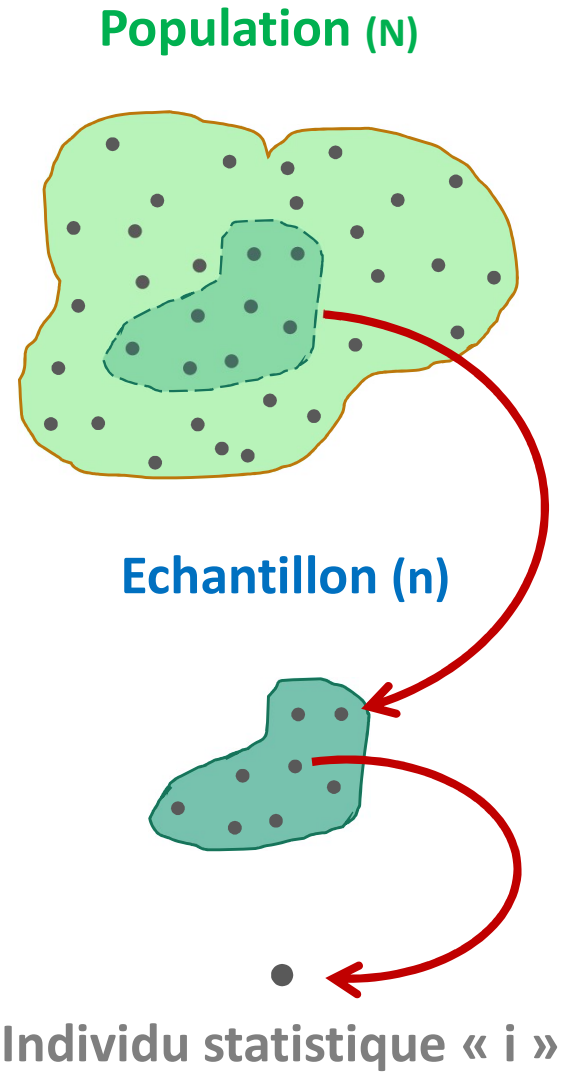
Rarement possible d'accéder aux N individus de la pop

➔ **échantillonnage** (aléatoire, même proba. sélection)

Ex : sélection de 200 boites de petit-pois produites par cette entreprise en déc 20

L'**échantillon :** sous-ensemble de n **individus statistiques** sélectionnés dans la population et sur lesquels seront réalisées les mesures/observations.

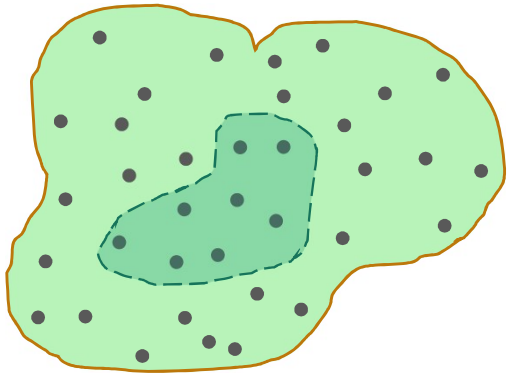
Il doit être **représentatif** de la population dont il est issu (mêmes caractéristiques –paramètres de position, dispersion -, mais effectif **n** < **N**) → **plan d'échantillonnage**.



Rappel : Statistique descriptive *vs* inferentielle

Question posée

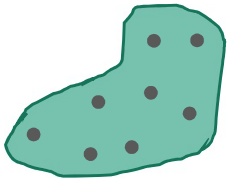
Population (N)



Echantillonnage aléatoire

Observation

Echantillon (n)



Si représentatif

Estimation

Conclusions

Stat. inférentielle :

Généraliser

(paramètres –inconnus-, conclusions tests…)

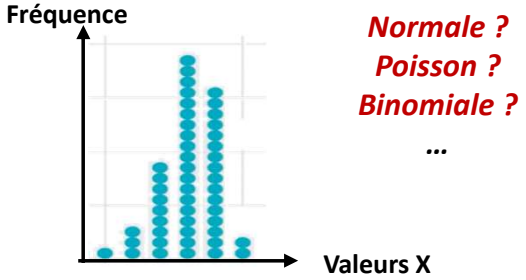
Stat. Descriptive :

Résumer

(distribution, paramètres,…)

Position — Moyenne ? Mode ? Médiane ?

Dispersion — Variance ? Ecart type ?



Quelle taille d'échantillon ?

Différentes méthodes selon paramètre à estimer (proportion, moyenne…) et marge d'erreur acceptée (α).

Roscoe (1975)

Quelques règles pour guider le **choix du nombre d'individus** à inclure dans un échantillon :

- 30 à 100 individus → approprié pour la plupart des recherches
- Si 1 échantillon doit être divisé en **sous-groupes**, chacun doit contenir > 30 individus
- Lors de recherches impliquant des **analyses multivariées** (ex. : régression multiple), l'échantillon devrait contenir au moins 10 fois plus d'individus qu'il y a de VI
- Pour les **recherches expérimentales étroitement contrôlées**, l'échantillon pourra ne contenir que 10 à 20 individus

Plan du cours

Partie 1. Démarche scientifique et introduction aux statistiques

1. Introduction : De la problématique à l'hypothèse de recherche

2. **Méthodologie de recherche : Protocole et planification**

   - Les variables
   - Population, échantillon, individu statistique
   - **Type d'investigation : Etude expérimentale vs observationnelle**
   - Introduction aux plans expérimentaux
   - Introduction aux stratégies d'échantillonnage

3. Vers l'analyse statistique des données

Type d'investigation : Approche **expérimentale**

Nécessairement 1 **VD** et au moins 1 **VI** (≥2 modalités) **identifiées en amont** collecte données.

But : tester l'existence et le sens de relations de **causalité** entre VI et VD, comparer des effets.

→ **Question précise** : *ECQ* X (VI) influence Y (VD)? Si oui, comment et de combien ?

Ex: comparer rendement maïs avec et sans engrais (→ pour quantifier : judicieux de tester ≠ doses d'engrais)

Consiste à:

- **Manipuler** les valeurs/modalités de ≥1 variables principales (VI fixes)
- **Contrôler** les variables secondaires (pour isoler les effets des V. Pples)
- **Randomiser** = répartir aléatoirement les individus statistiques (unités expérimentales) dans les conditions expérimentales à comparer
- **Mesurer l'effet** de cette manipulation sur la VD.

Peu être conduit ***ex-situ*** (en laboratoire) ou ***in-situ*** (sur le terrain).

Inconvénient : Prive le système étudié de nombreux phénomènes d'interaction potentiellement à l'œuvre dans la nature.

Sorce : Arvalis

Type d'investigation : Approche **observationnelle**

But : Observer des corrélations ou des co-occurrences entre variables.

→ **Question ouverte** : Qu'est ce qui est lié à quoi? Qu'est ce qui influence quoi?

Le scientifique ne peut ni manipuler ni contrôler les variables (VI aléatoires).

Ex: lien entre l'abondance carabes, diversité végétale et/ou la densité de haies observées dans champs tirés au sort dans 80?

Consiste à :

Recueillir des informations à partir d'un **échantillon aléatoire,** en un **temps et lieu donné**, et en **situation** la plus **naturelle** possible afin d'influencer le moins possible ce qui sera observé.

Nécessairement **sur le terrain** (où de multiples variables interagissent).

Inconvénient : Ne donne pas d'indication sur le sens des causalités ou leurs circuits (direct / indirect, uni / bidirectionnel).

Type d'investigation

… Chacun des types d'investigation a des points forts et des faiblesses …

Le choix de l'un ou l'autre dépend du ***type de question posée*** et de la ***nature des variables***.

Le **type de question** réfère à la **nature de la relation** que l'on veut mettre en évidence :
Relation de cause à effet **vs** Corrélations / co-occurrences

Agronomie, zootechnie, industrie… → adapté à expérimentation : possible de contrôler V. secondaires, pour isoler effets de V. principales et révéler relations causales (X implique Y).

Ecologie → contrôle V. secondaires extrêmement difficile, nombreuses V. parasites certaines pouvant agir à distance (ex: pollutions); mais si mesurés, on peut en tenir compte en modélisation (ex: covariables de l'ANCOVA).
En général on constate que plusieurs VI (X_1 à X_n) inter-reliés influencent la VD (Y) sans qu'1 seule ne puisse en être la cause directe et unique de variation (très rarement causalité).

Notions de validité interne et externe d'une étude

Le choix du type d'investigation (degrés de contrôle) va conditionner le niveau de validité de l'étude :

- **Validité interne** (Campbell & Stanley 1966) : niveau de confiance que le scientifique peut avoir en la validité de ses conclusions vis-à-vis des relations de cause à effet étudiées.
  (validité dite « interne » car elle ne porte que sur les échantillons étudiés)

Evaluation : Jusqu'à quel point puis-je être sûr que les variations de la VD que j'observe sont bien uniquement causées par les variations de(s) la V. principale(s) étudiés(s) (et pas les V. secondaires) ?

Maximisation : Nécessite un échantillonnage aléatoire des individus, leur randomisation dans les groupes à comparer (= modalités du F. principal), et que les V. secondaires puissent être contrôlées.

- **Validité externe** *:* degré de généralisation possible des résultats obtenus (applicables à d'autres situations, contextes, individus statistiques?).

→ 1 échantillonnage aléatoire est indispensable pour garantir bonne validité externe.

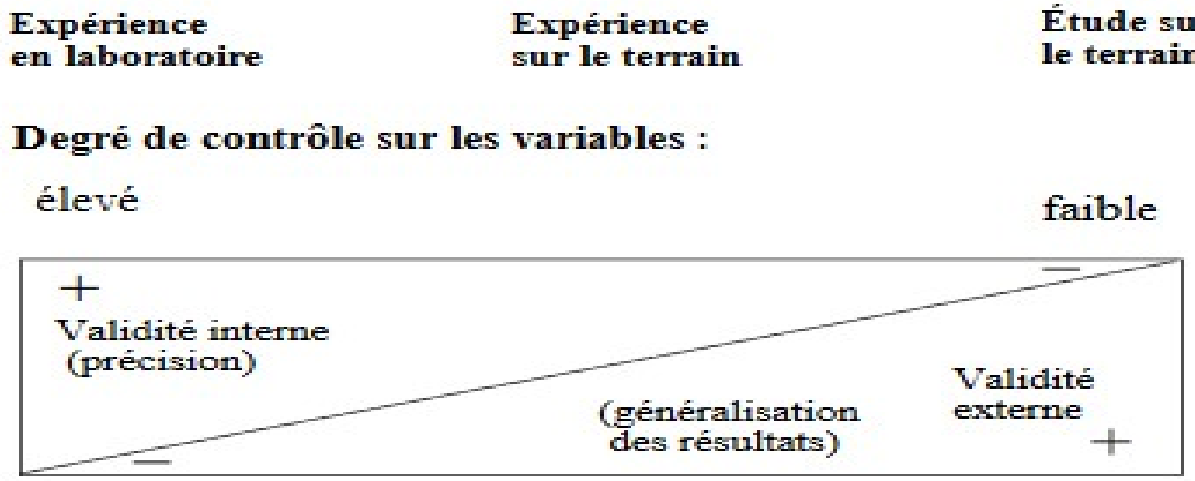
Notions de validité interne et externe d'une étude

Un compromis permanent : + la validité interne d'une recherche est élevée, + il y a de chances que sa validité externe soit faible.

Pourquoi?

Le niveau de validité interne dépend du niveau de contrôle.

Plus le niveau de contrôle est important, plus on s'éloigne des conditions naturelles (perte de réalisme), donc moins les conclusions seront généralisables.



Boire du thé avant le coucher

Une étude a été menée auprès d'un échantillon aléatoire d'adultes afin de tester l'influence du thé sur le temps d'endormissement.

Les résultats montrent que les personnes buvant une tasse de thé avant de se coucher ont tendance a s'endormir plus vite que celles n'en buvant pas.

De quel type d'étude statistique s'agit il?

A. Étude observationnelle ✓ **A**

B. Étude expérimentale

Média sociaux et bonheur

Des chercheurs ont aléatoirement réparti des individus volontaires entre 2 groupes :

- Les indiv du 1er devaient garder leurs habitudes d'utilisation des médias sociaux

- Ceux du 2nd n'avaient accès à aucun média social

Le but était d'étudier dans quel groupe les personnes avaient tendance à être plus heureuses.

De quel type d'étude statistique s'agit il?

A. Étude observationnelle

B. Étude expérimentale ✓ **B**

Plan du cours

Partie 1. Démarche scientifique et introduction aux statistiques

1. Introduction : De la problématique à l'hypothèse de recherche

2. **Méthodologie de recherche : Protocole et planification**

   - Les variables
   - Population, échantillon, individu statistique
   - Type d'investigation : Etude expérimentale vs observationnelle
   - **Introduction aux plans expérimentaux**
   - Introduction aux stratégies d'échantillonnage

3. Vers l'analyse statistique des données

Quel plan experimental mettre en oeuvre?

1. Les plans à 1 VI

Les plus simples = 1 VD (généralement quantitative) et 1 seule VI.

La **VI** est **qualitative** (catégorielle) et ses modalités (minimum 2) permettent de distinguer les groupes / conditions expérimentales /échantillons à comparer.
(*ex: si 3 modalités → 3 groupes*)

→ 2 types :

- **Plans à groupes (échantillons) indépendants**
- **Plans à groupes (échantillons) appariés (= non indépendants)**

Quel plan experimental mettre en oeuvre?

Exemple médical : *Une molécule A est suspectée d'avoir un effet sur le taux de cholestérol. On veut tester l'effet d'un traitement contenant cette molécule sur la cholestérolémie (mmol/l) de n patients.*

1. identifier VD et VI :

✓ **VD** = taux cholestérol, **VI** = traitement

2. Combien de modalités de la VI ? Lesquelles ?

✓ ***2*** : avec molécule A, sans molécule A (= contrôle)

3. Combien de groupes/échantillons à comparer ?

✓ ***2 :*** *effet du traitement (molécule A vs* contrôle*) sur la cholestérolémie*

Quel plan experimental mettre en oeuvre?

1.1. Plans à groupes (échantillons) indépendants

- Les individus statistiques qui composent chaque échantillon sont distincts et sans lien entre eux = indépendants (ex: pas de lien familial, spatial, temporel…)

- Chaque individu n'est mesuré qu'1 fois et n'est exposé qu'à 1 modalité de la VI

- Les échantillons peuvent être de tailles différentes

On compare généralement la moyenne de la VD obtenue pour chaque échantillon

On s'intéresse aux différences entre les échantillons (variabilité INTER)

Indiv	Contrôle	Trait. A
1	y1	
2	y2	
…	…	
40	y40	
41		y41
….		…
90		y90