

Méthodes Quantitatives

Séance 2 : Statistiques univariées

Distribution statistique et premières analyses
Application aux résultats de l'élection présidentielle 2022

Sciences Po

2024-2025

Plan de la séance

- 1 Concepts fondamentaux et données
- 2 Description des variables
- 3 Distribution statistique
- 4 Indicateurs de tendance centrale
- 5 Indicateurs de dispersion

Les données de l'élection présidentielle 2022

- **Source** : data.gouv.fr
 - Résultats du premier tour par département
 - Variables disponibles :
 - Inscrits, Abstentions, Votants
 - Blancs, Nuls, Exprimés
 - Voix par candidat
- **Précautions méthodologiques** :
 - Granularité départementale $\neq$ résultats individuels
 - Prise en compte de la taille des départements
 - Role de la participation

Qu'est-ce qu'une variable statistique ?

- **Types de variables :**

- **Qualitative** : modalités non numériques
 - Nominale (ex : nom du département)
 - Ordinale (ex : niveau de participation : faible/moyen/fort)
- **Quantitative** : valeurs numériques
 - Discrète (ex : nombre de votes)
 - Continue (ex : pourcentage de votes)

Importance de la description d'une variable

- **Pourquoi décrire les variables ?**

- Étape préliminaire indispensable de toute analyse
- Permet de détecter les erreurs et anomalies
- Aide à choisir les méthodes d'analyse appropriées

- **Dans la pratique scientifique**

- Présent en annexe des articles scientifiques
- Permet la reproductibilité des analyses
- Base de la validation des résultats

- **Points d'attention**

- Vérification du codage
- Traitement des valeurs manquantes
- Détection des valeurs aberrantes

Exemple concret : Table descriptive

Variable	Moyenne	Écart-type	N
Âge	42.13	15.73	2,796
Éducation (années)	12.45	3.78	2,796
Revenu (\$)	25,000	15,000	2,650

Table – Exemple de statistiques descriptives

- Notez les informations essentielles :
 - Tendance centrale (moyenne)
 - Dispersion (écart-type)
 - Taille d'échantillon (N)

Variables Quantitatives

Définition : Variables exprimées numériquement et pour lesquelles les opérations arithmétiques ont un sens

Variables continues

- Taille
- Poids
- Température
- Revenu exact

Variables discrètes

- Nombre d'enfants
- Nombre de pièces
- Effectifs
- Comptages

Caractéristiques importantes :

- Permettent le calcul de moyennes
- Autorisent les opérations arithmétiques
- Nécessitent des outils spécifiques de description

Variables Catégorielles (1/2)

Définition : Variables prenant un nombre fini de modalités

- **Variables nominales**

- Pas d'ordre naturel entre les modalités
- Exemples :
 - Sexe (H/F)
 - Couleur des yeux
 - Profession
 - État civil

- **Codage**

- Par lettres : H/F
- Par chiffres : 1/2 (codage arbitraire)
- Multiple : plusieurs variables dichotomiques

Variables Catégorielles (2/2)

Variables ordonnées

- Relation d'ordre naturelle entre modalités
- Exemple 1 : Niveau d'études
 - Primaire < Secondaire < Supérieur
- Exemple 2 : Fréquence
 - Jamais < Parfois < Souvent < Toujours

Variables intervalles

- Modalités définies par des bornes numériques
- Exemple : Tranches de revenu
 - < 1000€
 - 1000€ - 2000€
 - 2000€ - 3000€
 - > 3000€

De Quantitative à Catégorielle

Processus de catégorisation

• Méthodes courantes

- Découpage en classes d'égale amplitude
- Découpage en quantiles
- Seuils significatifs (théoriques ou pratiques)

• Exemples

- Âge → Classes d'âge
- Revenu → Tranches de revenu

• Avantages

- Simplification de l'information
- Communication plus claire
- Traitement des valeurs extrêmes

• Inconvénients

- Perte d'information
- Choix arbitraire des seuils
- Sensibilité aux bornes

De Catégorielle à Quantitative

Variables dichotomiques (0/1)

• Principe

- Transformation en variables indicatrices
- Une variable par modalité
- Codage 0/1

• Exemple : Niveau d'études

Originale :	Bac	Licence	Master
Indicatrices :			
Bac	1	0	0
Licence	0	1	0
Master	0	0	1

• Utilisation

- Analyses quantitatives
- Régressions
- Calculs de moyennes

Transformation des variables ordinales

En scores numériques

• Principe

- Attribution de valeurs numériques
- Respect de l'ordre
- Choix d'une échelle

• Exemple

Modalité	Score
Jamais	1
Parfois	2
Souvent	3
Toujours	4

• Avantages et limites

- + Permet des calculs
- + Conserve l'ordre
- - On peut interpréter les valeurs directement

Distribution statistique - Définitions

- **Distribution statistique** : répartition des valeurs d'une variable dans une population

Exemples : répartition de niveau d'abstention dans les départements, vote pour Emmanuel Macron, abstention aux précédentes élections...

- **Notations mathématiques** :
 - x_i : valeur de la variable pour l'observation i
 - N : effectif total

Application aux données électorales

Variables d'intérêt :

- **Taux de participation :**

$$\text{Taux de participation} = \frac{\text{Votants}}{\text{Inscrits}} \times 100\%$$

- **Scores des candidats :**

$$\text{Score candidat} = \frac{\text{Voix obtenues}}{\text{Exprimés}} \times 100\%$$

- **Taux de votes blancs/nuls :**

$$\text{Taux} = \frac{\text{Blancs} + \text{Nuls}}{\text{Votants}} \times 100\%$$

Classes statistiques

Pourquoi créer des classes ?

- Regrouper les valeurs proches
- Faciliter la lecture et l'analyse
- Faire apparaître des tendances

• Notations mathématiques :

- x_i : valeur de la variable pour l'observation i
- n_i : effectif de la valeur x_i
- $N = \sum n_i$: effectif total
- $f_i = \frac{n_i}{N}$: fréquence relative

• Propriétés :

- $\sum f_i = 1$ (ou 100% en pourcentage)
- Fréquences cumulées : $F_k = \sum_{i=1}^k f_i$

Représentation des distributions

Trois formes équivalentes :

① **Effectifs absolus** (n_i) :

Nombre de départements pour chaque intervalle de score

② **Fréquences relatives** (f_i) :

$$f_i = \frac{n_i}{N} \times 100\%$$

③ **Fréquences cumulées** (F_i) :

$$F_k = \sum_{i=1}^k f_i = \frac{\sum_{i=1}^k n_i}{N} \times 100\%$$

Exercice 1 : Création d'un tableau croisé dynamique

Étapes dans Google Sheets :

- 1 Fichiers > Importer > Importer > Sélectionner le fichier
- 2 **Créer une nouvelle feuille de calcul**
- 3 Insertion > Graphique > Type de Graphique > Histogramme
- 4 Configurer l'histogramme pour utiliser un nombre différents de classes

Premier aperçu des distributions

Questions à explorer :

• **Forme de la distribution :**

- Symétrique ou asymétrique ?
- Unimodale ou multimodale ?
- Présence de valeurs extrêmes ?

• **Comparaisons possibles :**

- Entre candidats
- Entre régions
- Avec le score national

• **Premières interprétations :**

- Homogénéité/hétérogénéité territoriale
- Identification des cas particuliers
- Hypothèses sur les facteurs explicatifs

Indicateurs de tendance centrale - Introduction

- **Objectif** : résumer une distribution par une valeur "représentative"
- **Trois indicateurs principaux** :
 - Mode
 - Médiane
 - Moyenne
- **Application électorale** :
 - Score "typique" dans un département
 - Comparaison avec le score national
 - Identification des départements représentatifs

Le mode

- **Définition** : valeur la plus fréquente dans une distribution
 - Pour une variable continue : classe modale
 - Peut être multiple (distribution bimodale)
- **Propriétés** :
 - Ne dépend pas des valeurs extrêmes
 - Pertinent pour les variables qualitatives
 - Sensible au choix des classes pour les variables continues

La médiane

- **Définition** : valeur qui partage la distribution en deux parties égales

$$P(X \leq m) = P(X \geq m) = 0.5$$

- **Propriétés** :

- Robuste aux valeurs extrêmes
- Existe toujours pour une variable ordonnée
- Pas influencée par les valeurs aberrantes

- **Dans Google Sheets** :

- =MEDIANE() ou =MEDIAN()
- Données doivent être ordonnées

La moyenne arithmétique

- **Définition** : somme des valeurs divisée par leur nombre

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i = \frac{\sum_{i=1}^k n_i x_i}{\sum_{i=1}^k n_i}$$

- **Propriétés** :

- Sensible aux valeurs extrêmes
- Tient compte de toutes les valeurs
- Centre de gravité de la distribution

- **Dans Google Sheets** :

- =MOYENNE() ou =AVERAGE()
- =SOMMEPROD() pour moyenne pondérée

Moyenne pondérée

- **Définition** : moyenne tenant compte de l'importance relative des observations

$$\bar{x}_w = \frac{\sum_{i=1}^N w_i x_i}{\sum_{i=1}^N w_i}$$

- **Application électorale** :

- Score national = moyenne pondérée par le nombre d'exprimés
- Score départemental moyen = moyenne simple des scores départementaux
- Différence = effet de la taille des départements

- **Dans Google Sheets** :

=SOMMEPROD(plage1, plage2) / SOMME(plage2)

Exercice 2 : Calcul des indicateurs

Objectif : Comparer les différents indicateurs de tendance centrale

Étapes :

- ① Calcul pour chaque candidat :
 - Score moyen départemental (moyenne simple)
 - Score national (moyenne pondérée)
 - Score médian
- ② Interprétation des différences :
 - Écart moyenne/médiane : asymétrie
 - Écart moyennes simple/pondérée : effet taille

Interprétation des résultats

Questions à explorer :

- **Comparaison des indicateurs :**
 - Quel indicateur est le plus pertinent ?
 - Que nous apprennent leurs différences ?
- **Géographie électorale :**
 - Départements "représentatifs" (proches de la moyenne)
 - Départements atypiques (éloignés de la moyenne)
- **Implications politiques :**
 - Structure du vote sur le territoire
 - Forces et faiblesses géographiques

Indicateurs de dispersion - Introduction

- **Objectif** : mesurer la variabilité des valeurs autour du centre
- **Deux types d'indicateurs** :
 - Absolus : dans la même unité que la variable
 - Étendue
 - Écart-type
 - Écart interquartile
 - Relatifs : sans unité
 - Coefficient de variation
 - Rapport interquartile

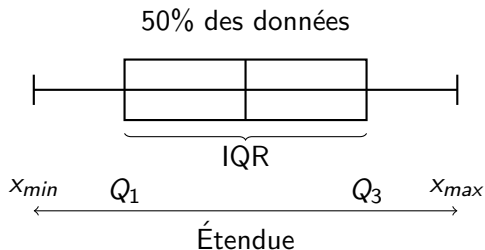
L'étendue et l'écart interquartile

- **Étendue** : différence entre maximum et minimum

$$E = x_{max} - x_{min}$$

- Très sensible aux valeurs extrêmes
 - Facile à calculer et interpréter
- **Écart interquartile** : écart entre Q3 et Q1 :
 $IQR = Q_3 - Q_1$
 - Contient 50% des observations centrales
 - Plus robuste aux valeurs extrêmes
- **Dans Google Sheets** :
 - Étendue : =MAX() - MIN()
 - Quartiles : =QUARTILE.INC()

Illustration de l'écart interquartile



- L'écart interquartile (IQR) contient les 50% des données centrales
- Il n'est pas affecté par les valeurs extrêmes
- Utile pour détecter les valeurs aberrantes : $x < Q_1 - 1.5 \times IQR$ ou $x > Q_3 + 1.5 \times IQR$

L'écart-type et la variance

- **Variance** : moyenne des carrés des écarts à la moyenne

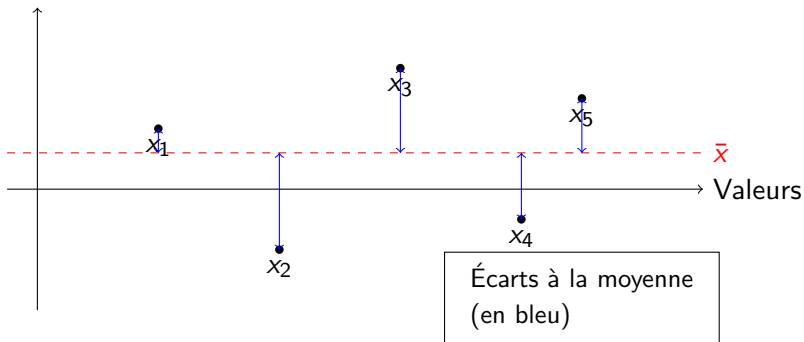
$$s^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

- **Écart-type** : racine carrée de la variance

$$s = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2}$$

- **Propriétés** :
 - Tient compte de toutes les observations
 - Même unité que les données pour l'écart-type
 - Sensible aux valeurs extrêmes

La variance : moyenne des écarts à la moyenne



- La variance est la moyenne des carrés de ces écarts
- Le passage au carré permet d'éviter que les écarts positifs et négatifs s'annulent

Les quantiles

- **Définition** : valeurs qui partagent la distribution en parts égales
- **Principaux quantiles** :
 - Quartiles : 4 parts (25%, 50%, 75%)
 - Déciles : 10 parts (10%, 20%, ..., 90%)
 - Centiles : 100 parts (1%, 2%, ..., 99%)
- **Dans Google Sheets** :
 - =QUARTILE.INC(plage, k)
 - =CENTILE.INC(plage, k)

Indicateurs relatifs

- **Coefficient de variation** : écart-type relatif $CV = \frac{s}{\bar{x}} \times 100\%$
 - Permet de comparer des variables différentes
 - Indépendant de l'unité de mesure
- **Rapport interquartile** : dispersion relative $RIQ = \frac{Q_3}{Q_1}$
 - Mesure l'asymétrie de la distribution
 - Robuste aux valeurs extrêmes

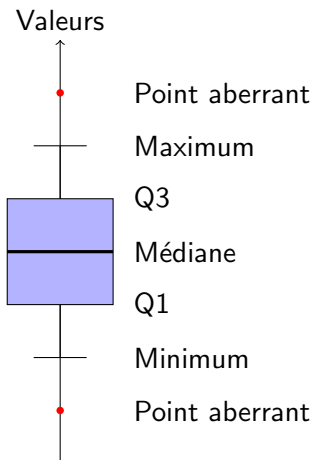
Exercice 3 : Analyse de la dispersion

Objectif : Comparer la dispersion des votes entre candidats

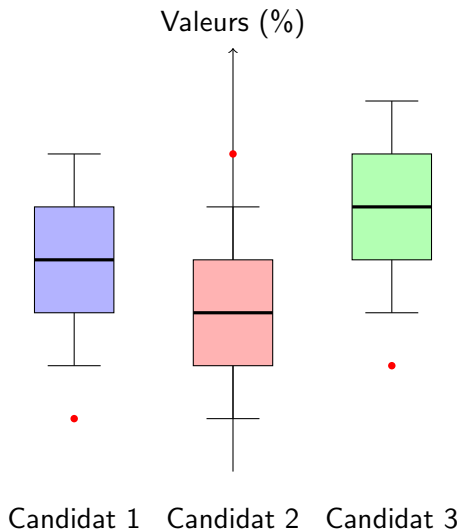
Étapes :

- ① Calculs de base :
 - Écart-type des scores
 - Quartiles
 - Coefficient de variation
- ② Interprétation :
 - Homogénéité/hétérogénéité géographique
 - Identification des valeurs extrêmes
 - Comparaison des candidats

Anatomie d'une boîte à moustaches



Boîtes à moustaches multiples



Boîte à moustaches

- **Composition :**

- Boîte : Q1 à Q3 (50% des données)
- Ligne centrale : médiane
- Moustaches : $1.5 \times IQR$
- Points : valeurs atypiques

- **Dans Google Sheets :**

- Insertion > Graphique
- Type : Boîte à moustaches
- Configuration des axes et légendes

Exercice pour vous entrainer : télécharger les données à un niveau plus bas, et produire le boxplot du vote pour un candidat, mais département par département.

Interprétation globale

Questions à explorer :

- **Comparaison des dispersions :**
 - Quel candidat a le vote le plus homogène ?
 - Quelles régions sont les plus atypiques ?
- **Implications politiques :**
 - Forces et faiblesses territoriales
 - Identification des bastions
 - Zones de fragilité
- **Pour aller plus loin :**
 - Comparaison avec d'autres élections
 - Analyse des facteurs explicatifs