Statistiques descriptives et Introduction à R

Partie 1 : Statistiques univariées

Edmond Sanou

Université d'Evry Val d'Essonne

Mars 2023

Sommaire

Sommaire

Sommaire

Objectifs du cours

- Familiarisation avec le vocabulaire de la statistique
- Description d'une série statistique
- Représentation graphiques de série statistique
- Résumés numériques d'une série statistique

Sommaire

Statistique versus statistiques

- La **statistique** renvoie à la discipline
- Une statistique ou les statistiques renvoient à un chiffre (indicateurs, résumés statistiques)

Statistique versus statistiques

- La **statistique** renvoie à la discipline
- Une statistique ou les statistiques renvoient à un chiffre (indicateurs, résumés statistiques)

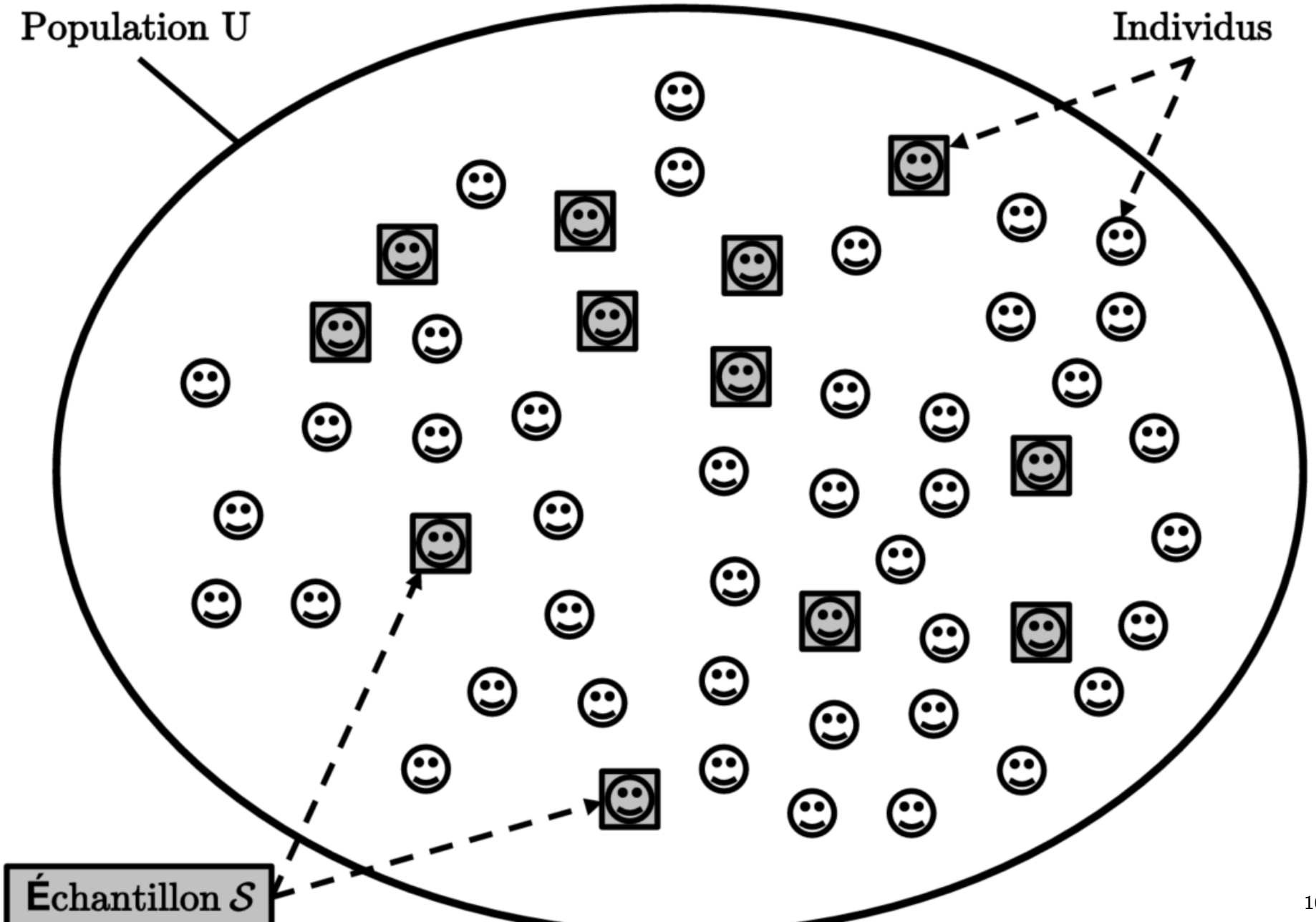
Vocabulaire

- On appelle **population** l'ensemble sur lequel porte l'actiité statistique.
  Exemples : étudiants en licence de maths de l'Université d'Évry en 2023, banques françaises au 1er janvier 2021, ...
- On appelle **unité statistique** ou **individu statistique** un objet qui présente une valeur pour un caractère étudié.
  Exemples : un étudiant inscrit en mathématiques à l'université d'Évry en 2023, une banque française au 1er janvier 2021, ...
- On appelle **échantillon** un sous-ensemble de la population.
  Exemples : un groupe d'étudiants, un petit nombre de banques, ...
- On appelle **cardinal** d'un échantillon ou d'une population la taille de cet échantillon ou de cette population.
- On appelle **variable** une caractéristique de l'individu à laquelle l'étude s'intéresse.
  Exemples : l'âge, le salaire, le sexe, la catégorie socio-professionnelle, la nationalité, le niveau d'étude, la note obtenue en statistiques descriptive, le statut fumeur ou non fumeur, ...

Vocabulaire

- On appelle **population** l'ensemble sur lequel porte l'actiité statistique.
  Exemples : étudiants en licence de maths de l'Université d'Évry en 2023, banques françaises au 1er janvier 2021, ...
- On appelle **unité statistique** ou **individu statistique** un objet qui présente une valeur pour un caractère étudié.
  Exemples : un étudiant inscrit en mathématiques à l'université d'Évry en 2023, une banque française au 1er janvier 2021, ...
- On appelle **échantillon** un sous-ensemble de la population.
  Exemples : un groupe d'étudiants, un petit nombre de banques, ...
- On appelle **cardinal** d'un échantillon ou d'une population la taille de cet échantillon ou de cette population.
- On appelle **variable** une caractéristique de l'individu à laquelle l'étude s'intéresse.
  Exemples : l'âge, le salaire, le sexe, la catégorie socio-professionnelle, la nationalité, le niveau d'étude, la note obtenue en statistiques descriptive, le statut fumeur ou non fumeur, ...

Vocabulaire (2)



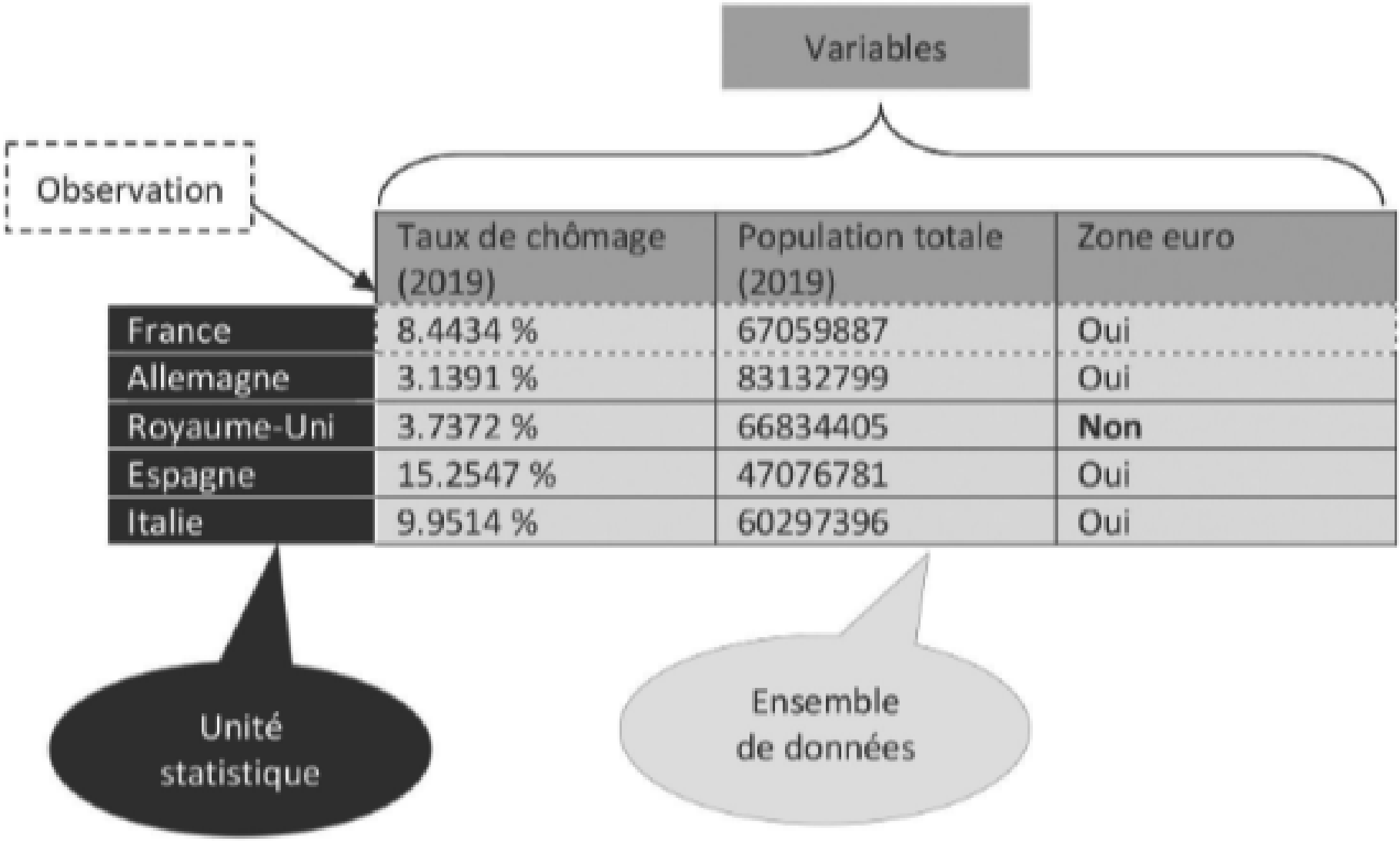
Représentation des données

Les données se présentent sous la forme d'un tableau où

- les lignes correspondent aux individus ou unités statistiques
- les colonnes correspondent à des variables

La cellule (i, j) contient la valeur de la variable j pour l'individu i.

Représentation des données (2)



Source : Banque mondiale (WDI, 2021)

Types de variables

Il existe deux types de variables :

- les **variables qualitatives**
- les **variables quantitatives**

Types de variables : variables qualitatives

Une **variable qualitative** est une variable dont les valeurs, appellées modalités, ne sont pas mesurables mais codifiées ou qualifiées. On distingue deux types de variables qualitatives :

- Les **variables qualitatives ordinales** qu'il est possible d'ordonner
  Exemples : évaluatioon d'un enseignement (pas satisfait, moyennement satisfait, ...)
- Les **variables qualitatives nominales** pour lesquelles l'ordre n'a pas de sens.
  Exemples : le genre, catégorie socio-professionnelle, nationalité,

Types de variables : variables qualitatives

Une **variable qualitative** est une variable dont les valeurs, appellées modalités, ne sont pas mesurables mais codifiées ou qualifiées. On distingue deux types de variables qualitatives :

- Les **variables qualitatives ordinales** qu'il est possible d'ordonner
  Exemples : évaluatioon d'un enseignement (pas satisfait, moyennement satisfait, ...)
- Les **variables qualitatives nominales** pour lesquelles l'ordre n'a pas de sens.
  Exemples : le genre, catégorie socio-professionnelle, nationalité,

Types de variables : variables quantitatives

Une **variable quantitative** est une variable présentant des valeurs numériques.
On distingue deux types de variales quantitatives :

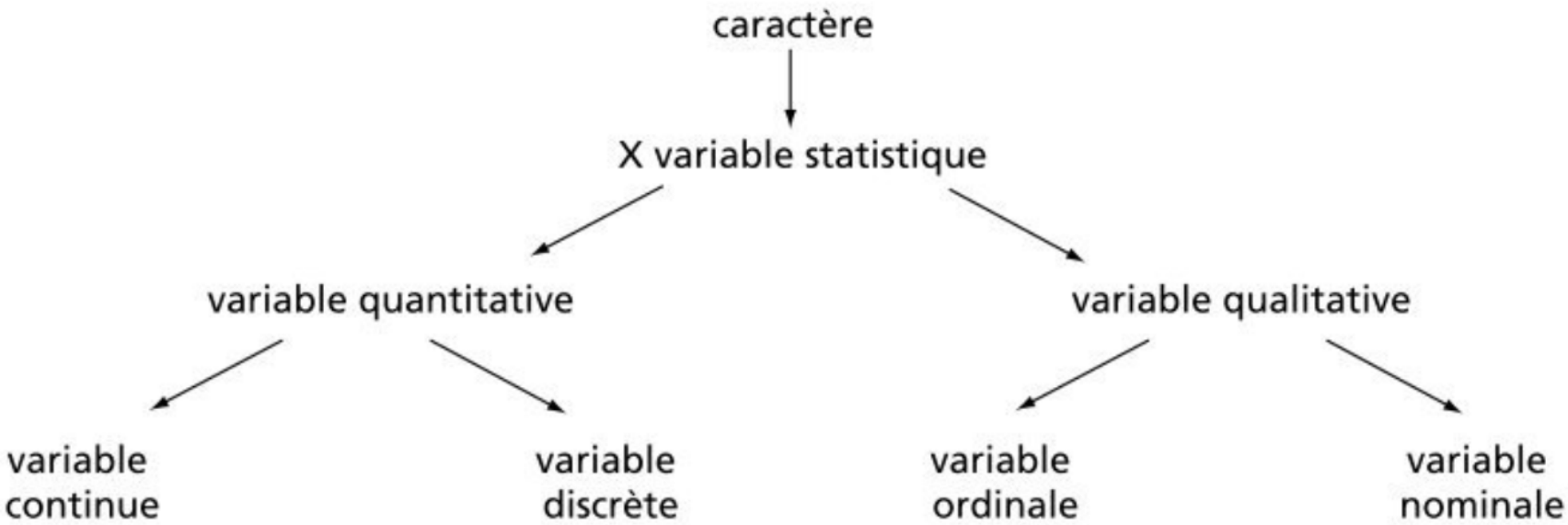
- Les **variables quantitatives discrètes**, souvent des valeurs entières
  Exemples : le nombre d'enfant dans un ménage, ...
- Les **variables quantitatives continues**,
  Exemples : la masse d'un individu, l'âge en mois ou en années, le salaire, la note obtenue en statistiques, ...

Types de variables : variables quantitatives

Une **variable quantitative** est une variable présentant des valeurs numériques.
On distingue deux types de variales quantitatives :

- Les **variables quantitatives discrètes**, souvent des valeurs entières
  Exemples : le nombre d'enfant dans un ménage, ...
- Les **variables quantitatives continues**,
  Exemples : la masse d'un individu, l'âge en mois ou en années, le salaire, la note obtenue en statistiques, ...

Types de variables : résumé



Exercice :

- Une variable avec des modalités numériques est-elle pour autant de nature numérique ?
- Peut-on transformer une variable quantitative en variable qualitative ?
- Echantillonnage d'une population de peupliers :
  Voici une liste de variables disponibles dans une étude portant les propriétés d'adaptation des peupliers à la sécheresse :
  - l'âge de l'arbre,
  - la région géographique,
  - les températures journalières,
  - l'intensité de la sécheresse (sévère, modérée, zone 100% irriguée),
  - le type de feuilles observé,
  - la présence/absence d'une infection chez la plante,
  - le niveau d'expression d'un gène (impliqué dans la résistance à une maladie de la plante).

    Quel est le type de chaque variable ?

Exercice :

- Une variable avec des modalités numériques est-elle pour autant de nature numérique ?
- Peut-on transformer une variable quantitative en variable qualitative ?
- Echantillonnage d'une population de peupliers :
  Voici une liste de variables disponibles dans une étude portant les propriétés d'adaptation des peupliers à la sécheresse :
  - l'âge de l'arbre,
  - la région géographique,
  - les températures journalières,
  - l'intensité de la sécheresse (sévère, modérée, zone 100% irriguée),
  - le type de feuilles observé,
  - la présence/absence d'une infection chez la plante,
  - le niveau d'expression d'un gène (impliqué dans la résistance à une maladie de la plante).

  Quel est le type de chaque variable ?

Exercice :

- Une variable avec des modalités numériques est-elle pour autant de nature numérique ?
- Peut-on transformer une variable quantitative en variable qualitative ?
- Echantillonnage d'une population de peupliers :
  Voici une liste de variables disponibles dans une étude portant les propriétés d'adaptation des peupliers à la sécheresse :
  - l'âge de l'arbre,
  - la région géographique,
  - les températures journalières,
  - l'intensité de la sécheresse (sévère, modérée, zone 100% irriguée),
  - le type de feuilles observé,
  - la présence/absence d'une infection chez la plante,
  - le niveau d'expression d'un gène (impliqué dans la résistance à une maladie de la plante).

  Quel est le type de chaque variable ?

Exercice R :

- Menu `Packages` : automatise la gestion et le suivi de bibliothèques de fonctions, permettant leur installation et leur mise à jour.
-

Exercice R :

- Installer la bibliothèque `BioStatR`

```
install.packages("BioStatR")
```

- Charger la bibliothèque

```
library(BioStatR)
```

- Afficher le jeu de données `Mesures`

```
Mesures
```

- Afficher les premières lignes

```
head(Mesures)
head(Mesures, 10)
```

- Afficher les dernières lignes

```
tail(Mesures)
```

Exercice R :

- Installer la bibliothèque `BioStatR`

```
install.packages("BioStatR")
```

- Charger la bibliothèque

```
library(BioStatR)
```

- Afficher le jeu de données `Mesures`

```
Mesures
```

- Afficher les premières lignes

```
head(Mesures)
head(Mesures, 10)
```

- Afficher les dernières lignes

```
tail(Mesures)
```

Exercice R :

- Description du fichier

```
str(Mesures)
```

- La classe `factor`

```
class(Mesures$espece)
levels(Mesures$espece)
```

Série statistique

Une **série statistique** est une suite de valeurs observées d'un caractère d'intérêt X.

- On note la série statistique $(x_1, x_2, \ldots, x_n)$ ou $\{x_1, x_2, \ldots, x_n\}$
- On parlera aussi de **vecteur des observations**

Sommaire

Sommaire

Tableaux statistiques et graphiques

Un tableau ou un graphique doit toujours avoir

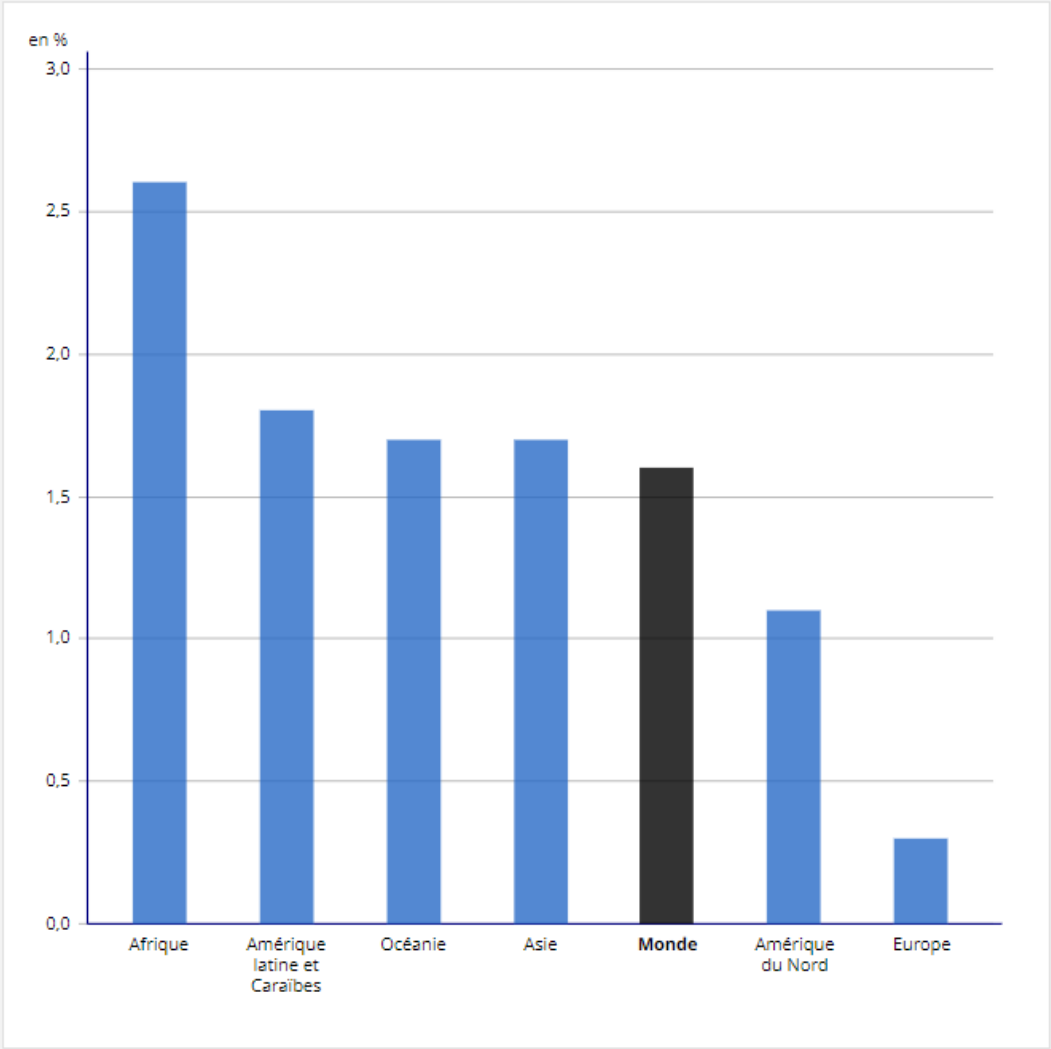
- Un titre (descriptif et informatif)
- La source, la date et le champ des données (éventuellement)
- Une note de lecture (éventuellement)
- Une référence dans le texte (éventuellement)

Remarques

- Ne pas oublier les légendes sur les axes et les unités
- Eviter les graphiques 3D et bien choisir les couleurs
- Les tableaux et graphiques doivent pouvoir être compris sans lire le texte.

Exemples de graphiques

Évolution de la population dans le monde entre 1960 et 2021



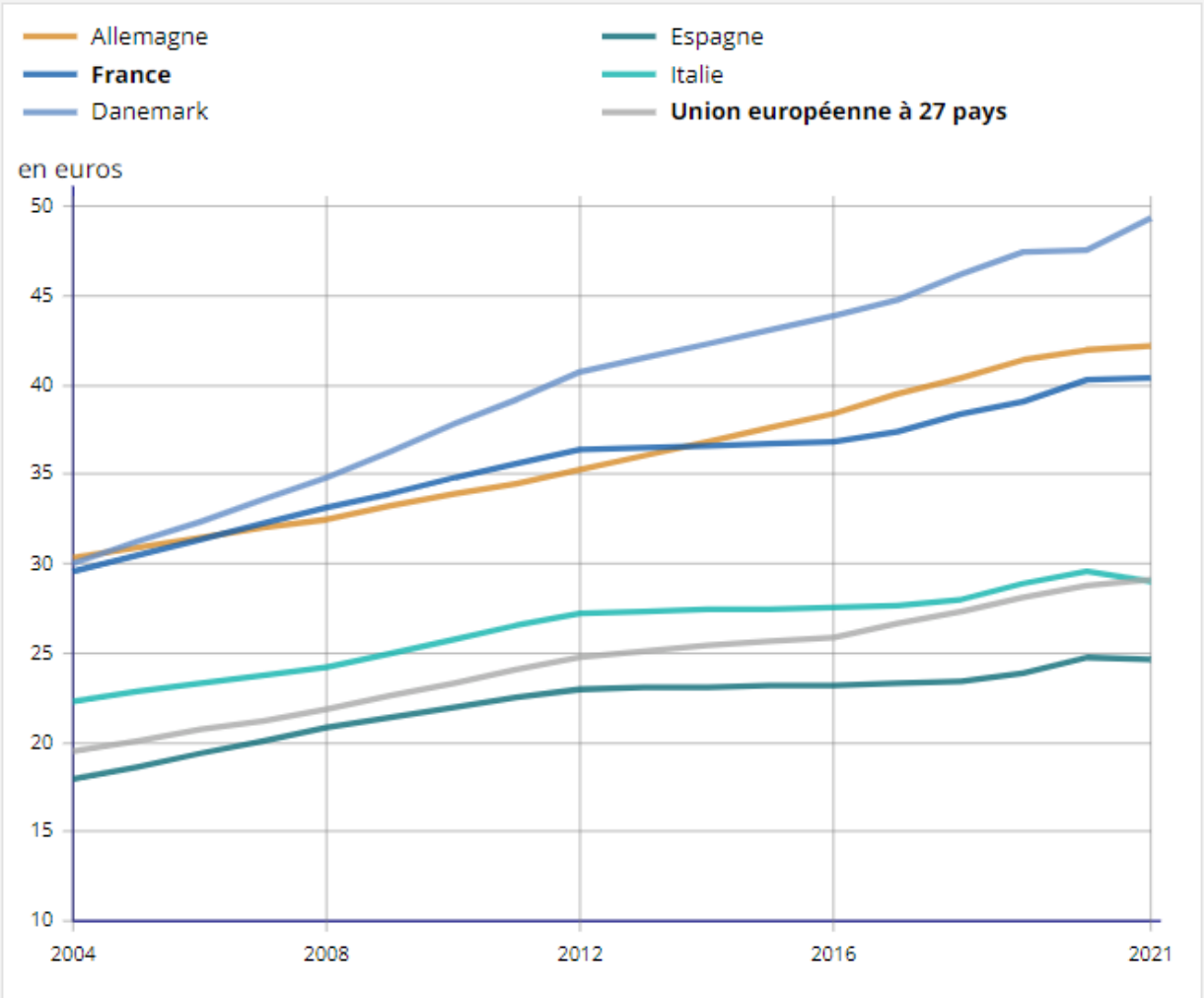
Note : évolution annuelle moyenne 2021/1960.

Lecture : entre 1960 et 2021, la population mondiale a augmenté de 1,6 % en moyenne par an.

Source : ONU (World Population Prospects 2022).

Exemples de graphiques (2)

Figure 2a - Coût de la main-d'œuvre dans l'industrie entre 2004 et 2021



Lecture : en 2021, dans l'Union européenne à 27 pays, le coût horaire de la main d'œuvre dans l'industrie pour les entreprises de 10 salariés ou plus s'élève en moyenne à 29,1 euros.

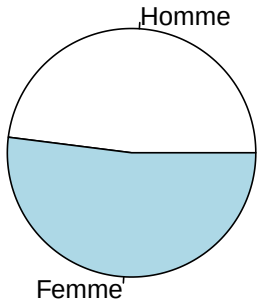
Champ : Union européenne à 27 pays, industrie (hors construction), entreprises de 10 salariés ou plus.

Source : Eurostat, annual labour cost data.

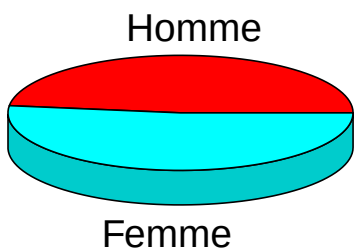
Exemples de graphiques (3)

Répartition hommes/femmes dans un échantillon

Bonne pratique



A éviter



Sommaire

Définition

On appelle **distribution** d'une variable qualitative X sa répartition en fonction de ses différentes modalités, autrement dit, le nombre d'observations de la série (effectif) pour chaque modalité du caractère.

- Il est souvent préférable de représenter les effectifs relatifs (ou fréquences) en mentionnant si possible l'effectif total de la série
- On distingue deux types de représentations, le tableau ou le graphique.

Notations

- Soit X une variable qualitative à k modalités $m_1, \ldots, m_k$
- Soit $(x_1, x_2, \ldots, x_n)$ la série statistique associée.
- On note n_j le nombre d'individus prenant la modalité m_j de X
- On a donc $n_j = \#\{i = 1, \ldots, n / x_i = m_j\}, \quad \forall j = 1, \ldots, k$ avec $\sum_{j=1}^{k} n_j = n$

La distribution de X est la suite des effectifs $(n_j), j = 1, \ldots, k$ ou des fréquences $(f_j), j = 1, \ldots, k$ avec $f_j = \frac{n_j}{n}$.

Notations (2)

Modalité de la variable	Effectifs	Fréquences
m_1	n_1	$f_1 = \frac{n_1}{n}$
m_2	n_2	$f_2 = \frac{n_2}{n}$
$\vdots$	$\vdots$	$\vdots$
m_j	n_j	$f_j = \frac{n_j}{n}$
$\vdots$	$\vdots$	$\vdots$
m_k	n_k	$f_k = \frac{n_k}{n}$
Total	n	1

Exemple

Intérêt pour une carte Ticket Restaurant

Salarié	Réponse	Salarié	Réponse
1	Très intéressé (e)	11	Assez intéressé (e)
2	Assez intéressé (e)	12	Assez intéressé (e)
3	Très intéressé (e)	13	Pas du tout intéressé (e)
4	Assez intéressé (e)	14	Pas du tout intéressé (e)
5	Peu intéressé (e)	15	Assez intéressé (e)
6	Très intéressé (e)	16	Très intéressé (e)
7	Assez intéressé (e)	17	Peu intéressé (e)
8	Très intéressé (e)	18	Très intéressé (e)
9	Peu intéressé (e)	19	Pas du tout intéressé (e)
10	Peu intéressé (e)	20	Très intéressé (e)

- Identifier la variable et ses modalités
- Quantifier le nombre d'individus qui partagent la même modalité
- Calculer les fréquences associées

Exemple sur R

- Calculer le tableau des effectifs de la variable `espece` dans le jeu de données `Mesures`.

```
table(Mesures$espece)
```

- Calculer les fréquences associées

```
table(Mesures$espece) / nrow(Mesures)

prop.table(table(Mesures$espece))
```

Représentation à l'aide d'un diagramme en colonnes (barres)

- Diagramme en barres verticales ou horizontales
- Suite de segments verticaux d'abscisses m_j dont la longueur est proprtionnelle à l'effectif (respectivement la fréquence) de la modalité m_j.

Représentation à l'aide d'un diagramme en colonnes (barres)

- Tracer le diagramme en barres des effectifs pour la variable espece du fichier Mesures

```
plot(table(Mesures$espece),
     lwd = 4,
     col = "red",
     xlab = "Nombre d'arbustes",
     ylab = "Effectif")
```

- Tracer le diagramme en barres des proportions pour la variable espece du fichier Mesures

```
plot(prop.table(table(Mesures$espece)),
     lwd = 4,
     col = "red",
     xlab = "Nombre d'arbustes",
     ylab = "Fréquences")
```

Représentation à l'aide d'un diagramme circulaire (camembert)

- Disque d'aire décomposée en secteurs circulaires représentant respectivement la part de chaque modalité
- L'angle au centre α_j pour la modalité m_j est donnée par $\alpha_i = 360 \times f_j$.
- Ne pas utiliser cette représentation lorsque la variable possède beaucoup de modalités

Représentation à l'aide d'un diagramme circulaire (camembert) (2)

- Représenter à l'aide d'un camembert la variable espece du jeu de données `Mesures`.

```
pie(table(Mesures$especе),
    col = c("green", "purple", "cyan", "blue"))
```