Analyse univariée (partie 1)

Vincent Audigier
vincent.audigier@lecnam.net

CNAM, Paris

STA101 2019-2020

Terminologie

- Population : groupe d'individus soumis à une étude
- Individu (statistique) : élément issu de la population
- Echantillon : partie d'une population

Pour chaque individu, on observe un ensemble de caractères $X_1, X_2, \ldots, X_j$ appelés **variables**

La valeur de la j variable observée sur le i-ème individu est notée x_{ij}

Typologie des variables

- variable qualitative : variable à valeurs non-numériques (où la moyenne n'a pas de sens). Ses valeurs sont appelées **modalités**
  - nominale (ou catégorielle) : absence d'ordre entre les modalités
  - ordinale : existence d'un ordre total
- variable quantitative : variable à valeurs numériques (où la moyenne a un sens)
  - continue : à valeurs dans un intervalle réel
  - discrète : dans le cas contraire

Données *ozone*

- Données climatiques et de pollution à l'ozone mesurées durant l'été 2001 à Rennes (112 individus)

	maxO3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	maxO3v	vent	pluie
20010601	87	15.60	18.50	18.40	4	4	8	0.69	-1.71	-0.69	84	Nord	Sec
20010602		17.00	18.40	17.70	5	5	7	-4.33	-4.00	-3.00	87	Nord	Sec
20010603	92	15.30	17.60	19.50	2	5	4	2.95	1.88	0.52	82	Est	
20010604	114	16.20	19.70	22.50	1		0	0.98			92		Sec
20010605	94	17.40	20.50	20.40	8	8	7	-0.50	-2.95	-4.33	114	Ouest	Sec
20010606	80	17.70	19.80	18.30	6	6	7	-5.64	-5.00	-6.00		Ouest	Pluie
...	...	...	...	...	...	...	...	...	...	...	...	...	

- maxO3, maxO3v : maximum d'ozone journalier et maximum de la veille
- T9, T12, T15 : température à 9h, 12h, 15h
- Ne9, Ne12, Ne15 : nébulosité à 9h, 12h, 15h
- Vx9 , Vx12, Vx15 : force du vent à 9h, 12h, 15h
- vent : direction du vent
- pluie : présence de pluie

Analyse univariée, bivariée, multivariée

- Analyse univariée : la description porte sur chacune des variables
- Analyse bivariée : la description porte sur des couples de variables
- Analyse multivariée : la description porte sur l'ensemble des variables du jeu de données

Analyse univariée

- L'analyse univariée d'une variable s'effectue différemment selon que celle-ci soit de nature qualitative (ordonnée ou non) ou quantitative (discrète ou continue)

- Ce type d'analyse est indispensable pour avoir une première idée de la distribution des variables, ainsi que pour identifier de potentielles "anomalies" dans les données

- Elle s'effectue par la présentation de tableaux ou de graphiques spécifiques

Variables qualitatives

On note $\mathcal{E} = \{m_1, ..., m_k\}$ l'ensemble des k modalités de la variable. Cette variable est observée sur un échantillon de n individus

On appelle **fréquence absolue** de la modalité m_q, le nombre total (effectif) n_q d'individus de l'échantillon pour lesquels la variable a pris la modalité m_q

$$n_q = \sum_{i=1}^{n} \mathbf{1}_{m_q}(x_i)$$

On appelle **fréquence relative** de la modalité m_q, la proportion d'individus à présenter cette modalité

$$f_q = \frac{n_q}{n}$$

Variables qualitatives : tableau

On s'intéresse à la variable vent du jeu ozone

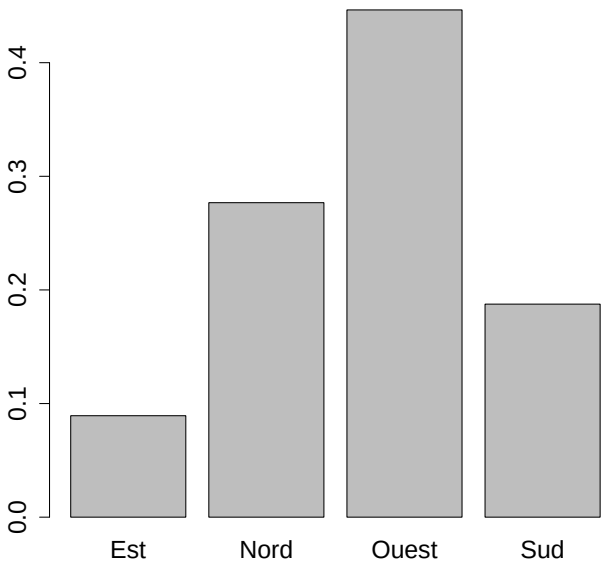
- $\mathcal{E} = \{\text{Est, Nord, Ouest, Sud}\}$
- La variable est résumée par le tableau suivant

	Est	Nord	Ouest	Sud	Somme
n_q	10	31	50	21	112
f_q (arrondi)	0.09	0.28	0.45	0.19	1.00

Table: Fréquences absolues et relatives pour la variable vent

Variables qualitatives : représentations graphiques

- Diagramme en barres ou “bar-plot” : à chaque modalité m_q, on associe un rectangle vertical dont **la hauteur est proportionnelle à la fréquence relative** f_q
- Diagramme circulaire : **angle proportionnel** à la fréquence relative



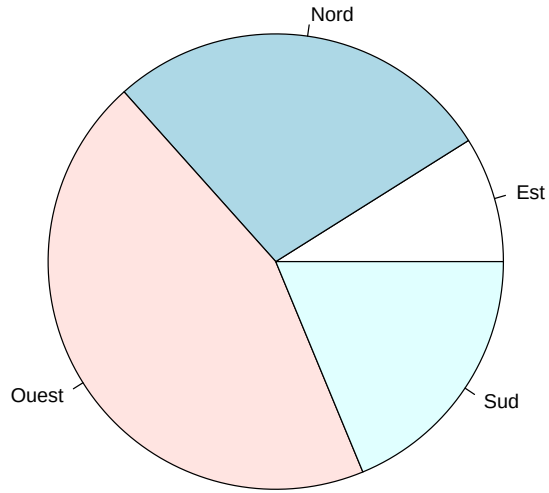


Figure: Diagramme en barres et circulaire pour la variable *vent*

Variables qualitatives ordinales

- Très courantes dans les questionnaires d'enquêtes d'opinion
- Exemple : *Compte tenu du travail que vous fournissez, diriez-vous que vous êtes*
  1. Très mal payé
  2. Plutôt mal payé
  3. Normalement payé
  4. Plutôt bien payé
  5. Très bien payé
- Représentation par tableau ou diagramme en barres en respectant l'ordre naturel des modalités

Variables quantitatives discrètes

- Une variable quantitative est dite **discrète** lorsqu'elle est à valeur dans un sous-ensemble dénombrable de $\mathbb{R}$
  $\mathcal{E} = \{e_1, \ldots, e_\ell\}$
- Exemple : la variable Ne9 est à valeur dans $\{0, 1 \ldots, 8\}$

	0	1	2	3	4	5	6	7	8
n_q	6	13	8	6	10	12	13	25	19
f_q (arrondi)	0.05	0.12	0.07	0.05	0.09	0.11	0.12	0.22	0.17

Table: Fréquences absolues et relatives pour la variable Ne9

$(f_1, \ldots, f_\ell)$ est appélé profil de la variable

Représentation graphique : diagramme en barres

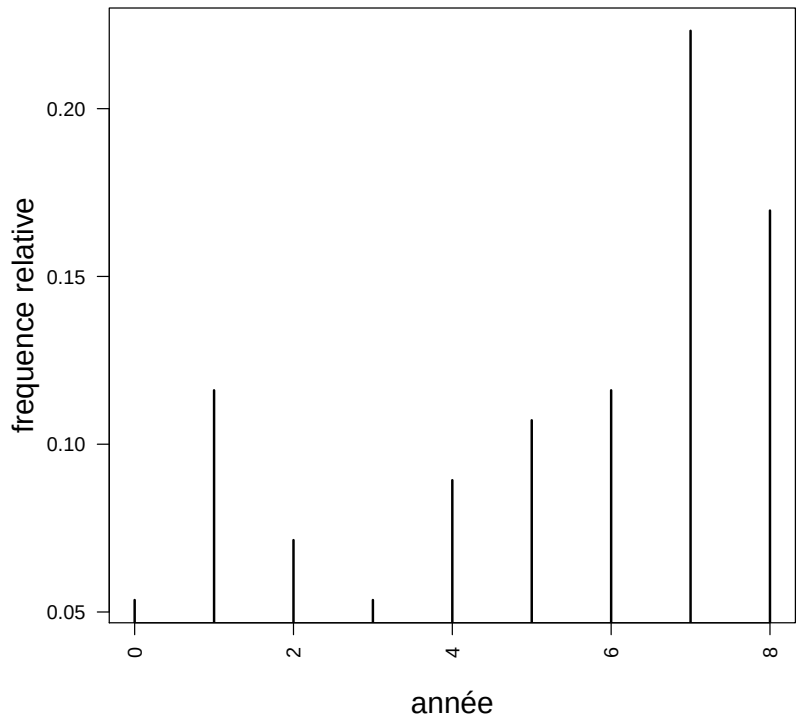


Figure: Diagramme en bâtons pour la variable Ne9

Représentation graphique : diagramme tige-et-feuille

- Diagramme en barre limité en présence d'un grand nombre de valeurs
- Principe de la représentation *tige-et-feuille* : représenter les valeurs elles mêmes en les regroupant par dizaine et en répétant le chiffre des unités en fonction de la fréquence absolue des valeurs

```
 4 | 25
 5 | 456799
 6 | 0033355667777899
 7 | 000011111222245666677778899 9
 8 | 0111233334447888
 9 | 02223467889
10 | 01116689
11 | 123344667
12 | 116
13 | 19
14 | 56699
15 | 3699
16 | 06
```

Figure: Exemple de représentation tige-et-feuille

Fréquences cumulées

- **fréquences absolues cumulées** : Pour $q' = 1, \ldots, \ell$

$$N_{q'} = \sum_{q=1}^{q'} n_q,$$

- **fréquences relatives cumulées** : Pour $q' = 1, \ldots, \ell$

$$F_{q'} = \sum_{q=1}^{q'} f_q,$$

	n_q	f_q	$N_{q'}$	$F_{q'}$
0	6	0.05	6	0.05
1	13	0.12	19	0.17
2	8	0.07	27	0.24
3	6	0.05	33	0.29
4	10	0.09	43	0.38
5	12	0.11	55	0.49
6	13	0.12	68	0.61
7	25	0.22	93	0.83
8	19	0.17	112	1.00

Table: Fréquences absolues et relatives (non cumulées et cumulées) pour la variable Ne9

Représentation graphique

Si la variable admet ℓ valeurs distinctes $\{e_1, ...e_\ell\}$, alors à partir de ses fréquences cumulées, on peut tracer la **fonction de répartition empirique** correspondante

$$F_X(x) = \begin{cases} 0 & \text{si } x < e_1 \\ F_{q'} & \text{si } e_q \leq x < e_{q+1} \\ 1 & \text{si } x \geq e_\ell \end{cases}$$

qui est une fonction en escalier continue à droite et limitée à gauche

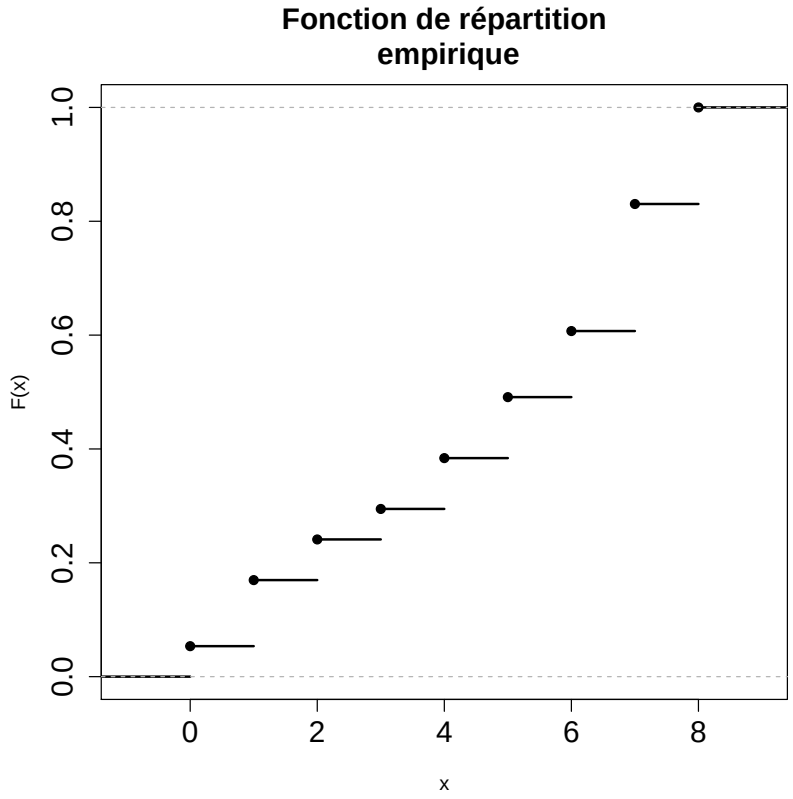


Figure: Fonction de répartition empirique pour la variable *Ne9*

Remarque : on parle de *diagramme cumulatif* pour les effectifs cumulés

Statistique d'ordre

- Une variable quantitative X est dite **continue** lorsqu'elle est à valeur dans $\mathbb{R}$ (ou un intervalle de $\mathbb{R}$).
- $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$
- on note $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ les valeurs ordonnées par ordre croissant
- $X_{(.)} = \begin{pmatrix} x_{(1)} \\ \vdots \\ x_{(n)} \end{pmatrix}$ est appelé **statistique d'ordre**

Exemple

On a mesuré la durée de vie (en h) de 10 ampoules identiques

	1	2	3	4	5	6	7	8	9	10
X	7.4	5.8	5.6	17.4	115.8	49.2	21.6	38.3	5.9	55.6
$X_{(.)}$	5.6	5.8	5.9	7.4	17.4	21.6	38.3	49.2	55.6	115.8

Vocabulaire

La représentation d'une variable quantitative continue nécessite un découpage en **classes**

- **classe** : soit a_0 et a_k deux réels et soit une partition de l'intervalle $]a_0, a_k]$ en k intervalles. On appelle classe tout intervalle de la forme $]a_{q-1}, a_q]$ $(1 \leq q \leq k)$ avec $a_0 < x_{(1)} < x_{(2)} < \ldots < x_{(n)} < a_k$ et $a_0 < a_1 < a_2 < \ldots < a_k$
- **amplitude** : On appelle amplitude (ou longueur) de la classe $]a_{q-1}, a_q]$ la différence $a_q - a_{q-1}$
- **effectif** : on appelle effectif de la classe $]a_{q-1}, a_q]$ le nombre n_q de valeurs de la série contenues dans cette classe

$$n_q = \sum_{i=1}^{n} \mathbf{1}_{]a_{q-1};a_q]}(x_i)$$

- **fréquence** : On appelle fréquence de la classe $]a_{q-1}, a_q]$ la proportion $f_q = n_q/n$ de valeurs de la série contenues dans cette classe

Variables continues : tableau

	n_q	f_q	N'_q	F'_q
(38.9,55.2]	4	0.04	4	0.04
(55.2,71.4]	29	0.26	33	0.29
(71.4,87.7]	32	0.29	65	0.58
(87.7,104]	18	0.16	83	0.74
(104,120]	13	0.12	96	0.86
(120,137]	4	0.04	100	0.89
(137,153]	6	0.05	106	0.95
(153,169.1]	6	0.05	112	1.00

Table: Fréquences absolues et relatives (non cumulées et cumulées) pour la variable max03

$a_0 = 38.9$ et $a_k = 169.1$, $k = 8$, $l = 16.275$

Choix des classes

- Nombre de classes :
  - peu de classes : perte d'information
  - beaucoup de classes : beaucoup de classe vides
  - pas de règle absolue
  - règle de Sturges ($k = 1 + ln(n)/ln(2)$)
- Choix de a_0 et a_k
  - $a_0 = x_{(1)} - 0.025(x_{(n)} - x_{(1)})$
  - $a_k = x_{(k)} + 0.025(x_{(n)} - x_{(1)})$
- Amplitude des classes
  - constante
  - variable (classes à effectif constant)

Histogramme

- La représentation graphique d'une variable continue est l'histogramme (on peut aussi représenter la fonction de répartition)
- L'**histogramme** est la figure constituée des rectangles dont les bases sont les classes et dont les **aires** sont égales aux fréquences de ces classes.
- Si toutes les classes ont même longueur, alors on construit un **histogramme à pas fixe**. Dans le cas contraire, on parle d'**histogramme à pas variable**.
- Le choix des classes (nombre et amplitude) influence l'allure de l'histogramme

Exemple : maxO3

Règle de Sturges : $n = 112$, donc $k = 8$ classes.
Comme $x_{(1)} = 42$ et $x_{(n)} = 166$, on obtient via la règle précédente $a_0 = 38.9$ et $a_k = 169.1$. Pour un histogramme à pas fixe avec 8 classes, on a $l = 16.275$.

	n_q	f_q (arrondi)	$h_q = f_q / l_q$
(38.9,55.2]	4.00	0.04	0.00
(55.2,71.4]	29.00	0.26	0.02
(71.4,87.7]	32.00	0.29	0.02
(87.7,104]	18.00	0.16	0.01
(104,120]	13.00	0.12	0.01
(120,137]	4.00	0.04	0.00
(137,153]	6.00	0.05	0.00
(153,169.1]	6.00	0.05	0.00

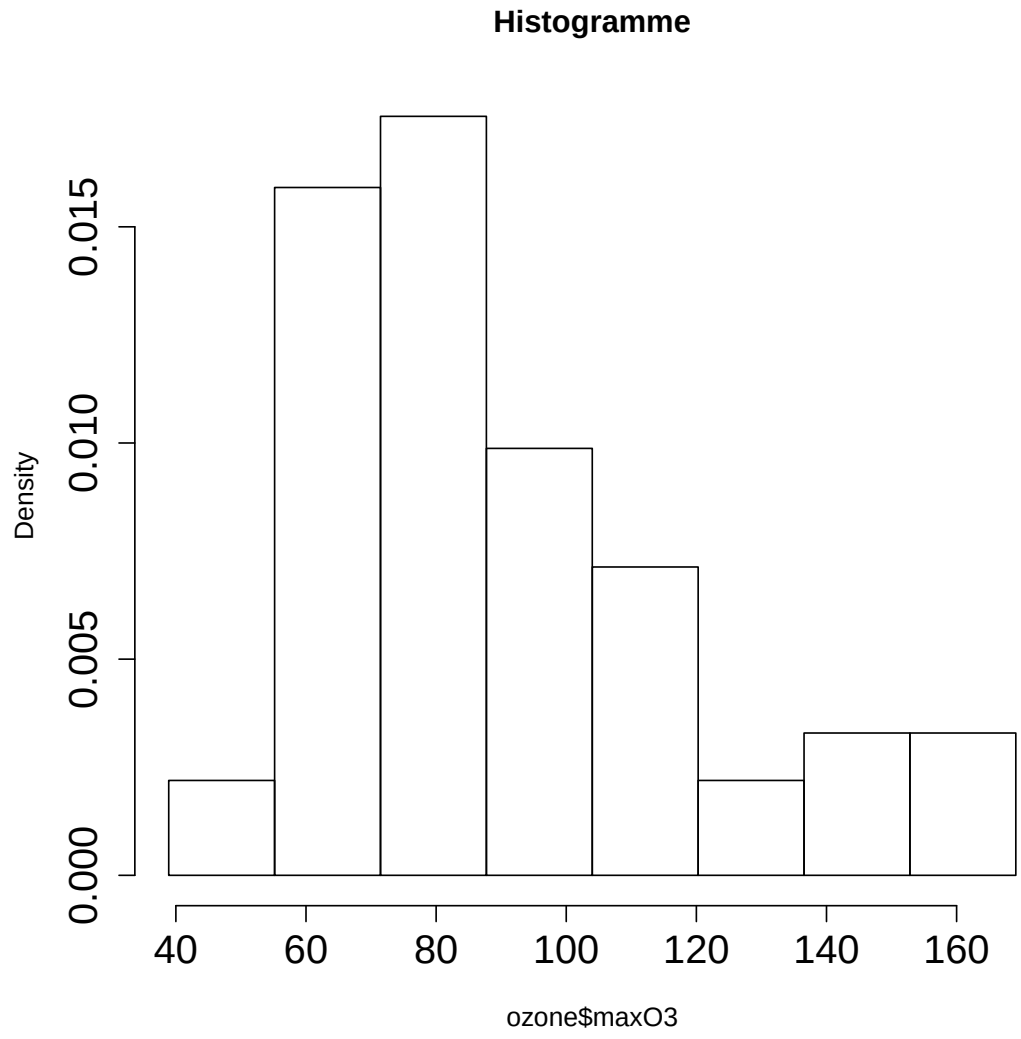


Figure: Histogramme à pas fixe pour la variable max03

Fonction de répartition empirique

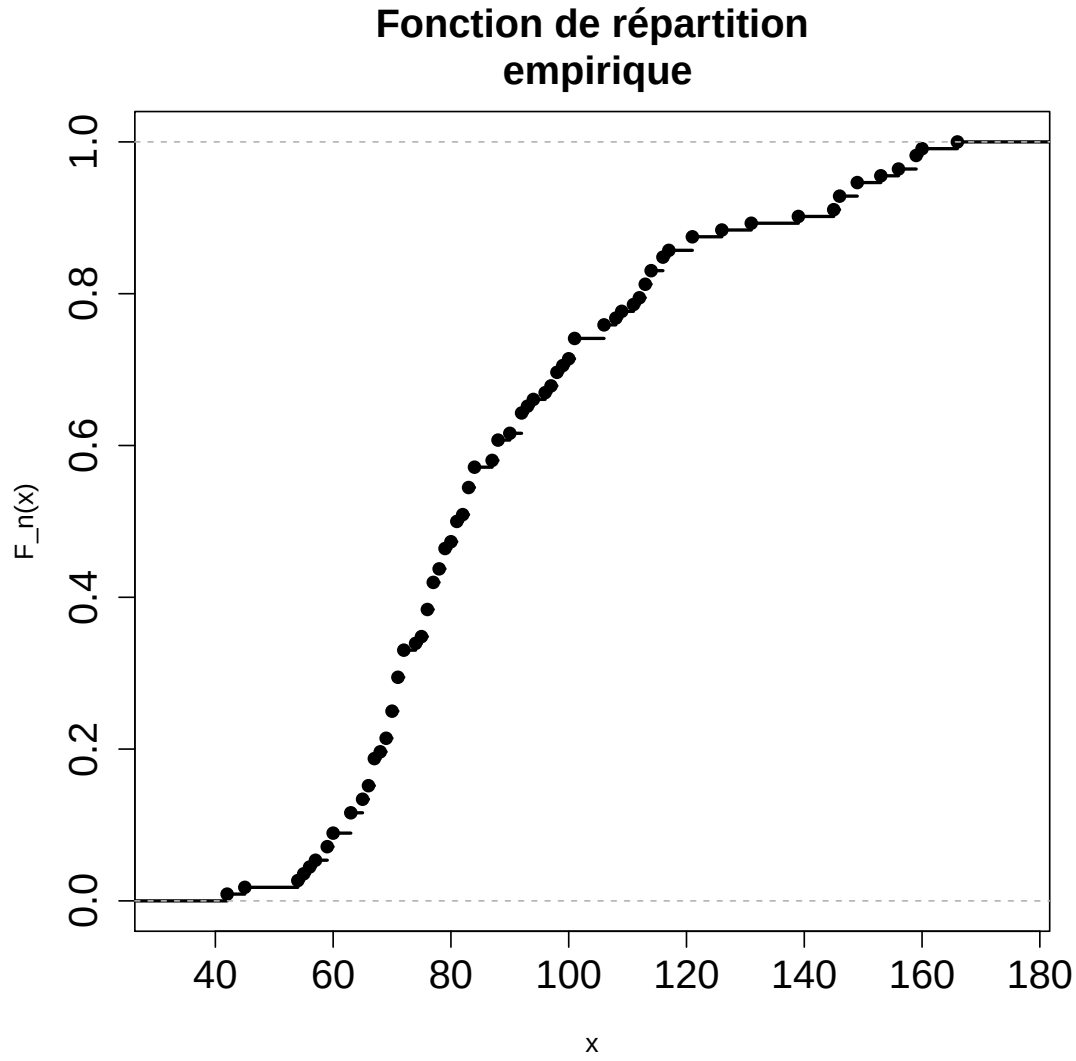


Figure: Fonction de répartition empirique pour la variable maxO3