Descriptions graphiques des distributions

Antonio Falcó

Modélisation

Definition du Modéle

le mot modèle est forgé sur le radical indo-européen **méd**-*d'où ont été dérivés les mots latins* **metiri** *(mesurer),* **modus** *(mesure imposée aux choses),* **modo** *(en restant dans la mesure),* **modestus** *(qui observe la mesure). Fidèle à son étymologie, le modéle se présente avec modestie comme un simple intermédiaire entre le scientifique et son objet d'étude qu'il aide à appréhender.*

Jean-Gabriel Ganascia *Le mythe de la Singularité: Faut-il craindre l'intelligence artificielle?* Editions du Seuil (2017).

La statistique s'intéresse á des populations. Le terme population est à comprendre dans un sens élargi.

Noté par Ω (ou population statistique): ensemble (au sens mathématique du terme) concerné' par une étude statistique. On parle parfois de champ de l'étude.

Exemples de populations:

- Les habitants d'une ville, d'une région, d'un pays.
- Les voitures qui circulent dans un pays
- L'ensemble des séjours hospitaliers pendant une année dans un hôpital
- L'ensemble des jets possibles d'une pièce de monnaie.

Les éléments d'une population sont appelés des unités d'observation.

On parle aussi d'individu qu'on note par $\omega \in \Omega$: tout élément de la population.

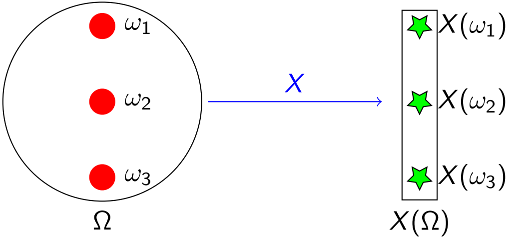
Ils peuvent être de différentes natures. Dans les exemples ci-dessus, on trouve les types suivants:

- Des personnes
- Des objets (voitures)
- Des unités abstraites (séjours hospitaliers, jets d'une piéce de monnaie)

Les unités d'observation possèdent des caractéristiques:

- Habitants: âge, nombre d'enfants, sexe, état de santé.
- Voitures: couleur, kilométrage, nombre de roues.
- Séjours hospitaliers: durée en jours, spécialité, coût.
- Jets d'une pièce: côté (pile ou face), bruit.

Ces caractéristiques sont appelées des variables qu'on note par X, Y, Z (car leur valeur varie d'une unité d'observation à l'autre). Les valeurs possibles d'une variable sont appelées ses modalités, qu'on note $X(\omega), Y(\omega), Z(\omega)$ avec $\omega \in \Omega$.

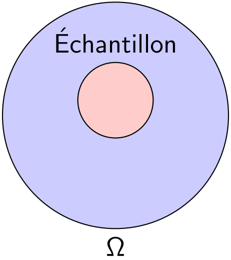

ω_1
ω_2
ω_3
Ω
X
$X(\omega_1)$
$X(\omega_2)$
$X(\omega_3)$
$X(\Omega)$

On distingue plusieurs types de variables:

- variable quantitative: les modalités sont des nombres qui expriment des quantités.
  - variable quantitative continue: les modalités sont des nombres réels, elles ne sont pas dénombrables (ex.: poids, taille)
  - variable quantitative discrète: les modalités sont dénombrables: nombres entiers, demi-entiers, etc (ex.: durées de déjours hostpitaliers en jours ou en demi-journées, nombre de fréres et soeurs)
- variable qualitative: les modalités ne sont pas des quantités numériques
  - variable qualitative catégorielle: les modalités sont des qualités (ex.: couleur des yeux, lieu de naissance)
  - variable qualitative ordinale: les modalités sont desqualités pouvant être ordonnées (ex.: qualité d'un film, d'un livre (bon, moyen, mauvais)).

En général, la population est trop grande pour qu'on puisse l'observer en entier, et on devra alors tirer un échantillon.

Échantillon: sous–ensemble de la population sur lequel sont effectivement réalisées les observations.



Enquête (statistique): opération consistant à observer (ou mesurer, ou questionner ...) l'ensemble des individus d'un échantillon.
Taille de l'échantillon n: cardinal du sous-ensemble correspondant.

Données (statistiques): ensemble des individus observés (échantillon), des variables considérées, et des observations de ces variables sur ces individus. Elles sont en général présentées sous forme de tableaux (individus en lignes et variables en colonnes) et stockées dans un fichier informatique. Lorsqu'un tableau ne comporte que des nombres (valeurs des variables quantitatives ou codes associés aux variables qualitatives), il correspond à la notion mathématique de matrice.

1	sexe	situation	the	cafe	taille	poids	age	viande	poisson	fruit_crus	fruit_legun	chocol	matgras
2	2	1	0	0	151	58	72	4	3	1	4	5	6
3	2	1	1	1	162	60	68	5	2	5	5	1	4
4	2	1	0	4	162	75	78	3	1	5	2	5	4
5	2	1	0	0	154	45	91	0	4	4	0	3	2
6	2	1	2	1	154	50	65	5	3	5	5	3	2
7	2	1	2	0	159	66	82	4	2	5	5	1	3
8	2	1	2	0	160	66	74	3	3	5	5	5	6
9	2	1	0	2	163	66	73	4	2	5	5	1	6
10	2	1	0	3	154	60	89	4	3	5	5	5	6
11	2	1	0	2	160	77	87	2	3	5	4	0	3
12	2	1	0	2	175	68	91	5	2	5	5	5	3
13	2	1	2	0	165	75	81	5	2	2	3	0	6
14	2	1	0	3	158	53	89	4	2	5	5	5	4
15	2	1	2	0	155	63	79	3	1	5	3	2	4
16	2	1	0	2	154	80	83	3	3	5	5	1	3
17	1	1	0	3	166	80	78	5	0	5	3	0	4
18	2	1	0	2	159	57	74	3	2	5	3	1	4
19	2	1	0	2	157	55	74	3	2	5	5	2	3
20	1	1	0	1	165	57	78	5	2	5	5	5	4
21	2	1	0	0	156	90	78	5	1	5	5	2	4
22	1	2	0	1	175	90	73	5	1	5	3	2	8
23	2	2	2	2	161	68	76	4	2	5	5	1	1
24	1	2	2	2	168	83	85	4	2	5	5	1	1
25	1	1	0	4	168	90	73	3	2	4	5	0	3
26	2	2	3	1	156	56	77	3	2	5	3	0	4
27	1	2	3	2	170	70	74	3	2	5	3	0	4

Tableaux des donneés $\equiv$ Matrices

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{im} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}$$

Lorsque l'on observe uniquement des variables numériques le tableau a la forme d'une matrice à n-lignes et m-colonnes de terme général $x_{ij} = X_j(\omega_i)$ il représente la **codification numérique** de la modalité associé à la variable X_j chez l'individu ω_i.

Notation matrice-vecteur

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{im} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots \mathbf{x}_m \end{pmatrix}$$

ou

$$\mathbf{x}_1 = \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{i1} \\ \vdots \\ x_{n1} \end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{i2} \\ \vdots \\ x_{n2} \end{pmatrix}, \ldots, \mathbf{x}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{ij} \\ \vdots \\ x_{nj} \end{pmatrix}, \ldots, \mathbf{x}_m = \begin{pmatrix} x_{1m} \\ x_{2m} \\ \vdots \\ x_{im} \\ \vdots \\ x_{nm} \end{pmatrix}$$

On adopera alors les notations suivantes:

- n pour la taille de l'échantillon
- Lettres majuscules pour les variables. Ex.: A pour l'âge, C pour la couleur des yeux.
- Lettres minuscules pour les valeurs observées des variables dans l'échantillon. Certaines de ces valeurs peuvent être égales. Ex.: $c_1, c_2, \ldots, c_n$ pour les couleurs des yeux des n individus de l'échantillon.
- Attention: on utilise la même notation pour désigner les modalités d'une variable. Toutes les modalités sont différentes. Ex.: $c_1 =$ brun, $c_2 =$ bleu, $c_3 =$ vert, $c_4 =$ noir, $c_5 =$ gris.

Exemple: étudiant(e)s de 1ére année

- Population: Ensemble des étudiant(e)s de 1ére année à CEU-UCH en 2010
- Unités d'observation: Chaque étudiant
- Variables:
  - Sexe, noté S: qualitative catégorielle
  - Taille en cm, notée T : quantitative continue
  - Poids en kg, noté P : quantitative continue
  - Nombre de fréres et soeurs, noté F, quantitative discrète
  - Couleur des yeux, notée C, qualitative catégorielle
- Modalités:
  - Sexe: $\{\text{femme}, \text{homme}\}$
  - Taille en cm: $[40, 280]$
  - Poids en kg: $[20, 400]$
  - Nombre de fréres et soeurs: $\{0, 1, \ldots, 50\}$
  - Couleur des yeux: $\{\text{brun}, \text{bleu}, \text{vert}, \text{noir}, \text{gris}\}$
- On a tiré un échantillon de taille $n = 45$.

Données:

T	P	S	F	C
180	70	h	2	brun
177	57	h	3	brun
180	60	h	1	bleu
180	66	h	0	brun
183	62	h	6	vert
184	68	h	0	brun
185	65	h	1	noir
184	72	h	2	brun
174	65	h	3	noir
180	72	h	1	brun
168	52	h	3	brun
180	75	h	0	bleu
183	75	h	2	brun
181	68	h	0	bleu
180	65	h	4	brun

T	P	S	F	C
190	66	h	1	brun
183	78	h	0	bleu
167	60	h	4	bleu
181	67	h	0	brun
179	98	h	2	brun
173	75	h	1	vert
170	68	h	1	gris
170	59	h	3	brun
183	72	h	2	bleu
179	73	h	3	vert
180	72	h	3	bleu
188	70	h	2	brun
176	65	h	1	vert
178	72	h	1	brun
185	71	h	1	bleu

T	P	S	F	C
168	52	f	0	brun
157	47	f	1	vert
167	53	f	2	vert
168	57	f	4	bleu
163	65	f	1	brun
167	60	f	2	brun
166	68	f	2	bleu
164	49	f	7	vert
172	57	f	3	brun
165	59	f	2	bleu
158	62	f	0	brun
161	65	f	1	brun
160	61	f	1	bleu
162	58	f	2	brun
165	58	f	5	brun

Soit X une variable qualitative et $\{x_1, x_2, ..., x_k\}$ l'ensemble de ses modalités. Pour un échantillon de taille n, soit n_i le nombre d'individus ayant la modalité x_i. On appelle

- fréquence absolue de x_i le nombre n_i
- fréquence relative de x_i le nombre $f_i = ni/n$
- distributionde fréquence de X l'ensemble des couples (x_i, n_i) ou des couples (x_i, f_i)

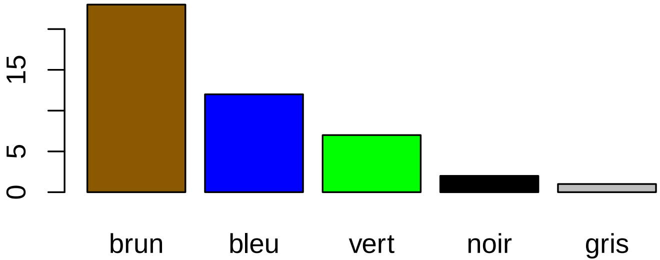
Exemple: distribution de fréquence de la variable couleur des yeux.

Modalité (c_i)	Fréquence absolue (n_i)	Fréquence relative $(f_i = n_i/n)$
brun	32	$0.7111 = 71.11$ %
bleu	12	$0.2667 = 26.67$ b%
vert	7	$0.1556 = 15.56$ %
noir	2	$0.04444 = 4.444$ %
gris	1	$0.02222 = 2.222$ %
Totaux	$n = 45$	$1 = 100\%$

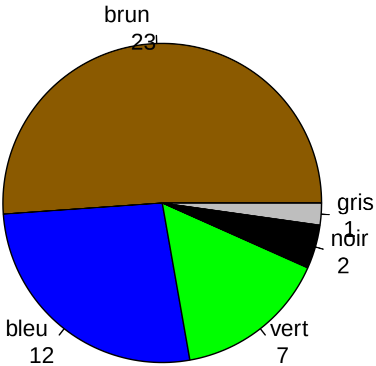
Propiétés: $\sum_{i=1}^{k} n_i = n_1 + n_2 + \cdots + n_k = n$ et $\sum_{i=1}^{k} f_i = f_1 + f_2 + \cdots + f_k = 1$.

Pour représenter graphiquement une distribution de fréquence, on peut utiliser

- un diagramme à barres:



- un diagramme en secteurs:



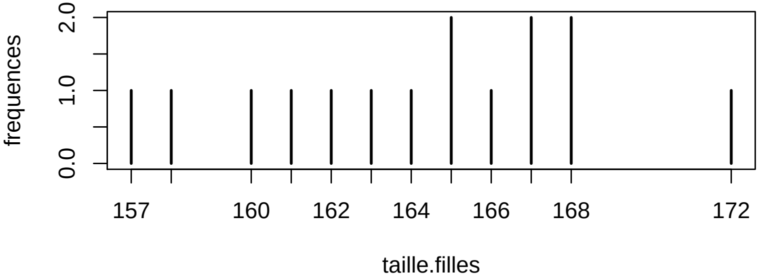
Nous allons distinguer trois cas:

1. Le nombre d'observations est petit ($n < 20$)
2. Le nombre d'observations différentes est petit
3. Le nombre d'observations est grand avec beaucoup d'observations différentes

1. Le nombre d'observations est petit ($n < 20$)

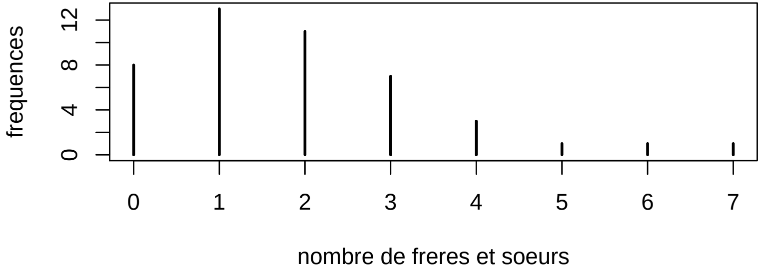
Ex.: Tailles des filles dans notre échantillon d'étudiant(e)s

Dans ce cas on peut simplement représenter les données sur un axe. Cette représentation permet de se faire une idée rapide de la forme de la distribution (symétrie, etc) et de repérer des éventuelles observations aberrantes (appelées outliers).



2. Le nombre d'observations différentes est petit

Ex.: Nombre de fréres et soeurs dans notre échantillon d'étudiant(e)s
Dans ce cas on procéde de façon similaire au cas d'une variables qualitative, avec un diagramme en barres qui tient compte de l'ordre des modalités.



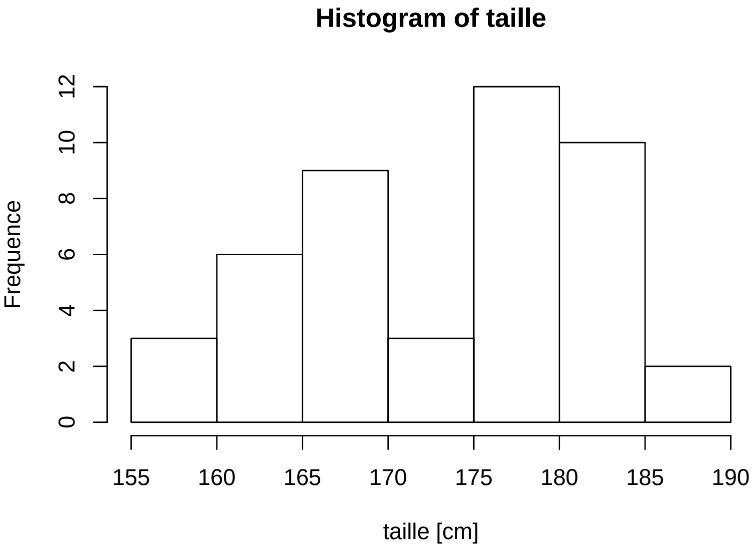
3. Le nombre d'observations est grand avec beaucoup d'observations différentes

Ex.: Tailles des étudiant(e)s

Dans ce cas on regroupe les données en classes de largeurs égales. On construit un graphique similaire à un diagramme en barres, où la hauteur des barres est égale au nombre d'observations dans la classe correspondante.

En régle générale, le nombre classes est compris entre 5 et 20.

Le graphique obtenu s'appelle un histogramme.

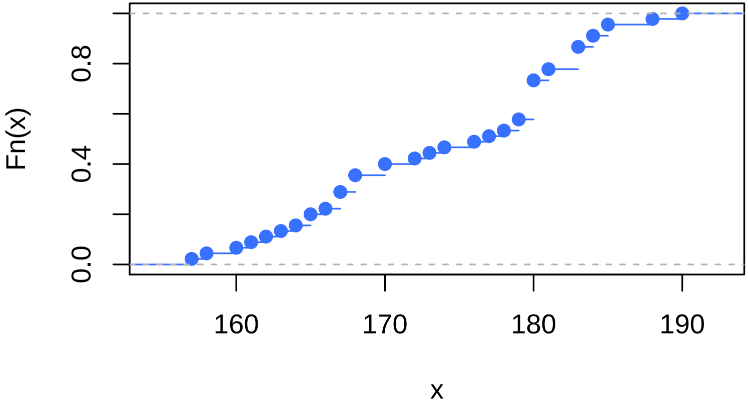

Histogram of taille
Frequence
0
2
4
6
8
10
12
155
160
165
170
175
180
185
190
taille [cm]

Fonction de distribution cumulative empirique

Pour des observations $x_1, ..., x_n$ d'une variable X, la fonction de distribution cumulative empirique, notée $F_n(x)$ est définie par

$$F_n(x) := \frac{\text{nombre de } x_i\text{'s} \leq x}{n}$$

Ex.: Tailles des étudiant(e)s Une fonction de distribution cumulative commence toujours à 0 et finit toujours à 1. Elle est toujours croissante.



La fonction de distribution cumulative empirique de la variable taille

x	$F_n(x)$

	x
157	0.0222222
158	0.0444444
160	0.0666667
161	0.0888889
162	0.1111111
163	0.1333333
164	0.1555556
165	0.2000000
166	0.2222222
167	0.2888889
168	0.3555556
170	0.4000000
172	0.4222222
173	0.4444444
174	0.4666667
176	0.4888889
177	0.5111111
178	0.5333333
179	0.5777778
180	0.7333333
181	0.7777778
183	0.8666667
184	0.9111111
185	0.9555556
188	0.9777778
190	1.0000000

La forme de la fonction de distribution cumulative est en général moins facile à interpréter que celle de l'histogramme.
Par contre, la fonction de distribution cumulative est utile pour certains calculs. Par exemple, pour trouver la proportion d'invividus mesurant entre 165 et 180 cm, il suffit de calculer

$$F_n(180) - F_n(165) = 0.73 - 0.20 = 0.53\,(53\%).$$