Statistique descriptive en SHS

Chapitre 2 : Graphes et distribution cumulée

Département MIAp - UFR 6 - UPV-UM3

Licence 1re année

Introduction

Il s’agit de **visualiser graphiquement** la **répartition des individus** sur les modalités de la variable.

A chaque type de variables correspond un graphe particulier.

Sommaire

Cas d'une variable qualitative nominale

La distribution de la variable X est fournie par le tableau :

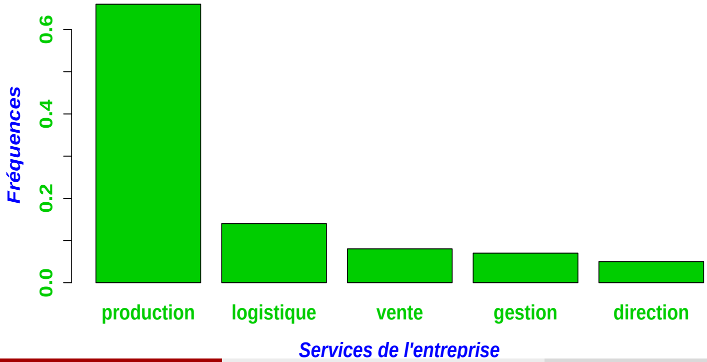
Variable	X				
Modalités m_k	m_1	m_2	...	m_C	Total
Effectifs n_k	n_1	n_2	...	n_C	n
Fréquences f_k	f_1	f_2	...	f_C	1

$\hookrightarrow$ **Distribution - diagramme en barres séparées**

Cas d'une variable qualitative nominale

Exemple : répartition dans les services de l'entreprise

Service	production	logistique	vente	gestion	direction
Effectifs	66	14	8	7	5
Fréquences	0.66	0.14	0.08	0.07	0.05



1. Graphe de la distribution
   - Cas d'une variable qualitative nominale
   - Cas d'une variable qualitative ordinale
   - Cas d'une variable quantitative discrète
   - Cas d'une variable quantitative continue

Cas d'une variable qualitative ordinale

La distribution de la variable X est fournie par le tableau :

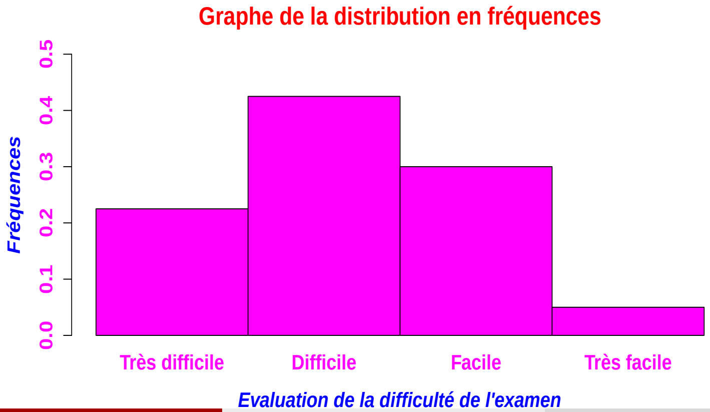
Variable	X				
Modalités m_k	m_1	m_2	...	m_C	Total
Effectifs n_k	n_1	n_2	...	n_C	n
Fréquences f_k	f_1	f_2	...	f_C	1

$\hookrightarrow$ **Distribution** - **diagramme en barres juxtaposées**
car nous envisageons ici l'*échelle continue sous-jacente* à la variable ordinale.

Cas d'une variable qualitative ordinale

Exemple : Évaluation de la difficulté d'un examen

Difficulté exam.	Très difficile	Difficile	Facile	Très facile	Total
Effectifs	9	17	12	2	40
Fréquences	0.225	0.425	0.300	0.050	1



1. Graphe de la distribution
   - Cas d’une variable qualitative nominale
   - Cas d’une variable qualitative ordinale
   - **Cas d’une variable quantitative discrète**
   - Cas d’une variable quantitative continue

Cas d'une variable quantitative discrète

Le tableau de la distribution d'une variable quantitative discrète présente donc les valeurs de la variable et les effectifs et fréquences correspondant(e)s :

Variable	X				
Valeurs v_k	v_1	v_2	$\ldots$	v_C	Total
Effectifs n_k	n_1	n_2	$\ldots$	n_C	n
Fréquences f_k	f_1	f_2	$\ldots$	f_C	1

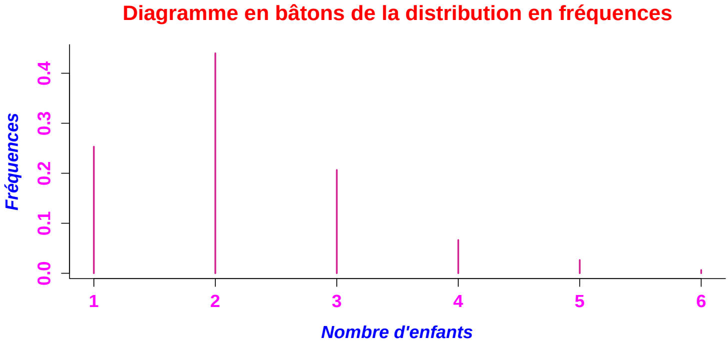
Exemple : Nombre d'enfants dans les familles

Variable	Nombre d'enfants						
Valeurs v_k	1	2	3	4	5	6	Total
Effectifs n_k	38	66	31	10	4	1	150
Fréquences f_k	0.253	0.440	0.207	0.067	0.027	0.007	1

Cas d'une variable quantitative discrète

Pour une variable quantitative discrète, les valeurs sont isolées les unes des autres.

$\hookrightarrow$ **Représentation graphique de la distribution** - **diagramme en bâtons**



Question : Pourquoi ne pas faire de barres mais bien des bâtons ?

1. Graphe de la distribution

Cas d'une variable quantitative continue

Comme pour toutes les variables, l'information sur les valeurs observées sont :

- soit des données brutes,
- soit un tableau de distribution.

Mais à la différence de toutes les autres variables, **il ne peut pas en théorie y avoir de valeurs répétées** dans l'observation d'une variable quantitative continue.
$\hookrightarrow$ pour regrouper les valeurs, on doit donc construire des **classes**
$\hookrightarrow$ on **perd alors de l'information** sur les valeurs réelles observées

Attention : Ce regroupement en classes est nécessaire pour construire une représentation graphique de la distribution.

Cas d'une variable quantitative continue

↪ Représentation graphique de la distribution - histogramme

La distribution de la variable X est fournie par :

Variable	X					
Classes	$[b_0 ; b_1[$	$[b_1 ; b_2[$	$\ldots$	$[b_{C-1} ; b_C[$		Total
Effectifs n_k	n_1	n_2	$\ldots$	n_C		n
Fréquences f_k	f_1	f_2	$\ldots$	f_C		1

Attention : Pour la représentation graphique de cette distribution, il est nécessaire de tenir compte des **amplitudes** des classes (ou **largeurs** des classes).

En effet, affirmer que 10 individus mesurent entre 155 cm et 165 cm ne signifie pas la même chose que d'affirmer que 10 individus mesurent entre 160 et 162 cm. Les effectifs sont identiques (les fréquences aussi) mais dans le second cas, il y a une **concentration** beaucoup plus forte d'observations.

Cas d'une variable quantitative continue

Nous allons donc calculer pour chaque classe :

- sa **densité de fréquence** :

$$\mathbf{d_k = \frac{f_k}{a_k}}$$

- on de manière équivalente (proportionnelle), sa **densité d'effectif**

$$\mathbf{d_k = \frac{n_k}{a_k}}$$

$\hookrightarrow$ que vaut le coefficient de proportionnalité entre les deux ?

On complète alors le tableau :

Variable	X				
Variable X	$[b_0; b_1[$	$[b_1; b_2[$	...	$[b_{C-1}; b_C[$	Total
Effectifs n_k	n_1	n_2	...	n_C	n
Fréquences f_k	f_1	f_2	...	f_C	1
Amplitudes a_k	a_1	a_2	...	a_C	
Densités d_k	d_1	d_2	...	d_C	

Cas d'une variable quantitative continue

Exemple : Prix des produits d'une grande surface

On donne dans le tableau ci-dessous la distribution et la distribution cumulée (en effectifs et en fréquences) de la variable X="Prix".

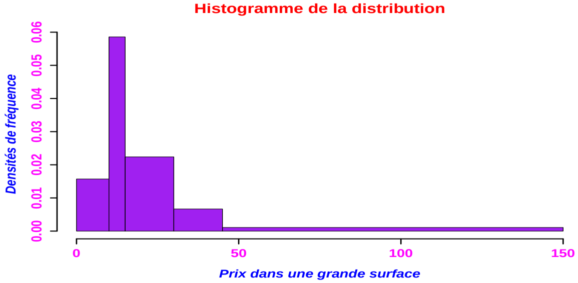
Variable	Prix					
Classes	]0 ; 10[	[10 ; 15[	[15 ; 30[	[30 ; 45[	[45 ; 150]	Tot.
n_k	110	205	235	70	80	700
f_k (%)	15.7	29.3	33.6	10.0	11.4	100

auquel on rajoute les amplitudes et densités :

Variable	Prix					
Classes	]0 ; 10[	[10 ; 15[	[15 ; 30[	[30 ; 45[	[45 ; 150]	Total
n_k	110	205	235	70	80	700
f_k (%)	15.7	29.3	33.6	10.0	11.4	100
a_k	10	5	15	15	105	/
d_k	1.57	5.86	2.24	0.67	0.109	/

Cas d'une variable quantitative continue

Exemple : Prix des produits d'une grande surface :



Remarque :

↬ Que vaut la surface totale d'un histogramme quand il est représenté en densité d'effectifs ?

↬ Que vaut la surface totale d'un histogramme quand il est représenté en densité de fréquences ?

Sommaire

2 Distribution cumulée (et graphe associée)
- Distribution cumulée d'une variable ordinale
- Distribution cumulée d'une variable discrète
- Distribution cumulée d'une variable continue

Distribution cumulée d'une variable ordinale

Le calcul des effectifs ou fréquences cumulés vise à évaluer la quantité d'individus "rencontrés" depuis la première modalité.

$\hookrightarrow$ pour une modalité fixée, on cherche à connaître la quantité d'individus sur toutes les modalités qui la précèdent et jusqu'à elle.

Attention : pour cela les modalités doivent être **ordonnées naturellement**.
$\hookrightarrow$ c'est le cas d'une **variable qualitative ordinale**
$\hookrightarrow$ mais on ne parlera donc **pas de distribution cumulée dans le cas d'une variable qualitative nominale**.

Pour une **variable qualitative ordinale**, on peut donc cumuler les effectifs et les fréquences pour construire la **distribution cumulée**.

Distribution cumulée d'une variable ordinale

Exemple : Évaluation de la difficulté d'un examen

Difficulté exam.	Très difficile	Difficile	Facile	Très facile	Total
Effectifs	9	17	12	2	40
Fréquences	0.225	0.425	0.300	0.050	1
Eff. cum.	9	26	38	40	
Fréq. cum.	0.225	0.650	0.950	1	

Distribution cumulée d'une variable ordinale

De façon générale pour une variable qualitative ordinale, les modalités sont rangées selon un ordre naturel :

$$m_1 < m_2 < ... < m_k < ... < m_C$$

Variable	X				
Modalités m_k	m_1	m_2	...	m_C	Total
Effectifs n_k	n_1	n_2	...	n_C	n
Fréquences f_k	f_1	f_2	...	f_C	1
Eff. cum. N_k	$N_1 = n_1$	$N_2 = n_1 + n_2$	...	$N_C = n$	
Fréq. cum. F_k	$F_1 = f_1$	$F_2 = f_1 + f_2$	...	$F_C = 1$	

$\hookrightarrow$ **Distribution cumulée** - L'échelle n'étant pas numérique, nous ne faisons **pas de représentation graphique** de cette distribution cumulée.

2. Distribution cumulée (et graphe associée)
   - Distribution cumulée d'une variable ordinale
   - Distribution cumulée d'une variable discrète
   - Distribution cumulée d'une variable continue

Distribution cumulée d'une variable discrète

Les valeurs étant numériques, elles peuvent être bien sûr rangées :

$$v_1 < v_2 < ... < v_k < ... < v_C$$

On forme alors le tableau contenant distribution et distribution cumulée :

Variable	X				
Valeurs v_k	v_1	v_2	$\dots$	v_C	Total
Effectifs n_k	n_1	n_2	$\dots$	n_C	n
Fréquences f_k	f_1	f_2	$\dots$	f_C	1
Eff. cum. N_k	$N_1 = n_1$	$N_2 = n_1 + n_2$	$\dots$	$N_C = n$	
Fréq. cum. F_k	$F_1 = f_1$	$F_2 = f_1 + f_2$	$\dots$	$F_C = 1$	

Attention : les cumuls se font au niveau des valeurs puisque les individus se situent exactement sur chaque valeur.

Distribution cumulée d'une variable discrète

Exemple : Nombre d'enfants dans les familles

Variable	Nombre d'enfants					
Valeurs v_k	1	2	3	4	5	6
Eff. cum. N_k	38	104	135	145	149	150
Fréq. cum. F_k	0.253	0.693	0.900	0.967	0.993	1

Distribution cumulée d'une variable discrète

À l'aide des fréquences cumulées, on définit **la fonction F** pour **n'importe quelle valeur x de $\mathbb{R}$** par :
"**la fréquence** des individus ayant une valeur **inférieure ou égale à x**".

Ainsi cette fonction est telle que :

$$\begin{array}{ll} \forall k \in \{1, .., C\} & F(v_k) = F_k \\ \forall x < v_1 & F(x) = 0 \\ \forall x \in [v_k ; v_{k+1}[& F(x) = F_k \\ \forall x \geq v_C & F(x) = 1 \end{array}$$

Exemple : Nombre d'enfants dans les familles

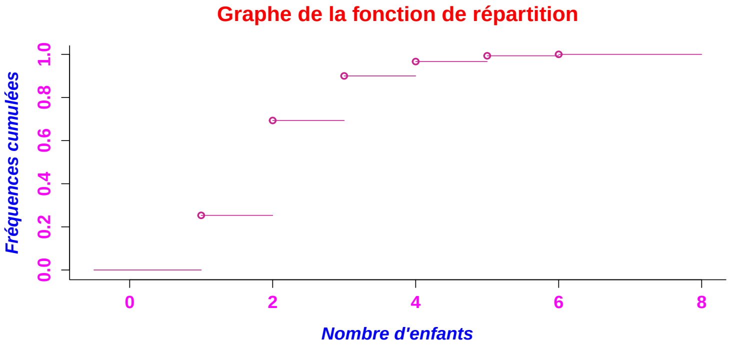
$$F(2.8) = 0.693$$

$\Rightarrow$ Cette fonction est appelée **fonction de répartition empirique**

$\hookrightarrow$ **Représentation graphique de la distribution cumulée** - **fonction en escalier**.

Distribution cumulée d'une variable discrète

Distribution cumulée :



Distribution cumulée d'une variable discrète

Cette fonction permet ensuite de répondre à des questions telles que :

- quel est le pourcentage d'individus dont la réponse est **plus petite que ...** ?
- combien d'individus ont une valeur **plus grande que** ... ?
- quelle est la fréquence d'individus dont la valeur est **comprise entre ... et ...** ?

2. Distribution cumulée (et graphe associée)
   - Distribution cumulée d’une variable ordinale
   - Distribution cumulée d’une variable discrète
   - Distribution cumulée d’une variable continue

Distribution cumulée d'une variable continue

Attention : N'ayant pas d'information sur les valeurs exactes observées, il n'y a qu'au niveau de la borne supérieure d'une classe que l'on est sûr d'avoir accumulé tous les individus de la classe !
$\hookrightarrow$ on rajoute donc effectifs cumulés et fréquences cumulées au niveau des bornes de classe.

Variable	X				
Variable X	$[b_0 ; b_1[$	$[b_1 ; b_2[$	$\ldots$	$[b_{C-1} ; b_C[$	Total
Effectifs n_k	n_1	n_2	$\ldots$	n_C	n
Fréquences f_k	f_1	f_2	$\ldots$	f_C	1

Bornes	b_0	b_1	b_2	$\ldots$	b_{C-1}	b_C
Eff. cum. N_k	0	N_1	N_2	$\ldots$	N_{C-1}	n
Fréq. cum. F_k	0	F_1	F_2	$\ldots$	F_{C-1}	1

$\hookrightarrow$ on trace alors la **fonction de répartition empirique** par un **graphe linéaire par morceaux**.

Distribution cumulée d'une variable continue

Exemple : Prix des produits d'une grande surface

On donne dans le tableau ci-dessous la distribution et la distribution cumulée (en effectifs et en fréquences) de la variable X="Prix".

Variable	Prix					
Classes	]0 ; 10[	[10 ; 15[	[15 ; 30[	[30 ; 45[	[45 ; 150]	Tot.
n_k	110	205	235	70	80	700
f_k (%)	15.7	29.3	33.6	10.0	11.4	100

Bornes	0	10	15	30	45	150
N_k	0	110	315	550	620	700
F_k (%)	0	15.7	45	78.6	88.6	100

Distribution cumulée d'une variable continue

La représentation graphique de la distribution cumulée :