CHAPITRE 1

Statistique descriptive : séries simples

La statistique descriptive consiste à **présenter et résumer des données**, récoltées sur une **population**, afin d'en rendre l'exploitation plus efficace.

Un élément de la population étudiée est appelé une unité statistique ou **individu**. Pour chaque individu on observe un ou plusieurs **caractères**. Dans une population donnée de taille N tout sous-ensemble de taille $n \leqslant N$ est un **échantillon**.

1.1 Séries statistiques

Quand on étudie un seul caractère, on parle de **série statistique simple** et quand on en étudie deux on parle de **série statistique double**. Chaque caractère a différentes **modalités** possibles : par exemple, pour la couleur, les modalités peuvent être rouge, jaune...
Les caractères sont :

- **qualitatifs** (nationalité, profession, couleur, sexe...)
- **quantitatifs**, lorsque les modalités sont mesurables (âge, taille...) Pour un caractère quantitatif, on parle des **valeurs** du caractère.

Caractère qualitatif

Les modalités d'un caractère qualitatif ne sont pas mesurables mais peuvent être décrites :

- dans une **échelle nominale** : les modalités sont exprimables par des noms et ne sont pas hiérarchisées. C'est le cas par exemple du sexe, de la couleur, des groupes sanguins... Remarquons aussi que des caractères comme le code postal, ou le numéro de sécurité sociale sont des variables qualitatives (bien qu'elles soient écrites avec des chiffres), car elles ne représentent pas des quantités.

- dans une **échelle ordinale** : les modalités traduisent le degré d'un état caractérisant un individu sans que ce degré soit défini par un nombre résultant d'une mesure. Les modalités sont alors hiérarchisées. Par exemple, lors d'une enquête de satisfaction, un client pourra être très satisfait, satisfait, peu satisfait ou insatisfait.

Soit x un caractère qualitatif. On notera par x_i ses modalités et par n_i l'effectif de chaque modalité. On a $\sum_i n_i = N$.

La **fréquence** relative à chaque modalité est $f_i = \frac{n_i}{N}$

Exemple

Un hypermarché a fait une enquête sur la satisfaction de ses clients à propos des horaires d'ouverture et de fermeture du magasin. Les résultats sont donnés dans le tableau 1.1. Les fréquences sont calculées dans la dernière ligne de ce tableau :

x_i	*Pas satisfait*	*Peu satisfait*	*Satisfait*	*Très satisfait*	*Total*
n_i	80	110	150	160	500
f_i	0.16	0.22	0.3	0.32	1

Tableau 1.1 : Enquête de satisfaction dans un hypermarché

Caractère quantitatif discret

Les caractères **discrets** prennent un nombre fini[1] de valeurs (qui sont toutes connues). Ce sont par exemple le nombre de frères ou sœurs, le nombre d'appareils électroménagers chez un couple...
On note par x_i l'une des valeurs prises et par n_i le nombre d'individus prenant cette valeur. On dit que n_i est **l'effectif** associé à la modalité x_i. On parle de la série :

$$(x_i, n_i), \text{ pour } i = 1, \ldots, k \tag{1.1}$$

si il y a k valeurs prises.

Exemple

On considère la liste des notes obtenues lors d'un concours pour $N = 50$ candidats. Soit (x_i, n_i) la série obtenue où : x_i est la note, n_i est le nombre de candidats ayant obtenu cette note, et i l'indice qui varie de 1 à 10. La série est représentée dans le tableau :

x_i	5	7	9	10	11	12	13	14	15	18
n_i	1	2	7	8	7	9	5	6	2	3

Tableau 1.2 : Notes obtenues lors d'un concours

Il apparaît par exemple, que 3 candidats ont eu la note 18, que les notes sont comprises entre 5 et 18, un seul candidat a eu la note de 5.

Caractère quantitatif continu

Les caractères **continus** peuvent prendre « toutes les valeurs »[2] d'un intervalle de $\mathbb{R}$. Ce sont par exemple la taille, le poids, ou le salaire d'un individu... Les résultats sont représentés par des intervalles, dits **classes** du type :

$$[a_i, a_{i+1}[$$

[1] Un caractère discret peut a priori prendre une infinité dénombrable de valeurs, mais sur l'échantillon ce nombre est fini.

[2] En réalité toute mesure a une précision limitée. Toutefois, on peut admettre en première approximation que les valeurs décrivent entièrement un intervalle.

On note par n_i l'effectif de cette classe $[a_i, a_{i+1}[$. On parle de la série

$$([a_i, a_{i+1}[, n_i), \text{ pour } i = 1, \ldots, k \tag{1.2}$$

Pour une classe $[a_i, a_{i+1}[$ on introduit :

- **Le centre**[3] $x_i = \frac{a_i + a_{i+1}}{2}$
- **L'amplitude**[4] égale à $a_{i+1} - a_i$

Exemple

On a relevé la taille en cm de $N = 25$ étudiants d'un groupe de TD. Après avoir rassemblé les données par classes d'amplitude 10 cm, on a obtenu la série $([a_i, a_{i+1}[, n_i)$ suivante :

classe	$[150, 160[$	$[160, 170[$	$[170, 180[$	$[180, 190[$
n_i	4	9	8	4
x_i	155	165	175	185

Tableau 1.3 : Les tailles regroupées en classes

Il y a 4 classes, i varie de 1 à 4. Avec les notations, on a par exemple, $a_1 = 150, a_2 = 160$, avec $n_1 = 4$, ce qui signifie que 4 étudiants mesurent entre 150 et 160 cm. On ne connait pas précisément leur taille, mais on en a un encadrement, donné par la classe.

Regroupement par classes.

Dans la pratique, les séries continues sont souvent obtenues à partir de séries discrètes ayant un grand effectif. Cette transformation se fait en regroupant les **données brutes** par classes. Par exemple, la série continue donnée par la tableau 1.3, a été obtenue à partir des 25 mesures suivantes :

[3] Dans certains ouvrages, le centre est noté c_i. Nous avons choisi la notation x_i pour garder une écriture uniforme qui s'applique aussi bien dans le cas discret que dans le cas continu.

[4] Parfois notée l_i.

152	156.5	158	159.5	160.5
162.5	164.5	166.5	167.5	168.5
168.5	169	169.5	170.5	172
173.5	174	174.5	176.5	178
179.5	180	181.5	186.5	189

La réalisation du tableau 1.3 a été effectuée en choisissant de regrouper les données en classes d'amplitude 10 cm. Les résultats sont alors plus facilement lisibles. En contrepartie, il y a une perte d'information car l'introduction des classes ne nous permet plus de retrouver les valeurs individuelles mesurées.

Notons aussi que le regroupement en classes n'est pas unique. Il dépend de divers choix : le nombre de classes, leurs amplitudes... Dans tous les cas, ces choix doivent être faits de manière judicieuse, pour pouvoir assurer une certaine « homogénéité » [5] à l'intérieur de chaque classe.

Effectifs et fréquences cumulés

Soit une série statistique de la forme (1.1) ou (1.2). On introduit les grandeurs suivantes :

- L'effectif cumulé croissant noté $\mathbf{n_i}\nearrow$

$$n_i\nearrow = n_1 + n_2 + \cdots + n_i$$

- L'effectif cumulé décroissant noté $\mathbf{n_i}\searrow$

$$n_i\searrow = n_i + n_{i+1} + \cdots + n_k$$

- Les fréquences cumulées

$$\mathbf{f_i}\nearrow = \frac{n_i\nearrow}{N} \text{ et } \mathbf{f_i}\searrow = \frac{n_i\searrow}{N}$$

[5] C'est à dire une bonne répartition. En particulier, on ne choisira pas de classe vide ou trop peu remplie.

Exemple

Poursuivons avec la série donnée par la tableau 1.2 et complètons celui-ci en calculant les effectifs et fréquences cumulés.

x_i	5	7	9	10	11	12	13	14	15	18
n_i	1	2	7	8	7	9	5	6	2	3
$n_i\nearrow$	1	3	10	18	25	34	39	45	47	50
$n_i\searrow$	50	49	47	40	32	25	16	11	5	3
f_i	0.02	0.04	0.14	0.16	0.14	0.18	0.1	0.12	0.04	0.06
$f_i\nearrow$	0.02	0.06	0.2	0.36	0.5	0.68	0.78	0.9	0.94	1
$f_i\searrow$	1	0.98	0.94	0.80	0.64	0.50	0.32	0.22	0.10	0.06

On remarque que la dernière case des effectifs cumulés croissants $n_i\nearrow$ est 50*, qui correspond à l'effectif total ; c'est aussi la première case de l'effectif cumulé décroissant $n_i\searrow$. De même la dernière case des fréquences cumulées croissantes $f_i\nearrow$ est* 1*, qui correspond à la fréquence totale ; c'est aussi la première case de la fréquence cumulée décroissante $f_i\searrow$.*

Exemple

Reprenons la série définie par la tableau 1.3 et complétons celui-ci avec les effectifs et fréquences cumulés.

classe	$[150, 160[$	$[160, 170[$	$[170, 180[$	$[180, 190[$
n_i	4	9	8	4
$n_i\nearrow$	4	13	21	25
$n_i\searrow$	25	21	12	4
x_i	155	165	175	185
f_i	0.16	0.36	0.32	0.16
$f_i\nearrow$	0.16	0.52	0.84	1
$f_i\searrow$	1	0.84	0.48	0.16

1.2 Représentations graphiques

Caractère qualitatif

Il existe une grande variété de représentations graphiques possibles. Nous nous intéresserons ici aux :

- **diagrammes à bandes ou tuyaux d'orgue**

  En général, les tuyaux d'orgue sont utilisés pour représenter les modalités qui peuvent être ordonnées. Ils sont composés d'un axe et d'une série de bandes horizontales ou verticales de même largeur représentant les modalités. Les longueurs des bandes sont proportionnelles aux effectifs ou aux fréquences.

- **secteurs circulaires**

  C'est un disque découpé en secteurs dont les aires (et donc les mesures des angles des secteurs angulaires) sont proportionnelles aux effectifs (ou aux fréquences) associés aux différentes modalités. Une fréquence de 100 % correspond à un angle de 360°.

- **graphiques en étoile** [6]

  Il y a autant d'axes que de modalités qui partent du point central ; les noms des modalités sont indiqués à la fin de chaque axe. Les valeurs de la série sont affichées sur chaque axe et reliées entre elles par des segments de droites formant un polygone. L'intérêt est que l'on peut représenter plusieurs séries dans un même graphique pour pouvoir les comparer.

[6] On les nomme aussi graphiques en toile d'araignée ou en radar ou de Kiviat.

Exemple

Traçons les graphiques en tuyaux d'orgue et en secteurs circulaires dans le cas de l'enquête sur la satisfaction des clients d'un hypermarché, dont les résultats sont dans le tableau 1.1. Pour le diagramme en secteurs circulaires, il faut calculer les angles correspondants à chaque modalité ; le calcul est fait dans le tableau suivant (chaque angle se calcule par $f_i \times 360$) :

x_i	*Pas satisfait*	*Peu satisfait*	*Satisfait*	*Très satisfait*	*Total*
n_i	80	110	150	160	500
f_i	0.16	0.22	0.3	0.32	1
angle	57.6°	79.2°	108°	115.2°	360°

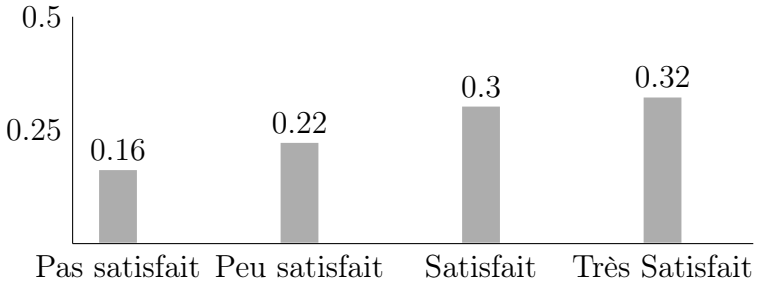

Figure 1.1 : Tuyaux d'orgue

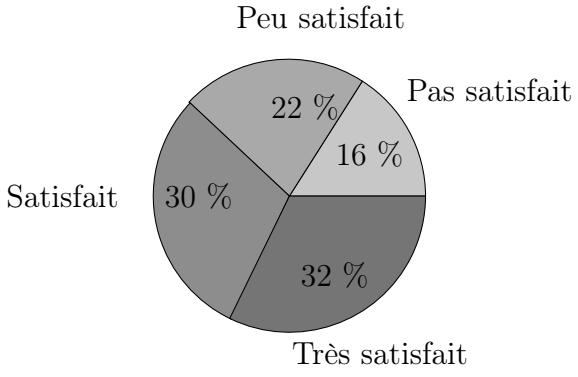

Figure 1.2 : Secteurs circulaires

.