Chapitre I. Statistiques Descriptives

Introduction

Ce chapitre a pour but de rappeler les notions élémentaires permettant de décrire de façon synthétique des données. Ainsi les notions habituelles de moyennes, écarts-types, médianes et les représentations graphiques usuelles (diagrammes en bâtons, diagrammes circulaires, histogrammes, box-plot) seront étudiées. L'accent sera mis la façon de calculer ces quantités mais également sur les interprétations précises qui leurs sont associées.

1 Terminologie statistique

Comme toute discipline, la statistique possède un langage spécifique qui possède parfois certaines nuances par rapport au sens commun. Pour éviter toute ambiguïté, voici les définitions qui vont être utilisées dans cet ouvrage :

Définition I.1 *Individu, population et échantillon :*

- ***Individu statistique*** *: On appelle individu statistique ou unité d'observation,* ***"l'objet"*** *observé lors d'une étude (enquête statistique ou expérimentation). Le terme historique d'individu fait référence aux sondages pour lesquels la théorie s'est initialement développée. Cela étant dit, par abus de langage un "individu statistique" n'est pas forcément une personne physique tel que cela est suggéré par le sens commun du mot "individu".*

- ***Population*** *: On appelle population statistique, un* ***ensemble homogène d'individus statistiques****. L'homogénéité se caractérise par l'ensemble des propriétés communes aux individus d'une même population statistique qui distingue la population de son extérieur. Par exemple, l'ensemble des étudiants inscrits à l'université de Saint-Etienne cette année constitue une population d'individus qui ont le point commun d'être "étudiants inscrits à l'université de Saint-Etienne cette année". Cela les distingue de catégories plus larges comme l'ensemble des étudiants en France ou la population mondiale.*

- ***Echantillon*** *: On appelle échantillon d'une population statistique, un* ***sous-ensemble*** *de cette population. En général, les échantillons sont de taille restreinte par rapport à la population étudiée dont l'étude entière demanderait trop de temps ou de ressources. Un des objectifs majeurs des statistiques est précisément d'obtenir des informations fiables sur la population entière à partir de la seule connaissance d'informations sur l'un de ses échantillons. Inversement la sous-discipline des statistiques appelée "échantillonage", qui ne sera pas abordée dans cet ouvrage, étudie les manières optimales de sélectionner un sous-ensemble de la population pour que celui-ci soit le plus représentatif possible de la population initiale.*

Voici quelques exemples de populations statistiques :

- L'ensemble des pièces produites par une machine.
- L'ensemble des départements français.
- L'ensemble des iris de l'espèce "Iris versicolor".[1]
- L'ensemble de tous les tickets d'une loterie.

Définition I.2 *Caractère et modalités d'un caractère*

- ***Caractères ou variables statistiques*** *: il s'agit de l'observation (ou de la mesure) de* ***certaines caractéristiques non homogènes*** *des individus d'une population statistique.*

[1] 1 Les "iris de Fisher" sont des données proposées en 1933 par le statisticien Ronald Aylmer Fisher comme données de référence pour l'analyse discriminante et la classification. Les données correspondent à 3 espèces de fleurs (*Iris setosa*, *Iris virginica*, *Iris versicolor*).

- ***Modalités d'un caractère*** *: Les modalités d'un caractère regroupent l'ensemble de toutes les possibilités d'observation pour ce caractère. Elles peuvent être en nombre fini ou non.*

Mathématiquement, les définitions précédentes reviennent à dire qu'un caractère statistique X est une variable représentant une application définie sur une population statistique Ω à valeurs dans l'ensemble des modalités M du caractère étudié :

$$\begin{array}{rcl} X : \Omega & \to & M \\ \omega_i & \mapsto & X(\omega_i) \end{array}$$

Exemples : Pour la population des étudiants inscrits cette année à l'université de Saint-Etienne on peut s'intéresser aux caractères suivants : âge, sexe, nombre d'années d'études antérieures, filières d'inscription, etc. Dans le cas du sexe, il n'y a que deux modalités (Homme-Femme). En revanche l'âge est un nombre entier pouvant aller potentiellement de 0 à 130 ans.
Pour la population des iris versicolor, on peut s'intéresser aux caractères comme la taille de la plante, la couleur de ses fleurs, la durée de floraison, etc. Dans le cas de la taille, les modalités potentielles sont l'ensemble des nombres réels positifs.

2 Les grands types de caractères statistiques

En statistique, on distingue classiquement des grandes familles de variables statistiques selon leurs modalités :

- Une variable (ou un caractère) statistique est dite **qualitative** si ses **modalités ne sont pas numériques**. Par exemple, le sexe, la couleur, etc. sont des variables qualitatives. Bien que non-numériques, les modalités peuvent parfois posséder un ordre intrinsèque (longueur d'onde pour les couleurs par exemple) qu'il peut être utile de conserver. On parle alors de variables **qualitatives ordinales** par opposition aux variables **qualitatives nominales** où aucun ordre naturel n'existe.

- Une variable statistique est dite **quantitative** si ses modalités potentielles sont numériques. On les distingue alors en 3 sous-catégories :

- Les variables **quantitatives finies** pour lesquelles le nombre de modalités potentielles est un nombre fini. (*Exemple* : résultat d'un lancer de dés)

- Les variables **quantitatives dénombrables** pour lesquelles le nombre de modalités potentielles est fini ou dénombrable. (*Exemple* : la durée de vie en jours d'une ampoule)

- Les variables **quantitatives non-dénombrables** pour lesquelles le nombre de modalités potentielles est non-dénombrable. Classiquement, il s'agit de variables dont la mesure donne des nombres réels. (*Exemple* : taille, poids, distance, etc.)

Remarque : En pratique, les données sont intrisèquement limitées par la précision de la mesure qui n'est effectuée qu'à un nombre fixé de chiffres significatifs. En théorie, cela ne donne toujours qu'un nombre fini de modalités possibles. Par exemple, la taille d'une personne est souvent donnée en cm avec 3 chiffres significatifs (par exemple 178 cm) ce qui ne fournit qu'un nombre fini de possibilités. Néanmoins, ces variables restent à classer dans la catégorie des variables non-dénombrables car une augmentation de la précision des mesures changerait le nombre des modalités. Le cas des bornes supérieures est aussi en pratique difficile à gérer. Par exemple, l'âge d'une personne ne possède pas de borne supérieure naturelle mais, en pratique, on sait qu'il ne va dépasser la valeur 130 que rarement. Même la durée de vie d'une ampoule est également limitée par la durée de l'expérience et il se peut que certaines ampoules soient encore en fonctionnement après le temps prévu par l'expérience.

3 Tri à plat des données

Le tri à plat d'un caractère statistique consiste à regrouper sous forme de tableau les différents effectifs obtenus pour chaque modalité rencontrée. Il est particulièrement utile lorsque les modalités sont en faible nombre. On présente en général les résultats sous la forme suivante :

Modalité	Effectif	Fréquence (%)
m_1	n_1	f_1
m_2	n_2	f_2
$\vdots$	$\vdots$	$\vdots$
m_p	n_p	f_p
Total	N	100(%)

La taille de l'échantillon (ou le nombre de personnes ayant répondu à cette question) est :

$$N = \sum_{i=1}^{p} n_p$$

Le nombre de modalités différentes est p et les fréquences (traditionnellement exprimées en %) se calculent par :

$$f_p = \frac{n_p}{N}$$

Notons que les effectifs contiennent un peu plus d'information que les fréquences car on ne peut obtenir les effectifs à partir des fréquences que si l'on connait la taille de l'échantillon étudié. Ce type de tri à plat peut être utilisé pour des variables **qualitatives ou quantitatives** tant que le **nombre de modalités** reste petit sinon la lecture devient compliquée. Pour les variables quantitatives présentant un nombre trop important de modalités, on peut procéder de la même façon en regroupant les modalités par classes. Par exemple (on utilisera dans cet ouvrage les conventions internationales de notation des intervalles : $[a, b)$ pour un intervalle fermé en a et ouvert en b) :

Age	Effectif	Fréquence (%)	Fréquences cumulées (%)
$[0, 5)$	n_1	f_1	f_1
$[5, 10)$	n_2	f_2	$f_1 + f_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[45, 50)$	n_{10}	f_{10}	$f_1 + \cdots + f_{10}$
50 ou plus	n_{11}	f_{11}	100%
Total	N	100(%)	

Lorsque les modalités présentent un ordre sous-jacent (par exemple si ce sont des nombres), il est parfois intéressant de calculer les **fréquences**

cumulées comme dans le tableau précédent (on peut également faire la même chose avec les **effectifs cumulés**). Ainsi la fréquence cumulée de la modalité m_i représente la probabilité empirique d'obtenir une valeur plus petite ou égale à m_i. La tracé de la courbe des fréquences cumulées correspond à la **fonction de répartition empirique.**

4 Grandeurs synthétiques

Pour les variables **quantitatives** (et uniquement celles-ci), on préfère souvent résumer les données concernant un caractère statistique à quelques grandeurs usuelles. Attention ces grandeurs ont un sens uniquement pour le cas de grandeurs dites **extensives**. C'est le cas de beaucoup de quantités (distance, poids, comptage, durée, surface, quantité de matière, etc.) mais pas de toutes (pression, taux bancaire, vitesse, etc.). Il faut donc bien vérifier si le type de caractère étudié est extensif ou non avant de calculer les quantités suivantes :

- **Moyenne empirique** : La **moyenne empirique** d'un caractère X est définie par :

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{p} n_i m_i = \sum_{i=1}^{p} f_i m_i$$

  On pondère ainsi chaque modalité (qui est un nombre puisque la variable est quantitative) par sa fréquence correspondante.

- **Ecart-type et variance empirique** : L'écart-type non-corrigé empirique, traditionnellement noté σ_X est défini par :

$$\sigma_X = \sqrt{\frac{1}{N}\sum_{i=1}^{p} n_i \left(m_i - \overline{X}\right)^2} = \sqrt{\frac{1}{N}\left(\sum_{i=1}^{p} n_i m_i^2\right) - \overline{X}^2}$$

  La variance étant toujours le carré de l'écart-type, on a la variance empirique non-corrigé σ_X^2 par :

$$\mathrm{Var_{nc}}(X) = \sigma_X^2 = \frac{1}{N}\sum_{i=1}^{p} n_i \left(m_i - \overline{X}\right)^2 = \frac{1}{N}\sum_{i=1}^{p} n_i m_i^2 - \overline{X}^2$$

  Comme on le verra dans cet ouvrage sur l'estimation de paramètres, ces quantités présentent un défaut qui nécessite une correction. On

définit ainsi les **écart-types et variances empiriques corrigés** notés s_X et s_X^2 par :

$$s_X = \sqrt{\frac{1}{N-1}\sum_{i=1}^{p} n_i \left(m_i - \overline{X}\right)^2} = \sqrt{\frac{1}{N}\left(\sum_{i=1}^{p} n_i m_i^2\right) - \frac{N-1}{N}\overline{X}^2}$$

La plupart des logiciels utilisent systématiquement par défaut la variance et l'écart-type corrigés lorsque l'on demande ces informations. Il est à noter que lorsque N est grand la différence est négligeable puisque l'on a :

$$s_X = \frac{N}{N-1}\sigma_X \;\Rightarrow s_X \overset{N\to\infty}{\to} \sigma_X$$

- **Médiane empirique** : La **médiane empirique** correspondant à la modalité pour laquelle la fréquence cumulée atteint ou dépasse les 50%. Lorsque l'on classe les données par ordre croissant, la médiane correspond à la valeur située au milieu. Dans le cas d'un échantillon de taille paire $N = 2k$, il existe plusieurs conventions selon les logiciels. La définition proposée ici donne l'entrée x_{k+1} lorsque celle-ci sont classées par ordre croissant. Certains logiciels préfèrent donner la moyenne entre x_k et x_{k+1}.

- **Moments empiriques** : Le **moment empirique d'ordre** k d'une grandeur X est défini par :

$$M_k = \frac{1}{N}\sum_{i=1}^{p} n_i m_i^k$$

Le moment d'ordre 1 correspond à la moyenne : $M_1 = \overline{X}$ tandis que le moment d'ordre 2 correspond à $M_2 = \sigma_X^2 + \overline{X}^2$. Les moments d'ordre supérieurs sont parfois utiles pour obtenir une description plus fine de la distribution des données.

- **Quantiles empiriques** : Soit $\alpha \in (0, 1)$. On appelle **quantile empirique d'ordre** α et l'on note q_α la plus petite modalité pour laquelle la fréquence cumulée dépasse ou égale α. Lorsque les données sont classées par ordre croissant, cela correspond à $x_{\lceil \alpha N \rceil}$. Comme dans le

cas de la médiane empirique, certains logiciels préfèrent effectuer une moyenne pondérée entre $x_{\lceil \alpha N \rceil}$ et $x_{\lfloor \alpha N \rfloor}$ lorsque αN n'est pas entier. Les quantiles empiriques les plus utilisés en pratique sont :

- Le 1$^{\text{er}}$ décile, noté D_1, correspond à $\alpha = \frac{1}{10}$.
- Le 1$^{\text{er}}$ quartile, noté Q_1, correspond à $\alpha = \frac{1}{4}$.
- La médiane qui correspond à $\alpha = \frac{1}{2}$ ainsi qu'au second quartile Q_2 et au cinquième décile D_5.
- Le 3$^{\text{ème}}$ quartile, noté Q_3, correspond à $\alpha = \frac{3}{4}$.
- Le dernier décile (le 9$^{\text{ème}}$), noté D_9, correspond à $\alpha = \frac{9}{10}$.

Remarque : Il existe également d'autres grandeurs qui peuvent avoir un intérêt dans certaines analyses comme par exemple l'asymétrie ou l'aplatisement. Dans tous les cas, il est bon de se souvenir que ces grandeurs ne contiennent que peu d'information sur les données initiales et sont très loin d'être toujours pertinentes. En particulier, **leur pertinence ne peut être évaluée qu'avec une représentation graphique adéquate** des données.

5 Représentations graphiques

Afin de visualiser la forme générale des mesures d'un caractère statistique X on a très souvent recours à des représentations graphiques qui permettent une lecture rapide et précise des données.

5.1 Diagramme circulaire

Le diagramme circulaire consiste à représenter un cercle coupé en secteurs angulaires dont l'angle est proportionnel à l'effectif de la modalité étudiée. Ce type de diagramme est pertinent lorsque le **nombre de modalités est faible** (2, 3 ou 4). Par ailleurs, il faut garder à l'esprit que **le cerveau humain est peu performant pour évaluer finement la largeur des secteurs angulaires** et qu'il est donc difficile d'évaluer l'écart relatif entre deux modalités sur un diagramme circulaire. **Il est donc souvent préférable d'utiliser un diagramme en bâtons à un diagramme ciculaire**.