TABLEAUX STATISTIQUES ET GRAPHIQUES

Dans ce cours, on considèrera des *échantillons* de taille N, c'est-à-dire N individus $\omega_1, \ldots, \omega_N$ issus d'une population Ω. On notera en majuscule (généralement X ou Y) les variables statistiques. Les modalités des variables statistiques seront notées avec la minuscule correspondante et indicées, s'il y a lieu, par le numéro de la modalité dans le cas discret ou de la classe (ie un ensemble de modalité) dans le cas continue. La modalité prise par la variable X pour l'individu ω_n sera notée $X(\omega_n)$.

1. TABLEAUX STATISTIQUES

1.1. **Tableau brut.** On considère une étude statistique portant sur un échantillon de taille N indivus. On mesure plusieurs variables statistiques, qui peuvent être qualitatives ou quantitatives. La récolte initiale des données conduit à un tableau brut de la forme

Individus	variable 1	variable 2	...
ω_1	$X(\omega_1)$	$Y(\omega_1)$	
ω_2	$X(\omega_2)$	$Y(\omega_2)$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$
ω_N	$X(\omega_N)$	$Y(\omega_N)$	

1.2. **Tableau de contingence.** Pour les variables *qualitatives*, on peut construire un tableau de contingence. Ce tableau résume comment une caractéristique dépend d'une autre. Pour des raisons pratiques, on se limite généralement au tableau de contingence de deux variables qualitatives X et Y de modalités respectives $(x_1, \ldots, x_i, \ldots, x_I)$ et $(y_1, \ldots, y_j, \ldots, y_J)$. Ce tableau donne le nombre d'individus possédant simultanément la modalité x_i de la variable X et la modalité y_j de la variable Y. Un tel tableau se présente sous la forme suivante, pour un échantillon de taille N :

X \ Y	y_1	...	y_j	...	y_J	
x_1	N_{11}		N_{1j}		N_{1J}	$N_{1\bullet}$
$\vdots$			$\vdots$			
x_i	N_{i1}	...	N_{ij}	...	N_{iJ}	$N_{i\bullet}$
$\vdots$			$\vdots$			
x_I	N_{I1}		N_{Ij}		N_{IJ}	$N_{I\bullet}$
	$N_{\bullet 1}$		$N_{\bullet j}$		$N_{\bullet J}$	N

où N_{ij} désigne le nombre de fois où X à pris la modalité x_i et Y la modalité j. Autrement dit, N_{ij} représente le nombre d'individus qui possède à la fois la caractéristique x_i et la caractéritique y_j. On définit les quantités

$$N_{i\bullet} = \sum_{j=1}^{J} N_{ij} \quad N_{\bullet j} = \sum_{i=1}^{I} N_{ij}$$

qui représentent respectivement le nombre d'individus qui possèdent la modalitée x_i et le nombre d'individus avec la modalitée y_j.

1.3. **Tableau des fréquences.**

1.3.1. *Variable discrète.* On considère une variable qualitative ou qualitative discrète X à valeurs dans $\{x_1, \ldots, x_K\}$. Si X est quantitative ou qualitative ordinale, on suppose ses modalités ordonnées tels que $x_1 < \ldots < x_K$. On commence par définir les quantités descriptives de bases.

Définition 1 (Effectif ou fréquence absolue)**.** *On appelle* effectif *ou* fréquence absolue *de la modalité x_k le nombre N_k d'individus qui ont pris la modalité x_k :*

$$N_k = \sum_{\omega \in \Omega | X(\omega) = x_k} X(\omega) \ .$$

Remarque 1. $\sum\limits_k N_k = N$.

Définition 2 (Fréquence relative)**.** *On appelle* fréquence relative *(ou simplement* fréquence*) de la modalité x_k le nombre f_k définit par :*

$$f_k = \frac{N_k}{N} \ .$$

Définition 3 (Fréquence relative cumulée)**.** *On appelle* fréquence relative cumulée *(ou simplement* fréquence cumulée*) de la modalité x_k le nombre $F_k^{(c)}$ définit par :*

$$F_k^{(c)} = \sum_{i=1}^{k} f_k \ .$$

Remarque 2. $F_K^{(c)} = \sum\limits_{k=1}^{K} f_k = 1$.

On peut aussi définir les effectifs cumulés :

Définition 4 (Effectifs cumulés)**.** *On appelle* effectif cumulé *de la modalité x_k le nombre E_k définit par :*

$$E_k = \sum_{i=1}^{k} N_k = N F_k^{(c)} \ .$$

La distribution des N observations de X peut être présentée sous la forme d'un tableau de fréquence où figurent, pour chaque modalité x_k, l'effectif N_k, la fréquence relative f_k et la fréquence cumulée $F_k^{(c)}$:

modalité	effectif	fréquence relative	fréquence cummulée
x_1	N_1	f_1	$F_1^{(c)}$
$\ldots$			
x_k	N_k	f_k	$F_k^{(c)}$
$\ldots$			
x_K	N_K	f_K	$F_K^{(c)}$

Les fréquences relative et cumulée peuvent être donnée sous forme de pourcentage.

1.3.2. *Variable continue.* Dans le cas où la variable X est continue, la réalisation d'un tableau de fréquence nécessite au préalable une répartition en *classes* des données. On doit définir *a priori* le nombre de classes K et l'amplitude (ou l'étendue) de chaque classe. Ce choix doit résulte d'un compromis entre deux objectifs antagonistes : résumer les données (K ne doit pas être trop grand) sans perdre l'information pertinente (K ne doit pas être trop petit). Pour ce faire, un moyen «simple» est de diviser l'étendue des données en plusieurs intervalles de même longueur, puis l'on regroupe les classes d'effectifs trop petit (ie moins de 5 individus). On peut utiliser une des deux règles suivantes pour déterminer le nombre de classes :

Règle de Sturge: $K = 1 + \frac{10}{3}\log_{10}(N)$

Règle de Yule: $K = 2.5N^{\frac{1}{4}}$

L'intervalle entre les classes est alors donné par

$$\text{Longueur de l'intervalle} = \frac{x_{\max} - x_{min}}{K} \ ,$$

où $x_{\max}$ (resp. $x_{\min}$) désigne la plus grande (res. la plus petite) valeur de prise par les $X(\omega)$, $\omega \in \Omega$.

On note $\mathcal{C}_k$ l'ensemble des individus qui appartiennent à la classe K.

Définition 5 (Amplitude)**.** *L'amplitude L_k de la classe k est donnée par*

$$L_k = \max\{X(\omega), \omega \in \mathcal{C}_k\} - \min\{X(\omega), \omega \in \mathcal{C}_k\} \ .$$

c'est-à-dire la longueur de l'intervalle.

On peut alors définir la *densité* d'une classe :

Définition 6 (Densité)**.** *La densité d_k de la classe k est donnée par $d_k = \frac{N_k}{L_k}$*

Ce découpage en classes permet de se ramener au cas discret décrit précédemment pour obtenir le tableau de fréquences, en adaptant directement les définitions vues précédemment.

2. Représentations graphiques

Lorsqu'on observe un caractère sur des individus, les tableaux de chiffres définis précédemment sont peu parlant. Ils sont cependant très utiles pour construire des graphiques divers, qui permettent d'un seul coup d'oeil d'avoir une idée de la manière dont se répartissent les individus.

2.1. **Variables qualitatives.** On considère une variable statistique qualitative X prenant K modalités $x_1, \ldots, x_k, \ldots, x_K$. La seule représentation qui nous intéresse est celle des effectifs N_k ou des fréquences f_k. On utilise le tableau de fréquence pour construire les graphiques définis par la suite.

2.1.1. *Diagramme en barre ou tuyaux d'orgue.*
- Les modalités de la variable sont placées sur une droite horizontale (**attention :** si la variable est nominale, ne pas orienter cette droite car les modalités n'ont pas de relation d'ordre).
- Les effectifs ou les fréquences sont placées sur un axe vertical. La hauteur du baton est proportionnelle à l'effectif.
- Les tuyaux ont une certaine épaisseur pour qu'il n'y ait pas de confusion avec les diagrammes en bâtons réservés à la variable quantitative discrète.
- Il doit y avoir un espace entre les tuyaux pour ne pas les confondres avec les histogrammes réservés aux variables quantitatives continues.

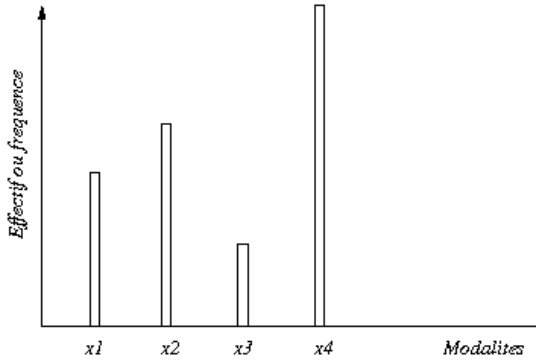

FIG. 1. Diagramme en barres

2.1.2. *Diagramme en secteurs ou «camenbert».*
- L'effectif total est représenté par un disque.
- Chaque modalité est représentée par un secteur circulaire dont la surface (pratiquement : l'angle au centre) est proportionnelle à l'effectif correspondant.

Si ce type de graphique est couramment utilisé dans les médias, c'est une très mauvaise représentation car il présente un risque d'interprétation : l'oeil distingue moins bien les différences entre secteurs (d'un camembert) qu'entre hauteurs (d'un diagramme en barre).

2.2. **Variables quantitatives.** Avant toute tentative de représentation, il y a lieu de distinguer entre variable discrète et variable classée (regroupements en classes). Si pour une variable continue le regroupement en classes est nécessaire, lorsque les modalités d'une variable discrète sont trop nombreuses il est préférable de regrouper des modalités pour obtenir une variable classée, afin que les graphiques synthétisent l'information et restent lisibles.

On considère une variable statistique quantitatives X prenant ses valeurs parmis K modalités ou classes $x_1, \ldots, x_k, \ldots, x_K$. On suppose les modalités (ou classes) ordonnées telles que $x_1 < x_2 < \ldots < x_K$. On utilise le tableau de fréquence pour construire les graphiques définis par la suite.

Deux types de graphiques sont intéressants à représenter :

(1) Les diagrammes différentiels qui mettent en évidence les différences d'effectifs (ou de fréquences) entre les différentes modalités ou classes.

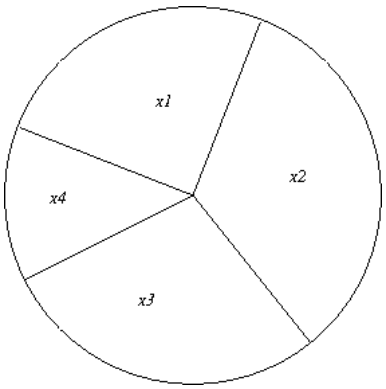

FIG. 2. Diagramme en secteurs

(2) les diagrammes cumulatifs qui permettent de répondre aux questions du style «combien d'individus ont pris une valeur inférieure (ou supérieure) à tant ?».

2.2.1. *diagrammes différentiels.*
Variables discrètes. Pour les caractères quantitatifs discrets, la représentation graphique différentielle est le *diagramme en bâtons* où la hauteur des bâtons correspond à l'effectif N_k (ou la fréquence relative f_k) associé à chaque modalité du caractère x_k.

- Les valeurs discrètes prises par les modalités sont placées sur l'axe des abscisses, ordonnées comme il se doit.
- Les effectifs ou fréquences sont placées sur l'axe des ordonnées.
- Les axes sont fléchés.
- La hauteur du baton est proportionnelle à l'effectif ou la fréquence.
- **Attention :** bien faire des batons et non des tuyaux ou des histogrammes.

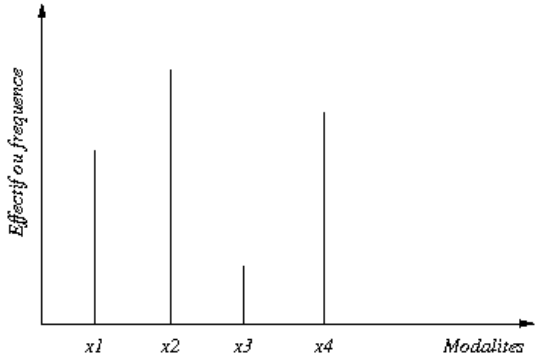

FIG. 3. Diagramme en batons

Variables continues. Lorsque les caractères quantitatifs sont continus, on utilise *l'histogramme*. Un histogramme est ensemble de rectangles contigus où chaque rectangle associé à chaque classe a *une surface* proportionnelle à l'effectif (fréquence)

de cette classe. Si les classes sont d'amplitudes égales, alors la hauteur des rectangles est proportionnelle à l'effectif de la classes. **Avant toute construction d'histogramme, il faut donc regarder si les classes sont d'amplitudes égales ou non**.

Les modalités (continues) sont représentés en abscisses. Le cas des classes d'amplitudes égales ne pose aucune difficulté car il suffit de reporter en ordonnée l'effectif (la fréquence). Si les classes sont d'amplitudes différentes, on reporte en ordonnée la densité $d_k = \frac{N_k}{L_k}$.

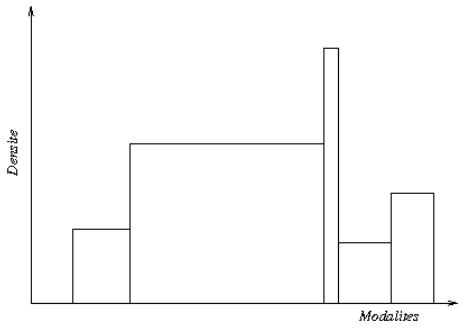

FIG. 4. Histrograme

2.2.2. *diagrammes cummulatifs.* Les diagrammes cummulatifs permettent de visualiser l'évolution des fréquences cummulées ou des effectifs cummulés. On utilise en général la *fonction de répartition empirique* dont la courbe correspond à l'évolution des fréquences cummulées. Elle se définie de la même manière pour les variables quantitatives continues ou discrètes.

Définition 7 (Fonction de répartition empirique)**.** *Soit X une variable statistiques quantitative observée sur un échantillon $\omega_1, \ldots, \omega_N$ de taille N issue d'une population Ω. On appelle fonction de répartition empirique la fonction*

$$\begin{aligned} \hat{F} \ : \ & \mathbb{R} \to [0,1] \\ & x \mapsto \frac{1}{N} \#\{i : X(\omega_i) < x\} \mathbf{1}_{[X(\omega_i), +\infty[}(x) \ . \end{aligned}$$

Pour tout réel x, $\hat{F}(x)$ est donc la proportion d'observations inférieurs ou égales à x. La fonction $\hat{F}$ est une fonction en escalier. Le calcul pratique de $\hat{F}$ s'effectue en ordonnant les N observations $X(\omega_1), \ldots, X(\omega_N)$ par ordre croissant. On note $x'_1, \ldots, x'_I$ les I valeurs distinctes obtenues et N_i l'effectif de x'_i. On a

$$\hat{F}(x) = \begin{cases} 0 & x < x'_1 \\ F_i^{(c)} = \frac{1}{N} \sum_{j=1}^{i} N_j & x'_i \leq x < x'_{i+1} \\ 1 & x \geq x'_I \end{cases}$$

On utilise la fonction de répartition empirique pour répondre aux questions du style : Quel est le nombre (ou le pourcentage) d'individus dont la valeur du caractère est inférieure ou égale à x ?

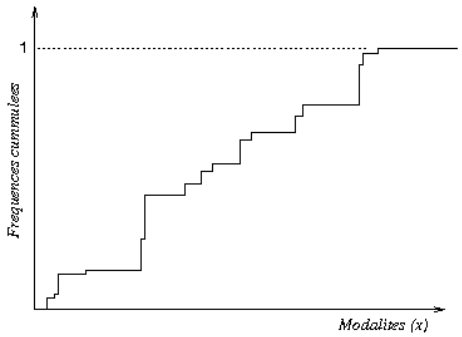

FIG. 5. Fonction de répartition empirique.