Chapitre 1 : Statistique descriptive

I. Introduction Les statistiques jouent un rôle intrinsèque en informatique et vice versa. Les statistiques sont utilisées pour l'exploration de données, la reconnaissance vocale, l'analyse de la vision et des images, la compression de données, l'intelligence artificielle et la modélisation des réseaux et du trafic. Un bagage statistique est essentiel pour comprendre les algorithmes et les propriétés statistiques qui forment l'épine dorsale de l'informatique.

II. Définition La statistique est l'ensemble des méthodes qui servent à organiser les épreuves fournissant des observations menant à collecter des données, à analyser celles-ci et à interpréter les résultats.

L'analyse statistique se subdivise en deux parties

1. *Statistique descriptive* : a pour but de décrire c-à-d de résumer ou représenter les données.

Questions typiques

*Représentation brute (série statistique)

*Représentation par tableaux

*Représentation graphique

*Résumés ou caractéristiques ou indicateurs numériques (Paramètres de position, de dispersion, de relation.)

2. *Statistique inférentielle* : l'ensemble des méthodes permettant de formuler un jugement. Elle nécessite des outils mathématiques plus pointus (théorie des probabilités).

III. Notions de bases

**POPULATION :* La collection d'objets ou de personnes étudiées (étudiants, ordinateurs, voitures…).

**INDIVIDU :* élément de la population étudiée. (un étudiant, un ordinateur, une voiture,…).

*ECHANTILLON : partie de la population étudiée. Le cardinal d'un échantillon est appelé taille de l'échantillon, noté n.

**VARIABLE (CARACTERE)* : propriété commune aux individus de la population, que l'on veut étudier.

Un caractère peut être :

a) qualitatif : on ne peut associer une valeur numérique (couleur des yeux, le processeur, type de voiture…).

Un caractère qualitatif incluse peut être :

*nominal : ses données consistent en des labels ou des noms (couleur des yeux,, type de voiture…)

*ordinal : désigne le rang (un ordre par convention) comme: un peu, moyen, beaucoup.

Exp :le processeur (i3, i5, i7)

b) quantitatif : peut prendre des valeurs numérique (poids, la quantité de mémoire vive, la vitesse du processeur, la capacité de stockage ,le prix,….).

Un caractère quantitatif peut être :

*Continu : peut prendre toutes les valeurs numériques d'un intervalle déterminer (taille, ...), il relève d'une mesure.

*Discontinu (discret) : ne peut prendre que des valeurs numérique isolées (nombre de pièces d'habitations, nombre de fruits endommagés…), il relève d'un comptage ou dénombrement.

*MODALITE l'une des formes particulières d'un caractère. La couleur des yeux est un caractère, ses modalités sont : bleu, vert, marron,… . Dans le cadre d'un caractère quantitatif on parle de VALEUR.

*EFFECTIF TOTALE (noté n) : La taille de l'échantillon, c'est le cardinal de l'échantillon.

*EFFECTIF PARTIELLE OU FREQUENCE ABSOLUE (noté n_i): Le nombre d'apparitions d'une modalité ou d'une valeur associé à un caractère dans un échantillon.

*FREQUENCE RELATIVE (noté f_i): La fréquence d'apparitions d'une modalité ou d'une valeur associé à un caractère dans un échantillon. $f_i = \frac{n_i}{N}$

*EFFECTIF CUMULE (noté n_i^{cum}): $n_i^{cum} = \sum_{j=1}^{i} n_j$. Il s'interprète par Le nombre d'individus qui ont la modalité ou valeur inférieure ou égale à la modalité correspondante.

*EFFECTIF CUMULE (noté f_i^{cum}): $f_i^{cum} = \sum_{j=1}^{i} f_j = \frac{n_i^{cum}}{n}$. Il s'interprète par La fréquence d'individus qui ont la modalité ou valeur inférieure ou égale à la modalité correspondante.

*POURCENTAGE :(exprimé en %): C'est une fréquence multiplié par 100.

IV. Représentation de données:

*SERIE STATISTIQUE: Une série statistique est la suite des modalités ou valeurs que prend un caractère au sein d'un échantillon.

*TABLEAU STATISTIQUE: En statistique, le tableau est un outil efficace pour présenter les données de façon structurée et plus lisible. Nous distinguons trois types de tableaux: **le tableau de données** (ou tableau élémentaire), **le tableau de distribution des effectifs** et enfin **le tableau de distribution des fréquences** (appelés aussi tableaux de dénombrement).

a) Le tableau de données La représentation brute des données n'est pas très lisibles. Ces informations le seront davantage dès qu'elles seront regroupées dans un tableau de données. C'est la raison pour laquelle, dans toute démarche statistique classique, les tableaux de données sont les premiers à être dressés. Ce sont les tableaux qui facilitent et rendent compte du dépouillement des données. Exp: lutilisation d'un fichier Excel,

Tout tableau est composé de lignes et de colonnes. Pour construire notre tableau de données il faut donc tracer des lignes et des colonnes. Les colonnes portent la liste des caractères étudiés et les lignes correspondent aux individus observés.

	Colonne 1 = Variable 1	Colonne 2 = Variable 2	Colonne 3 = Variable 3
Ligne 1 = Individu 1			
Ligne 2 = Individu 2		Case = Modalité ou Valeur	
Ligne 3 = Individu 3			

b) Tableau de Distribution des Effectifs : le tableau de distribution réorganise les données du tableau de données et les présente de manière plus claire et plus concises, sans rien perdre de l'information contenue dans la série statistique de départ. La construction du tableau des effectifs depend de la nature du caractère étudié. Elle se fait directement dans le cadre d'un caractère qualitatif ou quantitatif discret. En revanche, dans le cas d'un caractère quantitatif continu, la construction nécessite le passage par des classes où les données sont regroupées en en des intervalles semi ouverts, de nombre k donné par l'une des deux formules suivantes:

La règle de Sturge $k=1+3.3 \log(n)$

La règle de Yule $k=2.5\ (n)^{1/4}$

et La construction des classes se fait de la maière suivante:

1. On calcul L'étendue d'une série statistique $e = valeur\,maximale - valeur\,minimale$

2. On calcul la longueur l de la classe telle que $l > \frac{e}{k}$.

Le tableau des effectifs se compose d'une colonne présentant la liste des modalités (valeurs ou classes)du caractère étudié et l'autre colonne correspondant à l'effectif pour chaque modalité (valeurs ou classes).

modalité (valeurs ou classes)	Effectifs n_i
...	...

Remarque: De la même façon on définit le tableau de distribution des fréquences en remplaçant les effectif n_i par les fréquences f_i

*REPRESENTATION GRAPHIQUE: Il est nécessaire de dresser une représentation graphique afin de faire ressortir une partie de l'information des données pour quelles soient de plus en plus « parlantes ». Suivant la nature du caractère, le mode de représentation graphique va être différent: diagramme circulaire (nominal), diagramme en barre (ordinal), diagramme en batôns (discret) et l'histogrammes (continu).

*INDICATEURS NUMERIQUES (Paramètres):

a) les indicateurs de position (la tendence centrale):

Le mode: (noté m_o) le mode est la modalité ou valeur la plus fréquente ie celle qui correspond au plus grand effectif. Le mode n'est pas unique (dans le cas d'un caractère continue on parle de classe modale).

Exemple: la série statistique 2,2,5,7,9,9,9,10,10,11,12,18 a comme mode 9.
la série statistique 3, 5, 8, 10, 12,15, 16 n'a pas de mode.
la série statistique 2, 3, 4, 4, 4, 5, 5, 7, 7,7, 9 a deux mode 4 et 7. La série est appelée bimodale.

*Une série ayant un seul mode est appelée uni modale.

*Dans le cas d'une variable continue, on applique la formule suivante:

$$m_o = a_{i_0} + l\frac{\nabla_1}{\nabla_1 + \nabla_2}$$

avec:

a_{i_0} : est la limite inférieure de la classe modale,

l : est la longueur de la classe

∇_1 :la différence entre l'effectif (la fréquence) de la classe modale et celle d'avant.

∇_2 :la différence entre l'effectif (la fréquence) de la classe modale et celle d'aprés.

La médiane (noté m_e): c'est la valeur qui sépare une série statistique ordonnée en deux parties égales.

Exemple: la série statistique 3, 4, 4, 5, 6, 8, 8, 8, 10 a comme médiane la valeur 6.
la série statistique 5, 5, 7, 9, 11, 12,15,18 a comme médiane (9+11)/2= 10.

*La médiane correspond à la fréquence cumulée 0,5.

*Dans le cas d'une variable continue, on applique la formule suivante:

$$m_e = l_{i_e} + (l_{i_e+1} - l_{i_e}) \frac{0.5 * n - n_{i_e-1}^{cum}}{n_{i_e}^{cum} - n_{i_e-1}^{cum}}$$

avec:

l_{i_0} : est la limite inférieure de la classe modale,

l_{i_0+1} : est la limite supérieure de la classe modale,

n : est l'effectif

$n_{i_e-1}^{cum}$: est l'effectif cumulatif de la classe qui précède la classe médiane

$n_{i_e+1}^{cum}$:la différence entre l'effectif cumulatif de la classe qui succède la classe médiane.

La moyenne: Soit X un caractère discret ayant k valeurs $x_1, \ldots, x_k$. Sa moyenne $\bar{X}$ est donnée par:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{k} n_i x_i$$

*Si t X un caractère discret ayant k classes $[x_i, \ldots, x_{i+1}[$ $i = \overline{1,k}$. Sa moyenne $\bar{X}$ est donnée par:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{k} n_i c_i$$

où c_i centre de la classe $[x_i, \ldots, x_{i+1}[$, $c_i = \frac{x_{i+1} - x_i}{2}$.

b) les indicateurs de la dispersion :

La variance: Soit X un caractère discret ayant k valeurs $x_1, \ldots, x_k$ de moyenne $\bar{X}$. Sa variance est donnée par:

$$V_X = \frac{1}{n} \sum_{i=1}^{k} n_i (x_i - \bar{X})^2 \quad \text{ou} \quad V_X = \frac{1}{n} \sum_{i=1}^{k} n_i x_i^2 - \bar{X}^2$$

*Si t X un caractère discret ayant k classes $[x_i, \ldots, x_{i+1}[$ $i = \overline{1,k}$ de moyenne $\bar{X}$, Elle est donnée par:

$$V_X = \frac{1}{n} \sum_{i=1}^{k} n_i (c_i - \bar{X})^2 = \frac{1}{n} \sum_{i=1}^{k} n_i c_i^2 - \bar{X}^2$$

où c_i centre de la classe $[x_i, \ldots, x_{i+1}[$, $c_i = \frac{x_{i+1} - x_i}{2}$

L'écart-type: $\sigma_X = \sqrt{V_X}$

V. Fonction de répartition:

La fonction de répartition d'une variable statistique quantitative X est la fonction $F_X : \mathbb{R} \to [0,1]$ qui à tout x fait associer la quantité $F_X(x)$ désignant la fréquence (le pourcentage lorsque multipliée par 100) des individus pour lesquels la variable statistique X soit inférieur où égale à x . Son graphe est dit la courbe cumulative des fréquences.

*Cas d'une variable statistique discrète:

Soit X une variable statistique discrète ayant k valeurs $x_1, \ldots, x_k$. Sa fonction de répartition estdéfinie par :

$$F_X(x) = \begin{cases} 0 & si \quad x < x_1 \\ f_i^{cum} & si \quad x_i \leq x < x_{i+1} \\ 1 & si \quad x \geq x_k \end{cases}$$

Sa courbe est le graphe d'une fonction en escaliers.

Exemple:

Population: Les familles

L'échantillon: les familles d'un immeuble, n=64

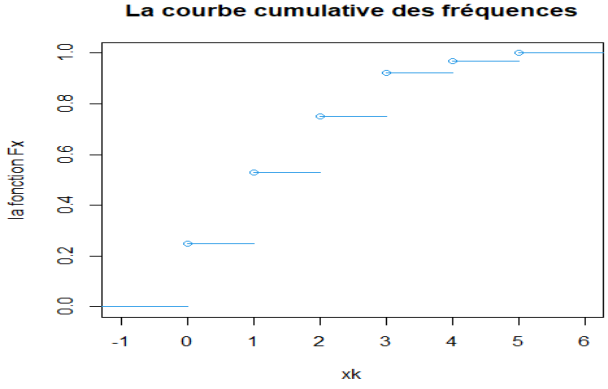
X : le nombre d'enfants par famille, une variable statistique discrète.

X	n_i	n_i	n_i^{cum}	f_i^{cum}
0	16	0.25	16	0.25
1	18	0.281	34	0.531
2	14	0.218	48	0.749
3	11	0.172	59	0.921
4	3	0.047	62	0.968
5	2	0.031	64	0.999
Total	64	1		

La fonction de répartition des famille selon le nombre des enfant est donnée par

$$F_x(x) = \begin{cases} 0 \; si \; x < 0 \\ 0.25 \; si \; 0 \leq x < 1 \\ 0.531 \; si \; 1 \leq x < 2 \\ 0.749 \; si \; 2 \leq x < 3 \\ 0.921 \; si \; 3 \leq x < 4 \\ 0.968 \; si \; 4 \leq x < 5 \\ 1 \; si \; x \geq 5 \end{cases}$$

et son graphe est



*Cas d'une variable statistique continue:

Soit X une variable statistique continue ayant k classes $[a_{i-1}, a_i], i=\overline{1,k}$. Sa fonction de répartition est:

$$F_X(x)=\begin{cases} 0 & si \quad x<a_0 \\ f_1\dfrac{x-a_0}{l} & si \quad a_0\le x<a_1 \\ f_i^{cum}+f_{i+1}\dfrac{x-a_i}{l} & si \quad a_i\le x<a_{i+1} \\ 1 & si \quad x\ge a_k \end{cases}$$

Sa courbe est le graphe d'une fonction continue croissante.

Exemple :Le taux de glucose sanguin déterminé chez 32 sujets est donné ci-dessous en g/l par la série: 0,85 - 0,95 -1,00 - 1,06 – 1,11 – 1,19 – 0,87 – 0,97 – 1,01 – 1,07 – 1,13 - 1,20 - 0,90 – 0,97 – 1,03 – 1,08 – 1,14 – 0,93 – 0,98 – 1,03 – 1,08 – 1,14 – 0,94 – 0,98 – 1,03 – 1,10 – 1,15 – 0,94 – 0,99 – 1,04 – 1,10 – 1,17.

Population étudié : sujets humains

L'échantillon sur lequel porte l'étude: n=32 sujets.

Le caractère étudié est Le taux de glucose sanguin. C'est un caractère quantitatif continu.

* D'après la formule de Yule donne dans ce cas

k=6=5 $(32)^{1/4}$

Etendue de la série: e= 1,20 – 0,85 en g/l = 0,35 g/l.

Classe en g/l	c_i	n_i	f_i	f_i^{cum}
[0,85 ; 0,91[	0,88	3	0,09	0,09
[0,91 ; 0,97[	0,94	4	0,13	0,22
[0,97 ; 1,03[	1,00	7	0,22	0,44
[1,03 ; 1,09[	1,06	8	0,25	0,69
[1,09 ; 1,15[	1,12	6	0,18	0,87
[1,15 ; 1,21]	1,18	4	0,13	1

La fonction de répartition et sa courbe cumulative des fréquences

$$F_X(x)=\begin{cases} 0 & si \quad x<0,85 \\ 0,09\dfrac{x-0,85}{0,35} & si \quad 0,85\le x<0,91 \\ 0,0,9+0,13\dfrac{0,x-0,91}{0,35} & si \quad 0,91\le x<0,97 \\ 0,22+0,22\dfrac{x-0,97}{0,35} & si \quad 0,97\le x<1,03 \\ 0,44+0,25\dfrac{x-1,03}{0,35} & si \quad 1,03\le x<1,09 \\ 0,69+0,18\dfrac{x-1,09}{0,35} & si \quad 1,09\le x<1,15 \\ 0,87+0,13\dfrac{x-1,15}{0,35} & si \quad 1,15\le x<1,21 \\ 1 & si \quad x\ge 1,21 \end{cases}$$