PHS : CH1 :

Statistique descriptive :

1] Def

* La statistique descriptive a pour but de résumer l'information contenue dans les données de façon à en dégager les caractéristiques essentielles sous une forme simple et intelligible
* Les deux principaux outils de la statistique descriptive sont :
  - les représentations graphiques
  - les indicateurs statistiques

2] Terminologie :

* **Données :** les mesures faites sur des individus issus d'une population.
* **Variables :** particularités des individus
* **Echantillon :** ensemble des individus.

* **recensement :** si l'échantillon est constitué de tous les individus de la population, on dit que l'on fait un recensement.

* **Sondage :** si l'échantillon n'est qu'une partie de la population, on parle de sondage.
* **Principe de sondage :** est d'étendre à l'ensemble de la population les enseignements tirés de l'étude de l'échantillon

* **statistique unidimensionnelle :** : on mesure qu'une seule variable sur les individus.

* **statistique multidimensionnelle :** : on mesure plusieurs variables sur les mêmes individus.

3] Variables discrètes / Continues :

* Les méthodes de représentation des données différent suivant la nature des variables étudiées

a] Variables statistique discrètes :

* Une Variable discrète est une Variable à valeurs dans un ensemble fini ou dénombrable.
* Variables quantitatives (ou numériques) : Sont des variables qui s'expriment par des nombres réels
* Variables qualitatives (ou catégorielles) : sont qui s'expriment par l'appartenance à une Catégorie

2°) Variables discrètes qualitatives

* Modalités : $E = \{e_1, \dots, e_n\}$: les n valeurs possible de la variable

* Fréquence absolue de la modalité e_j : nombre totale n_j d'individus de l'échantillon pour lesquels la variable a pris la modalité e_j.

$$n_j = \sum_{i=1}^{n} 1_{\{e_j\}}(x_i)$$

* Fréquence relative de la modalité e_j : pourcentage n_j/n d'individus de l'échantillon pour lesquels la variable a pris la modalité e_j.

* Représentations graphiques : Pour des V. qualitatives :
  - Diagramme en colonne ou en bâtons : à chaque modalité correspond un rectangle vertical dont la hauteur est proportionnelle à la fréquence relative de cette modalité.
  - Diagrammes Sectoriels ou Camembert : à chaque modalité correspond un secteur de disque dont l'aire (ou l'angle au centre) est proportionnelle à la fréquence relative de cette modalité

ii) Variables discrètes quantitatives :

* Seule différence : Ordre sur les modalités

b) Variables Continues :

* Déf : une Variable est continue est à valeurs dans un ensemble non dénombrable Ex : $\mathbb{R}$ ou $[a,b]$

Rq :

- Les représentations du type diagramme en bâtons sont sans intérêt car les données sont en général toutes distinctes, donc les fréquences absolues sont toutes égales à 1
- On a besoin lors d'ordonner les données : $x_1^*, x_2^*, \dots, x_n^*$.

* Histogrammes :

* classe : un intervalle $]a_{j-1}, a_j]$
  - Largeur de la classe : $h_j = a_j - a_{j-1}$
  - Effectif de la classe j : $n_j = \sum_{i=1}^{n} \mathbb{1}_{]a_{j-1}, a_j]}(x_i)$
  - Fréquence de la classe j : n_j/n

* Histogramme :

- est la figure constituée des rectangles dont les bases sont les classes $]a_{j-1}, a_j]$ et dont les aires sont égales aux fréquences de ces classes ie : la hauteur du $j^{ème}$ rectangle est $\frac{n_j}{n h_j}$.

* Règles conseillées :

- Règle de Sturges : le nombre de classe $k \simeq 1 + \log_2(n)$ n : nombre d'individus dans l'échantillon
- $a_0 = x_2^* - 0{,}025(x_n^* - x_1)$ et $a_k = x_n^* + 0{,}025(x_n^* - x_1^*)$.

Rq :

- $\forall j \in [\![1, n]\!]$ $h_j = cte = \frac{a_k - a_0}{k} \Rightarrow$ histogramme à pas fixe sinon histogramme à pas variable

* Propriétés :

i) Un histogramme fournit une estimation de la densité des observations

Preuve :

- Notons $\hat{f}$ la fonction en escalier constante sur les classes et $\forall x \in]a_{j-1}, a_j]$ $\hat{f}(x) = n_j / n h_j$
- Aire sous rectangle $= \frac{n_j}{n} = \int_{a_{j-1}}^{a_j} \hat{f}(x)\,dx$

$$\int_{a_{j-1}}^{a_j} \hat{f}(x)\,dx = \frac{n_j}{n}$$

or n_j/n le pourcentage d'observations appartenant à la classe j d'où $\frac{n_j}{n}$ c'est une estimation de la probabilité qu'une observation appartienne à cette classe d'où

$$n_j/n = \mathbb{P}(a_{j-1} < X \leq a_j) = \int_{a_{j-1}}^{a_j} f_X(x)\,dx$$

Ainsi $\int_{a_{j-1}}^{a_j} \hat{f}(x)\,dx = \int_{a_{j-1}}^{a_j} f_X(x)\,dx$

Donc :

* L'histogramme fournit une estimation de la densité des observations

$\Rightarrow$ L'estimation de la densité au pt $x \in]a_{j-1}, a_j]$, fait égale à $\hat{f}(x) = n_j / n h_j$

* L'allure de l'histogramme permettra donc de proposer des modèles probabilistes vraisemblables pour la loi de X en comparant la forme de $\hat{f}$ à celle de densité des lois de probabilité usuelles.

ii) si au lieu des effectifs n_j, On considère les effectifs cumulés $m_j = \sum_{\ell=1}^{j} n_\ell$, on construit un histogramme cumulé, qui fournit une estimation de la fonction de répartition de la variable étudiée