• Chapitre 0 : Généralités

→ Statistique descriptive : l'étude de la totalité de la population.

→ Statistique inferentielle : l'étude d'une partie de la populati[on] et puis généraliser.

↳ les avantages :

✓ Optimiser le coût et les ressources (humains, matériels)

✓ Prévoir dans le cas d'absence de toute la population

✓ Nécessité d'échantillonnage (vérifier la qualité des aliments).

* Définitions :

- La population : les éléments à étudier, un cible fini noté Ω

- Unité statistique : les éléments de Ω noté ω.

- Caractère statistique : C, X ou Y est un trait présent chez les éléments de Ω d'une manière :

↳ exhaustive : le caractère doit être présent chez tout les éléments de la population.

↳ incompatible : un élément doit avoir une valeur / modalité unique.

- Modalités : les différentes valeurs de C.

- Observation : les modalités avec répétition.

• exemple : étude des poids des étudiants d'une classe.

- population : les étudiants
- caractère : le poids
- modalité : valeur possible du caractère.

modalité → x_i	effectif n_i	f_i ~ fréquence relative
N		$\sum f_i = 1$

$f_i = \frac{n_i}{N}$ %

- **Preuve :**

soit l'application $f : \underset{\text{ensemble des étudiants}}{E} \longrightarrow \underset{\text{ensemble des poids.}}{F}$

→ si le caractère est exhaustif et incompatible alors f est définie.

exhaustif : $\equiv E = \bigcup_i A_i$ incompatible : $\equiv A_i \cap A_j = \{\emptyset\}$.

⇒ la distribution statistique peut être
- loi de probabilité.
- Application.
- partition de la population

* Typologie du caractère :

caractère statistique
- quantitatif
  - Discret (peu de modalité)
  - Continu (intervalle)
- qualitatif
  - nominal (pas d'ordre)
  - ordinal (relation d'ordre)
  - binaire (2 valeurs)

- Chapitre 1 : statistique descriptive univariée :

C'est la donnée d'une description statistique avec une seule variable ou caractère statistique à étudier.

- exemple : $\underbrace{\text{Poids}}_{\text{variable à expliquer}} = f^{\text{tion}}(\underbrace{\text{Taille, Age, Sexe, Région}}_{\text{variable explicative}})$

⇒ on cherche les variables explicatives qui interviennent et affectent plus la variable à expliquer.

① Représentation graphique :

• Caractère qualificatif :

On utilise principalement 3 types de représentation graphique : le diagramme en bâtons, tuyaux d'orgue et par secteurs.

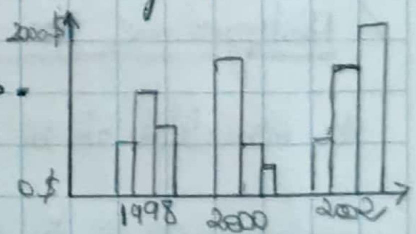
Lorsque le caractère étudié est la répartition géographique d'une population, la représentation graphique est un cartogramme.

↳ **Diagramme en bâtons :**

- Nous portons en abscisse les modalités, de façon arbitraire.
- Nous portons en ordonnée des segments dont la longueur est proportionnelle aux effectifs (ou aux fréquences) de chaque modalité.
- Nous appelons polygone statistique, ou diagramme polygonal, la ligne obtenue en joignant les sommets des bâtons.

↳ **Tuyaux d'orgue :**

- Nous portons en abscisses les modalités, de façon arbitraire.
- Nous portons en ordonnées des rectangles dont la longueur est proportionnelle aux effectifs, ou aux fréquences.



↳ **Secteurs :**

- Ces diagrammes conviennent très bien pour des données politiques ou socio-économiques.

Scanné avec CamScanner

- Caractère quantitatif :

Il existe deux types de représentation graphique :

✓ différentiel : correspond à une représentation des effectifs ou des fréquences.

✓ Intégral : correspond à une représentation des effectifs cumulés, ou des fréquences cumulées.

↳ Variable statistique discrète :

✓ Diagramme en bâtons, la différence avec le cas qualitatif consiste en ce que les abscisses ici sont des valeurs de la variable statistique.

✓✓ Courbe en escaliers des effectifs cumulés ou des fréquences

↳ Variable statistique continue :

Les observations sont regroupées en classes, chaque classe possède une certaine amplitude, qui est la longueur de I définissant la classe.

- Densité d'effectif : le rapport entre n_i et son amplitude.
- Densité de fréquence : le rapport entre f_i et son amplitude.

✓ Histogramme des densités : en abscisse les classes représentant les modalités et en ordonnées des rectangles dont la longueur est proportionnelle à la densité d'effectif ou de fréquence.

✓✓ Courbe cumulative des effectifs ou des moyennes

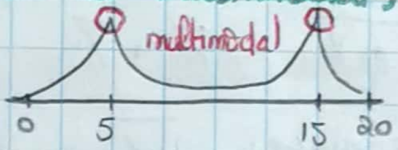
② Paramètres statistiques : qui caractérisent la distribution

↳ **De position :** (entre un max et min)

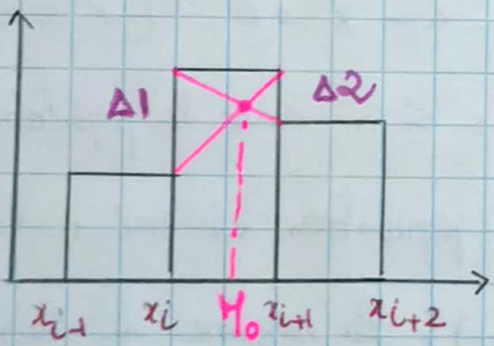
- Les paramètres qui permettent de savoir autour de quelles valeurs se situent les valeurs d'une variable statistique.

* **mode** : modalité qui a la plus grande fréquence

→ si une distribution est multimodal, donc la population est hétérogène.



* **La classe modale** : pour une variable quantitative continue.



$$\frac{M_0 - x_i}{\Delta 1} = \frac{x_{i+1} - M_0}{\Delta 2}$$

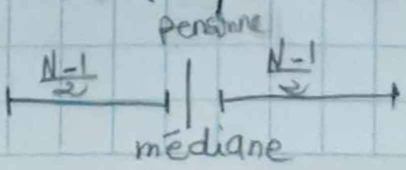
$$M_0 = x_i + \frac{\Delta 1}{\Delta 1 + \Delta 2}(x_{i+1} - x_i)$$

* **La médiane** : M_e c'est une modalité qui va diviser la population en deux parties :

1ère partie : modalité < médiane
2ème partie : modalité > médiane
} de même taille
$P = 14$

Elle est défini : si la description est quantitative, si la relation d'ordre existe → quantitative ordinal.

Si N est impaire :



* la moyenne : définie quand la distribution est quantitative.

$$\overline{X} = \frac{1}{N} \sum \overset{\text{modalité}}{X}(\omega) = \frac{1}{N} \sum n_i x_i = \sum f_i x_i$$

⇒ propriétés de la moyenne :

- Somme : X et Y 2 distrib stat quantitatif sur Ω alors X + Y est une distrib stat sur Ω

$$X : \Omega \longrightarrow \mathbb{R}$$
$$Y : \Omega \longrightarrow \mathbb{R}$$
$$X + Y : \Omega \longrightarrow \mathbb{R}$$

① $\omega \longrightarrow (X+Y)(\omega) = X(\omega) + Y(\omega)$

② $\overline{X+Y} = \overline{X} + \overline{Y}$

- Produit par un scalaire : $\overline{\lambda X} = \lambda \overline{X}$
- Ecart moyen à la moyenne : $\overline{X - \overline{X}} = 0 \quad (\overline{\overline{X}} = \overline{X})$

* la moyenne conditionnée : si on divise la population à des sous populations, chaque sous population est trié à l'aide d'une condition (Région, niveau, ...). La moyenne de chaque sous population est une moyenne conditionnée. et la moyenne de tous les moyennes conditionnées est la moyenne (générale).

- La moyenne c'est la moyenne des moyennes conditionnés pondérés par la taille des parties. $\overline{X}^j = \frac{1}{n_j} \sum n_{ij} x_i$

* la moyenne d'une variable continu :

x_i	n_i
[[	7
[[	30
[[	10

$\overline{X} = \frac{1}{N} \sum_{i=1}^{p} n_i \overline{X}^i$ avec $\overline{X}^i$ la marge à l'intérieur de $[x_i ; x_{i+1}[$

Scanné avec CamScanner

Selon l'hypothèse 1 : Dans chaque classe, toute les observations sont concentrées au centre de la classe.

et l'hypothèse 2 : la répartition des observations est uniforme.

$$\bar{x}^i = \frac{x_i + x_{i+1}}{2}$$

* Généralisation de la notion de moyenne :

↳ la φ-moyenne :

Soit X une variable stat réelle sur Ω : $\varphi : \mathbb{R} \longrightarrow \mathbb{R}$
application localement inversible

on peut calculer la moyenne de φ : $\bar{\varphi}(X) = \frac{1}{N} \sum n_i \varphi(x_i)$

$\varphi \circ X : \Omega \longrightarrow \mathbb{R}$ est une distrib stat réelle sur Ω

$\bar{X}^\varphi = \varphi^{-1}(\overline{\varphi \circ X})$ ⇒ la φ moyenne de X.

- exemple

1/ la moyenne arithmétique : $\varphi : x \longrightarrow x$ app identique

$$\bar{X}^\varphi = Id^{-1}\left(\frac{1}{N} \sum n_i \, Id(x_i)\right) = \bar{X}$$

2/ la moyenne quadratique : $\varphi : x \longrightarrow x^2$

$$\bar{X}^\varphi = \sqrt{\frac{1}{N} \sum n_i x_i^2} = \bar{X}_q$$

3/ la moyenne harmonique : $\varphi : x \longrightarrow 1/x$.

$$\bar{X}^\varphi = \left(\frac{1}{N} \sum \frac{n_i}{x_i}\right)^{-1} = \bar{X}_h$$

4/ la moyenne géométrique: $\varphi : x \longrightarrow \ln(x)$

$$\bar{x}^{\varphi} = \exp\left(\frac{1}{N}\sum n_i \ln(x_i)\right)$$
$$= \sqrt[n]{\prod_{i=1}^{p} x_i^{n_i}} = \bar{x}_g$$

* Propriétés de la φ-moyenne :

$$\bar{x}_h \leq \bar{x}_g \leq \bar{x} \leq \bar{x}_q$$

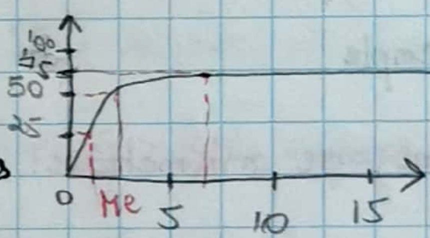
- Il y a égalité ssi toutes les valeurs de x sont égales.
- $\bar{x}_g$ est bien adaptée à l'étude des phénomènes de croissance
- $\bar{x}_h$ est utilisée pour les calculs d'indices économiques.

↳ de dispersion:

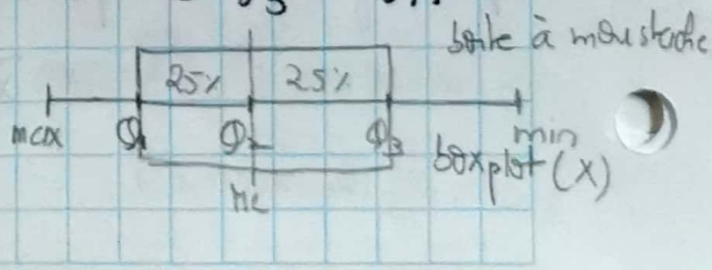
* étendue: $w = x_{max} - x_{min}$
* Quartile: généralisation de la médiane

• variable statistique continue:

on appelle quartiles les nombres Q_1, Q_2 et Q_3 pour lesquels les f cumulées de X sont respectivement 0,25 ; 0,50 ; 0,75.

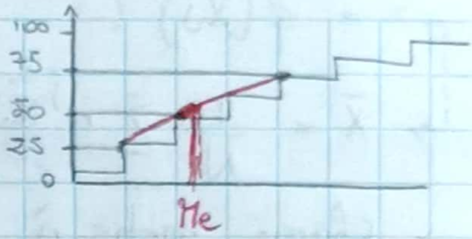


- les quartiles partagent l'étendue en 4 I qui ont le même effectif.
- le Q_2 est égal à la médiane
- L'intervalle interquartile est la différence entre Q_3 et Q_1, il contient 50% des valeurs de X.



• variable statistique discrète:

les quartiles seront déterminés par interpolation linéaire entre deux valeurs.



* percentiles : diviser I en 100 I_i de même effectif.

* déciles : partager l'étendue en 10 de même effectif.

* Ecart absolu moyen : $e = \frac{1}{N} \sum n_i \, | x_i - \overline{x} |$. difficile à manipuler

* Variance : La dispersion autour de $\overline{x}$

notation

$$V(x) = \delta^2(x) = \frac{1}{N} \sum n_i (x_i - \overline{x})^2$$

⇒ l'écart type : $\delta(x) = \sqrt{\frac{1}{N} \sum n_i (x_i - \overline{x})^2}$

König

⇒ formule de la variance : la moyenne des carrés - le carré de la moyenne : $V(x) = \delta^2(x) = \overline{x^2} - \overline{x}^2$

⇒ propriété de la variance :

- $V(x) \geq 0 \quad \forall x$
- $V(x) = 0$ ssi x est cte
- $\delta^2(ax + b) = a^2 \delta^2(x)$

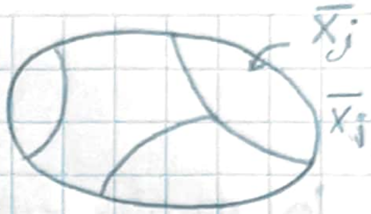
⇒ Inertie : C'est une généralisation de la variance, la moyenne des carrés des écarts à une référence a.

$$I_a(x) = \frac{1}{N} \sum n_i (x_i - a)^2$$

$$I_a(x) = \delta^2(x) + (a - \overline{x})^2$$

* Variance conditionnée :

X la distribution stat si : $\Omega = \cup X_j$



$\overline{X} = f(\overline{X_j})$?

$\hookrightarrow \overline{X} = \frac{1}{N} \sum n^j \overline{X_j}$

→ Chaque partie à une moyenne conditionnée et variance conditionnée, pour la moyenne $\overline{X}$ c'est la moyenne des moyenne conditionnée pondérée par l'effectif, Est-ce que c'est le cas de la variance ?

• $\delta^2(X) = f(\delta^2_j(X), \overline{X_j})$

$$\delta^2(X) = \underbrace{\frac{1}{N}\sum_1^q n_j\, \delta^2_j(X)}_{\delta^2_w(X)} + \underbrace{\frac{1}{N}\sum n_j(\overline{X_j} - \overline{X})^2}_{\delta^2_b(X)}$$

variance intraclasse (dispersion / à la $\overline{X}$)

variance interclasse. (dispersion entre # les classes)

⇒ c'est la moyenne des variances conditionnée + la variance des moyennes conditionnées.

• Classification : Quand X dépend de y mais X n'est pas numérique. $f(P, t)$ ↗ G ↘ F

• Corrélation : Lorsqu'une variable stat dépend d'une condition.

↳ Coefficient de corrélation de X en $(\cup \Omega_j)$:

pas de corrélation → $0 \leqslant R^2 = \frac{\delta^2_b(X)}{\delta^2(X)} \leqslant 1$ ← forte corrélation

* si $R = 1$: ⇒ $\delta^2_w(X) = 0$ → $\delta^2_j(X) = 0$ → X^j est ct

exemple : toute la population d'une classe a 18 en note → corrélation

* si $Q = 0 \Rightarrow \delta_b^2(X) = 0$

exemple : toutes les classes ont une moyenne = 18, X ne dépend pas de la classe.

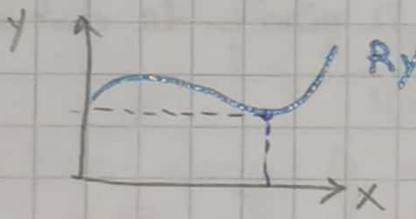
∘ Remarques : p 26   X et Y deux variables stat sur Ω

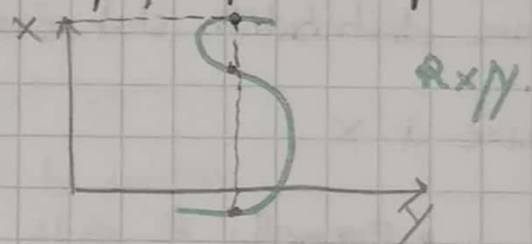
$$\left.\begin{array}{l}X \text{ une variable stat} \\ Y \text{ une partition}\end{array}\right\} \Rightarrow R_{X/Y} = \frac{\delta_b^2(X)}{\delta^2(X)}$$

dépend de X et Y car δ_b^2 c'est la variance des moyennes conditionnée

$$\left.\begin{array}{l}Y \text{ une variable stat} \\ X \text{ une partition}\end{array}\right\} \Rightarrow R_{Y/X} = \frac{\delta_b^2(Y)}{\delta^2(Y)}$$

* si X dépend de Y cela n'implique pas Y dépend de X





∘ $\delta^2(X) = \frac{1}{N} \sum n_i \, \delta_i^2(X) + \frac{1}{N} \sum n_i (\bar{X}_i - \bar{X})^2$

↳ selon hypo 1 : $\delta_i^2(x) = 0$ pas de dispersion.

$\delta^2(x) = \delta_b^2(\;)$.

↳ selon hypo 2 : $\delta_i^2(x) = \frac{(x_{i+1} - x_i)^2}{12}$

$\delta_i^2(x) = \overline{X^{j2}} - \overline{X^j}^2$

* Moments :

↳ moment d'ordre r : $m_r = \frac{1}{N} \sum [X(\omega)]^r = \frac{1}{N} \sum n_i x_i^r$

c'est la $(\bar{X})^r$

- si $r = 0$: $m_0 = 1$
- si $r = 1$: $m_1 = \overline{X}$
- si $r = 2$: $m_2 = \overline{X^2}$

↳ moment centré d'ordre r : $\mu_r = \frac{1}{N} \sum (X(\omega) - \overline{X})^r = \frac{1}{N} \sum n_i (x_i - \overline{x})^r$

- si $r = 0$: $\mu_0 = 1$
- si $r = 1$: $\mu_1 = 0$
- si $r = 2$: $\mu_2 = S^2(X) = m_2 - m_1^2$, le moment centré d'ordre 2 c'est la variance.

- Distribution statistique centrée : $\overline{X} = 0$

Soit X une distribution statistique réelle sur Ω alors $X - \overline{X}$ est une distribution stat réelle sur Ω nomée la dist stat associée à X.

- Exemple : le poids des étudiants $\overline{P} = 50$ kg

$P - \overline{P} =$ (pour $P = 50$ kg) 0

$P = 48 \rightsquigarrow -2$ devient le nouveau poids dans cette dist

$P = 52 \rightsquigarrow 2$ " " " " "

- Distribution centrée réduite :

$\frac{X - \overline{X}}{\sigma(X)}$ avec : $\overline{X} = 0$ et $\sigma(X) = 1$

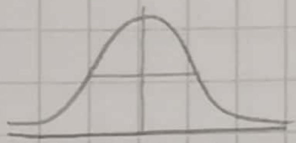
$V(ax+b) = a^2 V(X)$ $\quad V\left(\frac{X - \overline{X}}{\sigma(X)}\right) = \frac{1}{\sigma^2(X)} V(X) = 1$

↳ Ap

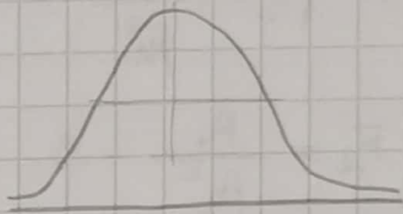
⇒ prq centrer et réduire une distribution ?

pour qu'on puisse la comparer à une référence.

• Exemple:



loi normal



si je souhaite comparer cette distn à la loi normal, on doit la redimensionner.

↳ De forme:

* Coefficient d'asymétrie:

① Pearson: intervenir le mode

$$P = \frac{\bar{X} - M_o}{s(X)}$$

le mode

l'écart type

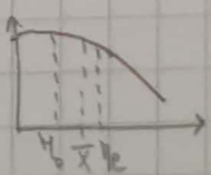
② Yule: intervenir la médiane et les quartiles

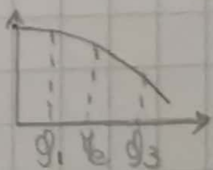
$$Y = \frac{Q_1 + Q_3 - 2M_e}{2(Q_3 - Q_1)}$$

③ Fisher: intervenir les moments centrés

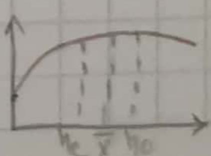
$$F = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{s^3(X)}$$

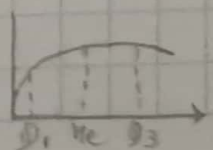
• Si le coefficient > 0 : on dit qu'il y a obliquité à gauche.





• Si le coefficient < 0 : on dit qu'il y a obliquité à droite





- Pearson basé sur les moments centrés :

nous donne la précision ← $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$ → tjrs positif

→ nul pour une dist à densité de fréquence symétrique, telle que la loi de Gauss

* Coefficient d'aplatissement :

① Pearson : $\beta_2 = \frac{\mu_4}{\mu_2^2}$

② Yule : $F_2 = \frac{\mu_4}{\mu_2^2} - 3$