Chapitre 3 : Les caractéristiques de dispersion

Objectif général du chapitre :

Acquérir l'habilité d'identifier et de différencier les caractéristiques de dispersion.

Introduction :

L'intérêt de ce chapitre sera analysé en suivant ces deux exemples :

n_i	1	1	1
x_i	9	10	11

Pour l'exemple 1 : $\overline{X}$ = 10. M_e = 10.

n_i	1	1	1
x_i	0	10	20

Pour l'exemple 2 : $\overline{X}$ = 10 M_e = 10.

On constate, alors que les deux distributions ont les mêmes paramètres de tendance centrale ($\overline{X}$ et M_e). Pourtant, elles sont assez différentes. En effet, l'étalement ou la dispersion des valeurs autours de la moyenne est plus grand pour la 2ème distribution. D'où, l'insuffisance des paramètres de tendance centrale et la nécessité de définir d'autres types de paramètres pour saisir la différence réelle entre deux ou plusieurs distributions (qui peuvent paraître en apparence semblables).

On va, donc, présenter quatre paramètres de dispersion suivant un ordre croissant de leur importance et de leur fiabilité, de point de vue précision dans l'analyse et l'interprétation des résultats.

I. L'étendue

Objectifs spécifiques :

- Maîtriser l'étendue.

Durée : 0,25 H.

Contenu :

L'étendue désigne la différence entre les deux valeurs limites d'une distribution (la modalité maximale est celle minimale). Il est noté par : $e = x_{max} - x_{min}$.

Ce paramètre est imparfait car il ne dépend que des deux valeurs extrêmes en ignorant l'intensité de la dispersion interne qui peut exister entre les autres modalités intermédiaires.

II. Les intervalles interquantiles

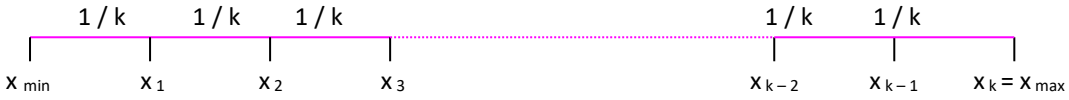
Objectifs spécifiques :

- Comprendre les intervalles interquantiles.

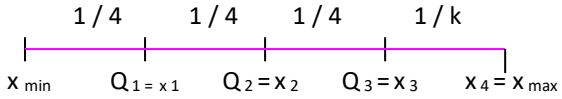
Durée : 1,45 H.

Contenu :

a/ Les quantiles : On appelle quantile d'ordre k, les valeurs de la variable qui partagent l'effectif total de la population en k sous-ensembles d'effectifs égaux (à condition que les observations soient ordonnées d'ordre croissant).



b/ Les quartiles : Sont des quantiles d'ordre k = 4. La distribution présente, alors, trois quartiles (k–1).



- 1er Quartile : On cherche Q_1, tel que : F (Q_1) = $\frac{1}{4}$: 25% des observations présentent une modalité du caractère inférieur à Q_1.
- 2ème Quartile : On cherche Q_2, tel que : F (Q_2) = $\frac{2}{4}$: 50% des observations présentent une modalité du caractère inférieur à Q_2 (Q_2 et M_e sont toujours confondus).
- 3ème Quartile : On cherche Q_3, tel que : F (Q_3) = $\frac{3}{4}$: 75% des observations présentent une modalité du caractère inférieur à Q_3.

L'intervalle interquartile : ($IIQ = Q_3 - Q_1$) :

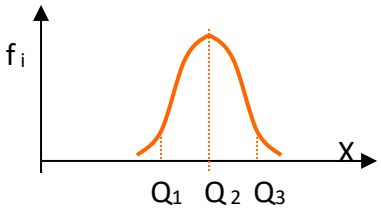
Il contient 50% des observations en laissant 25% à gauche de Q_1 et 25% à droite de Q_3.

Il permet de comparer la dispersion de deux ou plusieurs distributions, on dit, alors, plus IIQ est petit, plus la dispersion de la distribution correspondante est faible et plus sa population est homogène (et vis versa).

L'intervalle interquartile relatif : ($IIQR = \frac{IIQ}{M_e} = \frac{Q_3 - Q_1}{Q_2}$) : est utilisé pour interpréter la dispersion des distributions qui n'ont pas la même unité.

Remarque : on a les identités suivantes :

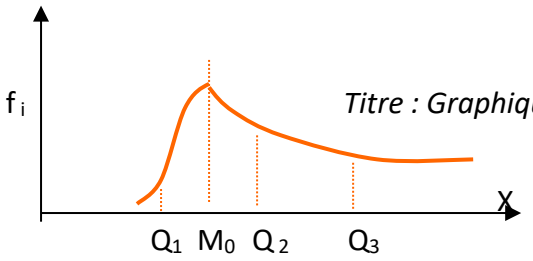
- Si la distribution est parfaitement symétrique : $Q_3 - Q_2 = Q_2 - Q_1$ ou bien, $\frac{Q_3 - Q_2}{Q_2 - Q_1} = 1$.



Titre : Graphique 1 : Distribution symétrique.

- Si la distribution est asymétrique à droite : $Q_3 - Q_2 > Q_2 - Q_1$ ou bien, $\frac{Q_3 - Q_2}{Q_2 - Q_1} > 1$.

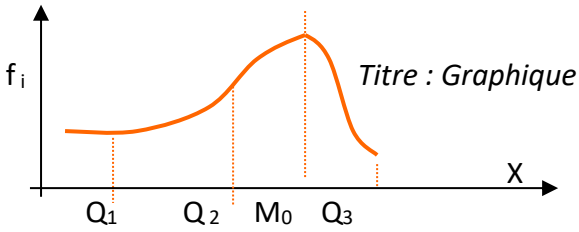
(L'étalement est à droite)



Titre : Graphique 2 : Distribution asymétrique à gauche.

- Si la distribution est asymétrique à gauche : $Q_3 - Q_2 < Q_2 - Q_1$ ou bien, $\frac{Q_3 - Q_2}{Q_2 - Q_1} < 1$.

(L'étalement est à gauche)



Titre : Graphique 2 : Distribution asymétrique à droite.

Exemple d'application : Cherchez IIQR de l'âge de ces salariés :

Age	f_i	FC ↑
[20, **25** [	0,15	**0,15**
[25, **30** [	0,11	**0,26**
[30, **34** [	0,13	**0,39**
[34, **50** [	0,21	**0,6**
[50, **56** [	0,2	**0,8**
[56, 60 [	0,2	1
Total	1	

Titre : Tableau 1 : La répartition des fréquences des salariés selon leurs âges.

- Q_1, tel que : $F(Q_1) = \frac{1}{4} = 0{,}25$, alors on : $Q_1 \in [\ 25, 30\ [$ et on peut écrire :

$$\begin{cases} 25 \rightarrow 0{,}15 \\ Q_1 \rightarrow 0{,}25 \\ 30 \rightarrow 0{,}26 \end{cases}$$

En appliquant l'interpolation linéaire, on obtient : $\frac{Q_1-25}{30-25} = \frac{0,25-0,15}{0,26-0,15}$, d'où, $Q_1 = 29,55$ ans : 25% des salariés présentent un âge inférieur à 29,55 ans.

- F (Q_2) = $\frac{2}{4}$ = 0,5 alors on : $Q_2 \in$ [34, 50 [et on peut écrire : $\begin{cases} 34 \rightarrow 0,39 \\ Q_2 \rightarrow 0,5 \\ 50 \rightarrow 0,6 \end{cases}$

En appliquant l'interpolation linéaire, on obtient : $\frac{Q_2-34}{50-34} = \frac{0,5-0,39}{0,6-0,39}$, d'où, $Q_2 = 42,38$ ans : 50% des salariés présentent un âge inférieur à 42,38 ans.

- F (Q_3) = $\frac{3}{4}$ = 0,75 alors on : $Q_3 \in$ [50, 56 [et on peut écrire : $\begin{cases} 50 \rightarrow 0,6 \\ Q_3 \rightarrow 0,75 \\ 56 \rightarrow 0,8 \end{cases}$

En appliquant l'interpolation linéaire, on obtient : $\frac{Q_3-50}{56-50} = \frac{0,75-0,6}{0,8-0,6}$, d'où, $Q_3 = 54,5$ ans : 75% des salariés présentent un âge inférieur à 54,5 ans.

IIQR = $\frac{Q_3-Q_1}{Q_2}$ = 0,58.

c/ Les déciles : Sont des quantiles de l'ordre k = 10, une distribution présente, alors, neuf déciles (k – 1).

- 1er Décile : On cherche D_1, tel que : F (D_1) = $\frac{1}{10}$: 10 % des observations présentent une modalité du caractère inférieur à D_1.
- 2ème Décile : On cherche D_2, tel que : F (D_2) = $\frac{2}{10}$: 20 % des observations présentent une modalité du caractère inférieur à D_2.
- 5ème Décile : On cherche D_5, tel que : F (D_5) = $\frac{5}{10}$: 50 % des observations présentent une modalité du caractère inférieur à D_5. (D_5 est toujours confondu avec M_e).
- Ième Décile : On cherche D_i avec i varie entre 1 et 9, tel que : F (D_i) = $\frac{i}{10}$: (10 × i)% des observations présentent une modalité du caractère inférieur à D_i.

Et ainsi, jusqu'à D_9.

- 9ème Décile : On cherche D_9, tel que : F (D_9) = $\frac{9}{10}$: 90 % des observations présentent une modalité du caractère inférieur à D_9.

L'intervalle inter décile : (IID = $D_9 - D_1$) : il contient 80 % des observations en laissant 10% à gauche de D_1 et 10% à droite de D_3.

Il permet de comparer la dispersion de deux ou plusieurs distributions, on dit, alors, plus IID est petit, plus la dispersion de la distribution correspondante est faible et plus sa population est homogène (et vis versa).

L'intervalle inter décile relatif : (IIDR = $\frac{IID}{M_e} = \frac{D_9 - D_1}{D_5}$)

d/ Les centiles : Sont les quantiles de l'ordre k = 100 ; une distribution présente, alors, 99 centiles (k – 1).

- 1er Centile : On cherche C_1, tel que : F (C_1) = $\frac{1}{100}$: 1 % des observations présentent une modalité du caractère inférieur à C_1.
- 2ème Centile : On cherche C_2, tel que : F (C_2) = $\frac{2}{100}$: 2 % des observations présentent une modalité du caractère inférieur à C_2.
- 50ème Centile : On cherche C_{50}, tel que : F (C_{50}) = $\frac{50}{100}$: 50 % des observations présentent une modalité du caractère inférieur à C_{50}. (C_{50} est toujours confondu avec M_e).
- Ième Centile : On cherche C_i avec i varie entre 1 et 99, tel que : F (C_i) = $\frac{i}{100}$: (i)% des observations présentent une modalité du caractère inférieur à C_i.

Et ainsi, jusqu'à C_{99}.

- 99ème Centile : On cherche C_{99}, tel que : F (C_{99}) = $\frac{99}{100}$: 99 % des observations présentent une modalité du caractère inférieur à C_{99}.

L'intervalle inter centile : (IIC = $C_{99} - C_1$) : il contient 98 % des observations en laissant 1% à gauche de C_1 et 1% à droite de C_{99}.

Il permet de comparer la dispersion de deux ou plusieurs distributions, on dit, alors, plus IIC est petit, plus la dispersion de la distribution correspondante est faible et plus sa population est homogène (et vis versa).

L'intervalle inter centile relatif : (IICR = $\frac{IIC}{M_e} = \frac{C_{99} - C_1}{C_{50}}$)

Remarques :

- Les quartiles, les déciles et les centiles se déterminent de la même manière que la médiane, soit analytiquement, soit graphiquement.
- La logique de cette section est présentée exprès de cet ordre, afin de montrer que l'analyse de ces intervalles diffère suivant leurs ordres : plus on augmente l'ordre k, plus on aboutit à un résultat meilleur et précis.

III. L'écart absolu moyen

Objectifs spécifiques :

- Détecter la formule appropriée pour déterminer l'écart absolu moyen.

Durée : 0,5 H.

Contenu :

L'écart absolu moyen noté par e :

a/ Définition : L'écart absolu moyen est la moyenne arithmétique des différences (en valeur absolue) entre chaque terme de la série et la valeur d'une caractéristique de tendance centrale (M_e ou $\overline{X}$). On distingue, alors, deux écarts :

- L'écart absolu moyen par rapport à la médiane : est une moyenne arithmétique des écarts entre les valeurs particulières de la variable et la médiane. On va distinguer trois formules suivant le cas :

♠ Des observations individuelles : $EAM_{(Me)}=\frac{1}{n}\sum_{i=1}^{n}|x_i-M_e|$:c'est un écart absolu moyen simple

♠ Une variable discrète : $EAM_{(Me)} = \frac{1}{n}\sum_{i=1}^{k}n_i|x_i-M_e|$: c'est un écart absolu moyen pondéré.

♠ Une variable continue : $EAM_{(Me)} = \frac{1}{n}\sum_{i=1}^{k}n_i|c_i-M_e|$: c'est un écart absolu moyen pondéré.

- L'écart absolu moyen par rapport à la moyenne : est une moyenne arithmétique des écarts à la moyenne. On va distinguer deux formules suivant le cas :

♠ Des observations individuelles : $EAM_{(\overline{X})} =\frac{1}{n}\sum_{i=1}^{n}|x_i-\overline{X}|$: c'est un écart absolu moyen simple

♠ Une variable discrète : $EAM_{(\overline{X})} = \frac{1}{n}\sum_{i=1}^{k}n_i|x_i-\overline{X}|$: c'est un écart absolu moyen pondéré.

♠ Une variable continue : $EAM_{(\overline{X})} = \frac{1}{n}\sum_{i=1}^{k}n_i|c_i-\overline{X}|$: c'est un écart absolu moyen pondéré.

b/ Exemple d'application : Calculez les deux écarts absolus moyens de cette distribution. La médiane est égale à 3.

X	f_i	FC↑	$f_i \times x_i$	$\lvert x_i - \overline{X} \rvert$	$f_i \lvert x_i - \overline{X} \rvert$	$\lvert x_i - Me \rvert$	$f_i \lvert x_i - Me \rvert$
1	0,06	0	0,06	2,34	0,14	2	0,12
2	0,26	0,06	0,52	1,34	0,34	1	0,26
3	0,35	0,32	1,05	0,34	0,12	0	0
4	0,11	0,67	0,44	0,66	0,072	1	0,11
5	0,05	0,78	0,25	1,66	0,083	2	0,1
6	0,17	0,83	1,02	2,66	0,452	3	0,51
Total	1		$\overline{X}$ = 3,34		EAM / $\overline{X}$ = 1,2		EAM / $_{Me}$ = 1,1

Titre: Tableau 2 : Calcul des caractéristiques de dispersion (EAM).

IV. La variance et l'écart type

Objectifs spécifiques :

- Mémoriser et appliquer les principes de calcul de la variance.

Durée : 2 H.

Contenu :

a) *Définition :* la variance d'une variable statistique X, notée V(X), est définie comme suit:

♠ Cas de données individuelles : $V(X) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{X})^2$

♠ Cas de variable discrète : $V(X) = \frac{1}{n} \sum_{i=1}^{k} n_i (x_i - \overline{X})^2 = \sum_{i=1}^{k} f_i (x_i - \overline{X})^2$

♠ Cas de variable continue : $V(X) = \frac{1}{n} \sum_{i=1}^{k} n_i (c_i - \overline{X})^2 = \sum_{i=1}^{k} f_i (c_i - \overline{X})^2$

Puisque V(X) est exprimée au carrée, donc, son unité est, automatiquement, le carré de celle de X. Alors, il est difficile de saisir la signification ou l'interprétation de V(X). D'où, on calcul :

σ (X) = $\sqrt{V(X)}$: c'est l'écart type de X. Il sert surtout pour comparer la dispersion entre 2 ou plusieurs distributions, on dit, alors :

- Plus σ est petit, plus la dispersion est faible dans cette distribution, et plus la population correspondante est homogène.
- Plus σ est grand, plus la dispersion est forte dans cette distribution, et plus la population correspondante est hétérogène.

b) La formule développée[1] de la variance : est plus facile à appliquer :

- Cas de données individuelles : $V(X) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{X})^2$
- Cas de variable discrète : $V(X) = \frac{1}{n}\sum_{i=1}^{k} n_i \times x_i^2 - (\overline{X})^2 = \sum_{i=1}^{k} f_i x_i^2 - (\overline{X})^2$
- Cas de variable continue : $V(X) = \frac{1}{n}\sum_{i=1}^{k} n_i c_i^2 - (\overline{X})^2 = \sum_{i=1}^{k} f_i c_i^2 - (\overline{X})^2$

c) Propriétés : on va citer deux principales propriétés :

❖ Propriété de l'union : Soit deux populations P_1 et P_2, d'effectifs respectifs n_1 et n_2, de moyennes arithmétiques respectives $\overline{X}_1$ et $\overline{X}_2$. Alors, la population totale $P = P_1 \cup P_2$, d'effectif total $n = n_1 + n_2$, et ayant pour moyenne arithmétique :

$\overline{X} = \frac{n_1 \overline{X}_1}{n} + \frac{n_2 \overline{X}_2}{n} = f_1 \times \overline{X}_1 + f_2 \overline{X}_2$, et, possède la variance totale suivante :

$$V(X) = \left[\frac{1}{n}\sum_{i=1}^{2} n_i \times V(X_i) \right] + \left[\frac{1}{n}\sum_{i=1}^{2} n_i \times \overline{(X_i)^2} - (\overline{X})^2 \right]$$

V(X) = variance intra-population (interne dans chaque population à part)+ variance inter-population (entre les populations).

Cette propriété peut être généraliser pour k populations unies.

❖ Propriété de la linéarité : soit deux variables statistiques X et X', tels que : $X_i = a x'_i + x_0$ quel que soit i. Alors, $V(X) = a^2 V(X')$ et $\sigma(X) = |a| \sigma(X')$

L'intérêt pratique de cette propriété est résumé dans deux avantages :

♪ Elle facilite le calcul de la variance, comme dans le cas de la série de salaire (cité au chapitre précédent).

♪ Elle facilite, aussi, le calcul de la variance en effectuant un changement de variable (on suit pratiquement la même démarche suivie pour le cas de la moyenne).

Exemple d'application : Calculer V (X) en utilisant le changement de variable.

X	n_i	c_i	c'_i	$n_i \times c'_i$	$n_i \times (c'_i)^2$
[50, 100 [	5	75	- 2	- 10	20
[100, 150 [	12	125	- 1	- 12	12
[150, 200 [	8	175	0	0	0
[200, 250 [	4	225	1	4	4
[250, 300 [	2	275	2	4	8
Total	31			- 14	44

Titre : Tableau 3 : Calcul de moyenne et de variance par changement de variable.

[1] C'est la formule de Konig Hugghens.

1ère étape : $c_0 = \frac{C_1 + C_k}{2} = \frac{C_1 + C_5}{2} = \frac{75+275}{2} = 175$, $a = 50$, $c'_i = \frac{C_i - C_0}{a} = \frac{C_i - 175}{50}$.

2ème étape : $V(X') = \frac{1}{n}\sum_{i=1}^{k} n_i c'^2_i - \left(\overline{X'}\right)^2 = \frac{44}{31} - \left(\frac{-14}{31}\right)^2 = 1{,}215$ et $\sigma(X') = 1{,}102$.

3ème étape : $V(X) = a^2 V(X') = 3\,037{,}5$ et $\sigma(X) = |a|\ \sigma(X') = 55{,}113$.

d/ Coefficient de variation : noté CV :

$CV = \sigma_X / \overline{X}$

L'utilité de ce coefficient réside dans deux choses :

- Il permet de comparer la dispersion des distributions de moyennes différentes, puisque σ_X ne suffit pas dans ce cas.
- Il permet de comparer la dispersion des distributions exprimées dans des unités de mesure différentes.

Ce coefficient est interprété de la même manière que σ_X : plus CV est petit, plus la dispersion de la distribution correspondante est faible, et plus sa population est homogène ; et vice versa.

Remarque : σ_X (d'unité de X) nous donne la dispersion absolue des distributions ; alors que, le CV (sans unité) nous donne leur dispersion relative.