Module statistique et probabilités _ partie 3
Zahra ROYER

→ *B-2- Caractéristiques de Dispersion*
Les caractéristiques de position ou tendance centrale renseignent seulement sur l'ordre de grandeur de la série. On les complète alors par les outils

- qui décrivent comment sont groupées ces informations
- ceux qui nous disent comment est la dispersion c'est-à-dire, sur les fluctuations des valeurs autour d'une valeur centrale, moyenne ou médiane ou autre

B-2-1 Les indicateurs utilisés sont

- ***L'intervalle de variation, l'étendue***
- ***Les intervalles interquantiles***
- ***L'écart absolu moyen***
- ***La variance- L'écart-type***
- ***Le coefficient de variation***
- ***Variance inter et intra populations.***

→ L'intervalle de variation, l'étendue
C'est la manière la plus simple d'appréhender la notion de dispersion.
L'étendue (ou l'intervalle de variation) est la différence entre la plus grande et la plus petite des valeurs de la série.
Elle ne présente que peut d'intérêt, car elle est fortement dépendante de l'échantillon utilisé.

→ Les intervalles interquantiles
Pour éviter d'effectuer des calculs sur des valeurs extrêmes souvent aberrantes, face au phénomène étudié, on choisit souvent, de les écarter de la série.
On perd de l'information mais on y gagne en homogénéité.
Question : pour des séries financières que signifie écarter les valeurs extrêmes ?

On utilise en général :

- L'intervalle interquartile $[Q_1 \quad , \quad Q_3]$ est celui qui contient approximativement 50% des valeurs ;
- L'écart interquartile est égal à $Q_3 - Q_1$
- L'intervalle interdécile $[D_1 \quad , \quad D_9]$ est celui qui contient approximativement 80% des valeurs.
- L'écart interdécile est égal à $D_9 - D_1$

→ *L'écart absolu moyen*

C'est une moyenne arithmétique d'écarts par rapport à une valeur centrale : moyenne ou médiane.

<u>Donc l'écart absolu moyen par rapport à la moyenne est</u>

$\bar{e}_{\bar{x}} = \dfrac{\sum_i n_i |x_i - \bar{x}|}{N}$. Noter qu'il faut prendre la valeur absolue des écarts.

<u>Et l'écart absolu moyen par rapport à la médiane :</u>

$$\bar{e}_{M_é} = \frac{\sum_i n_i |x_i - M_é|}{N}$$

<u>Or d'après les critères de YULE, comme la médiane, ces paramètres souffrent du fait qu'ils ne se prêtent pas aux calculs algébriques.</u>

Ce qui justifie l'usage courant et fréquent de la variance et l'écart-type.

→ La variance et l'écart type

La variance, notée V(X), d'une série statistique est la moyenne arithmétique des carrés des écarts à la moyenne ou encore la moyenne quadratique des écarts à la moyenne arithmétique.

$$V(X) = \frac{\sum_i n_i (x_i - \bar{x})^2}{N}$$

On peut calculer la variance en utilisant les fréquences fi et non les effectifs ni ; on a alors :

$$V(X) = \sum_i f_i (x_i - \bar{x})^2$$

→L'écart-type est défini comme étant la racine carrée de la variance :

$$\sigma_X = \sqrt{V(X)} = \sqrt{\frac{\sum_i n_i (x_i - \bar{x})^2}{N}}$$

On utilise le célèbre théorème de Kœnig pour les calculs de la variance :

$V(X) = \dfrac{\sum_i n_i x_i^2}{N} - \bar{x}^2$ Ainsi que l'écart type : $\sigma_X = \sqrt{\dfrac{\sum_i n_i x_i^2}{N} - \bar{x}^2}$

.

→ *Le coefficient de variation :*

Il correspond à l'écart type de la distribution exprimé en pourcentage de la moyenne de la distribution :

$CV = \frac{\sigma}{\bar{x}}$

Remarque :

→ L'écart-type satisfait assez bien à l'ensemble des conditions de YULE. Il se prête facilement aux calculs algébriques.

→ C'est l'indicateur utilisé en premier lieu en finance comme mesure de volatilité Ou encore intensité du risque financier, Plus l'écart-type est élevé, plus la dispersion est forte.

→ Cependant l'écart-type est très sensible aux valeurs extrêmes, de forte importance comme la moyenne, donc en particulier si la série est fortement asymétrique.

Pour une distribution dite "normale" loi de Laplace-Gauss parfaitement symétrique, on montre que :

- environ 68% des valeurs sont comprises entre $\bar{x} - \sigma$ *et* $\bar{x} + \sigma$
- environ 95% des valeurs sont comprises entre $\bar{x} - 2\sigma$ *et* $\bar{x} + 2\sigma$
- environ 99% des valeurs sont comprises entre $\bar{x} - 3\sigma$ *et* $\bar{x} + 3\sigma$

→ Comparaison de séries statistiques :

- On est souvent amené ou tenté de comparer des séries statistiques, chose qui s'avère vidée de sens lorsque les distributions ne sont pas mesurées dans la même unité.
- On transforme alors leurs caractéristiques de dispersion en valeur sans dimension. Et ainsi on utilise les nombre ci-dessus :

  - le coefficient de dispersion : $\frac{Q_3 - Q_1}{M_e}$ *ou* $\frac{D_9 - D_1}{M_e}$

  - le rapport interdécile : $\frac{D_9}{D_1}$

  - le rapport interquartile : $\frac{Q_3}{Q_1}$

  - le coefficient de variation : $CV = \frac{\sigma}{\bar{x}}$

→ Caractéristiques de Forme : on se limite aux deux mesures de forme les plus courantes.

→ Mesure de l'asymétrie ou de la dissymétrie

→ Mesure de l’ l'aplatissement

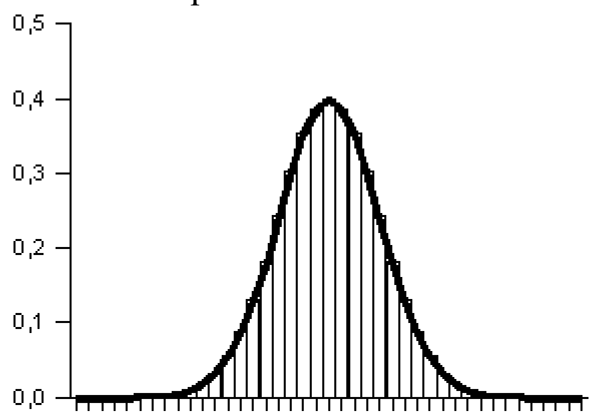
Ces caractéristiques sont basées sur l’étude de l'allure de la courbe de fréquences, ou de l’histogramme. Dans ce cadre on dégage 3 grandes familles d’allures

- les courbes symétriques
- les courbes asymétriques et obliques à gauche
- les courbes symétriques et obliques à droite

→ Les courbes symétriques :

Elles correspondent à des séries statistiques unimodales où : la moyenne arithmétique la médiane et le mode sont approximativement égaux :

$\bar{x} \approx M_e \approx M_0$, c’est la classique courbe en cloche :



Remarque :

- On reprendra cette catégorie de distribution, dite normale, plus amplement en cours de probabilités.
- On peut voir que la moyenne, la médiane et le mode sont identiques
- Le mot : « normale » ne signifie pas que les distributions qui n’ont pas cette allure sont « anormales »

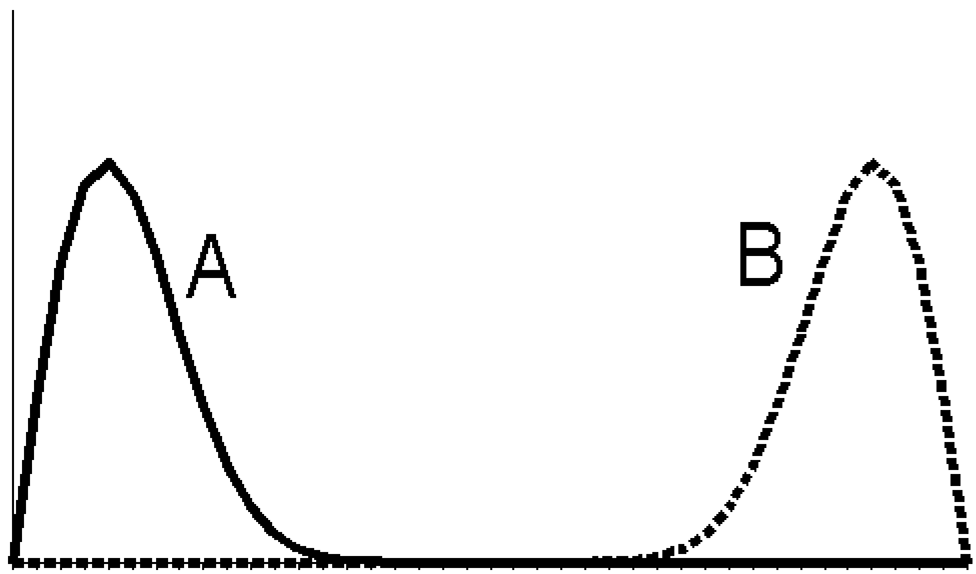
→ Les courbes asymétriques et obliques à gauche :

Lorsque qu'il y a concentration des données du côté gauche de la distribution, et une plus longue queue du côté droit, on dit que la distribution est positivement dissymétrique , c’est le cas pour la distribution A. Ce type d’allure correspond à des séries statistiques unimodales où : la moyenne arithmétique la médiane et le mode sont dans cet ordre

$M_0 \prec M_e \prec \bar{x}$; Cela correspond sur le graphique qui suit à la courbe A

→ Les courbes asymétriques et obliques à droite : Lorsque qu'il y a concentration des données du côté droite de la distribution,
et une plus longue queue du côté gauche, on dit que la distribution est positivement dissymétrique , c'est le cas pour la distribution B. Ce type d'allure correspond à des séries statistiques unimodales où : la moyenne arithmétique la médiane et le mode sont dans cet ordre $\bar{x} \prec M_e \prec M_0$; Cela correspond sur le graphique qui suit à la courbe B

→ Graphiquement : A est oblique à gauche et B oblique à droite



→ Les statisticiens Yule et Pearson ont défini une gamme de paramètres de mesure d'asymétrie tels que :

- Le coefficient de Yule :
  $$s = \frac{(Q_3 - M_é) - (M_é - Q_1)}{(Q_3 - M_é) + (M_é - Q_1)}$$
-

L'interprétation du coefficient de YULE, se fait comme suit :

- La courbe en cloche : la distribution est symétrique correspond au cas : $s = 0$
- La courbe est étalée à droite : la distribution est asymétrique oblique à gauche et correspond au cas $s > 0$, Courbe A
- La courbe est étalée à gauche : la distribution est asymétrique oblique à droite et correspond au cas $s < 0$, Courbe B

- Les coefficients de Pearson :

Nous les reprendrons avec l'étude des moments des variables aléatoires.

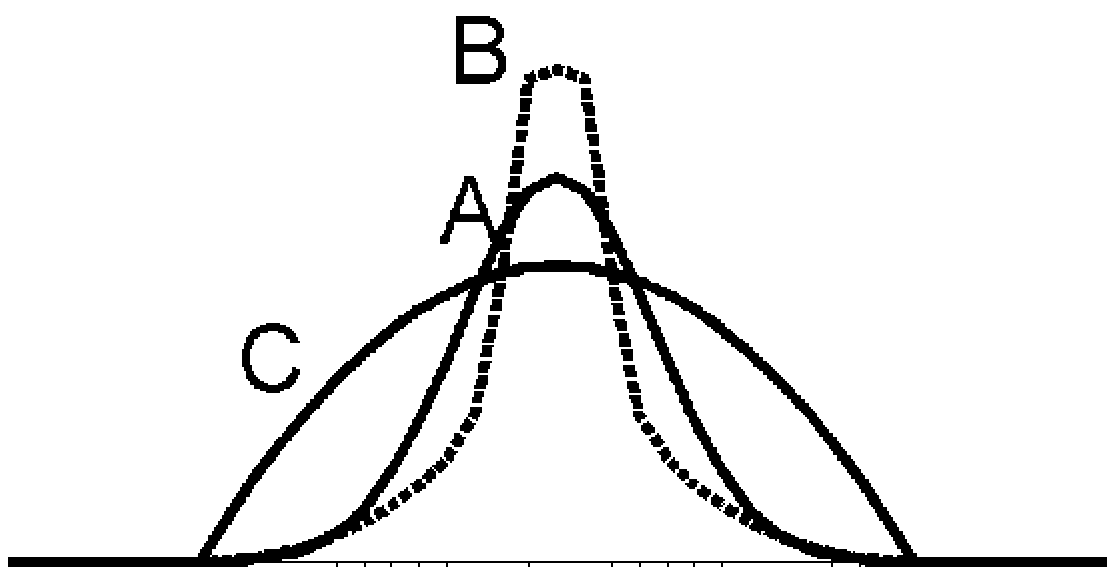
→ Mesure de l' l'aplatissement :

- *On prend comme référence la courbe de la distribution normale*
- *On observe si l'allure est plus ou moins pointue par rapport à la référence.*
- *On évalue la mesure de l'aplatissement en regardant la concentration des données autour du centre de la distribution par rapport à leur dispersion*

→ *A : courbe mésokurtique, elle est de la forme de la courbe normale.*

→ *B : courbe leptokurtique, elle est plus pointue et possède des queues plus longues que la courbe normale.*

→ *C : courbe platykurtique, elle est plus arrondie et possède des queues plus courtes que la courbe normale.*



Exemple de pièges de l'interprétation hâtive des paramètres calculés
→ Variances intra et inter-populations

→ *Dans le cadre des applications on est souvent amené à regarder la dispersion à l'intérieur d'une même population et sa relecture en fonction des diverses sous populations qui la composent. Par exemple : les catégories socioprofessionnelles composent des sous populations de la population française étudiée selon le revenu.*
→ *On rappelle que si une population P d'effectif* N *et de moyenne arithmétique* $\overline{x}$ *est constituée de deux sous population* P_1 *d'effectif* n_1*, de moyenne arithmétique* $\overline{x_1}$ *et* P_2 *d'effectif* n_2 *de moyenne arithmétique* $\overline{x_2}$ *alors :*

$$\overline{x} = \frac{n_1\,\overline{x}_1 + n_2\,\overline{x}_2}{N}.$$

On exprime alors la variance de la population P en fonction des variances des sous populations : mais pas seulement :
On admet le résultat suivant :

$$V(x) = \underbrace{\frac{n_1V(x_1) + n_2V(x_2)}{N}}_{\substack{\text{moyenne des variances} \\ \text{Variance intrapopulation}}} + \underbrace{\frac{n_1(\overline{x}_1 - \overline{x})^2 + n_2(\overline{x}_2 - \overline{x})^2}{N}}_{\substack{\text{variance des moyennes} \\ \text{Variance interpopulation}}}$$

On peut commenter la formule ci-dessus par :
La variance intrapopulations est la variance que l'on obtiendrait si toutes les sous populations avaient la même moyenne $\overline{x}$*, ce qui peut s'interpréter comme la dispersion globale, la variance inter-populations est nulle car :*
$\overline{x} = \overline{x_1}$ *et* $\overline{x} = \overline{x_2}$
Quand à la variance inter-populations peut être vue comme a variance que l'on obtiendrait si toutes les sous populations étaient homogènes, autrement dit chaque variable de chaque sous-populations était égale à sa moyenne, donc aucune dispersion intra populations.