Les caractéristiques de tendance centrale (1)

En s'intéressant aux paramètres de tendance centrale, le statisticien cherche à fournir une valeur qui puisse résumer toute la série statistique considérée. Yule, statisticien anglais du début du 20^{ième} siècle a décrit les propriétés souhaitables pour un indicateur de tendance centrale.

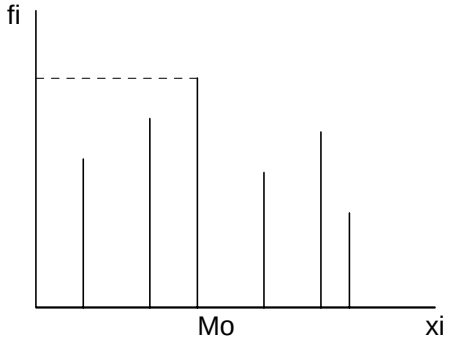
1. Etre défini de manière rigoureuse, i.e. ne pas être laissé à la simple appréciation de l'observateur.
2. Dépendre de toutes les valeurs réalisées.
3. Ne pas présenter un caractère mathématique trop abstrait pour que sa signification soit compréhensible.
4. Etre facile et rapide à calculer.
5. Etre peu sensible aux fluctuations d'échantillonnage.
6. Pouvoir se prêter au traitement algébrique (lorsqu'une variable est la composée de plusieurs autres variables).

1. Le mode

1.1 Définition

1.1.1 Dans une distribution non groupée

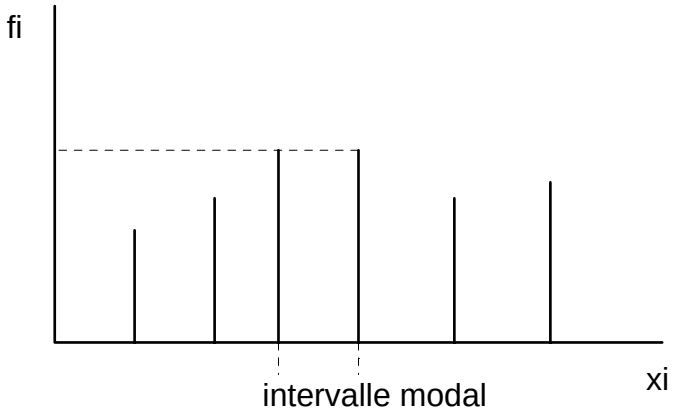
Le(s) ***mode***(s) d'une distribution non groupée est (sont) la (les) valeur(s) observée(s) de fréquence maximum. Le mode est donc la valeur observée qui apparaît le plus fréquemment.



Une distribution de fréquences possédant un seul maximum est qualifiée ***d'unimodale*** (distribution en cloche).

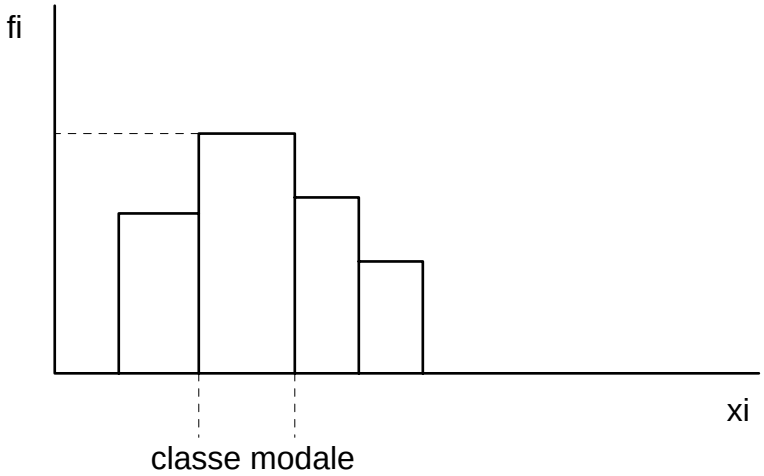
Le mode n'est pas forcément unique. Une distribution de fréquences possédant plusieurs maximums de fréquence est qualifiée de ***plurimodale***.

Si deux fréquences maximales identiques apparaissent successivement on définit un ***intervalle modal***.

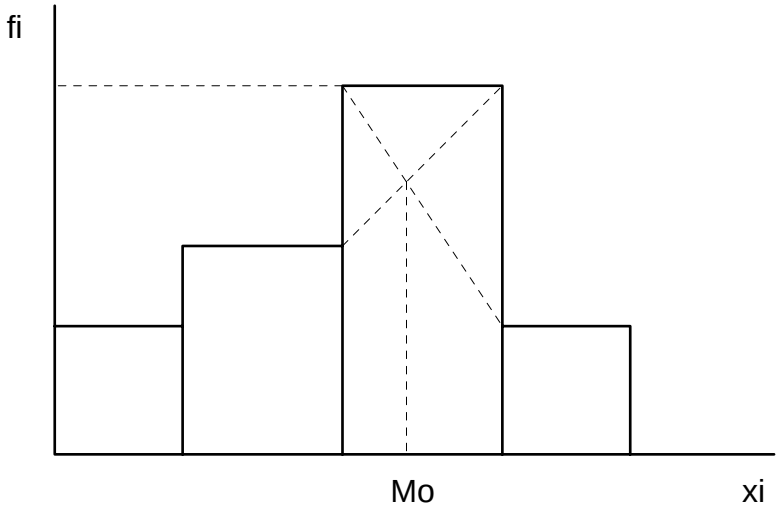


1.1.2 Dans une distribution groupée

La (les) ***classe***(s) ***modale***(s) d'une distribution groupée à intervalle de classe constant est (sont) la (les) classe(s) de fréquence maximum.



Si l'on désire une valeur réelle pour le mode on dispose de deux méthodes: On prend soit le centre de classe, soit on applique la méthode graphique des diagonales.



REMARQUE: Lorsque les amplitudes des classes sont non constantes, l'histogramme doit être construit avec les fréquences unitaires et le mode déduit comme précédemment.

1.2 Propriétés

Dans le cas de distributions unimodales symétriques, le mode se confond avec la médiane et la moyenne.

Comme pour la médiane, sa détermination ne dépend pas des valeurs extrêmes.

1.3 Utilisation

Le mode présente l'avantage d'être défini de façon objective, d'avoir une signification concrète et d'être d'un calcul simple.

Il ne dépend cependant pas de toutes les valeurs de la série statistique.

Il n'est pas forcément unique,

et peut correspondre à une valeur extrême aberrante.

2. La médiane

2.1 Définition

La ***médiane*** M_e est un paramètre de position tel que la moitié des observations lui est inférieure ou égale et la moitié supérieure ou égale. C'est donc la valeur du caractère qui partage la série d'observations, au préalablement rangée par ordre croissant ou décroissant, en deux sous-ensembles égaux.

2.1.1 Distributions non groupées

a. Les données sont individualisées.

On distingue le cas d'un nombre d'observation impair du cas d'un nombre d'observation pair.

Lorsque le nombre d'observation est impair, la médiane est parfaitement déterminée. C'est l'observation de rang $(n+1)/2$:

$$\mathrm{M_e} = \mathrm{x}_{(n+1)/2}$$

EXEMPLE: Si la série est (2, 4, 5, 7, 9), M_e=5

Lorsque le nombre d'observation est pair, la médiane n'est pas déterminée. Ce qui peut l'être n'est qu'un ***intervalle médian***: $\left[x_{\frac{n}{2}} ; x_{\frac{n}{2}+1} \right]$.

EXEMPLE: Si la série statistique est (2, 4, 5, 7, 9, 10), l'intervalle médian est [5, 7].

REMARQUE: Par convention parfois on pose que $\mathrm{M_e} = \frac{\mathrm{x}_{\frac{n}{2}} + \mathrm{x}_{\frac{n}{2}+1}}{2}$.

b. Les données ne sont pas individualisées.

Si les données ne sont pas individualisées, mais se présente comme un tableau de fréquences, la médiane se calcule par les fréquences cumulées.

En effet, puisque la moitié des observations lui est inférieure et l'autre moitié supérieure, la médiane est donc, par définition, telle que $\mathrm{F}(\mathrm{M_e}) = 1/2$. Dès lors, dans le cas

des distributions non groupées, la médiane peut être déterminée graphiquement en recherchant, sur le polygone de fréquences cumulées l'abscisse du point d'ordonnée *n*/2 ou 1/2 (ou 50), selon que l'on considère une distribution d'effectifs ou une distribution de fréquences.

La méthode de détermination est alors très simple. Dans le tableau de fréquence, on repère tout d'abord la valeur 0,5 ou n/2 dans la colonne des fréquences cumulées. Deux cas sont ici possible.

1. La valeur 50 apparaît dans la colonne. Dès lors, l'observation *x* qui correspond à cette valeur est la médiane.
2. La valeur 50 n'apparaît pas dans la colonne des fréquences cumulées directement mais se situe "entre deux lignes" de cette colonne. La médiane est alors la valeur *x* dans la colonne des observations correspondant à la ligne la plus basse de ces deux lignes.

REMARQUE: dans le deuxième cas, le mode présenté pour déterminer la médiane ne permettra pas à celle-ci de partager la série statistique en deux sous-ensembles égaux. Exiger cela de la médiane reviendrait à inventer une valeur (non observée par définition) qui satisferait cette propriété. On ne retrouvera pas cette limite avec les distributions groupées.

2.1.2 Distributions groupées

Dans le cas des distributions groupées la détermination de la médiane se fait en deux étapes. On détermine tout d'abord une ***classe médiane***. Celle-ci contient par définition la médiane. La classe médiane est déterminée à partir de la colonne des fréquences cumulées comme précédemment. Elle correspond à la classe contenant la valeur 50 pour la fréquence cumulée.

Une fois cette classe établie, on trouve la médiane qui est telle que la série est partagée en deux sous-ensembles. La détermination de cette valeur est possible dans la mesure où, lorsque les valeurs sont regroupées en classe, on ne sait pas exactement à la lecture des classes quelles sont les valeurs qui y ont été observées. Une marge de manœuvre nous est ainsi laissée.

Pour déterminer la médiane, et en faisant l'hypothèse d'une équi-répartition[1] des données, on peut utiliser alors la formule[2] suivante :

:

$$M_e = x_G + \alpha \cdot \left[\frac{50 - F(x_G)}{f_G} \right],$$

où x_G est la borne gauche de la classe médiane, α l'amplitude de la classe médiane (différence entre la plus grande et la plus petite des valeurs de la classe médiane), f_G sa fréquence, et $F(x_G)$ la fréquence cumulée correspondant à x_G.

Démonstration

La classe médiane est $[x_G ; x_D[$ et contient la fréquence cumulée égale à 50.

Si nous connaissions la fonction F(.) des fréquences cumulées nous pourrions calculer la médiane en la posant comme l'inconnue de l'équation $F(M_e) = 0{,}5$. Mais nous ne la connaissons pas. Pour la déterminer on va faire une hypothèse qui consiste à dire que *F*(.) est linéaire sur la classe médiane. En connaissant les deux point extrêmes appartenant à cette droite (les bornes de l'intervalle) nous pouvons calculer l'équation de la droite représentative par hypothèse de F(.).Une fois l'équation de *F*(.) trouvée, nous trouvons M_e en posant $F(M) = 0{,}5$.

Les points $(x_G ; F(x_G))$ et $(x_D ; F(x_D))$ appartiennent ainsi à la droite représentative de la fonction F(.) que nous recherchons. Ainsi :

$$\begin{cases} a \cdot x_G + b = F(x_G) \\ a \cdot x_D + b = F(x_D) \end{cases} \Leftrightarrow \begin{cases} b = F(x_G) - a \cdot x_G \\ a \cdot x_D + F(x_G) - a \cdot x_G = F(x_D) \end{cases} \Leftrightarrow \begin{cases} b = F(x_G) - a \cdot x_G \\ a(x_D - x_G) = F(x_D) - F(x_G) \end{cases}$$

[1] L' hypothèse d'équi-répartition des données revient à supposer que toutes les observations à l'intérieur d'une classe sont réparties uniformément

[2] Cette formule permet de calculer la médiane par interpolation linéaire

Dans cette écriture on reconnaît $\alpha = x_D - x_G$ l'amplitude de la classe médiane, ainsi que $f_G = F(x_D) - F(x_G)$ la fréquence associée à la classe médiane. De sorte que l'on peut écrire les deux équation comme :

$$\begin{cases} b = F(x_G) - a \cdot x_G \\ a\alpha = f_G \end{cases} \Leftrightarrow \begin{cases} b = F(x_G) - \dfrac{f_G}{\alpha} \cdot x_G \\ a = \dfrac{f_G}{\alpha} \end{cases}$$

La droite représentative par hypothèse de la courbe de fréquences cumulées sur la classe médiane est ainsi : $F = \frac{f_G}{\alpha} x + F(x_G) - \frac{f_G}{\alpha} \cdot x_G$.

M_e est la valeur x telle que :

$$\frac{f_G}{\alpha} M_e + F(x_G) - \frac{f_G}{\alpha} \cdot x_G = 0{,}5 \Leftrightarrow \frac{f_G}{\alpha} M_e = 0{,}5 - F(x_G) + \frac{f_G}{\alpha} \cdot x_G \Leftrightarrow M_e = \alpha \frac{0{,}5 - F(x_G)}{f_G} + x_G$$

Pour déterminer la médiane on peut ainsi utiliser alors la formule suivante:

$$M_e = x_G + \alpha \cdot \left[\frac{50 - F(x_G)}{f_G} \right]$$

où x_G est la borne gauche de la classe médiane, α l'amplitude de la classe médiane, f_G sa fréquence, et $F(x_G)$ la fréquence cumulée correspondant à x_G.

REMARQUES:

1. La détermination de la médiane n'est pas affectée par des classes d'amplitude inégales.
2. Si 50 apparaissait dans la colonne des fréquences cumulées, x_G serait pris, avec cette formule, pour la médiane.

2.2 Propriétés

Dans le cas de figure de distributions symétriques, la médiane est égale à la moyenne.

Contrairement à la moyenne, le résultat obtenu ne dépend pas des valeurs extrêmes (souvent douteuses).

2.3 Utilisation

La médiane n'a qu'un intérêt limité dans les cas :

1. D'une distribution de valeurs discrètes où l'on n'obtient qu'un intervalle médian. Plus cet intervalle est grand et moins il a de sens.
2. D'une distribution non groupée, lorsque la valeur 50 n'apparaît pas dans la colonne des fréquences cumulées. Dans ce cas, la convention utilisée pour déterminer la médiane ne permet pas d'obtenir une valeur pour celle-ci qui séparerait la distribution statistique en deux sous-ensembles égaux.

La médiane est par contre un paramètre de tendance centrale très adéquat dans le cas de distributions groupées puisqu'elle peut toujours être calculée, moyennant l'hypothèse de linéarité des fréquences cumulées sur la classe médiane.

3 – Les quantiles

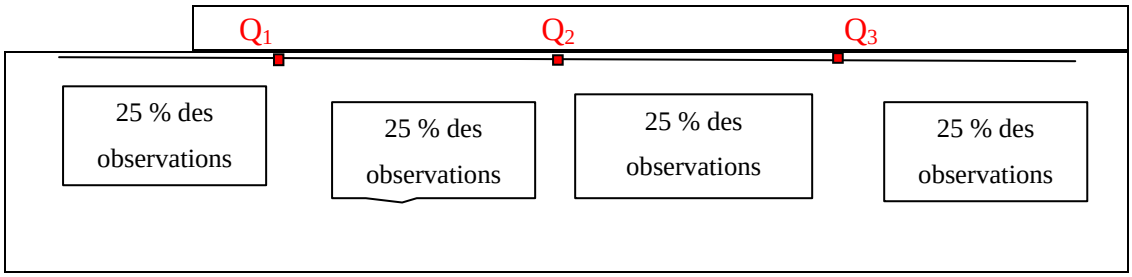
3.1 Définition

Il s'agit d'une généralisation de la médiane. Au lieu de s'intéresser à la valeur de la distribution qui partage la population en 2 parties égales, on s'intéresse aux valeurs qui partagent cette population en 4,10…100 parties égales.

3 découpages sont particulièrement utilisés :
-ceux qui partagent la population en 4 groupes d'effectif égal (quartiles)
-ceux qui partagent la population en 10 groupes d'effectif égal (déciles)
-ceux qui partagent la population en 100 groupes d'effectif égal (centiles ou percentiles)

3.2 Les quartiles

Les 3 quartiles Q_1, Q_2, Q_3 sont les valeurs de la distribution statistiques qui partagent la population en 4 groupes d'effectif égal. La définition d'un quartile est analogue dans son principe à celle de la médiane.
Pour des observations rangées par ordre croissant, les quartiles Q_1, Q_2, et Q_3 sont les valeurs de la distribution telles que :



Ainsi Q_1, Q_2, Q_3 sont les valeurs de la distribution pour lesquelles les fréquences cumulées sont égales à :

$F(Q_1)=0,25$

$F(Q_2)=0,50=M_e$

$F(Q_3)=0,75$.

La différence entre les valeurs du 3ème et du 4ème quartile

$Q_3 - Q_1$ est appelé intervalle interquartlie

L'intervalle interquartile contient donc 50% des observations

et en laisse 25 % à droite et 25% à gauche.

Les quartiles se déterminent de la même manière que la médiane, soit par le calcul (interpolation linéaire), soit graphiquement à partir des fréquences ou des effectifs cumulés.

3.3 Les déciles

Les 9 déciles sont les valeurs de la distribution statistique qui partagent la population en 10 groupes d'effectif égal. Leur définition est analogue dans son principe à celle de la médiane.

Pour des observations rangées par ordre croissant, les déciles sont les valeurs de la distribution telles que :

faire un schéma similaire au précédent avec 9 points D_1……..D_9

1/10 des observations sont inférieures à D_1

1/10 sont comprises entre D_1 et D_2

1/10 sont comprises entre D_2 et D_3

1/10 sont comprises entre D_3 et D_4

1/10 sont comprises entre D_4 et D_5

1/10 sont comprises entre D_5 et D_6

1/10 sont comprises entre D_6 et D_7

1/10 sont comprises entre D_7 et D_8

1/10 sont comprises entre D_8 et D_9

1/10 sont supérieures à D_9

Ainsi D_1, D_2,……D_9 sont les valeurs de la distribution pour lesquelles les fréquences cumulées sont égales à :

$F(D_1)=0,10$

$F(D_2)=0,20$

…….

$F(D_5)=0,50=M_e$

…….

$F(D_9)=0,90$

On peut définir différents intervalles interdéciles.

Par exemple, l'intervalle $D_9 - D_1$ contient 80% des observations et en laisse 10% à droite et 10% à gauche ;

Les déciles se déterminent de la même manière que la médiane, soit par le calcul (interpolation linéaire), soit graphiquement à partir des fréquences ou des effectifs cumulés.

3.4 les centiles

Les 99 centiles, ou percentiles, sont les valeurs de la distribution statistiques qui partagent la population en 100 groupes d'effectif égal. Les centiles sont déterminés lorsque la distribution comporte un grand nombre d'observations. La définition d'un centile est analogue dans son principe à celle de la médiane.

Pour des observations rangées par ordre croissant, les centiles sont les valeurs de la distribution telles que :

1/100 des observations sont inférieures à P_1

1/100 sont comprises entre P_1 et P_2

…….

1/100 sont comprises entre P_{98} et P_{99}

1/100 sont supérieures à P_{99}

Ainsi P_1, P_2,……P_{99} sont les valeurs de la distribution pour lesquelles les fréquences cumulées sont égales à :

$F(P_1)=0,01$

$F(P_2)=0,02$

…….

$F(P_{50})=0,50=M_e$

…….

$F(P_{99})=0,99$

On peut définir différents intervalles inter-centiles. Par exemple l'intervalle $P_{90} - P_{10}$ contient 80% des observations et en laisse 10% à droite et 10% à gauche ;

Les centiles se déterminent de la même manière que la médiane, soit par le calcul (interpolation linéaire), soit graphiquement à partir des fréquences ou des effectifs cumulés.