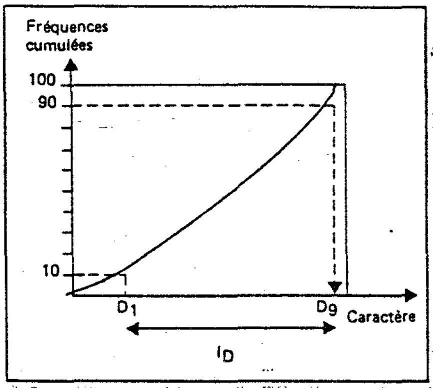
2

Les quantiles

La médiane est la valeur du caractère qui partage l'effectif étudié en deux parties égales. Mais, au lieu de deux, on aurait pu diviser en quatre, dix, cent, n parties égales ; les valeurs correspondantes du caractère sont appelées des quantiles (la médiane est un quantile d'ordre 2).

Les plus utilisés sont :
- les quartiles (quantiles d'ordre 4) qui divisent l'effectif en 4 parties égales. Le quartile Q_1 est tel qu'un quart des effectifs possède un caractère inférieur à Q_1 et trois quarts possèdent un caractère supérieur à Q_1 ;
- les déciles qui partagent l'effectif en 10 groupes égaux ;
- les centiles qui partagent l'effectif en 100 groupes égaux.

Notons que le nombre de quantile d'ordre n est égal à n – 1 : trois quantiles (Q_1, Q_2, Q_3) divisent l'effectif en quatre parties égales.



Les quantiles et plus précisément les intervalles interquantiles servent à mesurer la dispersion d'une distribution. Le plus souvent on utilise l'intervalle interdécile $I_D = D_9 - D_1$. Cet intervalle est tel que 80 % de la population est comprise entre les deux caractères. Cela permet d'avoir une idée de l'étendue d'une distribution en éliminant les valeurs extrêmes du caractère et facilite les comparaisons.

C.F.

La médiane

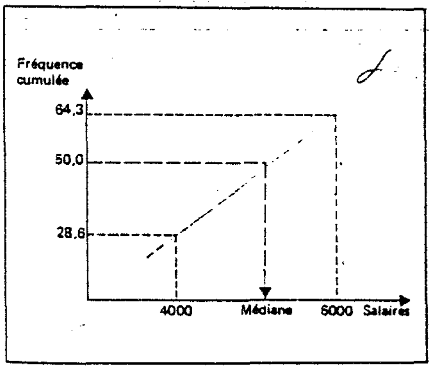
Autre indicateur de position centrale fréquemment utilisé, la médiane est la valeur du caractère qui partage en deux parties égales l'effectif total. Prenons l'exemple d'une répartition par taille d'un groupe : la taille médiane sera telle que 50 % des effectifs seront en dessous et 50 % au-dessus de ce chiffre.

Dans le cas où la variable est continue, c'est-à-dire partagée en classes, la détermination de la médiane se fera en deux temps. Dans une première étape, on cherche dans quelle classe on atteint la moitié des effectifs : ensuite on détermine graphiquement ou par calcul la valeur de la médiane. Soit la distribution de salaires suivante :

	Effectifs	Fréquence	Fréquence cumulée
de 2 000 à 4 000	20	28,6	28,6
de 4 000 à 6 000	25	35,7	64,3
de 6 000 à 8 000	15	21,4	85,7
8 000 et plus	10	14,3	100,0
	70	100,0	

28,6 % des effectifs reçoivent moins de 4 000 F ; 64,3 % moins de 6 000 F. Par conséquent, la 50e observation est atteinte dans la deuxième classe ; la médiane se situe donc entre 4 000 F et 6 000 F.

Pour la déterminer précisément, on suppose qu'à l'intérieur de la classe les salaires se répartissent uniformément selon le schéma suivant :



La médiane est telle que :

$$\frac{Me - 4\,000}{6\,000 - 4\,000} = \frac{50 - 28,6}{64,3 - 28,6}$$

(Théorème de Thalès)

$$Me = 2\,000 \times \frac{21,4}{35,7} + 4\,000$$

$Me = 5\,198,9$

La médiane est donc ici égale à 5 198,9.

C.F.

ou bien : 64,3 – 28,6 = 35,7
35,7 → 6000-4000 soit 2000
50 – 28,6 = 21,4 → $\frac{21,4}{35,7} \times 2000 = 1198$

Variance – Écart-type

Pour caractériser une distribution, on cherche généralement ses valeurs centrales (moyenne, médiane) et des mesures de sa dispersion, c'est-à-dire de la façon dont l'effectif est distribué autour de sa valeur centrale. L'indicateur de dispersion le plus souvent utilisé est l'écart-type, ou son carré, la variance.

Le principe de la variance découle d'une propriété mathématique de la moyenne. L'expression $(x - a)^2$ est minimale lorsque a est égal à la moyenne arithmétique.

La variance est définie comme la moyenne des carrés des écarts à la moyenne (notée $\bar{x}$) :

$$V = \frac{\sum_{i=1}^{p} n_i (x_i - \bar{x})^2}{N}$$

(l'écart-type est égal à la racine carrée de la variance : $\sigma = \sqrt{V}$).

Dans la pratique, on la calcule généralement à partir de sa formule développée :

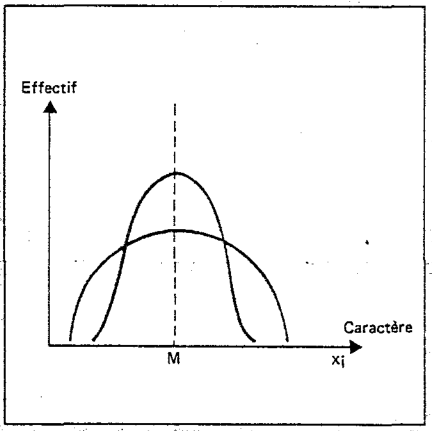
$$V = \frac{\sum n_i x_i^2}{N} - \bar{x}^2$$

Pour limiter les risques d'erreurs de calcul, il est préférable d'utiliser une présentation méthodique qui peut être de ce type :

Caractère x_i	Modalités du caractère	
Effectif n_i		N
Fréquence n_i / N		100 %
x_i^2		
$\frac{n_i}{N} \times x_i^2$		

L'écart-type sert habituellement à comparer deux distributions. Si elles sont de moyennes semblables et si l'écart-type de l'une est supérieur à l'autre, cela signifie que la première est beaucoup plus dispersée.

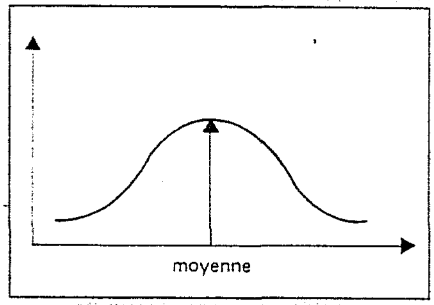
1. Exemple de deux distributions de même moyenne et de dispersion différente



La loi de Laplace-Gauss (ou loi normale) permet d'avoir une compréhension plus concrète de la signification d'un écart-type. Rappelons qu'une distribution statistique suit une loi de Laplace-Gauss lorsque la variable réalisée résulte de l'addition d'un très grand nombre de causes indépendantes de faible intensité. L'exemple classique est la distribution d'une caractéristique (le poids, la tenue en matières grasses, ...) d'une série d'objets fabriqués en série.

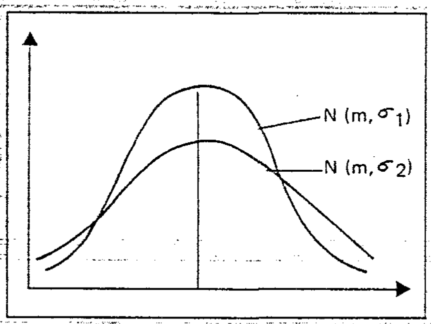
La distribution de Laplace-Gauss a cette allure :

2. Courbe de Laplace-Gauss



Elle est symétrique par rapport à sa moyenne ; sa forme est dépendante de la mesure de son écart-type :

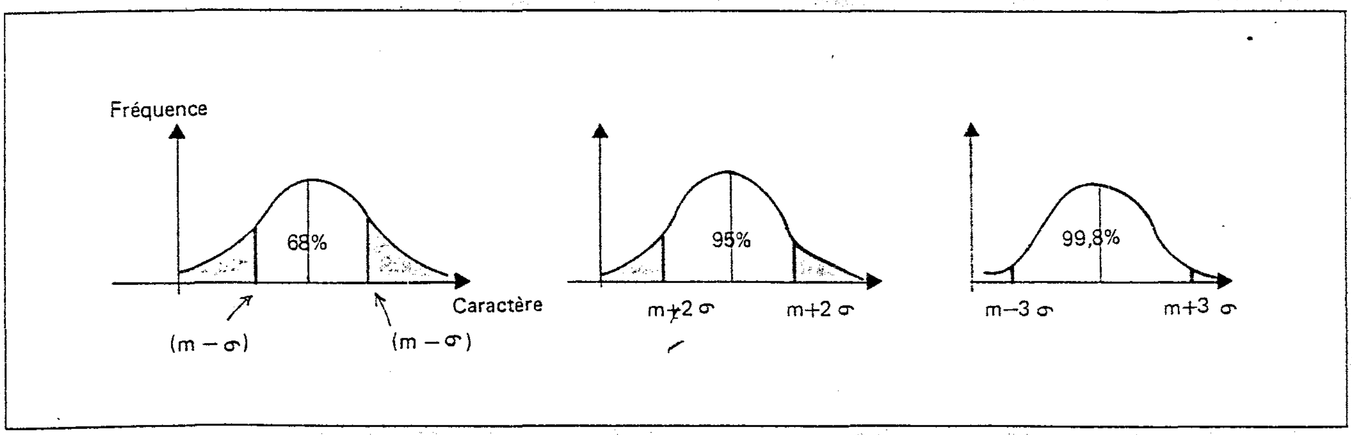
3. Deux courbes de Laplace-Gauss de même moyenne et d'écarts-types différents



On montre que 68 % des observations d'une distribution « normale » sont comprises entre la moyenne plus ou moins un écart-type, 95 % entre la moyenne plus ou moins deux écarts-types et 99,8 % entre la moyenne plus ou moins trois écarts-types (graphique 4).

C.F.

4. Les intervalles de confiance dans la loi de Laplace-Gauss



Le coefficient de concentration de Gini

Ce coefficient de concentration est obtenu à partir de la **courbe de Lorenz.** On mesure la concentration d'une distribution en la comparant à une distribution où la masse totale du caractère serait également répartie entre les individus.

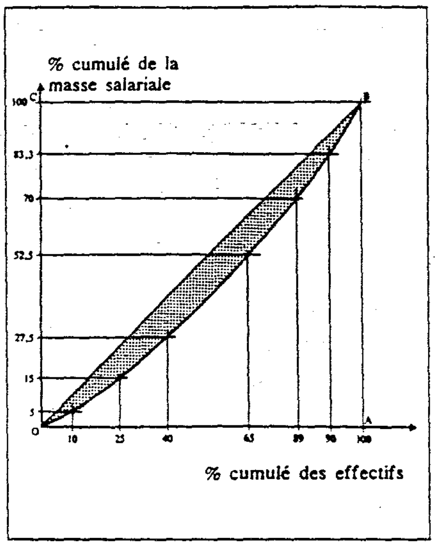
Dans une répartition totalement égalitaire un certain pourcentage des individus est titulaire d'un pourcentage égal de la masse totale du caractère : 10 % des individus sont titulaires de 10 % de la masse totale du caractère, 20 % des individus en détiennent 20 %, ...

Dire que 10 % de la population possède le tiers des revenus disponibles, c'est effectuer une comparaison de ce type.

La courbe de Lorenz (graphique ci-dessous) est **obtenue sur un graphique rectangulaire, en portant en abscisse le pourcentage cumulé des effectifs, et en ordonnée le pourcentage cumulé de la masse globale du caractère.**

Elle passe nécessairement par l'origine : 0 % de l'effectif perçoit 0 % de la sse salariale, et par le point B : 100 % l'effectif perçoit 100 % de la masse salariale.

Une distribution parfaitement égalitaire correspond à la diagonale OB du carré OABC. En tout point de cette diagonale *x* % des salariés perçoivent *x* % de la masse salariale.



La distribution la plus inégalitaire possible serait celle où un seul salarié percevrait la totalité de la masse salariale, les autres ne recevant aucun salaire. Sur le graphique, cette situation correspond à la courbe OAB formée par deux côtés du carré.

Toute distribution réelle doit nécessairement s'insérer entre ces deux situations extrêmes.

Plus la courbe de Lorenz est proche de la diagonale du carré, plus la concentration est faible, plus elle s'en éloigne, plus la concentration est importante.

La mesure de la concentration est obtenue en **rapportant la surface comprise entre la courbe et la diagonale,** appelée surface de concentration, à **la surface du triangle OAB.**

Ce rapport appelé coefficient de Gini est donc :

$$G = \frac{\text{surface de concentration}}{\text{surface du triangle OAB}}$$

Il varie entre 0 et 1 ; il est nul lorsque la courbe est confondue avec la diagonale ; il est égal à 1 lorsque la courbe est confondue avec les côtés OA et AB du triangle OAB.

La mesure pratique de ce coefficient est obtenue en mesurant les surfaces, le graphique ayant été tracé sur un papier millimétrique.

Cette mesure de la concentration est particulièrement adaptée pour mesurer les inégalités relatives, mais masque les inégalités absolues.

Si l'on augmente d'un même pourcentage tous les salaires, la courbe de Lorenz n'est pas modifiée. (*)

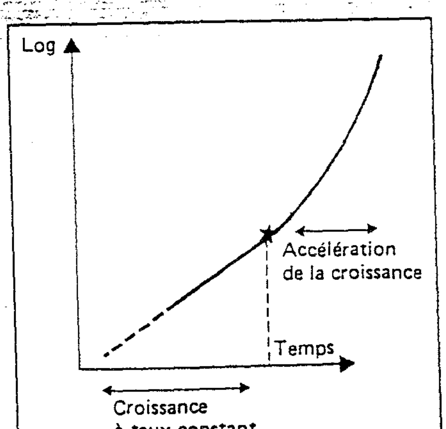
(*) Extrait choisi par la Rédaction des *Cahiers Français* dans Didier Schlacther, *De l'analyse à la prévision*, Ed. Ellipses, 1986.

Les graphiques logarithmiques

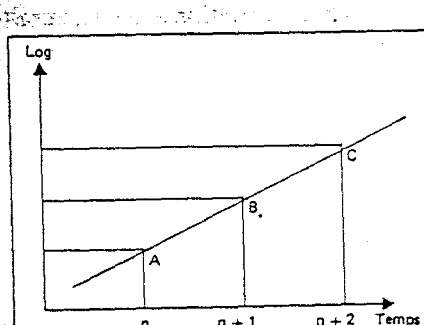
Pour représenter les séries chronologiques, on a très souvent recours à des graphiques logarithmiques : l'abscisse est à échelle arithmétique, l'ordonnée est logarithmique. Si l'on veut représenter une série statistique sur un tel graphique, il faut donc calculer le logarithme des valeurs observées (ou utiliser une feuille à graduation semi-logarithmique : la graduation imprimée correspond aux valeurs arithmétiques mais l'échelle est logarithmique).

Qu'apporte une telle représentation ? Elle est particulièrement appropriée pour l'étude des phénomènes de croissance. En effet, si deux grandeurs ont un même taux de variation relative, sur un graphique semi-logarithmique, elles auront des pentes égales, que l'on passe de 10 à 20 (+ 100 %) ou de 100 à 200 (+ 100 %). Si la série représentée est une droite, cela signifie que son taux de variation (croissance ou décroissance) reste le même de période en période.

Si la croissance s'accélère, la courbe aura l'allure suivante :



Lorsque la croissance est à taux constant, on peut lire sur le graphique le taux de variation associé en procédant comme suit :



sachant que $C = (1 + i)^2 A$

$\text{Log } C = \text{Log } (1 + i)^2 + \text{Log } A$

$2 \text{ Log } (1 + i) = \text{Log } C - \text{Log } A$

$= 2 - 1{,}2$ (par lecture du graphique)

$\text{Log } (1 + i) = 0{,}4$

$1 + i = 1{,}49$ (en utilisant une table des logarithmes)

soit un taux de croissance moyen de 49 %.

C.F.