Paramètres de position

On va parler ici des statistiques quantitatives. On veut les résumer par des nombres. On a deux types de nombres

Les **paramètre de position** : ce sont ceux qui définissent une notion de "valeur centrale" de la série. Ce sont les moyennes, le lode , la médiane.

Les paramètres de dispersion. Ils permettent de mesurer la dispersion de la série autour de la moyenne : une série dispsersée correpondra à une population inhomogène, une population concentrée (peu disersée) correspondra à une population homogène.

Les trois types de statistiques

1. **Les statistiques par individus**. On a la population $1, 2, \cdots, n$, et la statistique $x_1, x_2, \cdots, x_i, \cdots, x_n$.

   Le nombre x_i est la valeur du caractère associée à l'individu i.

   On notera symboliquement une telle statistique par

   $$(x_i)_{i=1\ldots n}$$

   ou simplement par

   $$\boxed{(x_i)}$$

2. **Les statistiques par valeurs**. Soient $y_1, \cdots y_p$ les p valeurs prises par la statistique.

   On dit qu'on a p **classes**
   On note n_i l'effectif associé à la classe y_i, et f_i la fréquence de cette même classe. On note enfin n l'effectif total. On a donc $n_1 + \cdots n_p = n$. Symboliquement, on notera ce type de statistique par

   $$\boxed{(y_i, n_i)}$$

   (présentation par effectifs), ou

   $$(y_i, f_i)$$

   (présentation par fréquences).

3. **Les statistiques par intervalles**. On a p classes représentées par des intervalles

   $$[a_1; a_2[,\ [a_2; a_3[,\ \cdots,\ [a_p; a_{p+1}]$$

On notera qu'on ouvre les intervalles à droite et qu'on les ferme à gauche. Toutefois, pour le dernier, on ferme en général aussi à droite.

Soit n_i l'effectif de la classe $[a_i; a_{i+1}[$. La statistique se note symboliquement

$$\boxed{([a_; a_{i+1}[, n_i)}$$

Moyennes arithmétiques

1. Statistique (x_i). On définit la moyenne par

$$\boxed{\overline{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n}(x_1 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i}$$

   Excel : =moyenne(zone des x_i)

2. Pour une statistique (n_i, y_i), cela devient

$$\boxed{\overline{y} = \frac{n_1y_1 + \cdots + n_py_p}{n} = \frac{1}{n}(n_1y_1 + \cdots + n_py_p) = \frac{1}{n}\sum_{i=1}^{p} n_iy_i}$$

   où $n = n_1 + \cdots + n_p$.

   Excel : =sommeprod(zone des n_i ; zone des y_i)/somme (zone des (n_i)

   Si on a calculé les fréquences f_i, on a les formules plus simples

$$\boxed{\overline{y} = f_1y_1 + \cdots + f_py_p = \sum_{i=1}^{p} f_iy_i}$$

   Excel : =sommeprod(zone des f_i ; zone des y_i)

3. Cas d'une statistique $([a_i, a_{i+1}[, n_i)$. Pour calculer la moyenne, on remplace chque intervalle $[a_i; a_{i+1}[$ par son centre c_i. La moyenne est donc celle de la startistique (c_i, n_i). On a donc

$$\boxed{\overline{y} = \frac{n_1c_1 + \cdots n_pc_p}{n} = \frac{1}{n}(n_1c_1 + \cdots + n_pc_p) = \frac{1}{n}\sum_{i=1}^{p} n_ic_i}$$

   Excel : même procédure que ci-dessus.

Moyennes pondérées. Soit une statistique individuelle (x_i). On peut calculer une moyenne arthmétique de cette série avec des pondération (α_i). Ce qui donne

$$\overline{x} = \frac{\sum_i \alpha_i x_i}{\sum_i \alpha_i}$$

Exemples.

1. Calcul d'une moyenne de fin d'année, les matières ayant chacune un coefficient.
2. Evaluation d'un stock au coût moyen pondéré.

Propriétés

Propriété 1. Soient X et Y deux statistiques sur la même population, qu'on peut additionner. On forme la statistique somme $X + Y$. Alors

$$\overline{x+y} = \overline{x} + \overline{y}$$

Autrement dit : la moyenne d'une somme est la somme des moyennes. Soit

α un nombre. On forme la statistique αX. Alors

$$\overline{\alpha x} = \alpha \overline{x}$$

Moyenne de moyennes. Soient deux populations 1 et 2, d'effectifs respectifs n_1 et n_2, avec $n = n_1 + n_2$. On mesure une même statistique sur les deux populations. Soient $\overline{x}$ la moyenne de la statistique sur l'ensemble de la population, $\overline{x_1}$ la moyenne de la population 1, et $\overline{x_2}$ la moyenne de la population 2. Alors, la moyenne générale est la moyenne des moyennes, pondérée par les effectifs des populations.

$$\overline{x} = \frac{n_1\overline{x_1} + n_2\overline{x_2}}{n_1 + n_2}$$

Ceci se généralise à p populations.

Moyennes géométriques

La moyenne géométrique est à la multiplication ce que la moyenne arithmétique est à l'addition. On suppose que les statistiques sont positives, c'est à dire que les valeurs prises sont toutes positives.

1. Statistique (x_i). On définit la moyenne géométrique par
$$\boxed{G_x = (x_1 \times \cdots \times x_n)^{1/n}}$$
Excel : =moyenne.geometrique(zone des x_i)
2. Statistique $(y_i; n_i)$. On définit la moyenne géométrique par
$$\boxed{G_y = (y_1^{n_1} \times \cdots \times (y_p^{n_p})^{1/n}}$$
où $n = n_1 + \cdots n_p$.
Excel : Il faut se débrouiller "à la main"...

Application : on utilise la moyenne géométrique pour les facteurs d'inflation ou de calitalisation.

Exemple : Un capital de 1000 euros est placé pendant 5 ans, au taux annuels de 3%, 2%, 4%, 5%, 6%. Taux global ? Taux moyen ?

Pour un taux α, on appelle $1 + \alpha$ le **facteur de capitalisation** associé.

Soient α et β les taux cherchés. Le taux global vérifie
$$1 + \alpha = (1 + 3\%)(1 + 2\%)(1 + 4\%)(1 + 5\%)(1 + 6\%)$$
donc $\alpha = 21,6\%$

Le taux moyen vérifie
$$(1 + \beta)^5 = (1 + 3\%)(1 + 2\%)(1 + 4\%)(1 + 5\%)(1 + 6\%)$$
donc le facteur de capitalisation moyen est la moyenne géométrique des facteurs de capitalisation.
$$(1 + \beta) = ((1 + 3\%)(1 + 2\%)(1 + 4\%)(1 + 5\%)(1 + 6\%))^{1/5}$$
donc
$$\beta = ((1 + 3\%)(1 + 2\%)(1 + 4\%)(1 + 5\%)(1 + 6\%))^{1/5} - 1$$

Médiane

Grossièrement, la médiane d'une statistique est la valeur de la statistique qui partage l'effectif en deux.

Exemple : si dans une entreprise le salaire médian est 1200 euros, ce la signifie que

$$\begin{gathered}50\% \text{ de la population a un salaire} < 1200 \text{ euros, et} \\ 50\% \text{ de la population a un salaire} \geq 1200 \text{ euros.}\end{gathered}$$

Calcul

1. Statistique (x_i).
   (a) Calcul à la main : on range les x_i par ordre croissant. S'il y a un nombre paire de données ($2p$), on prend la p-ième donnée, et s'il y a un nombre imair de données ($2p+1$), on prend aussi la p-ième donnée.

   Exemple : médiane de $\{1, 1, 3, 5, , 5, 6, 6\}$. Il y a 8 données, on prend la quatrième : 4.

   Exemple : médiane de $\{1, 1, 3, 4, 5, 5, 6, 8, 9\}$. Il y a 9=$2 \times 4 + 1$ données, on prend la quatrième : 5.

   Excel : =mediane(zone des x_i)

2. Pour une statistique (n_i, y_i), on calcule les fréquence cumulées croissantes, et on prend la première valeur pour laquelle le cumul est $> 50\%$.

   Excel : pas de procédure excel, on y va à la main. On peut faire une macro VBA.
3. Pour une statistique $(n_i, [a_,; a_{i+1})$, on détermine la classe médiane, c'est à dire la première classe au cours de laquelle on dépasse ou on égale 50%. Ensuite, on prend pour médiane le centre de la classe si on est pressé, sinon, on fait une intzerpolation linéaire.

Mode

1. Statistiques in dividuelles ou par valeurs : le mode est la valeur pour laquelle l'effectif est maximal.

   Excel : Excel ne calcule de mode que pour des statistiques individuelles. La syntaxe est

   =MODE (zone)

   Pour une statistique par valeur, on détermine le mode en regardant le tableau d'effectifs où le diagramme en bâtons. Le mode est la valeur correspondant au bâton le plus haut.
2. Statistiques par intervalles. On fait en sorte (voir technique des histogrammes) que les classes soient d'amplitudes égales.

   La classe modale est celle d'effectif maximum. Le mode est le centre de la classe. On peut raffiner un peu cette définition.