MESURES DE POSITION

• LE CONCEPT DE POSITION • DÉFINITIONS ET TYPES DE MOYENNES

• MOYENNE ARITHMÉTIQUE • LA MÉDIANE • LA DOMINANTE

• RÉSUMÉ DES DIFFÉRENTS TYPES DE MOYENNES ET LEURS USAGES

• FORMES DE DISTRIBUTIONS DE POPULATION LES PLUS FRÉQUENTES

À LA FIN DE CE CHAPITRE VOUS DEVEZ ÊTRE CAPABLE DE :

- CALCULER LA MOYENNE ARITHMÉTIQUE
- CALCULER LA MÉDIANE
- CALCULER LA DOMINANTE
- COMPRENDRE À QUEL MOMENT APPLIQUER LES DIFFÉRENTES MESURES DE POSITION

CHAPITRE 5

MESURES DE POSITION

Le concept de position

on veut réduire un peu plus l'information dans une distribution de fréquence

Au chapitre 3 on a vu comment on peut résumer les observations d'une variable en formant une distribution de fréquences. Cette distribution contient beaucoup d'informations sur la variable. Elle montre combien on a de valeurs hautes et basses, et en regardant des représentations graphiques on obtient une impression visuelle de la distribution de cette variable. Dans de nombreux cas c'est suffisant, mais on a souvent besoin de réduire encore plus les informations d'une distribution de fréquences.

comparaisons entre les distributions

Par exemple, si on prend deux distributions avec toutes leurs informations, on a du mal à faire la comparaison. C'est particulièrement vrai pour les personnes qui n'ont pas une bonne compréhension de la statistique. Comme les statisticiens, on devra souvent aider à des personnes qui n'ont pas l'habitude des distributions de fréquences, comme les administrateurs et les décideurs. Dans ces cas-là, on devra penser à d'autres mesures qui seront plus faciles à comprendre. Dans ce chapitre on verra comment calculer des valeurs qui peuvent représenter des caractéristiques ou des propriétés de la distribution d'une population étudiée. On peut alors utiliser ces valeurs pour faire des comparaisons et former la base de décisions plus complexes.

exemple

On va prendre un exemple très simple. On suppose qu'un ami veut connaître nos performances au collège. On peut choisir de rassembler toutes les notes et de dresser une distribution de fréquences simple comme celle-là :

Note	f (fréquence)
A	4
B	9
C	6
D	1
E	0

un rapport efficace et rentable

Ce tableau indiquer qu'on a eu quatre A, neuf B, six C et un D. Malheureusement, tous ces détails ne sont sûrement pas utiles pour répondre à la question de l'ami, et ils sont aussi fastidieux. De plus, la présentation des données sous cette forme rend la comparaison avec les autres étudiants difficile. Le mieux serait de prendre une ou deux mesures résumées des notes pour qu'elles puissent être interprétées vite et bien.

position ou rang comparatif

On voudra sûrement donner la position générale de la distribution des notes. On peut simplement dire verbalement que le travail au collège était légèrement en-dessous du niveau B. Si on veut être plus précis on peut même donner la note numérique (moyenne des points), un chiffre simple qui donne la position générale de ce jeu de notes. Dans les deux cas on résume les performances en se référant à un point central de la distribution qui sera représentatif des notes. Il serait clairement trompeur de décrire la performance générale comme étant de niveau A ou D, même si on a réellement reçu ces notes. C'est une des nombreuses situations qui requièrent l'utilisation d'une **mesure de position,** ou **tendance centrale**. C'est-à-dire une mesure simple qui essaye de décrire la position d'un jeu de données.

exemple

Prenons le cas auquel nous, individus impliqués dans un travail statistique, sommes souvent confrontés. Regardons le tableau suivant qui montre deux distributions de fréquences des revenus annuels des ménages dans différentes régions d'un pays :

Tableau 5.1 Comparaison entre deux distributions de fréquences

Revenus annuels des ménages			
Région A		Région B	
Revenus ($)	Fréquence (Nombre de ménages)	Revenus ($)	Fréquence (Nombre de ménages)
Moins de 500	137	Moins de 1.000	86
500–999	278	1.000–1.999	137
1.000–1.499	406	2.000–2.999	64
1.500–1.999	331	3.000–3.999	47
2.000–4.999	188	4.000–6.999	130
5.000–9.999	259	7.000–9.999	62
10.000–19.999	138	10.000 et plus	88
20.000 et plus	14		
Total	1.751		614

Source : fictive

moyenne et variabilité

Supposez que le gouvernement veuille comparer les revenus des ménages de la région A à ceux de la région B. Cette sorte d'analyse sera très importante lors des prises de décisions pour chaque région. Néanmoins, si on présente les données comme au tableau ci-dessus il sera difficile d'établir des comparaisons. On a les mêmes variables dans chaque cas, mais un nombre différent d'observations et différentes classes de revenus. Ce qu'on doit faire c'est examiner les distributions pour les deux régions et trouver un moyen de décrire certaines caractéristiques de chacune des régions, qui peuvent ensuite être comparées aisément. On peut choisir plusieurs caractéristiques différentes, mais dans la pratique on a tendance à se concentrer sur seulement deux : une valeur **moyenne** de la variable, et la **variabilité** (ou la progression) de la distribution. On choisit celles-là parce qu'elles ont une signification évidente et qu'elles ont tout ce qu'on peut vouloir décrire de la distribution. Ces deux mesures sont la base de presque toute analyse statistique, et on verra les moyennes (ou mesures de position) dans ce chapitre.

ASTUCE

MOYENNE ET **VARIABILITÉ** SONT LA BASE DE PRESQUE TOUTES LES ANALYSES STATISTIQUES

Définitions et types de moyennes

termes statistiques

Dans ce chapitre on se concentrera sur les observations de variables et on ne s'attachera qu'à une seule variable à la fois. Les observations seront dans leur état initial (pour le cas où les observations seraient groupées en distributions de fréquences, voir le chapitre 3 "Plus d'infos sur les distributions de fréquences". Si on veut être capable d´établir des formulations qui seront vraies pour tout jeu de données, on aura besoin de notations statistiques spéciales (ou de symboles) et de définitions. On peut utiliser certaines lettres et symboles pour remplacer certains composants. Cependant, on doit mentionner une ou deux choses avant de continuer à étudier les moyennes en détail.

x_n

Si on a un échantillon d'une population on utilise toujours la lettre *n* pour indiquer le nombre d'observations dans l'échantillon. Les valeurs des variables particulières observées sont indiquées par $x_1, x_2, \ldots, x_n$ (le symbole "…." veut dire "et ainsi de suite"). Ainsi, x_3 veut dire la troisième variable x dans l'échantillon. Par exemple, supposez que vous alliez à la pêche tous les jours pendant une semaine, et voici les chiffres des poissons que vous attrapez :

Tableau 5.2 Nombre de poissons capturés par jour

Observation	Jour	Nombre de poissons
x_1	1	11
x_2	2	5
x_3	3	3
x_4	4	17
x_5	5	12
x_6	6	9
x_7	7	6
Total		63

Source : données fictives

notation

Ainsi, x_3 est la troisième valeur de la variable "nombre de poissons capturés", soit $x_3 = 3$

deux types de population

La population d'où on sélectionne l'échantillon peut être de deux types. Premièrement elle peut être de taille fixe, c'est-à-dire que l'on peut compter tous les individus dans la population. Dans ce cas la population est appelée ""finie", et la taille de la population est indiquée par N. Les personnes vivant dans un pays, les entreprises opérant dans une île, et toutes les fermes dans une zone définie sont des populations finies. Le deuxième type de population n'a pas de limite de taille et on ne peut pas compter le nombre d'individus; une telle population est appelée "infinie". Tous les plants de taros qui pourraient être cultivés sur une île, tous les poissons qui pourraient être capturés dans une certaine zone de la mer, et tous les porcs qui pourraient jamais être élevés dans un pays sont des populations infinies.

différences entre des populations et des échantillons

On voudra faire la distinction entre des populations et des échantillons, car il y a des différences importantes. Quand on traite des données qui proviennent d'un échantillon de la population, la notation sera différente de celle de données provenant d'une population entière. Quand on traite de la population entière, on utilise des lettres de l'alphabet grec, comme μ (mu) et σ (sigma). Les valeurs calculées sont appelées les paramètres. Par contre, pour un échantillon, on utilise des lettres anglaises ordinaires pour représenter les valeurs calculées, et on appelle ces valeurs des estimations.

estimer les paramètres de population

Très souvent on n'a pas d'informations sur une population, on a plutôt une série d'observations qui proviennent d'un échantillon. Ce qu'on fait, c'est estimer les paramètres de population en calculant des estimations d'échantillons. On reviendra sur ce point quand on parlera des différents types de moyennes.

que veut dire "moyenne" ?

Le mot "moyenne" est très utilisé dans le language courant. Par exemple, les gens parlent souvent d'"homme moyen", une performance au-dessus de la moyenne, et une température au-dessous de la moyenne. Le mot "moyenne" est utilisé dans le sens "typique", "usuel" ou "normal". On utilise aussi beaucoup le mot moyenne en statistique, bien que le sens ne soit pas tout à fait le même.

exemple

La plupart des gens pensent que la moyenne d'un groupe de chiffres est le résultat de leur addition et ensuite de la division du résultat par le nombre de chiffres additionnés. Par exemple, la moyenne de 6,7; 9,6; 12,8; 13,0 et 15,9 serait :

$$\frac{6{,}7+9{,}6+12{,}8+13{,}0+15{,}9}{5} = \frac{58{,}0}{5} = 11{,}6$$

moyenne arithmétique

En fait on utilise différents types de moyenne en statistique, celui qui est décrit ci-dessus est connu plus exactement sous le terme **moyenne arithmétique**.

comparaisons entre différentes populations

Très souvent en statistique on veut établir des comparaisons entre différentes populations, en fait une grande partie de la théorie statistique est concernée par ce problème. Par exemple, on peut vouloir comparer les revenus de ménages de secteurs différents, l'incidence des caries entre différents groupes d'enfants scolarisés, ou les poids des poissons capturés à différentes époques de l'année.

ASTUCE

quand ON COMPARE DES POPULATIONS, il est beaucoup plus facile DE DÉTERMINER les DIFFÉRENCES avec LES **MOYENNES ARITHMÉTIQUES** QU'avec D'AUTRES TYPES DE MOYENNES.

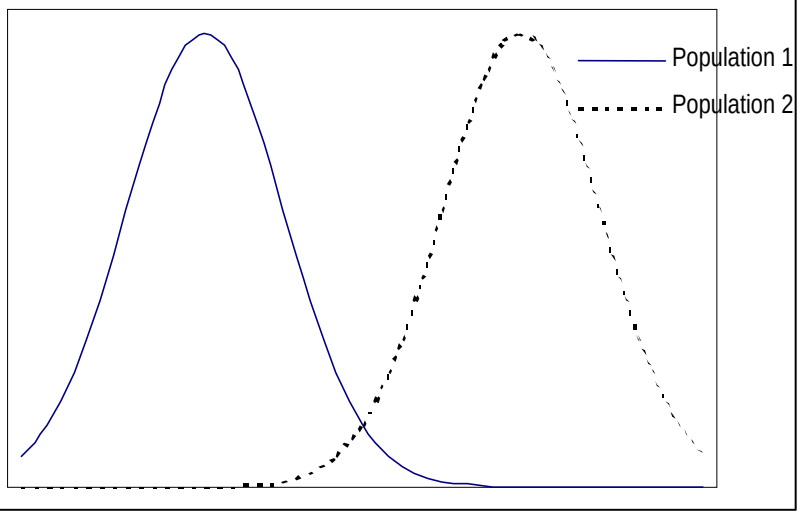
assurez-vous que les distributions sont similaires

Si deux populations ont des types de distribution très différents, la comparaison risque de s'avérer très difficile. Néanmoins, dans la plupart des cas on réalise que les deux distributions sont tout à fait similaires, et que les comparaisons sont faciles à faire. On veut alors trouver une ou deux façons de décrire la population en résumant la distribution par certaines caractéristiques.

une valeur résumée est une mesure de position

Par “mesure de position” d’une distribution on veut dire trouver une valeur qui résume d’une certaine manière la taille de toutes les valeurs différentes d’une distribution. D’autres termes sont utilisés pour décrire la même chose, et certains parlent de moyennes comme mesures de la tendance centrale, ou de mesures de position centrale; ce sont seulement deux manières différentes de parler de la même chose. On utilisera le terme de **mesure de position**. On peut voir au diagramme ci-dessous qu’en général les valeurs de la population 2 sont plus grandes que les valeurs de la population 1.

Figure 5.1 : Exemple de positions de deux populations



types de moyennes

Dans ce chapitre on verra trois types de mesure de position :

a. la moyenne arithmétique

b. la médiane

c. la dominante

données groupées

Pour illustrer ces moyennes on prendra des observations d’origine. Pour illustrer des moyennes groupées en distribution de fréquence, se référer à la fin de ce chapitre "Plus d’infos sur les mesures de position”. On parlera ici de deux autres mesures de position importantes, les quartiles et la moyenne géométrique.

Moyenne arithmétique

données quantitatives seulement

La moyenne arithmétique implique de faire des calculs sur les valeurs observées de la variable. Alors, la moyenne arithmétique s’applique seulement aux variables quantitatives (c’est-à-dire les variables qui ont des valeurs numériques). Si on observe un échantillon de n valeurs d’une variable particulière, on peut en dresser la liste comme suit : $x_1, x_2, \ldots, x_n$. Alors, la moyenne arithmétique de cet échantillon s’écrit sous la forme $\bar{x}$ (qui se prononce x barre) et se définit :

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Σ

La lettre grecque Σ (sigma majuscule) représente "la somme de". " *i=1* " et " *n* " au-dessus et en-dessous du signe Σ nous dit que la somme est de x_1 à x_n. La formule de $\bar{x}$ est simplement une manière abrégée d'écrire :

> *" La moyenne d'un nombre n de chiffres est la somme de tous les chiffres, divisée par n".*

exemple

Pour l'exemple des poissons du tableau 5.2, le calcul sera :

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{63}{7} = 9$$

Donc, la moyenne arithmétique du nombre de poissons capturés par jour est 9.

appliquer ensuite la moyenne aux données

À l'exemple ci-dessus la réponse à la question "Quelle est la moyenne arithmétique du nombre de poissons capturés par jour ?" était un chiffre qui pouvait réellement apparaître en réalité. On voit qu'au sixième jour le nombre de captures était égal à cette valeur "moyenne".

variables discontinues

Pourtant, pour les variables discontinues comme le "nombre de poissons capturés", quand on calcule la moyenne arithmétique et qu'on utilise le résultat en tant que moyenne, on n'a aucune garantie d'avoir un chiffre qui apparaîtrait naturellement. Prenons l'exemple suivant :

Tableau 5.3 Nombre de personnes dans chaque ménage dans un petit village

Nombre de ménages	Nombre de personnes
1	5
2	9
3	1
4	3
5	4
6	5
7	2
8	7
Total	36

Source : fictive

Dans ce cas, la moyenne arithmétique des valeurs observées est $^{36}/_{8} = 4{,}5$ personnes par ménage.

la moyenne peut prêter à confusion avec les données discontinues

Cette sorte de résultat est souvent déroutante. En réalité on ne peut pas avoir une demie-personne, alors comment peut-on dire que le nombre "moyen" de personnes est 4,5 ? Il faut garder en mémoire que la moyenne arithmétique est un concept artificiel. On l'utilise parce qu'elle est facile à calculer et à comprendre, et elle est mathématiquement pratique si on veut faire des calculs plus poussés.

Ce problème arrive seulement avec des données continues, car par définition avec les données discontinues toute valeur comprise dans l'étendue possible des valeurs peut apparaître en réalité. Il n'est pas difficile de comprendre ce qu'on veut dire par "la moyenne arithmétique de la taille d'un groupe d'hommes est 1,75 m".

On doit donc faire attention quand on parle de moyenne arithmétique d'un jeu de données discontinues. On doit réaliser que c'est un type de moyenne, une mesure de position de la population. Ce qui ne veut pas dire que cette valeur va certainement apparaître en réalité, elle peut ne pas apparaître du tout.

La médiane

la médiane partage un jeu de données en deux parts égales

La médiane est un autre type de moyenne, ou la mesure de position d'un jeu de chiffres, et en fait c'est un concept très simple. La **médiane** est la valeur qui partage un jeu de données en deux parts égales. C'est le chiffre du milieu du jeu de données lorsque celui-ci est mis en ordre.

Par exemple, on suppose qu'on a le jeu de données suivant :

14	9	16	3	1	7	5

mettre les observations en ordre

Pour trouver la médiane on doit mettre les observations par ordre de grandeur comme ceci :

1	3	5	7	9	14	16

la valeur du milieu

La médiane est la valeur du milieu, dans ce cas le 7. Comme on peut le voir, il y a autant de chiffres inférieurs à 7 (1,3 et 5) que de chiffres supérieurs à 7 (9, 14 et 16).

On voit aussi qu'avant que le jeu d'observations soit ré-organisé par ordre de grandeur, la valeur du milieu etait 3. Ceci **n'est pas** la médiane. Pour trouver la médiane, ou la valeur centrale, on doit d'abord classer les valeurs du jeu de données original par ordre de grandeur. Alors, si on a n observations, la médiane sera la valeur de l'observation $\frac{n+1}{2}$ de la liste classée.

nombre pair d'observations

On a un prolème si on veut déterminer la médiane d'un jeu d'observations qui n'a pas de valeur centrale, c'est-à-dire quand le nombre d'oservations dans le jeu de données est pair. On adopte une convention selon laquelle la médiane de la série :

1	3	5	7	9	14	16	21

est définie comme la moyenne arithmétique des deux valeurs du milieu, qui dans ce cas sont 7 et 9. La médiane est donc (7 + 9) / 2 = 8.

les valeurs aberrantes ne la modifient pas

Il est clair que si la médiane dépent de la valeur de l'observation du milieu dans un jeu de données, les valeurs extrèmement hautes, ou basses (**valeurs aberrantes**) n'ont aucun effet sur elle. Par exemple, prenez les informations suivantes sur la taille d'une plantation de cocotiers sur une île donnée, avec les valeurs en hectares :

1.3	1.3	1.5	1.7	2.0	2.1	2.3	2.7	2.8	3.0	3.7	5.0	5.5	7.0	120.1

la valeur la plus grande n'a aucun effet …

La dernière valeur représente la surface d'une plantation commerciale, tandis que les autres données sont celles de petites propriétés. La moyenne arithmétique des plantations est 10,8 ha, la médiane est 2,7 ha. Si la plantation commerciale est remplacée par une plantation d'une surface de 3,1 ha, la moyenne arithmétique devient alors 3,0 ha, alors que la médiane reste inchangée.

… pas le cas pour la moyenne arithmétique

La médiane n'est alors pas modifiée par les valeurs des observations très hautes ou très basses, alors que la moyenne arithmétique l'est. S'il y a un doute sur l'exactitude des observations à l'une ou l'autre des extrèmes sur l'échelle des mesures, la médiane est une meilleure "moyenne" que la moyenne arithmétique.

notation de l'échantillon

Pour un échantillon d'observations on désigne généralement la médiane par $\tilde{x}$ (prononcé x tilde). Il n'y a pas de symbole consacré pour désigner la médiane d'une population.

La dominante

un pic dans la fréquence

Si une série d'observations a un pic à un certain point de la fréquence, on dit qu'il y a une dominante à ce point. Comme vous le verrez aux exemples donnés à la fin de ce chapitre, la distribution d'une population peut être **unimodale** (avoir une dominante seulement), **bimodale** (avoir deux dominantes) ou **plurimodale** (avoir plusieurs dominantes); ou bien elle peut n'avoir pas de dominante du tout.

la dominante est un type de moyenne

Comme la moyenne arithmétique et la médiane, la dominante est un type de moyenne. C'est une mesure de position de la distribution.

données discontinues

Quand on a affaire à des observations d'échantillons, le concept de dominante associé aux fréquences de distribution est très utile. Pour une distribution discontinue, la dominante est la valeur qui survient le plus souvent. Par exemple, le nombre d'occupants (personnes) par ménage (tableau 5.3), la valeur dominante est 5 personnes par ménage. Cette valeur est celle qui apparaît le plus souvent parmi les autres valeurs.

données continues

Pour une distribution continue, trouver la dominante est plutôt compliqué. On verra donc ici seulement la **classe modale**. C'est la classe qui a la fréquence la plus haute. Aussi, au tableau 5.4 par exemple, 3,0 à 3,9 kg est la classe modale pour les données de poisson du tableau 3.2.

Tableau 5.4 Distribution de poids de poissons

Classe (kg)	Fréquence	Fréquence cumulée
2,0–2,9	7	7
3,0–3,9	19	26
4,0–4,9	16	42
5,0–5,9	12	54
6,0–6,9	6	60
7,0–7,9	3	63
Total	63	

Source : Tableau 3.2

exemple

Pour produire des résultats plus parlants, il vaut parfois mieux déterminer la dominante ou la classe modale d'une distribution de fréquences plutôt que calculer la moyenne arithmétique. Prenons par exemple les données suivantes. Dans une enquête sur les enfants scolarisés, on a compté le nombre de dents cariées de chaque enfant :

Tableau 5.5 Nombre de dents cariées

Nombre de dents cariées	Nombre d'enfants
0	113
1	156
2	37
3	21
4	11
5	7
6	3
7 ou plus	2
Total	350

Source : données fictives

moyenne contre dominante

La moyenne arithmétique des dents cariées par enfant est 1,16. Pour une personne qui ne fait pas de statistique c'est un chiffre qui n'a aucun sens. Il est probablement plus utile de dire que la dominante est une carie par enfant.

les données continues peuvent poser problème

On a souvent des difficultés à déterminer les classes modales pour les distributions de fréquences continues. La classe modale dépendra de la manière dont les classes sont définies, et pour les données qui ont une distribution relativement régulière entre les classes, un changement dans cette définition peut modifier la classe modale. Pour cette raison la dominante est d'un usage limité, on doit l'utiliser avec précaution.

Résumé des différents types de moyennes et leurs usages

moyennes différentes

Pour terminer ce chapitre, on va examiner les différents types de moyennes étudiés, résumer les avantages et inconvénients, et dans quelles circonstances ils doivent être utilisés.

Moyenne arithmétique

la moyenne

On en parle souvent sous le simple terme de moyenne, et c'est ce que la plupart des gens sous-entendent lorsqu'ils parlent de moyennes.

AVANTAGES	INCONVÉNIENTS
☺ elle est comprise par presque tout le monde;	☹ elle est modifiée par les valeurs extrèmes;
☺ elle est facile à calculer, et peut être déterminée dans tous les cas pour les données quantitatives;	☹ son chiffre peut ne pas exister en réalité;
☺ elle tient compte de toutes les valeurs dans un échantillon;	☹ elle peut ne pas réellement refléter la nature d'une distribution;
☺ elle est importante en mathématique, elle est facile à manipuler en algèbre;	☹ elle ne peut être utilisée qu'avec les données quantitatives.
☺ elle est particulièrement utile quand on compare des populations, ou qu'on estime des paramètres de populations à partir d'estimations d'échantillons.	

MÉDIANE

AVANTAGES	INCONVÉNIENTS
☺ elle est facile à calculer;	☹ elle n'utilise pas toutes les valeurs de l'échantillon;
☺ elle n'est pas modifiée par les valeurs extrèmes;	☹ elle peut produire un chiffre qui n'existe pas en réalité;
☺ elle peut être déterminée dans toutes les situations, excepté avec les données nominales.	☹ elle est difficile à traiter en algèbre.

distributions asymétriques

La médiane peut aussi être utilisée dans bien des situations. Néanmoins, sa meilleure utilisation est quand on traite d'observations qui ont des distributions asymétriques (voir le chapitre 6 pour les distributions asymétriques), car dans ces cas-là elle donne probablement plus d'informations que la moyenne.

DOMINANTE

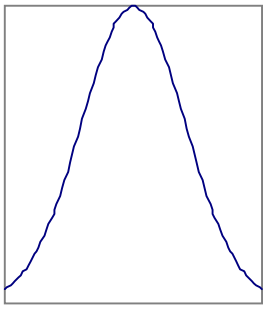
AVANTAGES	INCONVÉNIENTS
☺ elle est facile à calculer quand on a un nombre de données peu important;	☹ elle est difficile à calculer pour les distributions de fréquences continues;
☺ elle a un sens évident dans beaucoup de cas;	☹ quelquefois il peut arriver qu'elle n'existe pas;
☺ elle n'est pas modifiée par les valeurs extrèmes;	☹ pour une distribution de fréquences, elle dépend de l'intervalle de classe choisi;
☺ elle peut être utilisée avec les données nominales.	☹ elle n'est pas bien adaptée au traitement mathématique;
	☹ elle n'utilise pas toutes les données d'un échantillon.

informations sur la plus 'populaire'

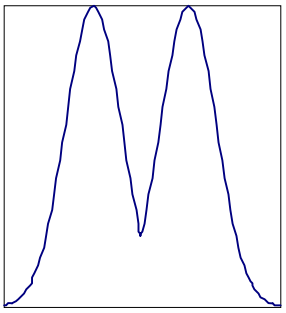
La dominante est utile dans le cas de distributions de fréquences, quand on a besoin d'informations sur la classe la plus courante ou la plus utilisée.

distributions de population les plus fréquentes

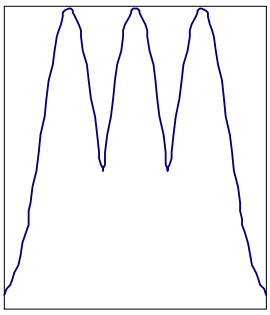
Distribution unimodale



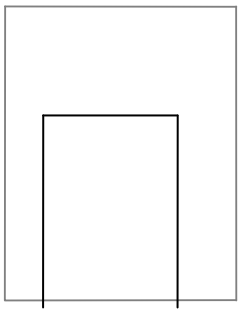
Distribution bimodale



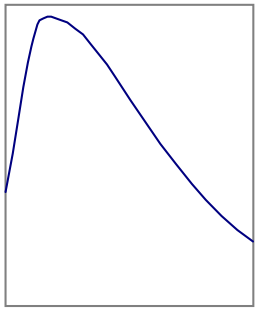
Distribution multimodale



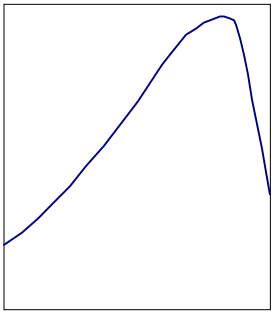
Distribution sans mode



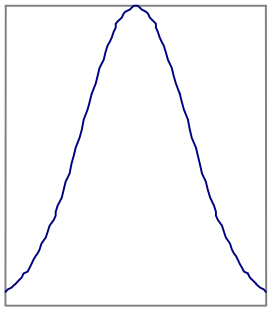
Distribution asymétrique à droite



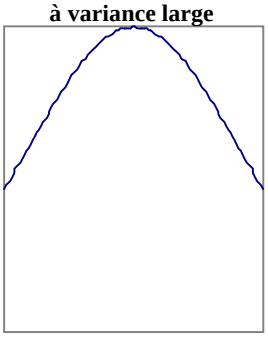
Distribution asymétrique à gauche



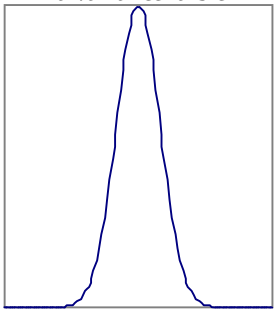
Distribution normale



Distribution à variance large



Distribution à variance faible



PLUS D'INFOS SUR LES MESURES DE POSITION

INTRODUCTION

données groupées

Au chapitre 5, on a vu la moyenne arithmétique et la médiane avec des observations d'origine (c'est-à-dire non groupées). Dans cette section on va traiter le cas où les observations sont groupées en distributions de fréquences. On verra aussi deux autres mesures de position utiles, les quartiles et la moyenne géométrique.

CALCULER LA MOYENNE ARITHMÉTIQUE D'UNE DISTRIBUTION DE FRÉQUENCES

données discontinues

prenons d'abord le cas d'une **distribution de fréquences discrète**. Les données du tableau 5.6 donnent le nombre de destinations de vacances aux Îles Cook en 1991.

Tableau 5.6 Nombre de destinations de vacances par nombre de personnes

Nombre de destinations	Personnes
1	1.811
2	683
3	342
4	273
5	137
6 ou plus	171
Total	3.417

Source : Enquête sur les visiteurs des Îles Cook, 1991,
Rapport d'enquête No. 13; COT, Tableau 23, p. 31.

calculer la moyenne

Si on veut calculer le nombre moyen de chambres par nombre de destinations on ne doit pas seulement aditionner le nombre de destinations et le diviser par le nombre de groupes. On a beaucoup plus de voyages à une destination que de voyages à 5 destinations, et on doit tenir compte de cette information. On doit également tenir compte du dernier groupe "6 ou plus". Alors que la fréquence de ce groupe est faible, on introduira une légère erreur si on suppose une moyenne de 7 destinations par voyage pour toutes les unités de cette classe. On calcule alors comme suit :

$$\text{Moyenne arithmétique } \bar{x} = \frac{(1\times 1.811)+(2\times 683)+(3\times 342)+(4\times 273)+(5\times 137)+(7\times 171)}{3.417} = \frac{7.177}{3.417}$$

$$= 2{,}1 \quad \text{destinations par voyage}$$

formule pour les distributions de fréquences

Tout comme on a une formule arithmétique pour un jeu de chiffres non groupés, on utilise certaines formules avec les distributions de fréquences. Dans ce cas on appelle le nombre de classes *K,* la valeur de chaque classe sera appelée *x* (c'est-à-dire que la valeur de la classe *I* sera appelée x_I) et la fréquence de chaque classe *f* (c'est-à-dire que la fréquence de la classe *I* sera appelée f_i).

La formule de la moyenne $\bar{x}$ est alors :

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$$

méthode

En d'autres termes, on doit effectuer les étapes suivantes quand calcule la moyenne d'une distribution :

(a) pour chaque classe, on calcule le produit de la fréquence, et la valeur de la classe (f_i x_I);

(b) on additionne les produits calculés en (a) sur toutes les classes k (Σf_i x_I);

(c) on calcule la fréquence totale en additionant les fréquences de chaque classe sur toutes les classes k (Σ f_i); et

(d) on divise le résultat calculé en (b) par le résultat calculé en (c).

exemple

Quand on traite une distribution continue on prend le centre de classe comme valeur x_I comme dans l'exemple suivant qui utilise les données du tableau 3.2 :

Tableau 5.7 Poids de 63 poissons (kg)

Classe (kg)	Centre de classe (x_i)	Fréquence (f_i)	Fréquence × Centre de classe ($f_i . x_i$)
2,0–2,9	2,45	7	17,15
3,0–3,9	3,45	19	65,55
4,0–4,9	4,45	16	71,20
5,0–5,9	5,45	12	65,40
6,0–6,9	6,45	6	38,70
7,0–7,9	7,45	3	22,35
Total		63	280,35

Source : Tableau 3.2

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{280,35}{63} = 4,45 \text{ kg}$$

limites de classe réelles

Notez que le le centre de l'étendue 2,0–2,9 est noté 2,45. C'est parce que l'étendue réelle est de 1,95 à 2,95, car on présume que les poids des poissons ont été arrondis à la décimale supérieure. Bien sûr, si on a les données d'origine (comme dans le cas des poissons), il vaut mieux calculer la moyenne directement à partir des données (le résultat est similaire dans ce cas : 4,42 à partir des données d'origine et 4,45 à partir de la distribution de fréquences).

exemple

Prenons quelques exemples de calcul. Voici la distribution des pourcentages des notes d'un examen de mathématique :

Tableau 5.8 Distribution des notes à un examen de mathématique par des étudiants

Notes %	Fréquence f_i	Centre de classe x_i	$f_i x_i$
0 à moins de 10	6	5	30
10 à moins de 20	14	15	210
20 à moins de 30	20	25	500
30 à moins de 40	35	35	1.225
40 à moins de 50	67	45	3.015
50 à moins de 60	86	55	4.730
60 à moins de 70	59	65	3.835
70 à moins de 80	37	75	2.775
80 à moins de 90	19	85	1.615
90 à 100	3	95	285
Total	346	–	18.220

Source : fictive

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = 18.220 / 346 = 52,7$$

La moyenne arithmétique des notes de cet examen est donc 52,7.

avantages et inconvénients

La moyenne arithmétique a beaucoup d'avantages en tant que moyenne. Elle est facile à calculer, elle existe toujours pour des données quantitatives, la plupart des gens la comprennent, et elle est facile à utiliser dans des travaux statistiques plus poussés. Elle a aussi pourtant quelques inconvénients et peut créer des problèmes dans certains cas. La valeur d'une moyenne arithmétique peut être sévèrement affectée par une ou deux valeurs élevées, ce qui peut arriver quand on a une distribution avec beaucoup de petites valeurs et quelques valeurs élevées. Dans ce cas l'utilisation de la moyenne arithmétique peut prêter à confusion.

exemple

Pour voir ce qui peut arriver, regardons le tableau suivant :

Tableau 5.9 Calcul de la moyenne arithmétique d'une distribution asymétrique

Revenus mensuels nets (dollars)	Fréquence	Centre de classe x_i	$f_i x_i$
0–50	42.872	25	1.071.800,0
51–150	1.213	100,5	121.906,5
151–250	1.591	200,5	318.995,5
251–350	1.861	300,5	559.230,5
351–500	1.383	425,5	588.466,5
501–700	827	600,5	496.613,5
701–900	607	800,5	485.903,5
900 et plus	737	(1.000)*	737.000,0
Total	51.091		4.379.916,0

Source : Solomon Islands Statistical Bulletin No. 18/95, Tableau 3.1.2, p8. – estimation du centre de classe

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{4.379.916}{51.091} = 85,7$$

pas une mesure instructive

Dans cet exemple, bien que les revenus moyens par ménage soient 85,70 dollars, on peut voir aussi que 85 pour cent des ménages touchent moins que cette valeur (ils ont un revenu moyen inférieur à 50 dollars). Cette forme de moyenne ne semble donc pas très représentative de la distribution, et il est trompeur de dire que la moyenne des revenus des ménages est de 85 dollars. Ceci arrive souvent dans les distributions de revenus et dépenses. La moyenne arithmétique des revenus n'est pas très instructive en tant que moyenne.

autres difficultés

La calcul du tableau précédent illustre aussi trois autres difficultées rencontrées avec les moyennes arithmétiques. Quand on traite une distribution de fréquences qui a une **classe ouverte**, comme la classe '900 et plus', on doit faire une estimation du centre de classe. Dans notre exemple on a utilisé 1.000 dollars, mais ce chiffre était pris au hasard et était peut-être faux. Si en réalité le centre de classe était 1.500 dollars, alors notre moyenne arithmétique serait 92,94 dollars de revenus par mois.

ATTENTION

LA MOYENNE ARITHMÉTIQUE PEUT VARIER SENSIBLEMENT EN FONCTION DE L'ESTIMATION DU CENTRE DE CLASSE.

les moyennes de classes réelles peuvent varier par rapport à celles des classes estimées

La deuxième difficulté avec les distributions qui ont des valeurs très élevées ou très basses (les distributions asymétriques) vient de l'utilisation des centres de classes comme estimations de moyennes. Dans l'exemple ci-dessus on prend les centres de classes comme estimations des moyennes, mais la moyenne réelle de chaque classe sera probablement inférieure à ces moyennes

estimées. Par exemple, pour la classe 51–150, la moyenne réelle est probablement bien moindre que 100,5 car la plupart des observations sont certainement groupées vers le bas de la classe. On a donc tendance à gonfler légèrement la moyenne.

perte d'informations des valeurs arrondies

La troisième difficulté qui peut survenir avec les moyennes arithmétiques est d'obtenir une valeur qui n'existe manifestement pas. Dans l'exemple du tableau 5.6, on a affaire à des données discontinues, mais on a obtenu une moyenne de 2,1 destinations. Bien sûr, on ne peut pas avoir une valeur de 2,1 destinations par voyage, et beaucoup de non statisticiens trouveraient de telles réponses extrèmement difficile à comprendre. On pourrait arrondir la réponse au chiffre entier le plus proche, 2 dans ce cas, mais en le faisant on pert un grand nombre d'informations.

distributions bimodales

Une autre difficulté survient avec les distributions dites 'bimodales' (c'est-à-dire, une distribution de fréquences qui a deux pics au lieu d'un). Ici la moyenne se trouvera probablement quelque part entre les deux pics, et ne sera pas représentative de la distribution (elle pourrait même donner une idée fausse de la distribution). Par exemple, un jeune homme peut être déçu si, après lui avoir dit que vous allez lui présenter trois jeunes filles célibataires séduisantes d'une moyenne d'âge de 20 ans, il réalise qu'une des jeunes filles a 1 an, une autre a 2 ans, et la troisième 57 ans !

CALCULER LA MÉDIANE D'UNE DISTRIBUTION DE FRÉQUENCES

deux méthodes

On peut calculer une distribution de fréquences de deux manières. La première, c'est d'utiliser une méthode de calcul directement à partir du tableau, la deuxième, c'est d'utiliser un diagramme précis de l'ogive. Pour illustrer la première méthode on utilise les données de poissons :

Tableau 5.10 Distribution des poids de poissons

Classe (kg)	Fréquence	Fréquence cumulée
2,0– 2,9	7	7
3,0– 3,9	19	26
4,0–4,9	16	42
5,0–5,9	12	54
6,0–6,9	6	60
7,0–7,9	3	63
Total	63	

Source : Tableau 3.2

trouver la valeur centrale dans les données groupées

Puisqu'il y a 63 observations et que la médiane est la valeur centrale, la médiane est la valeur de la 32ème observation dans la liste classée (rappelez-vous que l'observation médiane est l'observation :

$$\frac{(n+1)}{2} \quad \textbf{et pas} \text{ l'observation} \quad \frac{n}{2}$$

À la colonne de fréquence cumulée du tableau 5.10 on peut voir que la 32ème observation tombe dans la tranche 4,0–4,9 kilos (la classe médiane) La fréquence cumulée du groupe qui précède la classe médiane (appelée *c*) est 26, aussi on doit aller à l'observation :

$$\{\frac{(n+1)}{2}\} - c = \{\frac{(63+1)}{2}\} - 26 = \text{6ème observation}$$

dans la classe 4,0–4,9 kilos. Mais quelle valeur prend cette observation ?

Le tableau 5.10 montre qu'on a 16 observations dans la classe médiane, mais qu'on n'a aucune précision quant aux valeurs que prennent ces 16 observations. Elle pourraient avoir une valeur de 3,96 ou elles pourraient toutes être de 4,94.

ASTUCE

PAR CONVENTION ON PRÉSUME QUE LES VALEURS DE LA CLASSE SONT DISTRIBUÉES DE MANIÈRE RÉGULIÈRE DANS L'INTERVALLE DE CLASSE.

hypothèses

Pour ce faire, on suppose :

1 que les différences entre les observations classées sont égales,

2 que la différence entre la limite de classe inférieure réelle et l'observation la plus basse, et la différence entre la limite de classe supérieure réelle et l'observation la plus haute, sont chacune égales à la moitié des différences entre les observations.

exemple

Prenons un exemple simple. On suppose qu'on a une classe d'une fréquence de 5, et des limites réelles de classe de 0 et 10. On suppose que les 5 observations sont réparties uniformément à travers la classe, c'est-à-dire qu'elles ont les valeurs 1, 3, 5, 7 et 9. La différence entre les observations classées par ordre de grandeur est la même (2); la différence entre la limite inférieure réelle de classe et l'observation la plus basse est 1; la différence entre la limite supérieure réelle de classe et la valeur la plus haute est 1, 1 étant la moitié de la valeur de l'intervalle entre chaque observation. Notez que la différence entre les observations classées est égale à l'intervalle de classe (*i*) divisé par la fréquence de classe (*f*) (c'est-à-dire, $^{i}/_{f} = {}^{10}/_{5} = 2$, dans ce cas).

exemple des poissons

Si on revient à notre exemple des poissons, on a maintenant une méthode pour localiser la médiane. Voici les étapes principales :

méthode

- On part de la limite de classe inférieure réelle (appelée 'ℓ') de la classe médiane (dans notre exemple 'ℓ' = 3,95). Cette valeur représente *c* = 26 observations dans la liste mise en ordre.

- On additionne la moitié de la différence entre les observations classées par ordre de grandeur dans la classe médiane, pour localiser la première observation de la classe médiane :

$$(½) \times (^{i}/_{f}) = (½) \times (^{1}/_{16}) = {}^{1}/_{32} = 0{,}03125$$

Pour localiser la m ième observation de la classe médiane, on additionne m - 1 fois la différence entre les observations classées par ordre de grandeur.

$$(m - 1) \times (^{i}/_{f}) = (m - 1) \times (^{1}/_{16});$$

et, puisque la médiane est l'observation $\{\frac{(n+1)}{2}\} - c$ dans la classe médiane, on doit ajouter :

$$([\{\frac{(n+1)}{2}\} - c] - 1) \times (\frac{i}{f})$$

$$= ([\{\frac{(63+1)}{2}\} - 26] - 1) \times (\frac{1}{16})$$

$$= \frac{5}{16} = 0.3125$$

médiane

Donc, la médiane est :

$$3{,}95 + 0{,}03125 + 0{,}3125 = 4{,}29$$

$$\ell + (½) \times (^{i}/_{f}) + ([\{\frac{(n+1)}{2}\} - c] - 1) \times (\frac{i}{f})$$

formule

Grâce aux points (a) jusqu'à (d) ci-dessus, on peut voir que la médiane est donnée par la formule :

$$\tilde{x} = \ell + (½) \cdot (^{i}/_{f}) + ([\{\frac{(n+1)}{2}\} - c] - 1) \times (\frac{i}{f})$$

$$= \ell + (\frac{i}{f}) \cdot [(\frac{n}{2}) + (\frac{1}{2}) - c - 1 + (\frac{1}{2})]$$

$$= \ell + (\frac{i}{f}) \cdot [(\frac{n}{2}) - c]$$

symboles

ℓ est la limite réelle de la classe médiane

i est l'intervalle de classe de la classe médiane

f est la fréquence de classe de la classe médiane

c est la fréquence cumulée de la classe qui précède la classe médiane

n est le nombre d'observations

exemple

Si on applique cette formule aux poissons ça donne :

$$= 3{,}95 + \left(\frac{1}{16}\right).\left[\left(\frac{63}{2}\right) - 26\right]$$
$$= 4{,}29$$

distributions discontinues

Pour une distribution discontinue, le calcul est beaucoup plus simple. Par exemple, on prend les données sur les nombres de chambres par ménage. Elles ont une distribution de fréquences de :

Tableau 5.11 Nombre de destinations de vacances par nombre de vacanciers (Îles Cook, 1991)

Nombre de destinations	Personnes	Fréquence cumulée
1	1.811	1.811
2	683	2.494
3	342	2.836
4	273	3.109
5	137	3.246
6 ou plus	171	3.417
Total	3.417	

Source : Enquête sur les visiteurs des Îles Cook 1991, Survey Report No. 13; COT, Tableau 23, p. 31.

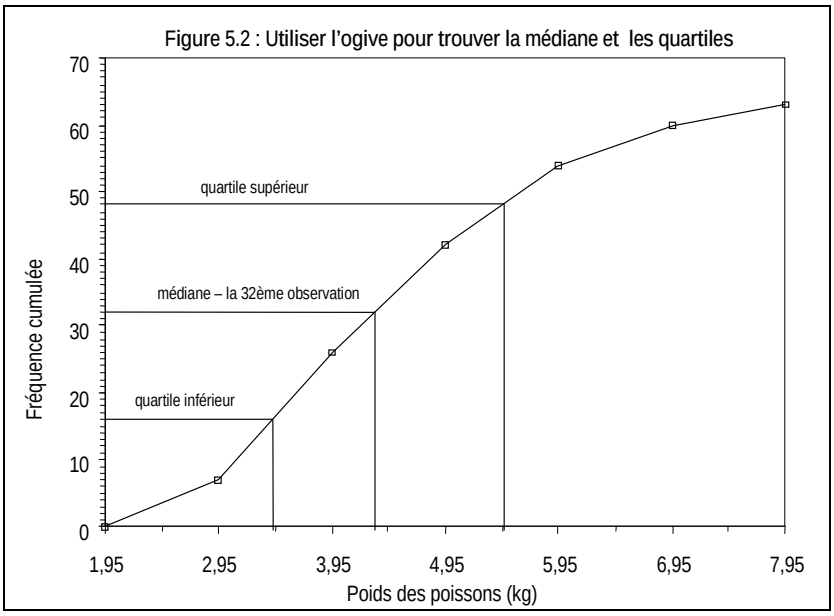
calcul de la médiane

Puisqu'on a 3.417 observations, la médiane est l'observation (3.417 + 1) / 2 = 1.709ème. Dans la colonne des fréquences cumulées on peut voir que la 1.709ème observation a la valeur 1, c'est-à-dire que la médiane des destinations par voyage est 1.

utiliser l'ogive

Une autre manière de déterminer approximativement la médiane se fait directement à partir de **l'ogive** de la distribution de fréquences. Regardons par exemple l'ogive 'moins que' des poissons :

Figure 5.2 : Utiliser l'ogive pour trouver la médiane et les quartiles



Source : Tableau 5.10

lire l'échelle horizontale

La médiane est la 32ème observation, donc on lit simplement sur l'échelle horizontale la position qui correspond à la fréquence 32 sur l'échelle verticale, comme sur la figure 5.2. On trouve que la médiane est 4,4 kilos, ce qui correspond à peu près à la valeur obtenue avec la première méthode.

on tend à surestimer

En déterminant la médiane à partir de l'ogive, on va surestimer légèrement la médiane. Quand on trace l'ogive, on suppose en effet que les données à l'intérieur d'une classe sont réparties uniformément à travers la classe, avec la valeur la plus large qui coïncide avec la limite supérieure réelle de classe, et avec la différence entre la limite inférieure réelle de classe et la valeur la plus basse qui soit la différence entre les valeurs des données. Ceci équivaut à remplacer le terme n/2 de la formule de la médiane utilisée précédemment par le terme (n+1)/2 . De toute manière, puisqu'on n'obtient qu'une approximation en déterminant la médiane à partir de l'ogive (en partie à cause du manque de précision du tracé de l'ogive), une surestimation de ce genre n'est pas très importante.

QUARTILES

quatre parts égales

La médiane est la valeur de l'observation qui divise la fréquence totale en deux parts égales. De la même manière, on peut déterminer d'autres valeurs qui divisent la fréquence en d'autres fractions. Les **quartiles**, comme leur nom le suggère, divisent la fréquence totale en quatre parts égales.

quartiles inférieur et supérieur

Le **quartile inférieur (ou premier quartile)** aura un quart des observations en-dessous de sa valeur, et trois quarts des observations en-dessus de sa valeur. Le **quartile supérieur (ou troisième quartile)** aura trois quarts des observations en-dessous de sa valeur et un quart des observations en-dessus de sa valeur.

formule

En termes de nombre d'observations on a alors :

le ***Quartile inférieur*** est la $\frac{n+1}{4}ème$ valeur de la liste classée

la ***Médiane*** est la $\frac{n+1}{2}ème$ valeur de la liste classée

le ***Quartile supérieur*** est la $\frac{3(n+1)}{4}ème$ valeur de la liste classée

exemple

Dans le cas où n est égal à 63, ça correspond à :

quartile inférieur 16ème valeur de la liste classée

médiane 32ème valeur de la liste classée

quartile supérieur 48ème valeur de la liste classée

Le premier quartile, ou quartile inférieur **(Q_1)** peut être calculé avec la formule :

$$Q_1 = \ell + \left(\frac{i}{f}\right).\left[\left\{\frac{(n-1)}{4}\right\} - c\right]$$

et le troisième quartile, ou quartile supérieur **(Q_3)** peut être calculé avec la formule :

$$Q_3 = \ell + \left(\frac{i}{f}\right).\left[\left\{\frac{(3n+1)}{4}\right\} - c\right]$$

la même logique que la médiane

Ces formules peuvent être dérivées en utilisant la même logique que celle utilisée pour dériver la formule de la médiane.

on peut aussi utiliser l'ogive

On peut ausi dériver des valeurs approximatives de quartiles à partir de l'ogive, comme on l'a fait pour la médiane. À la figure 5.2 le quartile inférieur a été trouvé en lisant sur l'échelle horizontale la position correspondant à la fréquence 16 de l'échelle verticale. Le quartile supérieur a été trouvé de la même façon, en retrouvant sur la barre des x la valeur de la 48ème observation. Ces valeurs sont respectivement 3,4 kilos et 5,4 kilos.

Les valeurs dérivées à partir de l'ogive sont des approximations pour la même raison que dans le cas des médianes.

LA MOYENNE GÉOMETRIQUE

formule

La **moyenne géométrique** de n nombres $x_1, x_2 \ldots x_n$ est définie par :

$$\textit{Moyenne géométrique} = (x_1 . x_2 . \ldots . x_n)^{1/n}$$

On écrit plus simplement :

$$\textit{Moyenne géométrique} = \left(\prod_{i=1}^{n} x_i\right)^{1/n}$$

produit

Le signe Π (la lettre grecque pi) signifie le **produit** d'un jeu de chiffres, de la même manière que Σ signifie la somme d'un jeu de chiffres. Par exemple, la moyenne géométrique de 3, 7 et 9 est :

$$(3 \times 7 \times 9)^{1/3} = (189)^{1/3} = 5{,}74$$

peut être difficile à calculer

La moyenne peut être assez difficile à calculer, en particulier pour une longue série de chiffres. Quelques calculettes ont des fonctions spéciales qui aident à faire ce calcul, et si un ordinateur est disponible les calculs sont très simples.

on ne peut pas effectuer le calcul avec certaines données

On ne peut pas calculer la moyenne géométrique si une des valeurs est négative ou égale à zéro. Pour cette raison l'usage de la formule est assez restreint, mais elle est très utile quand on travaille sur des indices et des taux d'accroissement.

exemple

Par exemple, on suppose qu'on a les augmentations annuelles de l'indice des prix d'un pays : de 1992 à 1993 16,7%, de 1993 à 1994 10,8%, de 1994 à 1995 2,1%, et de 1995 à 1996 4,5%. La moyenne annuelle de l'augmentation de l'indice des prix de 1992 à 1996 est donnée par la moyenne géométrique de ces chiffres, et pas la moyenne arithmétique. Le taux d'augmentation annuelle est calculé en calculant d'abord la moyenne géométrique des chiffres 1,167, 1,108, 1,021 et 1,045 :

$$\textbf{\textit{Moyenne géometrique}} = (1{,}167 \times 1{,}108 \times 1{,}021 \times 1{,}045)^{1/4} = 1{,}0838$$

utilisée pour les indices

Donc le taux d'augmentation moyen est 8,38%. Notez que la moyenne arithmétique donnerait un résultat de 8,53%.

La moyenne géométrique donnera toujours un résultat inférieur ou égal à la moyenne arithmétique. Elle on l'utilise en parallèle avec la moyenne arithmétique pour calculer les moyennes de taux d'accroissement composés dans le temps.

RÉSUMÉ DES DIFFÉRENTS TYPES SUPPLÉMENTAIRES DE MOYENNES

Quartiles

AVANTAGES	INCONVÉNIENTS
☺ ils sont assez faciles à calculer	☹ ce ne sont pas des 'moyennes', mais des mesures de position
☺ ils ne sont pas altérés par des valeurs extrêmes	☹ ils ne prennent pas en compte toutes les valeurs de l'échantillon
☺ ils peuvent être déterminés dans tous les cas, sauf pour les données nominales	☹ ils peuvent produire une mesure qui n'existe pas en réalité
	☹ ils sont difficiles à manipuler en algébre

☺☹

Moyenne Géometrique

AVANTAGES	INCONVÉNIENTS
☺ elle utilise toutes les données d'un échantillon	☹ elle peut être difficile à calculer
☺ elle est moins altérée par les valeurs extrêmes que la moyenne arithmétique	☹ on ne peut pas l'utiliser si on des valeurs négatives ou égales à zéro
☺ elle peut être manipulée par des formules d'algèbre	☹ elle est sérieusement altérée par les valeurs extrèmes
☺ elle est particulièrement utile pour calculer des moyennes de taux d'accroissement, des indices et des changements de pourcentages	☹ elle n'est en général pas bien comprise