M1/Statistique descriptive

3. LES CARACTÉRISTIQUES DE TENDANCE CENTRALE

Présentation générale

Pour résumer l'information statistique contenue dans un tableau de données, on calcule des paramètres qui permettent d'éclairer sur la position du noyau (centre de la distribution).

Ces paramètres sont appelés caractéristiques de position ou de tendance centrale.

Les plus souvent utilisés sont le mode, la moyenne arithmétique, la médiane (et plus généralement les quantiles).

Le mode

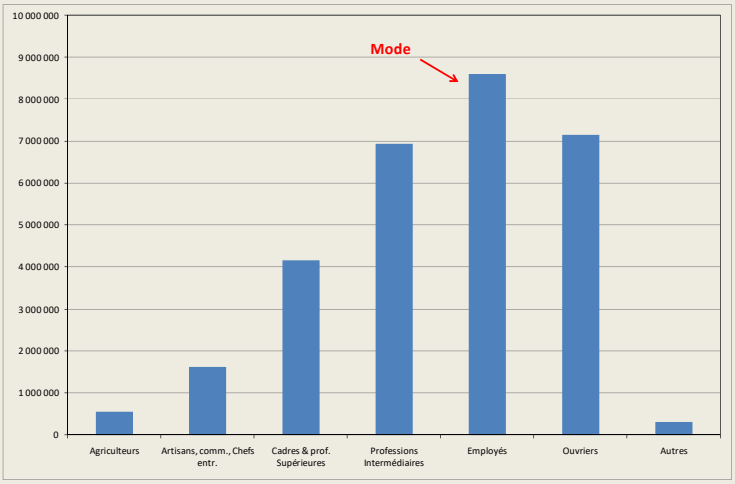
Le mode est la valeur de la variable qui a l'effectif (ou la fréquence relative) le plus grand. On le note Mo.

On peut le déterminer pour tous les types de variable : qualitative, quantitative discrète, quantitative continue.

Dans le cas d'une variable qualitative ou quantitative discrète, la détermination est simple : il s'agit de la modalité de la variable qui correspond à l'effectif maximal, soit la barre (cas d'un diagramme en barres) ou le bâton (cas d'un diagramme en bâtons) le plus long.

Le mode : cas d'une variable qualitative

Distribution de la population active selon la PCS (France métropolitaine, 2006)



M1_Analyse statistique (JFL) 4

Dans le cas d'une variable qualitative, les effectifs (ou fréquences relatives) conditionnent la hauteur des barres. En effet, la base de chaque barre est constante. Le mode est donc donné directement par la hauteur des barres.

Pour une variable quantitative discrète, le principe de détermination est le même. En effet, la représentation graphique est alors un diagramme en bâtons. Là aussi, la hauteur de chaque bâton est proportionnelle à l'effectif (ou à la proportion). Le mode correspond donc au bâton le plus important.

Le mode : cas d'une variable quantitative continue

Si la variable est quantitative continue, la détermination du mode peut être décomposée en plusieurs étapes :

1) Détermination de la classe modale ;
2) Détermination du mode au sein de la classe modale. Cette détermination peut être réalisée de manière graphique ou par le calcul.

Attention : on raisonne toujours à partir des effectifs ou fréquences relatives corrigées.

M1_Analyse statistique (JFL) 5

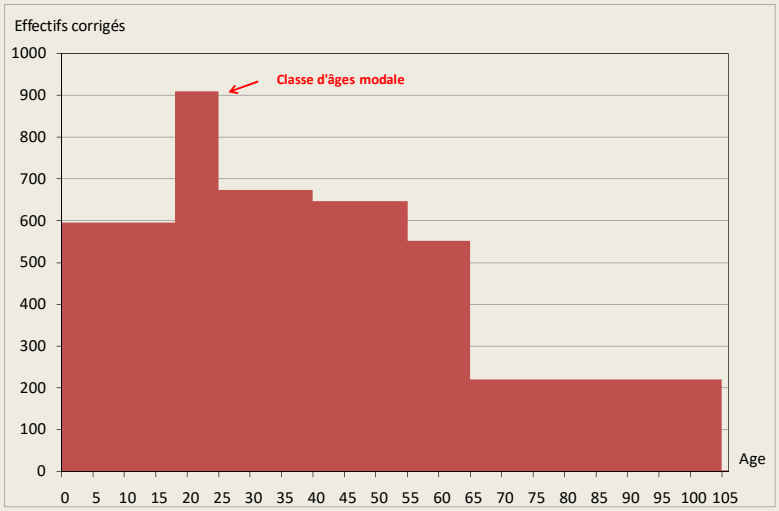
Quand la variable est quantitative continue, la détermination est aussi simple que pour les variables qualitatives ou quantitatives discrètes, à condition que les différentes modalités ont toutes la MÊME AMPLITUDE. C'est par exemple le cas lorsque l'on a une distribution par âge détaillé : chaque classe d'âge a une amplitude de 1 an.

Quand l'amplitude des classes sont différentes, alors il faut commencer par corriger les effectifs ou les proportions.

Le mode correspond alors à la modalité qui présente l'effectif corrigé le plus important.

Détermination de la classe modale pour une variable quantitative continue

Distribution de la population de la ville de Laval selon l'âge en 2006 (INSEE, 2009)



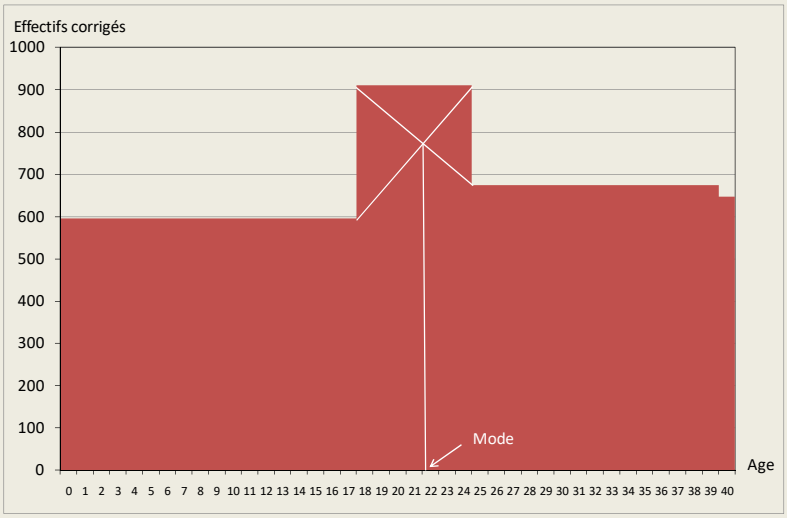
M1_Analyse statistique (JFL) 6

Dans cet exemple, on a représenté la structure par grands groupes d'âges de la population de la commune de Laval en 2006. Les effectifs qui figurent sur l'axe des ordonnées sont les effectifs corrigés, qui tiennent compte de l'amplitude inégale des classes d'âges initiales (0-17, 18-24, 25-39, 40-54, 55-64 et 65 ans et plus).

La classe d'âges modale est ici celle des 18-24 ans.

Détermination graphique du mode pour une variable quantitative continue

Zoom de la distribution de la population de la ville de Laval selon l'âge en 2006 (INSEE, 2009)



M1_Analyse statistique (JFL) 7

On a grossi ici la représentation graphique figurant sur la diapositive précédente.

La détermination graphique du mode, au sein de la classe d'âges modale, s'obtient en :

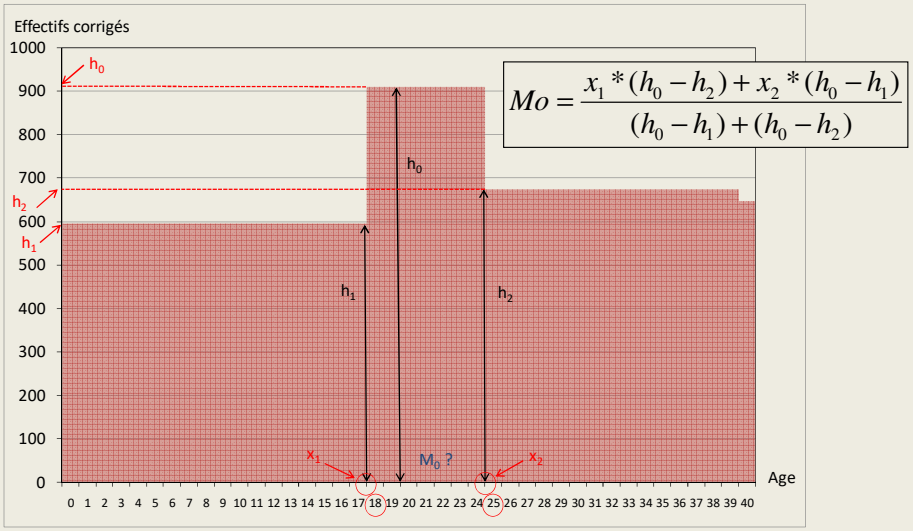
1) reliant le coin supérieur gauche de la classe d'âges modale au coin supérieur gauche de la classe d'âges suivante (ici les 25-39 ans) ;

2) reliant le coin supérieur droit de la classe d'âges modale au coin supérieur droit de la classe d'âges précédente (ici les 0-17 ans).

3) Le mode se situe sur l'axe des abscisses au point d'intersection d'une droite perpendiculaire à cet axe et passant par le point d'intersection des deux segments de droite précédemment définis.

Détermination par le calcul du mode pour une variable quantitative continue

Zoom de la distribution de la population de la ville de Laval selon l'âge en 2006 (INSEE, 2009)



M1_Analyse statistique (JFL) 8

Le mode peut également être défini au moyen d'un calcul.

Celui-ci est fondé sur un rapport de proportionnalité entre les hauteurs (h) et les abscisses (x).

On va considérer que le rapport entre les écarts entre, d'une part, h_0 et h_1 et, d'autre part, h_0 et h_2 est égal au rapport entre les écarts entre, d'une part, M_0 et x_1 et, d'autre part, M_0 et x_2.

On peut donc écrire :

$$\frac{(h_0 - h_1)}{(h_0 - h_2)} = \frac{(M_0 - x_1)}{(x_2 - M_0)}$$

$$(x_2 - M_0) * (h_0 - h_1) = (M_0 - x_1) * (h_0 - h_2)$$

$$x_2 * (h_0 - h_1) + x_1 * (h_0 - h_2) = M_0 * (h_0 - h_2) + M_0 * (h_0 - h_1)$$

$$M_0 = \frac{x_2 * (h_0 - h_1) + x_1 * (h_0 - h_2)}{(h_0 - h_1) + (h_0 - h_2)}$$

La médiane

La médiane est la valeur de la variable qui partage la série statistique ou la population en deux effectifs égaux. Elle est notée Me.

On peut déterminer la médiane pour :

- une série de données brutes ;
- une variable quantitative discrète ;
- une variable quantitative continue .

La médiane d'une série de données brutes (nombre pair de données)

(1) Données brutes

Hommes	e_0 en bonne santé	
Allemagne	58,8	
Autriche	58,4	
Belgique	63,3	
Bulgarie		
Chypre	63,0	
Danemark	67,4	
Espagne	63,2	
Estonie	49,5	
Finlande	56,7	
France	**63,1**	
Grèce	65,9	
Hongrie	55,0	
Irlande	62,7	
Italie	62,8	(e)
Lettonie	50,9	
Lituanie	53,4	
Luxembourg	62,2	
Malte	69,0	
Pays-Bas	65,7	
Pologne	57,4	
Portugal	58,3	
République tchèque	61,3	
Roumanie	60,4	
Royaume-Uni	64,8	(e)
Slovaquie	55,4	
Slovénie	58,7	
Suède	67,5	
UE (15 pays)	:	
UE (27 pays)	61,6	(e)

(2) Classement des données

Hommes	e_0 en bonne santé
Estonie	49,5
Lettonie	50,9
Lituanie	53,4
Hongrie	55,0
Slovaquie	55,4
Finlande	56,7
Pologne	57,4
Portugal	58,3
Autriche	58,4
Slovénie	58,7
Allemagne	58,8
Roumanie	60,4
République tchèque	61,3
Luxembourg	62,2
Irlande	62,7
Italie	62,8
Chypre	63,0
France	**63,1**
Espagne	63,2
Belgique	63,3
Royaume-Uni	64,8
Pays-Bas	65,7
Grèce	65,9
Danemark	67,4
Suède	67,5
Malte	69,0
Bulgarie	
UE (15 pays)	:
UE (27 pays)	61,6

On veut déterminer l'espérance de vie médiane en bonne santé des hommes au sein de l'UE à 27 pays en 2007 :

(1) On classe les données par ordre croissant.

(2) On dispose d'une série comportant 26 données (une donnée manquante).

(3) La médiane partage cette série en deux effectifs égaux, soit deux ensembles de treize données.

(4) Dans ce cas (nombre de données pair), la médiane est la moyenne arithmétique des données situées en les 13ème et 14ème positions).

(5) Me = (61,3 + 62,2)/2 = 61,8 ans

M1_Analyse statistique (JFL) 10

Comme la série de données est paire, la médiane se situe entre les deux valeurs qui partagent cette série en deux effectifs de pays égaux.

Dans ce cas, pour déterminer la médiane, on calcule la moyenne arithmétique de ces deux valeurs.

Notons ici que la médiane ne tient pas compte du poids démographique inégal de chaque pays au sein de l'UE à 27.

Il s'agit donc bien d'une médiane d'une série BRUTE de données.

La médiane d'une série de données brutes (nombre impair de données)

(1) Données brutes

Femmes	e_0 en bonne santé	
Allemagne	58,4	
Autriche	61,1	
Belgique	63,7	
Bulgarie		
Chypre	62,7	
Danemark	67,4	
Espagne	62,9	
Estonie	54,6	
Finlande	58,0	
France	**64,2**	
Grèce	67,1	
Hongrie		
Irlande	65,3	
Italie	62,0	(e)
Lettonie	53,7	
Lituanie	57,7	
Luxembourg	64,6	
Malte	70,8	
Pays-Bas	63,7	
Pologne	61,3	
Portugal	57,3	
République tchèque	63,2	
Roumanie	62,4	
Royaume-Uni	66,2	(e)
Slovaquie	55,9	
Slovénie	62,3	
Suède	66,6	
UE (15 pays)	:	
UE (27 pays)	62,3	(e)

(2) Classement des données

Femmes	e_0 en bonne santé
Lettonie	53,7
Estonie	54,6
Slovaquie	55,9
Portugal	57,3
Lituanie	57,7
Finlande	58,0
Allemagne	58,4
Autriche	61,1
Pologne	61,3
Italie	62,0
Slovénie	62,3
Roumanie	62,4
Chypre	62,7
Espagne	62,9
République tchèque	63,2
Belgique	63,7
Pays-Bas	63,7
France	**64,2**
Luxembourg	64,6
Irlande	65,3
Royaume-Uni	66,2
Suède	66,6
Grèce	67,1
Danemark	67,4
Malte	70,8
Bulgarie	
Hongrie	
UE (15 pays)	:
UE (27 pays)	62,3

On veut déterminer cette fois l'espérance de vie médiane en bonne santé des femmes au sein de l'UE à 27 pays en 2007 :

(1) On classe les données par ordre croissant.

(2) On dispose cette fois d'une série comportant 25 données (deux données manquantes).

(3) La médiane partage cette série en deux effectifs égaux, soit deux ensembles de douze données.

(4) Dans ce cas la médiane correspond à la valeur du pays classé au milieu de la distribution, soit en 13ème position : Chypre.

(5) Me = 62,7 ans

M1_Analyse statistique (JFL) 11

Ici le nombre de données est impair. De ce fait, le partage de cette série en deux effectifs de pays égaux laisse nécessairement un pays de côté, celui que se trouve juste sur la ligne de partage. C'est la valeur correspondant à ce pays qui devient la médiane de la série brute.

La médiane d'une série quantitative discrète

Distribution des familles ayant au moins un enfant de moins de 25 ans (France métropolitaine, RR 2006, source INSEE)

Familles avec enfant de moins de 25 ans	n_i	n_i cumulés croissants	n_i cumulés décroissants	f_i	f_i cumulés croissants	f_i cumulés décroissants
1 enfant	3 760 052	3 760 052	8 803 092	43%	43%	100%
2 enfants	3 399 998	7 160 050	5 043 040	39%	81%	57%
3 enfants	1 238 871	8 398 921	1 643 042	14%	95%	19%
4 enfants ou plus	404 171	8 803 092	404 171	5%	100%	5%
Ensemble	8 803 092		0	100%		0%

Dans le cas présent, la médiane correspond à un effectif de 4 401 546 familles. Les effectifs cumulés croissants indiquent que cette valeur est atteinte lorsque l'on agrège les familles avec un enfant et une partie de celles comptant 2 enfants. La médiane est donc ici 2 enfants.

On peut donc dire que la moitié des familles compte au moins deux enfants de moins de 25 ans, ou bien que la moitié des familles compte un nombre d'enfants inférieur ou égal à deux enfants.

M1_Analyse statistique (JFL) 12

Comme il s'agit ici d'une variable quantitative discrète, la médiane ne peut correspondre qu'à une des valeurs prises par la variable.

Les familles comptant 1 enfant représentent 43 % de l'effectif total de familles comptant en leur sein au moins un enfant de moins de 25 ans. Pour atteindre 50 % de la population des familles, il faut ajouter une partie des familles comptant deux enfants.

De ce fait, la médiane correspond ici à la modalité nécessaire à ce que le seuil de 50 % soit atteint, soit deux enfants.

Dans ce cas, la formulation en clair est légèrement nuancée :

50% des familles ont au plus deux enfants ;

50 % des familles ont au moins deux enfants.

La modalité médiane « deux enfants » fait donc partie des deux ensembles qu'elle partage pourtant.

La médiane d'une série quantitative continue

On peut déterminer à la fois graphiquement et par le calcul la médiane d'une série quantitative continue.

Cette détermination suppose au préalable de calculer les effectifs ou fréquences relatives cumulées (croissants ou décroissants).

Exemple : Distribution par âge de la population du monde en 2007

Groupe d'âges	Borne d'âge inférieur	Effectif (en milliers)	Effectif cumulé croissant	Effectif cumulé décroissant	% cumulée croissante	% cumulée décroissante
0-14 ans	0	1 838 700	0	6 810 000	0%	100%
15-64 ans	15	4 426 500	1 838 700	4 971 300	27%	73%
65+ ans	65	544 800	6 265 200	544 800	92%	8%
Total		6 810 000	6 810 000	0	100%	0%

M1_Analyse statistique (JFL) 13

Lecture du tableau :

Effectif cumulé croissant ($n_i cc$) = 1,8 milliards de personnes sont âgées de moins de 15 ans ; 6,2 milliards sont âgées de moins de 65 ans, etc. Ce « 6,2 milliards de personnes âgées de moins de 65 ans » correspond à la somme des personnes âgées de 0-14 ans et de 15-64 ans.

Effectif cumulé décroissant ($n_i cd$) = 6,8 milliards de personnes sont âgées d'au moins 0 ans (ou de 0 an et plus) ; 4,9 milliards de personnes sont âgées d'au moins 15 ans (ou de 15 ans et plus). Ce dernier effectif correspond à la somme totale des individus moins ceux âgés de 0-14 ans.

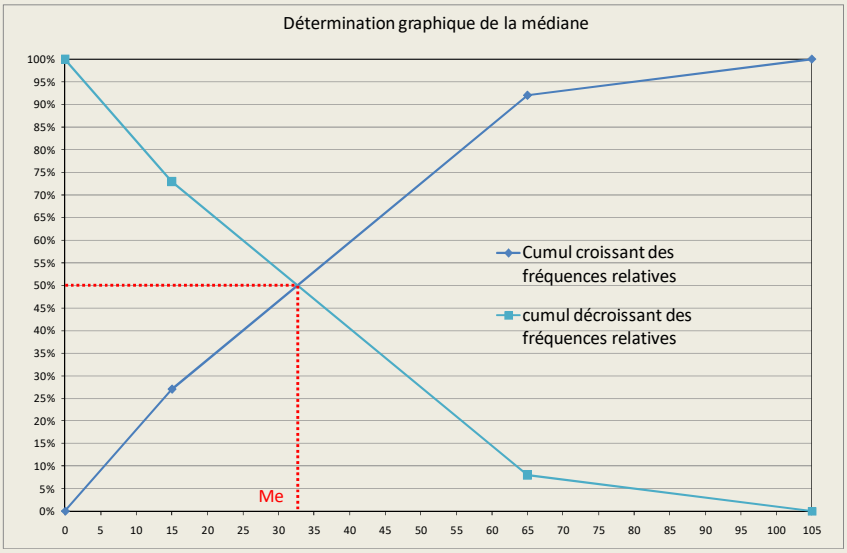
Proportion cumulée croissante ($f_i cc$) = 27 % des personnes sont âgées de moins de 15 ans, etc.

Proportion cumulée croissante ($f_i cc$) = 73 % des personnes sont âgées d'au moins moins 15 ans, etc.

Pour passer de $n_i cc$ à $f_i cc$, on effectue le calcul suivant : $f_i cc = \frac{n_i cc}{\sum n_i}$

L'opération est la même pour passer de $n_i cd$ à $f_i cd$.

Détermination graphique de la médiane d'une série quantitative continue



M1_Analyse statistique (JFL) 14

Graphiquement, la médiane d'une variable quantitative continue est l'abscisse qui correspond au point d'intersection entre la courbe des fréquences (absolues ou relatives) cumulées décroissantes et croissantes.

Dans le cas présent, l'âge médian est bien la valeur de la variable telle que la moitié des personnes a un âge inférieur à la médiane, tandis que l'autre moitié à un âge supérieur à cette même médiane.

Calcul de la médiane d'une série quantitative continue

Exemple : Distribution par âge de la population du monde en 2007

Groupe d'âges	Borne d'âge inférieur	Effectif (en milliers)	Effectif cumulé croissant	Effectif cumulé décroissant	% cumulée croissante	% cumulée décroissante
0-14 ans	0	1 838 700	0	6 810 000	0%	100%
15-64 ans	15	4 426 500	1 838 700	4 971 300	27%	73%
65+ ans	65	544 800	6 265 200	544 800	92%	8%
Total		6 810 000	6 810 000	0	100%	0%

La médiane est comprise entre 15 (x_{inf}) et 65 ans (x_{sup}). En effet, 27% (fcc_{inf}) de la population a moins de 15 ans, tandis que 92% (fcc_{sup}) de la population mondiale est âgée de moins de 65 ans.

$$Me = x_{\inf} + \frac{(50\ \% - fcc_{\inf})}{fcc_{\sup} - fcc_{\inf}} * (x_{\sup} - x_{\inf})$$

$$Me = 15 + \frac{(50\ \% - 27\ \%)}{(92\ \% - 27\ \%)} * (65 - 15) = 32,7$$

M1_Analyse statistique (JFL) 15

Pour calculer la médiane, on va considérer que la position relative de la médiane sur l'axe des âges, entre les bornes inférieure et supérieure de l'intervalle d'âge au sein duquel elle se trouve est la même que la position relative de la proportion cumulée correspondant à la médiane (50 %) par rapport aux proportions cumulées correspondant aux bornes inférieure et supérieure de l'intervalle d'âge au sein duquel la médiane se situe.

En conséquence :

$$\frac{M_0 - x_{\inf}}{x_{\sup} - x_{\inf}} = \frac{50\% - f_i cc_{\inf}}{f_i cc_{\sup} - f_i cc_{\inf}}$$

$$M_0 - x_{\inf} = \frac{50\% - f_i cc_{\inf}}{f_i cc_{\sup} - f_i cc_{\inf}} * (x_{\sup} - x_{\inf})$$

$$M_0 = x_{\inf} + \frac{50\% - f_i cc_{\inf}}{f_i cc_{\sup} - f_i cc_{\inf}} * (x_{\sup} - x_{\inf})$$

Généralisation : les quantiles

La médiane est le quantile d'ordre 50 %, c'est-à-dire qu'il partage la série statistique ou la population en deux effectifs de 50 % (donc en deux effectifs égaux).

Le quantile d'ordre 25% est le premier quartile (Q_1) : il partage la population en deux effectifs inégaux : 25 % des observations sont inférieures ou égales à Q_1 et 75% sont supérieures à Q_1.

Le quantile d'ordre 75 % est le troisième quartile (Q_3) : il partage la population en deux effectifs inégaux : 25 % des observations sont inférieures ou égales à Q_3 et 25% sont supérieures à Q_1.

Q_1, Q_2 (la médiane) et Q_3 sont des quartiles : les quartiles partagent la série statistique en quatre effectifs égaux comprenant chacun 25% de la population.

Il existe aussi les déciles (D_1 à D_9) qui partagent la population en dix groupes comprenant chacun 10 % de la population.

Enfin, les centiles (C_1 à C_{99}) partagent la population en cent groupes comprenant chacun 1% de la population.

M1_Analyse statistique (JFL) 16

La médiane est le quartile de rang 50 %.

Les quartiles, au nombre de trois (Q_1, Q_2, Q_3) partagent toute population en quatre groupes égaux regroupant chacun 25 % de la population (un quart, d'où le nom de quartile) :

- un quart de la population présente une valeur de la variable inférieure à Q_1;

- un quart de la population présente une valeur comprise entre Q1 et la médiane (Q_2) ;

- un troisième quart présente une valeur de la variable comprise entre la médiane et Q_3;

- enfin un quart de la population présente une valeur supérieure à Q_3.

Les quartiles sont eux même un cas particulier d'une forme plus générale : les quantiles.

La détermination des quantiles, qu'il s'agisse des quartiles, des déciles ou des centiles, se fait exactement comme pour la médiane. Lorsque le quantile se trouve au sein d'un intervalle (x_{inf};x_{sup}), on l'estime au moyen d'une interpolation linéaire.

Exemple :

$$\frac{Q_1 - x_{\inf}}{x_{\sup} - x_{\inf}} = \frac{25\% - f_i cc_{\inf}}{f_i cc_{\sup} - f_i cc_{\inf}}$$

$$Q_1 - x_{\inf} = \frac{25\% - f_i cc_{\inf}}{f_i cc_{\sup} - f_i cc_{\inf}} * (x_{\sup} - x_{\inf})$$

$$Q_1 = x_{\inf} + \frac{25\% - f_i cc_{\inf}}{f_i cc_{\sup} - f_i cc_{\inf}} * (x_{\sup} - x_{\inf})$$

La moyenne arithmétique

La moyenne arithmétique se calcule seulement pour les variables quantitatives, qu'elles soient discrètes ou continues.

On la note $\overline{x}$.

Il s'agit de la somme des valeurs observées divisée par le nombre d'observations.

Par exemple, si l'on cherche à déterminer le nombre moyen d'enfants par famille, il s'agit du rapport entre le nombre total d'enfants présents dans l'ensemble des familles divisée par le nombre de familles.

Autre exemple : l'âge moyen correspond au cumul des âges des individus composant une population donnée divisé par la taille de cette population (soit le nombre d'individus).

M1_Analyse statistique (JFL) 17

On peut prendre un exemple encore plus simple.

La note moyenne qu'un étudiant obtient dans le cadre de l'obtention d'un diplôme, correspond à la somme de ses notes rapportée au nombre de notes. Quand ces notes ont des coefficients différents, on pondère le poids de chaque note par son coefficient, puis l'on divise par la somme des coefficients.

Le calcul de l'âge moyen est identique :

L'âge correspond aux notes et l'effectif à chaque âge correspond pour sa part aux coefficients. Plus un groupe d'âge sera important, plus son poids (son « coefficient ») sera important dans le calcul de l'âge moyen.

Dans une moyenne, au numérateur figure donc toujours une somme, la somme des valeurs de la variable dont on cherche à calculer la moyenne. Au dénominateur, on a également toujours une somme : la somme des observations ou population totale.

Calcul de la moyenne arithmétique

Soit la variable X qui présente plusieurs modalités x_i.

A chaque modalité correspond un effectif n_i.

La somme des valeurs observées est égale à la somme des produits $n_i * x_i$.

Nombre d'enfants X	Effectif n_i	ni*xi
$x_1 = 1$	n_1	$n_1 * x_1$
$x_2 = 2$	n_2	$n_2 * x_2$
.	.	.
.	.	.
.	.	.
$X_p=p$	n_p	$n_6 * x_6$
Somme	N	$\sum n_i*x_i$

Et la moyenne s'écrit :

$$\bar{x} = \frac{\sum_{i=1}^{p} n_i * x_i}{\sum_{i=1}^{p} n_i}$$

M1_Analyse statistique (JFL) 18

Dans le cas d'une variable quantitative discrète, on multiplie la valeur de chaque modalité par l'effectif correspondant, puis on somme l'ensemble de ces produits.

Dans l'exemple ci-dessus, n_1*x_1 correspond au nombre d'enfants présents dans des fratries de 1 enfants (nombre d'enfants uniques).

n_2*x_2 correspond au nombre d'enfants vivant au sein de fratrie de deux enfants, etc.

Au total, au numérateur, on additionne tous les enfants quelle que soit la taille de la fratrie.

Puis, pour obtenir la moyenne, on divise par le nombre de fratries, ici le nombre de familles (puisque toutes les familles ont au moins un enfant dans ce cas précis).

Calcul de la moyenne arithmétique dans le cas d'une variable continue (modalités en classes)

Calcul de l'âge moyen de la population mondiale en 2007

Groupe d'âges	Borne inférieure de la classe d'âges	Effectif	centre de classe	$n_i * x_i$
0-14 ans	0	1 838 700 000	7,5	13 790 250 000
15-64 ans	15	4 426 500 000	40,0	177 060 000 000
65 ans et +	65	544 800 000	85,0	46 308 000 000
	105			
Total		6 810 000 000		237 158 250 000
Age moyen				34,8

M1_Analyse statistique (JFL) 19

Quand la variable continue a été discrétisée et est présentée sous forme de classes, il faut au préalable définir un centre de classe, qui est la valeur moyenne que prennent les individus regroupés au sein de cette classe.

Par exemple, dans l'exemple ci-dessus, on considère que les individus regroupés au sein de la tranche d'âges 0-14 ans ont en moyenne 7,5 ans [(0+15)/2)].

Le fait de prendre le centre de classe minimise l'erreur moyenne commise en faisant une telle approximation.

Quand des classes sont ouvertes (cas des 65 ans et +), on ferme la classe de façon plausible, sur la base d'informations complémentaires (ici l'âge à partir duquel les effectifs deviennent négligeables).

Une fois les centres de classes définis, le calcul est le même que celui énoncé dans la diapositive précédente.

Une moyenne arithmétique parmi d'autres : l'espérance de vie (e_0)

Extrait de la table de mortalité de la France 2003-2005

Âge x	S(x)	D(x)
0	100 000	387
1	99 613	30
2	99 583	21
3	99 562	16
…	…	…
50	95 155	395
51	94 760	414
52	94 346	437
…	…	…
101	1 351	389
102	962	255
103	707	197
104	510	510

Source : INSEE

$$\bar{x} = \frac{\sum_{i=0}^{105} (x_i + 0{,}5) * D_i}{\sum_{i=0}^{105} D_i}$$

M1_Analyse statistique (JFL) 20

Une table de mortalité comporte notamment deux séries, celle des survivants à chaque âge et celle des décès de la table (décès entre deux âges exacts).

La série des survivants correspond à une série de fréquences cumulées décroissantes : la série des décès cumulés décroissants.

La série des décès est une distribution des individus selon l'âge au décès. La population est ici la cohorte initiale composée de 100 000 personnes dont on mesure le calendrier d'extinction en l'absence de tout autre phénomène que celui étudié, le décès.

La variable d'analyse est l'âge, variable quantitative continue.

On peut donc calculer la moyenne que prend cette caractéristique : l'âge moyen au décès. Il s'agit de la moyenne des âges au décès pondérés par le nombre de personne décédées à chacun des âges.

On fait l'hypothèse d'une répartition uniforme des décès entre deux âges exacts : ils surviennent donc en moyenne à x_i + 0,5 an.

$$\bar{x} = \frac{\sum_{i=0}^{104} (x_i + 0{,}5) * D_i}{\sum_{i=0}^{105} D_i}$$

Une moyenne arithmétique parmi d'autres : l'espérance de vie (e_0)

$$\bar{x} = \frac{\sum_{i=0}^{105}(x_i + 0{,}5) * D_i}{\sum_{i=0}^{105} D_i}$$

$$\bar{x} = e_0 = \frac{\sum_{i=0}^{105}(x_i + 0{,}5) * (S_i - S_{i+1})}{\sum_{i=0}^{105} D_i}$$

$$\bar{x} = e_0 = \frac{0{,}5*(S_0 - S_1) + 1{,}5*(S_1 - S_2) + 2{,}5*(S_2 - S_3) + \ldots + 103{,}5*(S_{103} - S_{104}) + 104{,}5*(S_{104} - S_{105})}{S}$$

$$e_0 = \frac{0{,}5*S_0 + S_1 + S_2 + S_3 + \ldots + S_{103} + S_{104} - 104{,}5*S_{105}}{S_0}$$

$$e_0 = 0{,}5 + \frac{\sum_{i=1}^{104} S_i}{S_0}$$

M1_Analyse statistique (JFL) 21

On peut exprimer cet âge moyen autrement.

Les décès entre les âges i et i+1 correspondent à la différence entre les survivants aux âges i et i+1 :

$D_i = S_i - S_{i+1}$

On remplace dans la formule classique de la moyenne les D_i par cette dernière expression.

On développe cette nouvelle formule, ce qui permet ensuite de simplifier cette expression.

La somme des survivants à chaque âge au numérateur correspond au cumul des années vécues par les personnes appartenant à cette cohorte fictive : chaque personne intervient autant de fois dans le calcul qu'elle cumule d'années. Par exemple, une personne qui décède entre 3 et 4 ans sera comptabilisée parmi les survivants à 1 an, 2 ans et 3 ans, soit trois fois. Elle décèdent en moyenne à 3,5 ans : c'est la raison pour laquelle on ajoute 0,5 an à ces trois années.

Remarque : on considère que les décès au-delà de 105 ans sont rares. Dès lors, on ferme arbitrairement la série à 105 ans. De ce fait, $S_{105} = 0$