Table des matières

Table des figures

Chapitre 1

Statistique desciptive à une dimension

La statistique descriptive c'est un ensemble de méthode et de technique qui consistent à classer, synthéser et présenter de données par des tableaux, des graphes et des paramètres afin de s'en faire une idée plus précise, de les rendre expoitables et d'en tirer le maximum d'informations.

1.1 Préparation d'une étude statistique

1.1.1 Population/Echontillon/Individu

La population statistique est l'ensemble sur lequel porte l'étude ou la prévision,et **l'échantillon** représente la fraction de cette population qui est réellement observée

ou étudiée.

- La notion **d'individu** est très large : les éléments d'un échantillon ou d'une population sont appelés généralement des individus, cependant cette notion peut être remplacé par plusieurs dénominations : unité statistique, sujet, objet, élément, observation, mesure, doses,. . ...ext.

1.1.2 Variable /Modalité/ Nature

Toute étude statistique s'intéresse à un ou plusieurs caractères d'une population. Ce sont les **variables statistiques**.Toute valeur pouvant être prise par une variable s'appelle **modalité** de cette variable.

Reconnaître la nature des variables est important puisque leur traitement en dépend.

Nature

Une variable pouvant prendre n'importe quelle valeur d'un intervelle est dite continue.

Par contre, une variable ne pouvant prendre que des valeurs isolées est dite discrètes. Dans ces deux cas la variable mesure une certaine quantité, on dit qu'elle est quantitative. Mais, il existe des variables qui ne sont pas quantitatives, on dit qu'elles sont qualitatives.

Exemple 1.1 - *Le poids des vaches d'un troupeau donné est une variable (quantita-*

tive) continue.

- Le nombre de vache, du troupeau précédent, atteints d'une certaine maladie est une variable (quantitative) discrète.

- La couleur des yeux des étudiants de 1ére année sciences vétérinaires est variable qualitative.

1.1.3 Collecte des données

Les données peuvent être, parfois, obtenues par recensement. C'est à dire que toutes les unités de la population sont observées. Mais la plus souvent ceci n'set pas possible ; soit le temps d'étude ne le permet pas, soit le coût de l'opération est trop élevé, soit c'est carrément impossible. Ainsi, par exemple, l'orsqu'on s'interesse à la durée de vie de lampes produites par une usine, il n'est pas

raisonnable de tester toutes les lampes produites. Autrement dit, l'échantillonage s'impose trés souvent. Il faut donc extraire de la population un sous ensemble qui sera observé et dont les résultats constituent l'échantillon.

1.2 Etude d'une sérié statistique : variable qualitative

Nous n'allons pas nous étendre beaucoup dans d'étude du variable qualitative. Nous nous contentrons, seulement, de voir comment :

- Calculer les fréquances relatives.

- Visualiser par des représentations graphiques

Sur un échantillon de taille n, supposons que l'on a obtenu k modalités différentes ($k \leq n$) puisque certaines observation peuvent se répéter. Nous notons par $(x_i)_{1\leq i\leq k}$ l'ensemble des observation deux à deux diférentes ainsi obtenu.

Il est alors utile de définir leurs fréquences afin de simplifier la présentation des données.

1.2.1 Distribution des fréquences

Définition 1.2 *- Nous appelons fréquence (absolue) de la donnée x_i,et on note n_i, le nombre de fois que cette donnée a été observée.*

-Nous appelons fréquence relative de la donnée x_i,et on note f_i le nombre donnée par :

$f_i = \frac{n_i}{N}$ *où n est le nombre total des données.*

Il est clair que $\sum_{i=1}^{k} n_i = N$ et $\sum_{i=1}^{k} f_i = 1$.

L'introduction des fréquences relatives permet de calculer des porcentages et de comparer des population différentes.

Exemple 1.3 *L'étude de l'état civil de 40 employés d'une entreprise a conduit aux résultats suivants :*

Numéro	*Etat civil*	*Numéro*	*Etat civil*
1	*Marié*	21	*Marié*
2	*Marié*	22	*Célibatoire*
3	*Célibataire*	23	*Marié*
4	*Divorcé*	24	*Veuve*
5	*Marié*	25	*Marié*
6	*Célibataire*	26	*Divorcé*
7	*Célibataire*	27	*Célibataire*
8	*Marié*	28	*Marié*
9	*Marié*	29	*Marié*
10	*Divorcé*	30	*Marié*
11	*Veuf*	31	*Divorcé*
12	*Marié*	32	*Marié*
13	*Célibataire*	33	*Célibataire*
14	*Marié*	34	*Célibataire*
15	*Marié*	35	*Marié*
16	*Marié*	36	*Marié*
17	*Marié*	37	*Divorcé*
18	*Célibataire*	38	*Divorcé*
19	*Célibataire*	39	*Marié*
20	*Célibataire*	40	*Célibataire*

La distribution des fréquences s'écrit :

$Modalités\ x_i$	$Fréquences\ n_i$	$Fréquences\ relatives\ f_i$
$Marié(e)$	20	0.5
$élibataire$	12	0.3
$Divorcé(e)$	6	0.15
$Veuf(ve)$	2	0.05
	$\sum n_i = 40$	$\sum f_i = 1$

Les modalités ont été classées, dans le tableau suivant l'ordre non croissant des fréquences.

L'intérét de résumer les données brutes par un tel tableau est clair, mais il est i,téressant de les visualiser par les graphes.

1.2.2 Représentation graphique

Diagramme en colonnes(tuyaux d'orgue)

A chaque modalité correspond un rectangle dont la hauteur est égale à la fréquence (absolue ou relative) associée à cette modalité. Les réctangles ont des largeurs égales et sont séparés les uns des autres par des distances égales. Pour l'exemple précédent, il vient

Diagrammes à secteurs (secteurs angulaires) : On utilise un disque que l'on partage en autant de secteurs que de modalités prises par notre variable. Les surfaces de ces secteurs sont proportionnelles aux fréquences des modalités. L'angle au centre d'un sécteur est égal à $f_i * 360$ où f_i représente la fréquence relative de la modalité par ce secteur.Pour notre exemple

1.3 Etude d'une sérié statistique : variable quantitative

Comme il a été déja expliqué, nous voulons rendre un ensemble de données expoitable.

Nous commencerons par les ranger par ordre non décroissant, dont l'utilité sera claire un peu plus loin.

Si la taille de l'échantillon est petite (infèrieure à 20 à titre indicatif), il est unitile de dresser le tableau de fréquences et de tracer un diagramme puisque cala n'ajoutera pas grand chose à la clarté de la présentation des données. Dans le cas où la taille de l'échantillon est grande, nous distingnons deux cas.

1.3.1 Données condensées (k est petit par rapport à N) ou Varible discrète

Dans ce cas, il devient utile de faire une étude du même genre que dans le cas qualitatif. Autrement dit, nous dressons le tableau de fréquences et nous traçons le diagramme en bâtons.

Distribution des fréquences

Chaque modalité est affecté de sa fréquence (absolue ou relative) qui est définie et notée exactement comme dans le cas qualitatif.

Exemple 1.4 *Une clinique a enregistré le nombre de frères et soeurs de chacun se patients atteints de la maladie contagieuse qu'est la varicelle duant l'année2000. Les données brutes obtenues sont :*

2	1	3	0	6	0	1	2	3	1
3	0	2	0	4	1	0	4	0	2
1	1	3	2	3	3	2	1	1	1
0	1	2	4	1	2	2	7	3	2
0	1	1	2	5	5	3	4	3	0
1	2	2	3	0	1	2	0	2	2

Ici $n = 60$ *et* $k = 8$ *(petit par rapport à n). La distribution des fréquences est*

données dans le tableau suivant :

x_i	n_i	f_i
0	11	0.183
1	15	0.250
2	16	0.267
3	10	0.167
4	4	0.067
5	2	0.033
6	1	0.017
7	1	0.016

1.3.2 Représentation graphique

La repréentation graphique est donnée par le diagramme en bâtons. Sur l'axe des abscisses sont représentées les modalités de la variable et sur l'axe des ordonnées sont représentées les fréquences. Partant de la valeur x_i, nous traçons un segment parallèlle à l'axe des ordonnées (bâton), de longeur égale à la valeur de la fréquence x_i.

Ceci nous amène à cumulet les fréquences et à introduire la notion suivante.

1.3.3 Fréquences cummulées

- Nous appellons fréquence (absolue) (resp relative) cummulée da la donnée x_i notée N_i (resp F_i),la somme de la fréquence (absolue)(resp relativee) de cette donnée et des fréquences (absolue) (resp relatives) des données qui lui sont infèrieures.

- La distribution des fréquences (absolue) (resp relative) cumulées est la donnée des différentes modalités affectées chacune de sa fréquences (absolue) (resp relative) cumulée.

Pour l'exemple précedent, la distribution des fréquences cumulées est donnée par le tableau suivant :

X_i	N_i	F_i
0	11	0.183
1	26	0.433
2	42	0.700
3	52	0.867
4	56	0.934
5	58	0.967
6	59	0.984
7	60	1

Nous déduisons, par exemple, que 86.7% des patients ont un nombre de frères et soeurs infèrieurs ou égal à 3.

La représentation graphique concernant les fréquences cumulées est donnée par le polygône des fréquences cumulées qui est construit en escalier. Nous plaçons les points dont les abcisses sont les modalités et les ordonnées sont leurs fréquences cumulées. Partant de chacun de ces points, nous traçons des segments horizontaux s'arrêtant en regard de la valeur de la modalité suivante (pour montrer que la variable

ne peut prendre aucune autre valeur entre ces deux modalités). puis nous joignons ces segments par des segments verticaux.

pour l'exemple précédent, nous obtenons le polygône des fréquences suivant :

L'utilisation des fréquences relatives, pour le polygône des fréquences cummulées, à l'avantage de n'avoir, en ordonnées, que des valeurs comprises entre 0 et 1 et permettre la comparaison entre des série différentes.

1.3.4 Variable continue ,données groupées (k est grand) :

Notamment lorsque la variable observée est continue les observations sont presque toutes différentes. Ce cas se produit aussi lorsque la variable est discrète mais que l'ensemble de ses modalités possède un grand effectif. Il est clair que dans ce cas une étude semblable à celle effectuée juste au dessus s'avère inutile.

Il est alors indiqué de regrouper les données en classes en prenant le soin que la précision que l'on perd ne nuise pas trop à l'étude. C'est sans doute qui se pose alors le choix de ces classes. Il n'y a pas de règle absolument stricte d'un tel choix mais certaines indications de nous guider.

Soient $([b_i, b_{i+1}[)_{0\leq i\leq m-1}$les m classes à définir.

- m varie en général entre 5 et 20 mais de préférence il se situe entre 6 et 12.

- m doit être choisi de sorte que les fréquences des classes ne soient pas toutes trés petites.

- Ces intervalles doivent contenir toutes les données. Pour cela la plus petite donnée

doit être plus grande que b_0 et la plus grande plus petite que b_m (chaque donnée se trouve alors dans une et une classe).

- Si possible prendre la même longueur pour toutes les classes, qu'on choisit de préférence un amplitude de 5, 10, 100, 1000 (pour faciliter les calculs que nous aurons à mener plus loin).

- Eviter la concentration des données aux bornes des intervalles car cela pourrait fausser certains calculs ultérieurs.

La formule de Strurges nous donne une indication sur le choix de m en fonction de n, elle est donnée par :

$$m = 1 + 3.322 \log_{10} n.$$

Le tableau suivant donne m (d'aprés la formule de sturges) pour les valeurs de n les plus utilisées dans la pratique.

n	m
$10 \prec n \leq 22$	5
$22 \prec n \leq 44$	6
$44\vert \prec n \leq 90$	7
$90\vert \prec n \leq 180$	8
$180 \prec n \leq 360$	9
$360\vert \prec n \leq 720$	10
$720 \prec n \leq 1000$	11

b_0 sera choisi le plus proche possible (en étant inférieure) de la plus petite des données.

Il reste alors à définir la longeur des intervalles. Une bonne idée sur ce nombre peut être donnée par le calcul de $\frac{E}{m}$ ou E est l'étendue, qui est égale à la différence entre la plus grande et la plus petite des observations.

Exemple 1.5 *Un club d'athlétisme a noté la taille en centimètres de tous ses membres. Les données, rangées en ordre croissant, sont :*

142.4	162.1	172.1	178.3	181.2	188.5
148.7	163.4	172.3	178.5	181.4	189.1
151.5	165.1	172.7	179.0	182.1	190.1
153.6	165.2	173.4	179.2	183.5	191.4
156.1	166.3	175.2	179.6	184.2	192.5
158.2	167.1	175.3	179.8	186.1	193.3
159.9	170.3	176.1	180.2	187.2	196.2
160.5	171.1	177.2	180.4	188.3	197.1
161.3	171.2	177.4	180.9	188.4	205.2

Ici $n = 54$ *et la règle de Sturges donne* $m = 7$.

L'étendue de la variable est $205.2 - 142.4 = 62.8$ *et* $\frac{62.8}{7} \backsimeq 9$. *On peut prendre la longueur des classes égale à 10 et la borne inférieure de la première classes égale à 140. Ceci nous conduit à la distribution des fréquences suivante (la fréquence d'une classe étant le nombre d'individus pour lesquels la variable prend ses valeurs dans cette classe).*

Classes	*Fréquences* n_i	*Fréquences relatives* f_i
$[140, 150[$	2	0.037
$[150, 160[$	5	0.093
$[160, 170[$	8	0.148
$[170, 180[$	18	0.333
$[180, 190[$	14	0.259
$[190, 200[$	6	0.111
$[200, 210[$	1	0.019
Totaux	54	1.000

1.3.5 Effectifs cumulés, fréquences relatives cumulées et diagramme intégral :

Les définitions de ces notions restent les même pour les variables statistiques continues comme pour les variables discrètes.

Nous avons, de ce fait :

L'effectif cumulé corresponddant à la valeur X_i *est le nombre des individus ayant un valeur inférieure ou égal* X_i.

Autrement dit, c'est la somme des effectifs qui se sont accumulés en atteignant cette valeur.

Ce qui s'écrit :

$$N_i = \sum_{p=1}^{i} n_p$$

De même,

La fréquence cumulée correspondant à la valeur X_i est la fréquence des individus ayant une valeur inférieure ou égal à X_i.

C'est à dire :

$$F_i = \sum_{p=1}^{i} f_p$$

Exemple 1.6 *Un club d'athlétisme*

1.3.6 Représentation graphique :

Diagramme différentiel

Dans le cas d'une variable statistique continue le diagramme différentiel s'appelle **histogramme**. Achaque classe est associé un rectangle dont la largeur est l'amplitude de la classse et dont la hauteur est le rapport de la fréquence sur l'amplitude ($\frac{n_i}{a_i} ou \frac{f_i}{a_i}$).

Diagramme intégral :

Avant de définir le diagramme intégral nous avons à apporter quelques précisions importantes.

Nous avons vu, lors de l'étude de la variable statistique discrète, que le diagramme intégral n'est autre que la courbe représentative de la fonction cumulative. Nous

avons constaté que c'est une courbe en escalier. Ce genre de croissance par sauts est du au fait qu'entre deux valeurs de la variable il n'y a aucune accumulation. Toute accumulation supplémentaire se fait lorsqu'on dépasse une valeur de la variable.

Dans le cas de la variable continue, cela ne se passe pas de la même manière . En effet, en allant d'une extrémité de la classe à l'autre on accumule les individus de cette classe. Il y a donc croissance de la fonction cumulative à l'intérieur même de la classe et non seulement aux extrémités. Comment se fait cette croissance ?

Nous avous supposé que la distribution est uniforme. Ce qui siginfie que la croissance est constante d'une extrémité de la classe à l'autre. Une croissance constante est synonyme de croissance linéaire ; c'est à dire suivant une droite.

La courbe cumulative est une ligne brisée faite de segment de droite joignants les points dont les cordonnées sont : en abscisse, les extrèmités des classes et en ordonnées les fréquences cumilées correspondantes à ces extrémités.

1.3.7 Caractéristiques de position centrale :

Nous avons déja donné des moyens de présenter des données quantitatives.

Mias cela ne suffit pas, nous avons aussi besoin d'avoir une idée sur l'ordre de grandeur des données et sur la position où elles semblent se rassembler. Cela est donné par les paramètres de tendence centrale.

La moyenne arithmétique

C'est un paramètre trés utilisé par tous. Pensez à vos moyennes durant vos années d'études et qui ont servi à juger vos aptitudes !. La moyenne est la valeur centrale autour de laquelle se répartissent les données. La moyenne d'une série $(x_i)_{1\leq i\leq n}$,noté par $\overline{x}$, se calcule par :

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

pour les données condensées par :

$$\overline{x} = \frac{\sum_{i=1}^{n} n_i x_i}{N} = \sum_{i=1}^{n} f_i x_i \text{ où } (x_1, x_2, ...x_k) \text{ sont les différentes modalités.}$$

Et pour les données groupées par :

$$\frac{\sum_{i=1}^{n} n_i c_i}{N} = \sum_{i=1}^{n} f_i c_i \text{ } où \text{ } (c_1, c_2, ...c_k) \text{sont les différents milieux des classes.}$$

Exemple 1.7 *Calculer la moyenne de la variable statistique dont le tableau de dis-*

tribtion de fréquences est donné ci dessous :

modalités x_i	*Fréquences* n_i	$n_i x_i$
12	3	36
14	6	84
16	10	160
18	16	288
20	11	220
25	6	150
29	3	87
$Totaux$	55	1023

$$\overline{x} = \frac{1023}{55} = 18.64$$

Mode

C'est la valeur qui se répète le plus dans l'échantillon, Autrement dit pour les données condensées, c'est la modalité qui a la plus frande fréquence. Quant au cas des données groupées en classes, on détermine d'abord la classe modale qui est la classe de plus grande fréquence puis le mode, noté M_0, Il se calcul sur l'histogramme par :

$$M_0 = b_{m_0} + (\frac{\triangle_1}{\triangle_1 + \triangle_2})L_{m_0}$$

où

b_{m_0} : est la borne inférieure de la classe modale ;

$\triangle_1$: est la différence entre la fréquence de la classe modale et la fréquence de la classe précédente ;

$\triangle_2$: est la différence entre la fréquence de la classe modale et la fréquence de la classe qui suit ;

L_{m_0} : est la longeur de la classe modale.

Remarque 1.8 *Le mode peut ne pas être unique. Certaines distribution peuvent présenter plusieurs valeurs dont les effectifs sont égaux et qui sont les plus grands de la distribution.*

Exemple 1.9 *-Données condensées :*

Modaliés	*Fréquences*
12	3
14	6
16	10
18	16
20	11
25	6
29	3

$M_0 = 18$ *car c'est la modalité qui a la plus grande fréquence.*

Exemple 1.10 - *Données groupées :*

Classes	*fréquences*
$[30, 40[$	4
$[40, 50[$	7
$[50, 60[$	11
$[60, 70[$	12
$[70, 80[$	8
$[80, 90[$	5

La classe modale est $[60, 70[$ *car elle a la plus grande fréquence.*

$b_{m_0} = 60,\ \triangle_1 = 1, \triangle_2 = 4, L_{m_0} = 10,\ d'où\ M_0 = 60 + \frac{1}{5} * 10 = 62$

La médiane C'est la valeur de l variable statistique qui partage la population en deux populations d'éffectifs égaux. Autrement dit, les individus appartenant à la première moitié de la population ont des valeurs infèrieures à la médiane. les individus appartenant à la dexième moitié lui ont des valeurs supèrieures.

Pour déterminer la médiane nous avons deux moyens :

- Utiliser la colonne des effectifs cumulés (resp fréquences absolues cumulées) : Il suffit de situer où se trouve $\frac{N}{2}$.

Reprenons l'exemple 1 : où se trouve $\frac{N}{2}$? $\frac{N}{2} = \frac{150}{2} = 75$

Nous le trouvons encadré par deux valeurs de l'effectif cumulé : N_{i-1} et N_i

Ces effectifs cumulés sont écrits sur les lignes. La valeur de la variable inscrite dans la case comprise entre ces deux lignes est la médiane.

- Utiliser la colonne des fréquenes relatives cumulées : Quand on se trouve réfère à la définition de la médiane M alors la fonction cumilative donne :

$$F(M) = 0.5$$

Ainsi, nous devons chercher les valeurs F_{i-1} et F_i qui encadrent 0.5

Nous pouvons aussi nous servir du diagramme intégral pour situer la médiane. Il suffit de déterminer l'abscisse du point qui a 0.5 comme ordonnée.

Dans le cas des données groupées il faut d'abord situer la classe de la médiane.

nous savons que la médiane est la valeur de la variable statistique qui partage la population en deux effectifs égaux. C'est la valeur M telle que $F(M) = 0.5$ où F est la fonction cumulative.

Ainsi, à la premiére extrémité de la classe où se trouve la médiane nous n'avons pas encore atteint 0.5 comme fréquence cumulée et à la deuxiéme extrémité nous auront dépassé 0.5.

La définition de la classe de la médiane est donc :

C'est la classe $[\mathbf{e}_{i-1}, e_i[$ telleque

$$F(\mathbf{e}_{i-1}) \prec 0.5 \; et \; F(\mathbf{e}_{i-1}) \succ 0.5$$

Une fois que la classe de la médiane est déterminée, nous pouvons calculer la médiane par interpolaion linéaire alors :

$$M = b_{m_0} + (\frac{0.5 - F_{_{m_e-1}})}{f_{m_e}})L_{m_e}$$

où

b_{m_0} : est la borne inférieure de la classe médiane ;

$F_{_{m_e-1}}$: est la fréquence relative cumulée de la classe qui précède la classe médiane.

f_{m_e} : est la fréquence relative de la classe médiane ;

L_{m_0} : est la longeur de la classe médiane.

Exemple 1.11 *1) Données condensées utilisées en exemple pour le calcul du mode :*

x_i	n_i	f_i	F_i
12	3	0.055	0.055
14	6	0.109	0.164
16	10	0.192	0.346
18	16	0.291	$0.637 \succ 0.5$
20	11	0.200	
25	6	0.109	
29	3	0.054	
Totaux	55	1.00	

$M = 18$ *puisque c'est la première valeur dont la fréquence relative cumulée dépasse 0.5.*

2) Données groupées suivantes :

Classes	*fréquences*	*fréquences relatives*	*fréquences relatives cumulées*
$[30, 40[$	4	0.085	0.085
$[40, 50[$	7	0.149	0.234
$[50, 60[$	11	0.234	0.468
$[60, 70[$	12	0.255	$0.723 \succ 0.5$
$[70, 80[$	8	0.170	
$[80, 90[$	5	0.107	

La classe médiane est $[60, 70[$ car c'est dans cette classe que F_i dépasse, pour la première fois la valeur 0.5, la médiane est :

$M = 60 + \frac{0.5-0.468}{0.255} * 10 = 61.3.$

Les quartiles Comme on a définit la médiane pour répartir la population en moitiés on peut définir des paramètres qui la répartissent en quarts. On les appelle les **quartilles**. Ils sont au nombre de trois.

Au premier quartile, qu'on note Q_1,s'accumule le quart de la population.

Au deuxième quartile, qu'on note Q_2,s'accumule les deux quarts de la population (et c'est donc la médiane).

Au troisième quartile, qu'on note Q_3,s'accumule les trois quarts de la population.

Pour déterminer les quartiles on procède de la même manière que pour la médiane.

Il suffit, pour le premier quartile, de situer où se trouve $\frac{N}{4}$ sur la colonne des effectifs cumulés (ou 0.25 sur la colonne des fréquences relatives cumulées).

Pour le troisième quartile, nous avons à situer $\frac{3N}{4}$ comme effectif cumulé (ou 0.75

comme fréquence relative cumulée).

Dans le cas des données groupées la procédure est la même que pour la médiane.

il s'agit d'abord d'identifier les classes où se trouve les quartiles et de les calculer ensuite par interpolation linéaire.

Comme le premier quartile Q_1cumule le quart de la population alors il se trouve dans la classe $[\mathbf{e}_{i-1}, e_i[$ qui vérifie les conditions :

$$F(\mathbf{e}_{i-1}) \prec 0.25 \; et \; F(\mathbf{e}_{i-1}) \succ 0.25$$

Pour calculer Q_1 nous allons utiliser la formule de l'interpolaion linéaire

$$Q_1 = b_{m_0} + (\frac{0.25 - F_{_{m_e}-1)}}{f_{m_e}})L_{m_e}$$

Le troisiéme quartile Q_3 cumule les trois quarts de la population et donc sa classe est celle qui vérifie

$$F(\mathbf{e}_{i-1}) \prec 0.75 \; et \; F(\mathbf{e}_{i-1}) \succ 0.75$$

La valeur de Q_3 et donc

$$Q_3 = b_{m_0} + (\frac{0.75 - F_{_{m_e}-1)}}{f_{m_e}})L_{m_e}$$

1.3.8 Caractéristiques de dispersion :

Si les paramètres de tendances centrale donnent une idée sur les valeurs centrales autour desquelles se répartissent les données, il nous faut plus que cela puisque des sériés trés différentes peuvent avoir la même moyenne. Ainsi pour avoir une idée plus précise sur les données il est intéressant de connaitre la dispersion de ces données autour de la moyenne.

Par exemple les deux sériés -1, 0, 1 et -100, 0, 100 ont toutes deux la même moyenne mais dans le premier exemple les données se reppochent beaucoup plus de la moyenne. Si par exemple, en plus de la connaissance de la moyenne on sait que les écarts des données à la moyenne sont petits on peut déduire que les données se concenant autour de la moyenne.

Variance

La variance de l'échantillon, notée s^2, est égale à :

$$var(x) = s^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{N}.$$

- En données condensées :

$$s^2 = \sum_{i=1}^{n} \frac{n_i(x_i - \overline{x})^2}{N}.$$

- En données groupées :

$$s^2 = \sum_{i=1}^{n} \frac{n_i(c_i - \overline{x})^2}{N}$$

ou c_i est le milieu de la $i^{éme}$ classe et n_i sa fréquence.

Remarque 1.12 *Pour le calcul, il est plus souvent pratique d'utiliser l'expression équivalente de la variance et qui est donnée par :*

$$s^2 = \sum_{i=1}^{n} \frac{n_i x_i^2}{N} - \overline{x}^2$$

1.3.9 Ecart type

La variance est un paramètre de dispersion trés utilisé mais son unité est le carré des unités des meusures, c'est pourquoi sa racine carré est introduite et s'appelle l'écart type, on note s

1.3.10 Etendue

C'est la différence entre la plus grande donnée et la plus petite est aussi un paramétre de dispersion .Seulement elle donne une idée, certes rapide, mais grossiére de la dispersion des données puis qu'elle ne tient compte que des données extrémes.

pour l'exemple pécédent : E=29-12=17.

1.3.11 Ecart interquartile

Il est donné par la différence entre les quartiles d'ordre 3 et d'ordre 1. Cet intervelle contient la moitié des observation qui se situent au centre de la série.

pour l'exemple precident

$E = Q_3 - Q_1 = 20 - 16$