Chap. : I) Introduction à la Statistique

1. Objet de la Statistique

"***Science qui a des procès pour recueillir, organiser, classer, présenter et interpréter les données***"

La statistique nous donne les techniques pour qu'on puisse avoir l'information sur les données, lesquelles sont souvent insuffisantes, vu qu'elles nous donnent l'information utile sur un problème qu'on est en train d'étudier, mais elles ne mettent pas en relief des aspects très importants.

L'objectif de la Statistique, c'est d'extraire l'information des données afin qu'on puisse mieux comprendre les situations qu'elles représentent.

2. Population et Echantillon

La notion d'ensemble est fondamentale dans la Statistique, concept pour lequel on utilise indifféremment les termes **Population** ou univers.

Collection d'unités individuelles, qui peuvent être des personnes ou des résultats expérimentaux, avec une ou plusieurs caractéristiques, que l'on prétend étudier.

3. Recensement et Sondage

Normalement le mot **recensement** est lié au comptage officiel périodique des individus d'un Pays ou d'une partie d'un Pays. Pourtant, ce mot renferme en soi-même un éventail plus vaste de situations. Ainsi, on peut dire que recensement c'est:

L'étude scientifique d'un univers de personnes, d'institutions ou d'objets physiques pour acquérir des connaissances, observant tous ses éléments, et faire des jugements quantitatifs sur les caractéristiques importantes de cet univers-là.

Pour la plupart des personnes le mot **recensement** est lié à l'énumération des éléments de la population d'un Pays. Au Maroc, de dix en dix ans, on réalise le Recensement Général de la Population. Le dernier a eu lieu en 2004

4. Statistique Descriptive et Statistique Inductive

D'après ce que l'on a dit avant, dans une **analyse statistique** il faut faire la distinction de deux phases:

La **première phase** où l'on cherche décrire et étudier l'échantillon:

Statistique Descriptive

Et la **deuxième phase** où l'on essaie d'arriver à des conclusions pour la population:

Statistique **Inductive**

5. Exemple

Le gérant d'une usine de détergents veut commercialiser un nouveau produit pour faire la vaisselle, alors il demande à une entreprise spécialisée dans les études du marché d' «estimer» le pourcentage de potentiels acheteurs de ce produit.

Population: ensemble de tous les agrégés familiaux du Pays.

Echantillon: ensemble de quelques agrégés familiaux, enquérais par l'entreprise.

Problème: on veut, à partir du pourcentage de réponses affirmatives données par les individus enquérais à propos de l'achat du nouveau produit, obtenir une estimation du nombre d'acheteurs chez la Population.

Chap. II) : Données, tableaux et graphiques

1. Types de données

On peut classifier les données qui forment l'Echantillon, ou les données échantillonnées, en deux types fondamentaux: **Données qualitatives et données quantitatives**

Données qualitatives

Données qualitatives	***Elles représentent l'information qui identifie une certaine qualité, catégorie ou caractéristique, ce qui n'est pas susceptible d'être mesuré, mais qui peut être classifié, ayant en soi-même plusieurs modalités.***

Données quantitatives

Données quantitatives	***Elles représentent l'information des caractéristiques susceptibles d'être mesurées, en se présentant avec de différentes intensités, qui peuvent être discrète (pas continue) - données discrètes, ou continue - données continues.***

2. Représentation graphique de données

Données discrètes

Données discrètes	***Ces données ne peuvent que prendre un nombre fini ou infini nombrable de valeurs distinctes, en présentant plusieurs valeurs qui se répètent – c'est le cas, par exemple, du nombre d'enfants d'une famille ou du nombre d'accidents, chaque jour, dans un carrefour quelconque.***

Données continues

Données continues	***Dans le cas d'une variable continue, celle-ci peut prendre toutes les valeurs numériques, entières ou non, comprises dans son intervalle de variation – on a par exemple le poids, la taille, etc.***

Chap. III) : Les caractères de position

I) le Mode (Moyenne de fréquence)

Le **mode** est la valeur du caractère statistique qui apparaît le plus fréquemment.
Exemple 1: note des élèves

notes x_i	**5**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**16**	**Total**
effectifs n_i	**1**	**1**	**2**	**4**	**3**	**2**	**1**	**1**	**1**	**16**

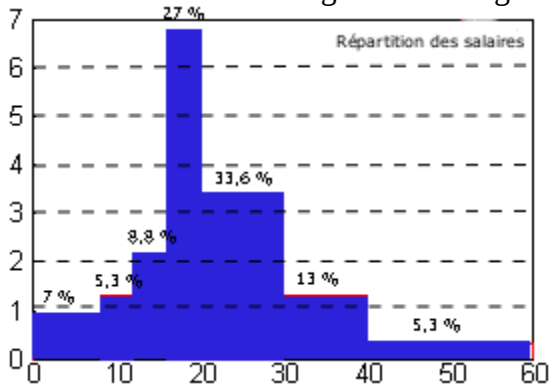
Le mode est 10.
Exemple 2: note des élèves

notes x_i	**5**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**16**	**Total**
effectifs n_i	**1**	**1**	**4**	**2**	**2**	**4**	**1**	**1**	**1**	**16**

Cette série est dite série bimodale car on voit apparaître deux modes: 9 et 12.
Dans le cas d'une variable continue, on peut entendre parler de classe modale qui serait la classe de plus grand effectif. Mais il faut se méfier de cette notion car, plus la classe est de grande amplitude, plus son effectif est important sans pour autant que cela soit significatif. Cette notion de classe modale définie par les effectifs de la classe n'a de sens que si les classes ont même amplitude. Si les amplitudes sont différentes, il faut aller chercher sur l'histogramme la classe associée au rectangle de plus grande hauteur.
Exemple : l'exemple développé dans Statistiques élémentaires continues conduit au tableau suivant: Répartition des revenus annuels en milliers d'Euros dans une population de 4370 personnes.

Salaires	**entre 0 (inclus) et 8 exclus**	**entre 8 (inclus) et 12 exclus**	**entre 12 (inclus) et 16 exclus**	**entre 16 (inclus) et 20 exclus**	**entre 20 (inclus) et 30 exclus**	**entre 30 (inclus) et 40 exclus**	**entre 40 (inclus) et 60 exclus**	**Total**
Effectifs	**306**	**231**	**385**	**1180**	**1468**	**568**	**232**	**4370**

L'observation de ce tableau laisse penser que la classe modale serait la classe [20; 30[. Mais une observation de l'histogramme corrige cette idée fausse:



La classe modale est la classe [16; 20[

II) la médiane (Moyenne de position)

La médiane est la valeur du caractère statistique qui coupe la population en deux populations de taille égale.

III) La Moyenne arithmétique

Cas de la série statistique discrète triée mais non regroupée

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

L'article Statistiques élémentaires discrètes explique cette formule.

Cas de la série statistique discrète regroupée

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

L'article Statistiques élémentaires discrètes explique cette formule.

Cas de la série continue

$$*\overline{x} = \frac{\sum_{i=1}^{N} n_i m_i}{\sum_{i=1}^{N} n_i} = \sum_{i=1}^{N} f_i m_i$$

L'article Statistiques élémentaires continues explique cette formule.

Stabilité par transformation affine

La moyenne est stable par transformation affine, c'est-à-dire : si yi = axi + b, si $\overline{x}$ est la moyenne de la série x alors la moyenne de la série y est $\overline{y} = a\overline{x} + b$.

Cette propriété est utile pour changer d'unité: si on connaît une moyenne de température en degré Fahrenheit, il est inutile de convertir toutes les valeurs en degrés Celsius pour calculer la moyenne en degrés Celsius, il suffit de ne convertir que la moyenne.

Il est aussi intéressant, pour limiter la taille des nombres, de partir d'un moyenne estimée et de calculer la moyenne des di = xi − Mestim.. Alors $\overline{x} = M_{estim.} + \overline{d}$

Découpage en sous-population

Si la population est découpée en deux sous-populations P_1 et P_2 de tailles n_1 et n_2, si la moyenne du caractère statistique pour la population P_1 est $\overline{x_1}$ et la moyenne pour la population P_2 est $\overline{x_2}$

alors la moyenne pour la population P est $\overline{x} = \frac{n_1\overline{x_1} + n_2\overline{x_2}}{n_1 + n_2}$.

Sensibilité aux valeurs extrêmes

La moyenne est sensible aux valeurs extrêmes ou aberrantes.

Exemple: dans une entreprise, 9 salariés sont payés 2000 Euros mensuels. Le patron se paie 22000 Euros mensuels.

Effectuer la moyenne dans ces conditions conduit à une valeur non représentative:

$\overline{x} = \frac{9 \times 2000 + 22000}{10} = 4000$ Euros.

Pour éviter ce genre de piège, il arrive que l'on tronque volontairement la population et qu'on élimine 10% des valeurs les plus basses et 10% des valeurs les plus hautes.

Les caractéristiques de concentration

L'étude de la concentration a pour objet de mettre en évidence et de mesurer les inégalités de répartition. Les domaines d'application sont nombreux en économie : concentration des salaires, concentration des revenus ou du patrimoine, concentration de l'emploi, concentration des vacances,… Ainsi, nous voulons savoir si des inégalités de répartition existent en ce qui concerne les nuitées personnelles à partir de l'enquête SDT de 1999.

Soit le tableau de fréquence suivant :

Nombre de nuitées pour motif personnel

Nombre de nuitées	Effectifs	Fréquences	ci
[0-1[	2627	26.3%	0
[1-5[	806	8.1%	2.5
[5-10[	1059	10.6%	7
[10-15[	1083	10.8%	12
[15-20[	818	8.2%	17
[20-30[	1302	13.0%	24.5
[30-40[	809	8.1%	34.5
[40-60[	847	8.5%	49.5
[60-100[	497	5.0%	79.5
[100-197[	152	1.5%	148
Total	10000	100.0%	

Source : SDT 1999

La population statistique correspond au panel de la Sofres 10000 personnes représentatives de la population française interrogées sur leur déplacements de l'année 1999. La variable étudiée est le nombre de nuitées personnelles, il s'agit d'une variable quantitative continue.
A partir de ce tableau, on peut s'interroger sur la répartition de la masse totale des nuitées personnelles. Quelle part de cette masse totale se partagent ceux qui sont partis plus de 60 nuitées dans l'année ?
Pour répondre à cette question, on va s'intéresser à la masse de nuitées, aussi appelée valeur globale. La valeur globale associée au couple (x_i,n_i) est le produit $n_i.x_i$. La valeur globale relative du même couple est $q_i=\frac{n_i.x_i}{\sum_{i=1}^{k} n_i.x_i}$, $\sum_{i=1}^{k} n_i.x_i$ représentant la masse totale de la variable étudiée. La valeur globale relative cumulée sera $Q_i=q_1+q_2+...+q_i=\sum_{j=1}^{i} q_j$

Evidemment, dans le cas d'étude d'une variable continue, on remplace xi par le centre de classe ci dans les formules. Si l'on complète notre tableau, nous aboutissons à :

Nombre de nuitées	Effectifs	Fréquences	ci	ni.ci	qi	ei	Fi	Qi
[0-1[	2627	26.3%	0	0	0.0%	0	0	0
[1-5[	806	8.1%	2.5	2015	1.0%	1	26%	0.0%
[5-10[	1059	10.6%	7	7415	3.7%	5	34%	1.0%
[10-15[	1083	10.8%	12	12998	6.5%	10	45%	4.7%
[15-20[	818	8.2%	17	13906	6.9%	15	56%	11.2%
[20-30[	1302	13.0%	24.5	31891	15.9%	20	64%	18.2%
[30-40[	809	8.1%	34.5	27901	13.9%	30	77%	34.1%
[40-60[	847	8.5%	49.5	41929	21.0%	40	85%	48.0%
[60-100[	497	5.0%	79.5	39541	19.8%	60	94%	69.0%
[100-197[	152	1.5%	148	22488	11.2%	100	98%	88.8%
Total	10000	100.0%		200083	100.0%	197	100%	100.0%

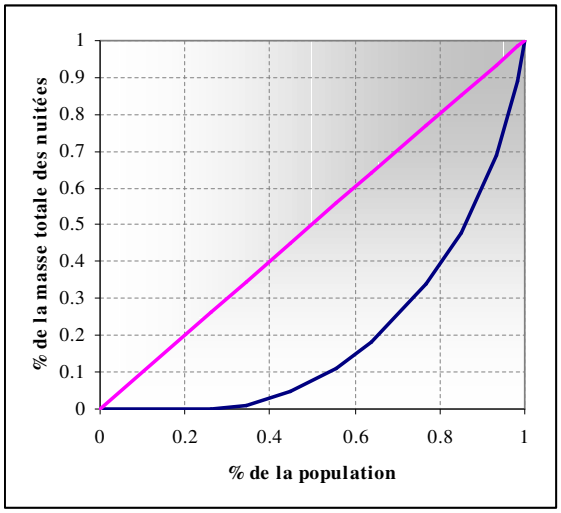
On s'aperçoit, par exemple, que :

- les 26% de la population qui ne partent jamais pour motif personnel représentent logiquement 0% de la masse des nuitées,
- les 1,5 % de la population qui partent plus de 100 nuitées par an représentent 11% des nuitées personnelles.

La valeur de la variable qui est associée à Qi=50% s'appelle la médiale. On la détermine de la même manière que la médiane.
Dans notre exemple, nous devons pratiquer une interpolation linéaire :
Me=10+5(50-45)/(56-15)=10,6

Mle=40+20(50-48)/(69-48)=42.
Son interprétation est la suivante : les personnes dont le nombre de nuitées personnelles annuelles est inférieur à 42 se partagent la moitié de la masse de nuitées effectuées en 1999.
On parle de concentration quand on observe de fortes disproportions entre la part en terme d'individus et la part en terme de masse totale de la variable.
L'écart entre la médiale et la médiane apporte une première information sur la concentration : la valeur est d'autant plus grande (par rapport à l'étendue) que la concentration est forte (la médiale est toujours supérieure ou égale à la médiane).
Dans notre exemple, l'écart est très important puisqu'il atteint 31,4, ce qui correspond à 16% de l'étendue.
Une autre manière de visualiser la concentration consiste à construire la courbe de concentration que l'on obtient en faisant correspondre F(x) (fréquences cumulées des individus) en abscisse à Q(x) en ordonnée (fréquences cumulées de la masse totale).
La courbe s'inscrit dans un carré de taille 1 et plus la courbe s'éloigne de la diagonale de ce carré, plus la distribution est inégalement répartie.
Pour mesurer la concentration, on peut donc aussi calculer la surface comprise entre la courbe et cette bissectrice.

Courbe de concentration



L'indice de concentration ou indice de Gini est le rapport de l'aire de la surface de concentration et de l'aire du triangle. Il existe différentes méthodes pour le calculer. La méthode des trapèzes nous donne la formule suivante :

$i_g = 1 - \frac{1}{100^2}\sum_{i=1}^{k} f_i(Q_i + Q_{i+1})$ pour Qi et fi exprimés en pourcentages et

$i_g = 1 - \sum_{i=1}^{k} f_i(Q_i + Q_{i+1})$ pour Qi et fi non exprimés en %.

Pour le calculer, il suffit de rajouter une colonne au tableau et d'en faire la somme :

Nombre de nuitées	Effectifs	Fréquences	ci	ni.ci	qi	ei	Fi	Qi	$(Q_i+Q_{i+1}).f_i$
[0-1[	2627	26.3%	0	0	0.0%	0	0	0	(0+0)x26.3=0
[1-5[	806	8.1%	2.5	2015	1.0%	1	26%	0.0%	(0+1)x8.1=8.1
[5-10[	1059	10.6%	7	7415	3.7%	5	34%	1.0%	(1+4.7)x10.6=61
[10-15[	1083	10.8%	12	12998	6.5%	10	45%	4.7%	172
[15-20[	818	8.2%	17	13906	6.9%	15	56%	11.2%	240
[20-30[	1302	13.0%	24.5	31891	15.9%	20	64%	18.2%	680
[30-40[	809	8.1%	34.5	27901	13.9%	30	77%	34.1%	664
[40-60[	847	8.5%	49.5	41929	21.0%	40	85%	48.0%	991
[60-100[	497	5.0%	79.5	39541	19.8%	60	94%	69.0%	785
[100-197[	152	1.5%	148	22488	11.2%	100	98%	88.8%	287
Total	10000	100.0%		200083	100.0%	197	100%	100.0%	3889

D’où $i_g = 1 - \frac{38,888}{100^2} = 0,61$, ce qui signifie que 61 % de la courbe de concentration correspond à 61 % de la surface du triangle.

Chap. IV) : Les indices

En général, on résume chaque série par sa moyenne, son écart-type ou encore des tracés graphiques.

Mais si on dispose de plusieurs séries de grandeurs élémentaires (en général des séries de prix ou des séries de quantités de biens), comment fait-on pour les résumer en une seule grandeur « synthétique »?

Exemple : Prix en DHs (par kg)

	1999	2000	2001	2002	2003
Oranges (par Kg)	2	2.3	2.5	2.7	2.9
Télévisions (unité)	3000	3200	2800	2750	2900
Eléctricité (en KW par heure)	3	3.5	2.2	3.2	3.4

Comment fait-on pour calculer un prix moyen 'agrégé' pour ces trois biens? Est-ce qu'on peut observer son évolution ?
Il existe des indicateurs « simples » (appelés indices élémentaires) qui sont un moyen synthétique permettant de faciliter la lecture de l'évolution des séries dans l'espace et le temps.
A partir de ces indicateurs on peut montrer qu'on peut obtenir des indicateurs (appelés indices synthétiques) qui peuvent décrire l'évolution d'une grandeur agrégée à partir de grandeurs simples.
1/ Les indices élémentaires

1.1/ Présentation

Définition : un indice élémentaire est un rapport entre deux valeurs d'une série à deux dates ou deux espaces différents.
Il est représenté par :

$$I_{c/r} = 100.\frac{V_c}{V_r}$$

avec V_c valeur observée à une date donnée ou dans un espace donné ;
V_r valeur de référence (à la date r ou dans un espace r).
En général,l'indice élémentaire relatif au temps s'écrie :

$$I_{t/t'} = 100.\frac{V_t}{V_{t'}}$$

L'indice est exprimé en pourcentage (d'où la multiplication par 100)

Exemple : PIB/habitant

Entre 2 dates : 1995 et 2000	Entre deux régions européennes : Bavière et Ile de France
$I_{2000/95} = 120$	$I_{Bav/IdF} = 130$

En 2000, le PIB/habitant était de 20% plus élevé qu'en 1995 ;
En Bavière (Allemagne), le PIB/tête est de 30% plus élevé qu'en Ile de France.

1.2/ Propriétés des indices élémentaires :

1.2.1/ La circularité

Un indice à la date t exprimé par rapport à une année de référence t', peut être décomposé en plusieurs indices élémentaires à des dates successives (ou à des dates intermédiaires) de la façon suivante :

$$I_{t/t'} = 100.\left[\frac{I_{t/t-1}}{100}.\frac{I_{t-1/t-2}}{100}....\frac{I_{t'+1/t'}}{100}\right]$$

Démonstration :

$I_{t/t'} = 100.\frac{V_t}{V_{t'}} = 100\left[\frac{V_t}{V_{t-1}}.\frac{V_{t-1}}{V_{t-2}}...\frac{V_{t'+1}}{V_{t'}}\right] = 100.\left[\frac{I_{t/t-1}}{100}.\frac{I_{t-1/t-2}}{100}....\frac{I_{t'+1/t'}}{100}\right]$ Donc, dès lors que nous observons des indices intermédiaires sur la période considérée nous pouvons en déduire un indice global. On dit alors que les indices élémentaires sont enchaînables.

Cas particuliers :

-Utile pour changement de base :

Soit deux indices $I_{t/0}$ et $I_{t'/0}$, exprimés en base 0 (année 0). On veut exprimer l'indice à la date t par rapport à la date t'. Donc, on veut effectuer un changement de base. Comment procède-t-on ? A l'aide de la formule générale:

$$I_{t/0} = 100.\left[\frac{I_{t/t'}}{100}.\frac{I_{t'/0}}{100}\right]$$

D'où :

$$I_{t/t'} = 100.\left[\frac{I_{t/0}}{I_{t'/0}}\right]$$

1.2.2- La réversibilité

Quand on inverse le rôle de la base de référence et celle de la valeur courante, l'indice élémentaire s'inverse à 10^4 près :

$$I_{r/c} = 10^4.\frac{1}{I_{c/r}}$$

Démonstration :

$$I_{r/c}.I_{c/r} = 100.\frac{V_r}{V_c}.100\frac{V_c}{V_r} = 10^4$$

3.3/ L'indice de Laspeyres

Quand on fait la moyenne arithmétique pondérée des indices élémentaires on obtient l'indice suivant.

Notons le :

$$L^{p}_{t/0} = \frac{\sum R_{i,0}.I_{i,t/0}(p)}{\sum_i R_{i,0}} = 100.\frac{\sum R_{i,0}\left(p_{i,t}/p_{i,0}\right)}{\sum_i R_{i,0}}$$

Cet indice exprime, simplement, une évolution moyenne des prix élémentaires. C'est un indice 'moyen' des prix Il accorde un poids (la recette) dans les recettes totales à chaque prix relatif. Mais, un indice des prix par rapport à une date 0 de référence, ne doit contenir qu'une variation des prix et non des quantités. D'où le maintien d'un poids à la date 0.

Mais les Recettes à la date t, peuvent s'écrire :

D'où $$L^{p}_{t/0} = 100.\frac{\sum (p_{i,0}.q_{i,0})\left(p_{i,t}/p_{i,0}\right)}{\sum_i (p_{i,0}.q_{i,0})} = 100.\frac{\sum (p_{i,t}.q_{i,0})}{\sum_i (p_{i,0}.q_{i,0})}$$

Cet indice des prix 'synthétique' est celui de Laspeyres des prix.

On remarquera que :

1/ seuls les prix varient dans cette relation

2/ L'indice de Laspeyres se réduit à un indice élémentaire quand il existe seulement un seul bien puisque :

De la même façon, on peut obtenir un indice de Laspeyres des quantités puisqu'il s'agit de la moyenne pondérée des indices de quantités :

D'où $$L^{q}_{t/0} = 100.\frac{\sum (p_{i,0}.q_{i,0})\left(q_{i,t}/q_{i,0}\right)}{\sum_i (p_{i,0}.q_{i,0})} = 100.\frac{\sum (p_{i,0}.q_{i,t})}{\sum_i (p_{i,0}.q_{i,0})}$$

Ici, l'idée serait de ne faire varier que les quantités élémentaires pour obtenir un indice des quantités synthétiques. Pour cela, les prix sont exprimés par rapport à la date de référence uniquement.

2.3/ L'indice de Paâsche

On peut aussi procéder à une moyenne harmonique des indices élémentaires pour définir un autre indice synthétique. Notons cet indice :

$$P^{p}_{t/0} = \left[\frac{\sum R_{i,t}\left[I_{i,t}(p)\right]^{-1}}{\sum R_{i,t}}\right]^{-1} = 100.\frac{\sum R_{i,t}}{\sum_i \frac{R_{i,t}}{\left(p_{i,t}/p_{i,0}\right)}}$$

Les caractéristiques de dispersion et de forme

Les indicateurs de tendance centrale ne rendent compte que d'une partie de la distribution statistique d'une population. Ainsi, une valeur moyenne de 9 à un examen n'a pas la même signification si les notes s'échelonnent de 6 à 16 ou si elles varient de 2 à 19... D'où l'intérêt d'étudier la dispersion des données autour de la tendance centrale. Cette dispersion peut être caractérisée par un certain nombre d'indicateurs.

L'étendue

L'étendue est la **différence entre la valeur la plus haute et la valeur la plus basse** de la distribution. Ainsi, dans une classe où les notes vont de 6 à 13, l'étendue est de 13-6=7. Lorsque les notes s'échelonnent de 2 à 19, l'étendue est de 17.
L'étendue ne fournit pas une bonne représentation de la dispersion, en effet cet indicateur dépend trop des valeurs extrêmes de la distribution.

L'écart interquantiles

Lorsque l'on considère un intervalle interquantiles, on se rapproche de la tendance centrale, on écarte donc les valeurs extrêmes pour s'intéresser à une certaine part de la population dispersée autour de la tendance centrale.
L'intervalle interquantiles pourra être un intervalle interquartile (ensemble des valeurs situées entre le premier quartile Q1 et le dernier quartile Q3), un intervalle interdéciles (ensemble des valeurs situées entre le premier décile D1 et le dernier décile D9), un intervalle intercentiles (ensemble des valeurs situées entre le premier centile C1 et le dernier centile C99).
L'écart interquantiles peut se définir comme la différence entre le dernier et le premier quantile d'une même catégorie. Ainsi, l'écart interquartile (Q3-Q1) comprend 50 % des observations, l'écart interdéciles comprend 80% des observations et l'écart intercentiles en comprend 98%.
Les caractéristiques de dispersion de cet indicateur sont assez imparfaites. Elles ne se prêtent que très mal aux calculs algébriques.
Sa détermination est d'autant plus imprécise que l'on se rapproche de valeurs extrêmes : il ne dépend en fait que du rang des observations, et non de leur valeur ou de leur écart relatif.

L'écart absolu moyen

Pour définir une mesure de la dispersion qui prennent en compte toutes les données, une solution consiste à calculer les écarts entre chaque valeur observée et une valeur centrale et d'en faire la moyenne arithmétique. Le problème qui se pose lorsque l'on essaie de calculer un écart moyen par rapport à la moyenne, c'est que les écarts positifs et les écarts négatifs s'annulent par définition même de la moyenne arithmétique. Il faut donc passer à la valeur absolue.

L'écart absolu moyen par rapport à la moyenne sera : $EAM_{\bar{x}}=\frac{1}{n}\sum_{i=1}^{k} n_i\left|x_i-\bar{x}\right|$

L'écart absolu moyen par rapport à la médiane sera : $EAM_{M_e}=\frac{1}{n}\sum_{i=1}^{k} n_i\left|x_i-M_e\right|$

Les écarts absolus se prêtent mal aux calculs algébriques du fait de la valeur absolue. On leur préfère donc une autre caractéristique de dispersion qui fait évite l'écueil cité précédemment en passant au carré plutôt qu'à la valeur absolue.

La variance et l'écart type

La variance est la moyenne des carrés des écarts à la moyenne. L'écart type est la racine carrée de la variance. L'écart type est l'indicateur le plus fréquemment utilisé. Il répond aux conditions a (définition rigoureuse), b (prise en compte de toutes les observations) et f de Yule (se prête au calcul algébrique). Par contre il n'est pas facile à calculer (d) et est sensible aux valeurs aberrantes qui interviennent par leur carré (e).
La formule diffère selon les cas comme dans le calcul de la moyenne arithmétique ou de la médiane.

1. **A partir d'un tableau élémentaire :**

La variance est donnée par la formule suivante : $V(X)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$

On peut aussi développer la formule de la variance et dans ce cas on arrive à la formule suivante : $V(X)=\frac{1}{n}\sum_{i=1}^{n} x_i^2-\bar{x}^2$

La variance étant une somme de carrés, les grandeurs qu'elle prend ne sont pas très faciles à interpréter. Pour avoir un indicateur exprimé dans les mêmes unités que les observations, on passe à l'écart type : $\sigma=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$

2. A partir d'un tableau de traitement

La variance de la série des x_i est : $V(X)=\frac{1}{n}\sum_{i=1}^{k} n_i(x_i-\bar{x})^2=\sum_{i=1}^{k} f_i(x_i-\bar{x})^2$ et la formule développée est alors : $V(X)=\frac{1}{n}\sum_{i=1}^{k} n_i x_i^2-\bar{x}^2=\sum_{i=1}^{k} f_i x_i^2-\bar{x}^2$.

L'écart typeσ, racine carrée de la variance : $\sigma=\sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i(x_i-\bar{x})^2}=\sqrt{\sum_{i=1}^{k} f_i(x_i-\bar{x})^2}$