Comprendre la statistique descriptive

Dans la statistique, il existe deux grands domaines d'application. D'une part, la statistique descriptive, où l'on prend en compte la totalité d'une population (au sens de la statistique : voir ci-dessous la définition) ; d'autre part la statistique inductive, où l'on s'intéresse à un sous-ensemble (un échantillon), censé être représentatif d'une population. Cet article d'initiation traite uniquement la statistique descriptive.

La **statistique descriptive** peut elle-même s'intéresser à des statistiques recueillies **à un moment donné**, mais elle peut également étudier **l'évolution des données dans le temps** (étude des séries temporelles – par exemple, des données recueillies avec les recensements de 1982, 1990, 1999, 2007 ou 2008).

Une fois des données collectées, il s'agit de mettre de l'ordre dans ces informations pour les rendre lisibles et leur donner un sens. C'est là qu'interviennent deux grandes techniques : les **tableaux de données** et les **représentations graphiques**.

Mais on ne peut pas en permanence conserver la totalité des informations. Il convient alors de choisir des caractéristiques typiques, des valeurs représentatives de la totalité des informations. Pour résumer l'information, on peut utiliser des **caractéristiques de position** (ou de tendance centrale) avec des indicateurs qui fixent l'ordre de grandeur des données : ce sont **le mode**, **la médiane**, **la moyenne**. Mais il existe également des **caractéristiques de dispersion**. On essaie alors d'analyser comment les données sont groupées ou au contraire dispersées autour de la médiane ou de la moyenne : les indicateurs s'appellent **les quantiles**, **la variance**, **l'écart-type**.

Définitions

Population : ensemble étudié par la statistique.
 ex. Étudiants inscrits à l'université.

Individus : éléments des ensembles ci-dessus.
 ex. M. Dupont qui est étudiant.

Caractères : ce qui permet de décrire les individus d'une population.

Une constante : caractère commun à tous les individus. C'est uniquement un constat, donc pas d'intérêt majeur en statistique.
 ex. Le nombre de pages d'une publication (toujours vingt par exemple).

Une variable : caractère qui peut prendre au moins deux modalités.
 ex. Le sexe pour les étudiants d'une université.

Modalités : les différentes positions prises par une variable.
 ex. « Hommes » et « Femmes » sont les deux modalités possibles pour la variable « sexe ».

1) Étude des variables et représentation graphique

Selon sa nature, une information statistique ne sera pas traitée et représentée de la même manière. On distingue ainsi les variables quantitatives (discontinues ou continues) et les variables qualitatives.

Variables quantitatives : les modalités se prêtent à des calculs qui sont significatifs. Par exemple, on peut calculer une moyenne.

- **Variables quantitatives discontinues :** elles se mesurent avec des nombres entiers ; ou en un nombre fini de modalités dans un intervalle donné. Exemple : l'âge (en années révolues).
- **Variables quantitatives continues :** les différentes valeurs d'une modalité sont en nombre infini dans un intervalle donné. Exemple : l'âge si on le calcule en années, mois, jours, etc.
 À noter : des variables par nature discontinue peuvent être assimilées à des variables continues pour la présentation des résultats. La distinction n'est pas stricte. Ce qui est important, c'est d'assimiler la logique interne et de justifier les choix que l'on effectue.

Variables qualitatives : les modalités ne se prêtent pas à des calculs significatifs. Par exemple, on ne peut pas calculer une moyenne sur la modalité « sexe », « nationalité », « profession », etc.

Le fait qu'une variable soit quantitative ou qualitative, et, si elle est quantitative, qu'elle soit discontinue ou continue, modifie le type de représentation graphique que l'on peut en faire.

Variable quantitative discontinue	Variable quantitative continue	Variable qualitative
Diagramme en bâtons	Histogramme	Diagrammes en tuyaux d'orgue, en barre ou circulaire

(Voir ci-après)

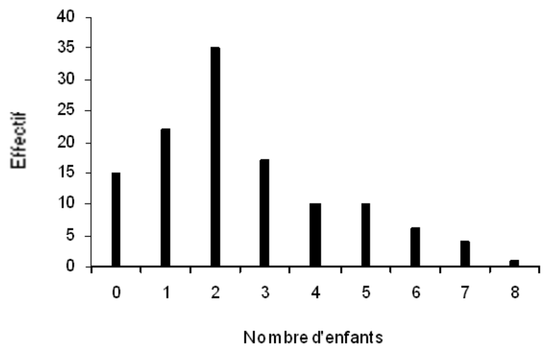
1.1. Variables quantitatives discontinues :

Soit 120 ménages : ils n'ont aucun enfant ou de 1 à 8 enfants. La variable est ici le nombre d'enfants. On la représente sur l'axe horizontal des abscisses. On utilise **un diagramme en bâtons** pour la représentation graphique des variables quantitatives discontinues : il représente les effectifs (ou les fréquences, soit par exemple des pourcentages), correspondant à chaque modalité, par des bâtons dont la longueur leur est proportionnelle.

Tableau n° 1 – Nombre d'enfants par ménage (1)

Nombre d'enfants (variable)	Nombre de ménages (effectifs)
0	15
1	22
2	35
3	17
4	10
5	10
6	6
7	4
8	1
Total	120

Graphique n° 1 – Nombre d'enfants par ménage



1.2. Variables quantitatives continues :

1.2.1. Classes d'amplitude :

Les modalités pouvant prendre un nombre infini de valeurs, on constitue des classes : par exemple, « 600-700 euros ».

À noter : par convention, et sauf exception (ex. l'âge), la classe inclut la valeur inférieure, mais exclut la valeur supérieure. Ainsi, la classe « 600-700 euros » se comprend comme suit :

[600-700 euros[
ou 600 à 699,99 euros
ou 600 à moins de 700 euros.

Les classes extrêmes sont généralement « ouvertes » (ex. « moins de 600 euros »). Cela permet d'utiliser un tableau quelle que soit la valeur la plus basse ou la plus haute (principe d'exhaustivité). Deux inconvénients : le manque de précision (ex. les revenus pouvant être négatifs) et les difficultés que cela entraîne pour certains calculs.

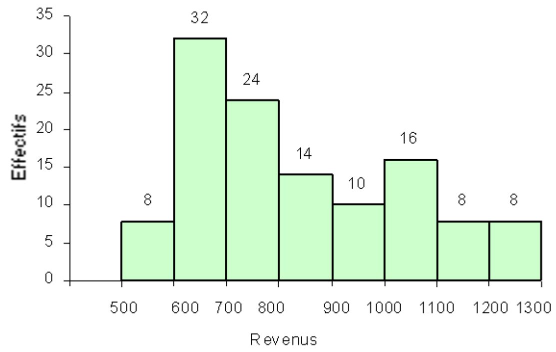
1.2.2. Classes d'amplitude égale :

La représentation graphique d'une variable quantitative continue répartie en classes est **un histogramme**. On représente les effectifs (ou les fréquences) par des rectangles juxtaposés dont les surfaces leur sont proportionnelles. Si les classes sont d'amplitude égale, la largeur des rectangles (sur l'axe horizontal des abscisses) est identique.

Tableau n° 2 – Nombre de ménages selon les revenus

Revenus (variable)	Nombre de ménages (effectifs)
Moins de 600 euros	8
600 à 700 euros	32
700 à 800 euros	24
800 à 900 euros	14
900 à 1 000 euros	10
1 000 à 1 100 euros	16
1 100 à 1 200 euros	8
1 200 euros ou plus	8
Total	120

Graphique n° 2 – Nombre de ménages selon les revenus



1.2.3. Classes d'amplitude inégale :

Si les classes sont d'amplitude inégale (ex. « moins de 1 mois » et « 1 à 3 mois »), il faut choisir une amplitude unitaire de référence (ex. 1 mois) et corriger tous les effectifs (ou fréquences) en les ramenant à ce qu'ils seraient s'ils correspondaient à cette amplitude de référence. Dans cette logique, l'échelle ne peut plus être utilisée et il faut réintégrer les nombres des effectifs (ou fréquences) à l'intérieur des rectangles.

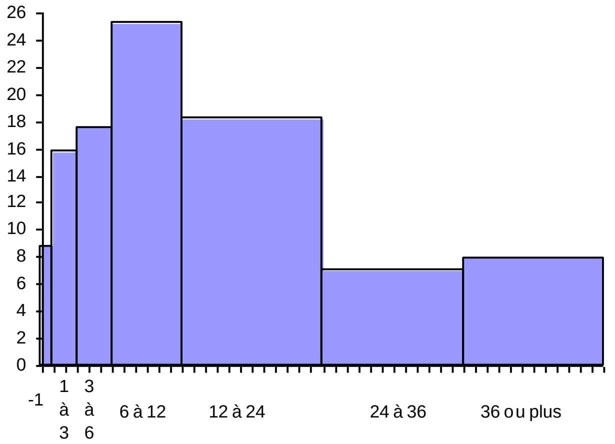
Tableau n° 3 – Nombre de demandeurs d'emploi selon l'ancienneté dans le chômage

Amplitude	Durée de chômage (variable)	Nombre de demandeurs d'emploi		
		Chiffres absolus (effectifs)	Valeurs relatives (fréquences)	Fréquences corrigées base 1 (fréquence / amplitude)
1	**Moins d'un mois**	118 236	8,73 %	8,73
2	**1 à 3 mois**	212 636	15,70 %	7,85
3	**3 à 6 mois**	236 367	17,46 %	5,82
6	**6 à 12 mois**	340 741	25,17 %	4,19
12	**12 à 24 mois**	245 341	18,12 %	1,51
12	**24 à 36 mois**	92 974	6,87 %	0,57
12	**36 mois ou plus**	107 668	7,95 %	0,66
Total		1 353 963	100,00 %	

(1) – Il s'agit ici d'un **tableau de distribution** car il porte sur des effectifs. S'il portait sur des fréquences (ex. des pourcentages), ce serait un **tableau de répartition**.

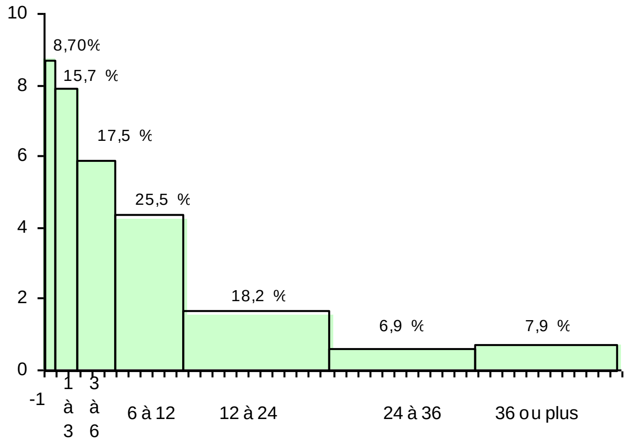
Graphiques n° 3 – Nombre de demandeurs d'emploi selon l'ancienneté dans le chômage (en mois)

a) Le graphique à ne pas faire :



Ce tableau fait ressortir une prépondérance de demandeurs d'emploi dont l'ancienneté d'inscription est comprise entre six et douze mois. Cette information est contraire à la réalité. Comme le prouve notre second graphique, ce sont les demandeurs d'emploi depuis moins d'un mois qui sont prépondérants.

b) Le graphique à réaliser :



1.3. Variables qualitatives :

Les modalités ne sont pas mesurables d'une façon significative. Les représentations graphiques possibles sont très diversifiées. Le principe est d'avoir des surfaces proportionnelles aux valeurs. Les trois formes de base : **les diagrammes en tuyaux d'orgue**, **en barre** ou bien **circulaires**.

Tableau n° 4 – Nombre de ménages selon la PCS [2] de la personne de référence

PCS	Chiffres absolus	Valeurs relatives
Cadres	15	12,5 %
Professions intermédiaires	27	22,5 %
Employés	30	25,0 %
Ouvriers	48	40,0 %
Total	120	100,0 %

Graphiques n° 4 – Nombre de ménages selon la PCS de la personne de référence age

Diagramme en tuyaux d'orgue

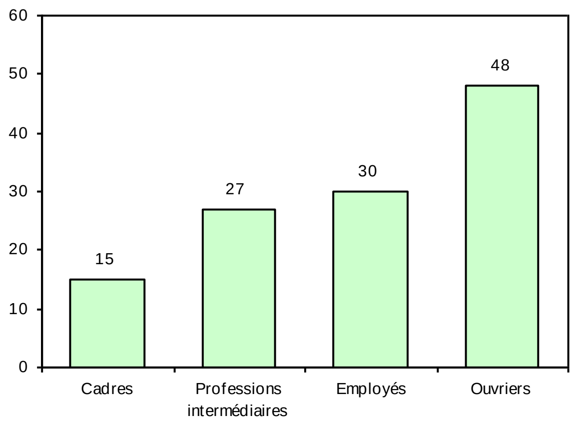


Diagramme en barre

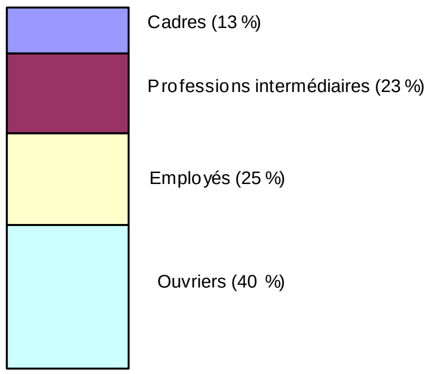
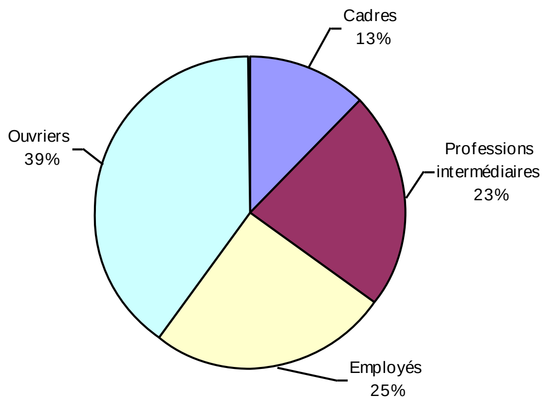


Diagramme circulaire



(2) – Professions et catégories socioprofessionnelles (soit PCS, et non plus CSP).

2) Indicateurs pour résumer l'information

Résumer l'information, c'est pouvoir disposer de données utiles sous forme d'un petit nombre de résultats. On perd inévitablement de l'information, mais on gagne en lisibilité et cela peut permettre d'établir des comparaisons. On distingue les caractéristiques de position (ou de tendance centrale) et les caractéristiques de dispersion.

2.1. Les caractéristiques de position (ou de tendance centrale) :

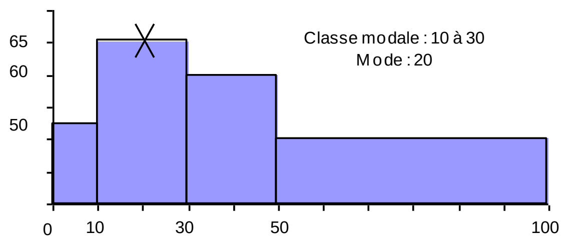
Il en existe trois : le mode, la médiane et la moyenne. Le mode s'applique aux variables quantitatives et qualitatives, mais la médiane et la moyenne ne peuvent pas s'utiliser pour les variables qualitatives.

2.1.1. Le mode d'une variable :

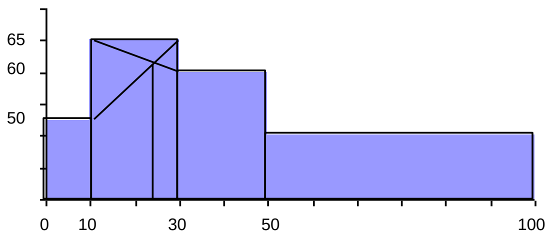
Le mode est la valeur d'une variable qui est la plus fréquente. Dans notre graphique n° 1 (nombre d'enfants par ménage), le mode est 2. Dans notre graphique n° 3 (nombre de demandeurs d'emploi selon l'ancienneté dans le chômage), le mode est « moins d'un mois » car on prend la valeur corrigée la plus élevée (et non pas la valeur observée). Enfin, dans les graphiques n° 4, ce sont les ouvriers qui constituent le mode.

Pour calculer le mode avec des variables quantitatives continues, on dispose de plusieurs solutions.

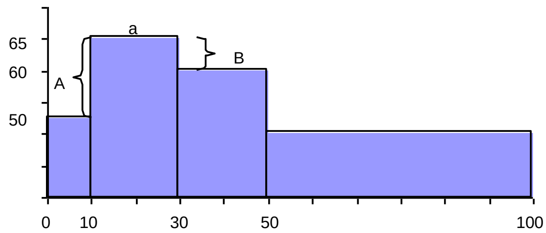
a) On prend le milieu de la classe modale.



b) On relie la borne gauche supérieure à la borne droite inférieure et la borne gauche inférieure à la borne droite supérieure. Le mode est la valeur au croisement des deux droites ainsi créées.



c) On peut enfin utiliser une formule mathématique.



$$\text{Mode} = \quad Xmo + \left(\frac{A}{A+B} \quad x\ a \right)$$

Xmo : valeur de la borne inférieure de la classe modale.
a : amplitude de la classe modale

soit :

$$10 + \left(\frac{(65\text{-}50)}{(65\text{-}50) + (65\text{-}60)} \quad x\ 20 \right) = 25$$

∗∗

Les avantages du mode :

- Il est universel : ainsi, c'est la seule façon de situer la position d'une variable qualitative.
- Il est relativement simple à calculer. De plus, il a une signification immédiate, très concrète.

Les inconvénients du mode :

- Il est relativement sensible au classement de la variable (par exemple, pour les variables quantitatives continues, il suffit de modifier les classes pour changer le mode). On peut donc modifier la valeur modale en jouant sur l'amplitude des classes (en particulier sur de tout petits effectifs).
- Le mode se prête extrêmement mal aux calculs algébriques. Une fois qu'on l'a calculé, on ne peut en faire rien d'autre.

2.1.2. La médiane (uniquement pour les variables quantitatives).

La médiane est la valeur d'une variable qui partage la population en deux effectifs égaux. Comment le partage s'effectue-t-il ? Cela implique qu'on a ordonné les valeurs selon un ordre croissant (ou décroissant).

a) Variables quantitatives discontinues :

Tableau n° 5 – Nombre d'enfants par ménage

Nombre d'enfants	**Nombre de ménages**	
	Chiffres absolus	Effectifs cumulés
0	15	15
1	22	37
2	35	72
3	17	89
4	10	99
5	10	109
6	6	115
7	4	119
8	1	120
Total	120	

Lecture des effectifs cumulés croissants : 37 ménages ont **au plus** un enfant (ou bien : 37 ménages ont **moins de** deux enfants).

Pour obtenir la médiane, il s'agit de bien lire le tableau : combien a d'enfants le 60e ménage (120/2) ? Comme 72 ménages ont au plus deux enfants, le 60e en a forcément deux. La médiane est donc « 2 ». On dira qu'il y a autant de ménages qui ont deux enfants ou moins, que de ménages qui ont deux enfants ou plus.

b) Variables quantitatives continues :

Tableau n° 6 – Nombre de ménages selon les revenus

Revenus	Nombre de ménages	
	Chiffres absolus	Effectifs cumulés
Moins de 600 euros	8	8
600 à 700 euros	32	40
700 à 800 euros	24	64
800 à 900 euros	14	78
900 à 1 000 euros	10	88
1 000 à 1 100 euros	16	104
1 100 à 1 200 euros	8	112
1 200 euros ou plus	8	120
Total	120	

40 ménages gagnent au plus 700 euros, et 64, au plus 800 euros. Le 60e gagne forcément entre 700 et 800 euros. La valeur médiane peut être fixée juste au milieu (par convention), soit 750 euros, ou bien on effectue une règle de trois en prenant pour hypothèse une répartition uniforme au sein de la classe de 700 à 800 euros.

$$24 \left(20 \left(\begin{array}{c} 40^{\text{ème}} \rightarrow 700 \text{ euros} \\ 60^{\text{ème}} \rightarrow ? \\ 64^{\text{ème}} \rightarrow 800 \text{ euros} \end{array} \right) ? \right) 100 \text{ euros}$$

soit :

$$\frac{100 \times 20}{24} = 83$$

Médiane = 700 + 83 = 783 euros

Ainsi, il y a autant de ménages qui gagnent moins de 783 euros que de ménages qui gagnent plus. Ou encore : un ménage sur deux gagne moins de 783 euros (ou plus de 783 euros).

La médiane ne dépend pas des valeurs elles-mêmes, mais de l'ordre dans lequel elles sont rangées, ce qui entraîne des conséquences positives et des conséquences négatives.

Les conséquences positives :

- Quand une variable se présente avec deux classes extrêmes ouvertes (ex. « moins de 600 euros » et « 1 200 euros ou plus »), si on ne connaît pas les valeurs minimales et maximales, la médiane est la caractéristique de tendance centrale la plus pertinente.
- De même, c'est la meilleure caractéristique quand on travaille sur des variables dont les valeurs sont de mauvaise qualité.
- Enfin, quand on travaille sur de petites populations, la médiane est la caractéristique la moins sensible au hasard.

Les conséquences négatives :

- Les comparaisons sont souvent difficiles à faire car la distribution des valeurs des modalités peut varier dans de très grandes proportions.
- La même valeur médiane peut correspondre à deux populations totalement différentes.

2.1.3. La moyenne (uniquement pour les variables quantitatives) :

La moyenne est le rapport entre la somme des valeurs, pondérées par leur effectif, et le nombre de valeurs. C'est donc la valeur la plus proche de l'ensemble des valeurs.

Tableau n° 7 – Notes obtenues par un étudiant dans différentes matières et calcul de la moyenne générale

Matières	Coefficients	Notes	Coefficients x notes
Statistiques	3	11	33
Sociologie	4	13	52
Politique	6	7	42
Histoire	2	6	12
Economie	2	12	24
Démographie	3	12	36
Total	20		199

La moyenne est égale à :

$$\frac{199}{20} = 9{,}95/20$$

Dans le cas de variables quantitatives continues, on borne les classes extrêmes et on prend, par convention, la valeur moyenne à l'intérieur de chaque classe. Ainsi, pour la classe « moins de 600 euros », on pourra retenir « 550 euros » (classe de 500 à 600 euros), et pour la classe « 600 à 700 euros », on retiendra « 650 euros ».

Les avantages de la moyenne :

- La moyenne tient compte de toutes les valeurs de la variable (valeur représentative de l'ensemble des valeurs).
- Elle se prête particulièrement bien aux calculs algébriques. Dans tous les cas, il est prudent d'y adjoindre un paramètre de dispersion (voir plus loin la variance et l'écart-type).

Les inconvénients de la moyenne :

- Chaque individu n'intervient pas de la même manière dans le calcul de la moyenne. Cela entraîne toute une série de déformations. Ainsi, il est prudent de ne pas utiliser des moyennes à des fins comparatives sans se demander si l'utilisation est possible. De plus, dès que l'on compare des moyennes, il faut se demander s'il n'y a pas des effets de structure qui les perturbent.

Pour l'illustrer, on peut reprendre l'exemple ci-dessus : l'étudiant a donc redoublé. Dans chaque matière, il obtient un point de plus. Seulement, les coefficients ont changé. Malgré ses progrès manifestes, il va encore redoubler.

Tableau n° 7 bis – Notes obtenues par un étudiant dans différentes matières et calcul de la moyenne générale

Matières	Coefficients	Notes	Coefficients x notes
Statistiques	2	12	24
Sociologie	3	14	42
Politique	8	8	64
Histoire	4	7	28
Economie	2	13	26
Démographie	1	13	13
Total	20		197

La moyenne est égale à :

$$\frac{197}{20} = 9{,}85/20$$

Autre illustration avec les salaires moyens dans deux entreprises.

Tableau n° 8 – Salaires moyens

Catégories	Entreprise A		Entreprise B	
	Effectifs	Salaire moyen	Effectifs	Salaire moyen
Techniciens	2/3	1 800 €	1/3	1 900 €
Ouvriers	1/3	1 200 €	2/3	1 300 €

Salaire moyen de l'entreprise A : 1 600 euros.
Salaire moyen de l'entreprise B : 1 500 euros.

Dans l'entreprise B, les techniciens gagnent plus cher que ceux de l'entreprise A et il en est de même pour les ouvriers. Pourtant, le salaire moyen de l'entreprise A est plus élevé que celui de l'entreprise B !

Pour éviter ce biais, on peut calculer des moyennes fictives mais qui permettent de réaliser des comparaisons : ainsi, on pourrait prendre comme hypothèse que les entreprises A et B ont chacune 50 % de techniciens et d'ouvriers ; ou bien que l'entreprise B a la même répartition de son personnel que l'entreprise A (dans ce cas, le salaire moyen est de 1 700 euros, et non plus de 1 500 euros). On utilise ainsi non plus une moyenne arithmétique, mais une moyenne comparative ou standardisée.

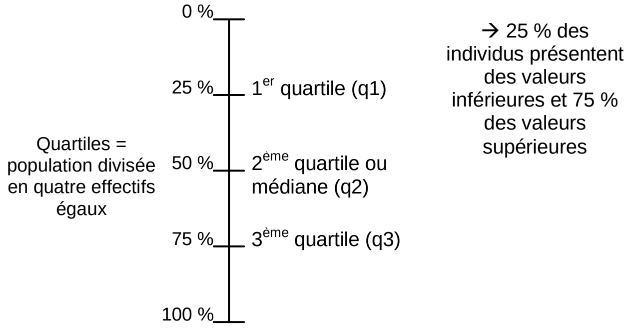
2.2. Les caractéristiques de dispersion :

Les caractéristiques de dispersion mesurent la répartition des valeurs les unes par rapport aux autres. On distingue le quantile, la variance et l'écart-type.

2.2.1. Le quantile :

Le quantile est une caractéristique de dispersion par rapport à la médiane. C'est une valeur de la variable telle qu'une certaine proportion de la population a une valeur qui lui est inférieure, et que la proportion complémentaire a une valeur supérieure. La médiane est en fait un quantile particulier pour lequel les proportions sont de 50 %.

Ainsi, pour le quartile, la population est divisée en quatre quarts ; de même, on peut définir des déciles, voire des centiles (respectivement dix dizaines et cent centièmes).



On peut calculer :

- L'intervalle (ou écart) interquartile : q3 – q1.
- L'intervalle interquartile relatif : $\frac{q3 - q1}{q2}$
- Le rapport interquartile : q3/q1.

2.2.2. La variance et l'écart-type :

La variance est égale à la moyenne des carrés des écarts à la moyenne. Elle mesure la fluctuation moyenne autour de la moyenne.

L'écart-type est la racine carrée de la variance. La moyenne est d'autant plus représentative que l'écart-type est faible

Comment calculer l'écart-type ? Soit les dix valeurs suivantes : 6, 10, 4, 4, 9, 6, 7, 8, 4, 8.

a) On calcule la moyenne :

$$\frac{66}{10} = 6{,}6$$

b) On calcule les écarts de chaque valeur à la moyenne ; on les élève au carré ; on calcule la moyenne de ces carrés (c'est-à-dire la variance) – cf. tableau ci-contre.

c) La racine carrée de la variance nous donne l'écart-type :

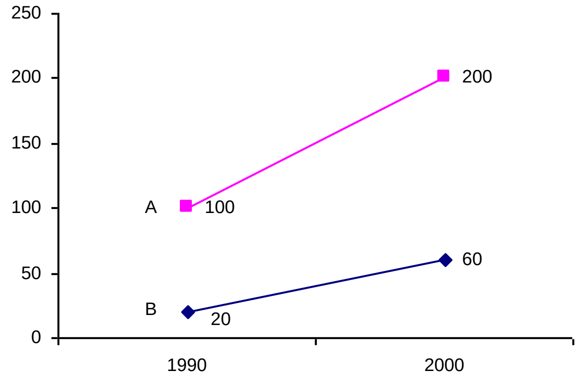
$$\sqrt{4{,}24} = 2{,}1$$

Valeurs	Ecarts à la moyenne	Carrés des écarts à la moyenne
6	6 - 6,6 = - 0,6	- 0,6 x - 0,6 = 0,36
10	10 - 6,6 = + 3,4	+ 3,4 x + 3,4 = 11,56
4	4 - 6,6 = - 2,6	- 2,6 x - 2,6 = 6,76
4	4 - 6,6 = - 2,6	- 2,6 x - 2,6 = 6,76
9	9 - 6,6 = + 2,4	+ 2,4 x + 2,4 = 5,76
6	6 - 6,6 = - 0,6	- 0,6 x - 0,6 = 0,36
7	7 - 6,6 = + 0,4	+ 0,4 x + 0,4 = 0,16
8	8 - 6,6 = + 1,4	+ 1,4 x + 1,4 = 1,96
4	4 - 6,6 = - 2,6	- 2,6 x - 2,6 = 6,76
8	8 - 6,6 = + 1,4	+ 1,4 x + 1,4 = 1,96
	NB. Le total des écarts à la moyenne est égal à zéro	Moyenne des carrés (variance) $\frac{42{,}40}{10} = 4{,}24$

3) Évolution des valeurs dans le temps

Nous avons vu que la statistique descriptive s'intéresse aux statistiques recueillies à un moment donné, mais elle peut également étudier l'évolution des données dans le temps. Nous nous contenterons ici d'aborder la différence entre une progression arithmétique (basée sur l'addition) et une progression géométrique (basée sur la multiplication) à partir d'un exemple concret.

Graphique n° 5 – Evolution A et B entre 1990 et 2000



A : 200 – 100 = 100
B : 60 – 20 = 40
A s'est accru davantage que B

mais,

A : 200 / 100 = 2
B : 60 / 20 = 3
B s'est accru davantage que A

On a raison dans les deux cas : tout dépend si l'on retient une variation absolue (1^er^ exemple) ou une variation relative (2^e^ exemple). Dans un tableau, il est souvent prudent de donner les chiffres correspondant aux deux types de variation.

Test : cherchez les anomalies !

NB. Ce tableau est destiné à une publication… Il contient une dizaine d'anomalies. Cherchez-les !

Répartition des étudiants selon la CSP des parents

	Nombre	Pourcentage
Agriculteurs exploitants	10	12,35 %
Artisans, commerçants et chefs d'entreprise	6	7,41 %
Cadres et professions intellectuelles supérieures	8	9,88 %
Professions intermédiaires	13	16,05 %
Employés	12	14,81 %
Salariés	20	24,69 %
Retraités	6	7,41 %
Autres	6	7,41 %