



Collège Saint-Barthélemy
Y. Michiels

Mathématique – Quatrième année

FASCICULE 3

Statistique descriptive à une variable



OBJECTIFS – UAA1 : Statistique descriptive à une variable

Connaitre

-  Expliquer le vocabulaire statistique.
-  Identifier les différents types de caractères statistiques et décrire les informations graphiques et numériques qui peuvent y être associées.
-  Expliquer pour quels usages sont requis les indicateurs de position et/ou de dispersion

Appliquer

-  Calculer ou estimer les indicateurs de position et de dispersion et les positionner sur un graphique.
-  Construire différents graphiques statistiques.
-  Extraire une information de graphiques et de tableaux statistiques.
-  Utiliser l'inégalité de Tchebychev.

Transférer

-  Choisir un support graphique, une valeur centrale, un indice de dispersion pour étudier une situation.
-  Critiquer des informations graphiques, numériques, textuelles, ...
-  Commenter des informations fournies sur un même sujet par différents supports.
-  Interpréter un résultat obtenu en lien avec le caractère étudié et le contexte.

EXPLORATION : Statistiques à une variable

Activité 1 : Présentation

La statistique est une branche des sciences mathématiques dont l'objet est la collection, l'analyse, l'interprétation et la présentation de données. Les outils statistiques peuvent s'appliquer dans toutes les disciplines :

- Les autorités administratives d'une université désirent étudier les taux d'échec (et de succès) des étudiants s'inscrivant en première année de médecine.
- Un grand recensement est organisé dans la ville d'Ath (Belgique) afin d'étudier la population (âge, sexe, emploi, ...).
- Le département des ressources humaines d'une grande entreprise souhaite évaluer l'opinion générale de ses employés face à un ajustement des horaires de la cafétéria.
- L'équipe de campagne d'un parti politique cherche à connaître les intentions de vote des habitants de la ville de Liège pour la prochaine élection communale.
- Un constructeur automobile souhaite évaluer la distance de freinage sur route sèche de son nouveau modèle de berline 5 places équipée d'un certain type de pneus.
- La haute autorité de surveillance des jeux de hasard estime que les montants pariés sur la rencontre de football opposant les clubs de Standard et d'Anderlecht sont anormalement élevés ; elle souhaite donc évaluer la probabilité de fraude.
- Dans un essai clinique, les autorités sanitaires souhaitent déterminer laquelle de deux procédures thérapeutiques est la plus efficace.

De multiples autres exemples sont imaginables. La statistique descriptive est la branche des statistiques qui regroupe les nombreuses techniques utilisées pour synthétiser, décrire, présenter et interpréter un ensemble conséquent de données.

Pour t'approprier le vocabulaire spécifique aux statistiques descriptives, regarde cette vidéo d'introduction :

<https://www.youtube.com/watch?v=VF6VqfknETg>

Activité 2 : Vocabulaire de base

Associe le mot ou le groupe de mots à la partie de phrase soulignée qui convient.

Classe →	← (1)
Effectif →	← (2)
Population →	← (3)
Modalité →	← (4)
Variable quantitative →	← (5)
Etude statistique →	← (6)
Fréquence →	← (7)
Variable qualitative →	← (8)

« L'asbl SPLJ (Sport Pour Les Jeunes) a **collecté, traité et interprété** (1) diverses données. Cette association a interrogé **1000 jeunes de 12 à 16 ans** (2). Une partie de l'étude porte sur le sport que ces jeunes pratiquent. Entre autres, elle a pu établir un top 10 des **sports les plus pratiqués** (3) ainsi qu'un **budget mensuel moyen consacré à la pratique du sport** (4). Parmi les sports les plus

pratiqués figure le **football** (5) qui ne compte pas moins de **491 adeptes** (6) chez les jeunes interrogés. Quant au budget, plus de la moitié de ces jeunes dépensent souvent **entre 15 et 40€ par mois** (7) pour leur inscription, leur abonnement, et le matériel spécifique. L'asbl a aussi pu constater **que 60 %** (8) de ces 1000 jeunes pratiquent leur sport au moins une fois par semaine. »



→ ***Théorie page 5***

Activité 3 : Création d'une enquête

1. INTRODUCTION

Une étude statistique naît d'un besoin d'information pour alimenter un processus de décision. Elle est présente dans de nombreux domaines : la recherche médicale et pharmaceutique, les tests de psychologie, les assurances... Que dire d'un monde politique privé de sondages d'opinion ? La statistique est un outil de communication universel qui permet de faire passer un message. Pour préparer puis réaliser une étude statistique, le chargé d'étude doit disposer de suffisamment d'éléments pour établir la liste des données qu'il va essayer de recueillir au moyen d'une enquête, d'un sondage, d'un questionnaire... Il doit avoir une idée des variables statistiques à définir ainsi que du type de population à observer. A la fin de la campagne de mesures, les données vierges de tout traitement sont disponibles.

2. VOCABULAIRE

2.1. Population

La **population** est l'ensemble des éléments auxquels se rapporte la recherche statistique (c'est donc l'ensemble des éléments que l'on va étudier).

Il convient d'être précis lorsqu'on la définit.

Exemple : définir une population comme « les véhicules en circulation » n'est pas assez complet. En effet, nous ne précisons pas quel type de véhicules, dans quel pays (ou sur quel territoire), en quelle année, etc. Une définition plus précise serait « les véhicules à quatre roues immatriculés et en circulation en Belgique au 1er janvier 2000 ».

2.2. Individu

Chacun des éléments de la population est appelé un **individu**.

Un individu peut donc être une personne, un objet, un fait.

Remarquons que cette notion est à écarter de celle connue dans le langage courant (où l'individu est assimilé à une personne). En effet, en statistiques, les individus peuvent être des objets. Ce sera par exemple le cas si nous étudions la population des véhicules automobiles en circulation en Belgique.

2.3. Échantillon

Lorsque la population est trop grande que pour être étudiée dans sa totalité, on décidera d'extraire de cette population un **sous-ensemble représentatif** de celle-ci. Ce sous-ensemble porte le nom d'échantillon.

L'**échantillon** est la partie de la population qui sera effectivement observée.

2.4. Caractère ou variable statistique

Sur base de l'échantillon (ou de la population complète), on décidera d'étudier certains critères. Ces derniers se nomment les **caractères statistiques** ou encore les **variables statistiques**. Nous pourrions étudier les caractères suivants pour les véhicules automobiles : la couleur, le degré de satisfaction du conducteur principal, le nombre d'airbags, la distance parcourue par le véhicule depuis sa première mise en circulation, etc.

Le **caractère** ou **variable statistique** est l'information commune recueillie à propos des individus.

2.5. Variable qualitative et variable quantitative

Etudions maintenant les réponses possibles pour chacune de ces variables :

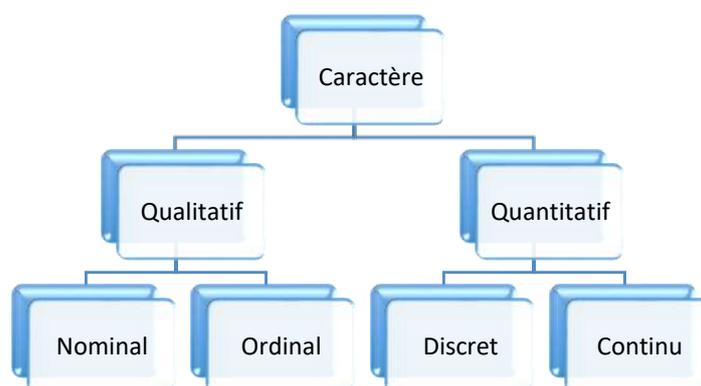
- Le caractère « *couleur* » peut prendre plusieurs valeurs : rouge, blanc, bleu, vert ... qui sont des **mots**. Ces différentes valeurs ne sont pas mesurables mais sont dénombrables. On appelle cela un **caractère qualitatif**.
Comme il n'y a pas de hiérarchie entre les couleurs (vert n'est pas plus grand ou mieux que blanc), on précisera que le caractère est **qualitatif nominal**.
- Le caractère « *degré de satisfaction du conducteur principal* » peut revêtir plusieurs valeurs : très satisfait, satisfait, peu satisfait ou insatisfait.
A l'instar du caractère « couleur », ce sont des **mots** et, en toute logique, nous dirons qu'il est **qualitatif**. Mais à l'opposé du précédent, un certain ordre, une certaine hiérarchie peut être établie entre les différentes réponses possibles ('satisfait' est mieux que 'peu satisfait'). Nous préciserons dès lors que la variable est **qualitative ordinale**. Dans le cas des variables qualitatives, les différentes valeurs que peuvent prendre la variable sont appelées les modalités.

Une variable est dite **qualitative** si ses modalités ne peuvent être mesurées ou repérées par un nombre. Elle peut être **qualitative nominale** ou **qualitative ordinale**

- Le caractère « *nombre d'airbags* » peut amener comme réponse : 0, 1, 2, 3, 4, ... qui sont des **nombres**. Il est donc possible de mesurer, de quantifier, le nombre d'airbags que possède un véhicule. Nous appelons une telle variable une **variable quantitative**. Cette dernière possède aussi une autre caractéristique : elle ne peut prendre qu'un **nombre fini de valeurs** dans l'intervalle (on imagine mal une voiture avec 2,7 airbags...). Dès lors, nous préciserons que cette variable est **quantitative discrète**.
- La variable « *distance parcourue par le véhicule* » peut amener comme réponse n'importe quel nombre entre 0 et... le maximum qu'un véhicule est capable de parcourir ! Tout comme le caractère précédent, les réponses fournies sont des nombres et le caractère sera donc quantitatif. Mais il n'est pas aisé de référencer toutes les réponses possibles tellement le choix est grand. On qualifiera ce caractère de **quantitatif continu**.

Une variable est dite **quantitative** si ses modalités sont mesurables ou repérables par un nombre. Les variables quantitatives sont divisées en deux sous-groupes :

- Une **variable quantitative** est dite **discrète** si les modalités sont des valeurs isolées ou peu nombreuses (ex. : nombre d'enfants d'une famille, nombre de voiture par famille,...).
- Une **variable quantitative** est dite **continue** si les modalités peuvent prendre n'importe quelle valeur comprise dans un intervalle ou un très grand nombre de valeurs (ex. : taille des adolescents de 15 ans, salaire brut des wallons de 50 ans,...).



2.6. Modalités d'un caractère ou valeurs de la variable statistique

Les **modalités** d'un caractère ou valeurs de la variable statistique sont les différentes valeurs ou formes que le caractère peut prendre.

Chaque individu présente une et une seule modalité du caractère.

- Exemples :
- Modalités du nombre de cigarettes : 0, 1, 2, 3, ...
 - Modalités de la couleur des voitures : bleue, rouge, noire, grise, ...
 - Modalités du nombre de paniers marqués : 0, 1, 2, ... 40, 41, ...
 - Modalités de la taille en cm : la taille peut varier, par exemple, entre 150 et 185 cm

2.7. Série statistique

La **série statistique** est l'ensemble des résultats obtenus par la collecte des observations, c'est-à-dire les valeurs prises par la variable statistique.

En résumé, on va considérer une **population** constituée d'**individus**. Souvent, on n'étudiera pas toute la population, mais seulement un **échantillon** représentatif de la population. Pour les individus de l'échantillon, on va observer un **caractère** (ou variable statistique) qui peut prendre plusieurs valeurs (**modalités**). Le résultat de cette collecte d'informations est une **série statistique**.

2.8. Quelques exemples :

- **Le nombre d'enfants par famille en Wallonie :**

Population :

Individu :

Caractère observé :

Modalités :

- **La taille en cm des garçons d'une classe de première année à Saint-Barthélemy :**

Population :

Individu :

Caractère observé :

Modalités :

- **La nationalité des habitants de Liège :**

Population :

Individu :

Caractère observé :

Modalités :

- **Le niveau de satisfaction d'un produit proposé par un supermarché :**

Population :

Individu :

Caractère observé :

Modalités :

Exemple 1 : Pour une variable qualitative nominale

Etude statistique

Voici le relevé du nombre de naissances enregistrées en décembre 2012 en Belgique (source SPF Intérieur – Registre National).

Population dans les grandes villes et chefs-lieux de province : nombre de naissances	
Situation 12 / 2012	
Grandes villes / Provinces	Nombre de naissances
Anvers / Anvers	591
Bruxelles / Bruxelles capitale	205
Louvain / Brabant flamand	76
Wavre / Brabant wallon	37
Bruges / Flandre occidentale	81
Gand / Flandre orientale	271
Mons / Hainaut	92
Liège / Liège	219
Hasselt / Limbourg	80
Arlon / Luxembourg	28
Namur / Namur	91

Population :

Individu :

Caractère observé :

Modalités :

Représentation des données sous forme de tableau

Modalités (x_i)	Effectifs (n_i)	Fréquences (f_i)
Anvers		
Bruxelles		
Louvain		
Wavre		
Bruges		
Gand		
Mons		
Liège		
Hasselt		
Arlon		
Namur		
	$n =$	

Remarque : comme la variable est qualitative, effectifs cumulés et fréquences cumulées n'ont pas de sens.

Exemple 2 : Pour une variable qualitative ordinale

Etude statistique

Voici une enquête réalisée auprès de 100 étudiants d'une université européenne sur leur avis pédagogique à l'égard d'un assistant.

TD très défavorable : 2 - **M** moyen : 54 - **F** favorable : 31 - **D** défavorable : 4 - **TF** très favorable : 9

Population :

Individu :

Caractère observé :

Modalités :

Représentation des données sous forme de tableau

Modalités (x_i)	Effectifs (n_i)	Fréquence (f_i)
TD (x_1)		
D (x_2)		
M (x_3)		
F (x_4)		
TF (x_5)		

Remarque : comme la variable est qualitative, effectifs cumulés et fréquences cumulées n'ont pas de sens.

Exemple 3 : Pour une variable qualitative discrète

Etude statistique

Voici les résultats (cotes sur 10) de 20 élèves d'une classe de 4ème lors de la dernière interrogation surprise de math :

10 - 8 - 0 - 6 - 5 - 9 - 9 - 2 - 4 - 7 - 7 - 1 - 5 - 8 - 7 - 4 - 9 - 7 - 6 - 7

Population :

Individu :

Caractère observé :

Modalités :

Représentation des données sous forme de tableau

Cote obtenue (/10)	Effectifs	Effectifs cumulés	Fréquences (%)	Fréquences cumulées (%)
0	1	1	5 %	5 %
1	1	2	5 %	10 %
2	1	3	5 %	15 %
3	0	3	0 %	15 %
4	2	5	10 %	25 %
5	2	7	10 %	35 %
6	2	9	10 %	45 %
7	5	14	25 %	70 %
8	2	16	10 %	80 %
9	3	19	15 %	95 %
10	1	20	5 %	100 %
	N = 20			

- Pour compléter la colonne « **effectifs** », il suffit de comptabiliser combien d'élèves ont obtenu la cote figurant dans la colonne « cote obtenue (/10) ».
Case grisée « 5 » des effectifs : « 5 élèves de la classe ont obtenu une cote égale à 5 sur 10 lors de l'interrogation de mathématique ».
- Pour compléter la colonne « **effectifs cumulés** », il suffit d'additionner tous les effectifs des cotes inférieures ou égales à la cote dont il est question. Pour la case grisée, on a additionné tous les effectifs des cotes inférieures ou égales à 5 sur 10 (2 + 2 + 0 + 1 + 1 + 1 = 7). La dernière valeur de la colonne des effectifs cumulés doit toujours correspondre à l'effectif de la population étudiée (dans notre exemple : « 20 » car il y a 20 élèves dans la classe).
Case grisée « 7 » des effectifs cumulés : « 7 élèves de la classe ont obtenu une cote inférieure ou égale à 5 sur 10 lors de l'interrogation de mathématique ». Et donc, par déduction : « 13 élèves de la classe ont obtenu une cote strictement supérieure à 5 sur 10 lors de cette même interrogation ».
- Pour compléter la colonne « **fréquences** », il suffit de diviser l'effectif de la ligne par l'effectif total de la population et de multiplier le résultat par 100 (pour obtenir un pourcentage). Pour la case grisée, on a réalisé le calcul suivant :

$$\frac{5 \text{ (nombre d'élèves ayant obtenu 7/10)}}{20 \text{ (nombre d'élèves dans la classe)}} \cdot 100$$

Case grisée « 25 » des fréquences : « 25 % des élèves de la classe ont obtenu une cote égale à 7 sur 10 lors de l'interrogation surprise ».

- Pour compléter la colonne « **fréquences cumulées** », il suffit d'additionner toutes les fréquences des cotes inférieures ou égales à la cote dont il est question. Pour la case grisée, on a additionné toutes les fréquences des cotes inférieures ou égales à 4 sur 10 (10 + 0 + 5 + 5 + 5 = 25 %). La dernière valeur de la colonne des fréquences cumulées doit toujours correspondre à 100 %.
Case grisée « 35 » des fréquences cumulées : « 35 % des élèves de la classe ont obtenu une cote inférieure ou égale à 5 sur 10 lors de l'interrogation surprise ». Et donc, par déduction : « 75 % des élèves de la classe ont obtenu une cote strictement supérieure à 4 sur 10 lors de cette même interrogation ».

Les effectifs cumulés et les fréquences cumulées permettent une **lecture du tableau** qui peut parfois s'avérer intéressante :

Exemple 4 : Pour une variable qualitative continue

Etude statistique

Dans un immeuble, voici les surfaces habitables en m^2 de chaque appartement mesurées avec une valeur approchée à $1 m^2$ près :

30, 148, 125, 120, 91, 70, 71, 50, 50, 80, 99, 102, 62, 40, 40, 61, 59, 49, 44, 36, 60, 85, 76, 57, 35, 31, 92, 105, 85, 50, 30, 110, 131, 95, 75, 98, 58, 36, 62, 88, 63, 64, 69, 49, 45, 68, 69, 48, 47, 69

Population :

Individu :

Caractère observé :

Modalités :

Ici, comme il y a beaucoup d'informations différentes, nous allons regrouper les données en **classes** ou intervalles. En principe, les classes peuvent être de largeurs différentes mais, par facilité, nous travaillerons avec des classes de même largeur.

Mesure de la surface (en m^2)	[30,50[[50,70[[70,90[[90,110[[110,130[[130,150[
Nombre d'appartements						

Nous allons représenter chaque intervalle (ou modalité) par sa valeur centrale. Le **centre d'une classe** est la moyenne arithmétique des extrémités de la classe, par exemple, le centre de la classe [30,50[est $\frac{30+50}{2} = 40$.

Représentation des données sous forme de tableau

Classes (C_i)	Centres (c_i)	Effectifs (n_i)	Fréquences f_i	Effectifs cumulés (N_i)	Fréquences cumulées (F_i)
[30,50[
[50,70[
[70,90[
[90,110[
[110,130[
[130,150[
		$n =$			

3.2. Représentation des données sous forme de graphique

Une représentation graphique permet souvent de visualiser très rapidement un certain nombre d'informations et d'avoir une vue globale de l'étude.

La représentation graphique des résultats peut se faire de différentes manières en fonction du type de variable et des renseignements souhaités.

3.2.1. Pour une variable qualitative nominale

Diagramme circulaire

Ce diagramme est constitué de secteurs circulaires (quartiers de tartes) de telle sorte que l'aire de chaque secteur soit proportionnelle à l'effectif (ou la fréquence) de la modalité correspondante. L'angle au centre pour chacun des secteurs est de $f_i \times 360^\circ$.

Ce diagramme n'impose pas d'ordre sur les modalités. Il est donc idéal dans le cas d'une variable qualitative.

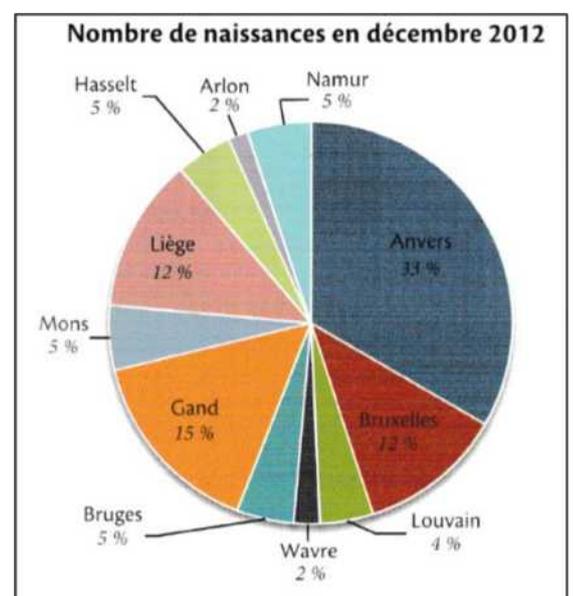
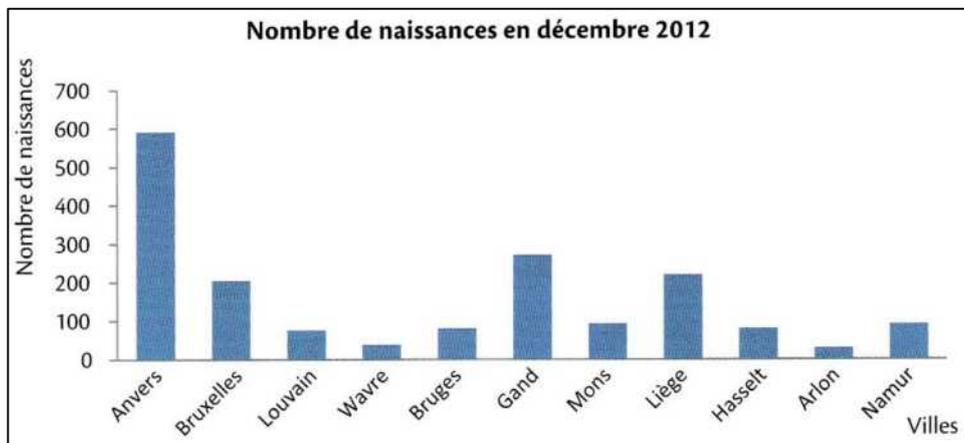


Diagramme en bâtonnets des effectifs ou des fréquences

Chacune des modalités est représentée par un rectangle de même base, de hauteur proportionnelle à son effectif (ou sa fréquence).



3.2.2. Pour une variable quantitative discrète

Diagramme en bâtonnets des effectifs ou des fréquences

Ce diagramme se réalise dans un repère. En abscisse, nous portons les différentes valeurs du caractère, en ordonnée, l'effectif (ou la fréquence) de chaque modalité.

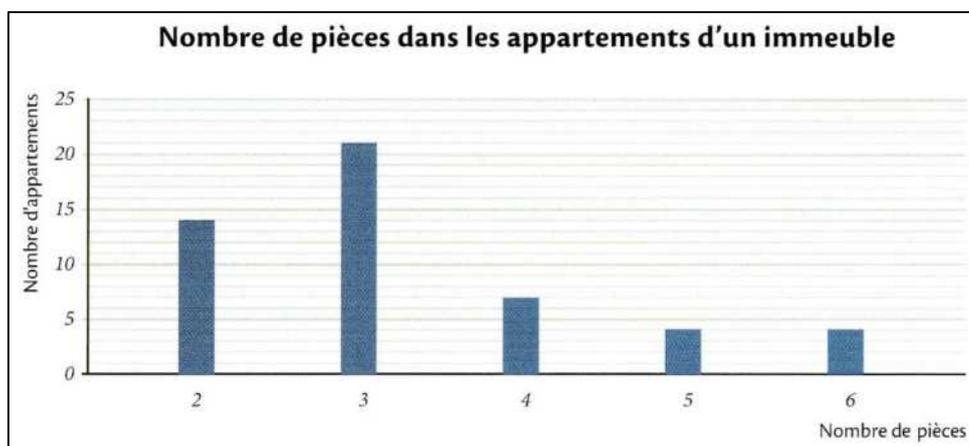
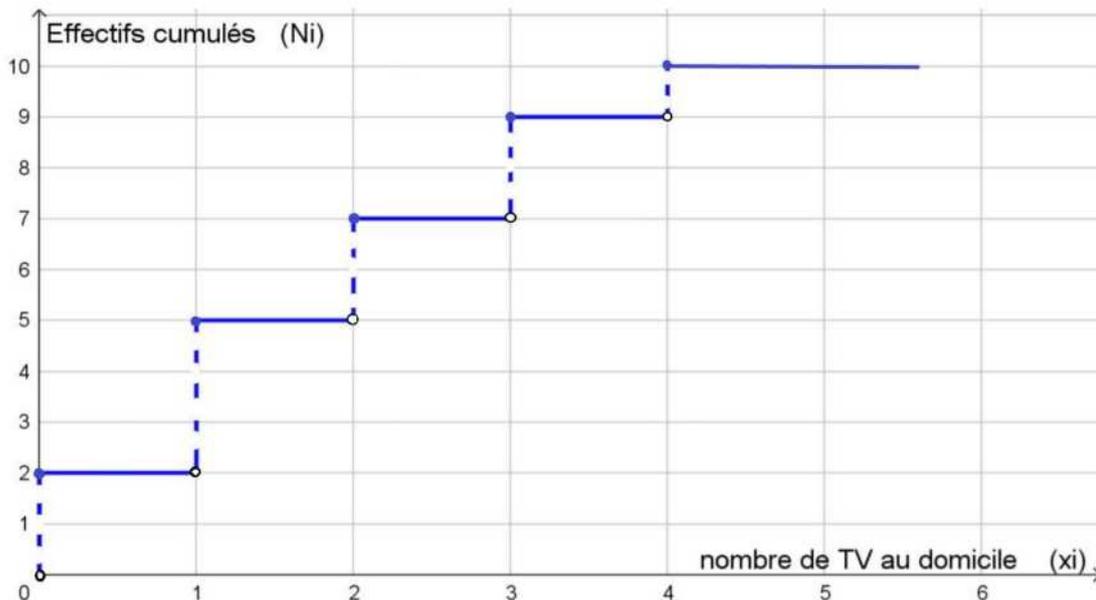


Diagramme en escaliers des effectifs ou des fréquences cumulé(e)s

On peut aussi représenter les effectifs (ou les fréquences) cumulé(e)s dans un diagramme. Lorsqu'on étudie un caractère discret, le diagramme des effectifs (ou des fréquences) cumulé(e)s se nomme **diagramme en escaliers**.

On porte les modalités en abscisses et les effectifs (ou fréquences) cumulé(e)s (N_i ou F_i) en ordonnées. On représente les points de coordonnées $(x_i; N_i)$ ou $(x_i; F_i)$ et on les relie entre eux par un segment horizontal puis vertical, ce qui fait apparaître un escalier. On relie le premier point à l'axe des x par un segment vertical.

Voici un exemple dans le cas de l'étude du nombre de TV chez 10 membres du personnel d'une entreprise :



Le digramme en escaliers permet de déterminer la médiane et les quartiles (voir plus loin).

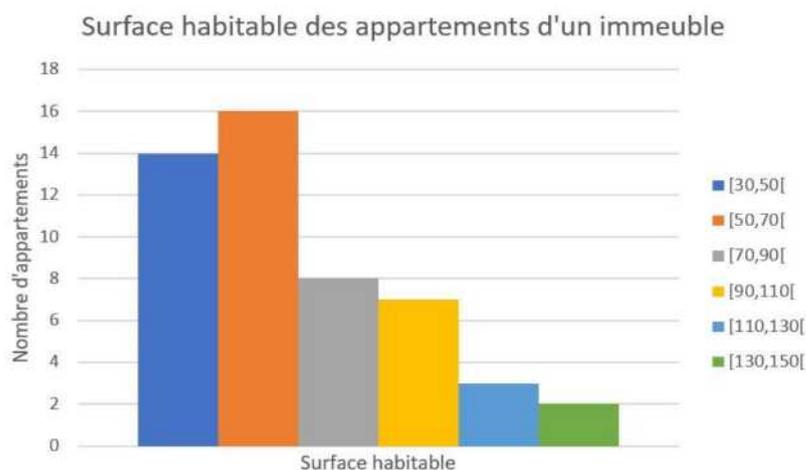
3.2.3. Pour une variable quantitative continue

Histogramme et ogive des effectifs (ou fréquences) cumulé(e)s

Lorsque le caractère quantitatif est continu et que les modalités sont regroupées en classes, la série statistique est représentée par un histogramme.

Celui-ci se représente dans un repère. Chaque classe est associée à un rectangle dont la base est délimitée par les extrémités de la classe. L'aire de chaque rectangle est proportionnelle à l'effectif (ou la fréquence) de la classe. Si toutes les classes ont la même amplitude, la hauteur de chaque rectangle est proportionnelle à l'effectif (ou à la fréquence) de la classe.

Voici l'histogramme représentant les données de l'exemple 4 :

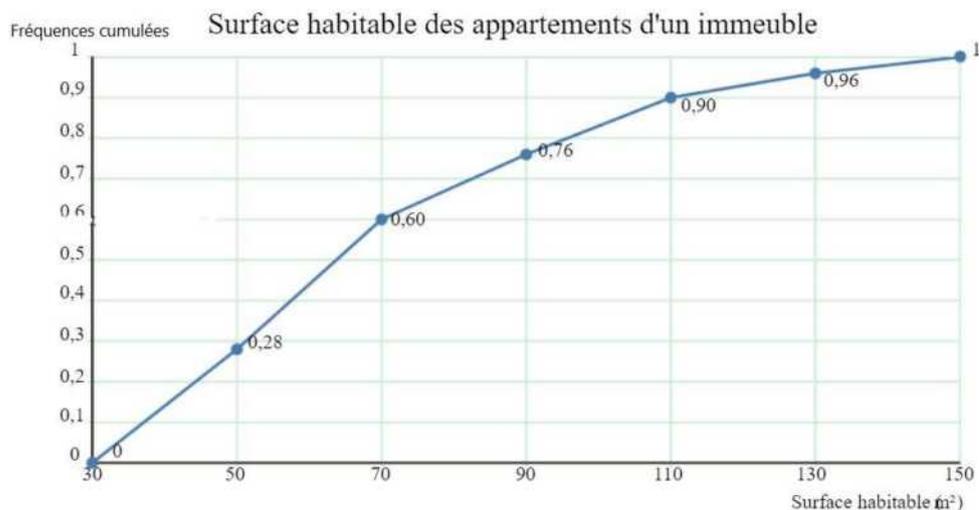


Remarque : Dans le cas où les largeurs de classes ne sont pas constantes, la hauteur des rectangles se calcule comme suit : hauteur = $\frac{n_i}{l_i}$ où l_i = largeur de la classe C_i .

Dans le cas continu, on peut aussi représenter les effectifs (ou fréquences) cumulé(e)s. Cette fois-ci, on ne tracera plus des escaliers mais des segments de droite.

On porte les classes en abscisses et les effectifs (ou fréquences) cumulé(e)s (N_i ou F_i) en ordonnées. On représente les points de coordonnées (borne supérieure de la classe, N_i) ou (borne supérieure de la classe, F_i) et on les relie entre eux par des segments de droites.

Voici l'ogive des fréquences cumulées de l'exemple 4 :



4. PARAMETRES DE POSITION

Pour résumer l'information contenue dans un ensemble de données statistiques, nous allons établir des paramètres de position. Ceux-ci sont soit des paramètres de valeur centrale (mode, moyenne ou médiane) ou des paramètres qui divisent la série en groupes de même effectif (quartiles, déciles, percentiles,...).

Vidéo à regarder : <https://www.youtube.com/watch?v=ulx2xvdwll0>

4.1. Le mode

L'expression « être à la mode » donne une idée de ce que représente le « mode » en statistique :

Le mode M_0 d'une série statistique est la ou les modalité(s) du caractère observé ayant le plus grand effectif.

Pour l'exemple 1 (Naissance), $M_0 =$ _____ car _____

Pour l'exemple 3 (Test), $M_0 =$ _____ car _____

Lorsque la série est regroupée en classes, on ne parlera plus de mode mais de **classe modale** (classe ayant le plus grand effectif).

Pour l'exemple 4, $M_0 = [50,70[$, la surface de la majorité des appartements de l'immeuble est supérieure à 50 m² et strictement inférieure à 70 m².

Remarque : Lorsque la série statistique comprend deux modes, on dira qu'elle est bimodale. Si elle comprend plus de deux modes, on dira qu'elle est plurimodale.

On peut facilement repérer le(s) mode(s) (ou la(les) classe(s) modale(s)) dans les diagrammes en barres, en bâtonnets et dans les histogrammes. Il suffit de regarder la (ou les) colonne(s) la (les) plus haute(s).

4.2. La moyenne arithmétique

La moyenne est généralement la caractéristique qui représente le mieux le « centre » de la distribution de la série statistique. Elle se prête facilement au calcul algébrique et sa signification est très concrète. C'est pour cela qu'elle est la plus utilisée. La moyenne est obtenue en additionnant toutes les données et en divisant le résultat par le nombre de données. Attention, elle ne peut être calculée que pour des modalités quantitatives. En effet, on doit additionner des nombres pour calculer la moyenne.

Exemple : Reprenons l'exemple introductif sur les tests.

La moyenne arithmétique est :

$$\frac{10 + 8 + 0 + 6 + 5 + 9 + 9 + 2 + 4 + 7 + 7 + 1 + 5 + 8 + 7 + 4 + 9 + 7 + 6 + 7}{20} = \frac{121}{20} = 6,05$$

Il est plus simple de calculer cette moyenne en groupant les résultats par modalités :

$$\frac{0.1 + 1.1 + 2.1 + 3. + 4.2 + 5.2 + 6.2 + 7.5 + 8.2 + 9.3 + 10.1}{20} = 6,05$$

Au lieu d'additionner les données, on multiplie la modalité par son effectif (le nombre de fois où la modalité apparaît).

La **moyenne** \bar{x} d'une série statistique est la moyenne arithmétique des valeurs de la variable prises par tous les individus de l'échantillon.

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + \dots + x_k n_k}{N} = \sum_{i=1}^k \frac{x_i \cdot n_i}{N} \quad \text{où } N \text{ est l'effectif total et } k \text{ le nombre de modalités}$$

Pour calculer aisément cette moyenne, on peut rajouter une colonne $n_i \cdot x_i$ dans le tableau recensé.

Modalités	Effectifs	Fréquences	Effectifs cumulés	Fréquences cumulées	Produit ($n_i \cdot x_i$)
0	2	20	2	20	$(n_1 \cdot x_1) = 0.2 = 0$
1	3	30	5	50	$(n_2 \cdot x_2) = 1.3 = 3$
2	2	20	7	70	$(n_3 \cdot x_3) = 2.2 = 4$
3	2	20	9	90	$(n_4 \cdot x_4) = 3.2 = 6$
4	1	10	10	100	$(n_5 \cdot x_5) = 4.1 = 4$
	N=10				17

Donc, $\bar{x} = \frac{17}{10} = 1,7$

Remarque : La moyenne est un paramètre central qui est fortement influençable par des valeurs extrêmes. En effet, si une modalité est beaucoup plus petite ou beaucoup plus grande que les autres, la valeur de la moyenne va considérablement diminuer ou augmenter.

Remarque :

Pour une variable quantitative **continue** :

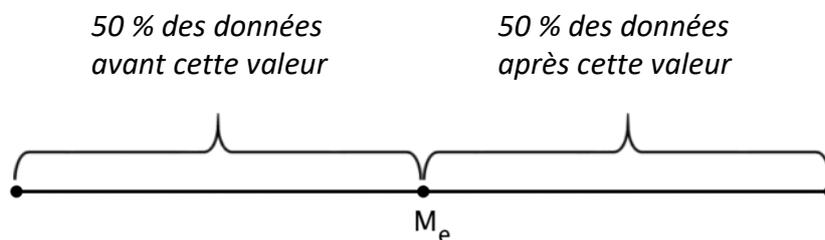
$$\bar{x} = \frac{c_1 n_1 + c_2 n_2 + \dots + c_k n_k}{N} = \sum_{i=1}^k \frac{c_i \cdot n_i}{N} \quad \text{où } N \text{ est l'effectif total et } k \text{ le nombre de classes de centre } c_i$$

Moyenne de l'exemple 4 :

4.3. La médiane (pour des variables quantitatives uniquement)

La médiane est un autre concept pour caractériser la tendance centrale d'une série statistique.

La **médiane** M_e d'une série statistique est la valeur de la variable qui partage cette série statistique préalablement classée par ordre croissant ou décroissant en deux groupes de même effectif.



Pour une variable quantitative **discrète** il suffit de classer les données en ordre croissant ou décroissant et de prendre l'élément du milieu.

- Si la série comprend un nombre impair d'éléments, il existe un élément qui occupe la place du milieu de la liste : c'est la médiane.

Ex : La série 12, 13, 13, **15**, 15, 16, 17 comprend 7 éléments ; sa médiane est le quatrième élément, c'est-à-dire $M_e = 15$.

- Si la série comprend un nombre pair d'éléments, la médiane est la moyenne arithmétique des deux éléments du milieu de la liste.

Ex : La série 12, 13, 13, **14**, **15**, 15, 16, 17 comprend 8 éléments ; sa médiane est la moyenne arithmétique entre le quatrième et le cinquième élément, c'est-à-dire $M_e = \frac{14+15}{2} = 14,5$.

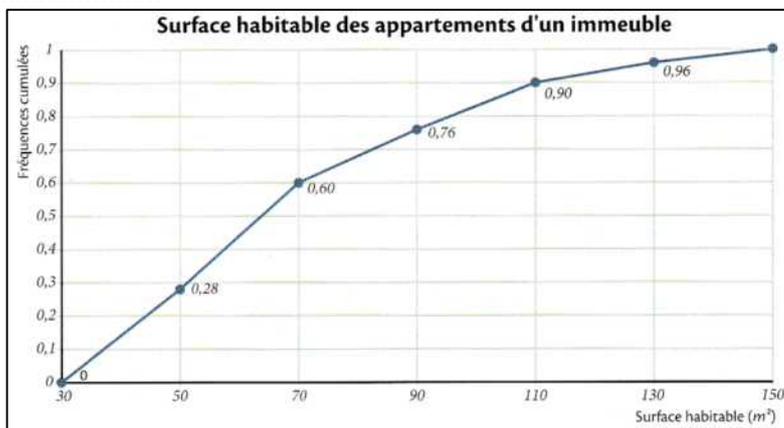
Exemple : Etude statistique p. 11 :

$M_e = \dots\dots\dots$

Interprétation :

Pour une variable quantitative **continue**, la détermination de la médiane se fait graphiquement. La médiane est l'abscisse du point d'intersection du polygone des effectifs cumulés et de la droite d'équation $y = \frac{n}{2}$. Sur le diagramme des fréquences cumulées, il faut considérer la droite $y = 0,5$.

Exemple : Etude statistique p. 13 :



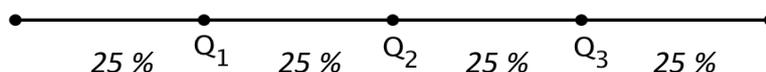
$M_e = \underline{\hspace{2cm}}$

Interprétation :

4.4. Les quartiles (pour des variables quantitatives uniquement)

Au lieu de séparer la série statistique en deux groupes de même effectif, on peut la partager en quatre groupe : **les quartiles**.

Les **quartiles** Q_1, Q_2, Q_3 d'une série statistique sont les valeurs de la variable qui partagent cette série statistique préalablement classée par ordre croissant ou décroissant en quatre groupes de même effectif.



Remarque : Il est évident que $Q_2 = M_e$

Un quart des observations sont donc inférieures à Q_1 , un quart des observations sont comprises entre Q_2 et Q_3 et le dernier quart des observations sont supérieures à Q_3 .

Les quartiles se déterminent avec les mêmes méthodes que la médiane après avoir ordonné la série statistique par ordre croissant ou décroissant.

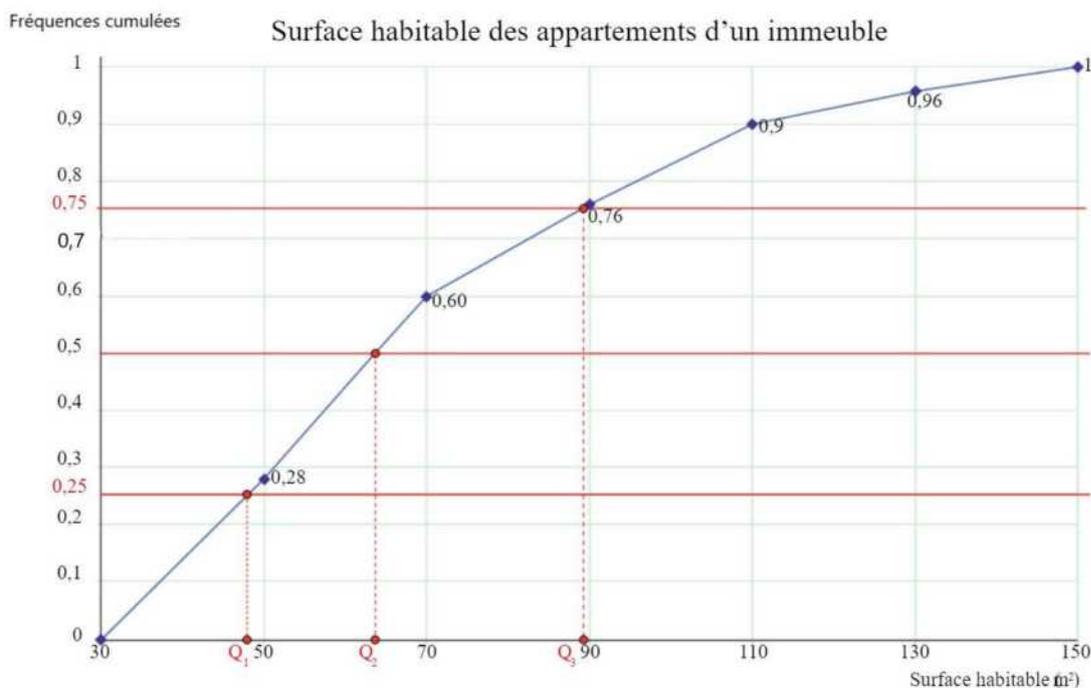
Pour une variable quantitative **continue**, la détermination des quartiles se fait graphiquement sur le diagramme des effectifs cumulés ou des fréquences cumulées.

Le premier quartile Q_1 est la valeur de l'abscisse du point dont l'ordonnée vaut 0,25 sur le diagramme des fréquences cumulées.

Le deuxième quartile Q_2 est la valeur de l'abscisse du point dont l'ordonnée vaut 0,5 sur le diagramme des fréquences cumulées (c'est la médiane).

Le troisième quartile Q_3 est la valeur de l'abscisse du point dont l'ordonnée vaut 0,75 sur le diagramme des fréquences cumulées.

Exemple : Etude statistique p. 13



$Q_1 =$ _____

$Q_2 =$ _____

$Q_3 =$ _____

Interprétation :

5. PARAMETRES DE DISPERSION (POUR DES VARIABLES QUANTITATIVES UNIQUEMENT)

Pour bien analyser une situation donnée, il faut tenir compte des paramètres de position mais également des paramètres de dispersion. Montrons par deux exemples que ces paramètres doivent être interprétés ensemble sous peine de mal interpréter la situation réelle.

Exemple 1 : La moyenne est influencée par toutes les valeurs et malheureusement très sensible aux valeurs extrêmes, au point d'en perdre parfois une bonne partie de sa représentativité, surtout sur des échantillons de petite taille.

Ainsi, la moyenne des 6 salaires mensuels suivants 1100€ - 1400€ - 1500€ - 1700€ - 2100€ - 12000€ est égale à 3300€ (!), alors qu'un seul salaire dépasse cette moyenne.

Exemple 2 : Prenons les résultats sur 20 de deux élèves lors de 12 tests

- élève 1 : 6, 7, 9, 9, 10, 10, 10, 10, 11, 11, 13, 14
- élève 2 : 1, 2, 2, 5, 6, 10, 10, 10, 16, 18, 20, 20.

Calculons le mode, la moyenne et la médiane des deux élèves.

Elève 1 : $\bar{x} =$
 $M_e =$
 $M_o =$

Elève 2 : $\bar{x} =$
 $M_e =$
 $M_o =$

Constatation :

Il faut donc compléter les informations précédentes en indiquant la façon dont la série s'étend autour des paramètres de position.

5.1. L'étendue

L'**étendue** E d'une série statistique est la différence entre la plus grande et la plus petite des valeurs observées.

Exemple : Reprenons les résultats des 12 tests des deux élèves

- Pour l'élève 1, $E = \dots \dots \dots$
- Pour l'élève 2, $E = \dots \dots \dots$

L'avantage de l'étendue est qu'elle est très facile à calculer et à interpréter. L'inconvénient réside dans le fait qu'elle est très influençable par une valeur extrême.

5.2. L'intervalle interquartile

L'intervalle interquartile d'une série statistique est l'intervalle $[Q_1, Q_3]$.

La différence entre le troisième et le premier quartile $Q_3 - Q_1$ est l'écart interquartile.

Au vu de la définition des quartiles, on peut en déduire que l'intervalle $[Q_1, Q_3]$ contient 50% des observations de la série statistique dont il est question. Le rôle de l'EIQ est donc le même que celui de l'étendue, cependant l'influence des valeurs extrêmes est réduite. On a donc une information concernant la moitié des valeurs les plus centrales de la série.

Dans l'exemple des 2 élèves, on obtient :

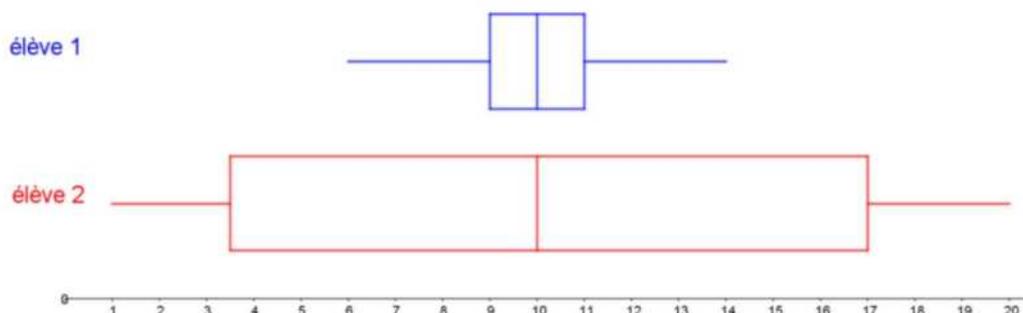
- Pour l'élève 1 : Intervalle interquartile = $[9;11]$ et l'écart interquartile = $11-9 = 2$.
- Pour l'élève 2 : Intervalle interquartile = $[3,5;17]$ et l'écart interquartile = $17-3,5 = 13,5$.

5.3. La boîte à moustaches

Pour représenter graphiquement la distribution d'une variable statistique, nous construisons un diagramme en boîte, appelé "**boîte à moustaches**", de la façon suivante :

- sur un axe où sont représentées les valeurs du caractère, nous plaçons les quartiles ainsi que les valeurs minimales et maximales de la variable ;
- un rectangle de largeur arbitraire et de longueur égale à l'écart interquartile est alors construit parallèlement à cet axe.

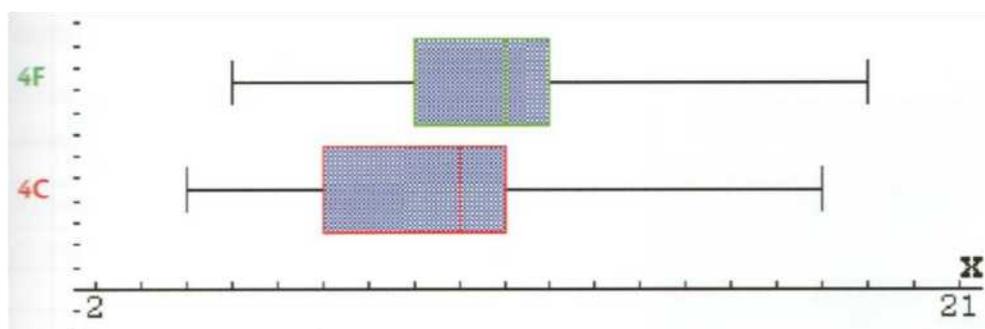
Dans l'exemple des 2 élèves, il est intéressant de comparer les 2 boîtes à moustaches.



On y retrouve l'égalité des médianes mais les nettes différences entre les étendues et les intervalles et écarts interquartiles.

Exercice :

D'après les informations contenues dans les boîtes à moustaches, précisez si les affirmations sont vraies ou fausses :



- 1) En 4C, le quart des élèves ont moins de 7.
- 2) En 4F, la moitié des élèves a entre 8 et 12.
- 3) En 4C, le niveau des élèves est plus homogène.
- 4) En 4F, il y a à peu près autant d'élèves qui ont entre 5 et 9 que d'élèves qui ont entre 12 et 19.
- 5) En 4C, la moyenne de la classe est de 10.

5.4. La variance et l'écart-type

La variance et l'écart type permettent de mesurer la « dispersion » des valeurs de la série autour de la moyenne. Si les valeurs de la série possèdent une unité, l'écart type s'exprime dans la même unité.

La **variance** V d'une série statistique est la moyenne arithmétique des carrés des écarts entre chaque observation et la moyenne.

$$V = \frac{n_1(x_1 - \bar{x})^2 + n_2(x_2 - \bar{x})^2 + \dots + n_k(x_k - \bar{x})^2}{n}$$

où les x_i sont remplacés par les c_i dans le cas d'une variable quantitative continue.

La variance peut aussi se calculer avec la formule plus pratique :

$$V = \frac{n_1x_1^2 + n_2x_2^2 + \dots + n_kx_k^2}{n} - \bar{x}^2$$

La variance est la caractéristique la plus utilisée pour mesurer la dispersion ou l'étalement des données autour de la moyenne. Pour des raisons de concordance des unités, nous calculerons l'écart-type qui, comme la moyenne, s'exprime dans la même unité que les données.

L'**écart-type** σ d'une série statistique est la racine carrée de la variance.

$$\sigma = \sqrt{V}$$

Une faible valeur de l'écart-type signifie qu'il y a une forte accumulation des observations autour de la moyenne. Une grande valeur de σ traduit un étalement considérable des observations autour de la moyenne.

Exemple : reprenons les résultats sur 20 de deux élèves lors de 12 tests

- élève 1 : 6, 7, 9, 9, 10, 10, 10, 10, 11, 11, 13, 14
- élève 2 : 1, 2, 2, 5, 6, 10, 10, 10, 16, 18, 20, 20.

Calculons la variance et l'écart-type des deux élèves.

Elève 1 : $V =$

$\sigma =$

Elève 2 : $V =$

$\sigma =$

Interprétation :

Pour réaliser ces calculs, on peut rajouter une colonne dans le tableau recensé avec $n_i \cdot x_i^2$, on divisera la somme des éléments de cette colonne par **N** et on retirera le carré de la moyenne du résultat obtenu (2^{ème} formule).

Pour l'exemple des 2 élèves :

Eleve 1						
x_i	n_i	f_i en %	N_i	F_i en %	$n_i x_i$	$n_i x_i^2$
6	1	8	1	8	6	36
7	1	8	2	16	7	49
9	2	17	4	33	18	162
10	4	33	8	66	40	400
11	2	17	10	83	22	242
13	1	8	11	91	13	169
14	1	8	12	99	14	196
	12	99			120	1254

$$V = \sigma^2 = \frac{1254}{12} - 10^2 = 4,5 \text{ d'où } \sigma = 2,12$$

Eleve 2						
x_i	n_i	f_i en %	N_i	F_i en %	$n_i x_i$	$n_i x_i^2$
1	1	8	1	8	1	1
2	2	17	3	25	4	8
5	1	8	4	33	5	25
6	1	8	5	41	6	36
10	3	25	8	66	30	300
16	1	8	9	74	16	256
18	1	8	10	82	18	324
20	2	17	12	99	40	800
	12	99			120	1750

$$V = \sigma^2 = \frac{1750}{12} - 10^2 = 45,83 \text{ d'où } \sigma = 6,77$$

Remarque : Compte tenu des arrondis, la somme des fréquences exprimées en % n'est pas égale à 100.

5.5. Inégalité de Tchebychev

Pour une série statistique de moyenne \bar{x} et d'écart-type σ et pour tout nombre réel $k > 0$, la proportion des données comprises dans l'intervalle $[\bar{x} - k\sigma, \bar{x} + k\sigma]$ est supérieure ou égale à $1 - \frac{1}{k^2}$.

Exemples :

- Supposons que la moyenne des notes obtenues à l'examen de mathématique en juin soit de 60 sur 100 avec un écart-type de 5.

Combien d'élèves ont obtenu une note entre 50 et 70 ?

Nous devons donc prendre $k = \dots$

-
- Quel sera l'intervalle dans lequel on retrouvera au moins 90% des résultats des élèves ?

On recherche donc la valeur de k telle que $1 - \frac{1}{k^2} = 0,9$

Nous pouvons donc conclure que 90% des notes obtenues par les élèves se situent dans l'intervalle :

6. EXERCICES – A REALISER DANS TON CAHIER D'EXERCICES

Exercice 1

Lors d'une récréation, on interroge les élèves d'une classe à propos du nombre de fruits qu'ils ont mangé la veille. On obtient les résultats suivants :

3	1	0	2	2	4	5	2	2
3	3	1	1	4	0	2	5	1
1	2	2	3	4	0	1	1	2

a) Complète le tableau suivant :

Modalités x_i	Effectifs n_i	Fréquences f_i en %	Effectifs cumulés N_i	Fréquences cumulées F_i en %		
0						
1		(2)				
2						
3			(3)			
4				(4)		
5	(1)					

b) Que vaut l'effectif total ? Quelle est la signification dans ce contexte ?

c) Explique la signification dans le contexte des nombres que tu as inscrits dans les cases grisées

- (1) _____
- (2) _____
- (3) _____
- (4) _____

d) Quel est le mode de la série statistique ? Que signifie ce résultat dans le contexte ?

e) Quelle est la médiane de la série statistique ? Explique. Que signifie-t-elle dans le contexte ?

f) Calcule la moyenne en expliquant et utilisant la 6^{ème} colonne du tableau (tu arrondiras au 100^{ème}). Que signifie ce résultat dans le contexte ?

g) Quel est le pourcentage de membres de la chorale ayant mangé au moins 4 fruits ? Explique le calcul effectué.

h) Trace le diagramme en bâtonnets des effectifs.

i) Trace le diagramme des fréquences cumulées, détermine graphiquement la valeur de la médiane et vérifie qu'elle est bien égale à la valeur trouvée au point e).

Détermine graphiquement Q_1 et Q_3 . Déduis-en l'intervalle interquartile, l'écart interquartile et la boîte à moustaches.

j) Dans un diagramme en secteurs de disque, quelle serait l'amplitude de l'angle au centre de la partie correspondant à 3 fruits ? Explique.

k) Quelle est l'étendue de la série ?

l) Calcule la variance en expliquant et utilisant la 7^{ème} colonne de droite du tableau (tu arrondiras au 100^{ème}). Que vaut l'écart-type de la série ?

Exercice 2

On relève le moyen de déplacement principal (correspondant au plus grand nombre de km parcourus) utilisé par les élèves d'une classe pour venir à l'école.

bus	train	bus	à pied	bus	voiture
bus	vélo	bus	voiture	à pied	voiture
voiture	bus	train	bus	bus	train
bus	voiture	bus	à pied	bus	vélo

- Dresse un tableau dans lequel tu placeras les modalités, les effectifs et les fréquences.
- Que vaut l'effectif total ? Quelle est sa signification dans le contexte ?
- Quel est le mode de cette série statistique ? Que signifie ce résultat dans le contexte ?
- Trace le diagramme en barres des fréquences (e) Trace le diagramme en secteurs des effectifs.

Exercice 3

Dans une classe de l'école, on demande à chaque élève le nombre de ses frères et sœurs. On obtient le tableau suivant :

0	2	1	2	0	1	3	2	1	2
1	0	1	1	2	3	5	1	0	2

- Quelle est la population étudiée ?
- Quel est le caractère étudié ? Est-il qualitatif ou quantitatif ?
- Quelles sont les modalités ?
- Trace le diagramme en bâtonnets des effectifs et le diagramme en escaliers des effectifs cumulés.
- Détermine le mode, la moyenne, la médiane, les quartiles, l'étendue et l'écart-type de la série.

Exercice 4

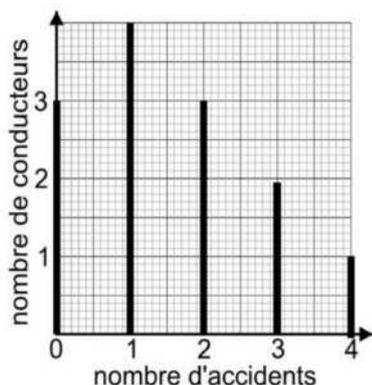
On note la couleur des véhicules garés sur le parking d'un centre commercial. On obtient le tableau suivant :

Couleur	blanc	bleu	gris	jaune	rouge	vert
Nombre	23	8	37	2	17	3

- La variable étudiée est-elle qualitative ou quantitative ?
- Trace le diagramme en barre des effectifs.
- Détermine, si possible, le mode, la médiane, la moyenne et l'écart-type de la série.

Exercice 5

On a recensé le nombre d'accidents de la route d'un ensemble de conducteurs de 19 ans.



- Dresse un tableau reprenant les modalités, les effectifs, les fréquences en %, les effectifs cumulés et une colonne te permettant de calculer la moyenne.
- Combien de personnes ont été interrogées ?
- Quel est le mode de la série ? Que signifie-t-il dans le contexte ?
- Combien de conducteurs ont fait moins de 3 accidents ?
- Quelle est, arrondie au 100ème, la moyenne du nombre d'accidents ?

Exercice 6

On souhaite étudier les écarts de temps qui séparent les coureurs du vainqueur au classement général d'un Tour de France. Voici la série statistique obtenue pour les coureurs figurant dans le classement général (excepté le vainqueur); le temps est donné à la minute près :

6 - 10 - 11 - 12 - 13 - 14 - 17 - 18 - 18 - 21 - 23 - 25 - 32 - 32 - 34 - 34 - 36 - 43 - 46 - 48 - 50 - 52 - 52 - 56 - 60 - 66
67 - 69 - 69 - 77 - 77 - 78 - 80 - 81 - 82 - 83 - 84 - 84 - 90 - 92 - 93 - 94 - 96 - 100 - 102 - 103 - 104 - 106 - 111
111 - 115 - 117 - 119 - 120 - 122 - 124 - 130 - 130 - 131 - 131 - 133 - 136 - 138 - 141 - 143 - 144 - 147 - 147
148 - 148 - 149 - 149 - 152 - 152 - 154 - 155 - 156 - 156 - 156 - 156 - 158 - 158 - 160 - 160 - 160 - 161 - 161
164 - 164 - 166 - 166 - 166 - 168 - 170 - 171 - 171 - 171 - 172 - 173 - 173 - 174 - 175 - 176 - 178 - 178 - 180
180 - 182 - 182 - 183 - 184 - 184 - 184 - 185 - 187 - 187 - 188 - 188 - 194 - 197 - 199 - 200 - 201 - 203 - 221 - 223 - 227

- Dresse le tableau recensé de cette série statistique.
- Dessine l'histogramme des effectifs et le diagramme des fréquences cumulées.
- Recherche les paramètres de position (mode, médiane, moyenne).
- Recherche les paramètres de dispersion (étendue, intervalle et écart interquartiles, variance et écart-type).
- Représente les valeurs extrêmes ainsi que les quartiles et la médiane sur une droite graduée (boîte à moustaches).

Exercice 7

On demande à 80 cyclistes le nombre de km qu'ils ont parcourus en un jour. Les résultats obtenus sont les suivants :

5	20	80	35	8	44	80	28	114	60
60	101	45	102	67	116	48	79	14	57
61	39	7	40	105	30	42	119	90	90
100	48	49	50	119	67	78	65	60	118
117	32	87	54	32	54	93	92	36	60
76	74	5	110	99	97	86	19	42	58
21	39	82	40	60	24	59	100	113	79
17	114	107	72	10	77	84	74	36	45

- Dresse le tableau recensé de cette série statistique.
- Dessine l'histogramme des effectifs et le diagramme des fréquences cumulées.
- Recherche les paramètres de position (mode, médiane, moyenne).
- Recherche les indices de dispersion (étendue, intervalle et écart interquartiles, variance, écart-type).
- Représente les valeurs extrêmes ainsi que les quartiles et la médiane sur une droite graduée (boîte à moustaches).