Statistique

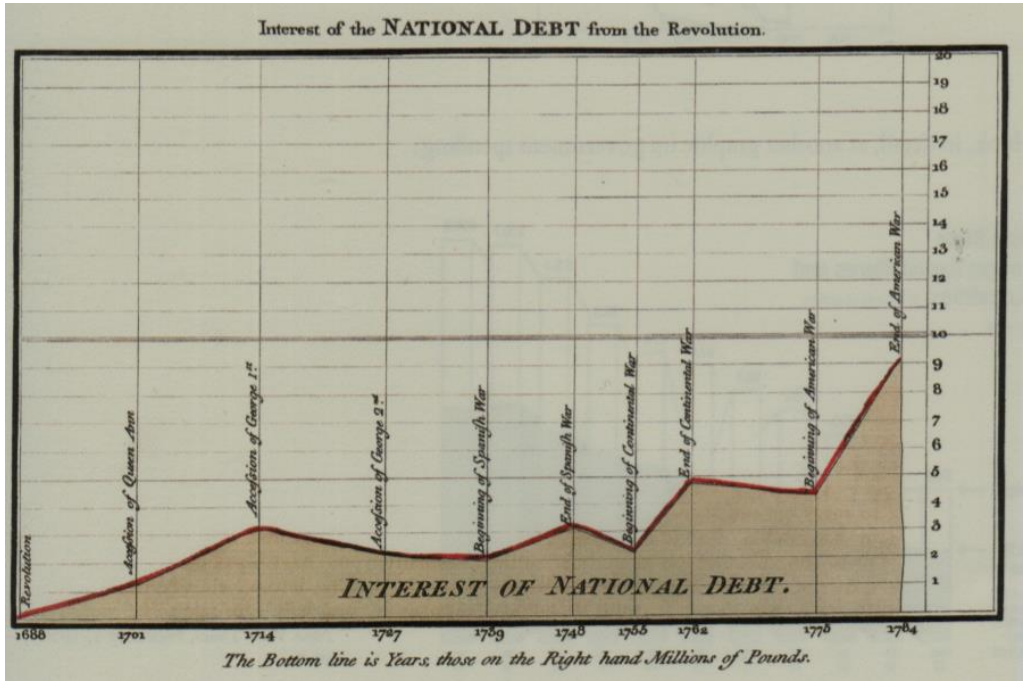
Introduction

La statistique s'est développée pour faire face à « l'avalanche de chiffres » que la réalité fournissait.

Les premiers paramètres de position utilisés furent le mode et le « milieu de l'intervalle défini par les valeurs extrêmes ». D'illustres savants contribuèrent au développement de la statistique :

- **Tycho Brahé** (moyenne arithmétique),
- **Roger Cotes** (moyenne pondérée),
- **Gauss** (écart-type)
- …

Le développement des graphiques statistiques furent l'œuvre de ingénieur et économiste écossais **William Playfair**.



Vocabulaire

- **Population :** tout ensemble d'objets de même nature.

Ces objets présentent tous un certain caractère qu'il s'agit d'étudier pour en révéler les tendances principales.

- **Caractère :** propriété que l'on étudie sur la population ; il peut être qualitatif ou quantitatif.

Exemples : durée de vie d'ampoules, dureté des eaux de distribution en Wallonie,…

- **Échantillon :** pour une population trop vaste pour être étudiée dans son ensemble, il faut prélever au hasard un échantillon de taille suffisante pour pouvoir tirer des conclusions sur la population totale

Vocabulaire

Un caractère peut être **discret** ou **continu** :

- Il est **discret** s'il ne peut prendre que des valeurs réelles isolées

Exemple : nombre d'enfants d'une famille,

- Il est **continu** s'il peut prendre toutes les valeurs réelles d'un certain intervalle

Une **série statistique** est un ensemble de valeurs collectées portées sur une liste

Exemple : nombre de voix obtenues par les candidats lors d'élections communales…

Démarche statistique

- **Collecte des informations :**

phase préparatoire et probablement la plus délicate : elle consiste à définir la population, choisir le caractère étudié, vérifier si l'échantillon éventuellement choisi est ou non représentatif de la population. C'est la qualité de cette collecte que dépendra la validité des résultats trouvés

- **Analyse des informations recueillies :**

déterminer un certain nombre de caractéristiques mathématiques relatives à la série statistique étudiée

- **Interprétation des résultats obtenus :**

tirer des conclusions sur l'étude faite et suggérer des décisions à prendre

Collecte de données

C’est le relevé méthodique de la valeur d’un caractère commun à tous les éléments d’une population

Pour chacun de ceux-ci il faudra toujours préciser :

- Les limites exactes de la population afin d’éviter les intrus et les absents
- Le caractère étudié
- Les conditions d’observation et de mesure afin de traiter tous les éléments de la même manière

Il faut éviter toute ambiguïté sans quoi les résultats perdent toute signification

Collecte de données

- Lorsque la population est relativement restreinte, il est possible de réunir des renseignements concernant tous les éléments.

- Par contre, lorsqu'elle est importante, il faudra se contenter d'étudier les éléments d'un sous-ensemble de la population, que l'on appelle échantillon

Ce sera souvent le cas…

1. Étude d’une série statistique : cas discret

Exemple 1 : Dans une classe de 26 élèves, la maîtresse a relevé les notes suivantes :

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

Nous avons ci-dessus une liste brute des points obtenus par les élèves

- Le caractère étudié est la note obtenue par chaque élève
- Le caractère est quantitatif discret

Afin d'y voir plus clair, on regroupe les notes dans un tableau...

- Il y a **7 observations** possibles (0, 1, 2, 3, 4, 5 et 6) ; on numérote de 1 à 7 ces observations dans la première colonne
- 0, 1, 2, 3, 4, 5 et 6 sont les **valeurs possibles x_i** des ces observations ; on notes ces valeurs dans la deuxième colonne

La différence entre les deux valeurs extrêmes s'appelle **l'étendue de la série**

On aura donc : $x_1 = 0$; $x_2 = 1$; … ; $x_6 = 5$; $x_7 = 6$

- Certaines valeurs de x_i reviennent plusieurs fois ; le nombre de fois que la valeur x_i apparaît est nommé **effectif n_i** de cette valeur. On note l'effectif pour chaque valeur dans la troisième colonne

	Notes	*Élèves*
Observations i	Valeurs x_i	Effectif n_i
1	0	1
2	1	1
3	2	2
4	3	5
5	4	7
6	5	8
7	6	2
		Effectif total : $n = \sum_{i=1}^{7} n_i = 26$

- L'effectif total n est le nombre d'éléments de la population, ici $n = 26$

Exercice 1 : avec les données du tableau ci-contre, calculez les expressions suivantes :

a. $\sum_{i=2}^{5} x_i$ **b.** $\sum_{k=1}^{6} n_k$ **c.** $\sum_{i=1}^{4} n_i x_i$

d. $\sum_{i=1}^{4} n_i \sum_{j=1}^{4} x_j$

- On peut également définir **l'effectif cumulé v_i de x_i** comme étant la somme des effectifs de toutes les valeurs de x_i inférieures ou égales à x_i.

Exemple : dans notre tableau, combien d'élèves ont obtenu une note inférieure à 3?

Il s'agit des cas où la valeur x_i est égale à 0, 1 ou 2. On note la réponse :

$$n_1 + n_2 + n_3 = \sum_{i=1}^{3} n_i$$

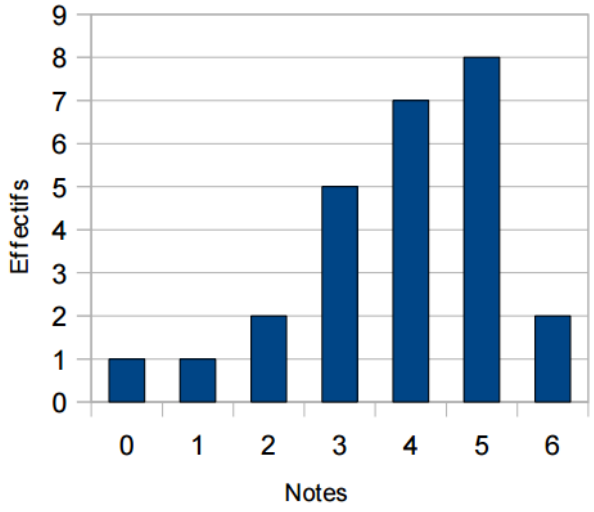
$$\boxed{v_3 = 4}$$

→ Donc dans 4 cas sur 26, les élèves sont en échec.

Représentation graphique

En statitique, les deux représentations graphiques les plus courantes sont l'**histogramme** (diagramme en bâtons) et le **diagramme à secteurs** (communément appelés « **camemberts** »).

Les deux graphiques suivants sont dessinés d'après les données présentées dans le tableau de l’exemple



Histogramme

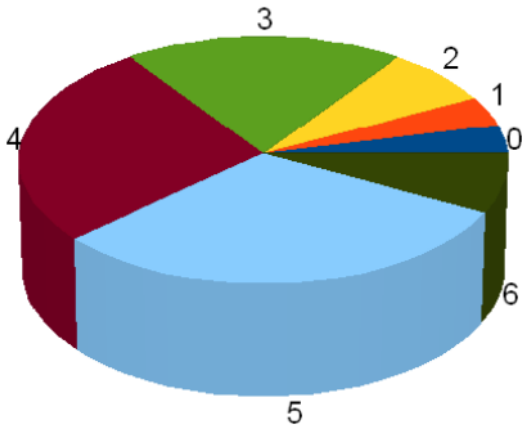


Diagramme à secteurs

Moyenne (mesure de tendance centrale)

La moyenne est la plus connue des mesures de tendance centrale. Elle s'obtient en divisant la somme des valeurs par le nombre de valeurs (n)

$$\bar{x} = \frac{\sum_{i=1}^{7} n_i x_i}{n}$$

En utilisant les données du tableau, on trouve :

$$\bar{x} = \frac{1 \cdot 0 + 1 \cdot 1 + 2 \cdot 2 + 5 \cdot 3 + 7 \cdot 4 + 8 \cdot 5 + 2 \cdot 6}{26} = \frac{100}{26} = 3{,}846$$

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

Variance et écart-type (mesure de dispersion)

Si l'on désire se faire une idée de la manière dont les valeurs du caractère s'écartent de la moyenne x de ce caractère, on calcule la moyenne des écarts quadratiques :

$$v = \frac{\sum n_i (x_i - \bar{x})^2}{n} = \frac{\sum n_i x_i^2}{n} - \bar{x}^2$$

v est la **variance** de l'échantillon. L'**écart-type** σ est la racine carrée de la variance

$$\sigma = \sqrt{v}$$

Variance et écart-type (mesure de dispersion)

En utilisant les données du tableau, on trouve :

$$\bar{x} = \frac{100}{26} = 3{,}846\ ; \qquad v = \frac{438}{26} - 3{,}846^2 = 16{,}846 - 14{,}793 = 2{,}053$$

$$\Rightarrow \sigma = \sqrt{v} = 1{,}433$$

Remarque : quand on calcule la variance d'un échantillon (et non de la population entière), le dénominateur est $n-1$.

Exercice : Les trois élèves suivants ont 4 de moyenne. Et pourtant, ils sont très différents. Calculez l'écart-type de leurs quatre notes. Que constatez-vous ?

a. 4 4 4 4 **b.** 2 2 6 6 **c.** 2 3 5 6

Médiane (mesure de tendance centrale)

On trie tout d'abord les n valeurs par ordre croissant :

0 1 2 2 3 3 3 3 3 4 4 4 **4 4** 4 4 5 5 5 5 5 5 5 5 6 6

La médiane est simplement la valeur qui se trouve au milieu :

$$\tilde{x} = x_{\frac{n+1}{2}}$$

Si n est pair, on prend la moyenne des deux valeurs du milieu :

$$\tilde{x} = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) = \frac{1}{2}(x_{13} + x_{14}) = \frac{4+4}{2} = 4$$

Remarque : La médiane n'est pas affectée par les valeurs extrêmes de la distribution.

Intervalle semi-interquartile (mesure de dispersion)

Méthode de calcul

1. Trier les données dans l'ordre croissant.
2. Diviser les données en deux groupes de taille égale : le groupe A avant la médiane et le groupe B après la médiane (si l'échantillon de départ a une taille impaire, rajouter la médiane en tête du groupe B).
3. Calculer la médiane du groupe A, que l'on appellera Q_1.
4. Calculer la médiane du groupe B, que l'on appellera Q_3.
5. L'intervalle semi-interquartile (isi) vaut : $$isi = \frac{Q_3 - Q_1}{2}$$

Intervalle semi-interquartile (mesure de dispersion)

Pour notre exemple :

Groupe A

0 1 2 2 3 3 3 **3** 3 4 4 4 4

$Q_1 = 3$

Groupe B

4 4 4 5 5 5 **5** 5 5 5 5 6 6

$Q_3 = 5$

$$isi = \frac{5-3}{2} = 1$$

Remarque : par convention $Q_2 = \tilde{x}$

Mode (mesure de tendance centrale)

Le mode est par définition la valeur la plus fréquente dans une série de données.

En lisant le tableau de notre exemple, on constate que le mode vaut 5

Remarque : Le mode n'est pas affecté par les valeurs extrêmes de la distribution. Selon la série de données, il peut y avoir plusieurs modes.

Exercices

1. Lors d'une journée, on a relevé les âges de 20 personnes venant se présenter à l'examen théorique du permis de conduire :

18 19 19 23 36 21 57 23 22 19
18 18 20 21 19 26 32 19 21 20

Calculez la moyenne, la médiane, le mode, la variance, l'écart-type et l'intervalle semi-interquartile de ces valeurs.

2. Au laboratoire de physique, une série de mesures de l'accélération de la pesanteur terrestre a donné les résultats suivants :

9,95 9,85 10,13 9,69 9,47 9,98 9,87 9,46 10,00

Calculez la moyenne et l'écart-type des résultats

Exercices

3. Le professeur de maths m'a dit : « C'est bien ; disons plutôt que c'est pas mal : tu as 4.5 de moyenne sur les cinq notes du semestre ». Sachant qu'aux quatre premières j'ai eu 5.2, 3.1, 4.4 et 4.2, quelle est ma note à la dernière épreuve ?

4. 41.250.000 personnes d'un pays ont atteint leur taille définitive (1,67 mètres en moyenne). Si l'on vous dit que, dans ce pays, la femme moyenne mesure 1,61 mètres et l'homme moyen 1,74 mètres, sauriez-vous en déduire de combien le nombre de femmes dépasse le nombre d'hommes dans ce pays ?

2. Étude d'une série statistique : cas continu

Lorsqu'il y a trop de valeurs discrètes, ou lorsque le caractère de la population est de nature continue, on regroupe les valeurs en classes de même amplitude.

Exemple 2 : Lors d'une course de vitesse, les 40 participants ont mis les temps ci-contre pour effectuer le parcours

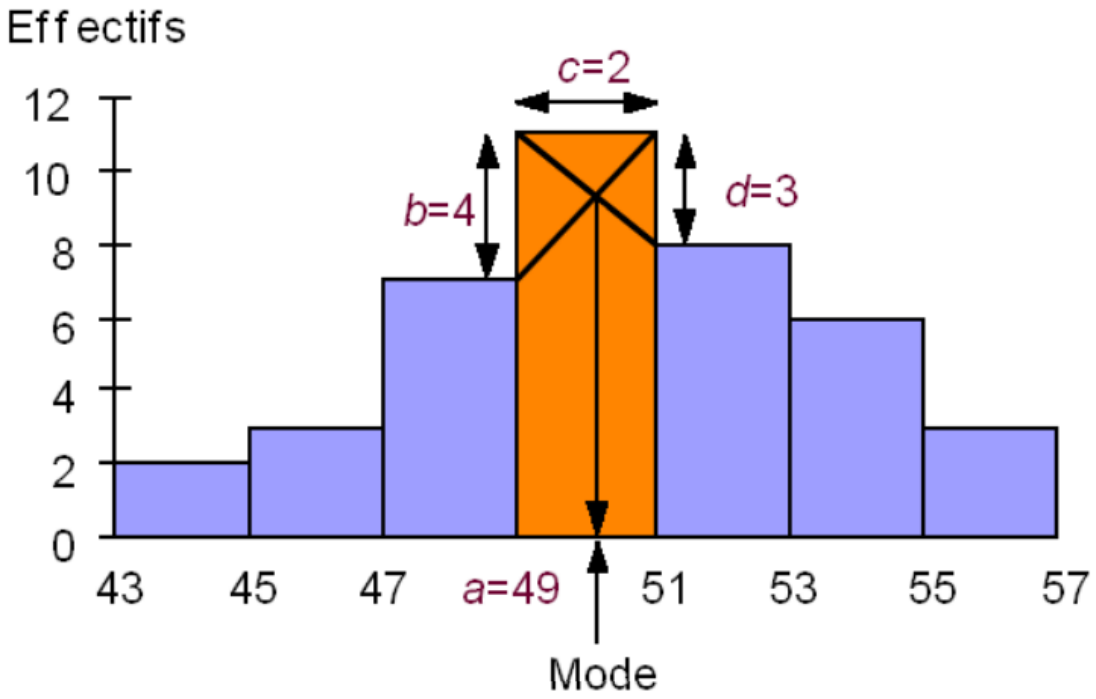
On représente ces données par un histogramme dans lequel chaque classe (ici d'amplitude 2) se voit attribuer un rectangle dont l'aire est proportionnelle à l'effectif de la classe.

Temps (*classes*)	Centres des classes x_i	Effectifs n_i
[43-45[	44	2
[45-47[	46	3
[47-49[	48	7
[49-51[	50	11
[51-53[	52	8
[53-55[	54	6
[55-57[	56	3
		$n = 40$

Mode

Dans le cas continu, le mode se trouve dans la classe ayant le plus grand effectif (la classe modale).

Il se calcule sur l'histogramme ainsi : $\mathbf{mode} = \mathrm{a} + \mathrm{c} \cdot \frac{b}{b+d}$



Ci-contre : $\mathbf{mode} = 49 + 2 \cdot \frac{4}{4+3}$

Fréquences et fréquences cumulées

Il est souvent intéressant de faire figurer dans un tableau statistique, pour chaque valeur (ou pour chaque classe) x_i que peut prendre le caractère, la proportion f_i des individus qui présentent cette valeur x_i. Ces proportions sont appelées **fréquences**.

Si n est l'effectif total, alors par définition $f_i = \frac{n_i}{n}$

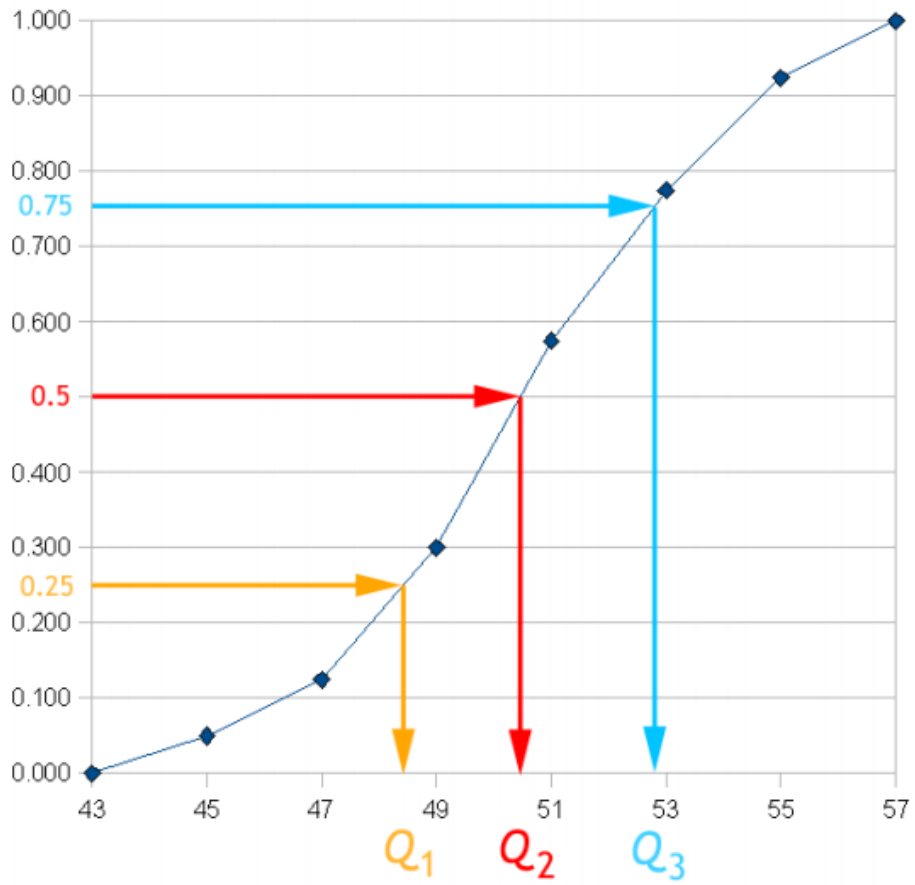
La fréquence cumulée $F(x)$ est la proportion des individus qui présentent des valeurs x_i inférieures ou égales à x. Elle se calcule en additionnant toutes les fréquences f_i correspondant aux x_i tels que $x_i \leq x$.

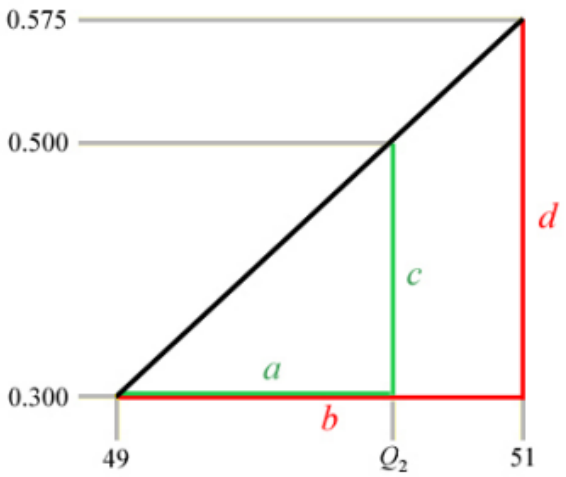
Si on reprend le tableau précédent :

Temps (*classes*)	Centres des classes x_i	Effectifs n_i	Fréquences f_i	Fréquences cumulées $F(x_i + 1)$
[43-45[	44	2	2/40 = 0,050	2/40 = 0,050
[45-47[	46	3	3/40 = 0,075	5/40 = 0,125
[47-49[	48	7	7/40 = 0,175	12/40 = 0,300
[49-51[	50	11	11/40 = 0,275	23/40 = 0,575
[51-53[	52	8	8/40 = 0,200	31/40 = 0,775
[53-55[	54	6	6/40 = 0,150	37/40 = 0,925
[55-57[	56	3	3/40 = 0,075	40/40 = 1,000
		$n = 40$	$\Sigma = 1$	

Polygone des fréquences cumulées

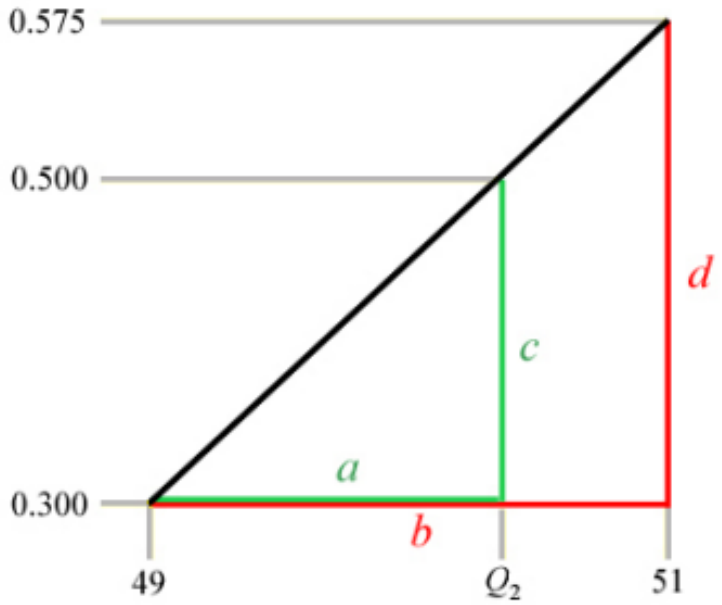
On porte en graphique les fréquences cumulée en fonction de la classe :





Médiane

La médiane se calcule en utilisant le polygone des fréquences cumulées. Il faut repérer quel segment coupe la droite horizontale d'ordonnée 0,5, puis calculer la médiane par proportionnalité (grâce au théorème de Thalès).



Dans l'exemple :

$$\frac{a}{b} = \frac{c}{d} \Rightarrow \frac{Q_2 - 49}{51 - 49} = \frac{0{,}5 - 0{,}3}{0{,}575 - 0{,}3}$$

$$\Rightarrow Q_2 = 49 + 2 \cdot \frac{0{,}2}{0{,}275} = 50{,}45$$

Intervalle semi-interquartile

F étant la fonction représentative du polygone des fréquences cumulées, on appelle respectivement premier, deuxième et troisième quartile les valeurs Q_1, Q_2 et Q_3 telles que

$$F(Q_1) = \frac{1}{4}; F(Q_2) = \frac{2}{4}; F(Q_3) = \frac{3}{4}$$

On voit que l'intervalle $[Q_1 ; Q_3]$ contient le 50% des valeurs de l'échantillon.

L'intervalle semi-interquartile est égal, par définition, à la moitié de la longueur de cet intervalle :

$$isi = \frac{Q_3 - Q_1}{2}$$

Moyenne et écart-type

Dans le cas continu, la moyenne et l'écart-type se calculent comme dans le cas discret en utilisant comme valeurs les centres de classes. Ces mesures changeront légèrement selon la manière dont on aura formé les classes.

Remarque : Si on utilise la moyenne pour mesurer la tendance centrale, on lui associera l'écart-type pour mesurer la dispersion. Si par contre on utilise la médiane, on lui associera l'intervalle semi-interquartile.

Exercices

Lors d'un contrôle de police sur l'autoroute, un agent a relevé les vitesses suivantes (arrondies à l'entier inférieur ou égal) :

117	134	130	113	127	125	98	110	124	122	126	101
106	121	121	104	124	117	109	128	134	146	111	139
123	124	130	123	120	133	111	143	145	111	110	119
114	104	126	99	140	105	119	134	128	119	137	109
122	130	92	104	113	130	120	84	166	138	129	119

a. Regroupez les données par classes : [80-90[, [90-100[
b. Dessinez le diagramme à secteur correspondant
c. Calculez le mode la médiane et l’isi

Les salaires mensuels (en francs suisses) payés aux ouvriers d'une entreprise se répartissent comme suit :

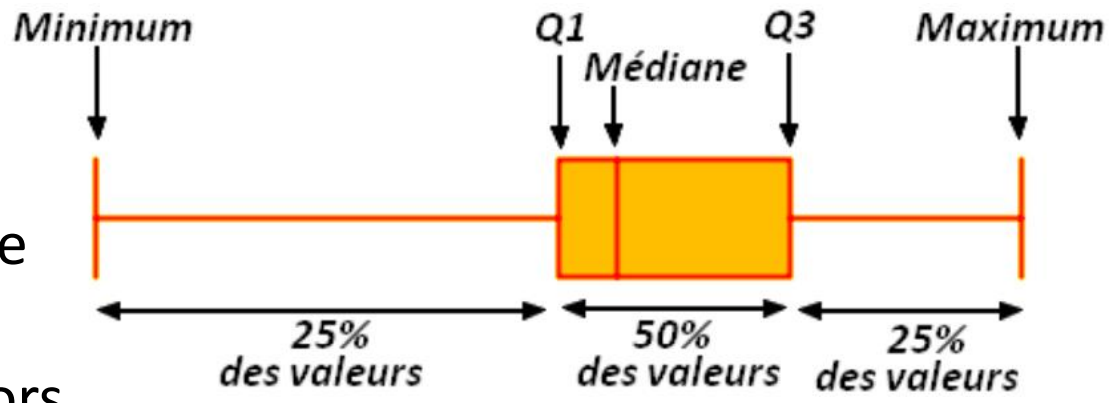
- 4 ouvriers gagnent entre 2400 et 2700 francs
- 21 ouvriers gagnent entre 2700 et 3000 francs
- 104 ouvriers gagnent entre 3000 et 3300 francs
- 163 ouvriers gagnent entre 3300 et 3600 francs
- 121 ouvriers gagnent entre 3600 et 3900 francs
- 57 ouvriers gagnent entre 3900 et 4200 francs
- 22 ouvriers gagnent entre 4200 et 4500 francs
- 10 ouvriers gagnent entre 4500 et 4800 francs

a. Faites un tableau en vous inspirant du tableau 3
b. Dessinez l'histogramme et le polygone des fréquences cumulées.
c. Calculez le mode, la médiane et l'intervalle semi-interquartile.
d. Calculez le salaire mensuel moyen et l'écart-type.

La boîte à moustache

Dans les représentations graphiques de données statistiques, la **boîte à moustaches** ou **diagramme en boîte** ou encore **diagramme à pattes** est un moyen rapide de figurer le profil essentiel d'une série statistique.

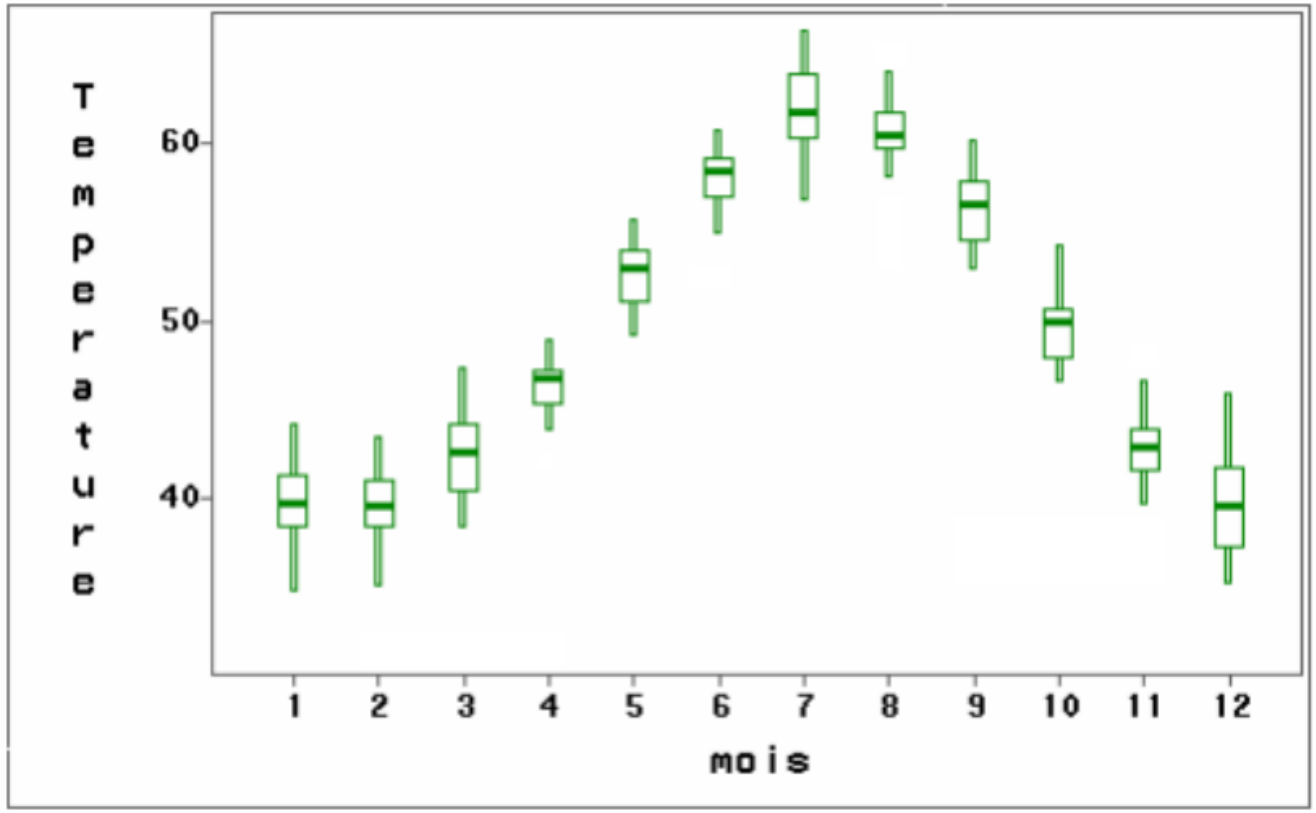
La boîte à moustaches résume certaines caractéristiques de position du caractère étudié. Ce diagramme est utilisé par exemple pour comparer un même caractère dans deux populations de tailles différentes. Il s'agit d'un rectangle allant du premier quartile au troisième quartile et coupé par la médiane. Ce rectangle suffit pour le diagramme en boîte. On ajoute alors des segments aux extrémités menant jusqu'aux valeurs extrêmes.



La boîte à moustache

Ces boîtes à moustaches peuvent aussi être dessinées verticalement.

Exemple : Soit la série des températures mensuelles moyennes à Nottingham de 1920 à 1939. Ces données ont été regroupées par mois et représentées sous forme de boîtes à moustaches :



Autres moyennes : moyennes géométrique

$$\bar{x} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

On peut l'illustrer avec le cas suivant : si l'inflation d'un pays est de 5% la première année et de 15% la suivante, l'augmentation moyenne des prix se calcule grâce à la moyenne géométrique des coefficients multiplicateurs 1,05 et 1,15 soit une augmentation moyenne de 9,88%.

Moyenne pondérée

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

Les p_i sont les poids de chaque valeur.

Un prof qui donne différents poids à ses épreuves utilisera la moyenne pondérée