Chapitre 10

Statistiques descriptives

10.1. Vocabulaire

Une enquête statistique consiste à observer une certaine population (élèves d'une classe, personnes âgées de 20 à 60 ans dans une région donnée, familles dans une région donnée, exploitations agricoles, appartements, travailleurs, ...) et à déterminer la répartition d'un certain *caractère statistique* (note obtenue, taille, nombre d'enfants, superficie, nombre de pièces, secteur d'activité, ...) de cette population.

Lorsque le caractère statistique est un nombre (taille, note, nombre d'enfants, ...), on parle de *caractère quantitatif*. Quand ce caractère n'est pas chiffré (langue parlée, secteur d'activité, couleur, ...), on parle de *caractère qualitatif*.

Lorsqu'un caractère statistique quantitatif prend un nombre fini raisonnable de valeurs (note, nombre d'enfants, nombre de pièces, secteur d'activité, ...), le caractère statistique est *discret*.

Lorsqu'un caractère statistique quantitatif peut prendre des valeurs multiples (taille, superficie, salaire, ...), le caractère statistique est considéré comme *continu*. L'étude des caractères continus constitue un magnifique sujet invoquant les fonctions, mais dans ce cours, on va répartir les données dans des *classes* de même *amplitude*.

10.1.1. Exemples

Un vendeur de voitures d'occasion a écoulé 80 voitures au mois de janvier. L'étude de trois caractères statistiques a permis d'établir trois tableaux.

1. Pour le premier tableau, le caractère statistique étudié est la marque des voitures vendues. Il est qualitatif et discret.

Marque des voitures	A	B	C	Autres
Effectif	18	28	10	24

2. Quant au deuxième tableau, le caractère statistique étudié est la puissance des voitures (en chevaux vapeurs, notés CV). Il est quantitatif et discret.

Puissance en CV	50	70	90	110
Effectif	20	35	15	10

3. Pour le troisième tableau, le caractère statistique étudié est le prix de vente. Il est quantitatif et continu.

   Les intervalles $[3000; 5000[$, ... , $[9000; 11000[$ sont appelés les *classes* du caractère étudié.

Prix de vente [CHF]	[3000 ; 5000[	[5000 ; 7000[	[7000 ; 9000[	[9000 ; 11000[
Effectif	17	30	21	12

   L'*amplitude* de la classe $[3000; 5000[$ est donnée par la différence des bornes de l'intervalle : $5000 - 3000 = 2000$. Le *centre* ou la *valeur centrale* de cette classe est donnée par la moyenne des deux bornes : $\frac{3000+5000}{2} = 4000$.

10.1.2. Fréquences

La *fréquence* d'une valeur x d'un caractère est $f_x = \frac{n_x}{N}$, où n_x est l'effectif de la valeur x et N l'effectif total. Lorsque les valeurs du caractères sont dans des classes, alors n_x est égal aux nombres de valeurs se trouvant dans la classe qui contient la valeur x.

Exemples

On reprend les tableaux précédents relatifs à nos 80 voitures.

Marque	Effectif	Fréquences
A	18	$\frac{18}{80} = 0.225$
B	28	$\frac{28}{80} = 0.350$
C	10	$\frac{10}{80} = 0.125$
Autres	24	$\frac{24}{80} = 0.300$
Total	80	1

CV	Effectif	Fréquences
50	20	$\frac{20}{80} = 0.2500$
70	35	$\frac{35}{80} = 0.4375$
90	15	$\frac{15}{80} = 0.1875$
110	10	$\frac{10}{80} = 0.1250$
Total	80	1

Prix	Effectif	Fréquences
$[3000, 5000[$	17	$\frac{17}{80} = 0.2125$
$[5000, 7000[$	30	$\frac{30}{80} = 0.3750$
$[7000, 9000[$	21	$\frac{21}{80} = 0.2625$
$[9000, 11000[$	12	$\frac{12}{80} = 0.1500$
Total	80	1

Remarque

La somme des fréquences est toujours égale à 1. En effet, la somme des fréquences est égale à la somme des effectifs sur toutes les valeurs possibles divisé par l'effectif total.

10.2. Représentations graphiques

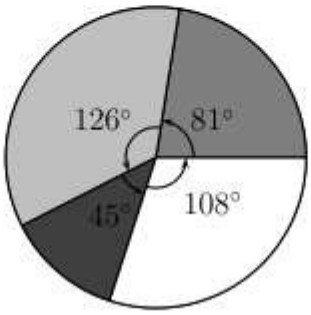
10.2.1. Diagramme à secteurs

Un *diagramme à secteurs* (ou camembert) est notamment utilisé pour représenter une série statistique de caractère qualitatif. L'aire d'un secteur (donc la mesure de l'angle au centre) est proportionnelle à l'effectif ou à la fréquence.

Reprenons le premier exemple relatif aux marques des voitures vendues. On partage un disque en secteurs dont l'angle au centre est proportionnel à l'effectif de la valeur correspondante du caractère. Le total des angles (360°) correspond à l'effectif total (80). Ainsi, pour la marque A :

$$\alpha_A = \underbrace{\frac{18}{80}}_{\text{fréquence}} \cdot 360^\circ = 81^\circ \quad \alpha_B = \underbrace{\frac{28}{80}}_{\text{fréquence}} \cdot 360^\circ = 126^\circ \quad \alpha_C = 45^\circ \quad \alpha_{\text{Autres}} = 108^\circ$$

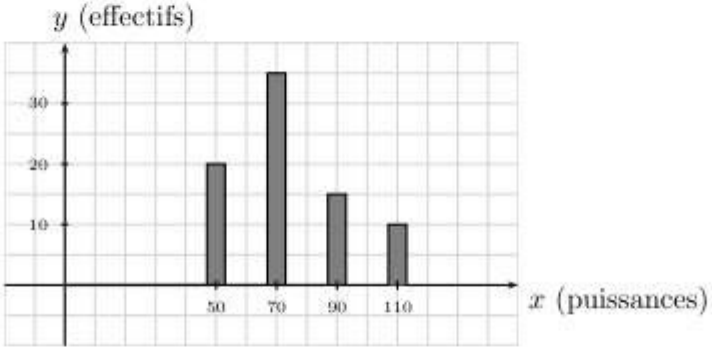
On trace les secteurs correspondants.



10.2.2. Diagramme en bâtons

Un *diagramme en bâtons* est notamment utilisé pour représenter une série statistique de caractère quantitatif discret. La hauteur d'un bâton est proportionnelle à l'effectif ou à la fréquence de la valeur du caractère.

Reprenons le deuxième exemple relatif aux puissances des voitures vendues. On trace des bâtons dont les hauteurs sont proportionnelles aux effectifs.

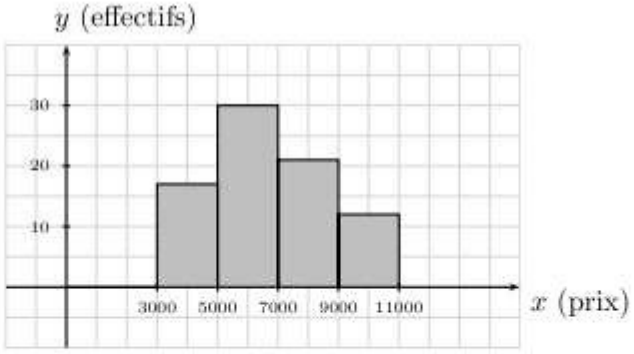


Remarque : Le caractère étudié étant discret, les bâtons ne doivent pas se toucher.

10.2.3. Histogramme

Un *histogramme* est notamment utilisé pour représenter une série statistique dont les valeurs sont regroupés en classe. L'aire d'un rectangle est proportionnelle à l'effectif ou à la fréquence de la classe.

Reprenons le troisième exemple relatif aux prix des voitures vendues. On trace des rectangles qui se touchent et dont les aires sont proportionnelles aux effectifs.



Remarque : lorsqu'on dessine un histogramme, les rectangles se touchent.

10.3. Effectifs et fréquences cumulés

Reprenons le deuxième exemple relatif aux puissances des voitures vendues et construisons le tableau des effectifs cumulés et des fréquences cumulées.

i	Puissance en CV x_i	Effectif n_i	ECC	ECD	Fréquence f_i	FCC	FCD
1	50	20	20	80	0.2500	0.2500	1.0000
2	70	35	55	60	0.4375	0.6875	0.7500
3	90	15	70	25	0.1875	0.8750	0.3125
4	110	10	80	10	0.1250	1.0000	0.1250

ECC signifie *effectif cumulé croissant* alors que ECD signifie *effectif cumulé décroissant*. L'effectif cumulé croissant donne le nombre de voitures dont la puissance est inférieure ou égale à x_i alors que l'effectif cumulé décroissant donne le nombre de voitures dont la puissance est supérieure ou égale à x_i.

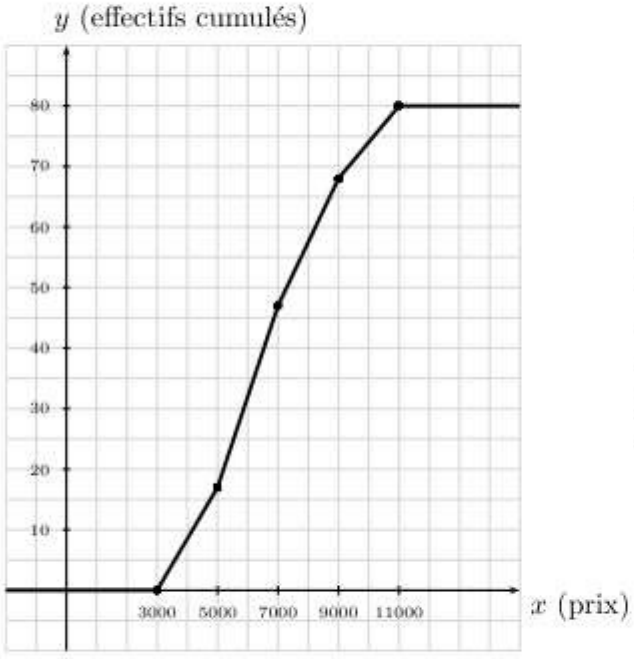
De même, FCC signifie *fréquence cumulée croissante* et donne la proportion de voitures dont la puissance est inférieure ou égale à x_i. FCD signifie *fréquence cumulée décroissante* et donne la proportion de voitures dont la puissance est supérieure ou égale à x_i.

10.3.1. Le polygone des effectifs cumulés croissants

Lorsque le caractère statistique est considéré comme continu, on peut parler du *polygone des effectifs cumulés croissants*. C'est un polygone formé de segments de droite. Chaque extrémité de segment (sauf le premier) a pour abscisse la borne supérieure d'une classe et pour ordonnée l'effectif cumulé croissant correspondant.

Exemple

Reprenons notre exemple de prix d'une voiture parmi les 80 de notre vendeur.



Prix	Effectifs	ECC
$[3000, 5000[$	17	17
$[5000, 7000[$	30	47
$[7000, 9000[$	21	68
$[9000, 11000[$	12	80
Total	80	

10.4. Paramètres de position

On cherche à caractériser chaque série statistique par un seul nombre, pour donner un ordre de grandeur des valeurs de la série. Ces valeurs sont des *paramètres de position* des séries.
Il est important de préciser que les paramètres calculés à partir de données réparties dans des classes (cas continu) dépendent du choix des classes et n'ont un sens que si on suppose que dans chaque classe les données sont réparties de manière uniforme.

10.4.1. Le mode

Le mode d'une série statistique discrète

Le *mode* d'une série statistique discrète est la valeur du caractère qui a le plus grand effectif. Il est possible qu'une série statistique admette plusieurs modes (en cas d'égalité).

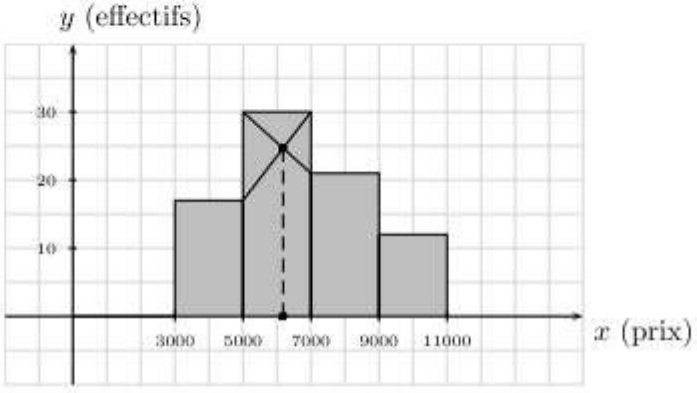
Exemple

Reprenons le cas des puissances de voitures. Il y a 35 voitures d'une puissance de 7 CV. Donc, le mode est de 7 CV.

Le mode d'une série statistique continue

La *classe modale* d'une série statistique continue est la classe qui contient le plus grand effectif. Il est possible qu'une série statistique admette plusieurs classes modales (en cas d'égalité).
Le *mode* est contenu dans chaque classe modale. Deux procédés permettent de déterminer le mode.

1. Il est possible de considérer le milieu de la classe modale.
2. On peut aussi utiliser les classes adjacentes, en trouvant le mode de manière graphique comme suit.



Exemple

Dans le cas des prix de voitures, on peut voir que 30 voitures coûtent entre 5000 et 7000 francs. Donc, la classe modale est l'intervalle [5000; 7000[. Selon le premier procédé, le mode est de 6000 francs. Selon le deuxième procédé, le mode est d'environ 6180 francs (on peut bien sûr calculer la valeur exacte de ce mode en cherchant la première coordonnée du point d'intersection des deux bouts de droites dessinés ou en utilisant un argument sur les triangles semblables).

10.4.2. La médiane

La médiane d'une série statistique discrète

La *médiane* d'une série est la valeur du caractère telle que le nombre de valeurs qui lui sont inférieures est égal au nombre de valeurs qui lui sont supérieures.

Exemple

Reprenons le cas des puissances de voitures. Il y a 80 voitures. La médiane correspond à 70 CV. En effet, il y a 40 voitures qui ont 70 CV ou moins et 40 voitures qui ont 70 CV ou plus.

Cas délicat

Ce cas ne se produit que lorsque le nombre d'observation est pair. Dans ce cas, la médiane est la valeur moyenne des deux extrémités autour du milieu (pour 80 voitures, on calcule la moyenne entre le nombre de chevaux de la 40^e et de la 41^e voiture).

Supposons que les voitures soient réparties de la manière suivante.

Puissance en CV	50	70	90	110
Effectif	20	20	20	20

Dans ce cas les 40 premières voitures ont 70 CV ou moins et les 40 dernières ont 90 CV ou plus. La médiane est ainsi à 80 CV.

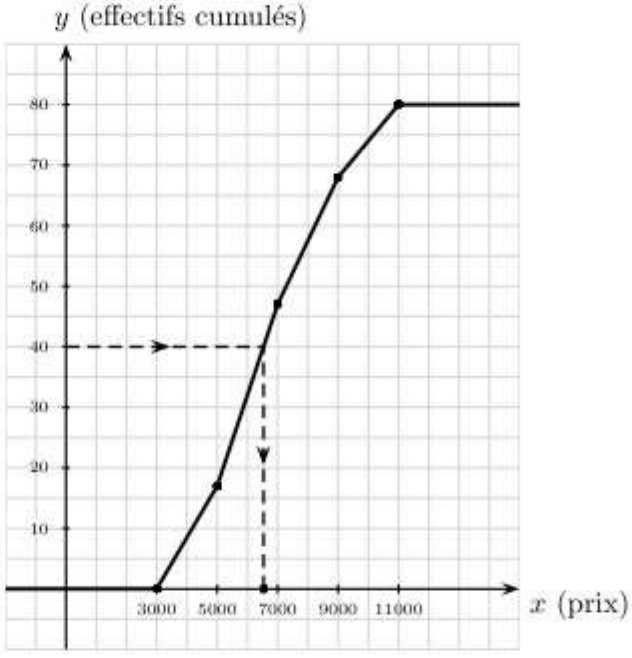
La médiane d'une série statistique continue

La médiane est définie de la même manière, mais la méthode de calcul change. Reprenons notre exemple de prix de voitures.

Vu qu'il y a un nombre pair de voitures, la médiane sera la moyenne de la valeur de la 40^e et de la 41^e voiture. Néanmoins, la plupart du temps, on se permet d'estimer la médiane en ne cherchant que la valeur correspondant à la 40^e voiture.

Graphiquement, on cherche la valeur sur l'axe des x qui correspond à un effectif cumulé de 40 à partir du polygone des effectifs cumulés croissants. On estime ainsi la médiane vers 6500 francs.

On peut être plus précis en raisonnant sur le segment de droite sur lequel se trouve la médiane.



En effet, sur ce segment de droite lorsqu'on monte de 30 (différence entre 17 et 47), on avance de 2000 (différence entre 5000 et 7000). Ainsi, lorsqu'on monte de 23 (différence entre 17 et 40), on avance de $2000 \cdot \frac{23}{30} \cong 1533.35$ francs.

La médiane est donc d'environ $5000 + 1533.35 = 6533.35$ francs.

10.4.3. La moyenne

La *moyenne arithmétique* d'une série statistique est :

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \cdots + n_p x_p}{N}$$

où p est le nombre de valeurs différentes possibles, n_i est l'effectif de la i-ième valeur x_i et N est l'effectif total. Dans le cas d'une série statistique continue, on choisit le centre de la i-ième classe pour x_i.

Exemples

Reprenons le cas des puissances et des prix de nos 80 voitures.

CV x_i	Effectif n_i	$n_i x_i$
50	20	1′000
70	35	2′450
90	15	1′350
110	10	1′100
Total	80	5′900

prix	x_i	Effectifs n_i	$n_i x_i$
[3000, 5000[	4'000	17	68′000
[5000, 7000[	6'000	30	180′000
[7000, 9000[	8'000	21	168′000
[9000, 11000[	10'000	12	120′000
Total		80	536′000

Ainsi la moyenne des puissances vaut $\bar{x} = \frac{5'900}{80} = 73.75$ CV.
La moyenne des prix vaut $\bar{x} = \frac{536'000}{80} = 6'700$ francs.

10.5. Paramètres de dispersion

On cherche à donner pour chaque série un nombre qui indique la dispersion des valeurs du caractère autour de la moyenne. Ces valeurs sont des *paramètres de dispersion* des séries.
Il est important de préciser que les paramètres calculés à partir de données réparties dans des classes (cas continu) dépendent du choix des classes et n'ont un sens que si on suppose que dans chaque classe les données sont réparties de manière uniforme.

10.5.1. L'étendue

L'*étendue* d'une série statistique est la différence entre la plus grande et la plus petite valeur de la série.

Exemples

1. Reprenons le cas des puissances de voitures. La plus petite valeur du caractère est 50, la plus grande est 110. L'étendue de la série est donnée par la différence, à savoir 60 CV.
2. Dans le cas des prix de voitures, l'étendue vaut $11'000 - 3'000 = 8'000$ CHF.

10.5.2. La variance et l'écart type

La variance

La *variance*, notée V, d'une série statistique est la moyenne des carrés des écarts à la moyenne, notée $\bar{x}$.

$$V = \frac{n_1(x_1 - \bar{x})^2 + n_2(x_2 - \bar{x})^2 + \cdots + n_p(x_p - \bar{x})^2}{N}$$

On peut aussi utiliser une formule équivalente plus facile à calculer.

$$V = \frac{n_1 x_1^2 + n_2 x_2^2 + \cdots + n_p x_p^2}{N} - \bar{x}^2$$

Preuve

On développe les identités remarquables et on simplifie.

$$\begin{aligned}
V &= \frac{n_1(x_1 - \bar{x})^2 + n_2(x_2 - \bar{x})^2 + \cdots + n_p(x_p - \bar{x})^2}{N} \\
&= \frac{n_1(x_1^2 - 2x_1\bar{x} + \bar{x}^2) + n_2(x_2^2 - 2x_2\bar{x} + \bar{x}^2) + \cdots + n_p(x_p^2 - 2x_p\bar{x} + \bar{x}^2)}{N} \\
&= \frac{n_1 x_1^2 + n_2 x_2^2 + \cdots + n_p x_p^2}{N} - \frac{2n_1 x_1\bar{x} + \cdots + 2n_p x_p\bar{x}}{N} + \frac{n_1\bar{x}^2 + \cdots + n_p\bar{x}^2}{N} \\
&= \frac{n_1 x_1^2 + n_2 x_2^2 + \cdots + n_p x_p^2}{N} - 2\bar{x} \cdot \underbrace{\frac{n_1 x_1 + \cdots + n_p x_p}{N}}_{\text{c'est la moyenne } \bar{x}} + \bar{x}^2 \cdot \underbrace{\frac{n_1 + \cdots + n_p}{N}}_{\text{ce terme vaut 1}} \\
&= \frac{n_1 x_1^2 + n_2 x_2^2 + \cdots + n_p x_p^2}{N} - 2\bar{x}^2 + \bar{x}^2 \;=\; \frac{n_1 x_1^2 + n_2 x_2^2 + \cdots + n_p x_p^2}{N} - \bar{x}^2
\end{aligned}$$

□

L'écart type

L'*écart type* σ est la racine carrée de la variance : $\sigma = \sqrt{V}$.

Exemples

Reprenons le cas des puissances et des prix de nos 80 voitures.

CV x_i	Effectif n_i	$n_i x_i$	$n_i x_i^2$
50	20	1′000	50′000
70	35	2′450	171′000
90	15	1′350	121′500
110	10	1′100	121′000
Total	80	5′900	464′000

prix	x_i	Effectifs n_i	$n_i x_i$	$n_i x_i^2$
$[3000, 5000[$	4′000	17	68′000	272′000′000
$[5000, 7000[$	6′000	30	180′000	1′080′000′000
$[7000, 9000[$	8′000	21	168′000	1′344′000′000
$[9000, 11000[$	10′000	12	120′000	1′200′000′000
Total		80	536′000	3′896′000′000

Ainsi la variance des puissances vaut $V = \frac{464'000}{80} - (\frac{5'900}{80})^2 \cong 360.9375$ CV^2. L'écart type vaut $\sigma \cong 18.9984$ CV.
La variance des prix vaut $V = \frac{3'896'000'000}{80} - (\frac{536'000}{80})^2 \cong 3'810'000$ CHF^2. L'écart type vaut $\sigma \cong 1951.92$ CHF.

10.6. Un problème d'examen et sa correction

10.6.1. Problème

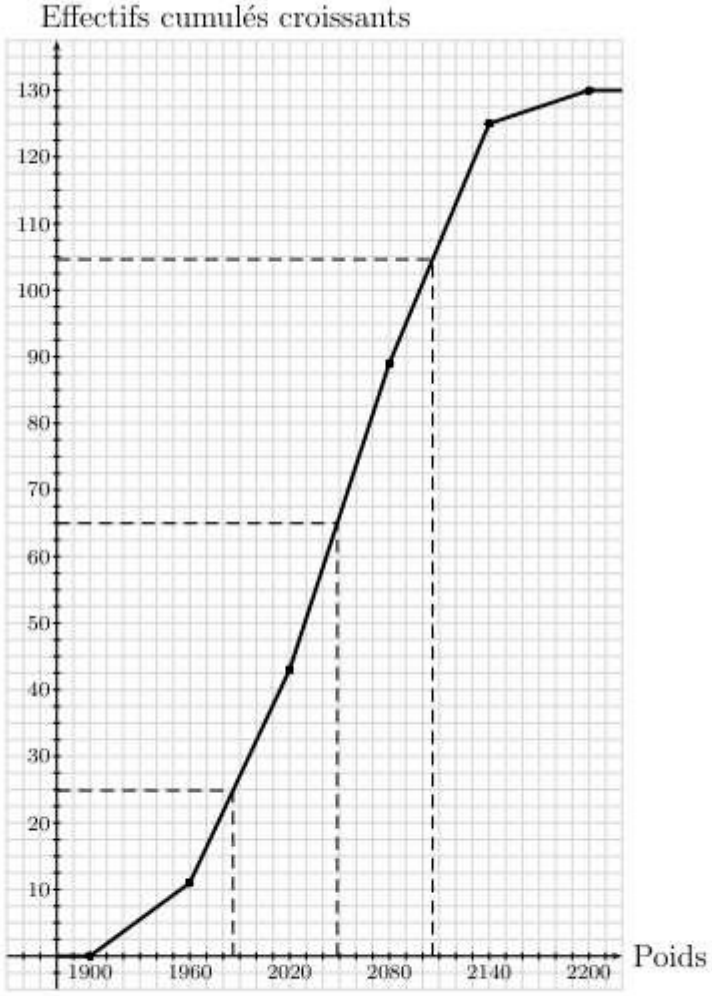
Une machine remplit automatiquement des paquets de café marqués 2 kg. On pèse les 130 paquets d'un carton prêt à être expédié. Les résultats suivants sont obtenus.

Poids en g	[1900 ;1960[	[1960 ;2020[	[2020 ;2080[	[2080 ;2140[	[2140 ;2200[
Effectif	11	32	46	36	5

1. Dessiner le polygone des effectifs cumulés croissants.
   Unités : 2 carreaux = 5 paquets pour les effectifs et 1 carreau = 10 g pour les poids.
2. Déterminer graphiquement la médiane de la distribution, puis calculer sa valeur exacte. Donner l'interprétation de la médiane dans la situation présente.
3. À l'aide de votre calculatrice, calculer la moyenne $\bar{x}$ et l'écart type σ de cette distribution (arrondir au gramme).
4. Utiliser le graphique pour déterminer le nombre de paquets qui ont un poids dans l'intervalle $I =]\bar{x} - \sigma; \bar{x} + \sigma[$.
5. La machine est considérée comme bien réglée si la moyenne est comprise entre 1980 g et 2050 g et si au moins 66% des paquets ont un poids dans l'intervalle I. Est-elle déréglée ?

Corrections

1. Voici le polygone des effectifs cumulés croissants.



2. On voit sur le polygone que la médiane vaut environ 2050.

   Sur le morceau de droite de la classe $[2020; 2080[$ (c'est la classe médiane), on voit que l'on monte de $89-43=46$ et que l'on avance horizontalement de $2080-2020=60$.

   La médiane correspondant à 65 paquets. Ainsi, on monte de 22 à partir de 43 (pour arriver à 65). Le facteur de proportionnalité est de

$$22 = 46 \cdot \underbrace{\frac{22}{46}}_{\text{facteur de proportionnalité}}$$

   Ainsi, si on monte de 22, on avance horizontalement de

$$60 \cdot \underbrace{\frac{22}{46}}_{\text{facteur de proportionnalité}} \cong 28.70 \text{ grammes}$$

   Donc la médiane se situe en environ $2020 + 28.70 = 2048.70$ grammes.

3. Voici le tableau permettant de calculer la moyenne, la variance et l'écart type.

classes	n_i	x_i	$n_i x_i$	$n_i x_i^2$
$[1900, 1960[$	11	1930	21′230	40′973′900
$[1960, 2020[$	32	1990	63′680	126′723′200
$[2020, 2080[$	46	2050	94′300	193′315′000
$[2080, 2140[$	36	2110	75′960	160′275′600
$[2140, 2200[$	5	2170	10′850	23′544′500
totaux	130		266′020	544′832′200

   Ainsi la moyenne vaut :

$$\bar{x} = \frac{11 \cdot 1930 + 32 \cdot 1990 + 46 \cdot 2050 + 36 \cdot 2110 + 5 \cdot 2170}{130} = \frac{266'020}{130} \cong 2046 \text{ g}$$

   Et la variance vaut :

$$\begin{aligned} V &= \frac{11 \cdot 1930^2 + 32 \cdot 1990^2 + 46 \cdot 2050^2 + 36 \cdot 2110^2 + 5 \cdot 2170^2}{130} - (\bar{x})^2 \\ &= \frac{544'832'200}{130} - \left(\frac{266'020}{130}\right)^2 \cong 3642 \text{ g}^2 \end{aligned}$$

   Finalement, l'écart type vaut $\sqrt{V} \cong 60$ grammes.

4. L'intervalle est $I =]\bar{x} - \sigma, \bar{x} + \sigma[=]1986, 2106[$. Les valeurs étant reportées sur le polygone, on voit qu'il y a environ 80 paquets dans cet intervalle.

5. La moyenne se trouve bien entre 1980 et 2050 grammes.
   Il y a donc 80 paquets dans l'intervalle I. Ce qui correspond à

$$\frac{80}{130} \cdot 100 \cong 61.54\%$$

   La condition n'est donc pas satisfaite.