Statistiques

Table des matières

I Activité 1 page 62

Pour Alix :

Épreuve	1	2	3	4	5	6	7
Coefficient	9	7	3	5	3	3	2
Note d'Alix	11,5	8	12	12	5	10	10
Produits	103,5	56	36	60	15	30	20

La somme des produits vaut 320,5.
La somme des coefficients est 32
Moyenne 10,015625
La moyenne d'Alix est $\overline{x} \approx \dfrac{320,5}{32} \approx \boxed{10,01}$.

Pour Emma :

Épreuve	1	2	3	4	5	6	7
Coefficient	7	9	3	5	3	3	2
Note d'Emma	12	8	12	12	5	10	10
Produits	84	72	36	60	15	30	20

La moyenne d'Emma pour l'école UCE est $\overline{x} = \dfrac{317}{32} \approx 9,91$; elle n'est pas reçue.

Pour UCA : $\overline{x} = \dfrac{325}{32} \approx 10,16$; elle est reçue.

II Vocabulaire

Le mot statistiques vient du mot latin status (état) .
On recueille des données et on essaye d'en tirer des renseignements exploitables.
On utilise le vocabulaire suivant pour décrire cette étude :

- **Série statistique** : ensemble des valeurs collectées.
- **Population** : ensemble sur lequel porte l'étude statistique.
- **Individus** : éléments qui composent la population.
- **échantillon** : partie de la population.
- **Caractère étudié** : propriété que l'on observe sur les individus. Les différentes valeurs obtenues sont appelées **valeurs du caractère** ou **modalités**, souvent notées x_1, x_2,… x_p dans le cas de valeurs discrètes.
  On distingue deux types de caractères.
  - ◇ Un caractère peut être **qualitatif** (situation de famille, sexe...).
  - ◇ Un caractère peut être **quantitatif**. Il est dit **discret** lorsqu'il ne prend que des valeurs isolées (nombre d'enfants, notes dans une classe...). Il est dit **continu** lorsqu'il peut prendre théoriquement toutes les valeurs d'un intervalle (taille, temps d'écoute...) ; dans ce cas, les valeurs sont regroupées en sous-intervalles appelés des **classes**.
- **Effectif** : pour une valeur du caractère (modalité ou classe), on appelle effectif le nombre d'individus de la population ayant cette valeur. On note souvent n_1, n_2,… n_p les effectifs respectifs des modalités x_1, x_2,… x_p.
- **Effectif total** : nombre total d'individus de la population (ou de l'échantillon). Il est égal à $n_1 + n_2 + \ldots + n_p$, souvent noté N.
- **Fréquence** : Pour une valeur du caractère (modalité ou classe), on appelle fréquence le quotient de l'effectif de cette valeur par l'effectif total. On note souvent f_1, f_2,… , f_p les fréquences respectives des modalités x_1, x_2,… x_p, donc :
  $$f_1 = \frac{n_1}{N},\ f_2 = \frac{n_2}{N}, \ldots,\ f_p = \frac{n_p}{N}.$$
  On en déduit que $0 \leq f_1 \leq 1$, $0 \leq f_2 \leq 1, \ldots, 0 \leq f_p \leq 1$, et $f_1 + f_2 + \ldots f_p = 1$.
- **Valeurs extrêmes** : valeurs minimales et maximales d'un caractère quantitatif.
- **Effectif cumulé** : Pour une valeur x d'une série statistique quantitative, l'effectif cumulé croissant (respectivement décroissant) de x est la somme des effectifs des valeurs inférieures (respectivement supérieures) ou égales à x.
- **Fréquence cumulée** : pour une valeur x d'une série statistique quantitative, la fréquence cumulée croissante (respectivement décroissante) de x est la somme des fréquences des valeurs inférieures (respectivement supérieures) ou égales à x.

III Variables discrètes (ndicateurs de position)

III.1 Mode

Définition

Le mode d'une série statistique est la valeur du caractère ayant l'effectif le plus grand

III.2 Moyenne (pondérée)

Définition

Soit une série statistique dont les valeurs du caractère sont $x_1, x_2, \cdots, x_k$ et $n_1, n_2, \cdots, c_k$ effectifs associés. La moyenne de la série statistique, notée $\bar{x}$, a pour valeur :

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \cdots + n_k x_k}{n_1 + n_2 + \cdots + n_k}$$

Remarque : on peut écrire cette formule de façon symbolique :

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

où la lettre grecque Σ (sigma) signifie somme.

Conséquence :

Lorsqu'on présente la série statistique en donnant la liste de toutes les valeurs, alors la moyenne est $\frac{x_1 + x_2 + \cdots + x_N}{N}$

Théorème

Si on appelle f_i la fréquence de la valeur x_i , alors :

$$\bar{x} = f_1 x_1 + f_2 x_2 + \ldots + f_k x_k.$$

Théorème (Linéarité de la moyenne)

Soit k un nombre réel. Soit $x_1, x_2, \cdots, x_n$ les valeurs du caractère d'une série statistique et $\bar{x}$ leur moyenne. Alors :

- la moyenne de la série $kx_1, kx_2, \cdots, kx_n$ est $k\bar{x}$;
- la moyenne de la série $x_1 + k, x_2 + k, \cdots, x_i + k, \cdots, x_n + k$ est $\bar{x} + k$.

Exemples :

- Si la moyenne au contrôle de biologie dans une classe est de 8 sur 20 et que le professeur décide d'augmenter toutes les notes de 10%, alors la nouvelle moyenne est de 8,8.
- Si la moyenne au contrôle d'histoire-géographie est de 8,7 sur 20 et que le professeur décide d'ajouter 1 point à tous les élèves, alors la nouvelle moyenne est de 9,7.

III.3 Médiane

Définition

La médiane d'une série statistique est le nombre tel que :
50 % au moins des individus ont une valeur du caractère inférieure ou égale à ce nombre et 50 % au moins des individus ont une valeur supérieure ou égale à ce nombre.

Médiane d'un caractère quantitatif discret

On considère une série statistique dont les valeurs du caractère sont rangées par ordre croissant, chacune de ces valeurs figurant un nombre de fois égal à son effectif.

- Si le nombre N de données est impair, donc de la forme $N = 2p + 1$, la médiane est le terme du milieu, c'est-à-dire le rang de terme $p + 1$.
- Si le nombre de données est pair, donc de la forme $N = 2p$, la médiane est la demi-somme des termes de rangs p et $p + 1$.

III.4 Quartiles

Définition

Le premier quartile d'une série statistique, noté Q_1 est la plus petite valeur de la série, rangée par ordre croissant, tel que 25 % des valeurs de la série soient inférieures ou égales à Q_1.
Le troisième quartile d'une série statistique, noté Q_3 est la première valeur de la série, rangée par ordre croissant, tel que 75 % des valeurs de la série soient inférieures ou égales à Q_3.

Remarque : Q_1 est la valeur x_i de la série dont l'indice i est le premier entier supérieur ou égal à $\frac{N}{4}$ (si N est l'effectif de la série).
Q_3 est la valeur x_i de la série dont l'indice i est le premier entier supérieur ou égal à $\frac{3N}{4}$ (si N est l'effectif de la série).

⚠ : la calculatrice permet de calculer la médiane mais ne calcule pas les bonnes valeurs pour les quartiles (elle applique une autre définition que celle du cours).

IV Variables discrètes : Indicateurs de dispersion

IV.1 Étendue

Définition

L'étendue d'une série statistique est la différence entre les valeurs extrêmes du caractère.

IV.2 Intervalle interquartile et écart interquartile

Définition

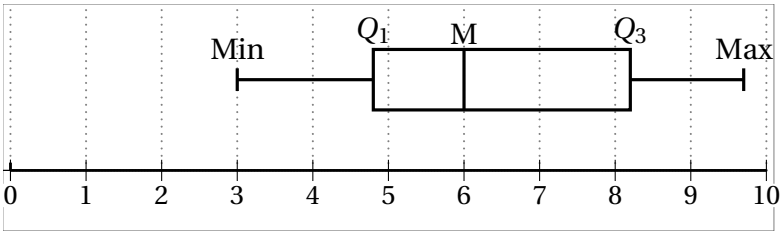
Si Q_1 et Q_3 sont les premier et troisième quartiles d'une série statistique, on nomme intervalle interquartile l'intervalle $[Q_1\ ;\ Q_3]$ et $Q_3 - Q_1$ est l'écart interquartile.

IV.3 Diagramme en boîte (ou diagramme de Tukey ou boîte à moustaches)

Les deux quartiles Q_1, Q_3, la médiane M d'une série statistique, associés aux valeurs extrêmes (minimum et maximum) permettent d'appréhender certaines caractéristiques de la répartition des valeurs.

Exemple :

Voici la série des températures (en degré Celcius) relevées sous abri à différents moments de la journée. Elles sont classées par ordre croissant.
3 ; 3,8 ; 4,5 ; 4,8 ; 5 ; 5,5 ; 5,7 ; 5,8 ; 6,2 ; 7 ; 7,3 ; 8,2 ; 9 ; 9,2 ; 9,5 ; 9,7
Les valeurs extrêmes sont 3 et 9,7.
La médiane vaut 6 (moyenne entre 5,8 et 6,2).
Le premier quartile est $Q_1 = 4,8$; le troisième quartile est $Q_3 = 8,2$. Le diagramme en boîte est alors :



Les diagrammes en boîte servent à faire des comparaisons de deux séries statistiques.

Exemple :

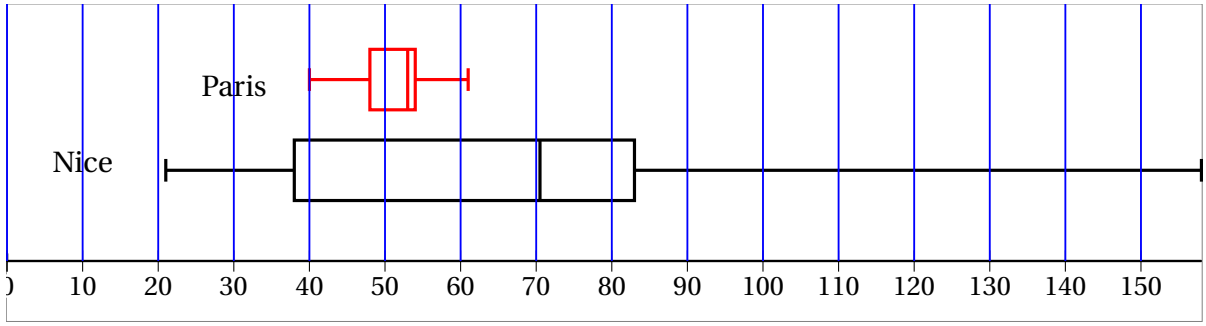
Les séries suivantes donnent les précipitations moyennes mensuelles en millimètres à Nice et à Paris :

Mois	J	F	M	A	M	J	J	A	S	O	N	D
Nice	67	83	71	70	39	37	21	38	83	109	158	92
Paris	53	48	40	45	53	57	54	61	54	50	58	51

Pour effectuer la comparaison, on va ranger chaque série par ordre croissant : (se fait à la calculatrice !)
Nice : 21 ; 37 ; 38 ; 39 ; 67 ; 70 ; 71 ; 83 ; 83 ; 92 ; 109 ; 158
Paris 40 ; 45 ; 48 ; 50 ; 51 ; 53 ; 53 ; 54 ; 54 ; 57 ; 58 ; 61

Pour Nice, on a : Min =21 ; Max = 158 ; $Q_1 = 38$; $M_1 = 70,5$ et $Q_3 = 83$
Pour Paris, on a : Min =40 ; Max = 61 ; $Q_1 = 48$; $M_1 = 53$ et $Q_3 = 54$

Diagrammes en boîtes :



Les précipitations sont plus régulières tout au long de l'année à Paris (série moins dispersée).
La totalité des valeurs de la série des précipitations à Paris est comprise entre le premier quartile et la médiane de la série des précipitations à Nice.
Pour la ville de Nice, plus de la moitié des mois ont des précipitations supérieures au maximum de Paris.

IV.4 Variance ; écart type

1. Considérons deux groupes d'élèves, l'un de dix élèves et l'autre de huit élèves ; leurs notes de mathématiques à un contrôle sont :

Première série :

note x_i	1	2	3	17	20
effectif n_i	3	1	1	1	4

Deuxième série :

note x_i	8	10	11	12
effectif n_i	1	2	4	1

La moyenne de la première série est : $m_1 = \dfrac{n_1x_1 + \cdots + n_5x_5}{n_1 + \cdots + n_5} = \dfrac{105}{10} = 10,5.$
La moyenne de la deuxième série est : $m_2 = \dfrac{84}{8} = 10,5.$
Les deux moyennes sont **égales** ; pourtant, la répartition des notes n'est pas du tout la même.

Il faut donc trouver un moyen de mesurer la dispersion des nombres autour de la moyenne.

Un premier moyen est l'étendue, mais ce n'est pas très fiable.

Nous pourrions calculer les écarts par rapport à la moyenne puis en faire la moyenne : Regardons ce que cela donne pour une série générale : Les valeurs x_i sont affectées d'une coefficient n_i et l'effectif total est N. La moeyenne est $\overline{x}$.

$$\frac{\sum_{i=0}^{i=p} n_i\left(x_i - \overline{x}\right)}{N} = \frac{\sum_{i=0}^{i=p} n_i x_i - \sum_{i=0}^{i=p} \overline{x}}{N} = \frac{\sum_{i=0}^{i=p} n_i x_i}{N} - \frac{N\overline{x}}{N} = \overline{x} - \overline{x} = 0.$$

Le problème de ce calcul est que l'on trouve systématiquement 0, donc cela ne permet pas d mesurer la dispersion (il y a compensation entre les valeurs supérieures à la moyenne et celles qui lui sont inférieures).

Nous allons voir un deuxième moyen, qui est l'écart type. L'idée est d'éviter les compensations, en n'ayant que des valeurs positives.

Définition (variance)

Soit une série statistique donnée par le tableau :

Valeur du caractère	x_1	x_2	$\cdots$	x_p	Total
Effectif	n_1	n_2	$\cdots$	n_p	N

La moyenne de cette série est : $\overline{x} = \dfrac{n_1x_1 + n_2x_2 + \cdots + n_Px_p}{N}$.

La **variance** est le nombre V défini par :

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

V est donc la **moyenne des carrés des écarts entre chaque valeur x_i et la moyenne**.

On dit aussi moyenne des carrés des écarts à la moyenne.

Autre formulation de la variance :

Pour chaque indice i, on a : $(x_i - \overline{x})^2 = x_i^2 - 2x_i\overline{x} + \overline{x}^2$.

En remplaçant dans le calcul de la variance chaque $(x_i - \overline{x})^2$ par ce que l'on vient de trouver, on obtient :

$$V = \frac{1}{N}\left[n_1\left(x_1^2 - 2x_1\overline{x} + \overline{x}^2\right) + n_2\left(x_2^2 - 2x_2\overline{x} + \overline{x}^2\right) + \cdots + n_p\left(x_p^2 - 2x_p\overline{x} + \overline{x}^2\right)\right]$$

$$= \frac{1}{N}\left[n_1x_1^2 + n_2x_2^2 + \cdots + n_px_p^2 - 2n_1x_1\overline{x} - 2n_2x_2\overline{x} - \cdots - n_px_p\overline{x} + n_1\overline{x}^2 + n_2\overline{x}^2 + \cdots + n_p\overline{x}^2\right]$$

$$= \frac{1}{N}\left[n_1x_1^2 + n_2x_2^2 + \cdots + n_px_p^2 - 2\overline{x}(n_1x_1 + n_2x_2 + \cdots + n_px_p) + \overline{x}^2\left(n_1 + n_2 + \cdots + n_p\right)\right]$$

$$= \frac{1}{N}\left[n_1x_1^2 + n_2x_2^2 + \cdots + n_px_p^2 - 2\overline{x} \times N\overline{x} + \overline{x}^2N\right]$$

$$= \frac{1}{N}\left[n_1x_1^2 + n_2x_2^2 + \cdots + n_px_p^2 - N\overline{x}^2\right]$$

$$= \frac{n_1x_1^2 + n_2x_2^2 + \cdots + n_px_p^2}{N} - \overline{x}^2.$$

donc :

$$\boxed{\mathbf{V = \frac{n_1x_1^2 + n_2x_2^2 + \cdots + n_px_p^2}{N} - \overline{x}^2}}$$

Symboliquement, on écrit $\boxed{V = \overline{x^2} - \overline{x}^2}$.

Exemple : pour la deuxième série de notes :

note x_i	8	10	11	12
x_i^2	64	100	121	144
effectif n_i	1	2	4	1

$$V = \frac{(1 \times 64) + (2 \times 100) + (4 \times 121) + (1 \times 144)}{8} - 10,5^2 = \frac{892}{8} - 10,5^2 = 111,5 - 110,25 = 1,25.$$

2. La variance est homogène aux carrés des valeurs de la série. Pour avoir une grandeur homogène aux valeurs de la série, on définit **l'écart type** de la série par : $\boxed{\sigma = \sqrt{V}}$.
   L'écart type est la racine carrée de la variance.

   Exemple : pour la première série de notes, on a : $V = \dfrac{1905}{10} - 10,5^2 = 80,25$.
   L'écart type de la première série est $\sigma = \sqrt{V} = \sqrt{80,25} \approx 8,96$.
   Celui de la deuxième série est $\sigma = \sqrt{1,25} \approx 1,118$.

   L'écart type de la première série est plus grand que celui de la deuxième série : les notes sont plus dispersées dans le premier cas que dans le second.

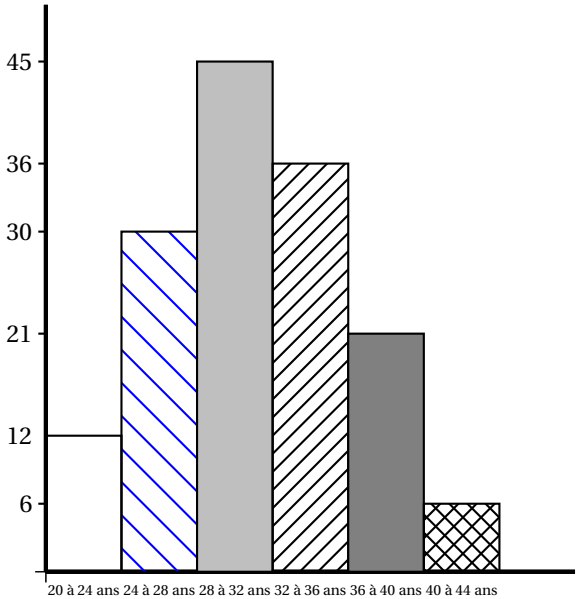
Remarque : La variance et l'écart-type se calculent facilement à la calculatrice, puisqu'elle donne la somme des valeurs et la somme des carrés des valeurs ainsi que l'effectif total.

V Variables continues

VI Histogrammes

Rappel : un histogramme est une représentation graphique sous la forme de rectangles, dont **l'aire** est proportionnelle aux effectifs.
Il y a deux cas possibles :

- Si le pas est constant (largeur des rectangles identiques), les aires des rectangles sont proportionnelles aux hauteurs des rectangles.
  Exemple : Brevet Polynésie juin 2007



(a) Compléter le tableau ci-dessous

Âge	$20 \leqslant$ âge < 24	$24 \leqslant$ âge < 28	$28 \leqslant$ âge < 32	$32 \leqslant$ âge < 36	$36 \leqslant$ âge < 40	$40 \leqslant$ âge < 44	Total
Centre de la classe	22						
Effectifs							
Fréquences en %							

(b) Quel est le pourcentage des employés qui ont strictement moins de 36 ans ?

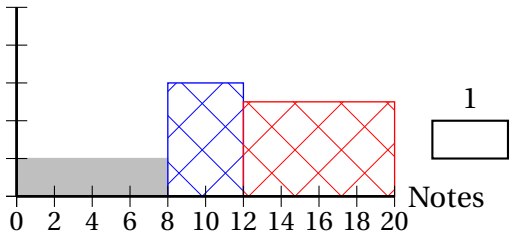
(c) Calculer l'âge moyen d'un employé de cette entreprise.

- Si le pas n'est pas constant, les hauteurs des rectangles ne sont plus proportionnelles aux aires de ceux-ci. C'est le cas quant les données sont réparties par classes, avec des largeurs d'intervalles différentes.
  Exemple :
  Considérons la répartition des notes de 10 élèves.

Classe	[0 ; 8[	[8 ; 12[	[12 ; 20[
Effectif	2	3	5

On commence par choisir une unité sur l'axe des abscisses et une unité d'aire. Par exemple : 2 cm^2 pour un effectif de 1.

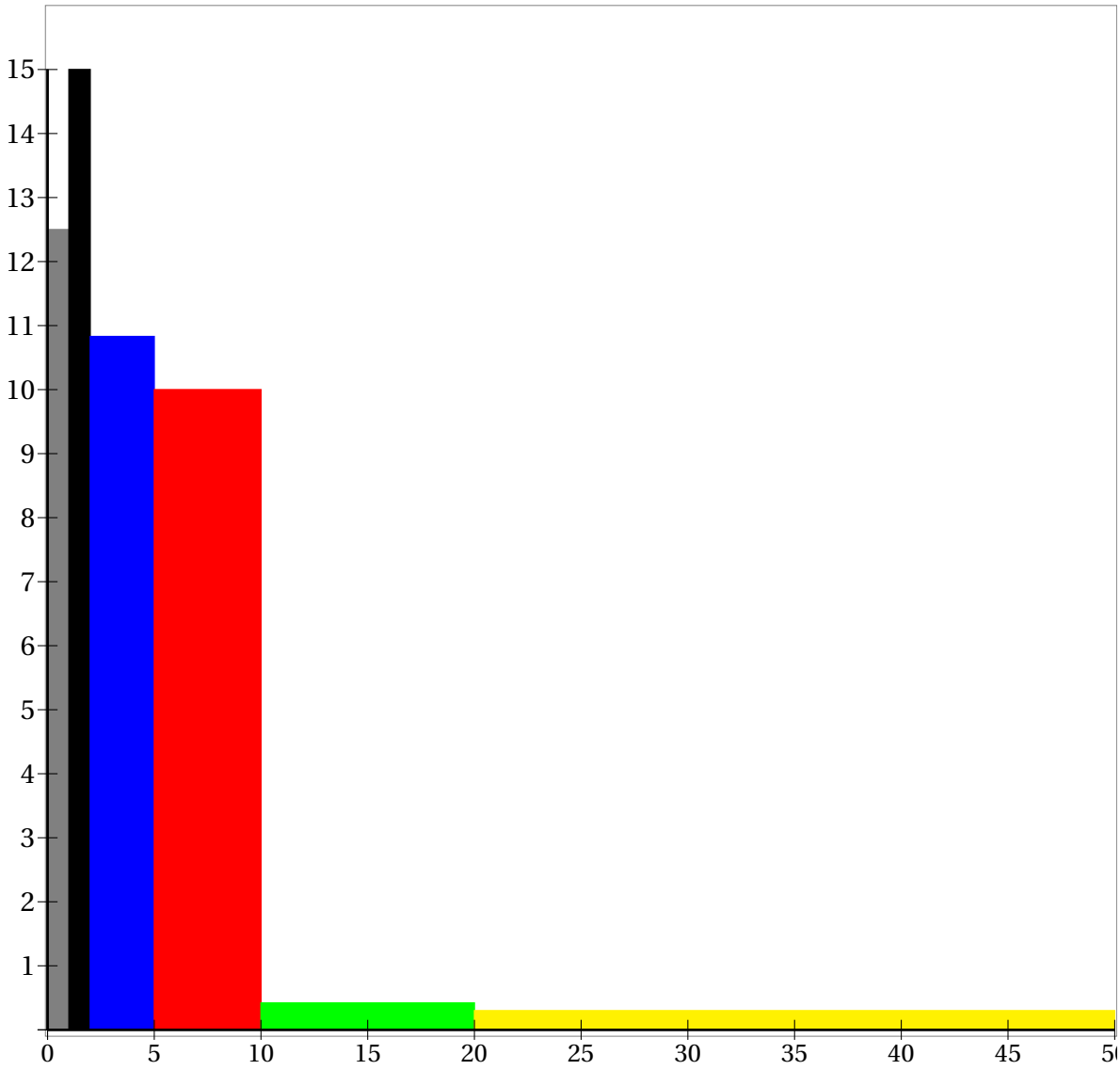
Classe	[0 ; 8[	[8 ; 12[	[12 ; 20[
Effectif	2	3	5
Aire en cm^2	4	6	10
Largeur en cm	4	2	4
Hauteur du rectangle	1	3	$\frac{5}{2} = 2,5$



Autre exemple :
On a relevé les distances domicile-lieu de travail pour les salariés d'une entreprise.

Distance (en km)	[0; 1[	[1; 2[	[2; 5[	[5; 10[	[10; 20[	[20; 50[
Effectifs	30	36	78	120	10	24

Construire l'histogramme relatif à cette série.



Exercice :

Le tableau ci-dessous représente la répartition des durées de 70 films (en min).

Durée en minutes	[100 ; 120[	[120 ; 160[	[160 ; 180[	[180 ; 260[
Effectifs	20	30	10	10

Représenter la situation par un histogramme.
On commencera à 100 sur l'axe des abscisses ; échelle : 1 cm pour 20 minutes en abscisses ; aire : 1 cm^2 pour 5 unités.