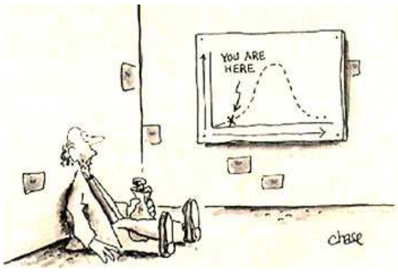
Cinquième Leçon

STATISTIQUES



A Les variables

A 1 Vocabulaire

Lors une enquête statistique, on recueille un certain nombre de caractéristiques ou de facteurs qu'on nomme en fait **variables** car elles... varient d'une observation à l'autre.
L'individu (personne, objet, entreprise, pays, etc.) sur lequel on mesure la variable est appelé **unité statistique**. L'ensemble de toutes les unités statistiques constitue la **population**. Cet ensemble peut parfois différer de l'ensemble des unités statistiques « interrogées » qui est alors appelé **échantillon**.

A 2 Variables qualitatives

Une variable est qualitative si ses différentes formes sont des catégories, des attributs comme par exemple : la spécialité choisie en terminale, le sexe, la couleur des chaussettes, degré de satisfaction à l'égard du professeur de maths, la région d'origine de son arrière grand-père, etc.

Les différentes formes prises par la variable sont appelées **modalités**. Par exemple, la variable « degré de satisfaction à l'égard du professeur de maths » peut prendre quatre modalités : satisfait, très satisfait, extrêmement satisfait, super-mega-plus satisfait.

A 3 Variables quantitatives

Une variable est quantitative si elle s'exprime sous la forme d'une valeur numérique : revenu, taille du lob de l'oreille gauche, âge, température, temps consacré quotidiennement à l'étude des mathématiques, etc.
Les variables quantitatives sont dites **continues** si elles peuvent couvrir toutes les valeurs d'un intervalle comme par exemple : la taille (on ne grandit pas d'un centimètre instantanément...), l'âge, etc.
Les autres variables sont dites **discrètes** comme par exemple le nombre de frères et sœurs, nombre de croissants au beurre consommés au petit déjeuner, etc.

B Tableaux et graphiques

B 1 Variables qualitatives ou quantitatives discrètes (peu de valeurs distinctes)

B 1 a Tableau des effectifs/ des fréquences

Le ministre de la justice de Syldavie a emprisonné 75 personnes la semaine dernière car elles n'avaient pas suffisamment déposé d'offrandes au Grand Protecteur Syldave (le GPS). Voici le tableau donnant le sexe des personnes emprisonnées :

m	f	m	m	m	f	f	f	m	f	m	m	m	m	f	f	f	m	m	f	f	f	f	m	f
f	m	m	m	m	m	f	m	f	f	m	f	f	f	f	m	m	m	f	f	f	m	m	f	f
f	f	f	f	f	f	f	f	m	m	m	f	m	m	f	m	f	m	m	m	f	m	m	m	f

FIGURE 1 – Sexe des 75 personnes ayant été emprisonnées (*Source : ministère syldave de la justice - BO n°57 p. 42*)

Cette série comporte 75 données qui prennent deux modalités : masculin ou féminin. Ce tableau est peu pratique à interpréter sous cette forme. Mieux vaut dresser un **tableau des effectifs** :

Sexe	Nombre d'emprisonnés
Masculin	
Féminin	
Total	75

FIGURE 2 – Répartition des 75 emprisonnés selon le sexe (*Source : ministère syldave de la justice - BO n°57 p. 42*)

Il peut même être encore plus parlant de dresser le **tableau des fréquences** :

Sexe	Pourcentage d'emprisonnés (%)
Masculin	
Féminin	
Total	100

FIGURE 3 – Répartition en pourcentage des 75 emprisonnés selon le sexe (*Source : ministère syldave de la justice - BO n°57 p. 42*)

B 1 b Représentations graphiques

Les résultats d'un sondage sur l'opinion qu'ont les Syldaves sur leur Grand Protecteur ont été regroupés dans le tableau suivant :

Degré de satisfaction	Pourcentage de la population
Extrêmement satisfait	50,9 %
Tout à fait satisfait	34,3 %
Totalement satisfait	10,9 %
Très satisfait	3,9 %
Total	100 %

FIGURE 4 – Répartition en pourcentage des Syldaves selon leur degré de satisfaction à l'égard du GPS (*Source : ministère syldave de la justice - BO n°58 p. 554*)

Diagramme en bâtons En face de chaque modalité représentée régulièrement en abscisses on trace un « bâton » de longueur proportionnelle à la fréquence :

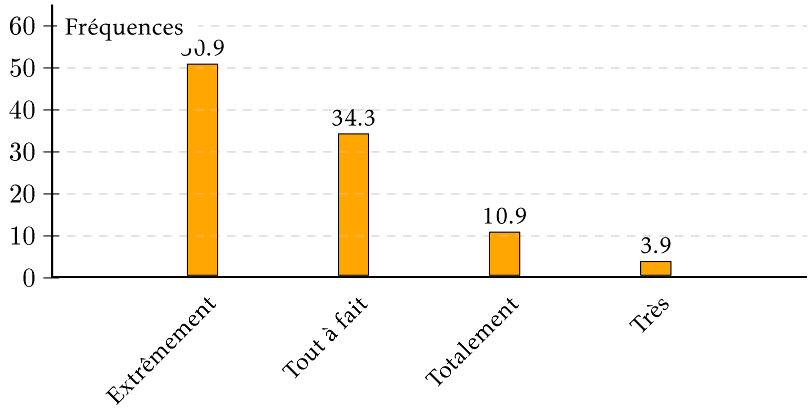


FIGURE 5 – Répartition en pourcentage des Syldaves selon leur degré de satisfaction à l'égard du GPS (*Source : ministère syldave de la justice - BO n°58 p. 554*) - Diagramme en bâtons

Diagramme à secteurs (« camembert ») On découpe un disque en autant de secteurs qu'il y a de modalités. La part de chaque secteur correspond à la fréquence relative de la modalité qu'il représente, sachant que 100 % correspond bien sûr à 360 °.
On remplit alors le tableau suivant :

Degré de satisfaction	Pourcentage de la population	Angle
Extrêmement satisfait	50,9 %	
Tout à fait satisfait	34,3 %	
Totalement satisfait	10,9 %	
Très satisfait	3,9 %	
Total	100 %	360°

FIGURE 6 – Répartition en pourcentage des Syldaves selon leur degré de satisfaction à l'égard du GPS (*Source : ministère syldave de la justice - BO n°58 p. 554*) - Calcul des angles du diagramme circulaire

Et on obtient le camembert suivant :

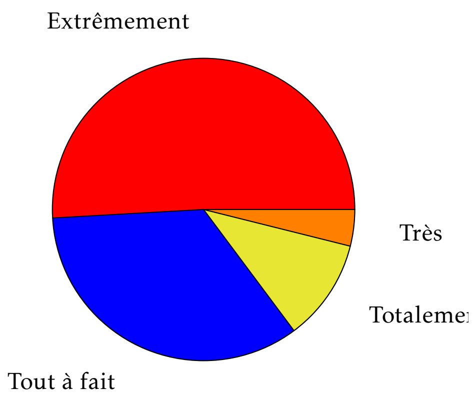


FIGURE 7 – Répartition en pourcentage des Syldaves selon leur degré de satisfaction à l'égard du GPS (*Source : ministère syldave des la justice - BO n°58 p. 554*) - Diagramme circulaire

On peut préférer d'ailleurs un demi-camembert en remplissant le tableau précédent avec 180°comme angle correspondant à 100 % :

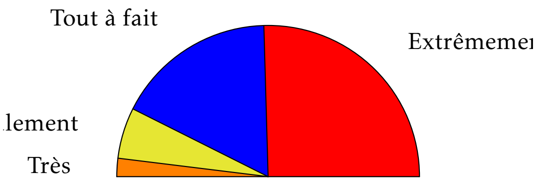


FIGURE 8 – Répartition en pourcentage des Syldaves selon leur degré de satisfaction à l'égard du GPS (*Source : ministère syldave de la justice - BO n°58 p. 554*) - Diagramme semi-circulaire

B 2 Variables quantitatives discrètes (grand nombre de valeurs) ou continues

B 2 a Groupement par classes

Voici un tableau donnant le nombre annuel de disques vendus par la femme du Grand Protecteur de la Syldavie de 1960 à 2008 :

163	169	183	224	140	163	172
104	121	141	139	135	126	163
127	141	136	95	128	95	133
92	119	103	150	149	110	148
153	156	151	109	143	123	75
189	126	173	171	128	153	142
148	112	112	150	121	162	120

FIGURE 9 – Nombre annuel de disques vendus par la femme du Grand Protecteur de la Syldavie de 1960 à 2008 (*Source : ministère syldave de la culture - BO n°51 p. 1254559*)

Il est totalement inutile de comptabiliser les occurrences d'un nombre donné de vente. Il est beaucoup plus parlant de regrouper les résultats par classe.
Nous pouvons par exemple ici regrouper les modalités dans des intervalles de largeur 25 en commençant par 75.
Par convention, on utilise des intervalles fermés à gauche et ouverts à droite. Ici, on dénombre 4 modalités m telles que $75 \leqslant m < 100$. On remplit alors le tableau suivant :

Ventes annuelles de disques	Nombres d'années
$[75\,;100[$	4
$[100\,;125[$	
$[125\,;150[$	
$[150\,;175[$	
$[175\,;200[$	
$[200\,;225[$	
Total	49

FIGURE 10 – Répartition de 49 années selon le nombre annuel de disques vendus par la femme du Grand Protecteur de la Syldavie de 1960 à 2008 (*Source : ministère syldave de la culture - BO n°51 p. 1254559*)

On peut dresser le même type de tableau avec les fréquences :

Ventes annuelles de disques	Pourcentage d'années
[75; 100[	
[100; 125[	
[125; 150[	
[150; 175[	
[175; 200[	
[200; 225[	
Total	100%

FIGURE 11 – Répartition en pourcentage de 49 années selon le nombre annuel de disques vendus par la femme du Grand Protecteur de la Syldavie de 1960 à 2008 (*Source : ministère syldave de la culture - BO n°51 p. 1254559*)

B 2 b Histogramme avec classes de même amplitude

Ici, on va représenter dans un repère les données d'un des tableaux précédents. Comme les classes sont de même amplitude, cela ressemblera à un diagramme en bâtons... dont les bâtons sont collés.

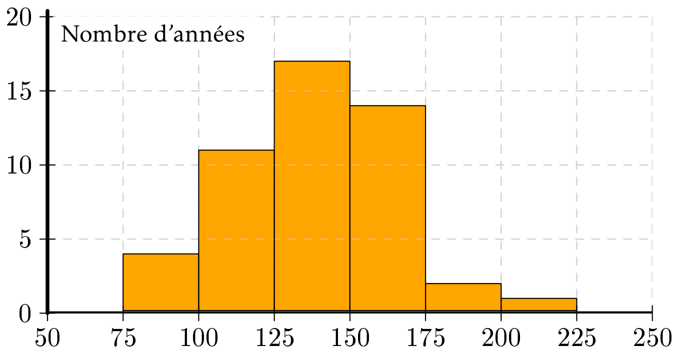


FIGURE 12 – Répartition de 49 années selon le nombre annuel de disques vendus par la femme du Grand Protecteur de la Syldavie de 1960 à 2008 (*Source : ministère syldave de la culture - BO n°51 p. 1254559*) - Histogramme

B 2 c Histogramme avec classes d'amplitudes différentes

Considérez le tableau suivant :

Âge	Nombre de femmes
15-20	202
20-25	204
25-35	359
35-45	338
45-65	304
65-75	10
Total	1 414

FIGURE 13 – Répartition des femmes du harem du GPS selon l'âge(*Source : ministère syldave du travail- BO n°7*)

Les classes ici ne sont plus de même amplitude. L'usage est de représenter chaque classe par des rectangles d'aires pro-

portionnelles aux effectifs de chaque classe. Par exemple, la 3e classe a un plus grand effectif mais est deux fois plus étendue. Elle est donc équivalente à deux classes d'amplitude 5 et d'effectif $\frac{359}{2}$.
On appelle **densité** le rapport

$$\frac{\text{Effectif}}{\text{Amplitude}}$$

Ce résultat est équivalent à

$$\text{Effectif} = \text{Amplitude} \times \text{Densité}$$

L'effectif correspond donc à l'aire du rectangle :

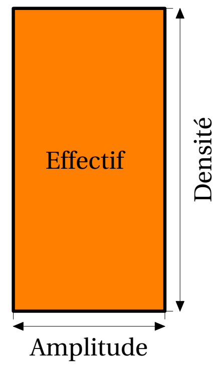


FIGURE 14 – Construction d'un histogramme à pas non constant

On remplit donc le tableau suivant :

Intervalles	[15 ; 20[	[20 ; 25[	[25 ; 35[			
Effectifs	202	204	359			
Amplitudes	5	5	10			
Densités	40,4					

FIGURE 15 – Répartition des femmes du harem du Grand Protecteur de la Syldavie selon l'âge(*Source : ministère syldave du travail- BO n°7*) - Tableau de construction de l'histogramme

On en déduit l'histogramme suivant :

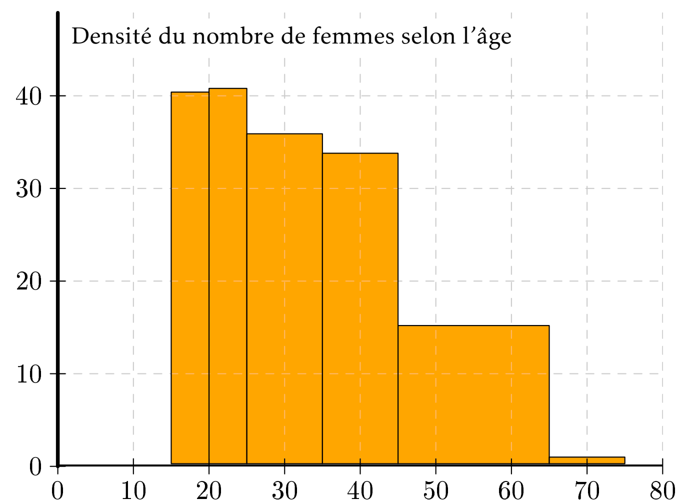


FIGURE 16 – Répartition des femmes du harem du Grand Protecteur de la Syldavie selon l'âge(*Source : ministère syldave du travail- BO n°7*) -Histogramme

C Tableaux à double entrée

Lorsqu'on effectue une enquête, on peut être amené à recueillir des renseignements sur plus d'une caractéristique des unités statistiques.
On a par exemple étudié la répartition des détenus d'une prison syldave selon l'ancienneté et le sexe :

Ancienneté \ Sexe	Masculin	Féminin	Total
Moins de 5 ans	20	20	
5 à 10 ans	20	10	
10 à 15 ans	25	15	
15 ans et plus	75	25	
Total			

FIGURE 17 – Répartition des détenus d'une prison syldave selon l'ancienneté et le sexe (*Source : ministère syldave des loisirs - BO n°0*)

Il est plus parlant de passer aux fréquences :

Ancienneté \ Sexe	Masculin	Féminin	Total
Moins de 5 ans			
5 à 10 ans			
10 à 15 ans			
15 ans et plus			
Total			100 %

FIGURE 18 – Répartition en pourcentage des détenus d'une prison syldave selon l'ancienneté et le sexe (*Source : ministère syldave des loisirs - BO n°0*)

On aurait pu également répartir les détenus masculins et féminins selon leur ancienneté :

Ancienneté \ Sexe	Masculin	Féminin
Moins de 5 ans		
5 à 10 ans		
10 à 15 ans		
15 ans et plus		
Total	100 %	100 %

FIGURE 19 – Répartition en pourcentage des détenus d'une prison syldave par sexe selon l'ancienneté(*Source : ministère syldave des loisirs - BO n°0*)

Ou encore répartir les détenus par catégorie d'ancienneté selon le sexe :

Sexe / Ancienneté	Masculin	Féminin	Total
Moins de 5 ans			100 %
5 à 10 ans			100 %
10 à 15 ans			100 %
15 ans et plus			100 %

FIGURE 20 – Répartition en pourcentage des détenus d'une prison syldave par catégorie d'ancienneté selon le sexe (*Source : ministère syldave des loisirs - BO n°0*)

Chacun de ces tableaux met en valeur des phénomènes différents.

D Mesures de tendance centrale

D 1 Le mode

D 1 a Définition

Définition V - 1 : mode

Mesure qui correspond à la valeur ou à la modalité la plus fréquente. S'il y a plusieurs mode, on dit que la distribution est multimodale.

C'est la mesure la plus simple à évaluer.

D 1 b Données groupées par valeurs ou modalités

Si on reprend l'exemple du tableau 4 page 2, le mode est « extrêmement satisfait ».

D 1 c Données groupées par classes

Reprenons cette fois le graphique 12 page 5. La classe modale correspond à l'intervalle [125 ; 150[. Si on veut un mode plutôt que la classe modale, on prendra le milieu de la classe modale : ici 137,5.

D 1 d Distribution polymodale

Considérons le nouvel exemple suivant :

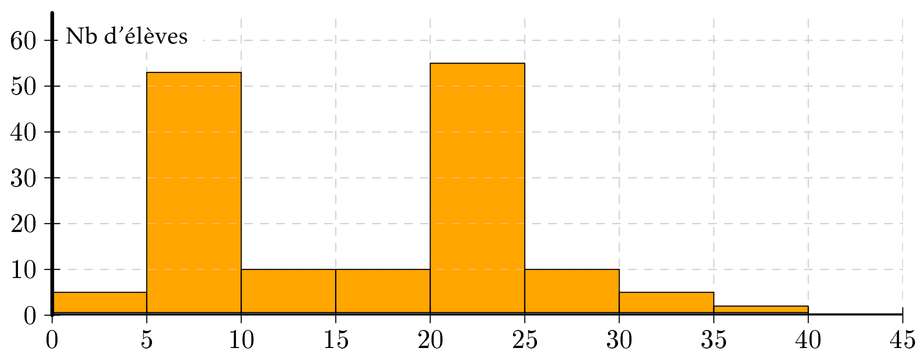


FIGURE 21 – Répartition des prisonniers syldaves selon la durée de leur peine (en années)

Ce graphique présente deux classes modales : [5 ; 10[et [20 ; 25[.

D 2 La médiane

D 2 a Conventions

Dans toute la suite, on étudiera une population, notée E, et une variable statistique quantitative X définie sur E.

Exemple V - 1 :

Si on étudie par exemple le nombre de poupées Barbue que possèdent les élèves de la classe de 2^{nde} 9
- E est l'ensemble des élèves de la classe
- X est la fonction qui, à un élément de E, associe le nombre de poupées Barbue qu'il ou elle possède.

Si E possède n éléments, on notera $\mathrm{V} = \{x_1, x_2, \cdots, x_{n-1}, x_n\}$ l'ensemble **ordonné par valeurs croissantes** des valeurs prises par X.

Remarque V - 1 :

Notez bien que certains éléments de V peuvent être égaux : en effet, deux élèves différents peuvent avoir le même nombre de poupées Barbue.

D 2 b Définition

Définition V - 2 : médiane

La médiane M_e est un nombre tel que :
- **au moins** 50% des éléments de V sont inférieurs à M_e,
- **au moins** 50% des éléments de V sont supérieurs à M_e.

Exemple V - 2 : calcul de la médiane

- $\mathrm{V} = \{10, 20, 20, 50, \mathbf{50}, 80, 80, 90, 370\}$
- ▷ $\mathrm{M}_e = 50$
- $\mathrm{V} = \{10, 20, 20, 50, \mathbf{80}, 80, 90, 370\}$
- ▷ $\mathrm{M}_e = \dfrac{50+80}{2} = 65$

D 2 c Cas des données groupées par valeurs

Considérons par exemple la répartition des ministres syldaves selon le nombre d'années qu'ils ont étudié après leur Brevet des Collèges :

Nombre d'années après le Brevet	Nombre de ministres	Effectifs cumulés croissants
0	69	69
1	31	100
2	15	
3	6	
4	3	
5	1	

FIGURE 22 – Répartition des ministres syldaves selon le nombre d'années qu'ils ont étudié après leur Brevet des Collèges

Il y a un nombre impair de données (125). La médiane correspond donc à la donnée de rang 63 (62 après, 62 avant). Grâce à la troisième colonne, on trouve que la 63^{e} valeur vaut 0. Ainsi la médiane $M_e = 0$. On en déduit qu'au moins 50 % des ministres n'ont pas poursuivi d'étude après le Brevet.

Il peut être plus pratique de travailler avec les fréquences :

Nombre d'années après le Brevet	Nombre de ministres	Pourcentage des ministres	Fréquence	Fréquences cumulées croissantes
0	69	69		
1	31	100		
2	15			
3	6			
4	3			
5	1			

FIGURE 23 – Répartition en pourcentage des ministres syldaves selon le nombre d'années qu'ils ont étudié après leur Brevet des Collèges

Il suffit alors de regarder à quelle valeur correspond la fréquence cumulée 50 %.

D 2 d Cas des données groupées par classes

Lorsque les données sont regroupées par classes, on peut, soit procéder graphiquement, soit effectuer un calcul. Dans les deux cas il s'agira d'une approximation.
Considérons l'exemple suivant :

Nombre d'heures	Nombre d'élèves	Pourcentage des élèves	Pourcentage cumulé des élèves
0-5	300		
5-10	420		
10-15	500		
15-20	330		
20-25	250		
25-30	160		
30-35	40		
Total	2 000	100 %	

FIGURE 24 – Répartition et répartition cumulée des 2 000 élèves des lycées Perrin et Goussier de Rezé selon le nombre d'heures de travail personnel par mois.

Méthode graphique On trace le *polygone des fréquences cumulées croissantes :*

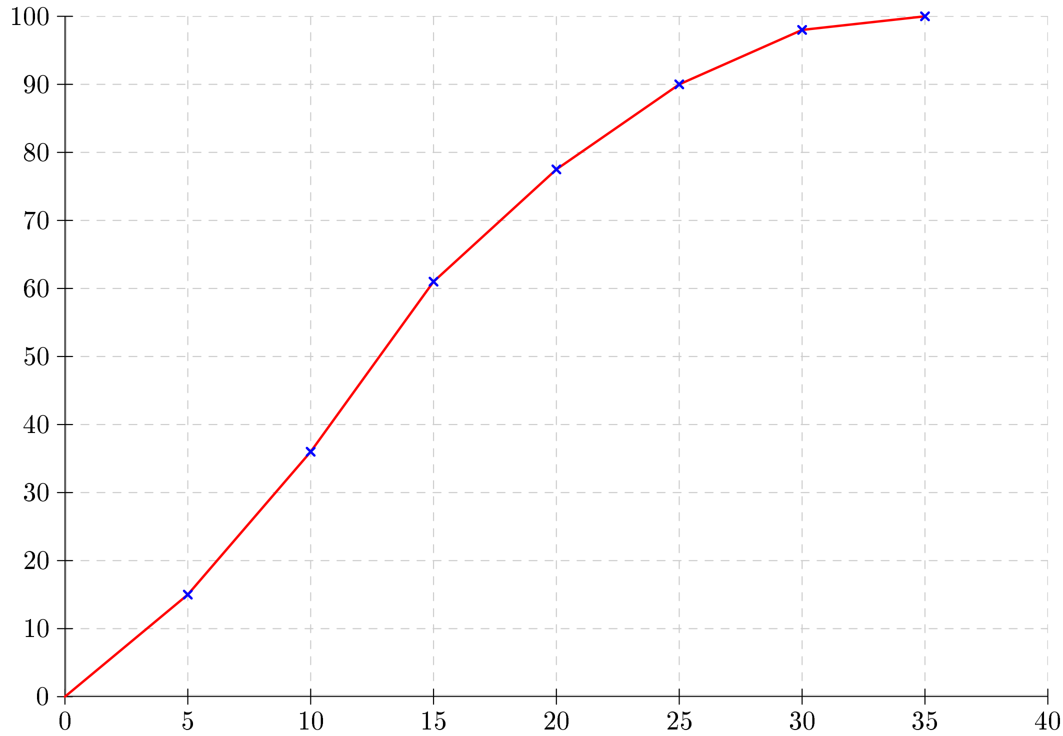


FIGURE 25 – Polygone des fréquences cumulées croissantes des 2 000 élèves des lycées Perrin et Goussier de Rezé selon le nombre d'heures de travail personnel par mois.

À chaque borne supérieure des classes, on fait correspondre la fréquence cumulée et on relie les points par des segments (cela constitue une approximation mais le fait de regrouper les élèves en classe en était déjà une).
Comme la médiane correspond à 50 % des effectifs, on lit sur le graphique l'abscisse du point du polygone d'ordonnée 50 : environ 13.

Méthode analytique P est le point de la droite (M_2M_3) d'ordonnée 50. On détermine donc une équation de la droite (M_2M_3) et on calcule l'abscisse cherchée.
Ici $M_2(10\,;36)$ et $M_3(15\,;61)$. L'équation réduite de (M_2M_3) est de la forme $y = ax + b$ avec a le coefficient directeur.

$$a = \frac{y_{M_3} - y_{M_2}}{x_{M_3} - x_{M_2}} = \frac{61-36}{15-10} = 5$$

L'équation devient donc $y = 5x + b$. Pour déterminer b, il suffit d'utiliser le fait que M_2 appartient à (M_2M_3) donc vérifie son équation :

$$36 = 5 \times 10 + b \iff b = 36 - 50 = -14$$

Finalement, l'équation réduite cherchée est $y = 5x - 14$. Il reste à chercher le point de cette droite d'ordonnée 50 :

$$50 = 5x_p - 14 \iff x_p = \frac{50+14}{5} = 12{,}8$$

On trouve donc $M_e \approx 12{,}8$.

D 3 Moyenne

D 3 a Données non groupées

Soit $x_1, x_2, \ldots, x_n$ une série statistique. La moyenne $\overline{x}$ est l'unique valeur que devrait prendre chacune des données pour que la somme des données soit préservée.
En d'autres termes, on cherche $\overline{x}$ tel que

$$x_1 + x_2 + \cdots + x_n = \underbrace{\overline{x} + \overline{x} + \cdots + \overline{x}}_{n \text{ termes}}$$

On en déduit que :

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

On utilise souvent le symbole Σ pour représenter une somme (sigma est la lettre grecque correspondant à notre S).
Ainsi $\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5$.
La formule de la moyenne devient donc :

$$\overline{x} = \frac{\sum_{i=1}^{n} x_1}{1} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Faut-il vraiment un exemple ?...
Bon, voici le relevé des pourcentages de réussite lors du dernier exercice de tir des généraux de l'armée syldave :

12 22 32 2 24 2 5 33

Quel est le pourcentage moyen de réussite ?

D 3 b Données groupées par valeurs

Reprenons le tableau 22 page 9. Si nous voulons calculer le nombre moyen d'années d'étude après le brevet des ministres syldaves il faudrait additionner 69 zéros, 31 un, 15 deux, 6 trois, 3 quatre, 1 cinq et diviser le tout par le nombre total de ministres à savoir 125 d'après ce que nous venons de voir.
Nous pouvons simplifier les choses en notant e_i l'effectif de la valeur x_i :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} e_i \cdot x_i$$

Pour les ministres, $\overline{x} = \frac{1}{125}(69 \times 0 + 31 \times 1 + 15 \times 2 + 6 \times 3 + 3 \times 4 + 1 \times 5) = \cdots$

D 3 c Données groupées par classes

Regrouper des données par classe constitue déjà une approximation mais a le désavantage de nous empêcher d'utiliser nos moyens de calculer la moyenne. Pour y remédier, nous allons effectuer une approximation supplémentaire en prenant comme représentant d'une classe son **milieu**.
Reprenons le tableau 13 page 5 en le complétant :

Âge	Milieu de classe (m_i)	Nombre de femmes
15-20		202
20-25		204
25-35		359
35-45		338
45-65		304
65-75		10
Total		1 414

FIGURE 26 – Répartition des femmes du harem du Grand Protecteur de la Syldavie selon l'âge avec le milieu des classes

Nous pouvons alors trouver une approximation de l'âge moyen des femmes du Harem :

$$\overline{x} \approx \frac{1}{n}\sum_{i=1}^{n} m_i e_i = \frac{202 \times 17,5 + \cdots}{1\,414} = \cdots$$

E Quartiles et diagrammes en boîte

E 1 L'idée

Pour avoir une idée un peu plus précise de la série statistique étudiée, on voudrait séparer notre population en 4 groupes au lieu de 2 comme cela a été fait avec la médiane.
On a donc envie de calculer les médianes des parties basses et hautes.
On a également comme cahier des charges d'avoir au moins 25% des valeurs prises par x inférieures au premier quartile Q_1 et au moins 75% des valeurs prises par x inférieures au troisième quartile Q_3.

E 2 Expérimentons

Exemple V - 3 : huit éléments

$V = \{10, 20, 30, 40, 50, 60, 70, 80\}$

On peut séparer l'effectif en quatre groupes de même effectif égal à 25% de l'effectif total donc ça « colle »

$$V = \{10, 20, 30, 40, 50, 60, 70, 80\}$$

On peut prendre
- $Q_1 = \frac{20+30}{2} = 25$ et 25% des effectifs ont une valeur inférieure à Q_1 (10 et 20 c'est-à-dire 2 sur 8)
- $M_e = \frac{40+50}{2} = 45$ et 50% des effectifs ont une valeur inférieure à M_e (10, 20 30 et 40 c'est-à-dire 4 sur 8)
- $Q_3 = \frac{60+70}{2} = 65$ et 75% des effectifs ont une valeur inférieure à Q_3 (6 valeurs sur 8)

et Q_1 et Q_3 sont bien les médianes respectives des parties basses et hautes.

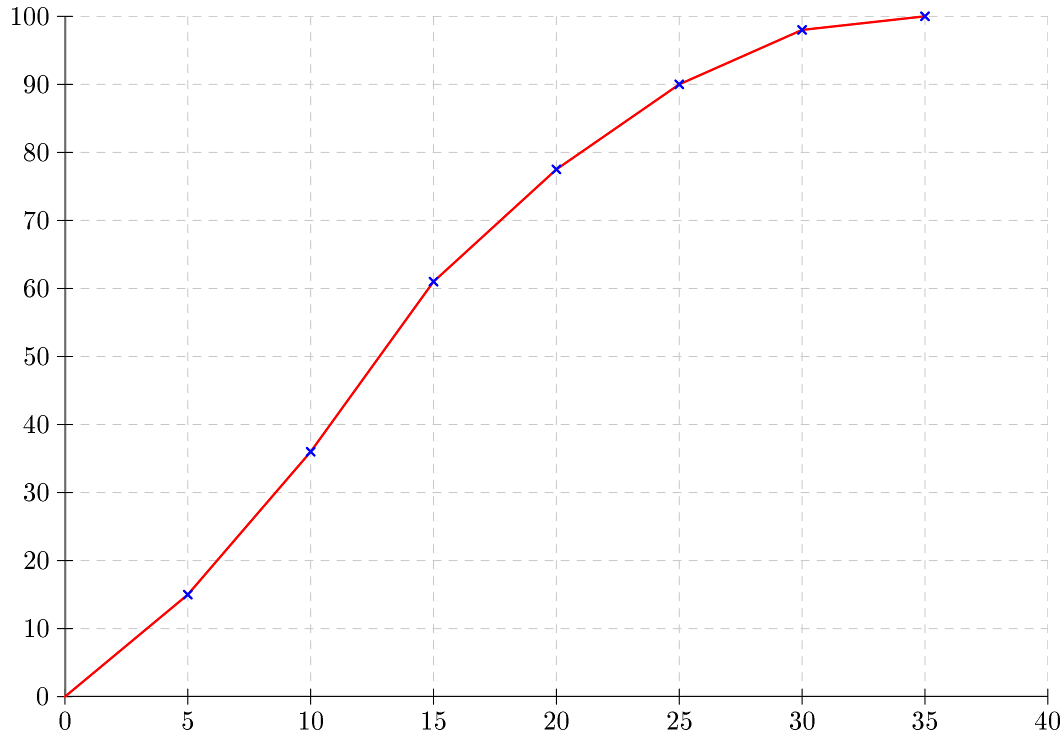
E 3 Définissons

Définition V - 3 : quartiles

Le premier quartile est obtenu en prenant la médiane de la sous-série contenant les observations dont le rang est strictement inférieur à celui de la médiane (*la partie basse*) pour autant qu'au moins 25% des observations soient inférieures ou égales à cette valeur.
Sinon, il faut inclure la médiane dans la partie basse.
Le troisième quartile est obtenu en prenant la médiane de la sous-série contenant les observations dont le rang est strictement supérieur à celui de la médiane (*la partie haute*) pour autant qu'au moins 75% des observations soient inférieures ou égales à cette valeur.
Sinon, il faut inclure la médiane dans la partie haute.

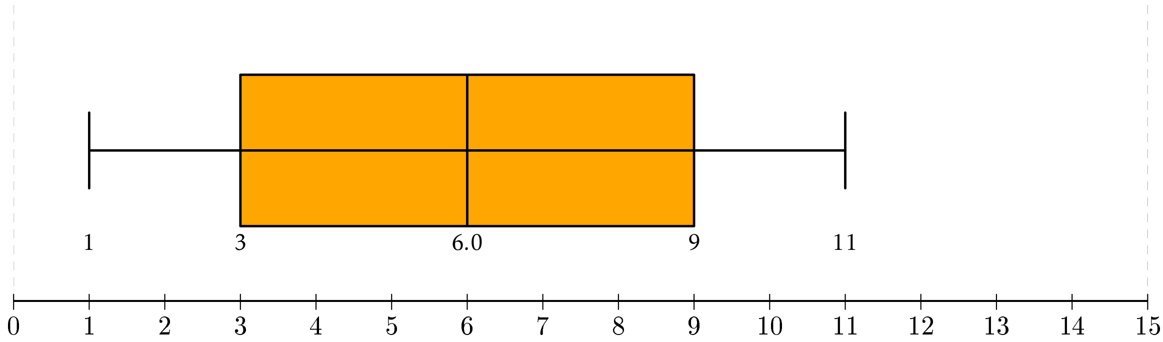
E 4 Lecture graphique dans le cas des données groupées en classe

Complétons le graphique 25 page 11 qui avait été obtenu grâce au tableau 24 page 10 :



E 5 Boîte à moustache

Après avoir calculé les quartiles, on peut les regrouper dans un tableau. Il est toutefois plus parlant de dresser un diagramme en boîte, ou diagramme de Tukey ou encore boîte à moustache. Reprenons pour cela l'exemple **??** page ??



Ce diagramme est dû au statisticien américain John TUKEY (1915 - 2000) qui a également dit :

> *Dans un monde où le prix du calcul continue à diminuer rapidement, alors que le prix de la démonstration d'un théorème est stable ou augmente, les principes élémentaires de l'économie indiquent que nous devrions utiliser une part de plus en plus grande de notre temps à faire des calculs.*

F Mesures de dispersion

F 1 Écart interquartile

Continuons à exploiter nos quartiles. Environ 50 % de la population a une modalité entre Q_1 et Q_3 : en observant la boîte à moustache décrivant une série statistique, on peut affiner sa description grâce à **l'écart interquartile**.

Définition V - 4 : écart interquartile

La distance entre Q_1 et Q_3 est appelé écart interquartile.

Observons deux exemples :

Exemple V - 4 : comparaison des écarts interquartiles

Voici un tableau donnant les notes au Bac de deux classes :

Notes	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Effectifs classe A	0	0	0	1	1	0	2	3	2	9	5	9	5	1	2	1	0	2	1	1	0
Effectifs classe B	0	5	5	5	5	0	0	0	0	0	3	0	5	5	0	2	0	5	0	5	0

Voici les boîtes à moustaches correspondant :

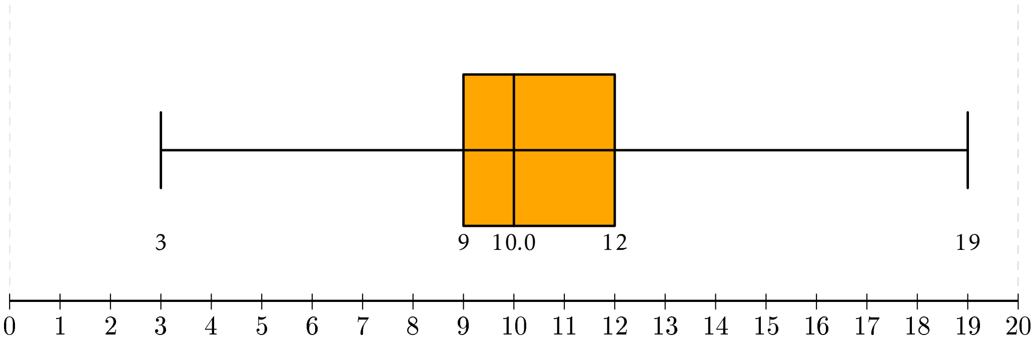


FIGURE 27 – Notes au Bac de la classe A

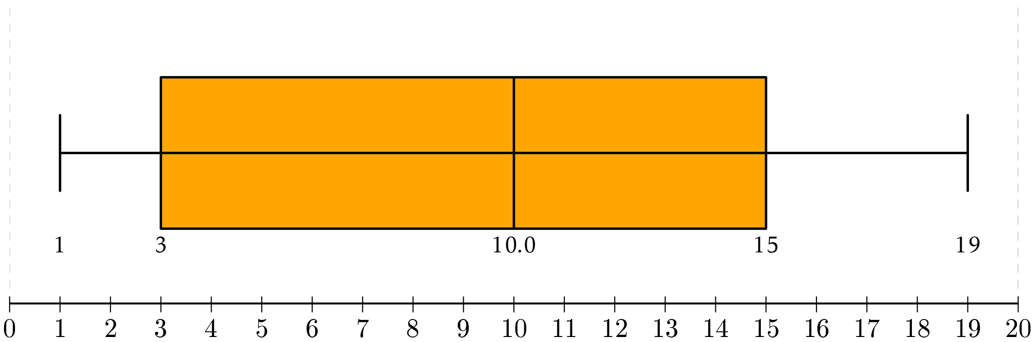


FIGURE 28 – Notes au Bac de la classe B

Dans les deux cas la médiane vaut 10 : cela signifie qu'au moins 50 % des élèves de chaque classe a eu moins de la moyenne.

Cependant, les écarts interquartiles sont nettement différents : $Q_3 - Q_1 = 3$ dans la classe A mais $Q_3 - Q_1 = 12$ dans la classe B. La classe B est nettement plus hétérogène en terme de résultats. En effet, dans la classe A, 50 % environ des élèves ont eu entre 9 et 12 alors que dans la classe B la moitié des notes se situent entre 3 et 15.

F 2 Variance et écart-type

La dispersion peut également se mesurer autour de la moyenne.
Considérons une série statistique quelconque :

Valeurs x_i	0	1	2	3	4	7
Effectifs n_i	1	2	2	4	3	2

FIGURE 29 – Série statistique quelconque

À partir de ces données, calculez la moyenne $\overline{x}$ puis remplissez le tableau suivant proposant deux façons de « mesurer » pour chaque valeur « l'éloignement » par rapport à $\overline{x}$.

$x_i - \bar{x}$						
$(x_i - \bar{x})^2$						

FIGURE 30 – Essais de mesures de dispersion par rapport à la moyenne

Calculez dans chacun des trois cas l'éloignement moyen, c'est-à-dire la moyenne des écarts...

On peut visualiser les écarts sur le schéma suivant :

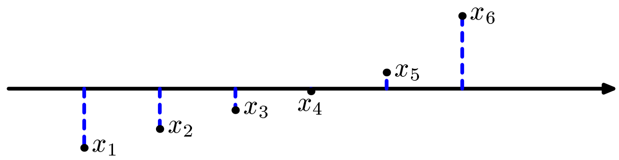


FIGURE 31 – Visualisation de l'écart par rapport à la moyenne

On peut prouver (à titre d'exercice...) que l'écart correspondant à la première ligne du tableau 30 est toujours nul. On préfère donc utiliser la moyenne des écarts de la deuxième ligne qu'on appelle **variance**.

Définition V - 5 : variance

On appelle variance d'une série quelconque à caractère quantitatif discret le nombre :

$$V = \frac{1}{N}\sum_{i=1}^{p} n_i(x_i - \overline{x})^2 = \sum_{i=1}^{p} f_i(x_i - \overline{x})^2$$

en notant n_i les effectifs et f_i les fréquences.

Remarque V - 2 : cas d'un regroupement en classe

Dans le cas d'un regroupement en classe, on considère, comme pour le calcul de la moyenne (cf D 3 c page 12), le milieu des classes.

La variance est homogène au carré des valeurs x_i : on préfère donc en prendre la racine carrée pour revenir à une grandeur homogène aux valeurs mesurées. L'ÉCART-TYPE A DONC LA MÊME UNITÉ QUE LA POPULATION !

Définition V - 6 : écart-type

L'écart-type d'une série est la racine carrée de la variance. On le note souvant σ.

Reprenez les séries de l'exemple 4 page 15 et calculez les moyennes et écarts-type. On peut visualiser cette dispersion en traçant un diagramme à bâton et en mettant en évidence l'intervalle $[\overline{x}-\sigma\,;\,\overline{x}+\sigma]$:

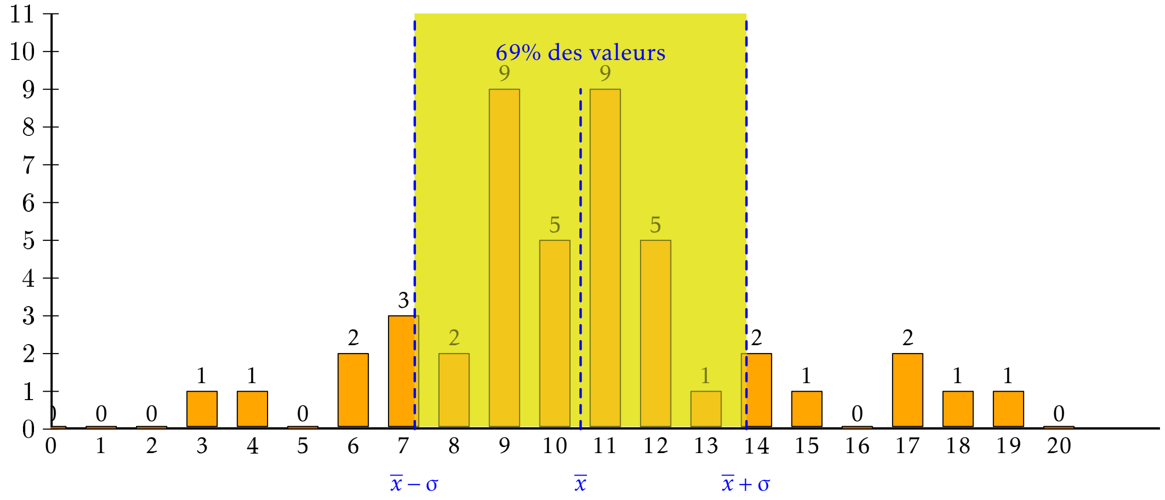


FIGURE 32 – Notes au Bac de la classe A

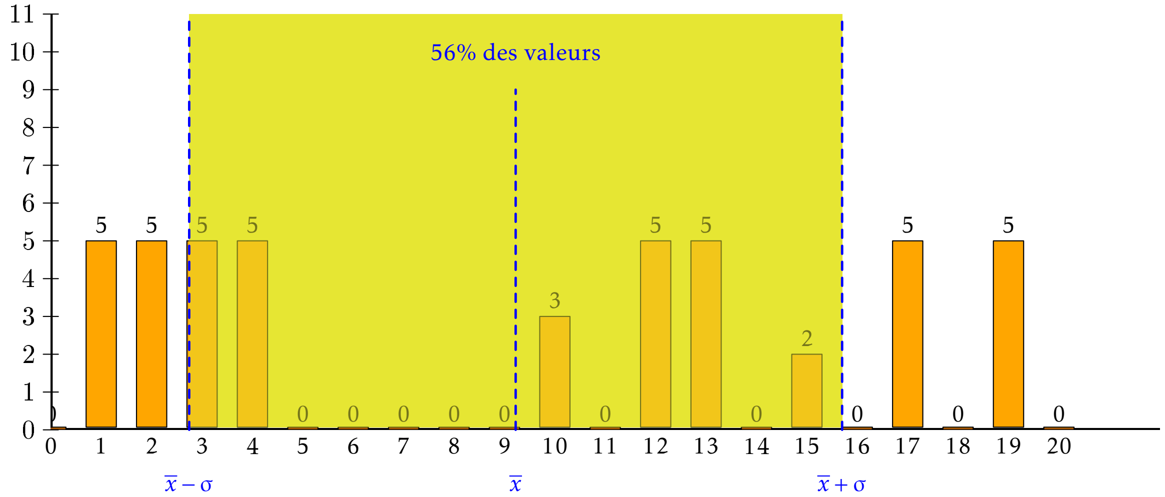


FIGURE 33 – Notes au Bac de la classe B

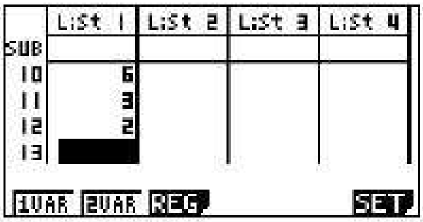
Propriété V - 1 : Distribution Gaussienne

Dans une distribution Gaussienne (ou normale), de moyenne $\overline{x}$ et d'écart-type σ, on (pas nous) démontre que :
- au moins 68% de la population se trouve dans l'intervalle $\left[\overline{x}-\sigma;\overline{x}+\sigma\right]$;
- au moins 95% de la population se trouve dans l'intervalle $\left[\overline{x}-2\sigma;\overline{x}+2\sigma\right]$.

G Utilisation de la calculatrice

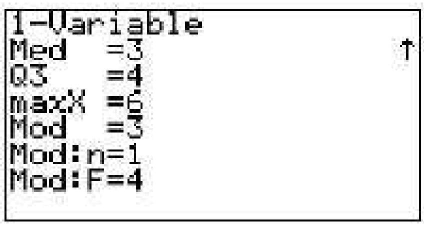
G 1 Casio

Vous rentrez les variables dans la liste 1 et les effectifs ou les fréquences dans la liste 2 puis vous lancez la commande CALC en appuyant sur F2.

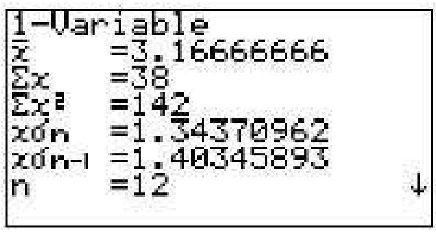


Pour vérifiez que la machine est bien configurée, il faut aller sur SET en appuyant sur F6
Il faut placer List1 devant 1Var XList et List2 devant 1Var Freq .

On revient à l'écran précédent avec EXIT. On lance la commande 1VAR en tapant sur F1 et ce qu'on recherche est affiché :

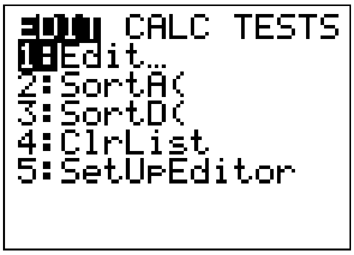


et en déscendant avec ▼

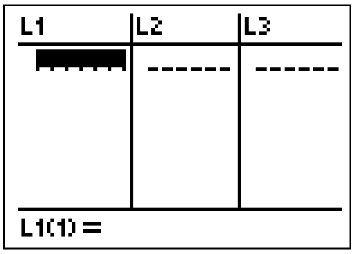


G 2 TI

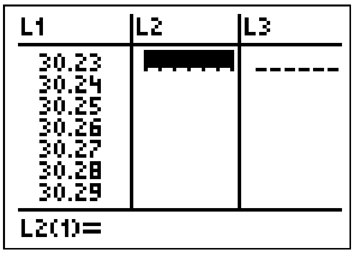
C'est plus simple...
D'abord, on rentre dans le module de statistique en tapant sur STAT et on aperçoit



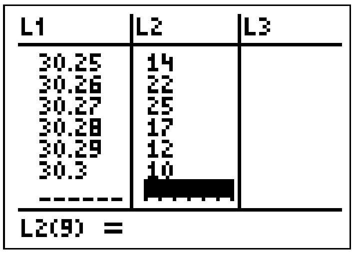
ensuite on va éditer la liste en appuyant sur 1.



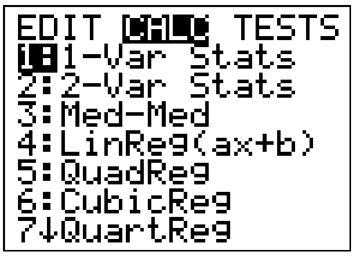
On entre les variables dans L1



et les effectifs ou les fréquences dans L2 . On passe à L2 avec ▸



On quitte ensuite cet écran avec 2nd QUIT puis on retourne dans le module statistique avec STAT.
Cette fois on va dans CALC avec ▸



On choisit 1-Var Stats

```
1-Var Stats █
```

puis on donne la colonne des variables en tapant [2nd][L1]

```
1-Var Stats L1
```

la virgule pour séparer [,]

```
1-Var Stats L1,
```

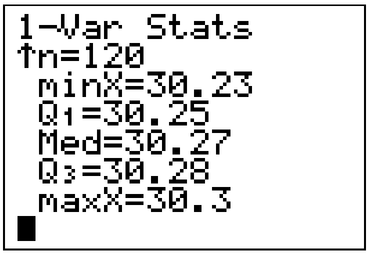
puis les fréquences ou les effectifs dans la deuxième colonne [2nd][L2]

```
1-Var Stats L1,L
2
```

et il ne reste plus qu'à taper [ENTER]

```
1-Var Stats
 x̄=30.26608333
 Σx=3631.93
 Σx²=109924.341
 Sx=.0194588915
 σx=.0193776432
↓n=120
█
```

et à faire défiler les résultats avec [▼]



H Des exercices

Exercice 1

1re Partie : Étude de l'évolution du prix du pétrole brut sur une période

La feuille de calcul de l'annexe 2 donne l'évolution entre le 2 mai 2008 (jour de cotation numéro 1) et le 27 juin 2008 (jour de cotation numéro 41) du cours du prix du baril de pétrole brut à New York et du cours de l'euro par rapport au dollar U. S, ce qui correspond à 41 jours de cotation. Les valeurs figurant dans les colonnes B à D sont arrondies au centième.

1. Le 2 mai 2008, le prix du baril de pétrole brut à New York était de 112,38 $.
Le cours de l'euro par rapport au dollar américain était de 1,55 ; cela signifie qu'un euro valait ce jour là 1,55 dollar américain.

a) Justifier que, le 2 mai 2008, le prix du baril était de 72,50 €.

b) Quelle formule alors saisir dans la cellule D3 pour obtenir, par recopie automatique vers le bas, les prix en euros du baril de pétrole ?

2. On répondra aux deux questions suivantes en utilisant les données de l'annexe 2.
Par quel coefficient, arrondi au millième, a été multiplié :

a) le prix du baril de pétrole en dollars entre le 2 mai 2008 et le 27 juin 2008 ? Arrondir à 0,001.

b) le prix du baril de pétrole en euros entre le 2 mai 2008 et le 27 juin 2008 ? Arrondir à 0,001.

3. a) Déduire de la question 2. les pourcentages d'évolution, arrondis au dixième, du prix du baril de pétrole en dollars, puis du prix du baril de pétrole en euros pendant la période du 2 mai 2008 au 27 juin 2008. *Arrondir à* 0,1 %

b) Comment expliquer le fait que ces pourcentages sont différents ?

2e Partie : Étude statistique :

1. On considère la série statistique construite avec les prix du baril de pétrole, arrondis à l'euro, entre le 2 mai 2008 et le 27 juin 2008 (41 journées de cotation).
Le tableau ci-dessous donne la répartition de ces prix arrondis à l'euro :

Prix	73	76	77	79	80	81	82	83	85	86	87	88	89
Effectifs	1	1	1	2	3	6	7	3	3	5	2	4	3

a) Déterminer la médiane, le 1er quartile et le 3e quartile de cette série.

b) Construire le diagramme en boîtes de cette série sur l'**annexe 1** (à rendre avec la copie).

2. Quel était le prix moyen, arrondi à l'euro, du baril de pétrole durant cette période ?

Annexe 2

	A	B	C	D
1	Journée	Prix du baril	Cours de	Prix du baril
2	numéro	en $	l'euro en $	en €
3	1 (2/05/08)	112,38	1,55	72,50
4	2	116,43	1,54	75,60
5	3	120,1	1,55	77,48
6	4	121,89	1,55	78,64
7	5	123,79	1,54	80,79
8	6	124,41	1,54	80,79
9	7	126,16	1,55	81,39
10	8	123,88	1,55	79,92
11	9	125,84	1,55	81,39
12	10	123,88	1,55	79,92
13	11	124,3	1,54	80,71
14	12	126,35	1,56	80,99
15	13	126,83	1,55	81,83
16	14	129	1,57	82,17
17	15	133,89	1,58	83,41
20	18	128,63	1,57	81,93
21	19	130,75	1,58	82,75
22	20	126,54	1,55	81,64
23	21	127,76	1,56	81,90
24	22	1027,6	1,55	82,32
25	23	124,3	1,54	79,30
27	25	128,18	1,56	82,17
28	26	137,86	1,58	87,25
29	27	134,66	1,56	84,85
31	29	136,75	1,56	87,66
32	30	136,72	1,54	87,48
33	31	134,72	1,54	87,48
34	32	133,94	1,55	86,41
35	33	133,8	1,55	86,32
36	34	136,27	1,55	87,92
37	35	131,82	1,55	88,55
38	36	134,84	1,56	86,44
39	37	137,25	1,55	88,55
40	38	137,06	1,56	87,86
41	39	134,6	1,57	85,73
42	40	139,67	1,58	88,40
43	41 (27/06/08)	140,7	1,58	89,05

Exercice 2

Un directeur de supermarché décide d'étudier le temps d'attente aux caisses de son établissement pour ajuster le nombre de caisses ouvertes à la demande. Pour cela, il interroge le lundi et le vendredi cent clients et note les temps d'attente approximatifs en minutes entières.

PARTIE A : Étude de l'échantillon du lundi

Le lundi, il obtient la répartition suivante :

Temps d'attente en caisse (en min)	1	2	3	4	5	6	7	8	9	10
Nombre de clients	14	13	23	9	14	8	12	4	1	2

1. Calculer le temps moyen d'attente aux caisses du supermarché pour l'échantillon étudié.

2. Déterminer la médiane et les quartiles de la série statistique des temps d'attente.

3. Construire sur la feuille annexe 2 le diagramme en boite de cette série.

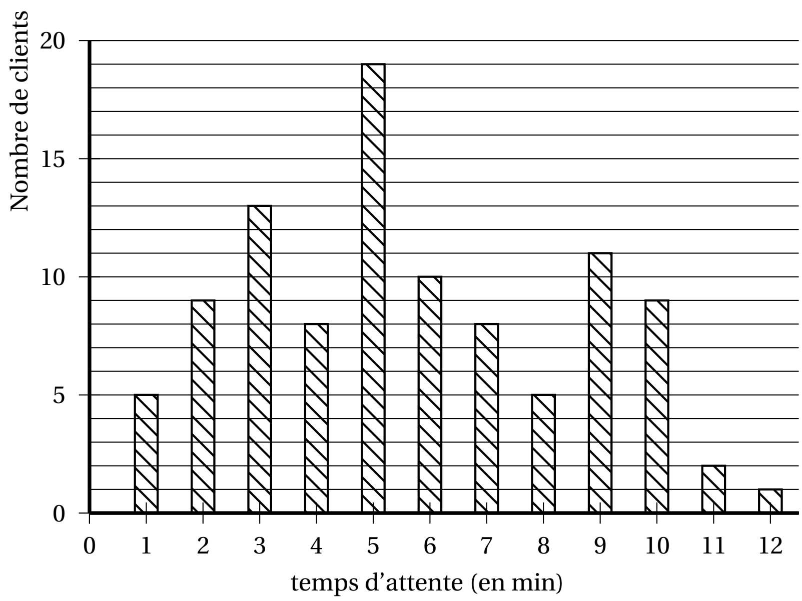
4. a) Son adjoint souhaite ouvrir une caisse supplémentaire si plus de 15 % des clients attendent 7 min ou plus en caisse. Doit-il ouvrir une nouvelle caisse le lundi ?
(On justifiera la réponse).

b) Le directeur décide d'ouvrir une caisse supplémentaire si le temps moyen d'attente aux caisses dépasse 5 min. Doit-il ouvrir une nouvelle caisse le lundi ?
(On justifiera la réponse).

PARTIE B : Étude de l'échantillon du vendredi

Le directeur décide de comparer les temps d'attente en début et en fin de semaine. Il a donc relevé le vendredi les temps d'attente aux caisses d'un échantillon de cent clients et obtient les résultats résumés dans le diagramme donné ci-dessous :

Temps d'attente le vendredi



1. Par lecture du diagramme, compléter le tableau donné en annexe 2.

2. Calculer le temps moyen d'attente aux caisses du supermarché le vendredi pour l'échantillon étudié (arrondi au dixième).

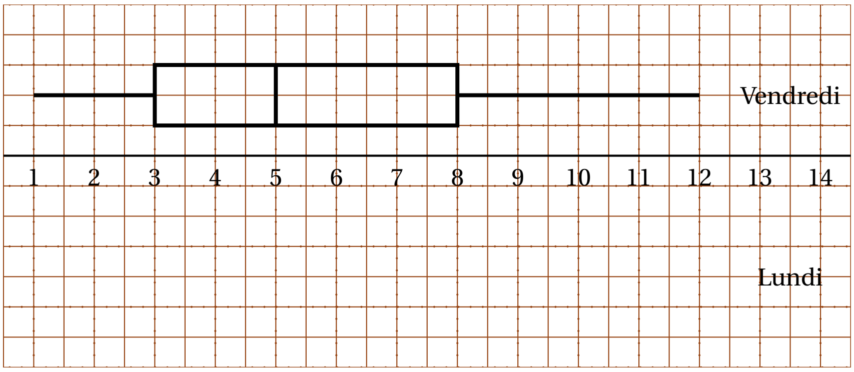
PARTIE C : Comparaison des deux échantillons

On a construit dans l'annexe 2 le diagramme en boîte de la série des temps d'attente aux caisses le vendredi. Dans un questionnaire, les clients qualifient d'acceptable un temps d'attente compris entre 2 et 6 minutes inclus.

Pour chacune des affirmations suivantes, dire si elle est vraie ou fausse en justifiant la réponse.

Affirmation A :
Le vendredi, la moitié des clients attendent cinq min ou plus de cinq min en caisse.
Affirmation B :
Le vendredi, un quart des clients attend au plus trois minutes en caisse.
Affirmation C :
Ily a autant de clients qui trouvent le temps d'attente acceptable le lundi que le vendredi.
À rendre avec la copie

Diagramme en boîte des séries



Partie B

Tableau de la série du vendredi

Temps d'attente en caisse (en min)	1	2	3	4	5	6	7	8	9	10	11	12
Nombre de clients	5	9		8		10				9	2	1

Exercice 3

Une entreprise qui produit du chocolat, fabrique des tablettes de 100 grammes. Au début de l'année 2010, elle décide de prélever un échantillon dans sa production afin d'en vérifier la masse.

Les résultats sont consignés en **annexe 2**.

1. a) Calculer la masse moyenne μ, exprimée en grammes, des tablettes de cet échantillon. (Arrondir au dixième)
   b) On admet que l'écart-type σ de cette série est environ égal à $1,6$.
   Déterminer le pourcentage des tablettes de chocolat dont la masse est dans l'intervalle $[\mu-2\sigma\ ;\ \mu+2\sigma]$.
   Ce résultat est-il en cohérence avec un modèle gaussien ? Expliquer pourquoi.
2. a) Déterminer la médiane et les quartiles de l'échantillon 2010.
   b) Dessiner le diagramme en boîte correspondant sur **l'annexe 3** en dessous de l'axe.
   Vous placerez en extrémités les valeurs minimum et maximum de la série.
   c) Un échantillon de même taille a été prélevé fin 2009, son diagramme en boîte se trouve également en annexe 3.
   Donner les valeurs du minimum, du maximum, des quartiles et de la médiane de l'échantillon 2009.

3. Les assertions suivantes sont-elles vraies ou fausses ? Justifier la réponse.

a) En fin 2009, environ trois-quarts des tablettes de chocolats avaient une masse supérieure à 98 g.

b) L'écart interquartile a été réduit de plus de moitié entre fin 2009 et début 2010.

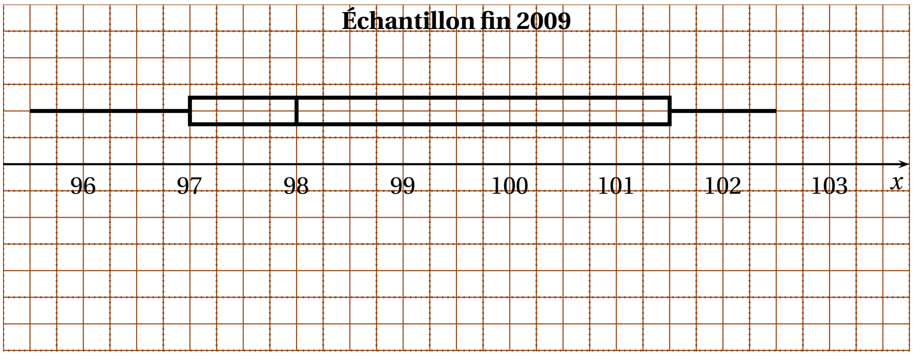
c) Le consommateur qui achète des tablettes produites par cette entreprise en fin 2009 peut se sentir lésé.

ANNEXE 2

Masse des tablettes de chocolat

Masse (en grammes)	96	97	98	99	100	101	102	103
Effectif	5	6	9	13	32	16	5	4

ANNEXE 3



Échantillon début 2010

Exercice 4

Un site de vente de livres par Internet désire réaliser une étude statistique de sa clientèle, afin de prévoir l'évolution de ses ventes pour les années à venir.
L'étude concerne l'âge de la clientèle. Pour répondre à cette question, les responsables de l'étude utilisent un échantillon de 2 100 clients, parmi les plus réguliers du site.
Les résultats sont résumés dans le tableau ci-dessous :

Classe d'âge	[18 ;20[	[20 ;24[	[24 ;30[	[30 ;36[	[36 ;46[	[46 ;56[	56 ans et +	total
Effectif	190	300	360	450	400	200	200	2 100

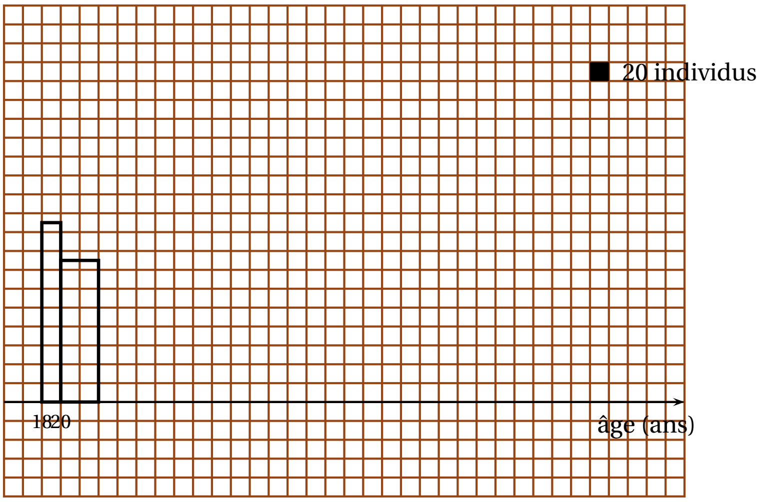
- On assimilera la dernière classe d'âge à l'intervalle [56 ; 76[.
- On fera l'hypothèse de l'uniforme répartition de l'effectif dans chaque classe d'âge.

1. À l'aide du quadrillage figurant en **annexe 2**, représenter ces données par un histogramme où un carreau représente 20 individus.

2. En utilisant les centres des classes pour valeurs du caractère, déterminer l'âge moyen m et l'écart type s de la série. On ne demande pas de justification, les valeurs seront arrondies au dixième.

3. Hachurer clairement l'histogramme pour faire apparaître l'effectif correspondant à la classe d'âge $[m - s\ ;\ m + s]$. Calculer le pourcentage de clients de cette classe d'âge par rapport à l'effectif de l'échantillon (arrondir à 1 %).

Annexe 2 à rendre avec la copie



Aide à la construction de l'histogramme (utilisation non obligatoire et donc tableau non évalué)

Classe d'âge	[18 ; 20[	[20 ; 24[	[24 ; 30[	[30 ; 36[	[36 ; 46[	[46 ; 56[	[56 ; 76[
Effectif	190	300	360	450	400	200	200
Nombre de carreaux	9,5	15					
Largeur	1	2					
Hauteur	9,5	7,5					

Exercice 5

Une association organise chaque année un tournoi comprenant à la fois des épreuves physiques et intellectuelles. Les inscriptions se font par équipe.

PARTIE 1 : étude de deux équipes

L'équipe des « lucioles » participe à cette manifestation ; les âges de ses 24 membres sont :
10 - 10 - 12 - 13 - 14 - 15 - 15 - 17 - 21 - 25 - 33 - 33 - 33 - 34 - 35 - 36 - 40 - 44 - 44 - 44 - 45 - 46 - 59 - 60

Dans cette partie 1, on ne demande pas de justifier.

1. Donner l'âge moyen des « lucioles » (arrondir à l'unité).
2. Donner la médiane Me, le premier quartile Q_1 et le troisième quartile Q_3 de cette équipe.
3. Tracer sur **l'annexe 1** le diagramme en boîte des âges des membres de cette équipe (comme pour le diagramme déjà tracé, les extrémités représentent le minimum et le maximum de la série).
4. On a représenté sur **l'annexe 1** le diagramme en boîte des âges des membres de l'équipe des « gazelles ».
   On considère dans cette question qu'une équipe est « jeune » si au moins 50 % de ses membres ont moins de 30 ans.
   Pour chacune des propositions suivantes, indiquer si elle est vraie, fausse ou indécidable (indécidable signifie ici que les informations données ne permettent pas de conclure si la proposition est vraie ou fausse).
   Pour cette question : on attribue $0,5$ *point par réponse exacte, on retranche* $0,25$ *point par réponse inexacte, l'absence de réponse n'est pas pénalisée. Un éventuel total négatif est ramené à* 0.
   a) Aucune gazelle n'a plus de 65 ans.
   b) Au moins la moitié des gazelles a plus de 25 ans.
   c) L'âge moyen des gazelles est compris entre 25 ans et 30 ans.
   d) L'équipe « les gazelles » est jeune.
   e) L'équipe « les lucioles » est jeune.

PARTIE 2 : étude de l'ensemble des participants

L'**annexe 2** représente une page automatisée de calcul :

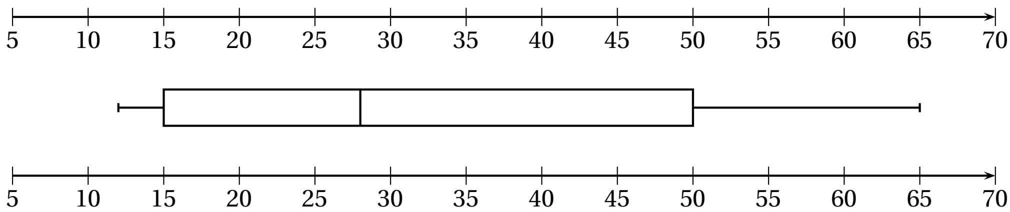
- Le tableau 1 de l'**annexe 2** donne les effectifs des participants en fonction de l'âge et du sexe.
- Le tableau 2 de l'**annexe 2**, calculé automatiquement à partir du tableau 1, indique les pourcentages par rapport au total général.
  La plage de cellules C11 : J13 est au format pourcentage arrondi à deux décimales.

1. Compléter la case D11.
2. Donner une formule qui, écrite dans la cellule C11 puis recopiée dans les cellules de la plage C11 : J13 du tableau 2 permet d'obtenir les pourcentages indiqués.
3. a) Calculer le pourcentage des femmes parmi les participants de la tranche 36-45 ans.
   b) Donner précisément la signification du pourcentage 12,83 % figurant dans la cellule F12.
4. Dans le tableau 3 de l'**annexe 2**, les cellules de la plage C18 : J20 sont au format pourcentage.
   Dans la cellule C18, on entre la formule **=C4/C$6** puis on recopie cette formule dans les cellules de la plage C18 : J20.
   Répondre aux questions suivantes sans justifier :
   a) Quelle est la signification du pourcentage calculé en C18 ?
   b) Quelle formule obtient-on en D19 ?

ANNEXE 1

Les lucioles

Les gazelles



ANNEXE 2

	A	B	C	D	E	F	G	H	I	J	K
1											
2						**TABLEAU 1**					
3			0-15	16-25	26-35	36-45	46-55	56-65	66 et plus	Total	
4		Homme	75	124	175	251	193	99	25	942	
5		Femme	46	83	135	203	73	83	17	640	
6		Total	121	207	310	454	266	182	42	1582	
7											
8											
9						**TABLEAU 2**					
10			0-15	16-25	26-35	36-45	46-55	56-65	66 et plus	Total	
11		Homme	4,74 %		11,06%	15,87%	12,20%	6,26%	1,58%	59,54%	
12		Femme	2,91%	5,25%	8,53%	12,83%	4,61%	5,25%	1,07%	40,46%	
13		Total	7,65%	13,08%	19,60%	28,70%	16,81%	11,50%	2,65%	100,00%	
14											
15											
16						**TABLEAU 3**					
17			0-15	16-25	26-35	36-45	46-55	56-65	66 et plus	Total	
18		Homme									
19		Femme									
20		Total									
21											