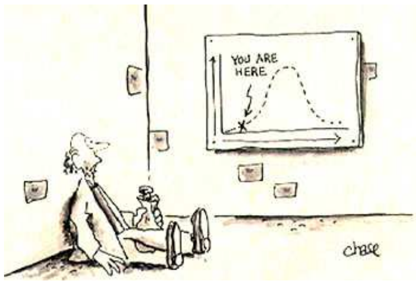
Deuxième Leçon

STATISTIQUES



I - Des mathématiques en sciences humaines et sociales

a. La démarche scientifique en sciences humaines et sociales

Les sciences humaines sont souvent amenées à tirer des enseignements d'observations de certains faits, tout comme les autres sciences. Par exemple, on peut, dans un cadre sociologique, étudier l'affirmation suivante : « les filles réussissent mieux leurs études que les garçons ».
On peut avoir une opinion sur cette affirmation fondée sur son expérience, le sens commun, ses propres valeurs. Il faudrait cependant être capable de l'étudier à partir de faits bien choisis. Quelles informations faut-il obtenir afin de porter un jugement éclairé sur cette affirmation ? Par exemple :
– quelles sont les sources de cette affirmation ?
– A-t-on des données pour le confirmer ?
– Comment ces données ont-elles été collectées ?
– Peut-on comparer les réussites scolaires ?
– etc.
En sciences physiques ou naturelles, par exemple, le scientifique procède souvent ainsi pour analyser un phénomène :
– observer et mettre en évidence sa régularité ;
– le décrire ;
– l'expliquer ;
– le prédire, le maîtriser.
Pour ce qui concerne les sciences humaines, la prédiction est plus difficile car il est délicat de prédire le comportement des individus, de l'économie, etc.

> L'utilisation massive de techniques statistiques (graphiques, indices, corrélation, régression, analyse de Fourier) dans les années 1920, appliquées à un objet d'étude omniprésent - le cycle économique -, a conduit à la multiplication des baromètres (et des instituts de conjoncture). Mais leur faillite, perceptible dès 1925, est flagrante en 1929.[1]

Et depuis 1929, les absences de prédictions des crises ultérieures s'est perpétuée...
Il est cependant possible de suivre une démarche scientifique en sciences sociales en formulant une problématique, en choisissant ou en élaborant des instruments de mesure, en collectant des données, en les traitant, en les analysant et en les diffusant. Une différence notable demeure : il est difficile d'adopter la démarche expérimentale des autres sciences. Par exemple, on ne va pas déclencher une guerre en France pour étudier ses effets sur la population ; on ne peut pas non plus diviser le prix du baril de pétrole par 100 pour vérifier certaines hypothèses... On favorise donc l'enquête en sciences sociales.

b. Les études statistiques

Définir la (les) statistique(s) relève de l'exploit : il existe au moins une centaine de définitions différentes. Nous nous contenterons de parler d'études statistiques dans un cadre mathématique. Nous nous bornerons même en classe de 1$^{\text{ère}}$ d'étudier la *statistique descriptive* qui a pour objet la représentation de données par des tableaux, des graphiques, des mesures. Vous

[1] in Michel Armatte et Alain Desrosières : Sous la direction de Jean-Pierre Beaud et Jean-Guy Prévost L'ère du chiffre. Presses de l'Université du Québec, 2000

abordererez l'an prochain de manière extrêmement ponctuelle la *statistique inférentielle* qui s'occupe de généraliser à toute une population des résultats observés sur un échantillon ou à les projeter dans le futur.

c. Acquérir un esprit critique

« La quantification constitue un moyen de dépasser l'intuition mais comme elle permet seulement de refléter la réalité sociale de façon parcellaire, il est recommandé d'exercer son sens critique et de s'interroger sur la manière par laquelle les données ont été obtenues ainsi que sur les raisons de leur publication[2] ».

On pourra à ce propos lire cet article paru dans le TIME le 26 avril 1993 :
« Damned lies and statistics » : http://www.time.com/time/magazine/article/0,9171,978314,00.html

II - Les variables

a. 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**.

b. 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.

c. 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.

III - Tableaux et graphiques

a. Variables qualitatives ou quantitatives discrètes (peu de valeurs distinctes)

i. 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 interrogées :

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

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

[2] in AMYOTTE, Luc Méthodes quantitatives. ERPI, Québec, 2002

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

FIG. 2 – Répartition des 75 emprisonnés selon le sexe (*Source : ministère syldave des 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

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

ii. 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 %

FIG. 4 – 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 en bâtons En face de chaque modalité représentée régulièrement en abscisses on trace un « bâton » de longueur proportionnelle à la fréquence :

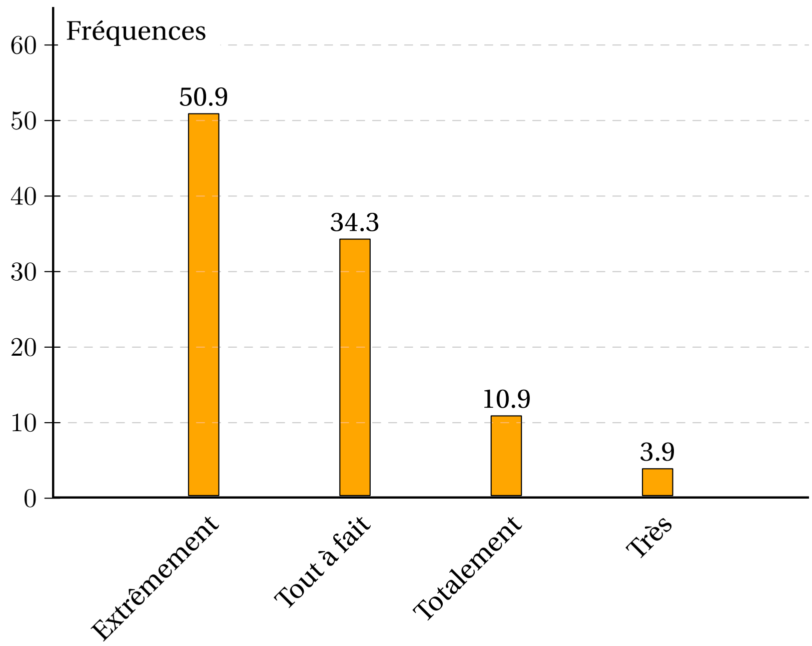


FIG. 5 – Répartition en pourcentage des Syldaves selon leur degré de satisfaction à l'égard du GPS (*Source : ministère syldave des la justice - BO n°58 p. 554*) - Diagramme en bâtons

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°

FIG. 6 – 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*) - Calcul des angles du diagramme circulaire

Et on obtient le camembert suivant :

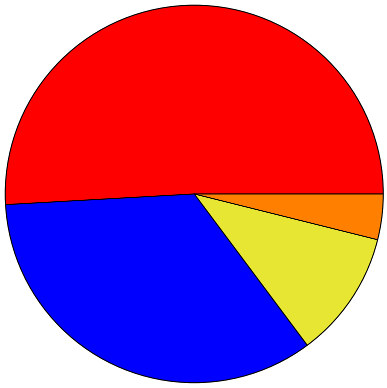


FIG. 7 – Répartition en pourcentage des Syldaves selon leur degré de satisfaction à l'égard du GPS (*Source : ministère syldave des la justice - BO n°58 p. 554*) - Diagramme circulaire

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

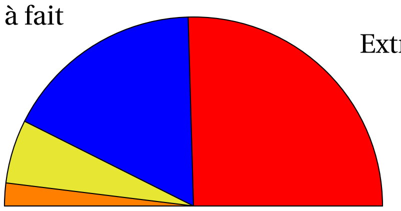


FIG. 8 – Répartition en pourcentage des Syldaves selon leur degré de satisfaction à l'égard du GPS (*Source : ministère syldave des la justice - BO n°58 p. 554*) - Diagramme semi-circulaire

b. Variables quantitatives discrètes (grand nombre de valeurs) ou continues

i. 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

FIG. 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

FIG. 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%

FIG. 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*)

ii. 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.

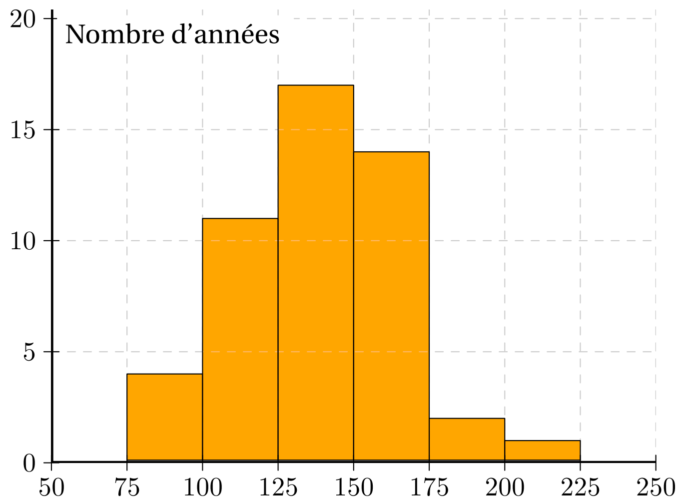


FIG. 12 – Répartition de 49 années selon le nombre annuel de disques vendus par la femme du Grand Protecteur de la Syldavie de 1960 à 2008 (*Source : ministère syldave de la culture - BO n°51 p. 1254559*) - Histogramme

iii. 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

FIG. 13 – Répartition des femmes du harem du Grand Protecteur de la Syldavie 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 proportionnelles aux effectifs de chaque classe. Par exemple, la 3^{e} 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 :

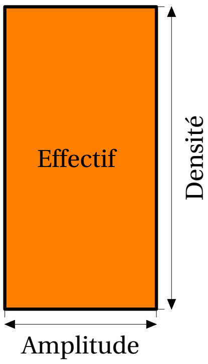


FIG. 14 – Construction d'un histogramme à pas non constant

On remplit donc le tableau suivant :

Intervalles	[15 ; 20[	[20 ; 25[	[25 ; 35[	[35 ; 45[	[45 ; 65[	[65 ; 75]
Effectifs	202	204	359	338	304	10
Amplitudes	5	5	10	10	20	10
Densites	40.4	40.8	35.9	33.8	15.2	1.0

FIG. 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 :

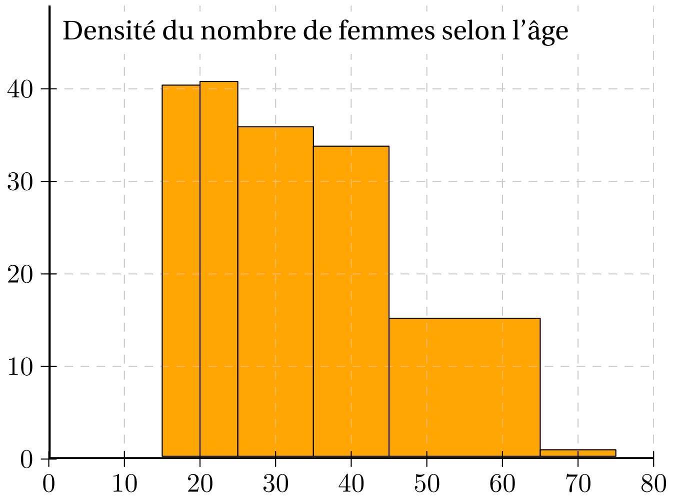


FIG. 16 – Répartition des femmes du harem du Grand Protecteur de la Syldavie selon l'âge(*Source : ministère syldave du travail- BO n°7*) -Histogramme

IV - 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			

FIG. 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 %

FIG. 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 %

FIG. 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 :

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

FIG. 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.

V - Séries chronologiques

Lorsqu'on étudie l'évolution d'une variable dans le temps, on obtient une série chronologique.

a. Mesures sur les séries chronologiques

Considérons par exemple les données suivantes :

Année	Nombre d'invités à la garden party du GPS
1987	6 491
1992	5 858
1997	5 042
2002	5 418
2007	4 744

FIG. 21 – Évolution du nombre d'invités à la garden party annuelle du GPS (*Source : ministère syldave du protocole - BO n°2*)

i. Variation absolue

Ici, la variation du nombre d'invités entre 1997 et 2007 a été de

$$\Delta V_a = V(2007) - V(1997) = 4744 - 5042 = -298$$

Définition 1 : Variation absolue

La variation absolue mesure l'augmentation ou la diminution de la valeur d'une variable entre deux moments donnés. Elle s'exprime dans les mêmes unités que la variable (sauf le cas particulier des pourcentages comme on le verra au ii.).

$$\Delta V_a = V(t_2) - V(t_1)$$

ii. Cas particulier des variables exprimées en pourcentage

Voici un tableau donnant l'évolution des suffrages obtenu par le GPS lors des précédentes élections :

Année	pourcentage des votants ayant voté pour le GPS
1948	99,89%
1968	99,91%
1988	99,97%
2008	99,92%

FIG. 22 – Évolution de la proportion des personnes ayant voté pour le GPS lors des 4 dernières élections dans l'ensemble des votants (*Source : ministère syldave des services secrets - BO n°56*)

Il serait faux de dire que l'augmentation des suffrages a été de 0,06 % entre 1968 et 1988 car les pourcentages donnés pour ces années correspondent à des nombres de votants sûrement différents. On parle plutôt d'une variations de 0,06 **points**.

variation de pourcentage

Lorsque la variable d'une série chronologique est en pourcentage, la variation s'exprime en POINTS.

iii. Variation relative

Définition 2 : Variation relative

La variation relative est le pourcentage de diminution ou d'augmentation de la valeur d'une variable entre deux moments donnés par rapport à la valeur initiale :

$$\Delta V_r = \frac{V(t_2) - V(t_1)}{V(t_1)}$$

Elle s'exprime en pourcentage.

Si on reprend le tableau 21 page précédente, la variation relative entre 1997 et 2007 est de :

$$\Delta V_r = \frac{4\ 744 - 5\ 042}{5\ 042} \approx -5{,}9\ \%$$

On observe donc une baisse de 5,9 % du nombre d'invités entre 1997 et 2007.

Remarque 1 : Variation relative de pourcentages

On peut mesurer des variations relatives de pourcentages car on divise des grandeurs de même nature. Par exemple, si on reprend le tableau 22 page ci-contre et qu'l'on calcule la variation relative de suffrages entre 1968 et 1988 :

$$\Delta V_r = \frac{99,97 - 99,91}{99,91} \approx 0,06\ \%$$

On observe alors une augmentation relative de 0,06 % de pourcentage de suffrages en faveur du GPS entre 1968 et 1988.

iv. Variation moyenne

Définition 3 : Variation moyenne

C'est la variation par unité de temps. Elle s'exprime en unités de la variable par unité de temps. On la calcule grâce à la formule :

$$\Delta V_m = \frac{V(t_2) - V(t_1)}{t_2 - t_1}$$

Reprenons une nouvelle fois le tableau 21 page 9 pour mesurer la variation moyenne du nombre d'invités à la garden party entre 1987 et 2007 :

$$\Delta V_m \frac{V(2007) - V(1987)}{2007 - 1987} = \frac{4\ 744 - 6\ 491}{20} \approx -87,4$$

On observe donc une baisse moyenne de 87,4 invités par an.

v. Comment commenter une série chronologique ?

On vous demande de commenter le tableau 21 page 9. Voici comment vous pourriez procéder en utilisant ce que nous venons de voir.

En Syldavie (le lieu), entre 1987 et 2007 (la période), le nombre annuel d'invités à la garden party du GPS a diminué de manière importante (le phénomène).
De 6 491 en 1987, le nombre d'invités est passé à 4 744 en 2007 (les valeurs extrêmes).
Ainsi, en 2007, on a dénombré 1 747 invités en moins qu'en 1987 (variation absolue), ce qui représente une baisse moyenne annuelle de 87,4 invités (variation moyenne), soit une baisse de 26,9 % pour cette période (variation relative).

b. Lissage par les moyennes mobiles

Voici un tableau qui donne l'extension de la banquise au minimum de septembre de 1979 à 2007 :

Année	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93
Extension ($\overline{\text{M}}\,\text{k}m^2$)	5,3	5,5	4,95	5,13	5,37	4,7	5	5,37	5,3	5,2	4,8	4,62	4,5	5,05	4,45
Année	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08
Extension ($\overline{\text{M}}\,\text{k}m^2$)	4,85	4,4	5,25	4,9	4,26	4,2	4,14	4,55	4,05	4,12	4,3	4,05	4,06	3,75	3,8*

FIG. 23 – Extension de la banquise au minimum de septembre de 1979 à 2007 en millions de km^2 (*Source :*[3])

Représentons cette évolution par un nuage de points :

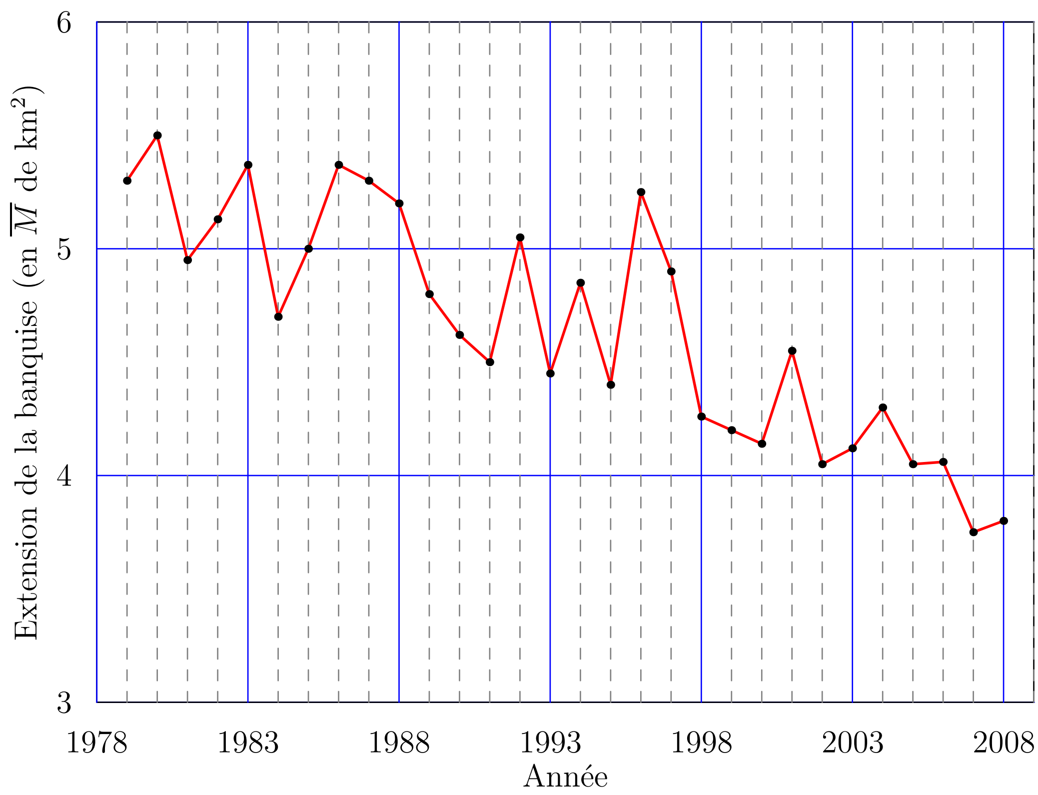


FIG. 24 – Extension de la banquise au minimum de septembre de 1979 à 2007 en millions de km^2 - Graphe

On « sent » une tendance à la baisse mais des variations saisonnières la masque. Pour y remédier, on peut remplacer chaque terme par sa moyenne avec les termes voisins.

Définition 4 : moyenne mobile

On appelle moyenne mobile d'ordre k la moyenne arithmétique d'un terme avec les k termes voisins.

EN PRATIQUE

- si k est impair, c'est-à-dire s'écrit sous la forme $k = 2p + 1$ alors on remplace chaque terme par la moyenne de ce terme, des p termes suivants et des p termes précédents ;
- Si k est pair, c'est-à-dire s'écrit sous la forme $k = 2p$, c'est moins évident : on remplace chaque terme par la moyenne de ce terme, des p termes suivants et des p termes précédents mais les termes extrêmes sont affectés d'un coefficient 1/2.

Définition 5 : Lissage par moyennes mobiles

Lisser une série chronologique c'est remplacer la série initiale par la série des moyennes mobiles.

Par exemple, reprenons les données du tableau 23 page précédente et effectuons un lissage par moyennes mobiles d'ordre 5. Il y a 30 termes au départ. L'ordre 5 étant impair, on va remplacer chaque terme par la moyenne de ce terme, des 2 précédents et des deux suivants : on est donc obligé de commencer par le troisième terme et de terminer par le vingt-huitième.

On remplacera alors t_3 par

$$t'_3 = \frac{t_1 + t_2 + t_3 + t_4 + t_5}{5} = \frac{5,3+5,5+4,95+5,13+5,37}{5} = 5,25$$

Remplissez le tableau suivant :

Année	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93
Extension ($\overline{\mathrm{M}}\mathrm{k}m^2$)	/	/	5,25	5,13	5,03	...	...	5.114	5.134	5.058	4,884	4,834	...	4,694	4,65
Année	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08
Extension ($\overline{\mathrm{M}}\mathrm{k}m^2$)	...	...	4,732	4,602	4,55	4,41	4,24	4,212	4,232	4,214	4,116	4,056	...	/	/

FIG. 25 – Extension de la banquise au minimum de septembre de 1979 à 2007 en millions de km^2 lissé par moyennes mobiles d'ordre 5

On obtient un nouveau graphique qui met plus nettement en évidence la décroissance :

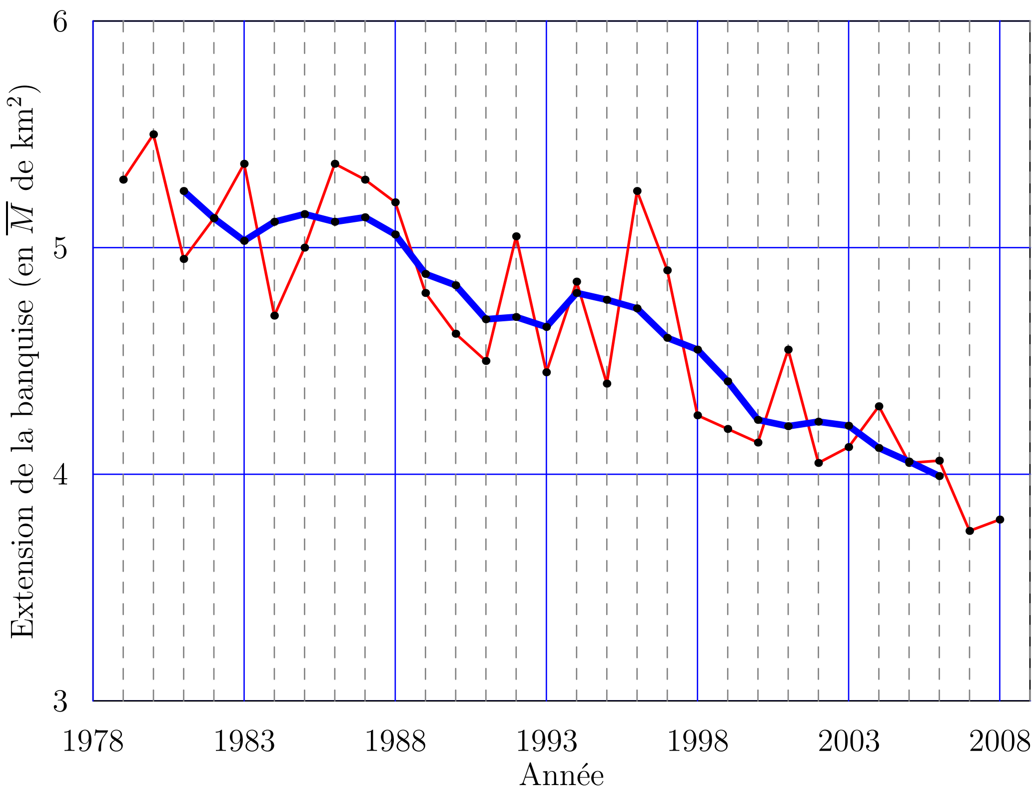


FIG. 26 – Extension de la banquise au minimum de septembre de 1979 à 2007 en millions de km^2

VI - Mesures de tendance centrale

a. Le mode

i. Définition

Définition 6 : 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.

ii. Données groupées par valeurs ou modalités

Si on reprend l'exemple du tableau 22 page 10, le mode est « extrêmement satisfait ».

iii. Données groupées par classes

Reprenons cette fois le graphique 12 page 6. 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.

iv. Distribution polymodale

Considérons le nouvel exemple suivant :

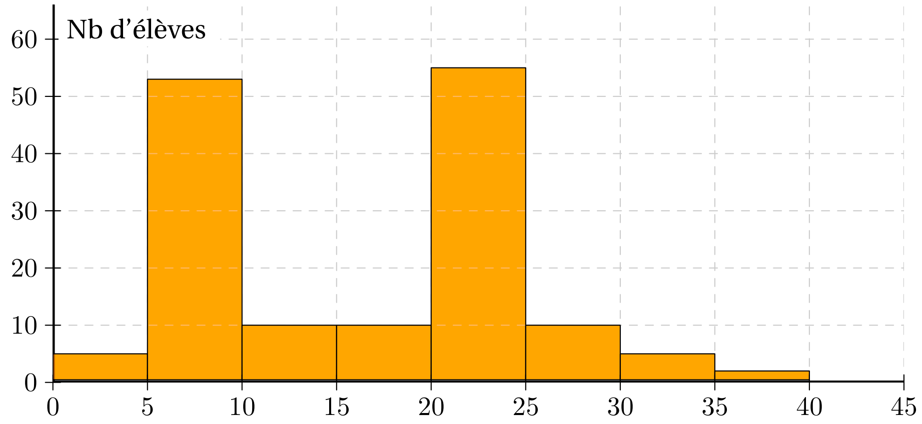


FIG. 27 – Répartition des prisonniers syldaves selon la durée de leur peine (en années)

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

b. La médiane

i. Conventions

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

Exemple 1 :

Si on étudie par exemple le nombre de poupées Barbue que possèdent les élèves de la classe de 1ère ES 1
- 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 2 :

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

ii. Définition

Définition 7 : 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 2 : calcul de la médiane

- $V = \{1,2,2,5,5,8,8,9,37\}$

▷ $M_e = 5$

- $V = \{1,2,2,5,8,8,9,37\}$

▷ $M_e = \dfrac{5+8}{2} = 6,5$

iii. 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	

FIG. 28 – 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			

FIG. 29 – 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 %.

iv. 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 %	

FIG. 30 – 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 :*

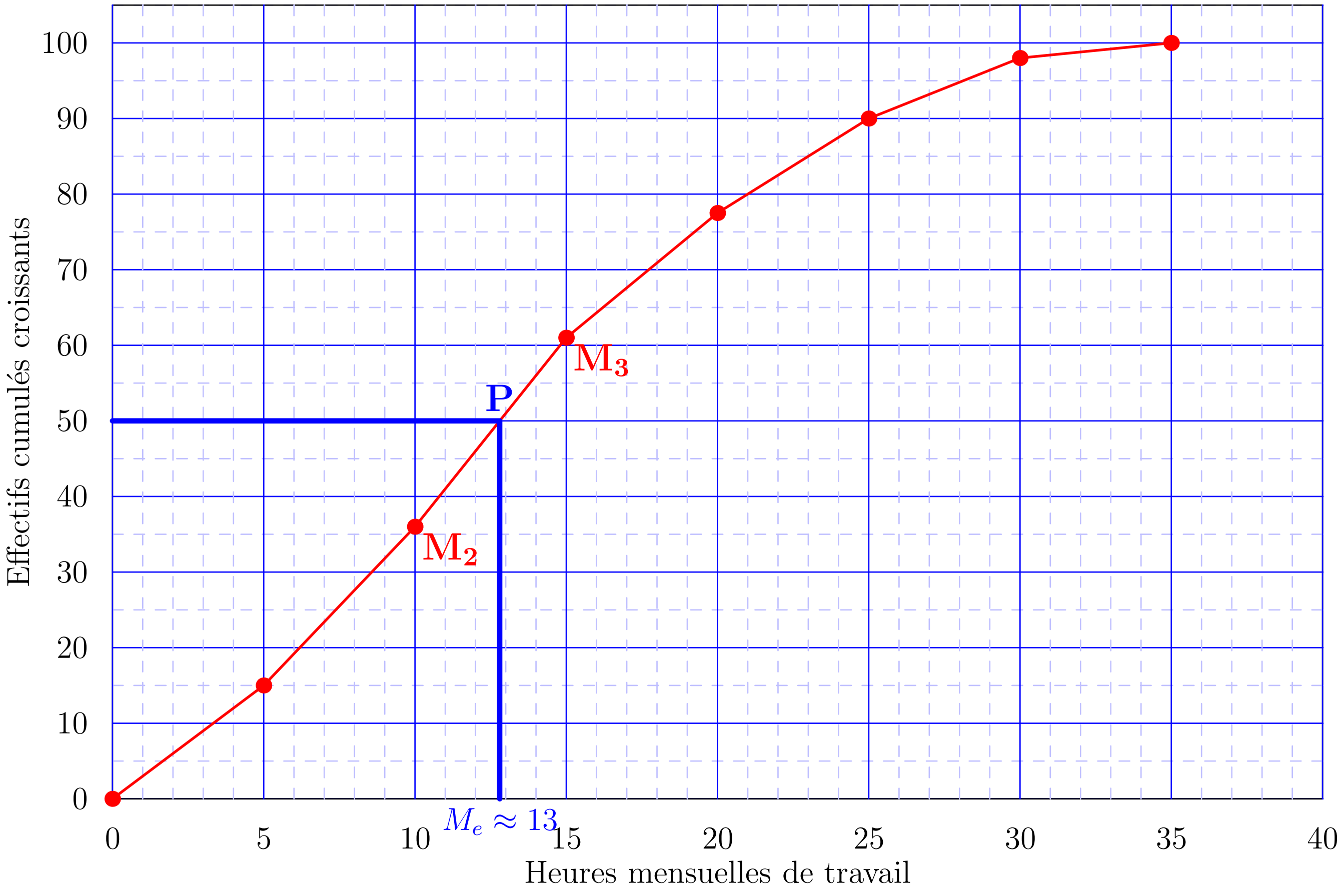


FIG. 31 – Polygone des fréquences cumulées croissantes des 2 000 élèves des lycées Perrin et Goussier de Rezé selon le nombre d'heures de travail personnel par mois.
À chaque borne supérieure des classes, on fait correspondre la fréquence cumulée et on relie les points par des segments (cela constitue 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 $M_e \approx 12,8$.

c. Moyenne

i. Données non groupées

Soit $x_1, x_2, \dots, x_n$ 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 $\sum\limits_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5$.
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 ?

ii. Données groupées par valeurs

Reprenons le tableau 28 page 15. 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, $\overline{x} = \frac{1}{125}(69 \times 0 + 31 \times 1 + 15 \times 2 + 6 \times 3 + 3 \times 4 + 1 \times 5) = \cdots$

iii. 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 7 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

FIG. 32 – 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 :

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

VII - Quartiles et diagrammes en boîte

a. 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.

b. Expérimentons

Exemple 3 : huit éléments

$V = \{1, 2, 3, 4, 5, 6, 7, 8\}$

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

$$V = \{1, 2, 3, 4, 5, 6, 7, 8\}$$

On peut prendre
- $Q_1 = \frac{2+3}{2} = 2,5$ et 25% des effectifs ont une valeur inférieure à Q_1
- $M_e = \frac{4+5}{2} = 4,5$ et 50% des effectifs ont une valeur inférieure à M_e
- $Q_3 = \frac{6+7}{2} = 6,5$ et 75% des effectifs ont une valeur inférieure à Q_3

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

Exemple 4 : dix éléments

$V = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$

On peut séparer l'effectif en parties hautes et basses

$$V = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

On peut prendre
- $Q_1 = 3$ et $3/10 = 30\% (\geqslant 25\%)$des effectifs ont une valeur inférieure à Q_1
- $M_e = \frac{5+6}{2} = 5,5$ et 50% des effectifs ont une valeur inférieure à M_e
- $Q_3 = 8$ et $8/10 = 80\% (\geqslant 75\%)$ des effectifs ont une valeur inférieure à Q_3

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

Exemple 5 : onze éléments

$V = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}$

On peut séparer l'effectif en parties haute et basse

$$V = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}$$

On peut prendre

- $Q_1 = 3$ et $3/_{11} \simeq 27\%(\geqslant 25\%)$des effectifs ont une valeur inférieure à Q_1
- $M_e = 6$ et 50% des effectifs ont une valeur inférieure à M_e
- $Q_3 = 9$ et $9/_{11} = 82\%(\geqslant 75\%)$ des effectifs ont une valeur inférieure à Q_3

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

Exemple 6 : neuf éléments

$V = \{1,2,3,4,5,6,7,8,9\}$

On peut séparer l'effectif en parties haute et basse avec la médiane au milieu :

$$V = \{1,2,3,4,5,6,7,8,9\}$$

et prendre $Q_1 = \frac{2+3}{2} = 2,5$, mais...
... seulement $2/_9 \simeq 22\%$ des effectifs ont une valeur inférieure à Q_1, ce qui est contraire à notre cahier des charges.
Dans ce cas particulier, on va inclure la médiane dans les parties basses et hautes pour éviter cet écueil.
Cela donne :

- Partie Basse = $\{1,2,3,4,5\}$ donc $Q_1 = 3$ et $3/_9 \simeq 33\% \geqslant 25\%$ des effectifs ont une valeur inférieure à Q_1, ce qui convient.
- Partie Haute = $\{5,6,7,8,9\}$ donc $Q_3 = 7$ et $7/_9 \simeq 78\% \geqslant 75\%$ des effectifs ont une valeur inférieure à Q_1, ce qui convient.

Remarque 3 : cycle

Il est aisé de constater que ces observations vont se répéter par cycle de longueur 4.
C'est quand V a un nombre d'éléments égal à un multiple de 4 plus 1 que nous devons être prudents.
D'ailleurs, toutes les machines à calculer ne s'accordent pas sur le calcul des quartiles.
La méthode que je vous propose est la plus cohérente et sera en accord avec la détermination graphique des quartiles que nous verrons bientôt.
Elle permet également d'avoir une définition rigoureuse comme nous allons le voir.

C'est quand V a un nombre d'éléments égal à un multiple de 4 plus 1 que nous devons être prudents.

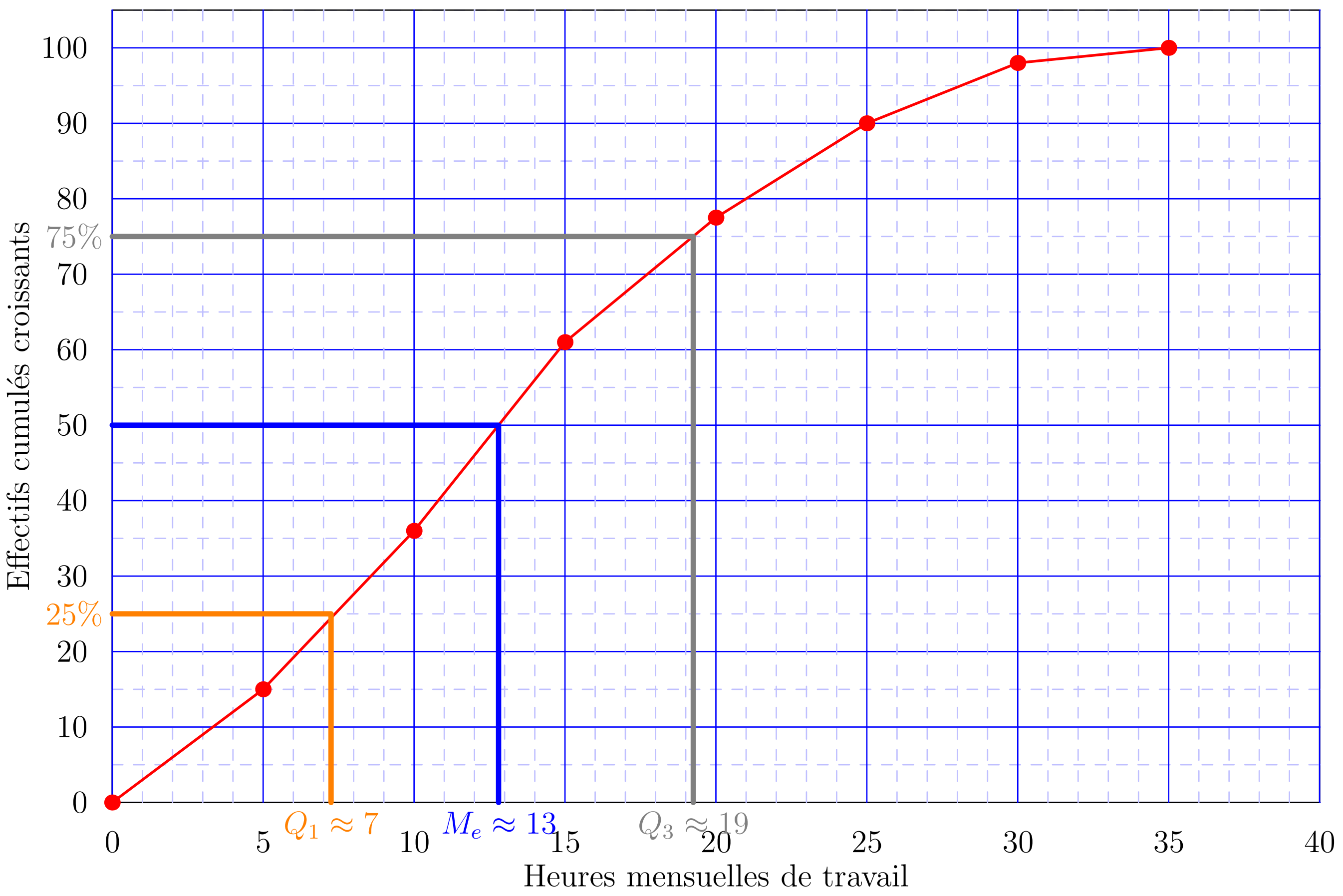
c. Définissons

Définition 8 : 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.

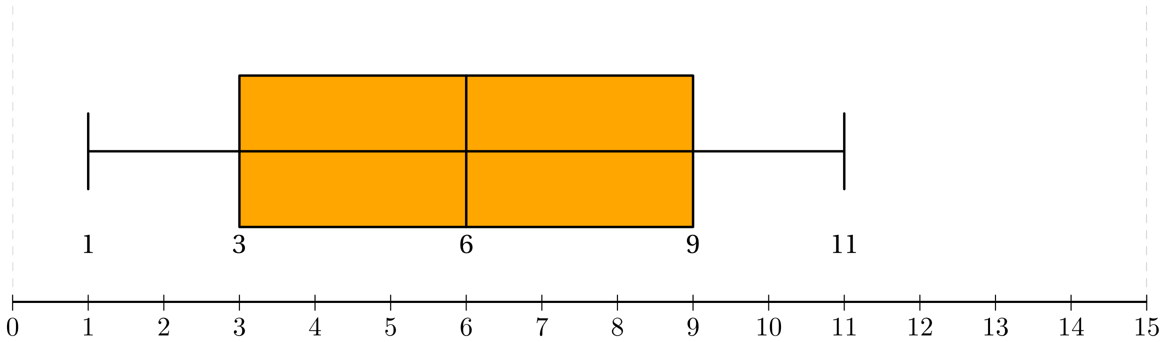
d. Lecture graphique dans le cas des données groupées en classe

Complétons le graphique 31 page 17 qui avait été obtenu grâce au tableau 30 page 16 :



e. 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 5 page 19



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.*

VIII - Mesures de dispersion

a. 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 9 : écart interquartile

La distance entre Q_1 et Q_3 est appelé écart interquartile.

Observons deux exemples :

Exemple 7 : 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 :

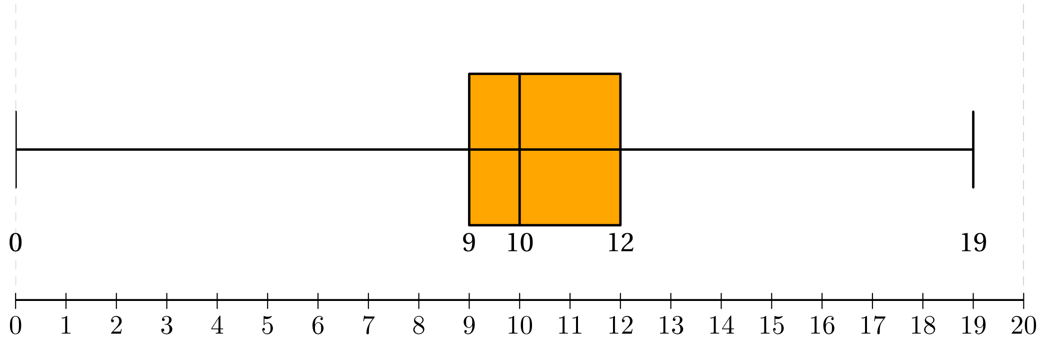


FIG. 33 – Notes au Bac de la classe A

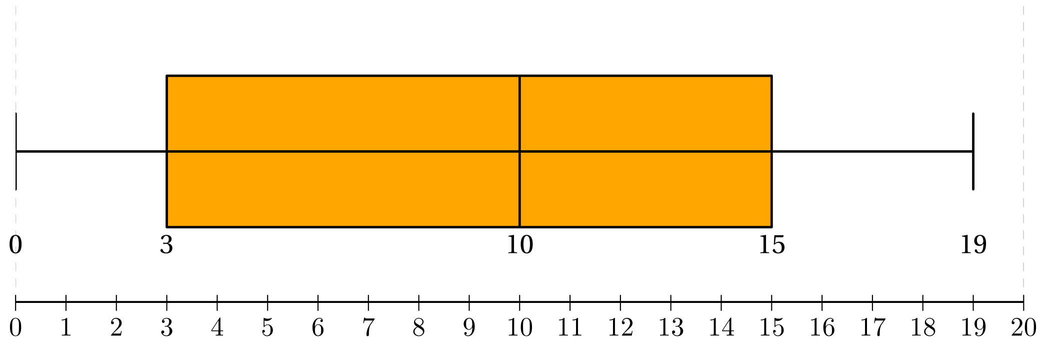


FIG. 34 – Notes au Bac de la classe B

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.

b. 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

FIG. 35 – 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 $\overline{x}$.

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

FIG. 36 – 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 :

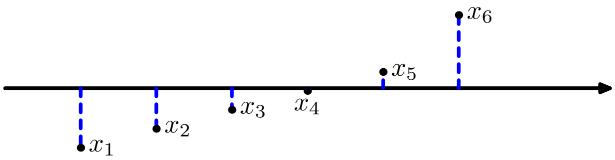


FIG. 37 – Visualisation de l'écart par rapport à la moyenne

On peut prouver (à titre d'exercice...) que l'écart correspondant à la première ligne du tableau 35 est toujours nul. On préfère donc utiliser la moyenne des écarts de la deuxième ligne qu'on appelle **variance**.

Définition 10 : 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 4 : cas d'un regroupement en classe

dans le cas d'un regroupement en classe, on considère, comme pour le calcul de la moyenne (cf iii. page 18), 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 11 : 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 7 page 22 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 $[\overline{x}-\sigma\,;\overline{x}+\sigma]$:

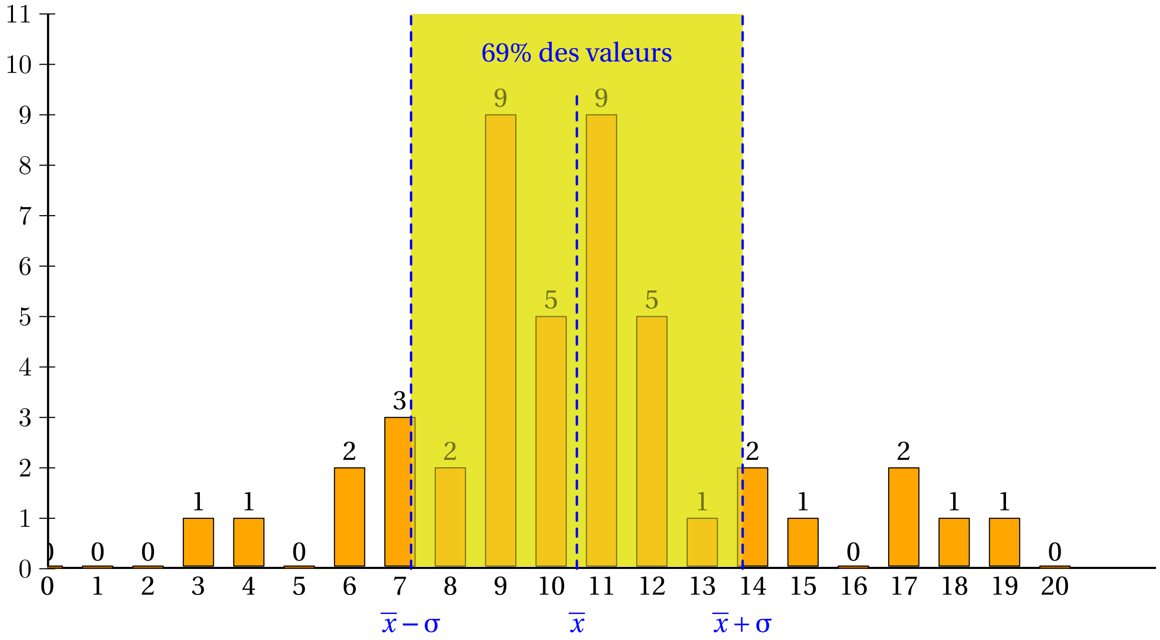


FIG. 38 – Notes au Bac de la classe A

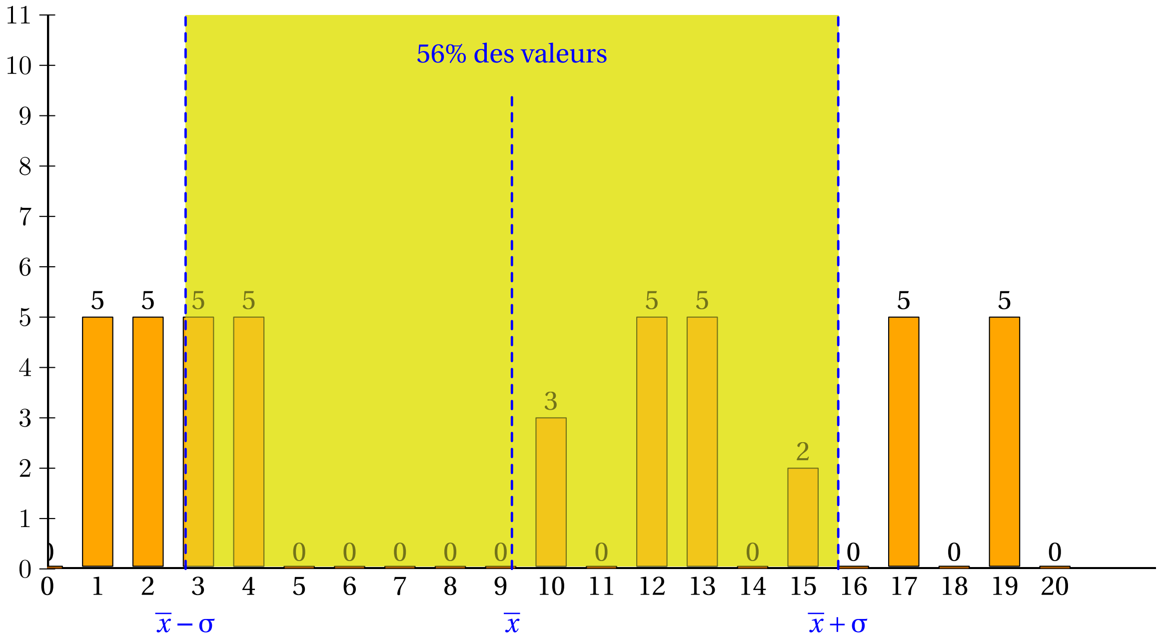


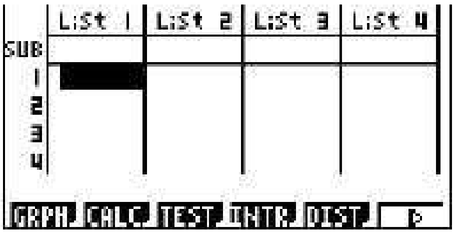
FIG. 39 – Notes au Bac de la classe B

IX - Utilisation de la calculatrice

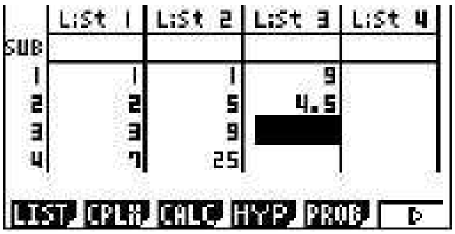
a. Casio

i. Manipulations de listes

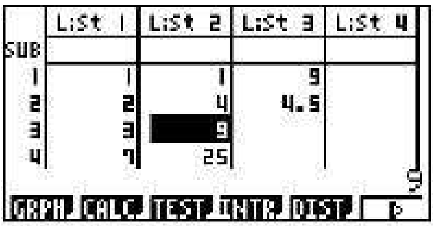
On commence par aller dans le menu STAT [STAT] et obtient cet écran :



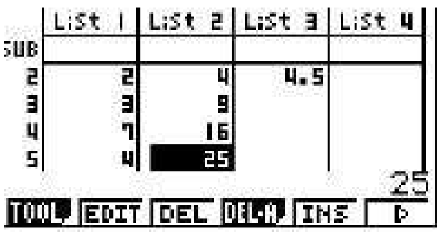
Les touches de curseur [▶] et [◀] permettent de choisir sa liste. On entre ensuite successivement les valeurs dans la liste choisie en validant à chaque fois avec [EXE].
Par exemple entrer cette suite de commandes :
[1] [EXE] [2] [EXE] [3] [EXE] [7] [EXE] [4] [EXE] [5] [EXE]
[▶]
[1] [EXE] [5] [EXE] [9] [EXE] [2] [5] [EXE]
[▶]
[9] [EXE] [4] [•] [5] [EXE]
permet d'afficher :



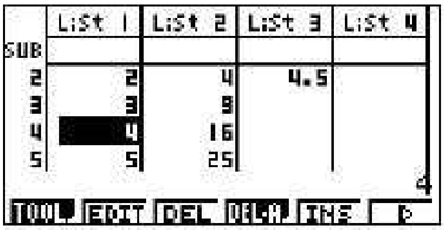
Pour *remplacer une entrée*, mettez-la en surbrillance, entrez la nouvelle valeur et validez :
[◀] [▲] [4] [EXE]
et on obtient :



Pour *insérer une entrée* en cas d'oubli, appuyez sur [F6] (qui correspond à la petite flèche pour avoir les autres icônes d'édition, mettez la cellule cible en surbrillance, appuyez sur [F5] qui correspond à l'INSertion puis entrez la nouvelle valeur et validez :
[F6] [▼] [F5] [1] [6] [EXE]

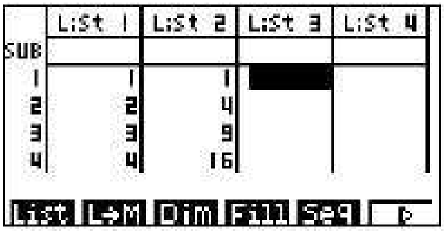


Pour *supprimer une entrée*, mettez-la en surbrillance, appuyez sur [F3] qui correspond à [DEL] (delete : effacer) :
◀ ▲ [F3]



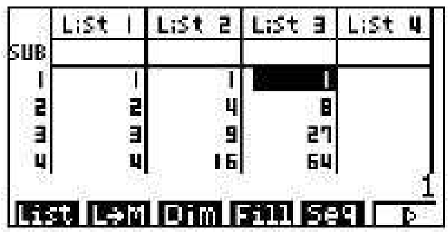
Pour *supprimer une liste*, mettez-la en surbrillance, appuyez sur [F4] qui correspond à [DEL-A] puis [F1] qui correspond à [YES] :
▶ ▶ [F4] [F1]

Pour *effectuer des opérations sur les listes* il faut accéder à un nouveau menu en tapant sur [OPTN] [F1] qui correspond à [List] :
[OPTN] [F1]



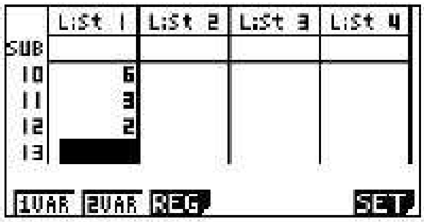
À l'aide des flèches de curseur, vous mettez la nouvelle liste en surbrillance et entrez à l'aide du menu [List] un calcul sur les listes déjà existantes. Par exemple pour obtenir le produit des deux premières listes dans la troisième on effectue :
▲ [F1] (List) [1] [×] [OPTN] [F1] (LIST) [F1] (List) [2] [EXE]

puis



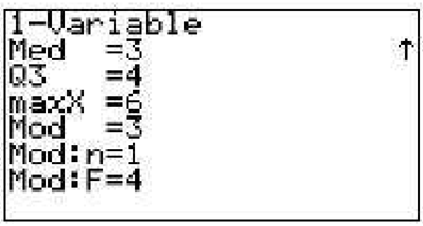
ii. Obtenir moyenne, médiane, quartiles, etc.

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.

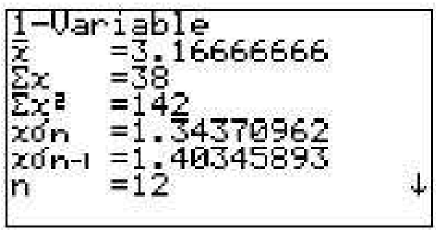


Pour vérifiez que la machine est bien configurée, il faut aller sur SET en appuyant sur F6
Il faut placer List1 devant 1Var XList et List2 devant 1Var Freq .

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



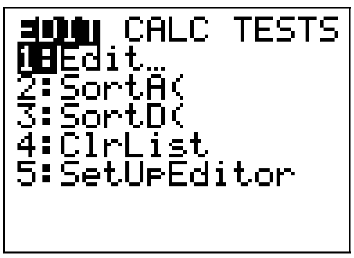
et en déscendant avec ⊙



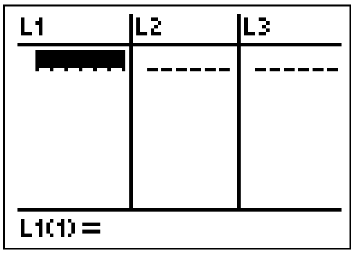
b. TI

i. Obtenir moyenne, médiane, quartiles, etc.

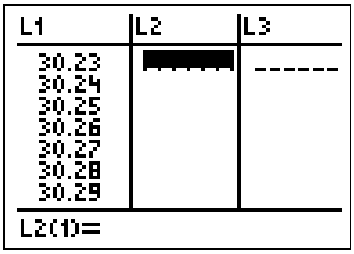
C'est plus simple...
D'abord, on rentre dans le module de statistique en tapant sur STAT et on aperçoit



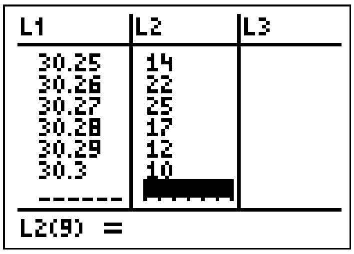
ensuite on va éditer la liste en appuyant sur [1].



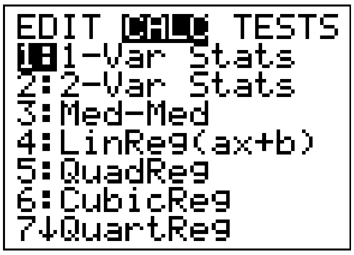
On entre les variables dans [L1]



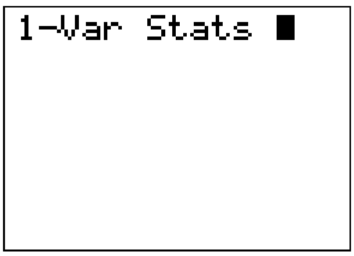
et les effectifs ou les fréquences dans [L2] . On passe à L2 avec [▸]



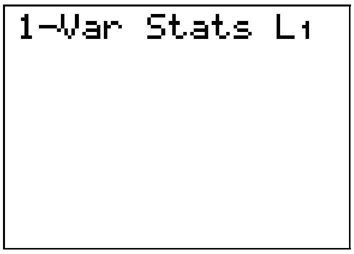
On quitte ensuite cet écran avec [2nd][QUIT]puis on retourne dans le module statistique avec [STAT].
Cette fois on va dans [CALC] avec [▸]



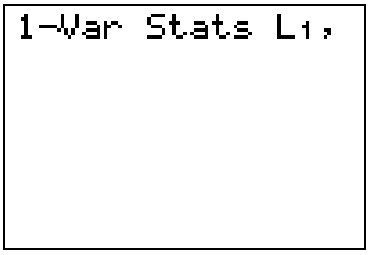
On choisit [1-Var Stats]



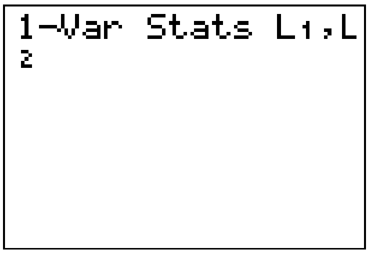
puis on donne la colonne des variables en tapant [2nd][L1]



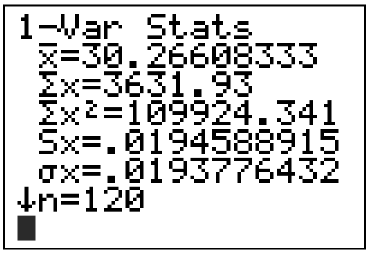
la virgule pour séparer [,]



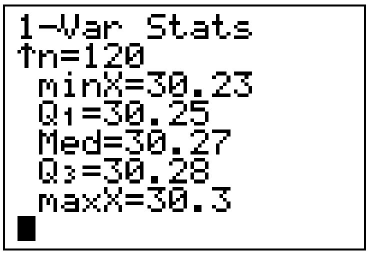
puis les fréquences ou les effectifs dans la deuxième colonne [2nd][L2]



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



et à faire défiler les résultats avec [▼]



ii. Manipulations de listes

c'est encore plus simple...
On se positionne sur une autre liste et [L1][×][L2][ENTER]et voilà...

Références

Sous la direction de Jean-Pierre Beaud et Jean-Guy Prévost : L'ère du chiffre. Presses de l'Université du Québec, 2000, Systèmes statistiques et traditions nationales

Amyotte, Luc: Méthodes quantitatives. ERPI, Québec, 2002

Gauvrit, Nicolas: Statistiques Méfiez-vous ! Paris: Ellispes, 2007

HENRY, Valéry: Une étude sur les quartiles d'une série statistique univariée. dans Bulletin Vert. APMEP, 2006, n°464

Malik, Daniel: La calculatrice TI-82 Stats.fr. Mai 2008, Lycée La Ramée ⟨URL: http://etablissements.ac-amiens.fr/0020049t/E

Urbana-Champaign, University of Illinois at: Department of Atmospheric sciences. septembre 2008, Polar Research Group ⟨URL: http://arctic.atmos.uiuc.edu/cryosphere/⟩

Table des matières

I - Des mathématiques en sciences humaines et sociales . . . 1
- a. La démarche scientifique en sciences humaines et sociales . . . 1
- b. Les études statistiques . . . 1
- c. Acquérir un esprit critique . . . 2

II - Les variables . . . 2
- a. Vocabulaire . . . 2
- b. Variables qualitatives . . . 2
- c. Variables quantitatives . . . 2

III - Tableaux et graphiques . . . 2
- a. Variables qualitatives ou quantitatives discrètes (peu de valeurs distinctes) . . . 2
  - i. Tableau des effectifs/ des fréquences . . . 2
  - ii. Représentations graphiques . . . 3
- b. Variables quantitatives discrètes (grand nombre de valeurs) ou continues . . . 5
  - i. Groupement par classes . . . 5
  - ii. Histogramme avec classes de même amplitude . . . 6
  - iii. Histogramme avec classes d'amplitudes différentes . . . 7

IV - Tableaux à double entrée . . . 8

V - Séries chronologiques . . . 9
- a. Mesures sur les séries chronologiques . . . 9
  - i. Variation absolue . . . 9
  - ii. Cas particulier des variables exprimées en pourcentage . . . 10
  - iii. Variation relative . . . 10
  - iv. Variation moyenne . . . 11
  - v. Comment commenter une série chronologique ? . . . 11
- b. Lissage par les moyennes mobiles . . . 11

VI - Mesures de tendance centrale . . . 13
- a. Le mode . . . 13
  - i. Définition . . . 13
  - ii. Données groupées par valeurs ou modalités . . . 14
  - iii. Données groupées par classes . . . 14
  - iv. Distribution polymodale . . . 14
- b. La médiane . . . 14
  - i. Conventions . . . 14
  - ii. Définition . . . 15
  - iii. Cas des données groupées par valeurs . . . 15
  - iv. Cas des données groupées par classes . . . 16
- c. Moyenne . . . 17
  - i. Données non groupées . . . 18

ii. Données groupées par valeurs 18
iii. Données groupées par classes 18

VII - Quartiles et diagrammes en boîte **19**
a. L'idée 19
b. Expérimentons 19
c. Définissons 20
d. Lecture graphique dans le cas des données groupées en classe 20
e. Boîte à moustache 21

VIII - Mesures de dispersion **22**
a. Écart interquartile 22
b. Variance et écart-type 23

IX - Utilisation de la calculatrice **25**
a. Casio 25
i. Manipulations de listes 25
ii. Obtenir moyenne, médiane, quartiles, etc. 27
b. TI 27
i. Obtenir moyenne, médiane, quartiles, etc. 27
ii. Manipulations de listes 29