STATISTIQUES

L'étude d'un **caractère** (caractéristique) d'une **population** (des objets) à partir de critères statistiques permet une représentation synthétique de l'information sous forme de tableaux par exemple, plus faciles à comprendre et à interpréter. Elle rend plus évidentes certaines tendances de la série, permet d'anticiper quelques réactions de la population (les objets !), d'essayer de prévoir la continuité du phénomène observé.
Remarque : une étude statistique correspond à une réduction de l'observation effective… synthétiser l'information c'est tronquer une partie de la réalité !

I. Caractéristiques de valeur centrale, de position :

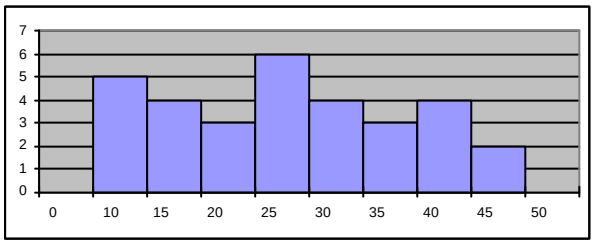
Notre habitude est de tout rattacher à la moyenne. Pourtant, ce n'est pas toujours le résultat le plus significatif… « un statisticien s'est noyé dans une mare d'une profondeur moyenne de 1 mm ! », ou, résultats de trois élèves : E_1 {10 ; 10 ; 10} ; E_2 {20 ; 10 ; 0} ; E_3 {5 ; 10 ; 15}. Ils ont tous la même moyenne. Leurs situations sont-elles identiques ?

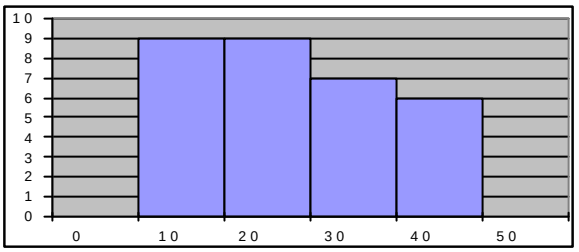
1. Le mode :

Def : j'appelle **mode** ou **dominante** d'une série statistique toute **valeur** ou **classe** ou **modalité** (il peut y en avoir plusieurs) d'**effectif** ou **fréquence** maximal. N'est intéressant que s'il est unique, et vraiment supérieur aux autres valeurs.

S_1 * : Classe :	[10;15[	[15;20[	[20;25[	[25;30[	[30;35[	[35;40[	[40;45[	[45;50]
Valeur :	5	4	3	6	4	3	4	2

* : Série 1





a) classe modale de la série ? rep : [25 ; 30[
b) regrouper en classes d'amplitude 10 (oui c'est fait, c'est le deuxième graphique !). Mode(s) ?

2. Médiane : (pour une série quantitative aux valeurs ordonnées).

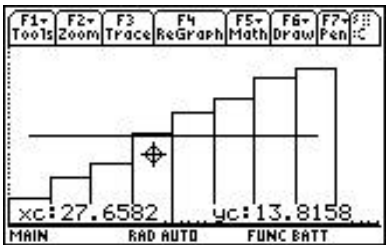
Def : j'appelle **médiane** tout nombre M tel que l'effectif des valeurs inférieures à M et l'effectif des valeurs supérieures à M ne dépasse pas la moitié de l'effectif total.

En gros, c'est la valeur pour laquelle se trouve la moitié de l'effectif total de chaque côté (dessous-dessus).

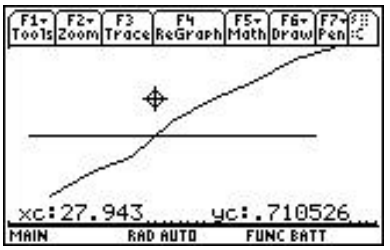
Ex : S_2 :{6 ; 8 ; 9 ; 10 ; 10 ; 11 ; 12 ; 13 ; 14 ; 15 ; 15} médiane 11 (5 avant, 5 après, et $5 \leq 11/2$).
S_3 :{5 ; 7 ; 8 ; 9 ; 9 ; 10 ; 10 ; 11 ; 12 ; 12} médiane tout nombre de [9 ; 10]. En général, le centre de l'intervalle est pris comme médiane (ici 9,5).

S_1 : Classe :	[10;15[	[15;20[	[20;25[	[25;30[	[30;35[	[35;40[	[40;45[	[45;50[
Effectifs :	5	4	3	6	4	3	4	2
S_{1a} : Valeur :	15	20	25	30	35	40	45	50
Effectifs cumulés >	5	9	12	18	22	25	29	31
Fréquences cumulées >	0,16	0,29	0,39	0,58	0,71	0,81	0,94	1,00
S_{1b} : Valeur :	10	15	20	25	30	35	40	45
Effectifs cumulés <	31	26	22	19	13	9	6	2
Fréquences cumulées <	1,00	0,84	0,71	0,61	0,42	0,29	0,19	0,06

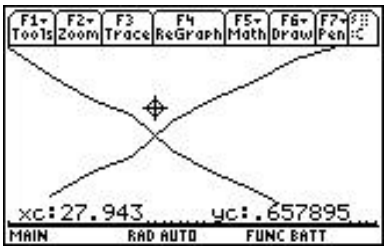
Dans la cas de données groupées par classes, construire un tableau comme ci-contre,
puis l'**histogramme** des effectifs ou des fréquences cumulées (fréquences mieux que effectifs car 50 % se voit bien), ou tracer les courbes des fréquences cumulées croissantes et décroissantes, l'intersection donnant la médiane.… suite ⇒



Effectifs cumulés et 1/2 effectifs d'où la médiane : dans la classe [25 ; 30[(curseur)



Fréquences cumulées > et 50 % d'où médiane à 27,9 environ (curseur) …
(valeurs x_i : bornes sup)



Fréquences cumulées > et < dont l'intersection se trouve à environ…
(valeurs x_i fréq. < : bornes inf)

Pour ceux qui aiment calculer, $M = B_{inf} + a_M \times \left(\dfrac{\frac{N}{2} - \sum n_{inf}}{n_M} \right)$ avec :

B_{inf} : borne inférieure de la classe médiane.
n_M : effectif de l'intervalle médian.
$\sum n_{inf}$:somme des effectifs des classes inférieures à l'intervalle médian.
a_M : amplitude de l'intervalle médian.

$M = 25 + 5 \times \left(\dfrac{31/2 - (5+4+3)}{6} \right) \approx 27{,}92.$

Remarque : la médiane est utile (représentative) dans les études statistiques sur les prix, salaires, ages, où la distribution n'est pas trop irrégulière.

3. **Moyennes :** (pour une série quantitative).

a) **Arithmétique** (pondérée) **:** celle bien connue !

N° valeur (i)	1	2	3	…	p
Valeur :	x_1	x_2	x_3	…	x_p
Effectif :	n_1	n_2	n_3	…	n_p

Je pose : $N = \sum_{i=1}^{p} n_i$. **Def :** J'appelle **moyenne** le nombre $\overline{x} = \dfrac{\sum_{i=1}^{p} n_i x_i}{N}$.

Ex : S_4={6 ; 8 ; 10 ; 10 ; 11 ; 12 ; 13 ; 14 ; 15 ; 15} moyenne de cette série ?

i	1	2	3	4	5	6	7	8	9
X_i	6	8	9	10	11	12	13	14	15
n_i	1	1	1	2	1	1	1	1	2

Commencer par faire le tableau statistique (ci-contre), puis calculer la moyenne :
$\overline{x}$=(6x1+8x1+9x1+10x2+11x1+12x1+13x1+14x1+15x2)/(11)≈9,82.

Pour 100 véhicules de 5 ans					
Distance parcourue en km	<85000	85000 à 95000	95000 à 105000	105000 à 1150000	>115000
Effectif :	10	18	40	20	12

Pour une série (S_5) donnée par classes, faire le tableau en considérant que chaque classe est équivalente à son centre. Pour les valeurs extrêmes, suivre l'énoncé, ou choisir en fonction des cas (le pif !).

i	1	2	3	4	5
$S5_a$: Valeur :	80000	90000	100000	110000	120000
Effectif :	10	18	40	20	12

Ici, précision de l'énoncé : « on considère que les classes extrêmes sont de même amplitude que les autres ».
$\overline{x}$=106 000. (VOUS devez écrire la formule, puis le résultat. Il sera **éventuellement** demandé quelques calculs).

b) **moyenne géométrique :** (valeurs positives)

Année (rang)	1	2	3	4	5	6	7	8	9	10
Pourcentage d'augmentation	1	3	2	3	8	1	5	2	3	2
Coefficient multiplicateur	1,01	1,03	1,02	1,03	1,08	1,01	1,05	1,02	1,03	1,04

Ci-contre (S_6) le tableau du pourcentage d'augmentation d'un produit sur dix ans. J'ai ajouté une ligne « coefficient multiplicateur » dont l'usage vous est familier… non ?

Quelle est la moyenne du pourcentage d'augmentation ? Je trouve 3 % ou 1,03 pour le coefficient.
Est-ce représentatif du phénomène observé ?

Je traduis : pour 100 F, si j'applique une augmentation de 3 % par an j'obtiens $100 \times (1{,}03)^{10} = 134{,}40$ F.

En réalité je devrais avoir $100 \times (1{,}01 \times 1{,}03 \times 1{,}02 \times 1{,}03 \times 1{,}08 \times 1{,}01 \times 1{,}05 \times 1{,}02 \times 1{,}03 \times 1{,}04) = 136{,}77$ F.

Ce que je cherche c'est donc « quel coefficient multiplicateur élevé à la puissance 10 donne le bon résultat ? ».

Nouvelle traduction : $100\times\left(1+\frac{x}{100}\right)^{10}=136{,}77$. Un détour par la leçon sur les logarithmes vous permettra de trouver seuls la jolie formule qui correspond à ceci : $G=\sqrt[\Sigma n_i]{\prod x_i^{n_i}}$

Ma calculatrice me dit : $G=\sqrt[10]{1{,}01^2\times1{,}02^2\times1{,}03^3\times1{,}04\times1{,}05\times1{,}08}=1{,}0318\ldots$

Remarque : sur la calculatrice, 'racine dixième' s'écrit '^ 1/10'.

La différence, pourrait sembler minime entre 1,03 et 1,0318. Pourtant, 136,77 ce n'est pas 134,40. Et appliquez donc ceci aux augmentations de budget d'un pays ! l'erreur serait énorme.

c) **<u>moyenne harmonique</u> :** (valeurs strictement positives)

Le problème : si je roule une heure à 90 kmh^{-1}, puis une heure à 60 kmh^{-1}, ma moyenne est de (90+60)/2=75 kmh^{-1}. Fastoche, MAIS, si je parcours 60 km à 90 kmh^{-1}, puis 80 km à 60 kmh^{-1} ? c'est loin d'être aussi simple.

La résolution passe par la définition de la vitesse moyenne, $V=\frac{\text{distance parcourue}}{\text{temps mis pour la parcourir}}$.

Il faut 40 minutes pour le premier tronçon, puis 1 h 20' pour le second (vous trouvez pareil ?), la distance totale étant de 140 km. D'où la moyenne V=70 kmh^{-1}.

La moyenne harmonique est $h=\frac{N}{\sum\frac{n_i}{x_i}}$. Ce qui se traduit par : la moyenne harmonique est l'inverse de la moyenne arithmétique des inverses (coefficientés) des valeurs. Appliquons :

Valeurs :	90	60
Effectifs :	60	80

La vitesse moyenne est $h=\frac{d_1+d_2}{\frac{d_1}{V_1}+\frac{d_2}{V_2}}=70$ kmh^{-1}.

d) **<u>moyenne quadratique</u> :** (éventuellement, valeurs coefficientées)

Petite formule : $M_Q=\sqrt{\overline{X^2}}=\sqrt{\frac{\sum n_i(X_i)^2}{N}}$ qui se raconte : c'est la moyenne des carrés des valeurs.

4. **<u>Quantiles</u> :** (série ordonnée rangée par ordre croissant)

<u>Quartiles :</u> comme la médiane découpe l'intervalle d'étude en deux classes de mêmes effectifs, les quartiles coupent l'intervalle en quatre parts égales. La médiane correspond de par sa définition au deuxième quartile.

Remarque : que penser des déciles, centiles, béciles, etc. ?

D'autres outils de position existent (qu'est-ce que la médiale ?), réservons les aux spécialistes.

II. <u>Les caractéristiques de dispersion</u> :

Considérons les notes d'un contrôle de deux moitiés d'une classe de 34 élèves, soit 17 élèves par groupe :
S_7 : {8 ; 9 ; 9 ; 10 ; 10 ; 10 ; 11 ; 11 ; 11 ; 11 ; 11 ; 12 ; 12 ;12; 13 ; 13 ; 14}
S_8 : {2 ; 5 ; 5 ; 7 ; 7 ; 8 ; 10 ; 10 ; 11 ; 11 ; 11 ; 12 ; 14 ; 16 ; 19 ; 19 ; 20}

S_9

V_1	8	9	10	11	12	13	14
n_i	1	2	3	5	3	2	1

S_{10}

V_1	2	5	7	8	10	11	12	14	16	19	20
n_i	1	2	2	1	2	3	1	1	1	2	1

Ces deux groupes ont même dominante 11 (ça se Woua dirait mon chien), même médiane 11, même moyenne 11 (ça c'est ma calculatrice qui le dit, écrans ci-joints). Seraient-ils identiques ?. A l'œil, ils semblent bien différents ! Les caractéristiques centrales ne sont pas suffisantes pour comparer ces séries. Nous allons étudier comment elles s'écartent ou se rassemblent eu égard à leur position moyenne. Rappelons que les calculatrices donnent les résultats. Il vous appartient de connaître les formules, savoir entrer les données, puis de les exploiter.

1. **Etendue :**

Def : J'appelle **étendue** d'une série statistique la différence entre la plus grande et la plus petite valeur du caractère.
Pour les deux séries précédentes : étendue S_9 : 14-8=**6**, étendue S_{10} : 20-2=**18**. C'est déjà ≠.

2. **Ecart moyen :** (n'est pas trop utilisé, les calculatrices et ordinateurs faisant le calcul du suivant)

Petite formule : $e = \frac{1}{N}\sum_{i=1}^{p} n_i \left| x_i - \overline{x} \right|$.

3. **Ecart type :** (à la moyenne)

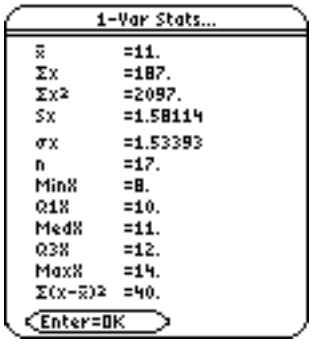
L'utilisation des valeurs absolues ne plait pas toujours…, élever au carré c'est obtenir forcément des nombres positifs. Mais pour obtenir la même dimension il faut prendre ensuite la racine carrée du résultat. D'où la démarche :

Je calcule la **Variance :** $V = \frac{1}{N}\sum_{i=1}^{p} n_i \left(x_i - \overline{x}\right)^2 = \frac{1}{N}\sum_{i=1}^{p} n_i {x_i}^2 - \overline{x}^2$, puis **l'écart type** : $s = \sqrt{V}$.

Dans la pratique, sauf à certains concours où il est obligatoire d'effectuer tous les calculs, j'écris les formules puis les résultats donnés par ma calculatrice.

Pour les deux séries précédentes (écrans page précédente), je lis Sx=1,58 pour la première, Sx=5,24 pour la seconde. (ATT : en réalité le premier s_x=1,534 et le second s_x=5,087 ! voir remarque qui suit).

Remarque : les machines (TI 89), conçues par de grosses têtes loin de l'autre côté de la mer, sont aux normes de ces gens là, ie, leurs considérations vont vers l'échantillon, ou si l'on préfère, la population étudiée sans ses éléments extrêmes (de extrémité NDR) considérés comme perturbateurs au comportement aberrant (ce sont eux qui le disent NDR) ! La machine nous donne des résultats, quelle note : s_x ou Sx.



Notre notation : 'écart type de la population $s_n(x)$ traduit s_x, et pour l'échantillon $s_{n-1}(x)$ traduit Sx. Déjà qu'ils calculent avec leurs pieds au lieu d'utiliser le mètre qui est pourtant l'unité légale…
Une correction est offerte sous forme d'une application 'flash' admirer l'écran (pas encore au point (XI 1999), ne veut pas tenir compte des effectifs !).

Utilité de l'écart type : il s'interprète en quelque sorte comme l'écart que l'on peut s'attendre à remarquer en général entre les individus étudiés et la moyenne. Un point remarquable étant qu'au moins 75 % de la population étudiée se trouve dans l'intervalle m-2 s_x et m+2 s_x où m est la valeur moyenne.

4. **Coefficient de variation :**

Def : c'est le rapport $\frac{\text{écart type}}{\text{moyenne aritmétique}}$. (Je ne chercherai pas **l'indice de concentration** !)

5. **L'intervalle interquartile :**

Def : l'intervalle interquartile est la différence entre le troisième et le premier quartile=Q_3-Q_1.

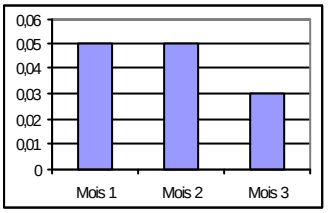
Utilisé dans certaines représentation graphiques qui suivent. Il concerne la moitié de l'effectif total.

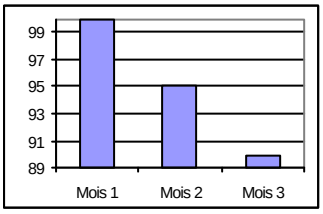
III. Représentations graphiques :

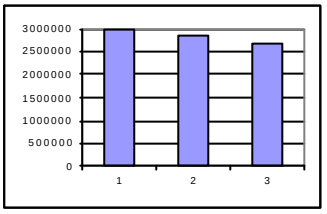
Remarque préliminaire : on assiste à une banalisation de l'erreur de la représentation graphique des séries statistiques. S'il est vrai que parfois l'œil appréhende mieux les tendances des phénomènes étudiés par l'utilisation de certaines représentations plutôt que d'autres, il faut quand même savoir quand l'utilisation de ces graphiques est légitime.

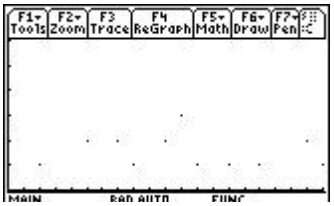
Par exemple : quand les objets étudiés sont des valeurs (phénomène discret, ponctuel), et non des intervalles ou classes (phénomène continu), il ne devrait pas être utilisé de courbe, reliant les points par des segments qui n'existent pas, ou des histogrammes en lieu et place de bâtons ou barres (largeur unitaire).

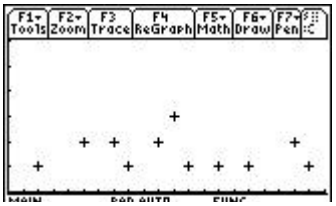
Il est vrai que nous ne sommes pas très bien servis par les logiciels tableur-gapheur (oui l'oubli du r est volontaire, il y en a trop dans cette phrase, et quand il y a trop d'r on s'enrhume) sur ordinateur, dont

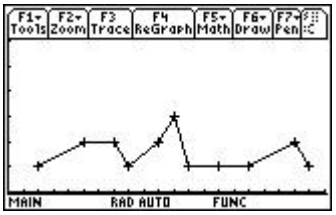


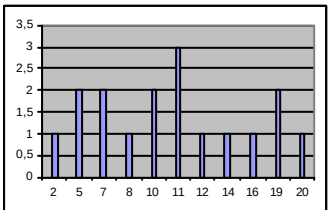


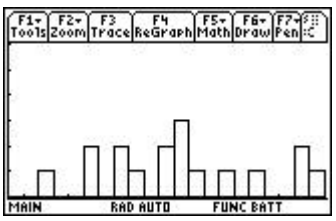


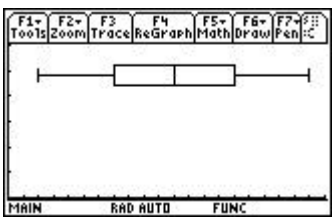


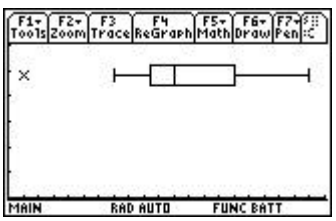












l'objectif est « faire beau», ou par l'utilisation des représentations graphiques prédéfinies sur les calculatrices (alors, ça bouge sur les stats chez les constructeurs ?).

En voir de toutes les formes et toutes les couleurs… Il suffit d'utiliser un tableur-gapheur pour s'en convaincre (liste ci-contre). Ce qui est remarquable c'est de constater l'usage qui est fait des différentes représentations graphiques. Choisir une représentation peu usitée, permet souvent de profiter de l'incompréhension du public pour faire passer quelques erreurs (volontaires ?) d'interprétation du phénomène présenté.

Exhiber un schéma (réducteur) vite passé comme point d'appui d'une information, donne l'impression de **preuve** du discours.

Or, la représentation choisie devrait être simple, claire, représentative du caractère étudié, accompagnée des informations (échelle par exemple) nécessaires à sa bonne interprétation.

Augmentation des prix exprimée en pourcentages.

Ci-contre et ci-après, « **vu à la télé** » deux graphiques (trois pour le même prix !). Le premier montre « la maîtrise de l'inflation » jugulée aux environs de 2,2 % sur un an, dont on ne montre que les pourcentages d'augmentation sur les trois derniers mois.

Baisse du chômage… information ?

Le second est là pour nous démontrer une baisse du chômage. Baisse sensible évidente du %.

Le troisième, non « vu à la télé », représente le nombre de chômeurs… baisse sensible, mais pas dans les mêmes proportions !

1. <u>Graphiques à représentation cartésienne</u> : par points, bâtons, barres, bandes (barres horizontales !), histogrammes, etc.

J'utilise les données de la série (ponctuelle) S_{10} (page 3) :

Des petits points, qui peuvent être grossis.

Courbe, ce qui ne devrait pas se faire pour des données ponctuelles.

Diagramme **bâtons** ou barres, plus joli sur le gapheur, MAIS il ne sait pas lui ce qu'est un intervalle, et donc respecter les dimensions (2 à 5 comme de 7 à 8).

Remarque : ils confondent histogrammes et bâtons ou barres, ainsi que barres et bandes…

Boites statistiques, ou boites à moustaches : Elles indiquent les valeurs minimales et maximales (trait vertical), la médiane à l'intérieur de l'intervalle interquartile.

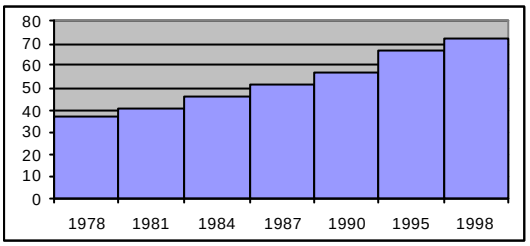
La deuxième boite à moustaches, exclu les valeurs extrêmes (petite croix) qui ne seront pas comptabilisées lors des calculs statistiques. A utiliser à bon escient.

Parfois, une distinction (normalisée ?) est faite entre les boites statistiques conservant toutes les valeurs, et les boites à moustaches « lissées », aux valeurs dites aberrantes (c'est parfois vrai !) supprimées, par le tracé aux contours rectangulaires pour les premières, et un contour arrondi, en parenthèses pour les secondes.

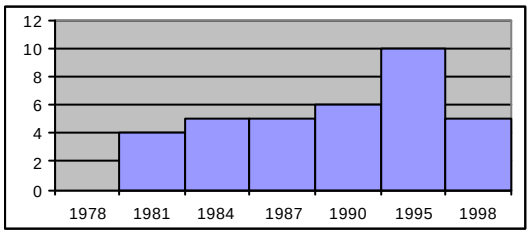
Histogrammes… il y a beaucoup à dire sur ceux-ci. Pas qu'ils soient importants, mais parce qu'ils représentent une erreur flagrante de traitement des gapheurs, et ne sont pas un point fort des calculettes. L'histogramme représente chaque valeur par une aire (surface) proportionnelle au couple intervalle-effectif.

Période :	0	1	2	3	4	5	6
Année :	1978	1981	1984	1987	1990	1995	1998
Intervalle (années) :	0	3	3	3	3	5	3
Taux : (%)	37	41	46	51	57	67	72
Augmentation du taux :	0	4	5	5	6	10	5

S_{11} est la série constituée par l'étude du taux de pénétration d'un appareil ménager dans les foyers français.
Année de base 1978. Une enquête permet de calculer ce taux tous les trois ans. Faute de crédits (!) il fut envisagé la suppression de cette enquête, entre 1990 et 1993. Ceci explique l'intervalle de cinq ans pour la 5ième période.



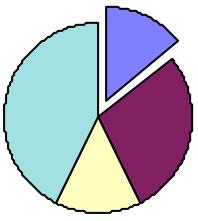
Le bel histogramme ci-contre montre le taux de pénétration pour chaque période. Rien d'anormal? (rép : pas de différence de largeur pour un intervalle 3 ans ou 5 ans !)



Celui-ci, représente la seule augmentation du taux pour chaque période. [1990 ; 1995[semble bénéficier d'une forte augmentation du taux de pénétration, donc des ventes. Les fabricants, les vendeurs ont du faire de bonnes affaires… peut-être que les ouvriers, les employés, leurs syndicats devraient se précipiter sur une demande d'augmentation ? Triste réalité. L'augmentation du taux sur 5 ans ne correspond qu'à la même augmentation sur la période précédente de 3 ans ! (calculez, plus haut j'ai donné les outils). Erreur inadmissible !
En effet, considérons les deux intervalles [1984 ; 1987[et [1987 ; 1990[. Ils ont même taux, 5 %. Si je colle les deux morceaux de trois ans l'un contre l'autre, j'aurai un intervalle de 6 années, **et la même hauteur**.

Une deuxième erreur est d'avoir comme unité sur l'axe vertical les effectifs…, en effet, si une hauteur doit apparaître, il faut d'abord la calculer. Nous avons aire du rectangle = base x hauteur = effectif. La base est donnée, c'est l'intervalle de la classe. La hauteur est le rapport effectif / (amplitude de l'intervalle).
Par exemple, hauteur du rectangle 1990 : (intervalle [1987 ; 1990[augmentation 6 %) h = 6 / 3 = 2. Exactement identique à celle du rectangle 1995 : 10 / 5 = 2 !
2 est d'ailleurs la plus grande valeur de la série (hauteur). Nous devrions avoir comme bornes sur [Oy) 0 et 2. Essais en solitaire autorisés.

2. Graphiques à représentation circulaire :

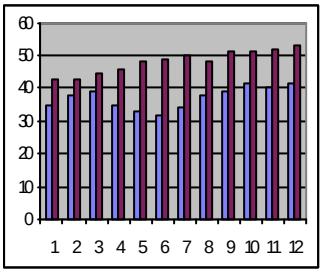


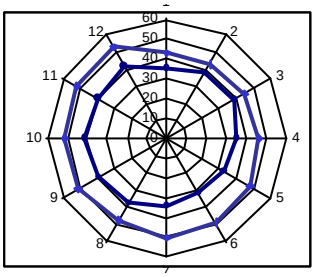
Ci-contre une représentation en **secteurs circulaires**, un « **camembert** » décoratif. Est-ce plus explicite que d'utiliser des barres ou des histogrammes ? C'est en tout cas une possibilité de représentation.

Petit détour par une représentations **polaire**, ou en « **radar** ».

Mois	1	2	3	4	5	6	7	8	9	10	11	12
Année 1	35	38	39	35	33	32	34	38	39	41	40	42
Année 2	43	43	45	46	48	49	50	48	51	51	52	53

S_{12} Tableau du chiffre d'affaire mensuel, en millier de Francs, d'une petite entreprise.
Quelle représentation choisir ?





A gauche un diagramme bâton où chaque série de valeur d'un mois d'une année est collée à celle de l'année suivante lui correspondant. C'est assez clair.
A droite, un diagramme polaire. Sans couleur, et un peu petit… la deuxième année s'enroule en spirale autour de la première.
Dans le cas de plusieurs années, et sous réserve d'avoir une augmentation entre chaque année, la représentation peut être intéressante.

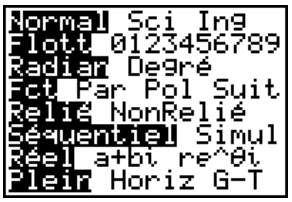
IV. Utilisation d'une calculatrice : (je fais avec ce que j'ai !)

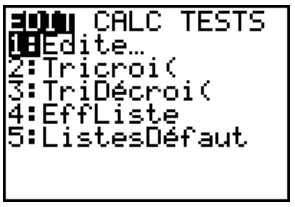
TI 80, 81,82,83 et 83+ sont du même type, avec quelques petites différences… TI 89 et 92 sont semblables.

Données de la série S_{11}. Je n'indique pas qu'il faut appuyer sur **Í** après chaque choix !

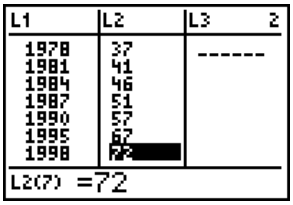
TI 83

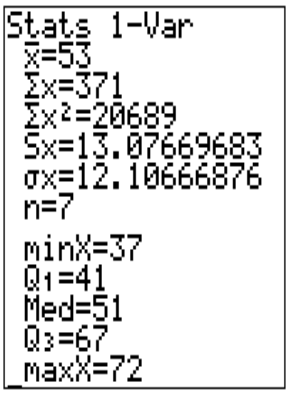
Vérifier d'être en mode Fonction, par **MODE**.
Passer en mode statistiques par la touche **STAT**.



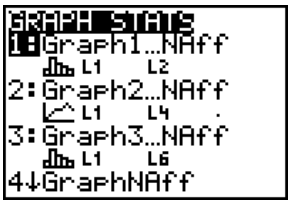


Dans l'éditeur de données statistiques, remplir.
Les résultats par **STAT** **CALC** L_2.

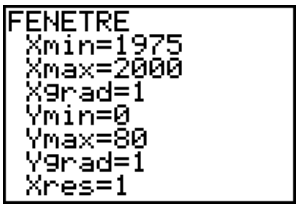


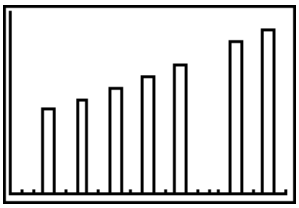


Passer dans l'éditeur graphique par **2nd** **STATPLOT**.
Choisir le premier graphe, remplir.



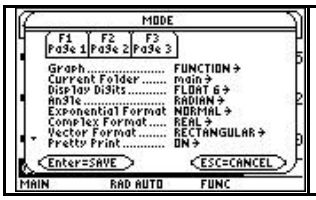
Fenêtre d'affichage **WINDOW**, remplir. Résultat.



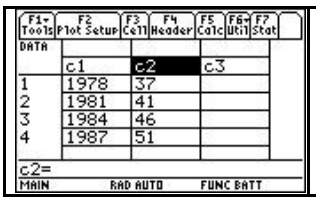


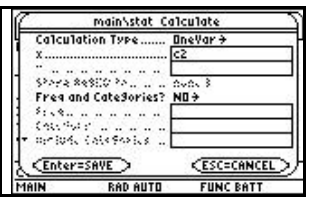
TI 89

Idem (⇐) !
Touche **APPS**, puis 6 et **NEW**. Donner un nom.

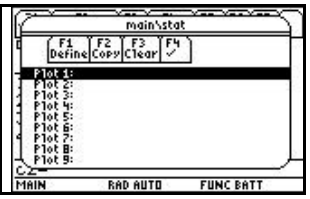


Idem (⇐) !
Pour les résultats, **F5** et compléter.

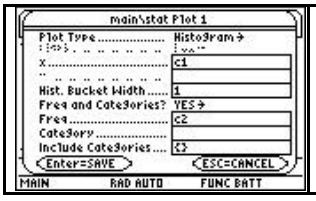


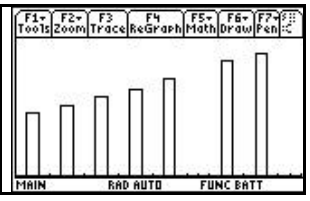


Les calculs. **F2** éditeur graphique, **F1** définir.



Remplir et résultat.





Il est ici possible de déterminer la largeur de l'intervalle.

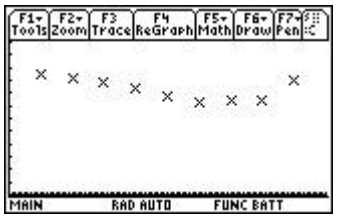
REM : sur TI 89 s_x s'obtient par ¥ c j S, ou d'utiliser l'application flash.

V. **Statistiques à deux variables (** statistiques doubles) **:**

Où l'on croise les phénomènes…
La question préoccupante de cette partie des statistiques est : peut-il y avoir un rapport, une **corrélation** entre les deux variables étudiées ?

Par exemple : « plus je roule vite et plus je consomme ». Série S_{13}.

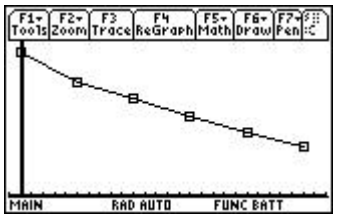
Je trace le « **nuage de points** » obtenu en plaçant, **dans un bon repère** , les points de coordonnées $(x_i ; y_i)$.



Vitesse en Km.h^{-1} (x_i)	10	20	30	40	50	60	70	80	90
Consommation en L/100 km (y_i)	9,4	9,2	8,9	8,3	7,8	7,2	7,4	7,5	9,0

Conclusion ? ça descend, ça remonte, mais ce n'est pas de type linéaire qui monte. Ce véhicule ne respecte pas la loi proposée.

« plus les années passent et plus les surfaces plantées en avoine diminuent en France ». Série S_{14}.



Année :	1970	1975	1980	1985	1990	1995
Rang : (xi)	0	5	10	15	20	25
Superficie (milliers hectares) (yi)	789	629	534	431	343	274

Une jolie droite (petit écart en première valeur, sans remise en cause de l'observation).

Comment vérifier qu'une relation existe ? la forme du nuage de points est déjà une indication, s'il **ressemble** à une droite, une parabole, une courbe logarithmique, exponentielle, etc., il faut rechercher une validation de l'impression oculaire (pour l'approximation de valeurs non indiquées).

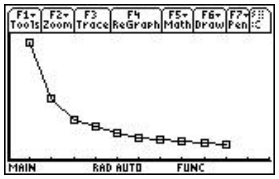
Nous connaissons la variance pour une série statistique simple.

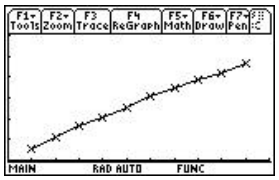
Donnons vie à la **covariance** : $s_{xy} = \frac{1}{N}\sum_{i=1}^{p} x_i y_i - \bar{x}\,\bar{y}$

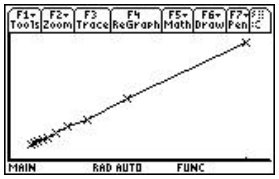
ainsi qu'au **coefficient de corrélation linéaire** **:** $r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$. Si $0{,}75 < r^2$ alors les deux séries présentent une bonne corrélation linéaire, et plus r^2 est proche de 1, plus il est légitime de rechercher cette corrélation linéaire entre les deux séries.

J'ai donc un bon outil dans la recherche d'une relation de proportionnalité (linéaire).
Et si ce n'est pas linéaire ? on se débrouille pour y revenir ! (enfin… bof, maintenant les outils de calcul sont performants)

Prix (en dollars) d'un microprocesseur d'ordinateur par périodes de trois mois :







Période	1	2	3	4	5	6	7	8	9	10
Prix	1720	902	595	489	399	325	288	256	238	214
1000x(1/Prix)	0,58	1,11	1,68	2,05	2,51	3,08	3,47	3,91	4,20	4,67
1/période	1,00	0,50	0,33	0,25	0,20	0,17	0,14	0,13	0,11	0,10

Je remarque : la première courbe présente une allure en 1/x. Ce qui se dirait : le prix du microprocesseur semble inversement proportionnel au temps qui passe (période).

D'où l'idée de prendre en ordonnée les inverses de la période (troisième ligne) (*), ou celle de prendre en abscisse les inverses de la période (quatrième ligne) (*).

La deuxième (ou troisième) courbe obtenue est suffisamment rectiligne pour conforter le statisticien (heureux ?) dans ses hypothèses.

(*) le choix de l'échelle du repère est primordial. [0 ; 11] et [0 ; 1850] pour la première courbe, [0 ; 11] et [0 ; 6] pour la seconde, [0 ; 1,1] et [0 ; 1850] en dernier.

VI. Utilisation des statistiques, l'estimation, l'ajustement :

Un problème important est de résumer une série statistique (ayant beaucoup de valeurs !)b et de pouvoir 'anticiper' une action, donc d'essayer de prévoir l'évolution du phénomène. La traduction des valeurs en points sur un graphique permet à l'œil d'interpréter une tendance, parfois une courbe (mathématique) connue (on dira dans ce cas que l'on procède par méthode analytique), permettant la prévision.

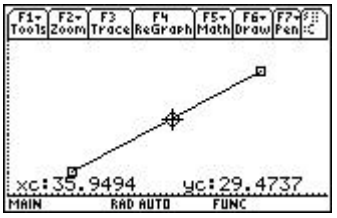
1. Séries simples : interpolation linéaire ou affine :

REMARQUE : il est d'usage d'indiquer une interpolation ou un ajustement '**linéaire**', alors qu'il s'agit d'un ajustement '**affine**'. Le vocabulaire mathématique est précis. L'usage interprète 'droite' avec 'linéaire'. En stat on dira 'ajustement linéaire', et en math on remplacera cet abus de langage usuel par 'ajustement affine'.

Lors des spéciales d'un rallye, les voitures ne partent pas avec le plein pour les 30 ou 40 km quelles doivent parcourir. L'essence pèse, et quelques litres en plus font des secondes bien lourdes !

Distance en km	20	50
Capacité nécessaire en L	16	42

Série S_{15}. Pour des courses similaires, l'équipe des mécaniciens a mesuré la consommation exacte de la voiture. Sachant que l'épreuve du jour fera 36 km, de quelle quantité d'essence faut-il remplir le réservoir ?



En considérant les consommations proportionnelles aux distances parcourues, il suffit de tracer la droite passant par les deux points de coordonnées (20 ; 16) et (50 ; 42), puis graphiquement estimer la valeur de l'ordonnée de la droite correspondant au point d'abscisse 36. Je lis (curseur) pour 36 km il faut environ 29,5 L.

Il est bien sur possible de vérifier ce résultat par le calcul. La pente de la droite est : $\frac{\Delta_x}{\Delta_y} = \frac{50-20}{42-16} = \frac{50-36}{42-y}$ d'où y = 448/15=29,87 à 10^{-2} près, donc 30 litres environ.

Remarque : une année, un certain ANDRUET, dans la dernière spéciale du rallye de Monte Carlo, alors qu'il avait course gagnée, s'est immobilisé à moins de 300 mètres de la ligne d'arrivée…

Inutile de préciser que l'estimation peut évidemment concerner un point extérieur au segment tracé !

2. Séries à deux variables :

Souvent les séries sont obtenues sur des périodes (régulières) de temps, des durées. On les nommera '**séries chronologiques**'. On y remarque quatre mouvements (ce n'est pas obligatoire, mais c'est ce que l'on peut constater) : un mouvement de longue durée (appelé 'trend') qui traduit l'allure général du phénomène, une composante cyclique (non obligatoire), des variations saisonnières (non obligatoire), des variations accidentelles (non obligatoire).

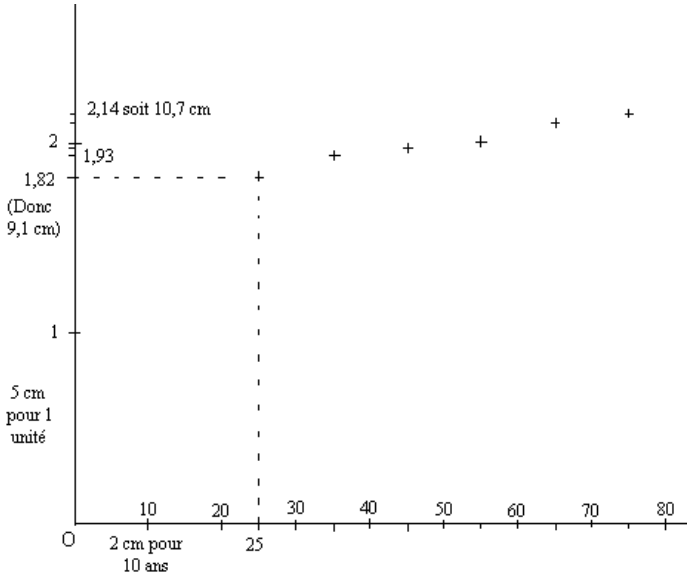
a) Ajustement linéaire par la méthode de Meyer :

L'idée : pour tracer une droite, deux points suffisent. Le point G_1 de la sous-série formée par la première moitié de la série et le point moyen G_2 du reste des points, permettent de tracer une droite d'estimation.
Une équation de la forme y=ax+b de cette droite d'estimation est rapidement déterminée, puisque celle-ci passe par G_1 donc $y_1=ax_1+b$, elle passe par G_2 donc $y_2=ax_2+b$. Nous obtenons un système (fastoche) de deux équations à deux inconnues.

Age (années)	25	35	45	55	65	75
Taux moyen	1,82	1,93	1,98	2,01	2,09	2,14

La série S_{16} donne le taux moyen de cholestérol dans le sang en fonction de l'age (à partir d'un pb bac SMS).

Commencer par en tracer le nuage de point, ce qui donne une idée du type (éventuel !) d'évolution du phénomène



Ci-contre le nuage de points obtenu dans le repère orthogonal où 2 cm représentent 10 ans, et pour les taux, l'unité graphique est de 5 cm. Par exemple, le nombre 1,82 sera à 5x1,82 = 9,1 cm.
Il est conseillé de travailler sur papier millimétré.

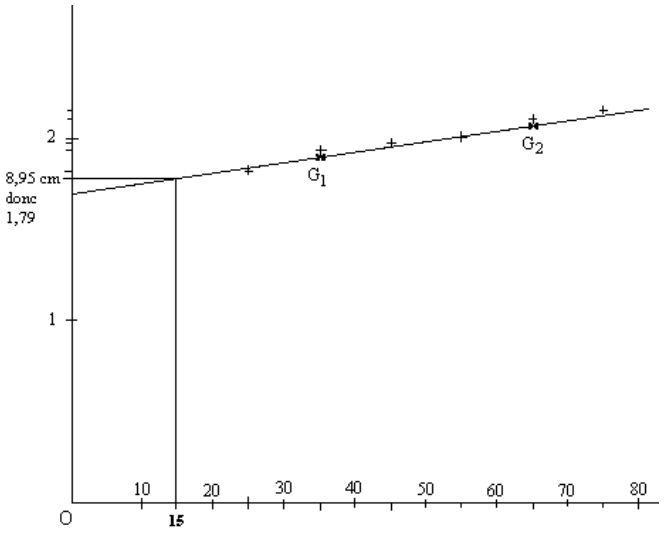
Soit G_1 le point moyen des trois premiers points, et G_2 le point moyen des trois derniers. Calculer les coordonnées de G_1 et G_2, puis tracer la droite (G_1G_2) sur le graphique.

Nous avons :

$$\bar{x}_{G_1} = \frac{25+35+45}{3} = 35 \text{ et } \bar{y}_{G_1} = \frac{1,82+1,93+1,98}{3} = 1,91$$

$$\bar{x}_{G_2} = \frac{55+65+75}{3} = 65 \text{ et } \bar{y}_{G_2} = \frac{2,01+2,09+2,14}{3} = 2,08$$

Ce qui donne : les coordonnées de $G_1\begin{vmatrix}35\\1,91\end{vmatrix}$ $G_2\begin{vmatrix}65\\2,08\end{vmatrix}$ les deux points à placer sur le graphique. Tracer la droite (voir ci-après).
La droite (G_1G_2) est tracée. Le travail d'estimation peut commencer.



Déterminer graphiquement le taux d'un individu de 15 ans.

'Monter' la verticale de 15. Du point d'intersection avec (G_1G_2) tracer l'horizontale. Mesurer l'ordonnée obtenue (je trouve 8,95 cm donc un taux de 8,95/5=1,79).

Donner une équation de (G_1G_2) sous la forme *y=ax+b.*
(à près 10^{-3} pour a et 10^{-2} pour b)

La droite passe par G_1 et G_2, donc a.35+b=1,91 et a.65+b=2,08. Je trouve (approximativement) a=0,0057 et b=1,712, donc je prends 0,006 pour a et 1,71 pour b. L'équation de la droite est donc : y=0,006x+1,71.

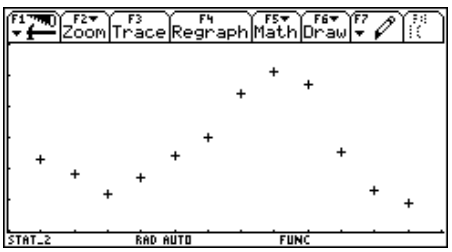
Retrouver par le calcul le taux d'une personne de 15 ans : je remplace : 0,006x15+1,71=1,8 (ce qui est proche de ce qui a été trouvé précédemment).

b) <u>Ajustement</u> (pas forcément linéaire) <u>par la méthode des moyennes périodiques</u> :

Le calcul des moyennes périodiques ou échelonnées est un simple résumé par intervalles de temps. Par exemple on remplace une vision mensuelle par les moyennes trimestrielles. Exemple : Pour mettre en place la tournée d'approvisionnement de distributeurs automatiques de jus de fruits (trop) sucrés, on recueille la fourniture mensuelle d'un groupe d'appareils sur 12 mois.

Mois :	1	2	3	4	5	6	7	8	9	10	11	12
Cannettes :	113	108	102	107	114	120	134	141	137	115	103	99

La série S_{17} donne le nombre de bouteilles fournies mensuellement.



Et le nuage de points qui en découle.

```
Prgm
dim(echel[1])/3»trim
newList(trim)»cc3
```

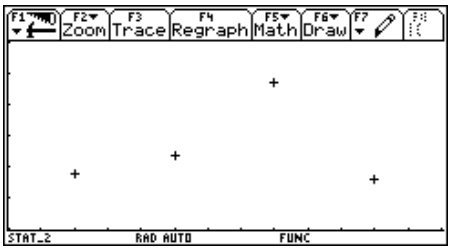
Le petit programme ci-contre se charge de me calculer les moyennes trimestrielles.

```
newList(trim)»cc4
gettime()»t1
seq(n,n,2,12,3)»cc3
For i,1,trim
   ((echel[2])[j],j,1+(i-1)*3,1+2+(i-1)*3)/(3.)»cc4[i]
EndFor
gettime()»t2
Disp t2-t1
EndPrgm
```

Le tableau des valeurs de la calculatrice se nomme ‘echel’, je donne le nom de trim au nombre de trimestres du tableau.

Remarque : la fonction ‘gettime()’ est spécifique de la V200, elle n’est pas utilisée sur la TI 89 ou 92.

Trimestre :	1	2	3	4
Affecté au mois	2	5	8	11
Cannettes :	107,67	113,67	137,33	105,67



On obtient le résumé (très résumé dis donc !) de la série précédente.

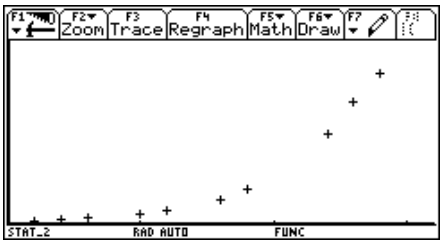
On peut remarquer une forte perte de données. Ce procédé est destiné à gommer les fluctuations, ce qu’il fait même un peu trop bien ici (c’est, en plus de sa simplicité, son avantage et inconvénient).

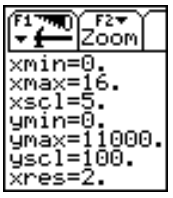
c) <u>Ajustement</u> (pas forcément linéaire) **<u>par la méthode des moyennes mobiles</u> :**

C’est une façon très intéressante et utile pour le ‘lissage’ des données ou de la courbe (nuage de points), les valeurs un peu extrêmes sont ramenées vers une position plus centrale.
C’est plus difficile à décrire qu’à faire… ‘un petit dessin vaut mieux qu’un long discours’ dirait mon inspecteur !

Série S_{18} donnant l’éloignement maximal en mètres d’un animal transporté en estive (pâturage d’été), autour de la bergerie.

Jour	1	2	3	5	6	8	9	12	13	14
Distance	140	195	270	520	750	1420	2015	5410	7300	9005

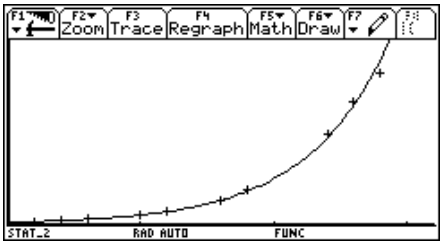


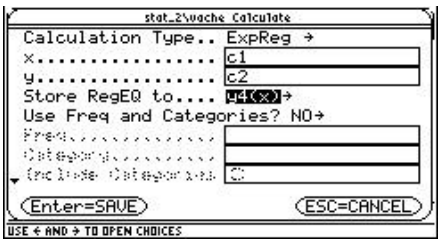


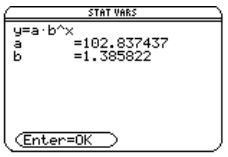
Le nuage laisse imaginer une exponentielle…

Que dit la calculatrice (elle fait ça si bien…!).

La fenêtre d’affichage.







Il suffit de demander…
(et de dire que le résultat de la fonction d’ajustement par le modèle cherché soit stocké dans l’une des variables graphiques).

remarque : le modèle ‘exponentiel’ est à retenir pour ce qui est mesuré ici. Par la suite, la distance maximale a tendance à se stabiliser, les animaux savent où trouver les compléments nutritifs (par exemple). Il est probable qu’une étude plus approfondie laisserait entrevoir une loi ‘logistique’.

La théorie : pour une série de n points (n=10 ici dans l’exemple), choisir un nombre p tel que 1<p<n, dans la pratique c’est fonction d’une grande habitude, ici je choisirais p=3 ou 4 ou 5 (4 retenu).

Remarque : prendre un nombre impair est préférable, car nous avons la valeur, puis autant de valeurs en dessous qu’au dessus. Dans le cas du calcul sur un nombre pair de valeurs (en bourse ils aiment), la moyenne serait à attribuer entre deux valeurs (souvent des dates journalières). Par exemple, la moyenne prise sur les mois 1, 2, 3 et 4 doit-elle être affectée au mois 2 ou 3 ? parfois on considérera l’affectation à la date intermédiaire, sachant que l’habitude de certains est de calculer la moyenne M_1 pour les mois 1, 2, 3, 4 puis M_2 pour 2, 3, 4, 5 et d’attribuer $(M_1+M_2)/2$ au mois 3.

Autrement dit… en pratique, si p est pair (p=2p'), on prend p+1 valeurs (la valeur, p'=p/2 valeurs avant et p' valeurs après) pondérées par les coefficients : ½ pour la première et dernière valeur, 1 pour les autres (la somme des coeffs est bien 2p !). C'est ce qui sera fait dans le deuxième exemple.

						Point moyen
Sous série 1	**1**	**2**	**3**	**5**		**2,75**
	140	**195**	**270**	**520**		**281,25**

						Point moyen
Sous série 2	**2**	**3**	**5**	**6**		**4**
	195	**270**	**520**	**750**		**433,75**

Calculer le point moyen des p premiers points, placer ce point sur le graphique. Refaire le calcul du point moyen en supprimant le premier point de cette sous série, tout en ajoutant le $(p+1)^{ième}$ point.
Dans mon exemple je passe de la sous série M_1, M_2, M_3, M_4, à la sous série M_2, M_3, M_4, M_5. Tableaux ci-joints pour les deux premières sous séries.

Recommencer jusqu'à ce que la sous série contienne le $n^{ième}$ et dernier point (oui c'est un peu long !).

Jour	**2,75**	**4**	**5,5**	**7**	**8,75**	**10,5**	**12**
Distance	**281,25**	**433,75**	**740**	**1176,25**	**2398,75**	**4036,25**	**5932,5**

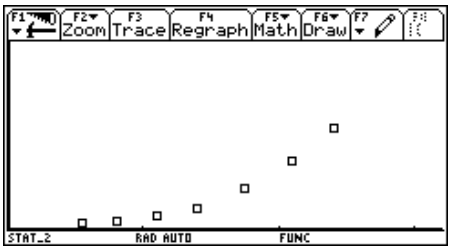
Résultats rassemblés dans le tableau ci-contre.

```
Prgm
dim(vache[1])-3»veau
newList(veau)»cc3
newList(veau)»cc4
gettime()»t1
For  i,1,veau
 ((vache[1])[j],j,i,i+3)/(4.)»cc3[i]
EndFor
For  i,1,veau
 ((vache[2])[j],j,i,i+3)/(4.)»cc4[i]
EndFor
gettime()»t2
Disp  t2-t1
EndPrgm
```

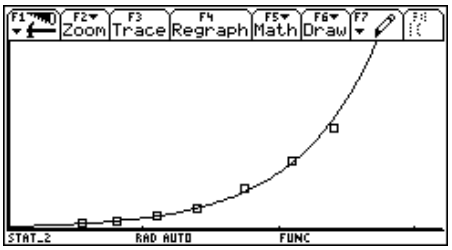
Evidement, je n'ai pas calculé toutes ces moyennes…

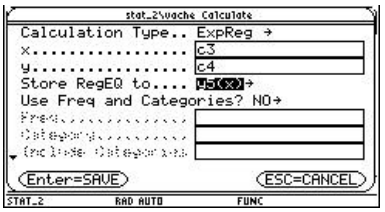
Le petit programme ci-contre l'a fait pour moi.

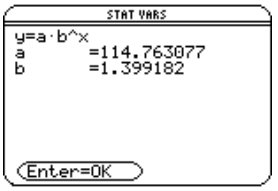
J'ai nommé 'vache' le tableau des données (se comporte comme une matrice), et 'veau' le nombre de moyennes à calculer en fonction des moyennes mobiles choisies.



Le nuage obtenu est alors remplacé, si possible, par une droite (méthode graphique le plus souvent), ou une autre courbe si cela s'impose. Ici le lissage en exponentielle est évident (à tracer à la main) de par ce qui a été vu précédemment (images ci-après).





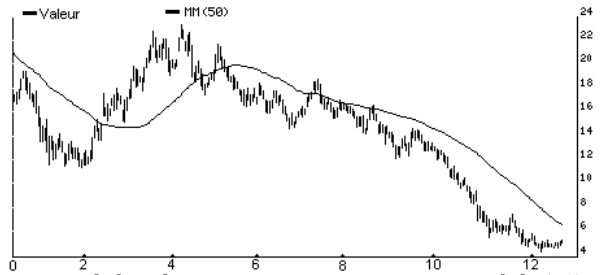


Remarque : cette méthode produit une 'petite' perte d'information. D'une part sur les premières et dernières valeurs, d'autre part en égalisant les valeurs (lissage). C'est un peu long à mettre en œuvre, mais très souvent utilisé (théories importantes en économie, bourse).

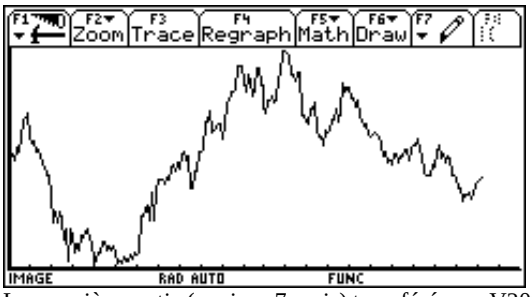
L'utilisation des moyennes mobiles permet généralement un bon lissage de la courbe, qui élimine les données 'extrêmes'. C'est un peu comme repasser l'allure de la courbe avec un gros feutre. Comme l'exemple suivant va le faire remarquer, c'est une méthode utile pour résumer une série statistique aux nombreuses données.

Un 'gros' exemple : le cours d'une valeur boursière sur presque une année (239 jours). J'utilise la TI V200 (ou la 92 même écran) car sur la 89 ou la 83 l'écran est plus petit. (Je fais en sorte qu'un pixel écran puisse contenir une information).

Je récupère une valeur boursière sur internet. Je n'ai pas les cours au jour le jour sur une grande période.

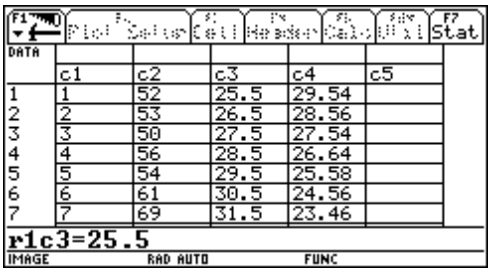


Le cours de la valeur sur internet et sa Moyenne Mobile (50).



La première partie (environ 7 mois) transférée sur V200.

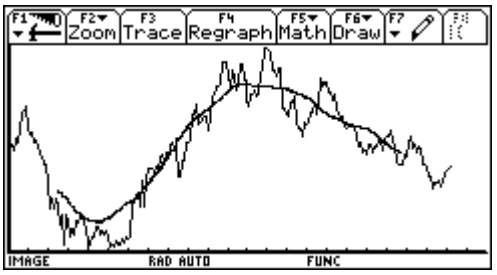
A partir de l'image, transférée sur la calculatrice (gros petit programme), je peux obtenir (toujours par programme maison et près de 10 minutes) les coordonnées des points de la courbe.



	c1	c2	c3	c4	c5
1	1	52	25.5	29.54	
2	2	53	26.5	28.56	
3	3	50	27.5	27.54	
4	4	56	28.5	26.64	
5	5	54	29.5	25.58	
6	6	61	30.5	24.56	
7	7	69	31.5	23.46	

Ce que je reporte dans l'éditeur de données, en colonnes C1 et C2.

Encore un programme… pour calculer les moyennes mobiles (sur 50 valeurs) en un peu plus de 14 minutes…

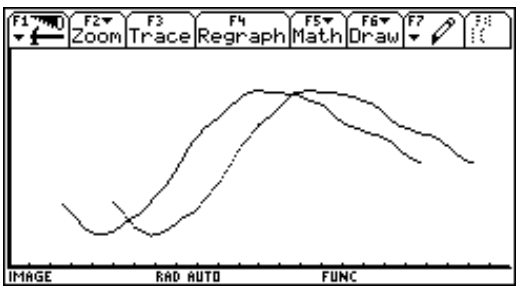


Et voici le résultat.

On remarque : sur l'image d'origine (internet) la courbe des moyennes mobiles commence dès l'origine (ils ont les cours précédents !). Mais aussi, elle va jusqu'à la dernière valeur (le cours du jour). Je m'interroge…
Et je me répond !

En bourse, ils calculent une moyenne mobile selon LEUR définition, pas celle mathématique. Une bonne raison est que la moyenne utilisée doit permettre des prises de décisions sur LE COUR DU JOUR. Que dit-elle cette définition ? Moyenne des 50 (ou autre !) derniers cours, y compris celui du jour. Je refais un programme (hé oui faut encore attendre environ ¼ d'heure).

Pour le début de la courbe, je n'avais pas les 50 cours précédents la première valeur, mais pour le reste, ça ressemble assez bien à ce que l'on peut voir sur la courbe internet.



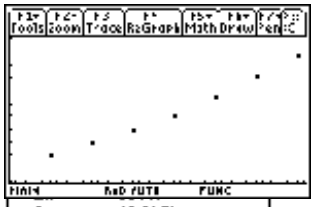
Dernier regard… les deux moyennes mobiles calculées se ressemblent fort, au décalage (normal !) près.

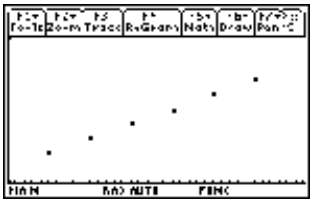
La moyenne mobile permet souvent d'obtenir une courbe de tendances 'rabotée', lissée. De là à en déduire le temps demain…

c) <u>Ajustement linéaire analytique, méthode des moindres carrés</u> :

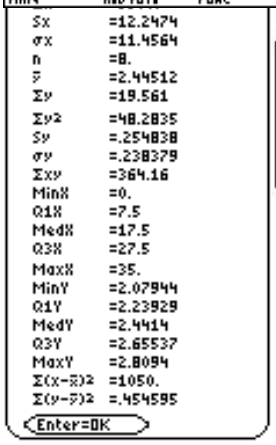
Jours (x_i)	0	5	10	15	20	25	30	35
Population en milliers (y_i)	8,0	8,9	9,9	11,0	12,0	13,5	15,0	16,6
Ln(y_i)	2,08	2,19	2,29	2,40	2,48	2,60	2,71	2,81

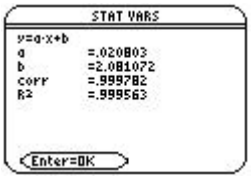
S_{18} est une série donnant la population d'une souche de culture dans une boite de Petri relevée tous les 5 jours.





Le premier écran montre le nuage de points, qui **attentivement observé** présente une courbure (logarithmique) plutôt qu'un bel aspect de points alignés. Raison de la troisième ligne du tableau et du deuxième écran. Nous travaillerons sur cette dernière série.





Connaissant les résultats des calculs statistiques de chacune des deux séries, moyenne, écart-type, etc. (beaux écrans de la calculatrice), nous calculons (revoir plus haut dans le texte) la covariance ainsi que le coefficient de corrélation. Il est très proche de 1, donc, à partir de la droite trouvée (bientôt), l'estimation qui sera faite dans un instant devrait être bonne.

Remarque : **la droite d'ajustement passe par le point moyen de la série.** D'où, équation de la droite y=ax+b d'estimation (de y en x) :

$$y-\bar{y}=a(x-\bar{x}) \text{ avec} a=\frac{\sigma_{xy}}{\sigma_x^{\,2}} \text{ et } b=\bar{y}-a\bar{x}$$

La calculatrice me dit a=0,02 b=2,08 et très bon r (respirez bien). A vérifier par calcul pour s'habituer.

Estimer alors la population p du 42ième jour. Il suffit de remplacer dans l'équation de la droite. y=0,02.42+2,08=2,92 qui est le logarithme de la population. D'où $p=e^{2,92}=18,54$ (milliers).
De même estimer quel jour la population dépasse 19 000. Comme ln(19)=2,94 nous cherchons 2,94=0,02.x+2,08. Je trouve x=43.

Remarque : les équations de droites trouvées par ces diverses méthodes, Meyer, ajustement graphique d'après les moyennes mobiles, droite des moindre carrés, ne sont pas forcément les mêmes…

Pourquoi des formules si compliquées (écart-type, méthode des moindres carrés) ? comme remarqué dans le calcul des écarts (moyen, type), il faut parfois compliquer pour simplifier…

Reprenons les écarts :
S'il semble intéressant de construire un outil mesurant la dispersion, l'écart que l'on peut s'attendre à trouver entre la moyenne et la valeur de l'individu choisi. Comment procéder ?

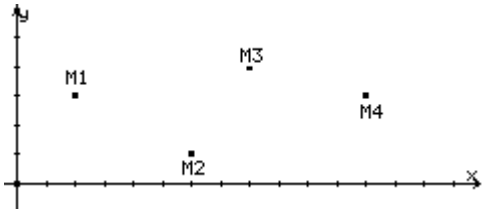
Première idée : la moyenne des écarts à la moyenne. Sombrons immédiatement : S_{19} série simple aux trois valeurs 0 ; 10 et 20. Moyenne 10. Ecarts à la moyenne –10 ; 0 ; 10. Moyenne de ces écarts : 0. Aux antipodes de nos espérances. Le résultat devrait se rapprocher davantage de 10. Nous avons utilisé une mesure algébrique de la différence à la moyenne, les positifs et les négatifs se neutralisent.

Deuxième idée : cherchons la moyenne des distances des valeurs à la moyenne, où seuls des réels positifs sont utilisés. Problème : comment s'écrit la distance entre la valeur x_i (individu i) et la moyenne ? $m-x_i$ ou x_i-m ? dans notre exemple : 0-m ou m-0, 10-m ou m-10, 20-m ou m-20 ? le souligné donne la bonne réponse.
La formule varie selon les cas, la valeur étant plus petite ou plus grande que la moyenne. Il faudrait passer par des valeurs absolues (berk paraît-il). La formulation de la distance moyenne serait alors $d_m=\frac{1}{N}\sum_{i=1}^{n}\left|x_i-m\right|$ (ici $\frac{1}{3}(10+0+10)=6,667$). Il parait que les valeurs absolues ne plaisent pas à tout le monde, elles fâchent certains. D'où la recherche d'une autre formulation, la bonne bien sûr, l'écart-type.

Après avoir remarqué qu'un carré est toujours positif (ou nul NdR), l'idée est de passer par la moyenne des carrés des distances à la moyenne. C'est la variance, dont la racine nous donne l'écart-type $s_x = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_i - m)^2}$ (ici 8,165).

Et y en a qui préfèrent calculer des tas de carrés, puis extraire une racine toute aussi carrée, plutôt que de jouer de la valeur absolue ? Faut croire. On remarquera que pour la série ci-dessus, l'écart serait deux fois de 10, que l'écart type nous offre 8,165 et la distance moyenne 6,667, ce qui n'est peut-être pas plus mal en comptant qu'un écart est nul.

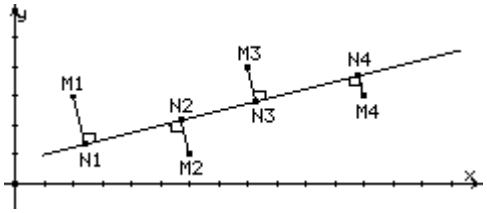
Le principe de « compliquer certaines parties pour en simplifier d'autres » est utilisé aussi dans le cas du choix de la méthode de la recherche d'une droite d'approximation d'un nuage de points à peu près linéaire.



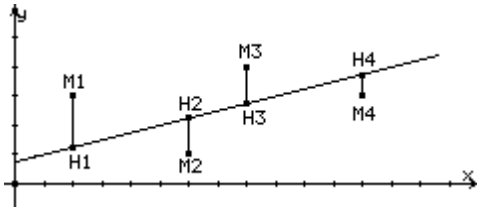
J'ai des points. Je cherche l'équation de la forme y=ax+b de <u>la meilleure</u> droite d'approximation de ce nuage.

Deux remarques (trois pour le même prix) :

- ✓ <u>la</u> droite : les différentes méthodes d'approximation, Meyer, moyennes mobiles, etc. trouvent en général des droites différentes,
- ✓ que signifie '<u>meilleure</u>' ? disons que le but recherché est de minimiser l'écart,
- ✓ même si un tracé de droite permet de raisonner, il faut se persuader qu'il s'agit d'une image virtuelle, fictive, puisque nous cherchons la droite, elle ne devrait être tracée qu'après en avoir déterminé une équation…



Première idée : le plus court entre un point et une droite, c'est la distance entre ce point et la droite. De chacun des points, je mène donc la perpendiculaire à la droite. Je me félicite de tracer virtuellement le segment que vous voyez joindre le point et son projeté orthogonal sur la droite qui n'existe pas encore.
Les géomètres (heureusement qu'ils sont là) proposent immédiatement une formulation du calcul de cette distance entre M_i $(x_i ; y_i)$ et H_i projeté orthogonal de M_i sur la droite d'équation y=ax+b : $d_i = \frac{|y_i - (ax_i + b_i)|}{\sqrt{a^2+1}}$. Remercions vivement le géomètre pour une si belle formule, et prenons notre courage à deux pieds pour fuir rapidement l'horrible vision de la moyenne de ces distances.
Dommage, le résultat eut été intéressant nonobstant la difficulté des calculs. Trop compliqué, passons.



Idée suivante… au lieu de la distance absolue (projeté orthogonal), prenons la distance verticale. La même abscisse caractérise des point de même verticalité. Grosse simplification de la formule. Petit appel au géomètre, qui répond que la distance entre M_i et N_i est $dv_i = y_i - y_n = y_i - (ax_i + b)$ <u>environ</u>. Environ, comment ça environ ? un compas dans l'œil, le géomètre arpente les distances verticales. 'Pour M_1 qui se trouve au dessus de la droite c'est bien $dv_1 = y_1 - (ax_1 + b)$, mais pour M_2, qui est plus bas, ce serait plutôt $dv_2 = (ax_2 + b) - y_2$'. Tournant son compas vers moi il ajouta : 'ne sachant où passe réellement la droite, je ne sais quel calcul choisir…'. 'De quel bois est-il fait ?' (le compas NdR), 'En hêtre je crois, hêtre ou pas hêtre telle est la question…'. Laissons le géomètre à ses réflexions métaphysiques.

Nous remarquons désormais une grande ressemblance avec ce qui précède, dans le calcul de l'écart. Pour éviter le problème d'un calcul de distances négatif, un petit coup de valeur absolue serait le bienvenu. Cela ne faisant pas l'unanimité, le joyeux calculateur matheux jouera plus volontiers du carré puis d'une petite extraction de racine, toute aussi carrée. C'est la méthode 'des moindres carrés'.
Alors, compliquées les formules, certes, mais pour la bonne cause.

Les images : merci aux tableur gapheur excel pour quelques graphiques, et aux calculatrices V200, TI 92, TI 89 pour tous les écrans qui illustrent ces quelques pages.