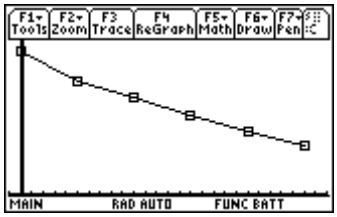
Statistiques à deux variables

I. **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 les années passent et plus les surfaces plantées en avoine diminuent en France ». Série S_{13}.

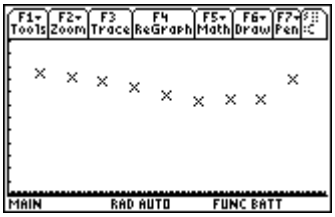
Je trace le « **nuage de points** » obtenu en plaçant, **dans un bon repère**, les points de coordonnées $(x_i ; y_i)$.



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).

« plus je roule vite et plus je consomme ». Série S_{14}.



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 affine, linéaire qui monte. Ce véhicule ne respecte pas la loi proposée.

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).

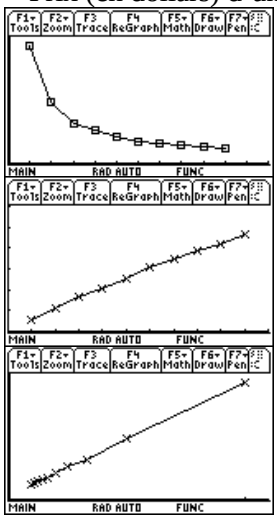
Nous connaissons la variance pour une série statistique simple.

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

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-affine).
Et si ce n'est pas linéaire (affine) ? 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 du prix (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.

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

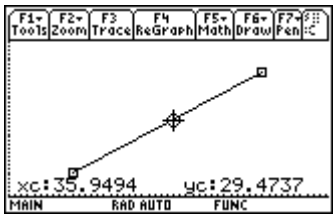
Un problème important est de résumer une série statistique (ayant beaucoup de valeurs !) et de pouvoir 'anticiper' une action, donc d'essayer de prévoir l'évolution du phénomène. La traduction des valeurs en points sur un graphique permet à l'œil d'interpréter une tendance, parfois une courbe (mathématique)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

1. Séries simples : interpolation linéaire : (ici il faut bien écrire **linéaire**)

Exemple : 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 : $\frac{\Delta_y}{\Delta_x} = \frac{42-16}{50-20} = \frac{42-y}{50-36}$ d'où $y = 42 - \frac{182}{15} = 29{,}86666... = 29{,}87$ 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 (manquait pas grand-chose !) à moins de 300 mètres de la ligne d'arrivée… il y en a qui ont dû entendre quelques mots gentils.

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 ou affine par la méthode de Meyer : (plus souvent **affine** !)

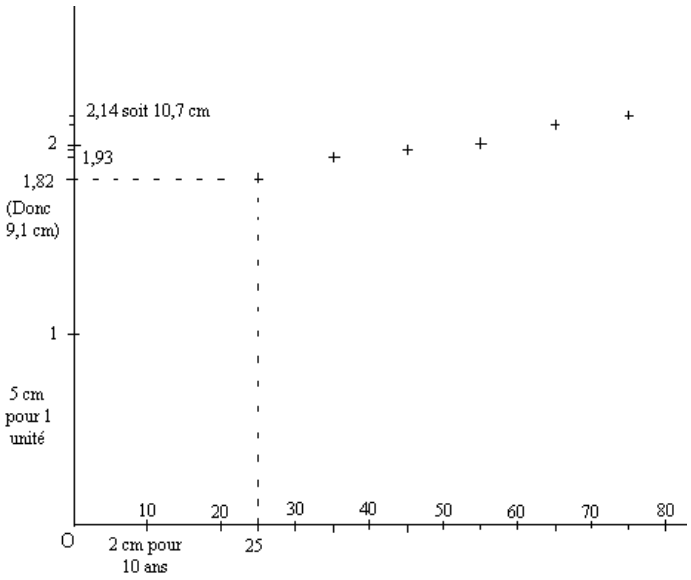
L'idée : pour tracer une droite, deux points suffisent. Le point G_1 de la sous-série formée par la première moitié de la série et le point moyen G_2 du reste des points, permettent de tracer une droite d'estimation.
Une équation de la forme $y=ax+b$ de cette droite d'estimation est rapidement déterminée, 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.

Remarque : en général on préfère résoudre le problème en écrivant $a = \frac{\Delta_y}{\Delta_x}$, puis pour obtenir b on écrit que la droite passe par G_1 donc $y_1=ax_1+b$, connaissant a. D'où $b= y_1-ax_1$.

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'âge (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 5×1,82 = 9,1 cm.
Il est conseillé de travailler sur papier millimétré.

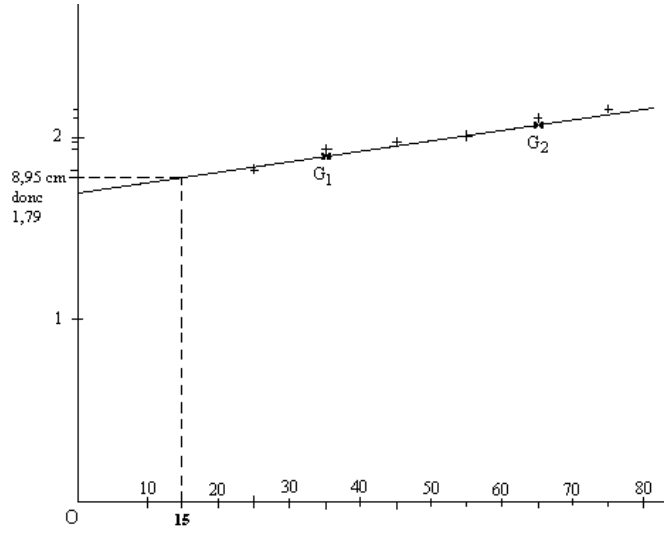
Soit G_1 le point moyen des trois premiers points, et G_2 le point moyen des trois derniers. Calculer les coordonnées de G_1 et G_2, puis tracer la droite (G_1G_2) sur le graphique.

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, puis tracer la droite (voir ci-contre).



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 10^{-3} près 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.

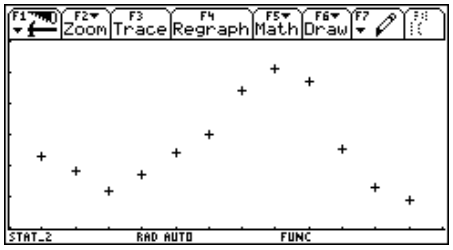
Retrouver par le calcul le taux d'une personne de 15 ans : je remplace : 0,006×15+1,71=1,8 (ce qui est proche de ce qui a été trouvé précédemment).

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
```

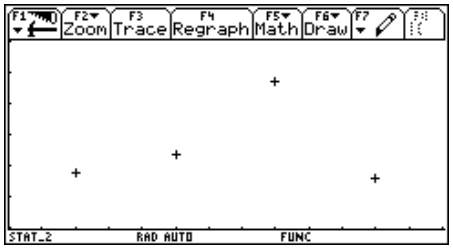
Le petit programme ci-contre se charge de me calculer les moyennes trimestrielles.

```
newList(trim)→cc4
gettime()→t1
seq(n,n,2,12,3)→cc3
For i,1,trim
  Σ((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.

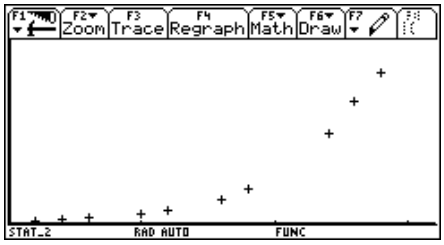
On peut remarquer une forte perte de données. Ce procédé est destiné à gommer les fluctuations, ce qu’il fait même un peu trop bien ici (c’est, en plus de sa simplicité, son avantage et inconvénient).

c) <u>Ajustement</u> <u>(pas forcément linéaire)</u> **<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.

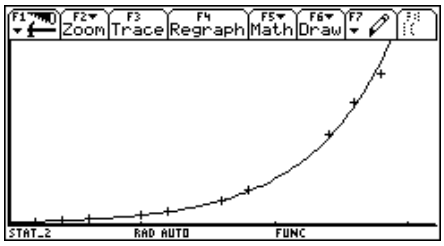
Jour	1	2	3	5	6	8	9	12	13	14
Distance	140	195	270	520	750	1420	2015	5410	7300	9005

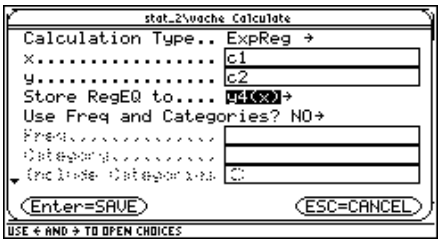


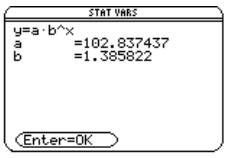
Le nuage laisse imaginer une exponentielle…

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

La fenêtre d’affichage.







Il suffit de demander… (et de vouloir 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**

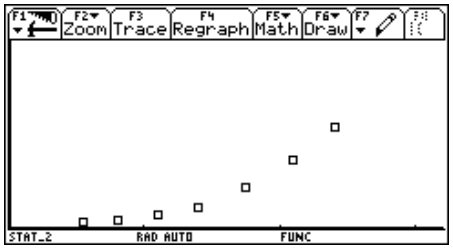
Résultats rassemblés dans le tableau ci-contre.

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

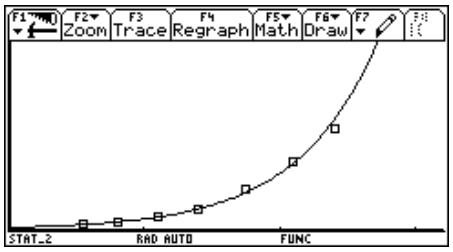
Evidement, je n'ai pas calculé toutes ces moyennes…

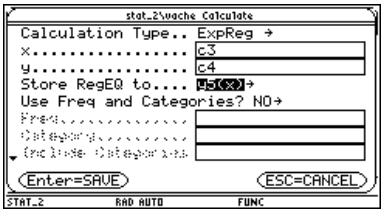
Le petit programme ci-contre l'a fait pour moi.

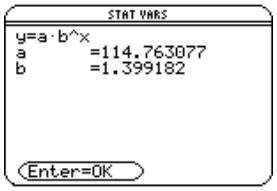
J'ai nommé 'vache' le tableau des données (se comporte comme une matrice), et 'veau' le nombre de moyennes à calculer en fonction des moyennes mobiles choisies.



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).





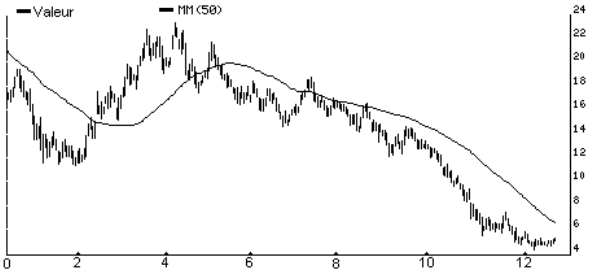


Remarque : cette méthode produit une 'petite' perte d'information. D'une part sur les premières et dernières valeurs, d'autre part en égalisant les valeurs (lissage). C'est un peu long à mettre en œuvre, mais très souvent utilisé (théories importantes en économie, bourse).

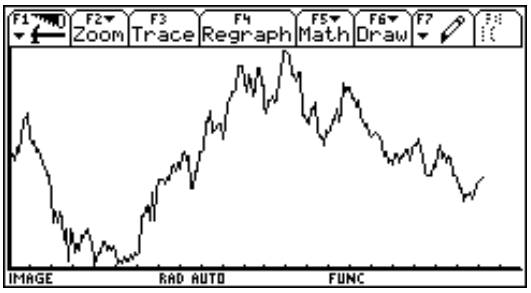
L'utilisation des moyennes mobiles permet généralement un bon lissage de la courbe, qui élimine les données 'extrêmes'. C'est un peu comme repasser l'allure de la courbe avec un gros feutre. Comme l'exemple suivant va le faire remarquer, c'est une méthode utile pour résumer une série statistique aux 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 d'où 239 données car 239 pixels).

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

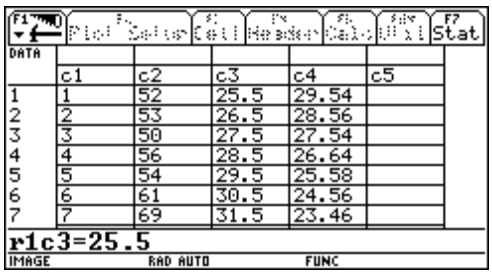


Le cours de la valeur sur internet et sa Moyenne Mobile (50).



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.


DATA

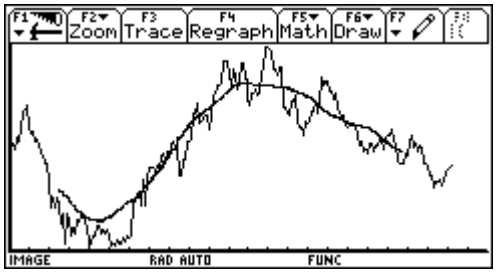
	c1	c2	c3	c4	c5
1	1	52	25.5	29.54	
2	2	53	26.5	28.56	
3	3	50	27.5	27.54	
4	4	56	28.5	26.64	
5	5	54	29.5	25.58	
6	6	61	30.5	24.56	
7	7	69	31.5	23.46	

r1c3=25.5

IMAGE RAD AUTO FUNC

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

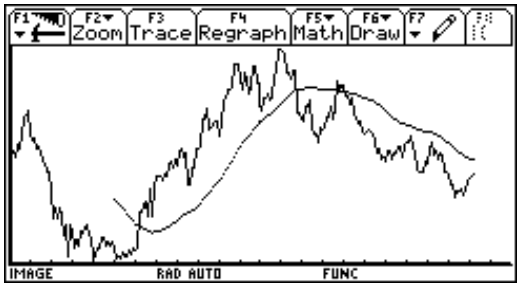
Encore un programme… pour calculer les moyennes mobiles (sur 50 valeurs) en un peu plus de 14 minutes…



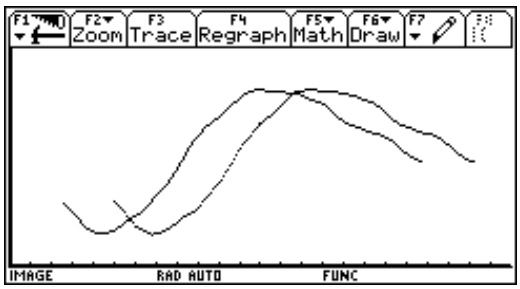
Et voici le résultat.

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.

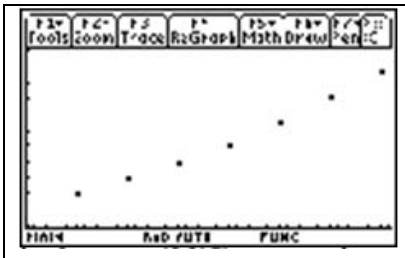
La moyenne mobile permet souvent d'obtenir une courbe de tendances 'rabotée', lissée. De là à en déduire le temps demain…

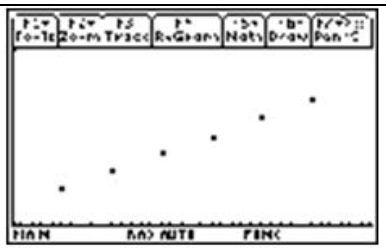
c) Ajustement affine dit « linéaire analytique, méthode des moindres carrés » :

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

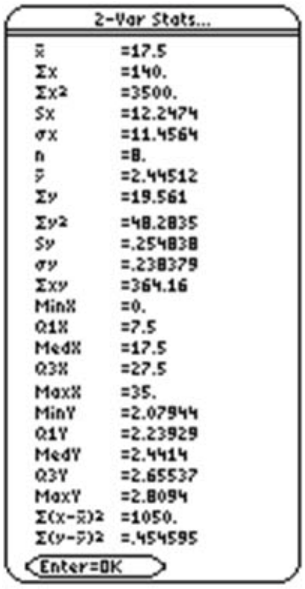
S_{20} est une série donnant la population d'une souche de culture

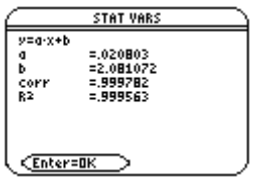
dans une boite de Petri relevée tous les 5 jours.





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. $y=0,02\times 42+2,08=2,92$ 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_{21} 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 $d_m=\frac{1}{N}\sum_{i=1}^{n}|x_i-m|$

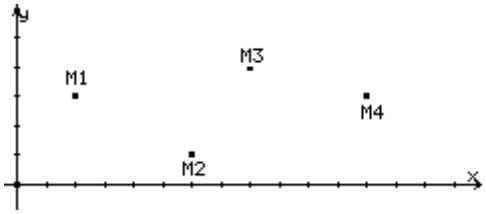
(ici $\frac{1}{3}(10+0+10) = 6{,}667$). Il parait que les valeurs absolues ne plaisent pas à tout le monde, elles fâchent certains. D'où la recherche d'une autre formulation, la bonne bien sûr, l'écart-type.

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

$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_i - m)^2}$ (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.

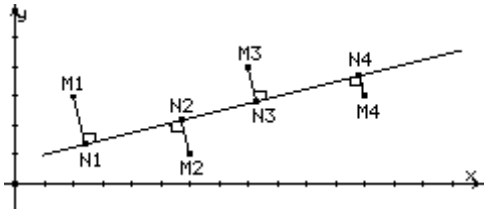
Le principe de « compliquer certaines parties pour en simplifier d'autres » est utilisé aussi dans le cas du choix de la méthode de la recherche d'une droite d'approximation d'un nuage de points à peu près affine (linéaire).



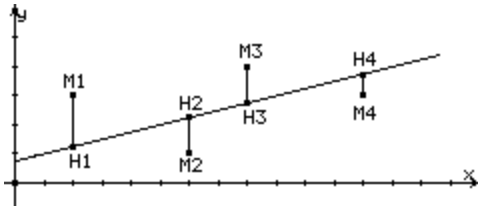
J'ai des points. Je cherche l'équation de la forme y=ax+b de <u>la meilleure</u> droite d'approximation de ce nuage.

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 H_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' « de x en y » on peut faire de même « de y en x » en considérant les écarts suivant l'horizontale et non plus selon la verticale. Je vous laisse reprendre les calculs.
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.