Chapitre 4 : Statistiques -04-11-11-
Terminale ES 2, 2011-2012, Y. Angeli

1. Rappels : statistiques à une variable

Un objectif des statistiques à une variable est d'étudier un *caractère* x mesuré sur un ensemble d'*individus*. Pour résumer une série de mesures $x_1, ..., x_n$, on donne souvent un couple de nombres : une valeur centrale et un paramètre de dispersion qui permet d'évaluer l'éloignement des termes de la série à la valeur centrale.

Vocabulaire

• L'ensemble des éléments étudiés est appelé la *population* (même s'il ne s'agit pas de gens !).
• Les différents éléments de cet ensemble sont appelés les *individus*.
• L'*effectif total* est le nombre N d'individus de la population.
• Le *caractère* ou la *variable statistique* x désignent la grandeur mesurée dans la population. Les valeurs du caractères pour les individus $1, 2, ..., N$ sont notées $x_1, x_2, ..., x_N$. Plusieurs des x_i peuvent être identiques.
• Les *modalités* $a_1, ..., a_k$ sont différentes valeurs prises par le caractère x (donc pas de répétition). L'*effectif* n_j d'une modalité a_j représente le nombre d'individus de la population dont le caractère vaut a_j. (c'est-à-dire le nombre de fois que a_j apparaît parmi les x_i). En particulier : $N = n_1 + ... + n_k$.

Le couple mode / étendue.

Le *mode* d'une série est la modalité la plus représentée, l'*étendue* est la différence entre la plus grande et la plus petite modalité.

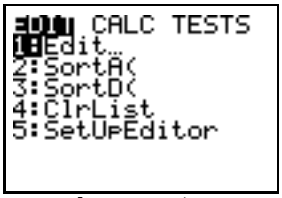
Remarque. Le seul intérêt de ces données est d'être simple à comprendre et à calculer.

Exemple. Soit la série de notes de 10 élèves :

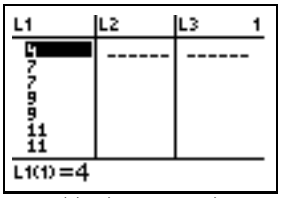
Élève n ^{o}	1	2	3	4	5	6	7	8	9	10
Note x_i	4	7	7	9	9	11	11	11	15	18

La population étudiée est le groupe d'élève, un individu de cette population est un élève et le caractère étudié est la note de l'élève. Cette série statistique a un effectif total de 10 (donc compte 10 termes x_i).
La série a 6 modalités sont 4, 7 (effectif 2), 9 (effectif 2), 11 (effectif 3), 15 et 18. Son mode est 11 (modalité la plus réprésentée), son étendue est 14 (valeur max moins valeur min).

Remarque. Pour entrer une série (x_i) dans une liste de la TI :



[STAT], Edit (1 :Edit)



$((x_i) \rightarrow L_1)$

2. Le couple moyenne / écart-type

L'avantage de la moyenne réside dans le fait qu'elle est linéaire : si une classe de 30 élèves a 10 de moyenne et une classe de 20 élèves a 12 de moyenne, on calcule la moyenne des 50 élèves par : $(30 \times 10 + 20 \times 12)/50 = 10,8$. Il n'est pas nécessaire de connaître les différentes notes de chaque classe.
L'inconvéniant de la moyenne est qu'elle est sensible aux valeurs extrêmes : dans une entrprise ou l'essentiel des salariés gagne le SMIC et un dirigeant un salaire extravagant, le salaire moyen peut être élevé.

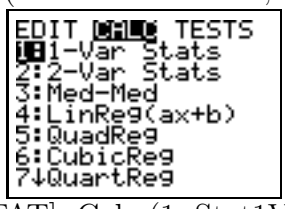
• La *moyenne* de la série est $\bar{x} = \dfrac{x_1 + ... + x_N}{N} = \dfrac{1}{N}\sum_{i=1}^{N} x_i = \dfrac{n_1 a_1 + ... + n_k a_k}{N} = \dfrac{1}{N}\sum_{j=1}^{k} n_j a_j$

• La *variance* V de la série est $V = \dfrac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \dfrac{1}{N}\left(\sum_{i=1}^{N} x_i^2\right) - \bar{x}^2 = \dfrac{1}{N}\sum_{i=j}^{k} n_j(a_j - \bar{x})^2$

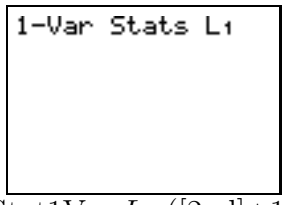
• L'*écart type* est $\sigma = \sqrt{V}$. (L'écart-type mesure la *dispersion* de la série autour de la moyenne)

Exemple. Dans l'exemple précédent : $\bar{x} = \frac{1}{10}(4 + 7 \times 2 + 9 \times 2 + 11 \times 3 + 15 + 18) = 10,2$.
$V = \frac{1}{10}((4 - 10,2)^2 + 2(7 - 10,2)^2 + 2(9 - 10,2)^2 + 3(11 - 10,2)^2 + (15 - 10,2)^2 + (18 - 10,2)^2) = 14,76$
$\sigma = \sqrt{V} \approx 3,84$.

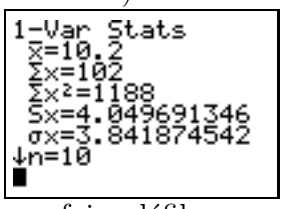
Si l'on a entré les différents x_i dans la liste L_1, la calculatrice donne $\bar{x}$ la moyenne, σx l'écart-type (et l'effectif total n, la somme des termes Σ et la somme des carrés Σx^2).



[STAT], Calc (1 :Stat1Var)



Stat1Var L_1 ([2nd]+1)



faire défiler.

3. Le couple médiane / écart-interquartile

L'inconvéniant de ces données est l'absence de linéarité (on ne peut calculer a priori la médiane de deux groupes en connaissant seulement la médiane de chacun des groupes). L'avantage est le peu de sensibilité aux valeurs extrêmes (dans l'exemple de l'entreprise, le salaire médian serait aussi égal au SMIC). On suppose la série des x_i rangée par ordre croissant.

• Une *médiane* M est telle que 50% au moins des caractères sont au dessus de M et 50% en dessous. Par convention : $M = \frac{1}{2}(x_{\frac{N}{2}} + x_{\frac{N}{2}+1})$ si N est pair, $M = x_{\frac{N+1}{2}}$ si N est impair.

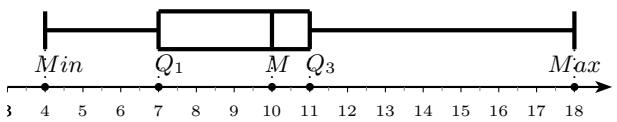
• Le *premier quartile* Q_1 est une médiane de la série des x_i où $i > \frac{N+1}{2}$.

• Le *troisième quartile* Q_3 est une médiane de la série des x_i où $i < \frac{N+1}{2}$.

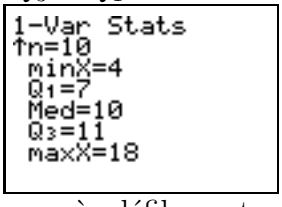
• L'*intervalle interquartile* est $]Q_1, Q_3[$. L'écart interquartile est $Q_3 - Q_1$.

Exemple. Dans l'exemple précédent :

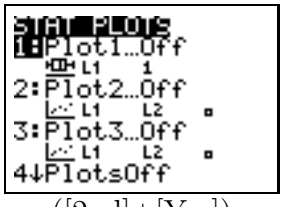
4 7 $\underbrace{7}_{Q_1=7}$ 9 $\underbrace{9 \,||\, 11}_{M=\frac{9+11}{2}=10}$ 11 $\underbrace{11}_{Q_3=11}$ 15 18



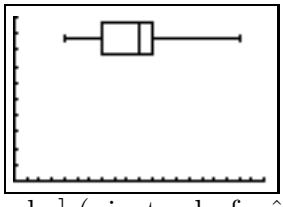
Donc $Q_3 - Q_1 = 4$.



après défilement



([2nd]+[Y=])


[Graphe] (ajuster la fenêtre)

4. Exemple : Structure de l'emploi du temps par genre

Le tableau suivant a été publié par l'organisme Eurostat le 8 mars 2006 (journée de la femme). Il présente la structure des emplois du temps des femmes et des hommes agés de 20 à 74 ans, en heures par jours, selon un sondage réalisé en 2005.

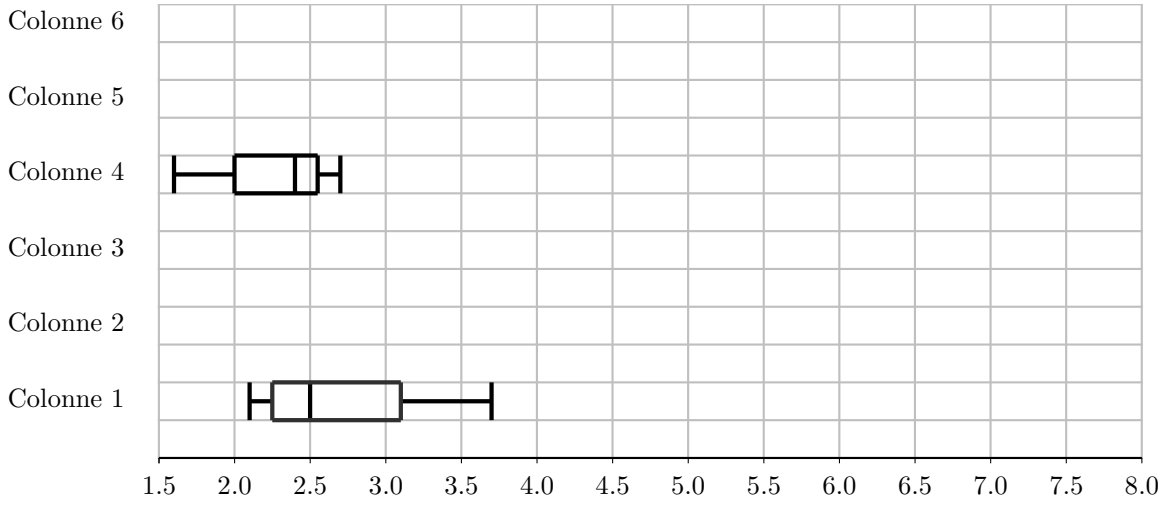
	Emploi, études		**Travail domestique**		**Travail total**		
	Femmes	Hommes	Femmes	Hommes	Femmes	Hommes	Différence
Belgique	2.1	3.5	4.5	2.6	6.6	6.1	0.5
Allemagne	2.1	3.6	4.2	2.4	6.3	6.0	
Espagne	2.4	4.7	4.9	1.6	7.3	6.3	
France	2.5	4.1	4.5	2.4	7.0	6.5	
Italie	2.1	4.4	5.3	1.6	7.4	6.0	
Lettonie	3.7	5.2	3.9	1.8	7.6	7.0	
Lituanie	3.7	4.9	4.5	2.2	8.2	7.1	
Hongrie	2.5	3.8	5.0	2.7	7.5	6.5	
Pologne	2.5	4.3	4.8	2.4	7.3	6.7	
Slovénie	3.0	4.1	5.0	2.7	8.0	6.8	
Suède	3.2	4.4	3.7	2.5	6.9	6.9	
Royaume-Uni	2.6	4.3	4.3	2.3	6.9	6.6	
Norvège	2.9	4.3	3.8	2.4	5.7	6.7	
Moyenne	2.7			2.3	7.2		
Variance	0.28			0.13			
Écart-type	0.53			0.36			
Minimum	2.1			1.6			
Quartile 1	2.25			2			
Médiane	2.5			2.4		6.6	
Quartile 3	3.1			2.55			
Maximum	3.7			2.7			

• **Travail payé et études des hommes** : Ranger les termes de la série de la seconde colonne par ordre croissant :

$\leqslant \quad \leqslant \quad \leqslant \quad \leqslant \quad \leqslant \quad \leqslant \quad \leqslant \quad \leqslant \quad \leqslant \quad \leqslant \quad \leqslant \quad \leqslant$

Moyenne = ...
Variance = ...

• Remplir les autres colonnes (calculatrice) et compléter le diagramme :



5. Statistiques à deux variables

On considère deux caractères d'une population, dont les valeurs sont $x_1, ..., x_N$ et $y_1, ..., y_N$.

• Dans le plan munis d'un repère orthogonal, l'ensemble des points $M_i(x_i, y_i)$ est le *nuage de points* associé à la série statistique à deux variables.

• Le *point moyen* associé à cette série est le point $G = (\bar{x}, \bar{y})$ où $\bar{x}$ et $\bar{y}$ sont les moyennes respectives des séries x et y.

• Rechercher un *ajustement affine* du nuage revient à déterminer une droite affine d'équation $y = ax + b$ aussi proche que possible des points du nuage.

• Étant donné un ajustement affine $\mathcal{D} : y = ax + b$, $r_i = y_i - (ax_i + b)$ représente, au signe près, l'écart entre le point $M'_i(x_i, ax_i + b)$ de $\mathcal{D}$ et le point du nuage $M_i(x_i, y_i)$. Le nombre r_i est appelé le *résidu* numéro i de l'ajustement.

• On mesure la qualité d'un ajustement affine par la somme des carrés des résidus :

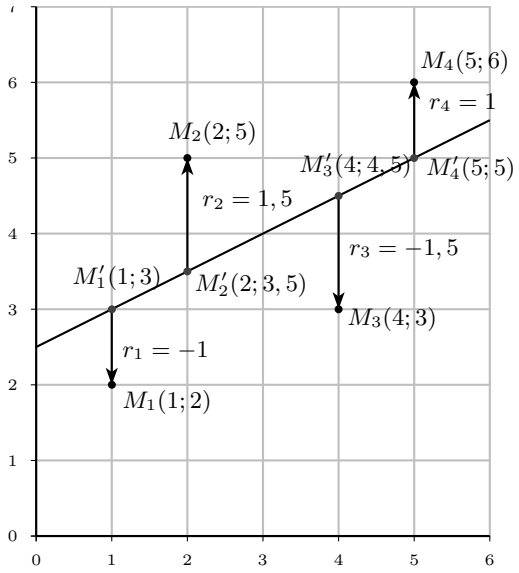
$$S = r_1^2 + r_2^2 + ... + r_N^2 = \sum_{i=1}^{N} (y_i - (ax_i + b))^2$$

Plus cette somme est petite, plus l'ajustement est précis.

Exemple. Soit la série double :

i	1	2	3	4
x_i	1	2	4	5
y_i	2	5	3	6

Le nuage des points $M_i(x_i; y_i)$ est représenté :
Le point moyen G a pour coordonnées $(3; 4)$ car :
$\bar{x} = \frac{1}{4}(1+2+4+5) = 3$ et $\bar{y} = \frac{1}{4}(2+5+3+6) = 4$.
On se propose d'ajuster le nuage par la droite d'équation $y = 0,5x + 2,5$.
Les résidus de cet ajustement sont :
$r_1 = y_1 - (ax_1 + b) = 2 - (0,5 * 1 + 2,5) = 2 - 3 = -1$
$r_2 = y_2 - (ax_2 + b) = 5 - (0,5 * 2 + 2,5) = 5 - 3,5 = 1,5$
$r_3 = y_3 - (ax_3 + b) = 3 - (0,5 * 4 + 2,5) = 2 - 3 = -1,5$
$r_4 = y_4 - (ax_4 + b) = 6 - (0,5 * 5 + 2,5) = 2 - 3 = 1$
La somme des carrés des résidus est :
$S = r_1^2 + r_2^2 + r_3^2 + r_4^2 = 6,5$



Théorème. Soit $x_1, ..., x_N$ et $y_1 ... y_N$ une série double. Si la variance de la série des x_i est non nulle, il existe un ajustement affine qui rend minimale la somme des carrés des résidus. ajustement Cet est la *droite des moindres carrés* (ou droite de régression linéaire) d'équation :

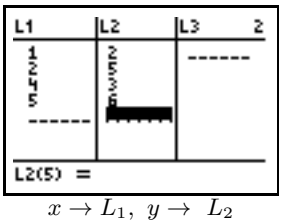
$$y = ax + b \text{ avec } a = \frac{\text{cov}(x, y)}{V(x)} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \text{ et } b = \bar{y} - a\bar{x}$$

Remarque. La droite des moindres carrés passe toujours par le point moyen G. Elle est le meilleur ajustement affine au sens où elle rend minimale la somme des carrés des résidus.

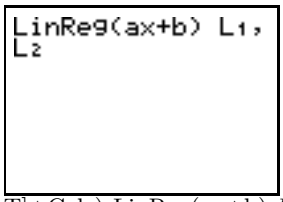
Exemple. Dans l'exemple précédent, déterminer l'équation de la droite des moindres carrés et montrer que la somme des carrés des résidus est inférieure à celle de l'ajustement proposé.

6. Utiliser la calculatrice

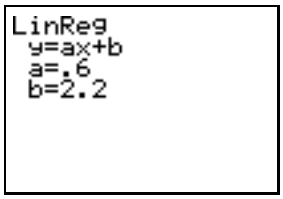
• Équation de la droite des moindres carrés :



$x \to L_1,\ y \to\ L_2$

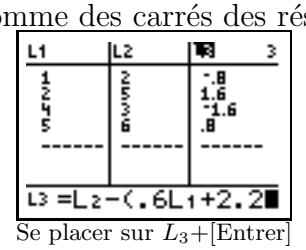


([STAT]+Calc) LinReg(ax+b) L_1, L_2

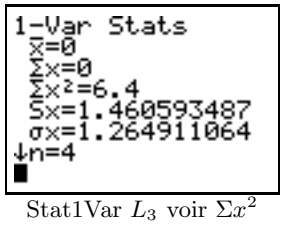


Résultat

• Somme des carrés des résidus.

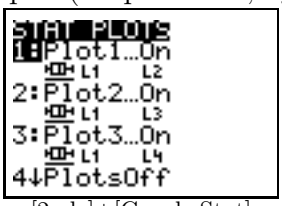


Se placer sur L_3+[Entrer]

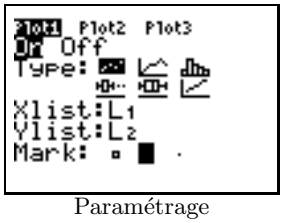


Stat1Var L_3 voir Σx^2

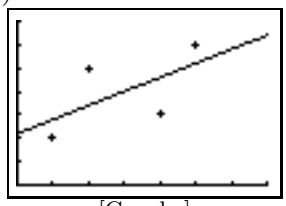
• Graphe (au préalable, ajuster la fenêtre et entrer $Y_1 = 0.6X + 2.2$)



[2nde]+[GrapheStat]



Paramétrage



[Graphe]

7. Exemple : inscrits au pôle emploi depuis septembre 2008

D'après l'INSEE, voici le nombre d'inscrits au pôle emploi (en milliers) entre septembre 2008 et août 2009 :

Mois	09-08	10-08	11-08	12-08	01-09	02-09	03-09	04-09	05-09	06-09	07-09	08-09
Mois n$^o x_i$	0	1	2	3	4	5	6	7	8	9	10	11
Inscrits y_i	3299	3348	3387	3438	3524	3605	3688	3786	3843	3851	3888	3923
Résidu r_i												

Les coordonnées du point moyen sont $(\bar{x}, \bar{y}) =$ .....

La droite d'ajustement des moindres carrés a pour équation : .....

La somme des carrés des résidus r_i est $\sum_{i=0}^{11} r_i^2 =$ .....

Combien de milliers d'inscrits peut-on prévoir pour septembre 2009 ? .....