CHAPITRE 6 : STATISTIQUES -28-01-11-
Première S 1, 2010-2011, Y. Angeli

1. Introduction, vocabulaire

Les statistiques sont une branche des mathématiques dont le but est de décrire des échantillons (importants) de données numériques par quelques grandeurs, en vue d'une interprétation. Une première approche consiste à dégager deux nombres à partir d'un échantillon : une valeur centrale et une valeur mesurant la dispersion de la série.

• La *population* désigne l'ensemble étudié. Un *individu* est un élément de cet ensemble, et l'*effectif total* N est le nombre d'individus de la population.
• Le *caractère* ou la *variable statistique* x désignent la propriété étudiée.
• Le terme x_i de la série statisitique est la valeur prise par le caractère x pour l'individu numéro i de la population.
• Les *modalités* $a_1, ..., a_k$ sont différentes valeurs prises par le caractère x.
• 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 : $N = n_1 + ... + n_k$
• La *fréquence* f_j d'une modalité a_j est le quotient $\frac{n_j}{N}$. C'est le pourcentage de la population dont le caractère vaut f_j.
• Le *mode* de la série est la modalité qui a le plus grand effectif.
• L'*étendue* est la différence entre la plus grande et la plus petite modalité.

Le mode et l'étendue sont un exemple naïf (et trop imprécis) de valeur centrale - paramètre de dispersion pour qualifier une série.

Exemple. Le tableau suivant recense les salaires annuels bruts en kilo euros de vingt salariés d'une entreprise :

Salaire x_i	21	22	22	81	37	24	24	23	23	24	23	22	22	23	37	23	21	23	24	24

Population= ....
Caractère= ....
Effectif total=...................
Mode=........................
Étendue=......................

Modalités						
Effectif						
Fréquence						
Cumul des Freq.						

2. Moyenne et écart-type

La moyenne et l'écart-type sont un exemple de valeur centrale et de paramètre de dispersion d'une série.

La *moyenne* de la série $x_1, ..., x_N$ dont les modalités sont $(a_1, n_1), ...(a_k, n_k)$ est
$$\bar{x} = \frac{x_1 + ... + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}\sum_{j=1}^{k} n_k a_k$$
La *variance* est la moyenne des carrés des écarts à la moyenne :
$$\mathrm{Var}(x) = \frac{(x_1 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{N} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\left(\sum_{i=1}^{N} x_i^2\right) - \bar{x}^2$$
L'*écart-type* est la racine carrée de la variance : $s(x) = \sigma(x) = \sqrt{\mathrm{Var}(x)}$

Exemple. Dans l'exemple du paragraphe 1, calculer :

$\bar{x} =$ .... $\mathrm{Var}(x) =$ .... $s(x) =$ .....
....

On suppose que le dirigeant gagne 41 kilo euros au lieu de 81. On a alors :

$\bar{x} =$ .... $\mathrm{Var}(x) =$ .... $s(x) =$ .....
....

On en conclue que la moyenne et l'écart-type sont sensibles aux valeurs extrêmes de la série.

Propriété. La moyenne et l'écart-type vérifient :
- ⋆ si la série x est composée des sous-séries y et z d'effectifs respectifs p et q alors on a $\bar{x} = \dfrac{p\bar{y} + q\bar{z}}{p + q}$.
- ⋆ Si $f : x \mapsto ax + b$ est une transformation affine, la série des $y_i = ax_i + b$ vérifie $\bar{y} = a\bar{x} + b$ et $s(y) = |a|s(x)$.

Preuve. Si on associe à chaque terme x_i de la série un point M_i d'abscisse x_i sur un axe gradué $(O; \vec{\imath})$, la moyenne $\bar{x}$ est l'abscisse de l'isobarycentre des M_i. La première formule est une conséquence de l'associativité du barycentre, et la seconde de son comportement par rapport aux homothéties et aux translations.

Exemple. La moyenne des salaires des 7 femmes travaillant dans l'entreprise de l'exemple 1 est de 30K€ par an. Quelle est la moyenne du salaire des hommes ? ....

3. Médiane et intervalle interquartile

La médiane et l'intervalle interquartile sont un autre exemple de valeur centrale et de paramètre de dispersion d'une série.

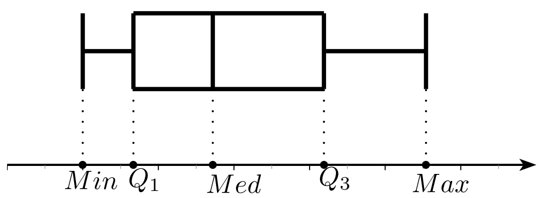
On considère une série $x_1, ..., x_N$ dont les termes sont rangés dans l'ordre croissant. Sa médiane est le terme central $x_{\frac{N+1}{2}}$ si N est impair ou la moyenne des deux termes centraux $x_{\frac{N}{2}}$ et $x_{1+\frac{N}{2}}$ si N est pair. Le *premier quartile* Q_1 est la médiane de la première moitié de la série (les x_i pour $i < (N+1)/2$) et le *troisième quartile* Q_3 est la médiane de la seconde moitié de la série (les x_i pour $i > (N+1)/2$). L'*intervalle interquartile* est la différence $Q_3 - Q_1$.

Exemple. Dans l'exemple du paragraphe 1,
$Med =$........... $Q_1 =$............. $Q_3 =$............. $Q_3 - Q_1 =$.......
On suppose que le dirigeant gagne 41 kilo euros au lieu de 81. On a alors :
$Med =$........... $Q_1 =$............. $Q_3 =$............. $Q_3 - Q_1 =$.......
On en conclue que la médiane et l'intervalle interquartile sont peu sensibles aux valeurs extrêmes.

Propriété. Soit x une série statistique de premier quartile Q_1, médiane M et troisième quartile Q_3. Soit $f : x \mapsto ax + b$ une fonction affine. Alors la médiane de la série des $y_i = ax_i + b$ est $aM + b$ et son intervalle interquartile $|a|(Q_3 - Q_1)$.

Preuve. si $a > 0$ la fonction affine est strictement croissante donc les y_i sont rangés par ordre croissant. La médiane est le terme $y_{\frac{N+1}{2}} = ax_{\frac{N+1}{2}} + b = aM + b$ si N est impair, et $(y_{\frac{N}{2}} + y_{1+\frac{N}{2}})/2 = (ax_{\frac{N}{2}} + b + ax_{1+\frac{N}{2}} + b)/2 = aM + b$ si N pair. On étudie de même les cas où $a < 0$ et $a = 0$.

Définition. Le *diagramme* en boîte d'une série statistique se construit ainsi :

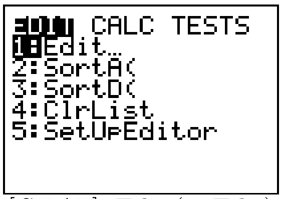


Remarque. Il existe plusieurs définitions de la médiane, des quartiles. Dans tous les cas, on peut dire que 25% des valeurs de la série sont inférieures où égales à Q_1, 50% des valeurs de la série sont inférieures où égales à la médiane et 75% des valeurs de la série sont inférieures où égales à Q_3. (Les quartiles divisent la série en quarts)

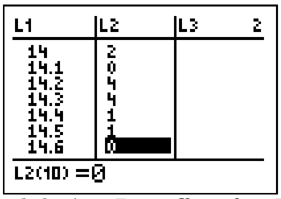
4. Utiliser la calculatrice

Exemple. Durant quatre cours de mathématiques, un professeur a fait tracer à chacun de ses élèves un carré de 10cm de côté et mesurer la diagonale. Ces mesures sont consignées ici :

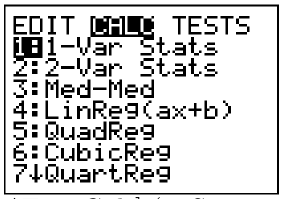
Modalité	13,7	13,8	13,9	14	14,1	14,2	14,3	14,4	14,5	14,6	N	Mode	Éte.	Moy.	Méd.
2nde1	1	1	2	2	0	4	4	1	1	0	16		0,8	14,14	14,2
2nde2	0	2	5	3	2	1	2	1	1	0					
2nde3	0	2	5	0	1	3	3	4	0	0					
2nde4	0	0	3	1	2	2	4	4	0	2					



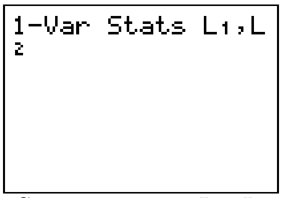
[$STAT$], $Edit$ (1 :Edit)



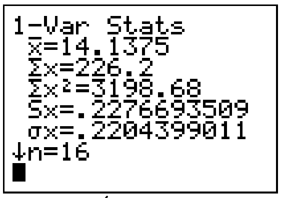
(Modalités→L1, effectifs→L2)



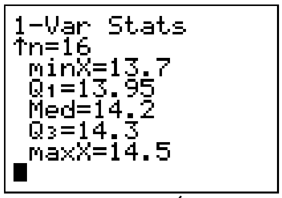
[$STAT \rightarrow Calc$] (1 :Stats-1-var)



$Stats - 1 - varL1, L2$



Écran 1



[Flèche bas] Écran 2

Exemple. Représenter le diagramme en boîte des secondes 2,3 et 4 :