Statistiques descriptives (M-2.1)

I. Séries statistiques à une variable

Exemple : on relève en bout de chaîne de production, les longueurs de 50 objets produits :

Longueur en cm : variable x_i	1,7	1,8	1,9	2	2,1	2,2	2,4	2,8
Effectifs correspondants : entier n_i	6	8	9	7	11	6	2	1

La population est l'ensemble sur lequel porte l'étude.
Exemple:
Un individu est un élément de la population étudiée.
Exemple:
Un échantillon est une partie de la population.
Exemple:
L'effectif total est le nombre d'individus de la population.
Exemple:

Le caractère est l'aspect que étudié dans la population. Si le caractère est un nombre, on parle de variable.
Exemple:
La série statistique est l'ensemble des valeurs collectées.
Exemple:
Les classes sont des intervalles regroupant plusieurs valeurs d'un caractère quantitatif.
Exemple:

L'effectif d'une valeur (ou d'une classe) du caractère est le nombre d'individus pour lesquels on observe cette valeur du caractère ou dont la valeur du caractère est incluse dans la classe. (on note souvent n_i l'effectif de la valeur x_i)

Longueur en cm regroupées en classes	$]1,4;1,8]$	$]1,8;2]$	$]2;2,2]$	$]2,2;2,8]$
Effectifs correspondants	14	16	17	3

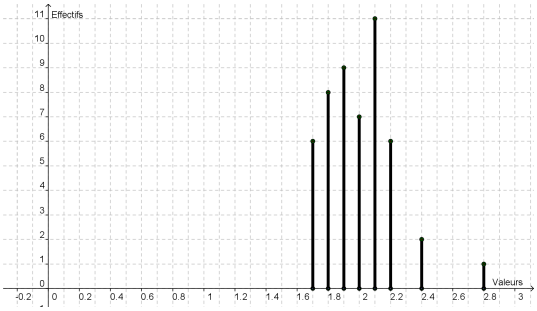
L'objectif des statistiques descriptives est de fournir des outils permettant de résumer une série statistique. On distingue en particulier les caractéristiques de tendance centrale (ou de position) et les caractéristiques de dispersion d'une série statistique à une variable.

1. Représentations graphiques des séries statistiques à une variable

Valeurs discrètes

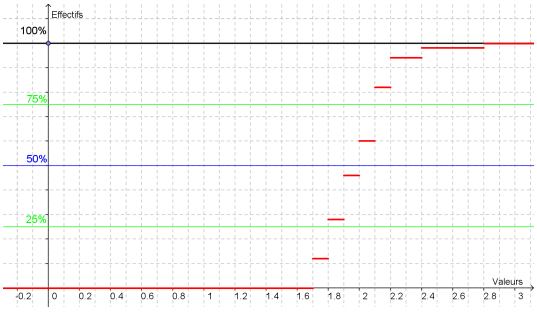
Un diagramme en bâton permet de représenter une série statistique ne prenant que quelques valeurs distinctes.
Chaque bâton a une hauteur proportionnelle à l'effectif du caractère.

- Les effectifs sont en ordonnée
- Les valeurs du caractère sont en abscisse



La courbe cumulative est la courbe représentative de la fonction en escalier F qui à tout réel a associe la fréquence de la classe $]-\infty;a]$.
Exemple : $F(1,89)=\ldots$; $F(2)=\ldots$

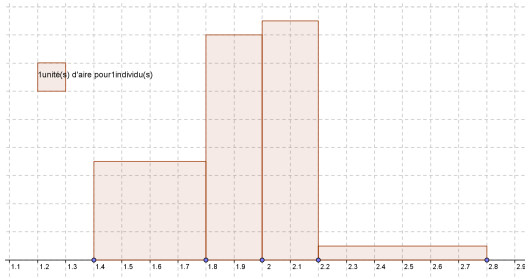
- Les fréquences sont en ordonnée
- Les valeurs du caractère sont en abscisse



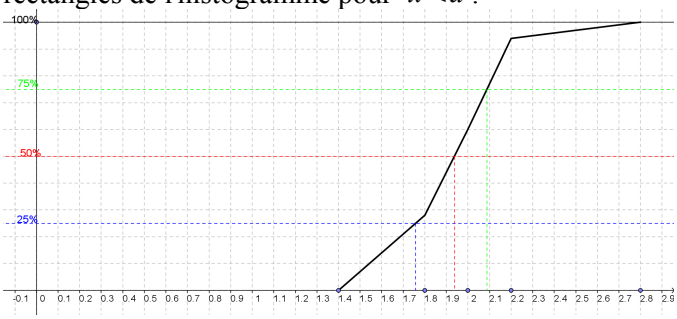
Valeurs regroupées par classes

Un histogramme permet de représenter une série statistique dont les valeurs sont regroupées en classes.
Chaque rectangle a une aire proportionnelle à l'effectif du caractère.

- Une unité d'aire est donnée pour les effectifs
- Seul l'axe des abscisses est tracé pour les valeurs du caractère



Le polygone des effectifs cumulés croissants est la courbe représentative de la fonction affine par morceaux qui à tout réel a associe la fréquence de la classe $]-\infty;a]$. Les effectifs sont estimés grâce aux sommes des aires des rectangles de l'histogramme pour $x \leqslant a$.



2. Caractéristiques de tendance centrale

La médiane

La valeur médiane d'une série statistique est une valeur du caractère souvent noté M_e telle que les effectifs des classes $]-\infty;M_e]$ et $[M_e;+\infty[$ soient au moins égaux à la moitié de la population totale.
On peut retenir: "la valeur médiane est une valeur du caractère qui partage l'effectif total en deux parties d'effectifs égaux"
Méthode : on ordonne tous les termes de la série par ordre croissant (chaque valeur est répétée un nombre de fois égal à son effectif):

- Si l'effectif total est impair : $N=2p+1$

$$x_1 \leqslant x_2 \leqslant \ldots \leqslant x_p \leqslant \left| x_{p+1} \right| \leqslant x_{p+2} \leqslant \ldots \leqslant x_{2p+1}$$

p plus petites valeurs | p plus grandes valeurs

donc : $\boxed{M_e=\frac{x_{p+1}}{2}}$

- Si l'effectif total est pair : $N=2p$

$$x_1 \leqslant x_2 \leqslant \ldots \leqslant x_p \left| \leqslant \right| x_{p+1} \leqslant \ldots \leqslant x_{2p}$$

p plus petites valeurs | p plus grandes valeurs

On fixe arbitrairement : $\boxed{M_e=\frac{x_p+x_{p+1}}{2}}$

Exemple :

Remarque : la médiane n'est pas sensible aux variations des valeurs extrêmes du caractère.

Pour une série statistique regroupée par classes, la médiane est déterminée grâce au polygone des effectifs cumulés croissants.

La moyenne

La moyenne d'une série statistique prenant p valeurs x_i avec les effectifs n_i est le nombre noté $\bar{x}$:

$$\boxed{\bar{x}=\frac{n_1\times x_1+n_2\times x_2+\ldots+n_p\times x_p}{n_1+n_2+\ldots+n_p}}$$

Remarques : on parle de moyenne pondérée par les effectifs (cf barycentres).
Exemple : $\bar{x}=\ldots$

Propriété : soit la fonction d qui à tout réel x associe la moyenne des carrés des écarts entre x et les valeurs de la série statistique :

$$d(x)=\frac{n_1(x-x_1)^2+n_2(x-x_2)^2+\ldots+n_p(x-x)^2}{n_1+n_2+\ldots+n_p}$$

Cette fonction d évalue une distance entre x est les valeurs de la série statistique.
La fonction d atteint son minimum en $x=\bar{x}$

Remarque : la moyenne est sensible aux variations des valeurs extrêmes du caractère.

Pour une série statistique regroupée par classes, la moyenne estimée en remplaçant les valeurs x_i par le centre des classes.
Exemple :

classes	$]1,4;1,8]$	$]1,8;2]$	$]2;2,2]$	$]2,2;2,8]$
centres				
effectifs	14	16	17	3

$\bar{x}=\ldots$

3. Caractéristiques de dispersion

Interquartiles

Le premier quartile q_1 est la plus petite valeur du caractère telle que la fréquence de la classe $]-\infty;q_1]$ dépasse 25%.
Le troisième quartile q_3 est la plus petite valeur du caractère telle que la fréquence de la classe $]-\infty;q_3]$ dépasse 75%.

Classes	$]-\infty;1,7]$	$]-\infty;1,8]$	$]-\infty;1,9]$	$]-\infty;2]$
Effectifs				
Fréquences				

Classes	$]-\infty;2,1]$	$]-\infty;2,2]$	$]-\infty;2,4]$	$]-\infty;2,8]$
Effectifs				
Fréquences				

Pour une série statistique regroupée par classes, les quartiles sont déterminés grâce au polygone des effectifs cumulés croissants.
L'intervalle interquartiles est l'intervalle $[q_1;q_3]$. Il contient au moins 50% des valeurs centrales de la série statistique.
L'écart inter-quartiles est la différence q_3-q_1.

Variance et écart-type

La variance V est la valeur du minimum de la fonction d.

$$\boxed{V=\frac{n_1(\bar{x}-x_1)^2+n_2(\bar{x}-x_2)^2+\ldots+n_p(\bar{x}-x_p)^2}{n_1+n_2+\ldots+n_p}}$$

"Moyenne des carrés des écarts à la moyenne"

Propriété : $\boxed{V=\frac{n_1\times(x_1)^2+n_2\times(x_2)^2+\ldots+n_p\times(x_p)^2}{n_1+n_2+\ldots+n_p}-(\bar{x})^2}$

"Moyenne des carrés moins le carré de la moyenne"

L'écart-type s de la série statistique est défini par $\boxed{s=\sqrt{V}}$
Remarques : la variance V est positive ou nulle car c'est la moyenne de carrés de nombres positifs. L'écart-type s est homogène aux valeurs de la variable de la série statistique. L'écart-type évalue l'écart moyen entre les valeurs de la série statistique et la moyenne. L'écart-type mesure donc la dispersion des valeurs autour de la valeur moyenne.
Exemple :
Ces calculs sont effectués à l'aide des modes statistiques des calculatrices travaillant sur les listes L_1 et L_2 :
Sur Casio :

```
1Var XList  :List1
1Var Freq   :List2
```

Sur TI :

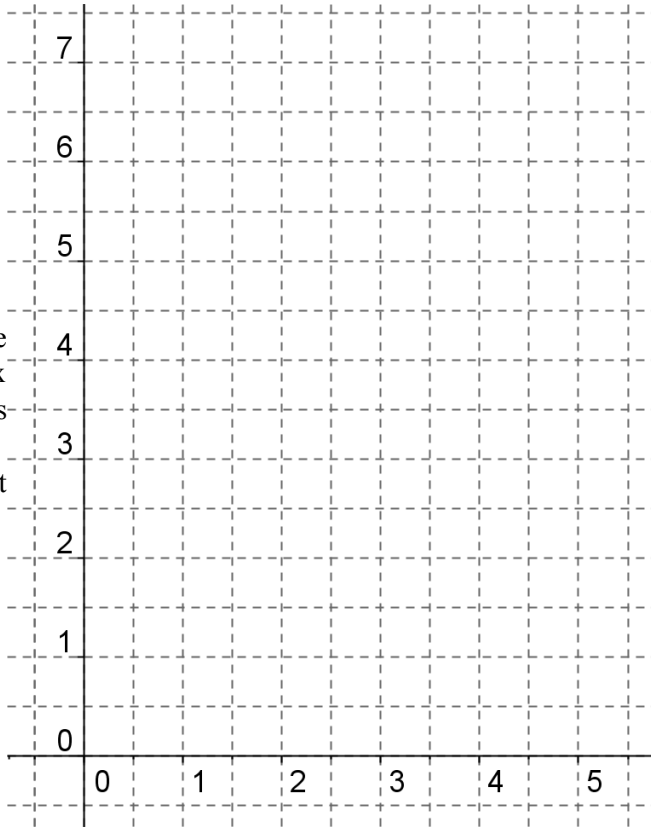
```
1-Var Stats L1,L
2
```

II. Séries statistiques à deux variables

Exemple : on mesure la longueur et la masse de 5 objets produits.

Longueur en cm : variable x_i	1,5	1,5	2	3,5	4
Masse en g : variable y_i	5	3	4	5	7

La série statistique à deux variables est l'ensemble des N couples de valeurs collectées $(x_i ; y_i)$.



1. Nuage de points

Définition d'un nuage de points : le plan étant muni d'un repère $(O;\vec{i};\vec{j})$, le nuage de points associé à une série statistique à deux variables x_i et y_i d'effectif total N est l'ensemble des N points de coordonnées $(x_1;y_1)$, $(x_2;y_2)$, ..., $(x_N;y_N)$.

Définition du point moyen : le point moyen G d'un nuage de point est l'isobarycentre des points constituants le nuage :

$$G\left(\frac{x_1+x_2+\ldots x_N}{N};\frac{y_1+y_2+\ldots y_N}{N}\right)$$

Exemple :

2. Ajustement affine

L'objectif est d'étudier la pertinence d'une relation affine entre les deux variables de la série statistique. D'un point de vue graphique, lorsque le nuage de point est allongé, il s'agit de déterminer la droite "approchant au mieux" les points du nuage.

Définition : la covariance de la série statistique double $(x_i;y_i)$ est le nombre réel noté $\text{cov}(x;y)$ défini par :

$$\text{cov}(x;y)=\frac{(x_1-\overline{x})\times(y_1-\overline{y})+(x_2-\overline{x})\times(y_2-\overline{y})+\ldots+(x_N-\overline{x})\times(y_N-\overline{y})}{N}$$

Remarque : soit le point $P_i(x_i;y_i)$ et le rectangle R_i ayant des cotés parallèles aux axes du repère et de diagonale $[GP_i]$ Alors le produit $(x_i-\overline{x})\times(y_i-\overline{y})$ est égal à l'aire algébrique du rectangle R_i : positive si la fonction affine f telle que $f(x_i)=y_i$ et $f(\overline{x})=\overline{y}$ est croissante, négative si f est décroissante.

La covariance est la moyenne des aires algébriques de tous les rectangles R_i elle permet donc d'évaluer les écarts entre les points du nuage et son point moyen mais ces écarts peuvent se compenser.

Contrairement à la variance, la covariance peut être négative.

Exemple :

Variable x_i	1,5	1,5	2	3,5	4
Variable y_i	5	3	4	5	7
Écarts $x_i-\overline{x}$					
Écarts $y_i-\overline{y}$					
Produit des écarts					

Donc : $\text{cov}(x;y)=\ldots$

Par ailleurs pour les variances : $V_x=\ldots$

$V_y=\ldots$

Droite de régression de y en x

L'objectif est d'approcher à l'aide d'une fonction affine les valeurs y_i en fonction des valeurs x_i, c'est-à-dire de déterminer deux réels a et b tels que $y_i\approx a x_i+b$

Les résidus sont alors les écarts $a x_i+b-y_i$ mesurés "verticalement" entre les points du nuage et la droite d'équation $y=ax+b$.

La somme quadratique des résidus évalue globalement la précision de l'ajustement affine des y_i en fonction des x_i

$$d=(ax_1+b-y_1)^2+(ax_2+b-y_2)^2+\ldots+(ax_N+b-y_N)^2$$

Droite de régression de x en y

L'objectif est d'approcher à l'aide d'une fonction affine les valeurs x_i en fonction des valeurs y_i, c'est-à-dire de déterminer deux réels a' et b' tels que $x_i\approx a' y_i+b'$

Les résidus sont alors les écarts $a' y_i+b'-x_i$ mesurés "horizontalement" entre les points du nuage et la droite d'équation $x=a' y+b'$ (si $a'\neq 0$, $y=\frac{1}{a'}x-\frac{b'}{a'}$).

La somme quadratique des résidus évalue globalement la précision de l'ajustement affine des x_i en fonction des y_i.

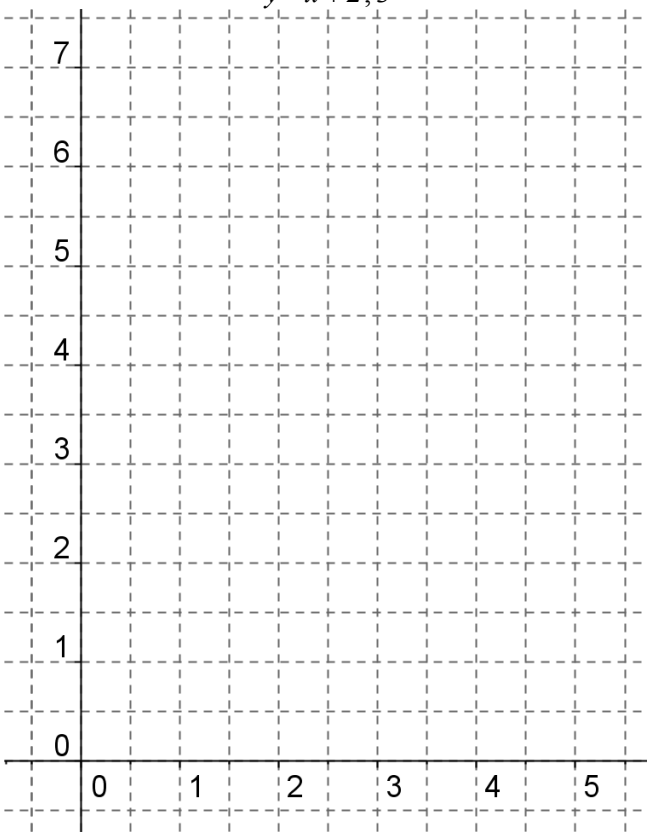
$$d'=(a' y_1+b'-x_1)^2+(a' y_2+b'-x_2)^2+\ldots+(a' y_N+b'-x_N)^2$$

Théorème : pour une série statistique à deux variables $(x_i ; y_i)$ donnée, il existe un unique couple de réels $(a ; b)$ minimisant le nombre $(x;y)$. La droite d'équation $y = ax + b$ alors appelée **droite de régression de y en x**, vérifie :

$$a = \frac{\text{cov}(x ; y)}{(\sigma_x)^2} \quad \text{et} \quad \overline{y} = a\overline{x} + b$$

Cette droite est aussi appelée "droite des moindres carrés".

Exemple : la droite de régression de y en x a pour équation $y = x + 2,3$



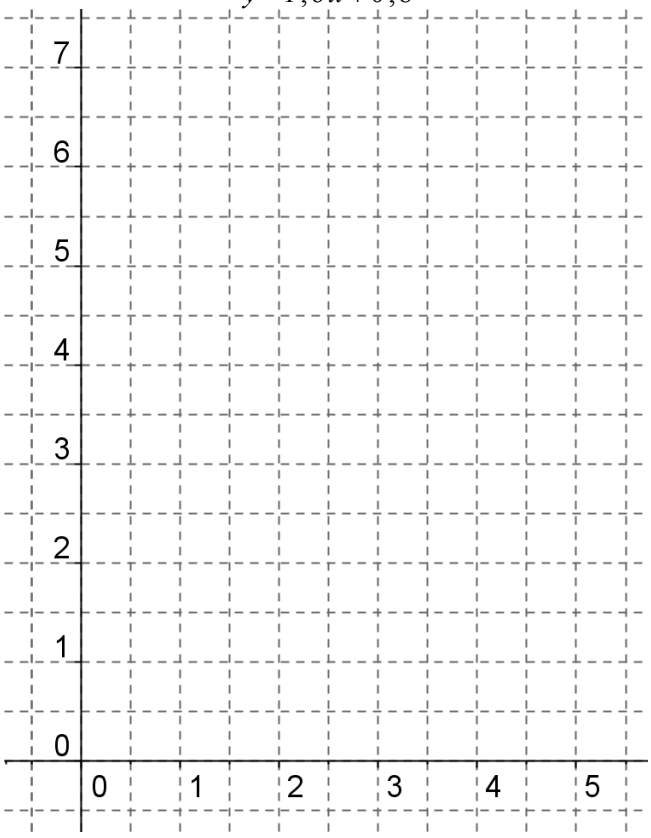
x_i	1,5	1,5	2	3,5	4
y_i	5	3	4	5	7
$ax_i + b$					
$(ax_i + b - y_i)^2$					

Somme quadratique des résidus :

Théorème : pour une série statistique à deux variables $(x_i ; y_i)$ donnée, il existe un unique couple de réels $(a' ; b')$ minimisant le nombre d'. La droite d'équation $x = a'y + b'$ alors appelée **droite de régression de x en y**, vérifie :

$$a' = \frac{\text{cov}(x ; y)}{(\sigma_y)^2} \quad \text{et} \quad \overline{x} = a'\overline{y} + b'$$

Exemple : la droite de régression de x en y a pour équation $y = 1,6x + 0,8$



x_i	1,5	1,5	2	3,5	4
y_i	5	3	4	5	7
$a'y_i + b'$					
$(a'y_i + b' - x_i)^2$					

Somme quadratique des résidus :

Sur Casio : CALC, REG, $\boxed{\times}$

Sur TI : CALC, 4:LinReg(ax+b)

Définition : **le coefficient de corrélation linéaire** de la série statistique $(x_i ; y_i)$ est le nombre réel $r = \frac{\text{cov}(x ; y)}{\sigma_x \times \sigma_y}$

Exemple : pour la série statistique précédente, le coefficient de corrélation linéaire est $r \approx 0,79$

Sur TI : TESTS, B:LinRegTTest...

Théorème : pour toute série statistique $(x_i ; y_i)$ le coefficient de corrélation linéaire r est compris entre -1 et 1 : $|r| \leqslant 1$

Interprétation : si le coefficient de corrélation linéaire est proche des valeurs extrêmes -1 ou 1, les variables x et y sont dites fortement corrélées. Cependant cette interprétation dépend du contexte et des objectifs.

⚠ Une forte corrélation entre deux variables ne signifie pas l'existence d'un lien de cause à effet entre ces deux variables.