BTS Mécanique et Automatismes Industriels

Statistique descriptive

Lycée Louis Armand, Poitiers, Année scolaire 2006/2007

Table des matières

Statistiques à une variable

Statistiques à deux variables

Statistiques : exercices

Statistiques à une variable

1. Vocabulaire

Une étude statistique porte sur un ensemble, appelé *la population*, (dont les éléments sont appelés *éléments*) et consiste à observer un même aspect (appelé *caractère*) sur chaque individu.

On appelle *échantillon* (ou *lot*) tout sous-ensemble de la population.

Un caractère peut être *discret* (nombre de frères et sœurs) ou *continu* (nombre d'heures passées devant la TV). Ce caractère peut également être *quantitatif* ou *qualitatif*, suivant qu'il est ou non mesurable.

L'*effectif* d'une valeur du caractère, c'est le nombre d'individus correspondant à cette valeur. Plus généralement, on appelle *classe* un sous-ensemble de la population correspondant à une même valeur ou à des valeurs « voisines » (le terme de voisinage est à définir au cas par cas) prises par le caractère. On appelle alors *effectif d'une classe* son nombre d'éléments.

La *fréquence* d'une classe, c'est l'effectif de cette classe divisée par l'effectif total de la population. Une fréquence est donc un nombre qui est toujours compris entre 0 et 1. On l'exprime souvent en pourcentage.

L'ensemble des classes constitue une *partition* de la population. Autrement dit :

- deux classes distinctes sont d'intersection vide,
- la population est contenue dans la réunion de toutes les classes.

1.1 - Moyenne arithmétique d'une série

• On appelle *moyenne arithmétique* de n nombres $x_1, x_2, \dots, x_n$ le quotient de la somme de toutes les valeurs par l'effectif total. On la note $\overline{x}$. On a ainsi

$$\boxed{\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}}$$

• Si l'on a p classes $x_1, x_2, \dots, x_p$, et que chaque classe x_i a un effectif n_i, alors la moyenne de la série est donnée par

$$\boxed{\overline{x} = \frac{n_1x_1 + n_2x_2 + \cdots + n_px_p}{n}}$$

où $n = n_1 + n_2 + \cdots n_p$ est l'effectif de la population totale.

• On suppose que l'on a p classes $[a_i, b_i[$, de centres respectifs $c_i = \frac{1}{2}(a_i + b_i)$, et que chaque classe a un effectif n_i. Si dans chaque classe $[a_i, b_i[$ les éléments sont : soit uniformément répartis, soit concentrés au milieu c_i, alors la moyenne de la série est donnée par

$$\boxed{\overline{x} = \frac{n_1c_1 + n_2c_2 + \cdots + n_pc_p}{n}}$$

où $n = n_1 + n_2 + \cdots n_p$ est l'effectif de la population totale.

Propriété : *Linéarité de la moyenne*

Soit a et b des constantes réelles. Alors quelle que soit la série de nombres réels $x_1, x_2, \dots x_n$, on a

$$\boxed{\overline{ax + b} = a\overline{x} + b.}$$

Autrement dit, $a\overline{x} + b$ est la moyenne de la série $ax_1 + b, ax_2 + b, \dots, ax_n + b$. ∎

1.2 - Médiane d'une série

Pour une série **ordonnée** quelconque, on appelle *médiane* et on note *Me* la valeur du caractère qui sépare l'ensemble de la population en deux parties de même effectif.

Par exemple, le salaire moyen en France sur l'année 1997 était de 11 000 F (je cite de mémoire) et le salaire médian était de 6 000 F (toujours de mémoire).

1.3 - Mode d'une série

On appelle *mode* (resp. *classe modale*) d'une série l'élément (resp. la classe) de la population correspondant au plus grand effectif.

1.4 - Écart absolu moyen par rapport à la moyenne

Soit une série de n nombres $x_1, x_2, \ldots, x_n$. On appelle *écart absolu moyen de la série par rapport à la moyenne* le nombre

$$\boxed{e = \frac{1}{n}\left(|x_1 - \overline{x}| + |x_2 - \overline{x}| + \cdots + |x_n - \overline{x}|\right)}$$

Remarque – En remplaçant $\bar{x}$ par Me, on définit de même l'*écart absolu moyen de la série par rapport à la médiane.*

1.5 - Variance, écart-type

• Pour une série $x_1, x_2, \ldots, x_n$ donnée, on définit la variance V

$$\boxed{V = \frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n} = \frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n} - \overline{x}^2}$$

• Si l'on a p classes $x_1, x_2, \ldots, x_p$, et que chaque classe x_i a un effectif n_i, alors la variance de la série est donnée par

$$\boxed{V = \frac{n_1(x_1 - \overline{x})^2 + \cdots + n_p(x_p - \overline{x})^2}{n} = \frac{n_1 x_1^2 + \ldots + n_p x_p^2}{n} - \overline{x}^2}$$

• On suppose que l'on a p classes $[a_i, b_i[$, de centres respectifs $c_i = \frac{1}{2}(a_i + b_i)$, et que chaque classe a un effectif n_i. Si, dans chaque classe $[a_i, b_i[$, **les éléments sont concentrés au milieu** c_i, alors la variance de la série est donnée par

$$\boxed{V = \frac{n_1(c_1 - \overline{x})^2 + \cdots + n_p(c_p - \overline{x})^2}{n} = \frac{n_1 c_1^2 + \ldots + n_p c_p^2}{n} - \overline{x}^2}$$

• Dans chacun des trois cas précédents, on définit l'écart-type σ par $\boxed{\sigma = \sqrt{V}}$.

2. Compléments : quartile et interquartile

On a vu que la médiane partage une population d'effectif n, ordonnée suivant les valeurs croissantes (ou décroissantes), en deux sous-populations de même effectif $n/2$.

On peut de la même façon, partager une population ordonnée en quatres sous-populations de même effectif $n/4$. Les nombres Q_1, Q_2 et Q_3 ainsi définis sont appelés les *quartiles*. À noter que l'on a $Q_2 = Me$ et que l'intervalle $[Q_1, Q_3]$ contient 50% des valeurs observées. Le nombre $Q_3 - Q_1$ est l'*interquaartile*. C'est un indicateur de dispersion.

Toujours de la même façon, on définit les *déciles* $D_1, D_2, \ldots, D_9$ et l'*interdécile* $D_9 - D_1$, ainsi que les *centiles* $D_1, D_2, \ldots, D_{99}$. et l'*intercentile* $D_{99} - D_1$.

3. Exemple

On considère la série

$$1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17.$$

Son effectif est 17, sa moyenne 9, sa variance 24, son écart-type est de $\sqrt{24} = 2\sqrt{6}$. Cependant que la médiane est 9, son premier quartile $4,5$ et son troisième quartile est $13,5$.

Statistiques à deux variables

1. Introduction. Définition

On observe que, dans certains cas, il semble exister un lien entre deux caractères d'une même population : par exemple entre l'épaisseur d'un mur et sa résistance thermique, entre la consommation et la vitesse d'une voiture, entre le temps de fonctionnement d'un appareil et la fréquence des avaries, etc. . .

Pour étudier d'éventuelles *corrélations*, on est amené à s'intéresser simultanément à deux caractères x et y d'une même population. On définit alors une *série statistique à deux variables x et y* prenant des valeurs $x_1, x_2, \ldots, x_n$ et $y_1, y_2, \ldots, y_n$.

2. Nuage de points

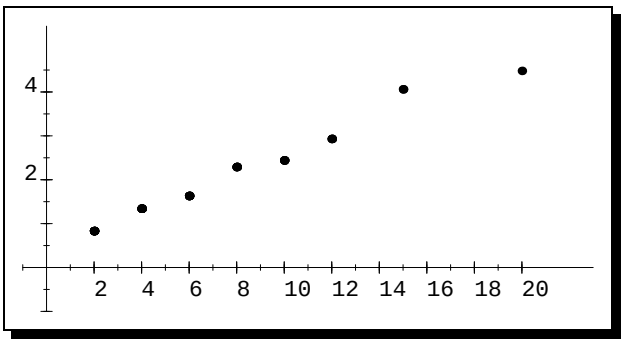
Le plan étant muni d'un repère orthogonal, on peut asspocier au couple (x_i, y_i) de la série statistique double le point M_i de coordonnées (x_i, y_i). L'ensemble des points M_i ains obtenu est appelé *nuage de points* représentant la série statistique.

Le nuage étant dessiné, et s'il existe une certaine *corrélation* entre les deux caractères étudiés, on peut essayer de trouver une fonction f telle que la courbe d'équation $y = f(x)$ passe « le plus près possible » des points du nuage. C'est le problème de l'*ajustement*.

Exemple (1) .

Le mur d'une habitation est constitué par une paroi en béton et une couche de polystyrène d'épaisseur variable x (en cm). On a mesuré, pour une même épaisseur de béton, la résistance thermique y (en $m^2 \cdot ^\circ / \text{watt}$) de ce mur pour différentes valeurs de x. On a obtenu les résultats suivants :

Épaisseur x_i	2	4	6	8	10	12	15	20
Résistance y_i	0,83	1,34	1,63	2,29	2,44	2,93	4,06	4,48



Au vu de ce nuage de points, on peut penser que, en première approximation, il est possible de tracer une droite D au voisinage de ces 9 points. On dit alors que l'on a un *ajustement affine*.

■

3. Point moyen

Lorsque l'on pense pouvoir réaliser un ajustement affine d'un nuage, il peut sembler intéressant, avant de tracer la droite, de placer le point dont l'abscisse est la moyenne des abscisses x_i et dont l'ordonnée est la moyenne des ordonnées y_i.

On appelle *point moyen* d'un nuage de n points $M_i(x_i, y_i)$ le point G de coordonnées

$$\boxed{x_G = \overline{x} = \frac{1}{n}(x_1 + x_2 + \cdots x_n)} \quad \text{et} \quad \boxed{y_G = \overline{y} = \frac{1}{n}(y_1 + y_2 + \cdots y_n).}$$

4. Ajustement affine

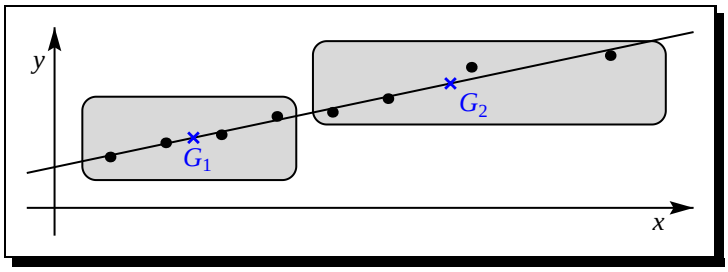
4.1 - Ajustement à la règle

On commence par représenter le nuage de points, puis on trace au jugé une droite D passant le plus près possible des points du nuage. Pour ce faire, on utilise une règle transparente et on la dispose suivant la direction constatée, en s'efforçant d'équilibrer les nombres de points situés de part et d'autre suivant les abscisses croissantes.

4.2 - Méthode de Mayer

On commence par trier les points selon leurs abscisses croissantes, puis on détermine la médiane des x_i afin de partager le nuage en deux parties ayant le même nombre de points. On détermine ensuite G_1 et G_2, les points moyens respectifs de chacune de ces parties. La droite (G_1G_2) est appelée *droite de Mayer* de la série statistique.

Il est à noter que la droite de Mayer d'un nuage passe toujours par le point moyen, de ce nuage.



4.3 - Les droites de régression

On considère une série statistique à deux variables représentée par un nuage justifiant un ajustement affine.

Soit D une droite d'ajustement et $M_i(x_i, y_i)$ un point du nuage. On note P_i le point de D d'abscisse x_i (fig. 1), et Q_i le point de D d'ordonnée y_i (fig. 2).

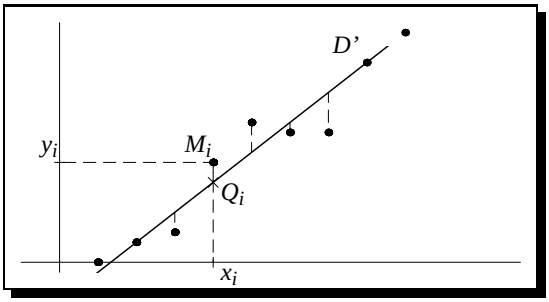


fig. 1

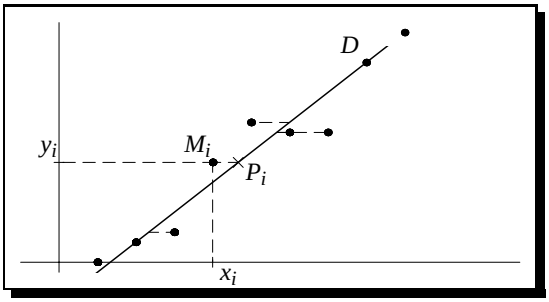


fig. 2

On appelle *droite de régression de y en x* la droite D telle que la somme

$$\sum_{i=1}^{n} M_iQ_i^2 = \sum_{i=1}^{n}[y_i - (ax_i + b)]^2 \qquad \text{soit minimale (fig. 1).}$$

On appelle *droite de régression de x en y* la droite D telle que la somme

$$\sum_{i=1}^{n} M_iP_i^2 \qquad \text{soit minimale (fig. 2).}$$

4.4 - Covariance d'une série statistique double

On appelle *covariance* de la série statistique double de caractères x et y le nombre réel

$$\boxed{\text{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}$$

où $\overline{x}$ et $\overline{y}$ désignent respectivement les moyennes arithmétiques des série statistiques à une variable x et y. On note aussi $\text{cov}(x, y) = \sigma_{xy}$.

On a une autre formule, plus commode pour les calculs :

$$\boxed{\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy}.}$$

4.5 - Équation des droites de régression

On admettra que :

• la droite de régression D de y en x a pour équation $y = ax + b$ où

$$\boxed{a = \frac{\sigma_{xy}}{[\sigma(x)]^2}} \quad \text{et où } b \text{ vérifie} \quad \boxed{\overline{y} = a\overline{x} + b}.$$

(La notation $[\sigma(x)]^2$ désignant la variance de la série statistique à une variable x.)

• la droite de régression D' de x en y a pour équation $x = a'y + b'$ où

$$\boxed{a' = \frac{\sigma_{xy}}{[\sigma(y)]^2}} \quad \text{et où } b' \text{ vérifie} \quad \boxed{\overline{x} = a'\overline{y} + b'}.$$

(La notation $[\sigma(y)]^2$ désignant la variance de la série statistique à une variable y.)

On remarque que les droites D et D' passent toutes deux par le point moyen $G(\overline{x}, \overline{y})$ du nuage.

5. Coefficient de corrélation linéaire

5.1 - Définition

Le *coefficient de corrélation linéaire* d'une série statistique double de variables x et y est le nombre r défini par

$$\boxed{r = \frac{\sigma_{xy}}{\sigma(x) \times \sigma(y)}.}$$

Ce coefficient sert à mesurer la qualité d'un ajustement affine. Par exemple, dans l'exercice précédent, on avait $[\sigma(y)]^2 = 76,04$, donc $r \approx 0,998$.

5.2 - Propriétés

- Le coefficient r est un nombre réel. De plus, comme $a = \sigma_{xy}/[\sigma(x)]^2$ et $a' = \sigma_{xy}/[\sigma(y)]^2$, on a la propriété

$$\boxed{\text{cov}(x, y),\ r,\ a \text{ et } a' \text{ sont de même signe.}}$$

- Le coefficient de corrélation linéaire r vérifie

$$\boxed{-1 \leqslant r \leqslant 1}$$

5.3 - Interprétation graphique

Dans tout ce paragraphe, on ne considère que les nuages de points « allongés », qui incitent à ajuster par une droite.

Le coefficient de corrélation r est lié, d'une part au coefficient directeur a de D, la droite de régression de y en x, d'autre part à $1/a'$, le coefficient directeur de D', la droite de régression de x en y. Les deux droites passant par le point moyen G, le coefficient r donne des indications sur l'angle des deux droites.

5.3.1 - Ajustement affine parfait

Dans le cas où $r^2 = 1$, on a $aa' = 1$, et donc $a = 1/a'$. Les coefficients directeurs de D et D' sont égaux. Les droites D et D' sont confondues (puisqu'elles ont déjà le point moyen G en commun), et on dit que l'on a un *ajustement affine parfait*.

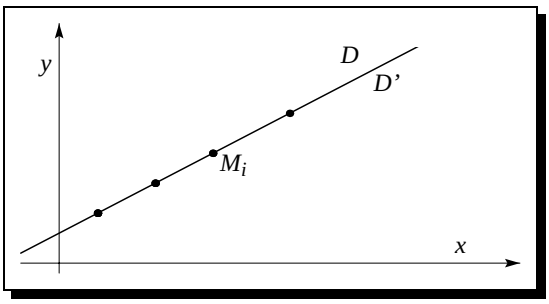


fig. 3 $r = 1, a > 0, a' > 0$

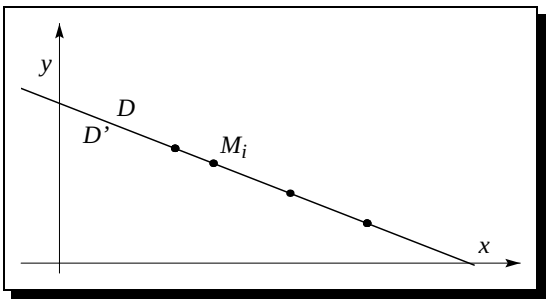


fig. 4 $r = -1, a < 0, a' < 0$

5.3.2 - Forte corrélation

Dans le cas où $|r|$ est « proche » de 1, alors les droites D et D' sont « proches » l'une de l'autre. On dit qu'il y a une « bonne » corrélation entre les deux caractères.

Les termes « proche » et « bonne corrélation » sont à définir au cas par cas. Dans le bâtiment par exemple, il arrive que l'on se contente de $|r| \approx 0,5|$, alors que dans la maintenance, on demande fréquemment $0,999 \leqslant |r| \leqslant 1$.

Attention : il ne faut surtout pas confondre forte corrélation et liaison de cause à effet : x_i et y_i peuvent, par exemple, mesurer deux effets d'une même cause. . .

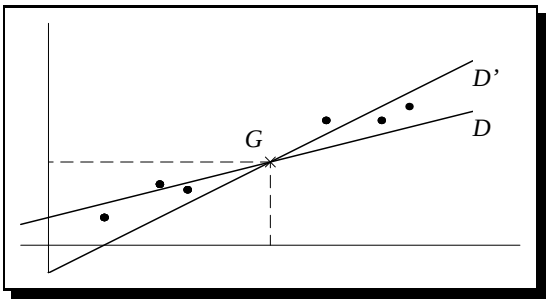


fig. 5 r proche de 1, $a > 0, a' > 0$

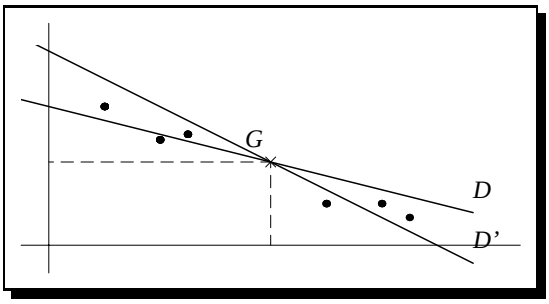


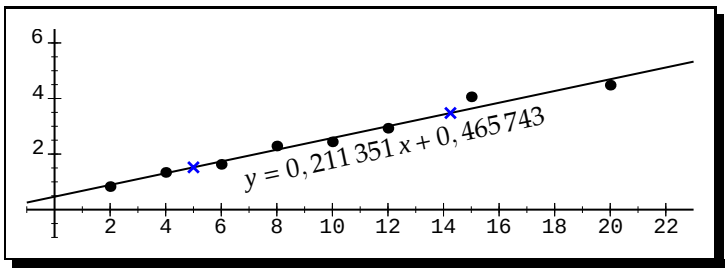
fig. 6 r proche de -1, $a < 0, a' < 0$

6. Un exemple complet

Prenons la série double définie ci-dessous :

x_i	2	4	6	8	10	12	15	20
y_i	0,83	1,34	1,63	2,29	2,44	2,93	4,06	4,48

La droite de Mayer a pour équation $y = 0,211\,351x + 0,465\,743$:



Le coefficient de corrélation est $r \approx 0,987\,635$.

La droite de régression des y en x a pour équation $y = 0,212\,527x + 0,454\,432$.

La droite de régression des x en y a pour équation $y = 0,217\,882x + 0,402\,89$.

Statistiques : exercices

Exercice 1 : Attention, statistiques !

Les résultats suivants ont été obtenus après une enquête en classe de Terminale :

	2000		2001	
	présentés	reçus	présentés	reçus
non-redoublants	22	12	15	8
redoublants	3	3	10	9
total	25	15	25	17

1. *a*) Vous êtes proviseur et, dans votre discours de fin d'année, vous voudriez dire la phrase suivante : « *L'année 2001 marque une progression dans la réussite au bac. Je félicite les professeurs de la classe* ». Peut-on justifier cette phrase au vu des résultats ?

b) Vous êtes délégué des élèves en Terminale, et vous avez envie de dire au conseil « *L'année 2001 marque une régression dans la réussite au bac. Je ne félicite pas les professeurs de la classe* ». Peut-on justifier cette phrase au vu des résultats ?

2. N'y a-t-il pas un paradoxe qui apparaît ? Lequel ?

3. Comment résoudre ce paradoxe ?

Exercice 2 : Ajustement affine, droite de Mayer

Le mur d'une habitation est constitué par une paroi en béton et une couche de polystyrène d'épaisseur variable x (en cm). On a mesuré, pour une même épaisseur de béton, la résistance thermique y (en $\mathrm{m}^2 \cdot {}^\circ/\mathrm{watt}$) de ce mur pour différentes valeurs de x. On a obtenu les résultats suivants :

Épaisseur x_i	2	4	6	8	10	12	15	20
Résistance y_i	0,83	1,34	1,63	2,29	2,44	2,93	4,06	4,48

1. Tracer le nuage de points représentant cette série.

2. On choisit, comme droite d'ajustement, la droite D passant par les points $A(6; 1,63)$ et $B(12; 2,93)$.

a) Représenter cette droite et en déterminer une équation.

b) Quelle résistance thermique peut-on espérer obtenir avec une couche de polystyrène de 18 centimètres d'épaisseur ?

3. *a*) Déterminer une équation de la droite de Mayer de cette série et représenter cette droite.

b) Reprendre la question **2.***b*)

Exercice 3 : Comparaison de positions, de dispersions

À un examen, les élèves A et B ont obtenu les notes suivantes :

$$A : \quad 7, 8, 11, 12, 13, 13, 13 \qquad \text{et} \qquad B : 4, 7, 9, 12, 13, 13, 19.$$

1. Comparer les modes, moyennes et médianes de ces deux séries.

2. Calculer, pour chacune de ces séries, l'écart absolu moyen par rapport à la moyenne.

3. Déterminer variance et écart-type pour chacune de ces séries.

Exercice 4 : Détermination de la médiane

Le responsable d'un magasin de petit outillage a relevé pendant une semaine le montant en francs des achats de 200 clients. Les résultats figurent dans le tableau suivant :

Montant des achats x_i	Nombre de clients n_i
[50, 150[	10
[150, 250[	22
[250, 350[	52
[350, 450[	62
[450, 550[	36
[550, 650[	14
[650, 750[	4

1. Quel est le pourcentage de clients dont le montant des achats est situé dans l'intervalle $[250, 550[$?

2. Dresser le tableau des fréquences cumulées croissantes de cette série statistique.

3. Représenter l'histogramme des fréquences cumulées croissantes de cette série statistique. Échelle : 1 cm pour 50 francs sur l'axe des abscisses et 1 cm pour 0, 1 sur l'axe des ordonnées.

4. Déterminer la moyenne $\overline{x}$ et l'écart-type σ de cette série.

5. On suppose que, dans chaque classe, les éléments sont répartis de manière uniforme. On peut alors remplacer l'histogramme par la ligne brisée définie par le point d'abscisse 50 et d'ordonnée 0 et chacun des sommets supérieurs droits des rectangles.

a) Tracer cette ligne brisée.

b) Par lecture du graphique, estimer le pourcentage de clients dont le montant est compris entre $\overline{x} - \sigma$ et $\overline{x} + \sigma$.

c) Déterminer par le calcul une valeur approchée à un franc près de l'abscisse du point I de la ligne brisée d'ordonnée $0,5$. Vérifier sur le graphique. Que représente cette abscisse ?

Exercice 5 : Influence des hypothèses sur un caractère continu

Dans cet exercice, on calcule la moyenne et l'écart-type de quatres populations différentes correspondant à un même tableau d'effectifs. On observera ainsi l'influence de la répartition des éléments de la population à l'intérieur de chaque classe.

On considère le tableau d'effectifs suivant :

Classe	$[0, 4[$	$[4, 8[$	$[8, 12[$
Effectif	4	4	4

1. On suppose que, dans chaque classe, tous les éléments sont situés au centre de la classe. La population est donc

$$2, \quad 2, \quad 2, \quad 2, \quad 6, \quad 6, \quad 6, \quad 6, \quad 10, \quad 10, \quad 10, \quad 10.$$

Calculer la moyenne $\overline{x}$ et une valeur approchée à 10^{-2} près de l'écart-type σ de cette première population.

2. On suppose que les éléments de chaque classe sont répartis uniformément de la manière suivante :

$$0,5\,; \quad 1,5\,; \quad 2,5\,; \quad 3,5\,; \quad 4,5\,; \quad 5,5\,; \quad 6,5\,; \quad 7,5\,; \quad 8,5\,; \quad 9,5\,; \quad 10,5\,; \quad 11,5.$$

Calculer la moyenne $\overline{x}'$ et une valeur approchée à 10^{-2} près de l'écart-type σ' de cette deuxième population.

3. On suppose que les éléments de chaque classe sont répartis de la manière suivante :

$$1, \quad 1, \quad 3, \quad 3, \quad 5, \quad 5, \quad 7, \quad 7, \quad 9, \quad 9, \quad 11, \quad 11.$$

a) Calculer la moyenne $\overline{x}''$ et une valeur approchée à 10^{-2} près de l'écart-type σ'' de cette troisième population.

b) Comparer $\overline{x}, \overline{x}', \overline{x}''$ d'une part, et $\sigma, \sigma', \sigma''$ d'autre part.

4. On suppose que, dans chaque classe, tous les éléments sont situés d'un même côté, et le plus loin possible du centre de la classe, c'est à dire que la population est :

$$0,\quad 0,\quad 0,\quad 0,\quad 4,\quad 4,\quad 4,\quad 4,\quad 8,\quad 8,\quad 8,\quad 8.$$

Calculer la moyenne $\overline{x}'''$ et une valeur approchée à 10^{-2} près de l'écart-type σ''' de cette quatrième population. Pouvait-on prévoir les valeurs de $\overline{x}'''$ et σ''' ?

Exercice 6 : Caractère qualitatif

On considère le tableau ci-dessous relatif aux ventes de voitures neuves en France en août 1996. La propriété étudiée est la marque : c'est un caractère qualitatif qui prend trois valeurs (ou *modalités*) permettant de définir trois classes avec leurs fréquences :

Classe	Renault	PSA	Étranger
Fréquence	$0,264$	$0,259$	$0,477$

1. Faire le diagramme à secteurs circulaire correspondant à cette série : chaque classe correspond à un secteur circulaire dont l'angle (ou l'aire) est proportionnel à l'effectif, donc à la fréquence de la classe.

2. Faire le diagramme en tuyaux d'orgues (ou en rectangles) correspondant à cette série : chaque classe est représentée par un rectangle de même largeur et de longueur proportionnelle à l'effectif, donc à la fréquence de la classe.

3. Faire le diagramme en bandes correspondant à cette série.

4. On voudrait représenter cette série par des diagrammes représentant des voitures. Comment procéder ?

Exercice 7 : Caractère quantitatif discret

Le responsable des ventes d'un fournisseur de matériel électronique a noté le niveau de la demande journalière pour un produit pendant cent jours ouvrables consécutifs :

Nombres x_i d'unités demandées par jour	Nombre n_i de jours où l'on a vendu x_i	Fréquence f_i
0	5	$0,05$
1	15	$0,15$
2	23	$0,23$
3	22	$0,22$
4	16	$0,16$
5	9	$0,09$
6	5	$0,05$
7 et plus	5	$0,05$

1. Ajouter sur ce tableau la colonne des fréquences cumulées croissantes (la **fréquence cumulée croissante** (resp. décroissante) d'une classe est la somme de la fréquence de cette classe et de toutes celles qui la précèdent (resp. qui la suivent). On définit de même les **effectifs cumulés**.

2. Tracer, pour cette série, le diagramme en bâtons des effectifs.

3. Comment passer du diagramme précédent au diagramme en bâtons des fréquences ?

Exercice 8 : Caractère quantitatif continu

Chez un fournisseur de matériaux pour le bâtiment, on a relevé les montants des achats effectués par mille clients au

cours d'un mois donné. On a obtenu les résultats suivants :

Montant des retraits (en francs)	Nombre n_i de clients	Effectif cumulé croissant
[0, 500[	5	5
[500, 1 000[	12	17
[1 000, 1 500[	33	50
[1 500, 2 000[	71	121
[2 000, 2 500[	119	240
[2 500, 3 000[	175	415
[3 000, 3 500[	185	600
[3 500, 4 000[	158	758
[4 000, 4 500[	122	880
[4 500, 5 000[	69	949
[5 000, 5 500[	35	984
[5 500, 6 000[	11	995
6 000 et plus	5	1 000
TOTAL	1 000	

Les classes ayant toutes la même amplitude 500, on convient d'assimiler la classe « 6 000 et plus » à la classe [6 000, 6 500[.

1. Tracer l'**histogramme des effectifs** correspondant à cette série.
2. On appelle **ligne polygonale des effectifs** la ligne brisée joignant les milieux c_i des largeurs supérieures des rectangles de l'histogramme des effectifs. Tracer cette ligne.

Exercice 9 : Droite de régression

Une entreprise vend des lots de circuits électroniques. Le tableau suivant indique le pourcentage y de circuits d'un lot qui ont une panne au cours de x semestres d'utilisation :

Nombre x_i de semestres	1	2	3	4	5	6	7	8	9	10
Pourcentage y_i	0	2	4	8	11	14	17	20	23	27

1. Représenter le nuage de points correspondant à cette série statistique.
2. Déterminer l'équation réduite de la droite de régression de y en x. Représenter cette droite.
3. En supposant que la tendance observée se poursuive, estimer le pourcentage$y$ de circuits d'un lot qui ont une panne au cours de douze semestres d'utilisation.

Remarque : Dans cet exemple, x en fonction de y n'a pas de relation concrète. On ne cherchera donc pas la droite de régression de x en y, bien que le calcul soit possible.

Exercice 10 : Ajustement affine par la méthode des moindres carrés

Dans cet exercice, tous les résultats numériques seront donnés par leu valeur décimale approchée à 10^{-3} près, obtenu directement avec une calculatrice.

L'étude, durant les cinq dernières années, du nombre de passagers transportés annuellement sur une ligne aérienne a

conduit au tableau suivant :

Année	Rang x_i de l'année	Nombre p_i de passagers
1992	1	7 550
1993	2	9 235
1994	3	10 741
1995	4	12 837
1996	5	15 655

1. On pose $y_i = \ln p_i$ où ln désigne le logarithme népérien.

a) Compléter, après l'avoir reproduit, le tableau suivant :

x_i	1	...
y_i	8,929	...

b) Représenter le nuage de points $M_i(x_i, y_i)$ dans un repère orthogonal du plan. Peut-on envisager un ajustement affine de ce nuage.

2. *a*) Déterminer, par la méthode des moindres carrés, une équation de la droite de régression D de x en y.

b) Déterminer le coefficient de corrélation r entre les deux variables y et x. Le résultat obtenu confirme-t-il l'observation faite au **1.** *b*) ?

c) Déduire du *a*) une expression de p en fonction de x.

d) En admettant que l'évolution constatée se poursuive les années suivantes, utiliser la relation obtenue au *c*) pour estimer le nombre de passagers transportés en 1998.

Exercice 11 : Méthode des moindres carrés : relation taille $\longleftrightarrow$ pointure

Sur un échantillon de vingt individus $x_1, x_2, \ldots, x_n$, appartenant à une même tranche d'âge, on a étudié les caractères *taille* t_i en mètres et *pointure des chaussures* p_i.

Les résultats obtenus sont les suivants :

$$\sum_{i=1}^{20} t_i = 34,28 \qquad \sum_{i=1}^{20} p_i = 848 \qquad \sum_{i=1}^{20} t_i^2 = 58,8614 \qquad \sum_{i=1}^{20} p_i^2 = 35\,996 \qquad \sum_{i=1}^{20} t_i \cdot p_i = 1\,445,18.$$

Dans ce qui suit, tous les résultats numériques seront données à 10^{-2} près.

1. Calculer le coefficient de corrélation linéaire de la série statistique en les variables t et p. Que peut-on en déduire ?

2. Déterminer par la méthode des moindres carrés la droite de régression de p en t permettant d'estimer la pointure d'un individu en fonction de sa taille.

3. Quelle pointure peut-on estimer pour un individu mesurant 1,83 m ?

Exercice 12 : Méthode des moindres carrés : charge de rupture d'un acier

Le tableau suivant donne les résultats obtenus à partir de 10 essais de laboratoire concernant la charge de rupture d'un en fonction de sa teneur en carbone.

Teneur en carbone x_i	70	60	68	64	66	64	62	70	74	62
Charge de rupture y_i (en kg)	87	71	79	74	79	80	75	86	95	70

1. Représenter graphiquement le nuage de points (x_i, y_i). On prendra 1 cm (ou 1 grand carreau) en abscisse pour une unité, en représentant les abscisses à partir de la valeur 60. En ordonnée, on prendra 1 cm (ou 1 grand carreau) pour 2 kg, en représentant les ordonnées à partir de 70.

2. Calculer les coordonnées du point moyen de ce nuage.

3. Déterminer la valeur approchée à 10^{-3} près du coefficient de corrélation linéaire de la série statistique de variables x et y. Interpréter le résultat.

4. *a*) Déterminer une équation de la forme $y = ax + b$ de la droite D de régression de y en x par la méthode des moindres carrés. On donnera des valeurs approchées des coefficients a et b à 10^{-3} près.

b) Tracer la droite D sur le graphique.

5. Un acier a une teneur en carbone de 77. Donner une estimation de sa charge de rupture.

Exercice 13 : Méthode des moindres carrés : Coefficient de diffusion

Lorsque l'on maintient en contact deux blocs de métal à haute température, les deux blocs se soudent au bout d'un certain temps, des atomes d'un bloc s'étant déplacés sur l'autre et réciproquement : on dit alors qu'il y a *diffusion*.

Le but de ce problème est d'étudier la variation du coefficient de diffusion D (exprimé en cm.s^{-1}) en fonction de la température T (exprimée en degrés Kelvin).

On étudie expérimentalement la diffusion de l'or irradié 198 dans l'or stable.

On pose

$$X = \frac{10^3}{T} \qquad \text{et} \qquad Y = -\log D$$

où log désigne le logarithme décimal (c'est à dire que $\log D = \frac{\ln D}{\ln 10}$ où ln désigne le logarithme népérien).

On obtient expérimentalement le tableau suivant :

X_i	0,8	0,9	1	1,1
Y_i	8,31	9,25	10,16	11,06

1. Représenter graphiquement le nuage de points (X_i, Y_i).

2. Calculer les coordonnées du point moyen de ce nuage.

3. Déterminer la valeur approchée à 10^{-3} près du coefficient de corrélation linéaire de la série statistique de variables X et Y.

4. *a*) Déterminer l'équation réduite de D, la droite de régression de Y en X par la méthode des moindres carrés. Les coefficients seront donnés à 10^{-2} près.

b) Tracer la droite D sur le graphique.

5. *a*) Déduire de la question précédente l'existence de deux réels strictement positifs α et β, que l'on déterminera avec deux chiffres significatifs, tels que

$$D = \alpha e^{-\beta/T}.$$

b) Dresser le tableau des valeurs de D (avec deux chiffres significatifs) associés aux valeurs de T suivantes : 900, 1 000, 1 100, 1 200 et 1 300.