Chapitre 7

Statistique descriptive

7.1 Série statistique à une variable

7.1.1 Une variable discrète

Un exemple

La *population* étudiée est un groupe de 50 élèves. Voici les notes obtenues à un contrôle de Mathématiques :

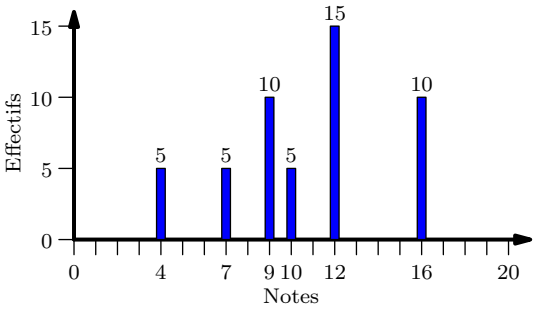
12 ; 12 ; 12 ; 12 ; 12 ; 4 ; 4 ; 4 ; 4 ; 4 ; 16 ; 16 ; 16 ; 16 ; 16 ; 16 ; 16 ; 16 ; 16 ; 16 ; 10 ; 10 ; 10 ; 10 ; 10 ; 7 ; 7 ; 7 ; 7 ; 7 ; 12 ; 12 ; 12 ; 12 ; 12 ; 9 ; 9 ; 9 ; 9 ; 9 ; 9 ; 9 ; 9 ; 9 ; 12 ; 12 ; 12 ; 12 ; 12.

Ceci s'appelle une *série statistique*. On dit que l'on étudie la *variable* « note ».

On peut ranger ces données dans un tableau :

note	4	7	9	10	12	16
effectif						
fréquence						

On représente ceci sous la forme d'un diagramme en bâtons.



Pour décrire la série des notes, on peut calculer sa *moyenne* :

$$\overline{x} = \frac{5 \times 4 + 5 \times 7 + 10 \times 9 + 5 \times 10 + 15 \times 12 + 10 \times 16}{5 + 5 + 10 + 5 + 15 + 10} = \text{——} =$$

La *médiane* de cette série est : la moitié des données est inférieure (ou égale) à et la moitié supérieure (ou égale) à .

Voici les notes obtenues par un deuxième groupe de 50 élèves au même devoir : 10 ; 10 ; 10 ; 10 ; 10 ; 12 ; 12 ; 12 ; 12 ; 12 ; 11 ; 11 ; 11 ; 11 ; 11 ; 11 ; 11 ; 11 ; 11 ; 11 ; 11 ; 11 ; 11 ; 11 ; 11 ; 11 ; 11 ; 11 ; 11 ; 11 ; 11 ; 10 ; 10 ; 10 ; 10 ; 10 ; 10 ; 10 ; 10 ; 10 ; 10 ; 10 ; 10 ; 10 ; 10 ; 10.

note	10	11	12
effectif			

La médiane est . La moyenne est la même :

$$\overline{x} = \text{————————————————————} =$$

Les deux moyennes sont égales. Cependant, la répartition des notes n'est pas la même. Pour le deuxième groupe, les notes sont plus regroupées autour de la moyenne.

Pour faire apparaître cette différence, on calcule l'*écart-type* σ (ou σ_n) des séries statistiques. On doit seulement savoir le trouver à l'aide de la calculatrice. Il n'est pas interdit de savoir que σ^2, appelé la *variance* V, est *la moyenne des carrés des écarts à la moyenne.*

Pour la première série :

$$\sigma^2 = \frac{1}{50}(5 \times (4-10{,}7)^2 + 5 \times (7-10{,}7)^2 + 10 \times (9-10{,}7)^2 + \\ + 5 \times (10-10{,}7)^2 + 15 \times (12-10{,}7)^2 + 10 \times (16-10{,}7)^2) = \frac{630{,}5}{50}$$

Ainsi ; $\sigma^2 = 12{,}61$ et $\sigma \simeq 3{,}55$

Deuxième série : la calculette donne : $\sigma^2 = 0{,}41$ et $\sigma \simeq 0{,}64$.

L'écart-type de la seconde série est plus faible : les notes sont moins dispersées autour de la moyenne. On dit que *l'écart-type est une caractéristique de dispersion. La moyenne est une caractéristique de position.*

Cas général

Voici une série de données x_1, x_2, …, x_n présentée sous forme de tableau (les x_i sont des nombres réels et les effectifs n_i sont des entiers naturels).

Valeurs de la variable	x_1	x_2	…	x_i	…	x_p
Effectifs	n_1	n_2	…	n_i	…	n_p

On peut calculer les fréquences

$$\boxed{f_i = \frac{n_i}{n} = \frac{\text{effectif}}{\text{effectif total}}} \quad \text{où} \quad n = n_1 + n_2 + \cdots + n_p$$

Moyenne :

$$\boxed{\overline{x} = \frac{n_1x_1 + n_2x_2 + \cdots + n_ix_i + \cdots + n_px_p}{n_1 + n_2 + \cdots + n_i + \cdots + n_p}}$$

Variance :

$$\boxed{V = \sigma^2 = \frac{1}{n}[n_1(x_1-\overline{x})^2 + \cdots + n_i(x_i-\overline{x})^2 + \cdots + n_p(x_p-\overline{x})^2]}$$

Écart-type :

$$\boxed{\sigma = \sqrt{V}}$$

Utilisation du symbole de sommation Σ :

$$n = \sum_{i=1}^{p} n_i \qquad \overline{x} = \frac{1}{n}\sum_{i=1}^{p} n_ix_i$$

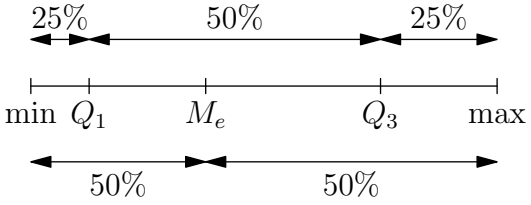
$$V = \sigma^2 = \frac{1}{n}\sum_{i=1}^{p} n_i(x_i-\overline{x})^2$$

Autres indicateurs

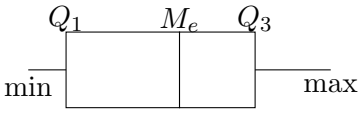
- Le *mode* est la valeur de la variable pour laquelle l'effectif est maximal.
- *Médiane* : La moitié (au moins) des valeurs est inférieure à la médiane M_e et la moitié (au moins) des valeurs est supérieure.
- Le *premier quartile* : c'est la plus petite donnée Q_1 telle qu'au moins 25 % des données soient inférieures ou égales à Q_1.

  Pour la première série de notes, c'est ; pour la deuxième, c'est .
- Le *troisième quartile* : c'est la plus petite donnée Q_3 telle qu'au moins 75 % des données soient inférieures ou égales à Q_3.

  Pour la première série de notes, c'est ; pour la deuxième, c'est .
- L'*écart interquartile* est le nombre $Q_3 - Q_1$.



On voit parfois des boîtes à moustaches :



- De même, les déciles partagent une série en 10 parties.
- *Étendue* : C'est la différence entre les valeurs extrêmes de la série.

7.1.2 Regroupement en classes

On traite ce cas comme une série à variable discrète en utilisant pour les calculs les centres de classes.

Exemple : le tableau suivant donne la répartition des prix de postes auto-radios relevés dans un catalogue. Effectuons une étude statistique complète de cette série.

Prix (euros)	[0 ;800[	[800 ;1200[	[1200 ;1600[	[1600 ;2000[	plus de 2000
Effectif	8	14	10	6	2

L'*histogramme* donne une représentation de la série. L'aire des rectangle est proportionnelle aux effectifs.

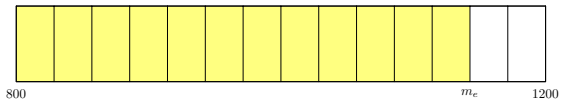
Calculons quelques valeurs caractéristiques (en prenant [2000 ;2400[pour la classe « plus de 2000 »).

- Moyenne =
- Variance =
- Écart-type =
- Mode : la classe [800 ;1200[est la classe modale. Le mode est 1000.

- Médiane : la médiane correspond à l'effectif 20. Elle se trouve entre 800 et 1200.

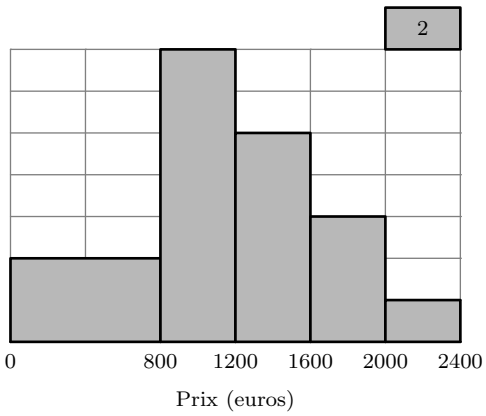
  On peut en calculer une valeur approchée, en supposant que les effectifs sont régulièrement répartis dans la classe où elle se trouve. Les valeurs 800 et 1200 correspondent respectivement aux effectifs 8 et 22, donc :

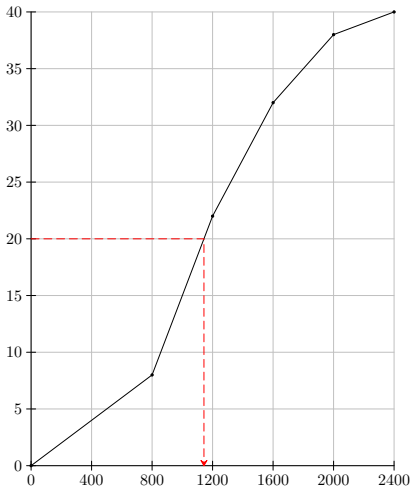
$$m_e = 800 + 400 \times \frac{20-8}{22-8} = 800 + 400 \times \frac{12}{14} \simeq 1143$$



  On peut aussi déterminer cette médiane graphiquement sur le polygone des effectifs cumulés (dé)croissants.

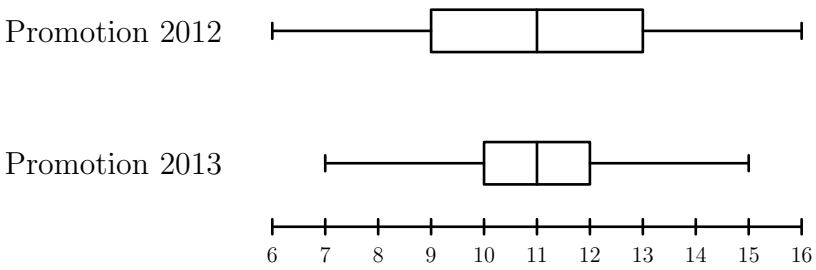
Prix (euros)	[0 ;800[	[800 ;1200[	[1200 ;1600[	[1600 ;2000[	plus de 2000
Effectif	8	14	10	6	2
Effectif cumulés ↑	8	22	32	38	40





Exercice 7.1. La directrice d'une école de journalisme a relevé les notes de français dans les dossiers d'admission des 33 élèves de chacune des promotions 2012 et 2013.

On donne les diagrammes en boîte des deux séries de notes.



1. Pour la promotion 2012, peut-on dire qu'environ 50% des élèves avaient une note de français comprise entre 9 et 13 ?
2. Comparer ces deux séries.

Exercice 7.2. On a relevé le nombre de buts marqués lors de chacun des dix matchs de football de la 19ème journée de championnat de ligue 1 de la saison 2012/2013.

Nombre de buts	0	1	2	3	4	5	6
Nombre de matchs	1	1	1	3	1	1	2

1. Déterminer le nombre moyen de buts marqués par match lors de cette journée de championnat.
2. Déterminer l'écart-type de cette série statistique.

Exercice 7.3. On donne dans le tableau ci-dessous les salaires mensuels nets, en milliers d'euros, des employés de deux PME comptant chacune dix salariés.

PME A	1,2	1,2	1,2	1,3	1,4	1,8	2,1	2,2	4	6,6
PME B	1,4	1,5	1,6	1,65	1,9	2,1	2,25	2,7	3	4,9

1. Déterminer le salaire moyen mensuel net pour chacune de ces PME. Que constate-t-on ?
2. Calculer l'écart-type de chacune des ces deux séries.
3. Comparer la répartition des salaires dans ces deux entreprises.

Exercice 7.4. On donne ci-dessous le nombre de demandes de cartes nationales d'identité traitées par jour par les services d'une commune du nord de la France durant le mois d'avril 2013.

Nombre de demandes	3	4	5	6	7	8	9	10	11	12	17	20
Effectifs (nombre de jours)	1	1	4	2	2	2	1	2	4	1	1	1

1. Déterminer la médiane et les quartiles, puis construire le diagramme en boîte de cette série.
2. Peut-on dire que le nombre de demandes traitées journellement dépasse 11 pendant la moitié des jours du mois d'avril ?

Exercice 7.5. Voici la répartition des ménages des secteurs HLM et libre selon le montant de leur loyer en janvier 2003.

Loyer (en € /mois)	Secteur HLM (en %)	Secteur libre (en %)
moins de 75	0,3	0,2
de 75 à 150	3,7	1,7
de 150 à 230	30,5	7,3
de 230 à 300	35,1	13,4
de 300 à 380	19,6	18,3
de 380 à 430	5,6	14
de 430 à 600	4,7	28,7
de 600 à 760	0,3	9,2
plus de 760	0,1	7,2

1. Construire l'histogramme du secteur libre.
2. Dans quelle classe se trouve le loyer médian pour chaque série ?
3. Calculer le loyer mensuel moyen dans le secteur HLM et dans le secteur libre.
4. Le secteur HLM représente 43,5% du secteur locatif. Calculer le loyer mensuel en France en 2003

(On prendra pour dernière classe la classe $[760; 920[$.)

Exercice 7.6. Un club de plongée compte 80 licenciés. Le tableau donne la fréquence des plongées effectuées par plongeur et par an.

Nombre de plongées	$[0;10[$	$[10;20[$	$[20;30[$	$[30;40[$	$[40;50[$	$[50;60[$
Fréquence	0,1	0,2	0,3	0,175	0,125	0,1

1. Quel est le nombre moyen de plongées effectuées par plongeur ?
2. Donner une estimation de la médiane de cette série.
3. L'année suivante, l'effectif du club est de 70 adhérents et le nombre moyen de plongées effectuées est 32. Quel est le nombre moyen de plongées par plongeur sur cette période de deux ans ?

7.2 Série statistique à deux variables

Dans ce paragraphe, on étudie deux variables numériques x et y dans une population de n individus. A chaque individu i ($1 \leqslant i \leqslant n$), correspond un couple $(x_i\,;\,y_i)$.

Nuage de points

Dans un repère du plan, les points $M_i(x_i;y_i)$ forment le nuage de points.

Point moyen

C'est le point de coordonnées $\boxed{(\overline{x};\overline{y})}$, avec :

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \quad \text{et} \quad \overline{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}$$

Ajustements affines

Un nuage de points étant donné, on peut chercher une droite passant « le plus près possible » de tous les points.

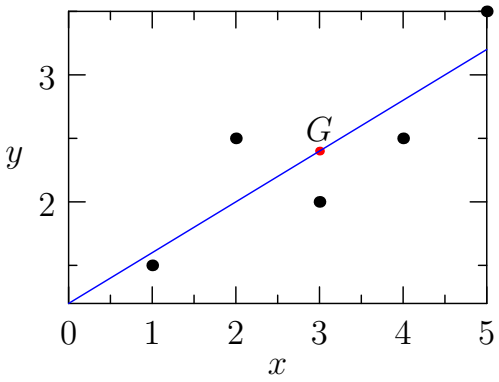
— La méthode au jugé consiste à tracer la droite à la main « au mieux ».[1]

— Méthode des moindres carrés : on cherche à minimiser la somme des carrés des écarts entre les ordonnées des points M_i et les points de la droite de même abscisse.
On démontre qu'il existe une unique droite d rendant minimale cette somme. Son équation est $y = ax + b$. a et b sont donnés par la calculette.
Cette droite passe par le point moyen. On l'appelle la **droite de régression**[2] de y en x.

Exemple

x	1	2	3	4	5
y	1,5	2,5	2	2,5	3,5

Le nuage de points est formé des points $M_1(1;1{,}5)$, $M_2(2;2{,}5)$, $M_3(3;2)$, $M_4(4;2{,}5)$, $M_5(5;3{,}5)$.



Le point moyen est $G(3;2{,}4)$.

La calculette nous dit que le droite d de régression de y en x a pour équation :

$$y = 0{,}4x + 1{,}2$$

G est bien sur d : $0{,}4 \times 3 + 1{,}2 = 2{,}4$.

1. Dans la méthode de Mayer, on coupe le nuage en deux parties contentant le même nombre de points (à un près). La droite d'ajustement passe par les points moyens des deux parties.

2. Francis Galton (1822-1911) a étudié les liens entre la taille y_i d'un individu et celle x_i de son père. La droite qu'il trace pour ajuster le nuage des points de coordonnées $(x_i\,;y_i)$ a un coefficient directeur positif et inférieur à 1. Il semble donc que les pères de grande taille ont des enfants de grande taille, mais en général inférieure à celle du père. Il y a régression du caractère « taille élevée » dans l'espèce, d'où l'expression « droite de régression ».

Corrélation

On calcule parfois le **coefficient de corrélation linéaire** r.
r est compris entre -1 et 1.

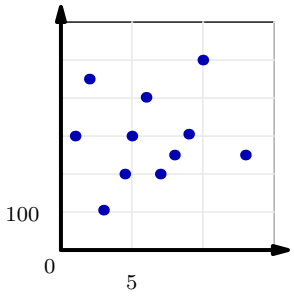
— Si $r = 1$ ou si $r = -1$, les points du nuage sont alignés.
— Si r est proche de 1 ou -1, on dit qu'il y a une bonne corrélation (et c'est tout). Il est raisonnable de déterminer un ajustement affine du nuage.
— Sinon, déterminer un ajustement affine est inutile : il faut alors se tourner vers des ajustements plus compliqués : à l'aide de polynômes, de fonctions exponentielles ...

Exemple : pour le nuage ci-dessus, $r \simeq 0{,}853$.

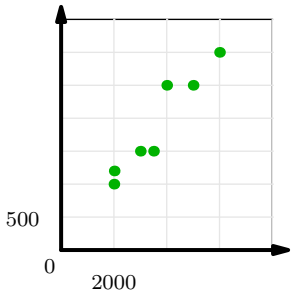
Exercice 7.7. Pour chaque question, une seule réponse est correcte.

1. Parmi les trois nuages de points suivants, indiquer celui pour lequel un ajustement affine semble judicieux.

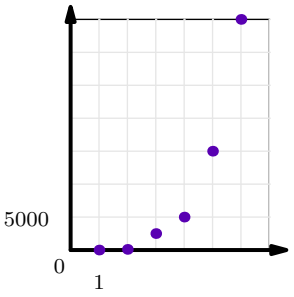
a .



b .

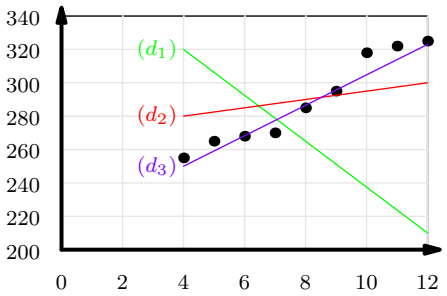


c .



2. Parmi les droites suivantes, quelle est celle qui réalise le meilleur ajustement affine du nuage ci-dessous ?

a. la droite (d_1) ? b. la droite (d_2) ? c. la droite (d_3) ?



3. Un particulier décide de changer, d'ici deux ou trois ans, son véhicule acheté en 2006. Souhaitant connaître le prix auquel il pourra le revendre, il consulte la cote argus de son véhicule et obtient le tableau suivant :

Année	2007	2008	2009	2010	2011	2012
Rang de l'année x_i	1	2	3	4	5	6
Cote y_i (en euros)	16 000	13 500	11 200	9 000	7 400	5 900

On précise que la cote argus est la valeur de revente du véhicule en fonction de l'année choisie pour la revente.

Pour estimer le prix de sa voiture en 2014, il procède à un ajustement affine par la méthode des moindres carrés à l'aide d'une calculatrice.

Après avoir arrondi les coefficients à la centaine la plus proche, une équation de la droite d'ajustement de y en x est

a. $y = -2100x + 17600$ b. $y = -2000x + 17600$ c. $y = -2100x + 17000$

4. L'estimation du prix du véhicule de la question 3 en 2014, selon le modèle précédent est :
   a. 1600 € b. 800 € c. 200 €

Exercice 7.8. Voici des données concernant 16 pays.[3]

Pays	Recettes publiques (en % du PIB 2001)	Taux de pauvres (en %)
Japon	29	11.6
USA	31	17
Australie	33	14.1
Espagne	37	10
Royaume-Uni	39	12.2
Canada	40	12.8
Pays-Bas	42	8
Allemagne	43	7.5
Italie	44	14
Belgique	47	8
France	47.5	8
Finlnde	49	5.5
Autriche	50	10.5
Danemark	53.5	9.2
Norvège	56	7
Suède	57	6.5

Sans représenter le nuage de points associé à ces données, dire si un ajustement affine semble pertinent. Si oui, déterminer les coordonnées du point moyen et l'équation de la droite de régression de y (les taux de pauvres) en x (les recettes publiques).

Exercice 7.9. Le tableau suivant donne l'évolution du nombre d'intérimaires travaillant dans une entreprise créée en 1978.

Année	1978	1983	1988	1993	1998	2003	2008	2013
Rang de l'année x_i	1	2	3	4	5	6	7	8
Nombre y_i d'intérimaires	15	30	55	80	105	130	165	180

1. Dans un repère orthogonal, représenter le nuage de points de coordonnées $(x_i; y_i)$ associé aux données du tableau. On choisira sur l'axe des abscisses 2 cm pour une unité et sur l'axe des ordonnées 1 cm pour 10 intérimaires.
2. Déterminer les coordonnées du point moyen G du nuage, puis placer G sur le graphique.
3. Soit (d) la droite passant par G et par le point $A(1\ ;\ 7{,}5)$. Tracer la droite (d) sur le graphique précédent.
4. Montrer que (d) a pour équation $y = 25x - 17{,}5$.
5. On prend la droite (d) comme droite d'ajustement du nuage.
   (a) Calculer à l'aide de l'équation de (d) une estimation du nombre prévsible d'intérimaires dans cette entreprise en 2018.
   (b) Utiliser le graphique pour retrouver le résultat précédent.

Exercice 7.10. Le tableau ci-dessous liste les classements de salaires et de stress pour des emplois sélectionnés aléatoirement. Le rang 1 pour les salaires correspond au salaire le plus bas et le rang 1 pour le stress correspond au stress le plus faible On cherche à savoir s'il y a une corrélation entre le salaire et le stress.

3. Source : International Centre for Prison Studies, Kings College, Londres. Ce sont des valeurs approchées lues sur un graphique.

Emploi	Rang du salaire	Rang du stress
Agent de change	9	9
Zoologiste	5	4
Ingénieur en électricité	8	5
CPE	6	7
Gérant d'hôtel	4	6
Employé de banque	1	3
Inspecteur de la sécurité	2	2
Economiste	3	1
Psychologue	7	8
Pilote de l'air	10	11
Trader à Wall Street	11	10

Exercice 7.11. Le tableau suivant donne l'évolution du prix d'un article entre le 1er janvier 2004 et le 1er janvier 2013.

Année	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Rang de l'année x_i	0	1	2	3	4	5	6	7	8	9
Prix y_i (en euros)	72	79	85	88	97	106	119	132	144	153

1. À l'aide de la calculatrice, déterminer une équation de la droite Δ d'ajustement de y en x par la méthode des moindres carrés (on aarrondira les coefficients au centième).
2. En utilisant une équation de Δ, estimer le prix de cet article au 1er janvier 2015 et estimer l'année au cours laquelle ce prix dépassera 200 euros.

Exercice 7.12. Afin d'étudier la relation qui pourrait exister entre l'âge et la pression sanguine, un médecin mesure sur 12 femmes d'âges (x) différents la pression sanguine systolique (y). Calculer le coefficient de corrélation linéaire.

x (ans)	56	42	72	36	63	47	55	49	38	42	68	60
y (mm Hg)	147	125	160	118	149	128	150	145	115	140	152	155

Exercice 7.13. L'INSEE fournit les valeurs du taux d'équipement en micro-ordinateur des ménages français pour la période 2004 à 2011, présentées dans le tableau suivant :

Années	2004	2005	2006	2007	2008	2009	2010	2011
Rang de l'année : x_i	1	2	3	4	5	6	7	8
Taux d'équipement : y_i	44,7	49,6	54,3	58,7	62,8	66,7	69,7	73,2

Le taux est estimé en fin d'année. Par exemple, 44,7 % des ménages français étaient équipés en micro-ordinateur fin 2004.

À partir de ces données, on souhaite effectuer des prévisions sur le taux d'équipement en micro-ordinateur des ménages français.

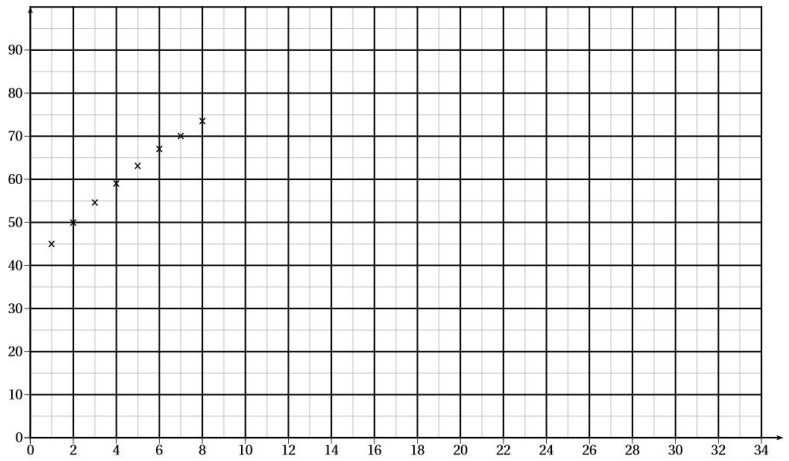
Partie A - Ajustement affine

Un nuage de points représentant la série statistique $(x_i\,;\,y_i)$ est donné en annexe 1.

1. (a) Un ajustement affine vous semble-t-il indiqué sur la période 2004 à 2011 ? Justifier.

   (b) Calculer le coefficient de corrélation linéaire, arrondi au millième, de cette série. Le coefficient calculé confirme-t-il la réponse à la question précédente ? Justifier.
2. Déterminer, à l'aide de la calculatrice, une équation de la droite de régression de y en x par la méthode des moindres carrés (les coefficients seront arrondis au centième). Tracer cette droite **sur l'annexe à rendre avec la copie**.

3. Quel taux d'équipement (arrondi au dixième) a-t-on avec cet ajustement pour 2012 ?
4. D'après cet ajustement, déterminer par le calcul, à partir de quelle année le taux d'équipement dépassera les 85 %.
5. D'après cet ajustement, déterminer à partir de quelle année le taux d'équipement atteindra-t-il les 100 %. Cela vous semble-t-il réaliste ?



Partie B - Ajustement proposé par un tableur

On décide d'utiliser la fonction « courbe de tendance » du tableur. Parmi les courbes proposées, on choisit celle représentant la fonction f définie sur $[1\ ;\ +\infty[$ par :

$$f(x) = -0{,}154x^2 + 5{,}45x + 39{,}36$$

$f(x)$ donne alors une estimation du taux d'équipement pour l'année de rang x. (le rang x est mesuré à partir de l'année 2003 : 2004 est l'année de rang 1).

1. Étude de la fonction f :
   (a) On admet que la fonction f est dérivable sur $[1\ ;\ +\infty[$ et on note f' sa fonction dérivée. Calculer $f'(x)$ et étudier le signe de $f'(x)$ sur $[1\ ;\ +\infty[$
   (b) En déduire le tableau de variation de cette fonction, en précisant la valeur du maximum et la limite de la fonction f en $+\infty$. Préciser le coefficient directeur de la tangente à la courbe représentative de f au point d'abscisse 1.
   (c) Tracer la courbe représentative de f dans le repère de **l'annexe à rendre avec la copie**.
2. À l'aide du graphique, déterminer sur quelle période le taux d'équipement dépassera les 85 %. On laissera apparents les traits de construction permettant de répondre à cette question.
3. (a) Montrer que l'équation $f(x) = 0$ possède une unique solution sur l'intervalle $[1\ ;\ +\infty[$.
   (b) Les prévisions à long terme effectués à l'aide de cet ajustement vous semblent-elles réalistes ?

Partie C - Avec une fonction logistique

On sait par expérience que, pour l'étude des taux d'équipement, une fonction logistique est souvent appropriée. Pour l'équipement en micro-ordinateur des ménages français, on décide d'utiliser la fonction g définie sur $[1\ ;\ +\infty[$ par

$$g(x) = \frac{100}{1 + 1{,}47e^{-0{,}17x}}$$

$g(x)$ donne alors une estimation du taux d'équipement pour l'année de rang x. (le rang x est mesuré à partir de l'année 2003 : 2004 est l'année de rang 1).

1. (a) Sachant que la limite de $e^{-0,17x}$ en $+\infty$ est 0, déterminer la limite de g en $+\infty$
   (b) Interpréter graphiquement ce résultat
   (c) Que peut-on déduire de ce résultat pour le taux d'équipement en micro-ordinateurs des ménages français ?
2. Avec ce dernier modèle, déterminer le taux d'équipement que l'on peut espérer atteindre en 2016.

Exercice 7.14. Le coefficient r indique seulement une relation de dépendance linéaire !

Choisir dix points sur la parabole d'équation $Y = X^2$. Calculer le coefficient de corrélation linéaire r.