Chapitre 12

Séries statistiques à deux variables

12.1 Introduction

On a étudié dans les classes antérieures des séries statistiques à une variable : série de notes, série de performances sportives, série de caractéristiques géographiques, économiques, ... L'objet du chapitre de la classe de terminale est d'étudier si deux variables sont dépendantes l'une de l'autre. On mesure par exemple aux mêmes dates le taux d'ozone dans l'air et le nombre d'entrée aux urgences pour des problèmes respiratoires. Les résultats de ces deux mesures ont-elles un lien l'un avec l'autre ?
Commençons par quelques rappels sur les séries statistiques à une variable...

12.1.1 Paramètre de position : la moyenne

Définition 12.1
La moyenne d'une série statistique est le quotient de la somme de toutes les valeurs de la série (comptées autant de fois que leur effectif) par l'effectif total. En considérant une série statistique de N observations où la variable x prend p valeurs notées x_1, x_2, ..., x_p, chacune ayant un effectif noté n_i, on a :

$$\overline{x} = \frac{\sum_{i=1}^{p} n_i x_i}{N}, \text{ où } \sum_{i=1}^{p} n_i x_i = n_1 x_1 + n_2 x_2 + \cdots + n_p x_p$$

12.1.2 Paramètres de dispersion : variance et écart-type

Définition 12.2
La *variance* est la moyenne des carrés des écarts à la moyenne. C'est un nombre positif.

$$V = \frac{\sum_{i=1}^{p} n_i \left(x_i - \overline{x}\right)^2}{N} = \frac{n_1 \left(x_1 - \overline{x}\right)^2 + \cdots + n_p \left(x_p - \overline{x}\right)^2}{N}$$

L'*écart type* d'une série statistique est égal à la racine carrée de la variance ; on le note σ :

$$\sigma = \sqrt{V} = \sqrt{\frac{n_1 \left(x_1 - \overline{x}\right)^2 + \cdots + n_p \left(x_p - \overline{x}\right)^2}{N}}$$

12.2 Série statistique à deux variables

Une série statistique à deux variables est obtenue en étudiant deux caractères d'une même population. On note x_i et y_i les valeurs de ces deux caractères pour l'individu i de la population.

Dans cette partie, on s'intéressera à des séries statistiques à deux variables x et y prenant chacune p valeurs x_1, x_2, ... x_p et y_1, y_2, ... y_p. On suppose que l'écart-type de chacune de ces deux séries n'est pas nul[1].

On présente généralement une série statistique à deux variables sous la forme d'un tableau :

Valeur de x	x_1	x_2	...	x_p
Valeur de y	y_1	y_2	...	y_p

12.2.1 Nuage de points

Définition 12.3
Si à chaque individu de la population on associe le point A_i de coordonnées $(x_i; y_i)$ dans un même repère, l'ensemble des points obtenus est appelé *le nuage de points* associé à cette série statistique.

Définition 12.4
En notant $\overline{x}$ et $\overline{y}$ les moyennes respectives des séries x et y, le point G de coordonnées $(\overline{x}; \overline{y})$ est appelé *point moyen* du nuage.

Exemple 12.1
On donne dans le tableau suivant la moyenne des températures dans un village canadien de la baie d'Hudson au cours des périodes d'été et la hauteur maximale de neige mesurée au cours des deux premières semaines d'avril.

Année	1968	1969	1970	1971	1972	1973	1974	1975	1976
Temp. ($^\circ$C)	11.8	12.5	13.7	14.0	12.5	14.1	12.1	14.1	14.4
Neige (cm)	30	64	58	81	112	28	91	53	76

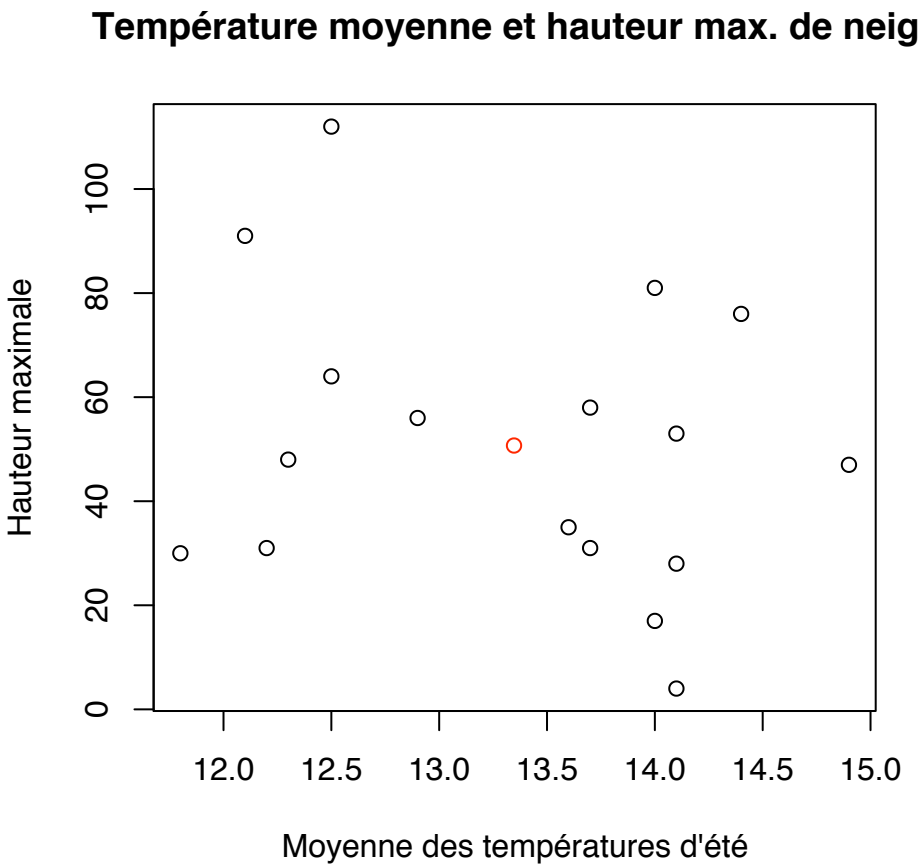
Année	1977	1978	1979	1980	1981	1982	1983	1984
Temp. ($^\circ$C)	12.2	12.3	13.6	14.0	14.9	12.9	13.7	14.1
Neige (cm)	31	48	35	17	47	56	31	4

La moyenne des températures au cours de ces 19 années est de 13,34 $^\circ$C ; la hauteur moyenne de neige est de 50,70 cm.

On construit[2] le nuage de points et le point moyen (en rouge) sur le graphique ci-après :

[1] Cela signifie que les x_i ne sont pas tous égaux et de même pour les y_i.

[2] Le graphique ci-après est obtenu avec le logiciel de statistiques « R » disponible gratuitement à l'adresse http ://www.R-project.org. Un court manuel de prise en main est disponible à l'adresse http ://reymarlioz.free.fr dans la rubrique « pour tous ».



12.2.2 Ajustement linéaire

Définition 12.5
Soit x et y deux séries statistiques de même effectif n. On appelle *covariance* de x et y et on note C_{xy} ou cov(x,y) le réel égal à :

$$C_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

Propriété 12.1
La covariance de deux séries x et y d'effectif n est égale à :

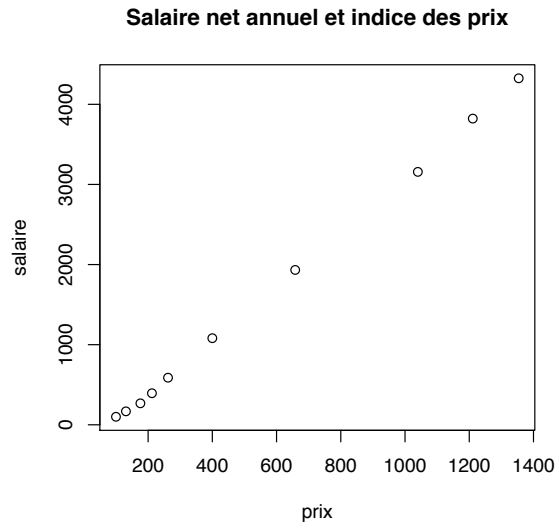
$$C_{xy} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \overline{xy}$$

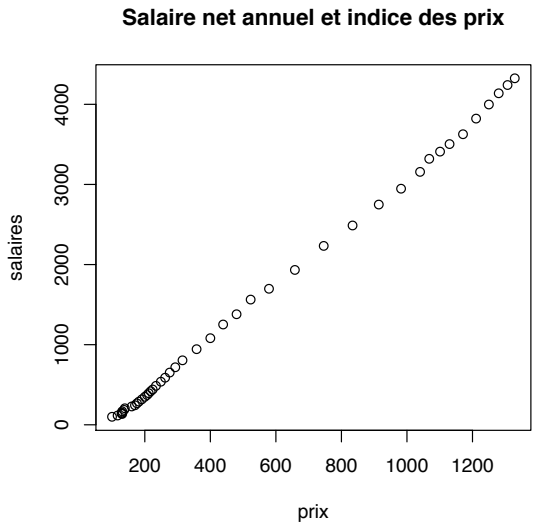
Exemple 12.2
En reprenant les données de l'exemple 12.1, on obtient : cov(températures,hauteurs)$= -5{,}8$.

Exemple 12.3
Dans ce deuxième exemple de série statistique à deux variables, le nuage de points a une forme particulière que nous allons étudier plus précisemment. Dans le tableau ci-après, on donne l'indice des prix et des salaires nets annuels pour quelques années de 1950 à 1994 (Base 100 pour les deux en 1950 - Source : Quid 2001).

Année	1950	1955	1960	1965	1970	1975	1980	1985	1990	1994
Prix	100	131	176	212	262	400	658	1040	1211	1329
Salaire	100	168	268	394	588	1081	1933	3157	3822	4325

On appelle x la série des prix et y la série des salaires. Le nuage de points est tracé ci-dessous (d'abord avec les données du tableau, puis avec toutes les données de 1950 à 1994) :

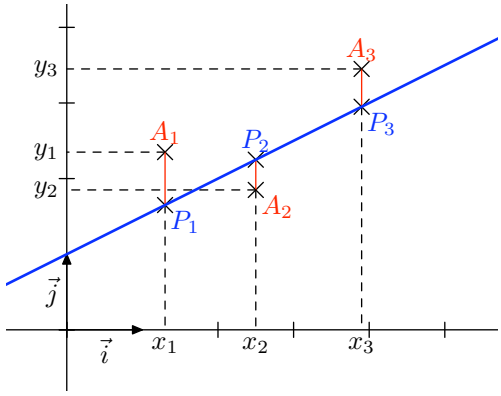




Les points sont presque alignés sur une droite. Nous allons tenter de tracer la droite qui passe « au mieux » parmi ces points.

Les deux théorèmes suivants sont admis :

Théorème 12.1
On considère une série statistique à deux variables x et y et on note A_i le point de coordonnées $(x_i ; y_i)$ pour chaque i, $1 \leq i \leq n$.
Il existe une unique droite D associée à ce nuage de points telle que la somme des « $A_iP_i^2$ » soit minimale. (Où P_i est le point de D de même abscisse que A_i).
Cette droite est appelée *droite des moindres carrés* associée au nuage de points (ou à la série statistique).



Théorème 12.2
On considère une série statistique à deux variables x et y.
– la droite des moindres carrés du nuage de points associé à la série passe par le point moyen ;
– son équation est $y = ax + b$ avec :

$$a = \frac{C_{xy}}{V(x)} \text{ et } b = \overline{y} - a\overline{x}$$

Remarque 12.1
La somme des carrés $A_iP_i^2$ est appelée *somme des carrés des résidus.*

Exemple 12.4
Reprenons les données de l'exemple 12.3.
Dans la calculatrice on rentre dans la `liste 1` la série des prix et dans la `liste 2` la série des salaires.
Pour les Casio : dans le menu `STAT`, sélectionner `SET`, puis pour l'option `2Var XList`, choisir `list1` et pour l'option `2Var YList`, choisir `list2`, et enfin `EXIT`.
En sélectionnant `REG`, puis `x`, on obtient les coefficients a et b de la droite des moindres carrés. ($a \approx 3{,}41$ et $b \approx -300$)
Après avoir appuyé deux fois sur `EXIT`, on sélectionne `2VAR`, on obtient les paramètres de la série : $\sum xy$, $\overline{x}$, $\overline{y}$, n. Ces paramètres permettent de calculer la covariance de la série à l'aide de la formule de la propriété 12.1 : on obtient $C_{xy} \approx 694\,011$. On retrouve a en divisant par $x\sigma n^2 (\approx 450{,}9^2)$: $a \approx 3{,}41$.
Pour les TI : pour obtenir les coefficient a et b, appuyer sur la touche `STAT`, puis dans le menu `CALC`, sélectionner `4:LinReg(ax+b)`. Saisir alors `L1,L2` et `ENTER` (`L1` s'obtient en appuyant sur `2NDE` et `1`).
Pour obtenir les paramètres statistiques de la série, appuyer sur la touche `STAT`, puis dans le menu `CALC`, sélectionner `2-Var Stats`. Saisir alors `L1,L2` et `ENTER`. Les paramètres statistiques s'affichent alors.

12.2.3 Ajustement non linéaire

Parfois, le nuage de points n'a pas la forme d'une droite mais plutôt d'une parabole, d'une fonction inverse, d'une fonction logarithme ou exponentielle, ... Dans ce cas on « transforme » une des deux variables à l'aide d'une fonction pour obtenir un nouveau nuage qui aura la forme d'une droite.

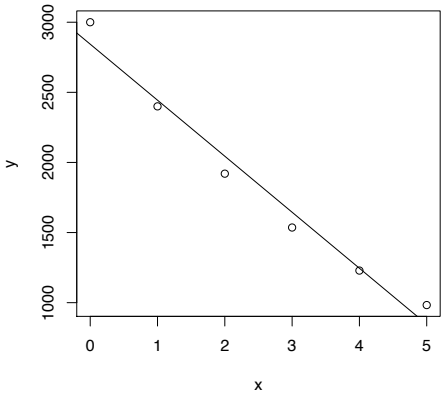
Exemple 12.5 (D'après bac ES. Amérique du Nord 2006)
Une machine est achetée 3 000 €. Le prix de revente y, exprimé en €, est donné en fonction du nombre x d'années d'utilisation dans le tableau suivant :

x_i	0	1	2	3	4	5
y_i	3 000	2 400	1 920	1 536	1 229	983

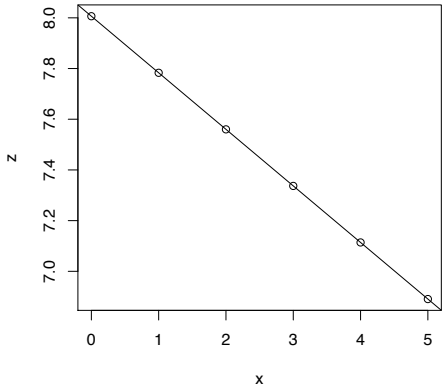
Le nuage de points associé à cette série est représenté ci-après.
À l'aide de la calculatrice on obtient l'équation de la droite d'ajustement : $y = -399x + 2\,843$.
En utilisant cet ajustement, on peut prévoir la valeur à la revente de la machine après 6 ans : $y = -399 \times 6 + 2843 = 449$€.
Cependant, le nuage de point associé ne semble pas s'ajuster correctement à l'aide d'une droite, mais ressemble plutôt à une courbe représentant une fonction exponentielle négative. On pose alors $z = \ln(y)$; on obtient le tableau suivant :

x_i	0	1	2	3	4	5
z_i	8	7,78	7,56	7,34	7,11	6,89

Le nouveau nuage de points est représenté ci-après, et la calculatrice nous donne comme droite de régression $z = -0{,}22x + 8{,}01$. L'estimation du prix de revente après 6 ans peut donc se calculer comme suit : $z = -0{,}22 \times 6 + 8{,}01 = 6{,}69$. On a alors $y = \exp(6{,}69) \approx 804$ €.
En réalité le prix de revente après 6 ans est de 780 €. Le second modèle d'ajustement semble donc plus pertinent.



Nuage des points $A_i(x_i; y_i)$.



Nuage des points $B_i(x_i; \ln(y_i))$.

12.3 Adéquation à une loi équirépartie

Dans les exercices de probabilité, on fait souvent l'hypothèse qu'on se place dans une situation d'*équiprobabilité* : on a un dé *équilibré*, on tire des boules *indiscernables* au toucher, ... Dans cette partie, on va chercher à déterminer si on peut, à la suite de répetitions d'une expérience aléatoire, accepter raisonablement l'hypothèse d'équiprobabilité.

Exemple 12.6
Supposons qu'on possède un dé à six faces numérotées de 1 à 6. Nous allons chercher à vérifier si on peut faire l'hypothèse[3] que ce dé est équilibré.
Pour cela, lançons 200 fois de suite ce dé. On obtient les résultats suivants :

face	1	2	3	4	5	6
effectif	42	43	37	27	30	21
fréquence	0,21	0,215	0,185	0,135	0,15	0,105

Ces fréquences ne sont évidemment pas égales aux fréquences théoriques ($\frac{1}{6}$ si le dé est équilibré), et à chaque nouvelle expérience de 200 lancers on obtient des fréquences différentes : c'est ce qu'on appelle les *fluctuations d'échantillonnage*.
Pour mesurer l'écart entre la distribution de fréquences ontenue et la loi de probabilité « théorique », on calcule le réel noté d^2_{exp} défini par :

$$d^2_{exp} = \sum_{i=1}^{6} \left(f_i - \frac{1}{6}\right)^2 = \left(f_1 - \frac{1}{6}\right)^2 + \left(f_2 - \frac{1}{6}\right)^2 + \left(f_3 - \frac{1}{6}\right)^2 + \left(f_4 - \frac{1}{6}\right)^2 + \left(f_5 - \frac{1}{6}\right)^2 + \left(f_6 - \frac{1}{6}\right)^2$$

On obtient $d^2_{exp} \approx 0{,}009\,63$.
Nous allons ensuite comparer ce « d^2_{exp} » aux résultats qu'on aurait obtenu avec un « vrai » dé équilibré : pour cela, on va lancer 200 fois de suite un dé équilibré[4] et calculer le « d^2 » correspondant.
On répète cette expérience avec le dé équilibré 1 000 fois. On simule donc 1 000 séries de 200 lancers d'un dé équilibré et pour chacune de ces séries de 1 000 lancers, on peut calculer la valeur

[3] On ne pourra jamais l'affirmer avec certitude, mais seulement l'accepter avec une marge d'erreur quantifiable.

[4] En fait, nous allons simuler ces lancers avec le logiciel R déjà évoqué.

correspondante de d^2. Finalement, on a une série de 1 000 valeurs de d^2. On peut déterminer le neuvième décile[5] de cette série ; on obtient $D_9 = 0{,}007\ 78$.
Cela signifie que 10% des valeurs de d^2 du dé équilibré sont supérieures à D_9 et que 90% des valeurs de d^2 du dé équilibré sont inférieures à D_9.
Le d^2_{exp} de départ (celui dont on ne sait pas s'il est équilibré) vaut 0,009 63 ; il est donc supérieur à D_9. On rejette[6] donc l'hypothèse que le dé est équilibré avec un risque de 10%.

Exemple 12.7
Nous allons utiliser le neuvième décile de la simulation précédente pour vérifier l'hypothèse qu'un deuxième dé est équilibré. La série de 200 lancers de ce deuxième dé donne les résultats suivants :

face	1	2	3	4	5	6
effectif	39	39	26	25	30	41
fréquence	0,295	0,295	0,13	0,125	0,15	0,205

On obtient pour ce dé $d^2_{exp} \approx 0{,}006\ 43$. Cette valeur est inférieure à D_9 ; on peut donc accepter avec un risque de 10% l'hypothèse d'équiprobabilité.

[5] C'est-à-dire la valeur pour laquelle 90% des valeurs de la série sont inférieures à ce neuvième décile (voir votre cours de première...).

[6] Attention cela ne signifie pas que le dé est truqué avec certitude !