6

Statistique à deux dimensions

A LA FIN DE CE CHAPITRE, IL FAUT ÃªTRE CAPABLE DE...

1. De caractÃ©riser un nuage de points ;
2. De dÃ©terminer les paramÃ¨tres d'une droite de Mayer ;
3. De dÃ©terminer les paramÃ¨tres d'une droite de rÃ©gression ;
4. De caractÃ©riser la qualitÃ© de la rÃ©gression.

THÉORIE

6.1 Introduction

Nous avons déjà précédemment étudié des éléments de statistique descriptive. Ainsi, lorsqu'on étudie un caractère d'une population donnée, nous sommes capables de :

- représenter les données numériques sur différents types de diagrammes : bâtonnets, histogrammes, quartiers de tartes, ...
- calculer des valeurs typiques (moyenne, mode, médiane) et étudier la manière dont les valeurs observées se situent par rapport à celles-ci (indices de dispersion : étendue, boîtes à moustache, écart-type, ...).

Nous allons maintenant aborder l'étude simultanée de deux caractères d'une même population et rechercher s'il y a une correspondance entre eux.

Exemples:

- existe-t-il un lien entre le fait de fumer et celui de développer un cancer des poumons ?
- existe-t-il un lien entre le poids et la taille des êtres humains adultes ?

Si ce lien existe et peut être déterminé, on pourra s'en servir pour faire des prévisions lorsque la relation qui les lie est significative. Notre travail se fera en plusieurs étapes :

- Nous rechercherons d'abord s'il existe un lien logique entre les variables dont on étudie le comportement.
  Il peut y avoir plusieurs explications au fait que deux séries varient en même temps. Les relations de causalité peuvent se classer en deux catégories :
  — La relation de cause à effet

  Exemple:

  La variation du nombre de journées froides en hiver influence directement celle de la consommation de mazout pendant cette même période.

  — La relation de cause commune

  Exemple:

  Il est peu probable qu'il existe une relation entre la variation de la consommation de mazout de chauffage en hiver et celle de la vente de gants pendant la même période, mais les variations de ces deux caractères sont probablement le résultat de la variation de température à cette saison !

  Il peut arriver aussi que deux caractères semblent varier en même temps mais que ce soit accidentel : on parlera d'"association".
  Dorénavant, avant de nous lancer dans une étude plus approfondie, nous supposerons toujours qu'a été établie l'existence possible d'un lien logique suffisant entre les variables.
- L'analyse de régression a pour but de décrire la nature fonctionnelle de la relation entre ces variables (lorsque cette relation existe, bien sûr !). On pourra alors estimer la valeur d'une des variables à partir de l'autre.

- L'analyse de corrélation dont le but est de mesurer le degré d'association entre les variables.

6.2 Régression linéaire

6.2.1 Nuage de points

Exemple:

On a pris un échantillon de 10 enfants de 9 ans et on a mesuré leur taille et leur poids. Les résultats obtenus sont repris dans le tableau ci-dessous.

Enfants	Taille en cm	Poids en Kg
E_1	134	27,2
E_2	135	26,9
E_3	135	29,3
E_4	137	33,3
E_5	137	32,6
E_6	139	34,5
E_7	142	32,1
E_8	144	36,7
E_9	145	35,2
E_{10}	146	38,2

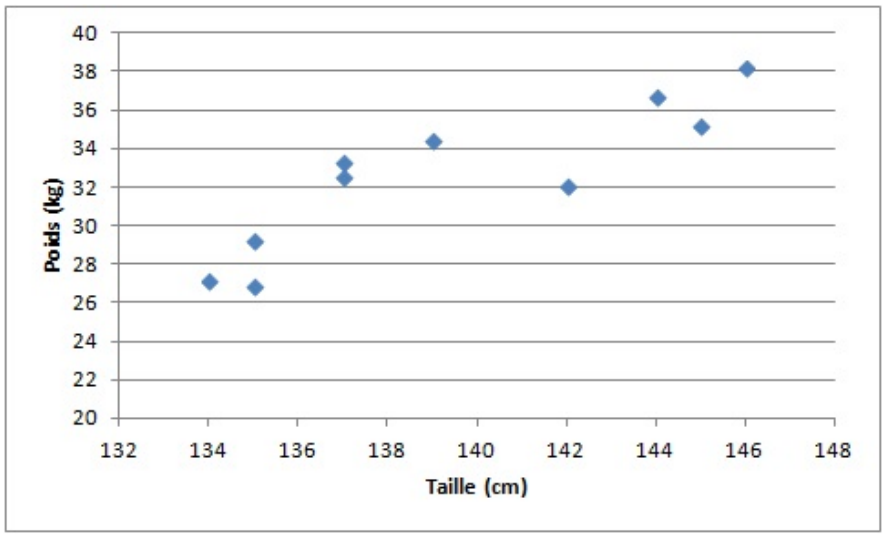
En observant ce tableau, on remarque que la taille et le poids des enfants varient plus ou moins dans le même sens, c'est-à-dire qu'à une taille élevée correspond un poids élevé.
Il nous faut à présent rechercher quelle est la correspondance (la plus exacte possible) entre ces deux caractères pour tenter de répondre à la question suivante : quel poids devrait avoir un enfant de 9 ans mesurant, par exemple, 1 m 40 ?
Attention ! Notre échantillon étant très petit, nos conclusions seront donc peu significatives !
À chaque enfant, nous associons les deux variables Taille et Poids. Elles seront considérées comme les coordonnées d'un point du plan rapporté à un système d'axes orthogonaux. La variable qui est censée exercer une influence sur l'autre est appelée variable indépendante et sera portée graphiquement sur l'axe Ox.
La variable qui doit être estimée est appelée variable dépendante et sera portée graphiquement sur l'axe Oy.
Ainsi, à tout enfant E_i nous avons associé un couple de coordonnées (x_i, y_i), $i = 1, ..., 10$.
Nous pouvons à présent représenter ces données statistiques sur un graphique.

THÉORIE

Définition: *Dans le plan muni d'un repère orthogonal, l'ensemble des points représentant les données de la série statistique est appelé nuage de points ou encore diagramme de dispersion.*

Remarque Nous parlons de repère orthogonal et non de repère orthonormé. En effet, les variables observées étant de natures différentes, il sera le plus souvent impossible de les représenter dans la même unité.
Pour cette même raison, l'intersection des axes ne sera pas nécessairement l'origine du plan.

Voici le nuage de points correspondant à l'exemple :



Un nuage de points n'est pas, en général, exactement le graphique d'une fonction. Le travail du statisticien consiste à rechercher la fonction dont le graphique s'approche au mieux des points du nuage. Cette recherche porte le nom d'ajustement statistique et la courbe celui de courbe d'ajustement. Dans ce chapitre, nous nous limiterons aux cas où les variables sont quantitatives et où le nuage laisse apparaitre une relation linéaire. Nous parlerons donc d'ajustement linéaire ou de régression linéaire.

6.2.2 Point moyen du nuage

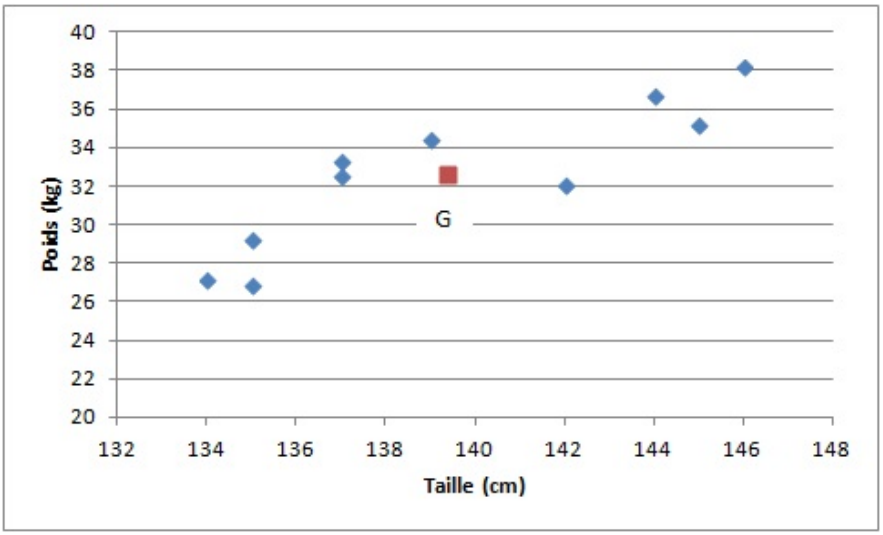
Définition: *Le point moyen du nuage ou barycentre est le point G de coordonnées $(\bar{x}, \bar{y})$ où $\bar{x}$ et $\bar{y}$ sont respectivement les moyennes arithmétiques des valeurs observées x_i et y_i $(i = 1, ...n)$.*

Rappels

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Exemple:

Pour notre exemple : $\bar{x} = 139,4$ et $\bar{y} = 32,6$. Le point moyen du nuage est $G(139,4; 32,6)$.



6.2.3 Ajustement linéaire et droite de régression

Méthode graphique

On peut tracer, au juger, une droite qui nous semble être au plus près des points du nuage. En prenant deux points de cette droite, on obtiendra son équation. L'inconvénient de cette méthode est son peu de précision, mais surtout le fait que des observateurs différents obtiendront des droites et donc des équations différentes.

Méthode des moyennes

Après avoir ordonné les points de la série de manière croissante selon les x_i on divise le nuage en deux parties contenant le même nombre d'éléments (à une unité près si les données sont en nombre impair).
Dans chaque sous-nuage ainsi obtenu, on calcule le point moyen.
La droite cherchée est celle qui passe par ces deux points moyens. On écrit son équation sous la forme $y = ax + b$. On l'appelle droite de MAYER.

Exemple:

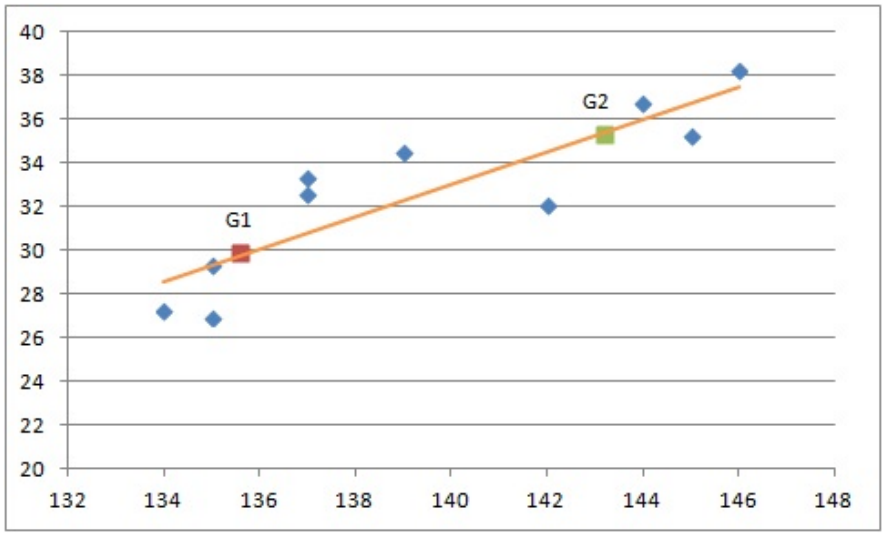
Pour notre exemple, prenons les caractères mesurés sur les cinq premiers enfants pour former le premier sous-nuage. On a :

$$\bar{x} = 135,6 \text{ et } \bar{y} = 29,86$$

Prenons les caractères mesurés sur les cinq derniers pour former le second sous-nuage. On a :

$$\bar{x} = 143,2 \text{ et } \bar{y} = 35,34$$

La droite de MAYER est la droite qui passe par les points (135,6 ; 29,86) et (143,2 ; 35,34).
Elle aura donc pour équation : $y = 0,72x - 67,91$
D'après cette relation, nous pouvons alors estimer qu'un enfant de 9 ans qui mesure 1 m 40 devrait peser 33,03 kg.



Méthode des moindres carrés

Cette méthode est la plus communément utilisée.
Nous recherchons toujours une droite d'équation $y = ax + b$. Pour chaque valeur x_i de la série des données, on peut calculer $(y_d)_i = ax_i + b$.
L'écart $[y_i - (y_d)_i]$ entre la valeur observée et la valeur calculée est appelé résidu.

La droite de régression est celle pour laquelle la somme des carrés des résidus est minimale, c'est-à-dire la droite d'équation $y = ax + b$ telle que

$$\sum_{i=1}^{n} [y_i - (ax_i + b)]^2 = \sum_{i=1}^{n} [y_i - (y_d)_i]^2$$

a la plus petite valeur possible.
On pourrait montrer que l'équation de la droite de régression est

$$y - \bar{y} = a(x - \bar{x})$$

où

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

est le coefficient de régression.

Remarques

- On peut améliorer l'expression de a en se rappelant que la variance de la série des valeurs de x se calcule par

$$V(x) = s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2$$

On peut alors écrire :

$$a = \frac{\sum_{i=1}^{n}(x_i.y_i) - n\bar{x}\bar{y}}{ns_x^2}$$

- On écrira plus souvent l'équation de la droite de régression sous la forme $y = ax + b$.
Dans ce cas, on montre que

$$b = \bar{y} - a\bar{x}$$

- Remarquons encore que la droite de régression passe par le point moyen du nuage . Cela se voit aisément sur la première forme de son équation.

Exemple:

Pour notre exemple :

Enfant	x_i	y_i	$x_i.y_i$	x_i^2
E_1	134	27,2	3644,8	17956
E_2	135	26,9	3631,5	18225
E_3	135	29,3	3955,5	18225
E_4	137	33,3	4562,1	18769
E_5	137	32,6	4466,2	18769
E_6	139	34,5	4795,5	19321
E_7	142	32,1	4558,2	20164
E_8	144	36,7	5284,8	20736
E_9	145	35,2	5104,0	21025
E_{10}	146	38,2	5577,2	21316
Totaux	1394	326	45579,8	194506

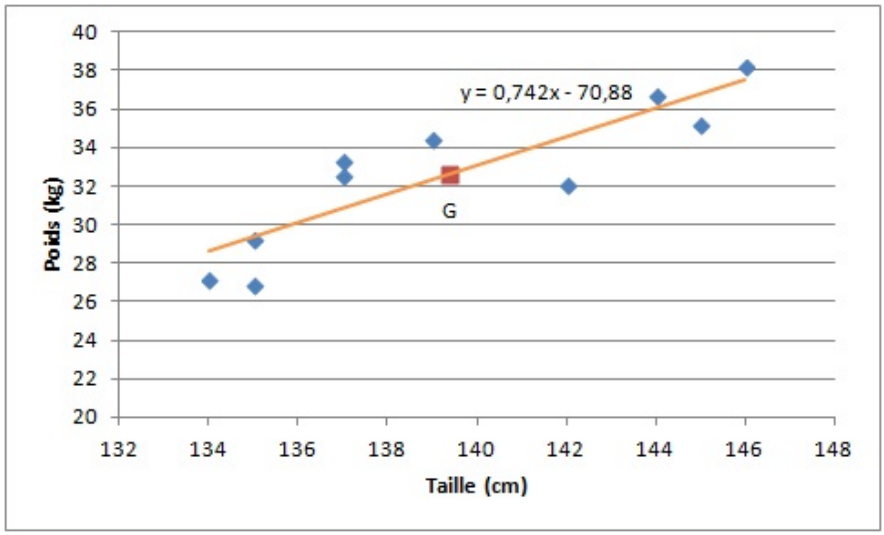
Ainsi $s_x^2 = 18,24$ et $a = 0,742$.
La droite de régression a pour équation :

$$y - 32,6 = 0,742.(x - 139,4)$$

ou

$$y = 0,742x - 70,88$$

Selon cette relation, nous pouvons alors estimer qu'un enfant de 9 ans qui mesure 1 m 40 devrait peser 33,04 kg.



6.3 Qualité de la corrélation

6.3.1 Covariance

Définition: *La covariance entre deux observations quantifient leurs écarts par rapports à leurs moyennes respectives. Elle est donnée par :*

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

Remarque | On peut démontrer que :

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$$

Remarquons qu'avec cette définition de la covariance, le coeficient de regression s'écrit :

$$a = \frac{s_{xy}}{s_x^2}$$

Exemple:

Pour notre exemple :

Enfant	x_i	y_i	$x_i.y_i$
E_1	134	27,2	3644,8
E_2	135	26,9	3631,5
E_3	135	29,3	3955,5
E_4	137	33,3	4562,1
E_5	137	32,6	4466,2
E_6	139	34,5	4795,5
E_7	142	32,1	4558,2
E_8	144	36,7	5284,8
E_9	145	35,2	5104,0
E_{10}	146	38,2	5577,2
Totaux	1394	326	45579,8

Ainsi $s_{xy} = 13,54$

6.3.2 Coefficient de correlation

Nous avons recherché une droite de régression des valeurs de y en fonction des valeurs de x, mais nous aurions pu faire le contraire et rechercher une droite de régression des valeurs de x en fonction de celles de y (pour autant qu'elle ait un sens).
Ces droites comprennent toutes les deux le point moyen du nuage.
Si les deux droites ainsi trouvées coïncident, c'est que la relation entre les variables est une relation fonctionnelle linéaire. Sinon, ces deux droites se coupent au point moyen du nuage, comme des ciseaux. Plus ces "*ciseaux*" sont fermés, plus grande est la dépendance entre les variables.
L'intensité de cette relation est mesurée par le coefficient de corrélation. Une bonne estimation de la dépendance des deux variables est donnée par le rapport :

$$r^2 = \frac{\sum_{i=1}^{n} [(y_d)_i - \bar{y}]^2}{\sum_{i=1}^{n} [y_i - \bar{y}]^2}$$

aussi appelé coefficient de détermination qui est le produit des pentes des deux droites.
Si, pour chaque valeur de x_i le résidu $[y_i - (y_d)_i]$ tend vers 0, c'est que les valeurs observées et les estimations données par la droite de régression sont très proches.
Dans ce cas, l'équation de la droite de régression reflète bien la liaison entre les variables.
Le coefficient de détermination tend vers 1.
Par contre, si les résidus sont grands, le coefficient de détermination tendra vers 0. Reprenons l'expression de r^2. On sait que $(x_i, (y_d)_i)$ est un point de la droite de régression donc

$$(y_d)_i - \bar{y} = a.(xi - \bar{x})$$

Ainsi

$$r^2 = \frac{\sum_{i=1}^{n} a^2.[x_i - \bar{x}]^2}{\sum_{i=1}^{n} [y_i - \bar{y}]^2} = \frac{a^2 \sum_{i=1}^{n} [x_i - \bar{x}]^2}{\sum_{i=1}^{n} [y_i - \bar{y}]^2} = a^2.\frac{s_x^2}{s_y^2}$$

De plus, puisque $a = \frac{s_{xy}}{s_x^2}$, on a :

$$r = \frac{s_{xy}}{s_x.s_y}$$

Définition: *Le coefficient de corrélation est le rapport*

$$r = \frac{s_{xy}}{s_x.s_y}$$

Le coefficient de corrélation est une quantité sans dimension qui varie entre - 1 et 1.
Si r= 0, il y a absence de corrélation entre les variables.
Si r= 1 ou r= - 1, la corrélation est parfaite.
Pour toutes les autres valeurs de r, selon qu'elles sont plutôt proches de 1 (ou de - 1) ou de 0, la corrélation sera forte ou faible.

Exemple:

Pour notre exemple, $s_{xy} = 13,54$, $s_x^2 = 18,24$. Nous pourrions calculer de même $s_y^2 = 13,18$ Dès lors $r = 0,873$.
On peut donc faire confiance à la relation trouvée entre les variables. Notre estimation du poids d'un enfant mesurant 1 m 40 est assez bonne.

Remarques

- Il serait peu légitime de faire des estimations pour des valeurs se situant en-dehors de l'étendue de l'échantillon observé.
  En effet, il est impossible de savoir si les valeurs observées en agrandissant l'étendue de l'échantillon auraient encore collé à l'équation de la droite de régression.
- Pour pouvoir faire des prévisions intéressantes et en tirer des conclusions fiables, on devra prendre un échantillon assez étendu et fourni. Plus le nombre d'observations est grand, plus on pourra faire confiance aux prévisions et aux conclusions.

Pour les séries statistiques suivantes, étudier complètement (graphiquement et par calcul) la corrélation entre les deux séries statistiques. On n'oubliera pas les conclusions.

1. Le tableau suivant donne les cotes aux deux premières interrogation de 10 élèves d'une classe.

Interrogation 1	6	5	8	8	7	6	10	4	9	7
Interrogation 2	8	7	7	10	5	8	10	6	8	6

2. On a mesuré la tension artériel de patient féminin en fonction de leur âge. La répartition est donnée dans le tableau suivant :

Age	56	42	72	36	63	47	55	49	38	52	68	60
Tension	14.7	12.5	16.0	11.8	14.9	12.8	15.0	14.5	11.5	14.0	15.2	15.5

3. Le tableau suivant représente le poids de 12 pères et de leurs fils aînés

Poids du pére	65	63	67	64	68	62	70	66	68	67	69	71
Poids du fils	68	66	68	65	69	66	68	65	71	67	68	70