Thème 1 Statistiques à deux variables

Fiche 1

Statistiques
Activités préparatoires

D'après Bréal ES, 2002.

I. La meilleure droite

Dans le plan muni d'un repère, on considère les trois points $A(2;6)$, $E(7;10)$ et $C(12;17)$ On cherche une droite d'équation $y = ax + b$ qui passe le plus près possible de ces trois points. Plus précisément, on veut que les ordonnées des trois points A', B' et C' de la droite, de mêmes abscisses que les points A, B et C respectivement, soient les plus proches possible des ordonnées des trois points A, B et C. Pour cela, on cherche à rendre le plus petit possible le nombre : $d^2 = AA'^2 + BB'^2 + CC'^2$.

1. Démontrer que $d^2 = (2a+b-6)^2 + (7a+b-10)^2 + (12a+b-17)^2$.
2. Calculer d^2 pour la droite d'équation $y = x+4$, puis pour la droite d'équation $y = 1,1x+3$. Quelle est la meilleure de ces deux droites ?
3. On appelle $\overline{x}$ la moyenne des abscisses de A, B, C et $\overline{y}$ la moyenne de leurs ordonnées. Calculer $\overline{x}$ et $\overline{y}$, puis placer le point G de coordonnées $(\overline{x}\,;\overline{y})$.
4. On se propose de trouver la meilleure droite parmi toutes les droites qui passent par G.
   (a) Démontrer qu'une telle droite a une équation de la forme $y = a(x-7)+11$.
   (b) Exprimer d^2 en fonction de a. Conclure.

II. Prévision

Au cours d'une période d'hiver, l'intendant d'un lycée a relevé, certains jours, la température extérieure moyenne de la journée et la consommation de fioul de la chaudière pendant la même journée. Il a obtenu le tableau suivant :

Température X (en °C)	-6	-5	-3	-2	-1	1	2	3	5	6
Consommation Y (enL)	400	390	360	330	310	290	260	250	200	190

Il voudrait estimer la consommation à prévoir pour une journée à -7 °C. Pour cela, il cherche une formule du type $y = ax+b$ qui exprime de façon approchée le lien entre la température et la consommation. Après avoir placé dans un repère les points de coordonnées $(X;Y)$, déterminer graphiquement une telle formule. Comment peut-on juger la qualité de l'approximation qu'elle fournit ?

Fiche 2

Statistiques

D'après Bréal 2002.

I. Étude de deux variables quantitatives sur une même population

On considère sur une même population deux variables quantitatives X et Y, pour lesquelles on dispose de n observations, résumées dans le tableau ci-dessous :

X	x_1	x_2	x_3	…	…	x_n
Y	y_1	y_2	y_3	…	…	y_n

L'objectif est de savoir s'il y a un lien entre les deux grandeurs X et Y et éventuellement de préciser ce lien.

I.1. Nuage de points

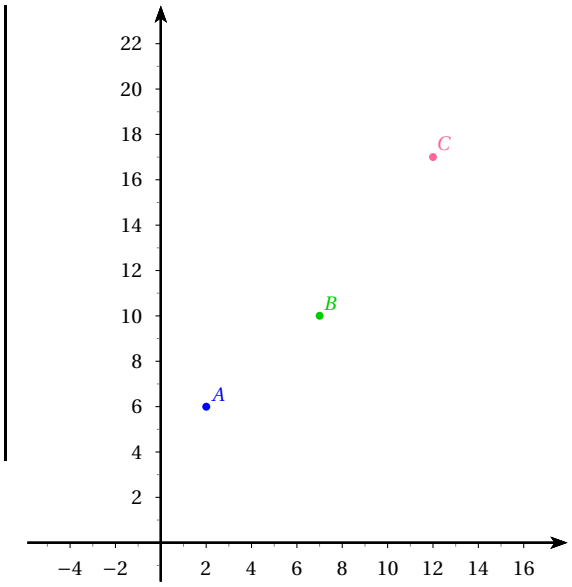
On peut représenter les données en plaçant dans un repère les points :

$$M_1(x_1\,;\,y_1)\,,\ M_2(x_2\,;\,y_2)\,,\ M_3(x_3\,;\,y_3)\,,\ \ldots,M_n(x_n\,;\,y_n).$$

On obtient ce qu'on appelle **un nuage de points**.
La forme de ce nuage donne une information importante : si les points sont proches d'une courbe d'équation $y = f(x)$, cela peut laisser penser que la valeur de X permet d'estimer celle de Y.

Des exemples pour comprendre

1. Nuage de points : $A(2;6)$, $E(7;10)$ et $C(12;17)$. Les points sont à peu près alignés. le nombre de points est trop faible pour que cela puisse avoir une signification statistique. Néanmoins, d'un point de vue géométrique, on peut rechercher une droite qui passe au plus près de ces trois points.



2. Au cours d'une période d'hiver, l'intendant d'un lycée a relevé, certains jours, la température extérieure moyenne de la journée et la consommation de fioul de la chaudière pendant la même journée. Il a obtenu le tableau suivant :

Température X (en °C)	-6	-5	-3	-2	-1	1	2	3	5	6
Consommation Y (enL)	400	390	360	330	310	290	260	250	200	190

Le nuage de points qui représente ces données montre des points à peu près alignés, ce qui laisse penser que la connaissance de X permet d'estimer la valeur de Y.

I.2. Point moyen

Pour avoir une idée de la tendance centrale de X et Y, on peut calculer leurs moyennes.

Définition 1
Étant donné un nuage de points, on appelle **point moyen** le point de coordonnées $(\overline{x}\,;\,\overline{y})$, où $\overline{x}$ et $\overline{y}$) sont les moyennes respectives des abscisses et des ordonnées des points du nuage.

Des exemples pour comprendre

1. Pour l'exemple 1, le point moyen a pour coordonnées $(7,11)$.
2. Dans l'exemple de la consommation de fioul le point moyen a pour coordonnées $(0,298)$.

I.3. Variances et covariance

pour étudier la dispersion de chaque variable X et Y, on peut calculer leurs variances :

$$V(X) = \frac{1}{n}\sum(x_i - \overline{x})^2$$

$$V(Y) = \frac{1}{n}\sum(y_i - \overline{y})^2$$

Mais il est inutile d'introduire une quantité qui fasse intervenir à la fois les valeurs de X et de Y.

Définition 2
On appelle covariance de X et Y le nombre :

$$cov(X,Y) = \frac{1}{n}\sum(x_i - \overline{x})(y_i - \overline{y})$$

Remarque
Considérons le repère dont l'origine est le point moyen et qui a les mêmes vecteurs de base que le repère initial. Dans ce repère, les coordonnées des points du nuage sont données par $x'_i = x_i - \overline{x}$ et $y'_i = y_i - \overline{y}$.
la covariance s'écrit alors

$$cov(X,Y) = \frac{1}{n}\sum x'_i y'_i.$$

Or dans le premier et le troisième quadrant $x'_i y'_i$ est positif, alors que dans le deuxième et le quatrième quadrant $x'_i y'_i$ est négatif.
Par suite :

1. si tous les points du nuage sont dans le premier et le troisième quadrant, la covariance est positive,
2. si tous les points du nuage sont dans le deuxième et le quatrième quadrant, la covariance est négative.

Mais la réciproque est fausse : quel que soit le signe de la covariance, il y a en général des points dans les quatre quadrants.

Des exemples pour comprendre

1. Pour l'exemple 1,
   $V(X) = \frac{50}{3}$, $V(Y) = \frac{62}{3}$ et $cov(X,Y) = \frac{55}{3}$.
2. Dans l'exemple 2,
   $V(X) = 15$, $V(Y) = 4896$ et $cov(X,Y) = -270$.

I.4. Ajustement affine

Lorsque les points du nuage paraissent presque alignés, on peut chercher une relation de la forme $y = ax + b$ qui exprime de façon approchée Y en fonction de X, autrement dit, une fonction affine f telle que l'égalité $y = f(x)$ s'ajuste au mieux avec les données. Graphiquement, cela signifie qu'on cherche **une droite qui passe au plus près de tous les points du nuage**.
Une telle relation permettrait notamment de faire des **prévisions**.
Pour mesurer la qualité d'une telle formule, on considère, pour chaque valeur x_i, la différence entre la valeur observée, c'est à dire y_i, et la valeur calculée par la formule, c'est à dire $ax_i + b$. on souhaite que toutes les différences :

$$y_i - ax_i - b,$$

appelées **erreurs**, ou résidus, ou perturbations, soient les plus petites possible.
La méthode la plus couramment employée dite méthode des moindres carrés, consiste à choisir a et b de façon que la **somme des carrés des résidus soit la plus petite possible**.

Remarque
Faire simplement la somme des résidus ne serait pas satisfaisant, car les erreurs positives et négatives peuvent se compenser même si la droite passe loin de tous les points.

Définition 3
Étant donné un nuage de points $(x_i\ ;\ y_i)$ et une droite d'équation $y = ax + b$, on appelle erreur quadratique totale la somme des carrés des résidus, c'est à dire e nombre :

$$E = (y_1 - ax_1 - b)^2 + (y_2 - ax_2 - b)^2 + \ldots + (y_n - ax_n - b)^2.$$

Des exemples pour comprendre

1. Dans le cas de l'exemple 1, la droite d'équation $y = x + 4$ fournit une erreur quadratique totale égale à $0^2 + 1^2 + 1^2$, c'est à dire 2.
   La droite d'équation $y = 1,1x + 3$ fournit une erreur quadratique totale égale à $0,8^2 + 0,7^2 + 0,8^2$, c'est à dire 1,77. La deuxième droite est donc meilleure que la première.

2. Dans le deuxième exemple, la droite d'équation $y = -15x + 300$ fournit une erreur quadratique totale égale à 1750.
   La droite d'équation $y = -20x + 300$ fournit une erreur quadratique totale égale à 1000. La deuxième droite est donc meilleure que la première.

Théorème 1
Étant donné un nuage de points $(x_i\ ;\ y_i)$, la droite pour laquelle l'erreur quadratique est la plus faible a pour équation $y = ax + b$, avec

$$a = \frac{cov(X,Y)}{V(X)}$$

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

Cette droite est appelée la droite de régression de Y en X.
Elle passe le point moyen. En effet : $\overline{y} = a\overline{x} + b$, donc les coordonnées $\overline{x}$ et $\overline{y}$ du point moyen vérifient l'équation de la droite.
a s'appelle le coefficient de régression. S'il est positif, Y varie dans le même sens que X, s'il est négatif, Y varie dans le sens contraire de X.

Remarques

— La méthode utilisée (ajustement affine par la méthode des moindres carrés) s'appelle aussi la régression linéaire. A proprement parler, celle-ci contient en plus une interprétation probabiliste des résidus. On ne développera pas ici.

— En 1877, le statisticien britannique Francis Galton (1822-1911), étudiant l'évolution de la taille des individus d'une génération à la suivante, a constaté une régression (c'est à dire une diminution) de la dispersion des tailles. Curieusement, le mot est resté attaché à la méthode des moindres carrés, alors que Galton n'avait pas utilisé cette méthode.

Des exemples pour comprendre

1. Dans le cas de l'exemple 1, on obtient pour coefficients $a = \frac{55}{50} = 1,1$ et $b = 11 - 1,1 \times 7 = 3,3$. La droite des moindres carrés a pour équation $y = 1,1x + 3,3$. L'erreur quadratique est alors égale à 1,5.
2. Dans le cas de l'exemple 2, $a = \frac{-270}{15} = -18$ et $b = 298 - 18 \times 0 = 298$. La droite des moindres carrés a pour équation $y = -18x + 298$. L'erreur quadratique est alors égale à 360.

Remarques

1. La formule trouvée n'a d'intérêt que si elle fournit une bonne approximation des données. La qualité de cette approximation peut être évaluée graphiquement : les points du nuage doivent apparaître presque alignés. il faut également que le nombre de points soit suffisant. Il existe des critères permettant de savoir, en fonction du nombre d'observations, si l'erreur quadratique totale est significativement élevé. mais cette étude n'est pas au programme.
2. Le fait que l'on trouve une relation de la forme $y = ax + b$ qui traduise de façon satisfaisante les données ne signifie pas nécessairement un lien de causalité entre X et Y, il peut se faire par exemple que X et Y soient toutes les deux conséquences d'une même cause Z. Ainsi, dans une station balnéaire, les ventes de crème solaire et de boissons fraîches sont fortement reliées, sans que l'on puisse dire pour autant que les unes soient les causes des autres ! Il est plus vraisemblable que l'ensoleillement en est une cause commune.

La formule permet d'obtenir des **prévisions**.

Des exemples pour comprendre

1. Dans le cas de l'exemple 2, pour une journée à 4řC, la consommation en fioul peut être estimée à $y = -18 \times 4 + 298$ soit $226L$ environ. Cette prévision est une **interpolation** car la valeur de x est à l'**intérieur** de l'intervalle dont les bornes sont la valeur minimale et la valeur maximale de x relevées.
2. Dans le cas de l'exemple 2, pour une journée à -7řC, la consommation en fioul peut être estimée à $y = -18 \times (-7) + 298$ soit $424L$ environ. Cette prévision est une **extrapolation** car la valeur de x est à l'**extérieur** de l'intervalle dont les bornes sont la valeur minimale et la valeur maximale de x relevées.

Remarque
Les prévisions obtenues sont à considérer avec prudence, car elles supposent que le lien observé (appelé corrélation) se maintienne au-delà du domaine des observations faites, ce qui n'est pas nécessairement le garanti. Ainsi dans l'exemple 2 (consommation de fioul et température extérieure) des travaux d'isolation peuvent remettre en cause les prévisions. Par ailleurs, la formule trouvée a un domaine de validité limité : dans le même exemple, il serait absurde de l'appliquer à une température de 20řC.

I.5. Un indicateur de la qualité de l'ajustement affine des moindres carrés

La covariance est sensible aux changements d'unités, par exemple si on multiplie par 1000 les valeurs de Y sans changer les valeurs de X, la covariance est multipliée par 1000. Ainsi, la valeur absolue de la covariance ne donne pas d'information sur la qualité de l'ajustement. D'où la définition :

Définition 4
On appelle coefficient de corrélation linéaire de X et Y, que l'on note r , le nombre défini par

$$r = \frac{cov(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}}$$

Comme, $\sqrt{V(X)}$ est l'écart type de X, noté $\sigma(X)$, on peut aussi écrire :

$$r = \frac{cov(X,Y)}{\sigma(X)\sigma(Y)}$$

Propriété 1
Le coefficient de corrélation linéaire est un nombre compris entre -1 et 1.

— si $|r| = 1$ alors la droite de régression passe par tous les points du nuage.

— si $r = 0$ alors il n y a pas de liaison linéaire entre X et Y.

Le coefficient de détermination

Définition 5
On appelle coefficient de détermination de X et Y, que l'on note R^2 , le nombre défini par

$$R^2 = r^2$$

Ce nombre désigne la part en % de la dispersion de la variable Y qui expliquée par la dispersion de la variable X. C'est donc un indicateur de la qualité du modèle. La part complémentaire étant liée aux aléas.

II. EXERCICES

Exercice 1. Représenter le nuage suivant et placer le point moyen :

X	1	3	5	7
Y	3	4	8	9

1. On propose l'ajustement $y = 2x - 2$. tracer la droite correspondante, et calculer l'erreur quadratique totale.
2. Même question avec l'ajustement $y = x + 2$. Cet ajustement affine est -il meilleur que le précédent ?
3. Démontrer que $V(X) = 5$, $V(Y) = 6,5$, $cov(X,Y) = 5,5$.
4. Déterminer et tracer la droite des moindres carrés et calculer l'erreur quadratique totale.

Exercice 2. Une entreprise souhaite mesurer l'impact de ses dépenses publicitaires sur son chiffre d'affaires. Elle dispose des renseignements suivants, observés sur les dix dernières années (en milliers d'euros).

Frais de pub X	1	3	5	7	9	11	13	15	17	19
Y	36	59	83	102	122	149	168	192	213	235

1. Représenter le nuage et placer le point moyen. Un ajustement affine vous paraît-il justifié ?
2. Déterminer (avec un tableur ou une calculatrice) la droite de régression de Y en X. Interpréter le signe de son coefficient directeur.
3. Calculer le coefficient de corrélation linéaire.
4. On envisage un budget publicitaire de 18000 euros. Estimer le chiffre d'affaires correspondant.

Exercice 3.
1. Écrire une fonction Python avec en arguments la liste des valeurs de X, donnant la moyenne de X.
2. Écrire une fonction Python avec en arguments la liste des valeurs de X, donnant la variance de X.
3. Écrire une fonction Python avec en arguments la liste des valeurs de X, et la liste des valeurs de Y, donnant la covariance de X et Y.
4. Écrire une fonction Python avec en arguments la liste des valeurs de X, et la liste des valeurs de y, le coefficient directeur de la droite des moindres carrés.

5. Écrire une fonction Python avec en arguments la liste des valeurs de X, et la liste des valeurs de y, l'ordonnée à l'origine de la droite des moindres carrés.
6. Écrire une fonction Python avec en arguments la liste des valeurs de X, et la liste des valeurs de y, le coefficient de corrélation linéaire de la droite des moindres carrés.

Exercice 4. Voici le tableau donnant la consommation finale d'électricité par secteur en TWh et en %

TWh	1970	1980	1990	2000	2008	2009	2010	2011	2012	2013	2014	2015
Sidérurgie	10,3	12,4	10,5	11,1	11,8	8,8	10,5	11,2	10,6	10,4	10,6	10,4
Industrie	62,1	83,0	105,0	127,4	120,9	108,1	110,5	108,0	107,8	106,9	106,0	105,7
Résid-Tert	41,1	105,3	172,2	236,5	288,8	289,5	302,8	292,3	300,4	305,7	293,9	298,6
Agric	2,6	4,2	5,0	6,0	6,6	7,4	7,6	8,0	8,4	8,8	8,2	8,1
Transp(urb+trains)	5,1	6,0	7,5	9,4	10,4	10,1	10,0	10,0	10,2	10,2	10,0	10,2
TOTAL	121,3	211,0	300,2	390,4	438,5	423,9	441,4	429,4	437,4	442,0	428,8	432,9

On se propose d'analyser si la crise de 2008 a eu un impact sur cette consommation. Proposer une démarche.