

XII. ANALYSE STATISTIQUE DES DÉPENDANCES.

XII. 1. Corrélation de deux variables aléatoires.

Au chapitre **VI** (variables aléatoires), nous avons introduit une grandeur appelée la *covariance* de deux variables aléatoires. Nous avons vu alors que, lorsque deux variables aléatoires X et Y sont stochastiquement indépendantes, leur covariance est nulle, mais que la covariance peut être nulle aussi pour des variables dépendant l'une de l'autre. Les trois tableaux du chapitre **VI** montrent trois possibilités pour la loi conjointe de deux variables aléatoires : (1) elles sont stochastiquement indépendantes ; (2) l'une est fonction de l'autre ; (3) elles ne sont pas stochastiquement indépendantes, et ne sont pas non plus fonction l'une de l'autre. Les trois exemples avaient été choisis de sorte que dans les trois cas, la covariance soit nulle. Ceci afin de bien montrer que la covariance n'est pas une mesure de la dépendance, qui serait d'autant plus grande que la dépendance entre les deux variables serait plus forte (si tel était le cas, la covariance serait nulle pour le tableau 1 seulement ; elle serait maximum pour le tableau 2 et entre les deux pour le tableau 3). En examinant les choses de près, il apparaît que la covariance ne mesure pas la dépendance en général, mais seulement la dépendance *linéaire*. Dans le tableau 2 du chapitre **VI**, Y dépend de X , mais pas linéairement : $Y = X^2$, donc Y est bien fonction de X , mais à une valeur donnée de Y correspondent deux valeurs opposées de X , qui s'annulent mutuellement dans le calcul de la covariance. On peut vérifier que si Y était une fonction linéaire (ou affine) de X , la covariance serait maximum. Si au lieu d'une dépendance linéaire rigoureuse, on avait une dépendance "floue" (mais linéaire), la covariance serait intermédiaire entre zéro et le maximum.

Afin de voir cela en détail, le mieux est d'introduire une troisième variable aléatoire $Z_\lambda = Y - \lambda X$, où λ est un paramètre réel fixé. Si la loi conjointe de X et Y est donnée, elle permet bien sûr de calculer la loi de Z_λ . Si par exemple la variance de Z_λ est nulle, cela signifie que Z_λ est, avec probabilité 1, c'est-à-dire avec certitude, égale à sa moyenne $m = \mathbf{E}(Z_\lambda)$. Mais comme $Z_\lambda = Y - \lambda X$, cela signifierait aussi que Y est avec certitude égale à $\lambda X + m$, donc que Y est une fonction affine, de pente λ , de X . Si la variance de Z_λ n'est pas nulle, mais petite, cela signifie que Z_λ s'écarte peu, ou ne s'écarte

qu'avec une faible probabilité, de sa moyenne m , et cela revient à dire que Y s'écarte peu, ou avec une faible probabilité, de $\lambda X + m$, ou encore, que la dépendance de Y par rapport à X est floue (comme dans le tableau 3), mais affine de pente λ . On peut donc dire que la variance de Z_λ est une mesure du degré de dépendance affine, pour une pente λ donnée, de Y par rapport à X . Voyons maintenant comment interpréter la covariance de X et Y dans ce contexte.

Désignons par $r_{j,k} = \mathcal{P}(X = x_j ; Y = y_k)$ la loi conjointe de X et de Y , par $a = \sum_{j,k} r_{j,k} x_j$ la moyenne de X , et par $b = \sum_{j,k} r_{j,k} y_k$ la moyenne de Y . Si on calcule la variance de Z_λ à partir de ces ingrédients, on obtient ceci :

$$\begin{aligned} \mathbf{Var}(Z_\lambda) &= \sum_{j,k} r_{j,k} [y_k - b - \lambda(x_j - a)]^2 \\ &= \sum_{j,k} r_{j,k} [(y_k - b)^2 - 2\lambda(x_j - a)(y_k - b) + \lambda^2(x_j - a)^2] \\ &= \mathbf{Var}(Y) - 2\lambda \mathbf{Cov}(X, Y) + \lambda^2 \mathbf{Var}(X) \end{aligned}$$

Or la variance est une grandeur qui est par nature positive, de sorte que, quelle que soit la valeur de λ , on aura toujours $\mathbf{Var}(Z_\lambda) \geq 0$; d'après ce qui précède, cela a pour conséquence que

$$\mathbf{Var}(Y) - 2\lambda \mathbf{Cov}(X, Y) + \lambda^2 \mathbf{Var}(X) \geq 0$$

Cette inégalité étant vraie quel que soit λ , on en déduit que nécessairement, dans tous les cas

$$\mathbf{Cov}(X, Y) \leq \sqrt{\mathbf{Var}(X) \cdot \mathbf{Var}(Y)} \quad (XII.1.)$$

en outre, on peut dire que, pour que $\mathbf{Var}(Z_\lambda) = 0$, il faut et il suffit que

$$\mathbf{Cov}(X, Y) = \sqrt{\mathbf{Var}(X) \cdot \mathbf{Var}(Y)} \quad (XII.2.)$$

et que

$$\lambda = \sqrt{\frac{\mathbf{Var}(Y)}{\mathbf{Var}(X)}} \quad (XII.3.)$$

Ainsi, $\mathbf{Cov}(X, Y)$ est maximum lorsque $\mathbf{Var}(Z_\lambda) = 0$, λ^2 étant égal à $\mathbf{Var}(Y)/\mathbf{Var}(X)$.

On peut donc résumer les choses ainsi : dans tous les cas la covariance de X et Y peut au maximum être égale à $\sqrt{\mathbf{Var}(X) \cdot \mathbf{Var}(Y)}$; elle *atteint* ce maximum lorsque Y est une fonction affine de X de pente (XII.3). Lorsque la corrélation n'est pas égale à ce maximum, mais en est proche, Y

est approximativement égale à $\lambda X + m$, ce qui veut dire que Y ne s'écarte sensiblement de $\lambda X + m$ qu'avec une faible probabilité (dépendance "floue").



figure 50

Ce graphique représente les valeurs prises par les deux variables aléatoires X et Y du tableau 2 du chapitre **VI** (les valeurs de X sont les abscisses des points et les valeurs de Y sont les ordonnées des points). Chaque point est affecté du poids correspondant à sa probabilité, donnée par la loi conjointe de X et Y : le point de coordonnées x_j, y_k a le poids $\mathcal{P}(X = x_j; Y = y_k)$ (mais les poids ne se voient pas sur le graphique). La droite de régression est représentée : c'est la droite la plus proche possible (en un certain sens, défini dans le texte) du nuage des points. Ici, bien que ce soit *la plus* proche, elle n'est pas proche car la corrélation entre les deux variables aléatoires est nulle.

La valeur maximum de la covariance donnée par (XII.2) dépend des variances de X et Y ; la covariance n'est donc pas la pure mesure du degré de dépendance linéaire entre X et Y , mais inclut aussi une mesure de leurs variances. Afin de séparer les deux, il est commode d'introduire la grandeur

$$\rho = \frac{\mathbf{Cov}(X, Y)}{\sqrt{\mathbf{Var}(X) \cdot \mathbf{Var}(Y)}} \quad (\text{XII.4.})$$

qu'on appelle le coefficient de corrélation de X et Y ; celui-ci est nécessairement compris entre -1 et $+1$, et il est égal à $+1$ si, et seulement si, l'égalité (XII.2.) est satisfaite, c'est-à-dire lorsque Y est une fonction affine de X . Le coefficient de corrélation mesure donc le degré de dépendance linéaire entre

X et Y , indépendamment de leurs variances respectives et sans préjuger de la pente λ . Ce degré est de 100% lorsque Y est une fonction affine de X , et est proche de 100% si Y est proche d'une fonction affine de X , ou du moins si Y ne s'écarte qu'avec une faible probabilité d'une fonction affine de X . Dans le cas du couple X, Y décrit sur le tableau 2 du chapitre **VI**, on a $Y = X^2$, c'est-à-dire que Y est fonction de X , mais le degré de dépendance *linéaire* de Y par rapport à X est nul. Cela peut se comprendre aisément : si on représente les valeurs que peut prendre (avec une probabilité non nulle) le couple X, Y par des points du plan ayant ces valeurs comme coordonnées, on obtient des points situés sur une parabole (voir figure 50) ; si Y était une fonction affine de X , les points seraient sur une droite. On peut dire que le degré de dépendance linéaire de Y par rapport à X est élevé si l'ensemble des points reste proche d'une certaine droite (par exemple sur la figure 51). Pour les points de la figure 50, aucune droite ne peut approcher correctement la parabole, et toutes les droites possibles sont également peu satisfaisantes ; c'est pourquoi on peut dire que le degré de dépendance linéaire de Y par rapport à X est nul.

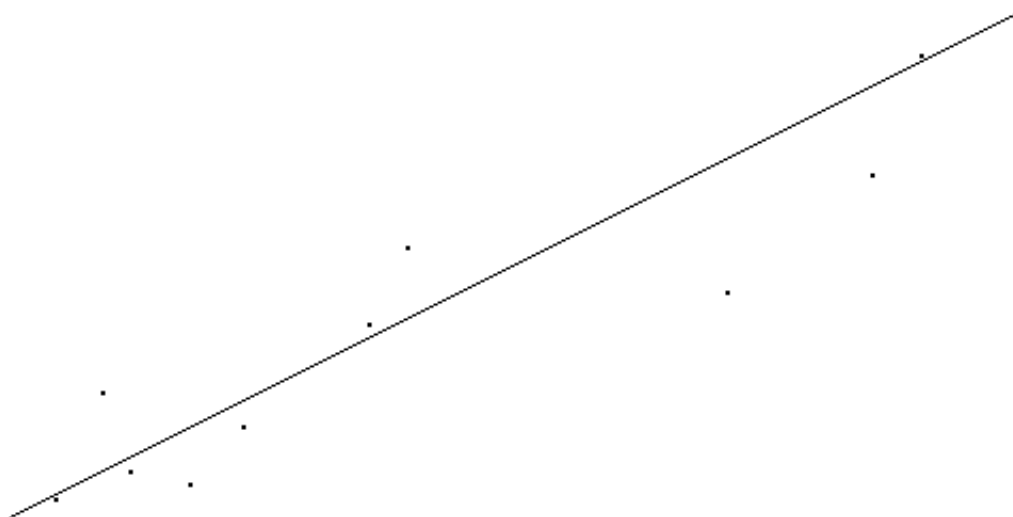


figure 51

Ce graphique montre un nuage de points correspondant aux valeurs qui seraient prises par deux variables aléatoires fortement corrélées (c'est-à-dire dont le coefficient de corrélation est proche de 1). Ici, contrairement à la figure 50, la droite de régression est réellement proche du nuage des points.

En général, pour un couple quelconque X, Y de variables aléatoires, il existe une valeur de λ qui rend minimum l'expression (XII.1.), c'est-à-dire qui rend minimum la variance de $Z_\lambda = Y - \lambda X$; la droite d'équation

$y = \lambda x + \mathbf{E}(Z_\lambda)$, correspondant à ce minimum, est appelée la *droite de régression* de Y par rapport à X . En interprétant la variance de Z_λ comme la somme des carrés des écarts des points par rapport à la droite (pondérés par les probabilités correspondantes), on peut dire que la droite de régression est la droite des “moindres carrés”. Si on calcule la valeur de λ correspondant à ce minimum, on obtient

$$\lambda = \rho \sqrt{\frac{\mathbf{Var}(Y)}{\mathbf{Var}(X)}}$$

ce qui redonne (XII.3) lorsque $\rho = 1$. L’ordonnée à l’origine de la droite de régression est simplement $\mathbf{E}(Z_\lambda) = \mathbf{E}(Y) - \lambda \mathbf{E}(X)$.

Il faut bien comprendre ceci : on peut toujours calculer la droite de régression, sauf si la variance de X est nulle. Par exemple, dans le cas du tableau 2 du chapitre VI, la droite de régression existe (elle est représentée sur la figure 50) : c’est la droite de pente 0 et d’ordonnée à l’origine 2. En effet $\lambda = \rho \sqrt{\mathbf{Var}(Y)/\mathbf{Var}(X)} = 0$ et $\mathbf{E}(Y) - \lambda \mathbf{E}(X) = 2$. Mais bien que cette droite existe, elle n’est pas proche de la parabole, de sorte qu’elle n’a rien à voir avec la dépendance de Y par rapport à X : elle minimise bien la somme des carrés des écarts, mais ce minimum n’est pas petit. Ainsi, pour n’importe quel couple X, Y de variables aléatoires (sauf si $\mathbf{Var}(X) = 0$), on peut trouver une fonction affine des moindres carrés, mais celle-ci ne représentera correctement la dépendance de Y par rapport à X que si le coefficient de corrélation est assez proche de 1.

Dans le cas du tableau 2 du chapitre VI, on devrait pouvoir dire que le degré de dépendance linéaire de Y par rapport à X est 0, mais qu’en revanche le degré de dépendance “quadratique” de Y par rapport à X est de 100%. Peut-on introduire des paramètres liés aux variables aléatoires X et Y , qui joueraient un rôle analogue à la covariance et au coefficient de corrélation, et qui permettraient de calculer quantitativement le degré de dépendance “quadratique” de Y par rapport à X ? La réponse est oui, mais il n’existe pas pour des fonctions non linéaires de procédé aussi simple que pour le cas linéaire. Nous reprendrons cela plus loin, dans la section 4 “régression non linéaire”).

XII. 2. Moyenne, variance, et covariance empiriques.

Dans le Calcul des probabilités, on calcule des probabilités à partir d’une invariance postulée. Les probabilités que l’on déduit ainsi par le calcul, sont appelées des *probabilités a priori*. Leurs valeurs sont exactes : par exemple on trouvera que la probabilité de tel ou tel événement est $\frac{31}{73}$ et non un nombre approximativement égal à $\frac{3}{7}$, 0.42 ou 0.425.

À l'inverse du Calcul des probabilités, la Statistique ne traite pas de probabilités a priori. Elle ne fait que mesurer sur des échantillons. Nous avons déjà discuté au chapitre **X** ce qu'elle mesure. Soit, par prélèvement d'un échantillon sur une population, elle permet (c'est la méthode du sondage) de mesurer approximativement les proportions exactes sur la population totale; soit elle permet, en répétant un grand nombre de fois une expérience reproductible (par exemple le lancer d'une pièce de monnaie) de mesurer approximativement les probabilités a priori de l'expérience. Si par exemple on a une pièce de monnaie non équilibrée, qui a un peu plus de chances de tomber sur pile que sur face, on peut difficilement trouver le "niveau où intervient le hasard pur" (celui-ci était facile à trouver si la pièce est symétrique, car justement la symétrie conduit à des invariances, mais sans symétrie, on ne peut plus). Dans ce cas on peut cependant mesurer a posteriori la probabilité de pile en lançant un grand nombre de fois la pièce et en comptant combien de fois la pièce est tombée sur pile. Pour connaître les fréquences moyennes d'accidents (et cette connaissance est nécessaire pour organiser des services d'assistance ou calculer le montant des primes d'assurance) on ne peut pas procéder autrement, car on est incapable de trouver une invariance a priori. Il en va de même lorsqu'on explore un domaine entièrement nouveau, où il existe une invariance sous-jacente, mais inconnue. Par exemple pour la statistique de Bose – Einstein, on ignorait au départ le principe d'invariance sous-jacent ("indiscernabilité des particules"). Celui-ci n'a pas été deviné à partir de rien par la seule logique: il a été déduit par induction à partir de résultats *statistiques* expérimentaux.

Ainsi la Statistique traite des données brutes, à travers lesquelles on cherche à détecter l'effet de probabilités inconnues. Les données brutes sont mesurées sur des objets, et un ensemble de données est mesuré sur un ensemble d'objets appelé un *échantillon*: les données sont des grandeurs se rapportant aux objets de l'échantillon. Par exemple l'échantillon est un ensemble de pièces manufacturées, disons pour fixer les idées cent cinquante cylindres en acier inoxydable. On peut mesurer les longueurs de ces cylindres, les diamètres, et les poids. Si on numérote (ne serait-ce que par la pensée) les cylindres de 1 à 150, appelons par exemple x_i la longueur du i^{e} cylindre, y_i son poids, et z_i son diamètre. Les trois grandeurs x_i , y_i , et z_i viennent donc du même objet $N^{\circ}i$.

On distinguera donc l'ensemble des *objets*, appelé échantillon, et les données numériques qui s'y rapportent, qui sont les *variables d'échantillon* ou *grandeurs empiriques*. Ainsi la longueur x , le poids y , et le diamètre z de l'exemple précédent sont trois variables empiriques relatives au même

échantillon. De même que le Calcul des probabilités traite essentiellement de variables aléatoires, la Statistique traite de variables d'échantillon. Au paragraphe 1 ci-dessus, nous avons vu comment la covariance de deux variables aléatoires permettait d'évaluer leur degré de dépendance linéaire mutuelle. Si on souhaite détecter sur un échantillon les effets d'une relation linéaire inconnue entre deux paramètres, on ne peut pas utiliser les grandeurs introduites au paragraphe 1, à savoir la covariance ou le coefficient de corrélation, puisque nous ne connaissons pas de probabilités a priori. C'est pourquoi la Statistique fait appel à des grandeurs de nature différente, liées à l'échantillon et non à des probabilités a priori (mais ayant un rapport avec les précédentes, que nous allons élucider), et qui sont la moyenne, la variance, ou la covariance *d'échantillon*. On dit aussi moyenne, variance, ou covariance *empiriques*. Par opposition, on appellera moyenne, variance, ou covariance *théorique* la moyenne, variance, ou covariance d'une variable aléatoire (telles qu'elles sont définies au paragraphe 1).

Pour un échantillon comportant n objets, la moyenne d'une variable d'échantillon x_i ($i = 1, 2, \dots, n$) est définie comme suit

$$M = \frac{1}{n} \sum_{i=1}^{i=n} x_i \quad (XII.5.)$$

Ceci est très facile à comprendre puisque c'est ainsi qu'on calcule les moyennes aux examens (l'échantillon est alors l'ensemble des copies). La variance d'échantillon de la variable x_i sera

$$P = \frac{1}{n-1} \sum_{i=1}^{i=n} (x_i - \bar{x})^2 \quad (XII.6.)$$

où $\bar{x} = M$ est la moyenne de la même variable. Notez bien le facteur $(n-1)$ et non $1/n$). La covariance d'échantillon de deux variables x_i et y_i se rapportant au *même* échantillon est

$$Q = \frac{1}{n-1} \sum_{i=1}^{i=n} (x_i - \bar{x})(y_i - \bar{y}) \quad (XII.7.)$$

où $\bar{x}$ est la moyenne des x_i et $\bar{y}$ celle des y_i . La covariance entre deux variables ne se rapportant pas au même échantillon n'a généralement pas de signification.

On notera tout particulièrement le coefficient $\frac{1}{n}$ pour les moyennes, mais $\frac{1}{n-1}$ pour les variances ou la covariance. Nous donnerons plus loin l'explication de ce choix qui au premier abord peut sembler bizarre.

La variance (théorique) d'une variable aléatoire et la variance (empirique) d'une variable d'échantillon sont des concepts différents, quoique apparentés. De même pour les moyennes ou les covariances. Du point de vue mathématique, d'abord, le calcul de la moyenne ou de la variance d'une variable aléatoire s'effectue en affectant à chaque valeur de la variable un *poids* égal à la probabilité de cette valeur ; par contre lorsqu'on calcule la moyenne ou la variance d'un échantillon, tous les points (ou coordonnées) de l'échantillon ont le même poids (par exemple $\frac{1}{n-1}$ pour la variance et $\frac{1}{n}$ pour la moyenne). Du point de vue pratique, ensuite, la variance d'une variable aléatoire se calcule d'après des probabilités a priori, tandis que la variance d'un échantillon se calcule d'après des résultats qui se sont produits, qui n'ont donc pas ou qui n'ont plus de probabilité.

Lorsqu'on dispose de données statistiques sur une population (par exemple la tension artérielle pour chaque personne d'un groupe de cent), on peut calculer la moyenne ou la variance de la tension artérielle pour cet échantillon. Il suffit d'appliquer une formule, et cela peut se faire sans se poser aucune question. Les problèmes commencent lorsqu'on se demande ce que représentent exactement les grandeurs ainsi calculées : quel est leur sens, quelle information peut-on en tirer concernant l'échantillon ? Et aussi : pourquoi la variance doit-elle être calculée avec le coefficient $\frac{1}{n-1}$ alors que la moyenne l'est avec le coefficient $\frac{1}{n}$?

Pour éclaircir ces mystères, il faut revenir à une remarque faite au chapitre **X** : à propos de l'expérience consistant à tester l'effet d'un médicament sur un groupe, nous avons distingué soigneusement, d'une part la probabilité a priori pour que le médicament agisse sur une personne donnée (cette probabilité résulte du hasard créé par le chaos des mécanismes moléculaires), et d'autre part la distribution des effets parmi les différentes personnes d'un groupe. On peut enregistrer les effets du médicament sur le groupe en notant pour chaque personne la baisse de tension sur les 24 heures qui suivent l'administration du médicament ; l'ensemble de ces données chiffrées constitue alors une variable d'échantillon : on pourra calculer la moyenne d'échantillon et la variance d'échantillon. Par ailleurs, on peut imaginer *pour chaque personne considérée séparément*, la variable aléatoire dont les valeurs sont les baisses de tension possibles avec les probabilités a priori correspondantes (ces probabilités sont évidemment difficiles à mesurer et impossibles à calculer a priori, et en outre dépendent du temps, mais cela importe peu pour comprendre le principe). On peut alors imaginer aussi ce que signifie la moyenne (ou espérance mathématique) de cette variable aléatoire, ou sa variance, et comprendre que ces grandeurs n'ont aucun rapport avec la moyenne ou la variance d'échantillon calculée sur le

groupe. Comme nous l'avons souligné au chapitre **X**, il n'y a aucune raison que la moyenne d'échantillon calculée sur le groupe soit égale à la moyenne de la variable aléatoire liée à une personne particulière, et de même pour la variance : cela provient de ce que les phénomènes se produisant dans le métabolisme d'une personne particulière ne peuvent pas influencer sur ce qui se passera dans le métabolisme des autres. Il s'agit de deux grandeurs qui n'ont aucun rapport entre elles. On peut même affirmer, en l'absence de phénomènes tels que la contagion ou des comportements de groupe avec incidences biologiques (tabagisme, alcoolisme, habitudes alimentaires, etc.) que certaines variables aléatoires liées au métabolisme d'individus différents sont stochastiquement indépendantes.

Il en va tout différemment si on considère un échantillon provenant de la répétition d'une *expérience reproductible*. Si la variable aléatoire (appelons-la X) est le résultat d'une expérience reproductible, elle aura une moyenne $m = \mathbf{E}(X)$ et une variance $v = \mathbf{Var}(X)$ (dites *théoriques*). Par exemple pour le jeu de pile ou face, X vaudrait 0 pour face et 1 pour pile, avec probabilité $\frac{1}{2}$, ce qui donnerait $m = \frac{1}{2}$ et $v = \frac{1}{4}$. Si on reproduit l'expérience un grand nombre de fois, la loi des grands nombres aura pour effet que la proportion de chacun des résultats possibles sera proche de la probabilité a priori ; par exemple, si on lance mille fois la pièce, on obtiendra environ 500 pile et 500 face (avec, comme nous l'avons déjà vu, une incertitude de l'ordre de ± 30). Si on considère les 1000 résultats comme un échantillon statistique, et qu'on calcule la moyenne et la variance *empiriques* pour cet échantillon, alors elles seront proches de la moyenne et de la variance théoriques de la variable aléatoire ; si on poursuit le lancement de la pièce, 2000 fois, 10 000 fois, 100 000 fois, etc., la taille de l'échantillon augmentera et la moyenne empirique sera de plus en plus proche de l'espérance mathématique m de la variable aléatoire X (respectivement : la variance empirique sera de plus en plus proche de la variance théorique v).

On peut donc dire ceci : la moyenne (resp. la variance) d'une grandeur empirique sur un échantillon $\mathcal{A}$ et la moyenne (resp. la variance) théorique d'une variable aléatoire X qui représente le résultat possible d'une expérience reproductible $\mathcal{E}$, sont deux grandeurs pratiquement identiques **lorsque l'échantillon $\mathcal{A}$ est constitué par la répétition d'un grand nombre de fois l'expérience $\mathcal{E}$** (ou encore : les grandeurs empiriques tendent vers les valeurs théoriques lorsque la taille de l'échantillon tend vers l'infini). Mais cette équivalence entre les grandeurs empiriques et théoriques n'est valable que pour des expériences reproductibles. Lorsqu'on fait appel à cette équivalence pour interpréter des données relatives à des organismes vivants, par exemple, il faut s'assurer que l'hypothèse de reproductibilité est à peu

près légitime.

Les définitions des grandeurs empiriques données en (XII.5.) et (XII.6.) ont été convenues délibérément de manière à satisfaire cette équivalence. En particulier, le mystérieux facteur $\frac{1}{n-1}$ qui apparaît dans ces expressions de la variance (ou de la covariance) empirique, n'a pas d'autre justification que d'être le facteur qui rend la grandeur empirique le plus proche possible de la grandeur théorique correspondante. Si on avait pris le facteur $\frac{1}{n}$ dans la variance, au lieu du facteur $\frac{1}{n-1}$, la grandeur empirique coïnciderait moins bien avec la grandeur théorique. Tout cela s'explique mathématiquement comme suit.

Ayant en vue le principe d'équivalence énoncé ci-dessus, considérons l'échantillon comme résultant de la répétition d'une même expérience, pour laquelle des probabilités a priori existent. Soit donc la variable aléatoire X dont les r valeurs x_j ($j = 1, 2, \dots, r$) sont prises avec probabilité p_j . Si on répète n fois l'expérience décrite par X , on obtiendra un certain nombre n_j de fois la valeur x_j , de sorte que $\sum_j n_j = n$. On peut alors former les grandeurs

$$S_0 = \sum_{j=1}^{j=r} n_j (x_j - m)^2$$

où m est l'espérance mathématique de la variable X , c'est-à-dire la moyenne théorique a priori, et

$$S_1 = \sum_{j=1}^{j=r} n_j (x_j - M)^2$$

où $M = \frac{1}{n} \sum n_j x_j$ est la moyenne empirique de l'échantillon. Si on part du principe que les résultats x_j sont tous des réalisations d'une même expérience reproductible, cela signifie que S_0 et S_1 sont des variables aléatoires qu'on peut écrire

$$S_0 = \sum_{i=1}^{i=n} (X_i - m)^2$$

$$S_1 = \sum_{i=1}^{i=n} (X_i - M)^2$$

où les X_i sont n variables aléatoires stochastiquement indépendantes et de même loi (celle de X), et où $M = \frac{1}{n} \sum_i X_i$. Cela n'est que l'expression mathématique de l'hypothèse que l'expérience est reproductible, et signifie simplement qu'on reproduit n fois, *indépendamment* et *à l'identique* la variable aléatoire X .

La différence essentielle entre S_0 et S_1 est que les variables aléatoires $X_i - m$ sont stochastiquement indépendantes, alors que les $X_i - M$ sont liées par la relation que leur somme est nulle. Entre S_0 et S_1 on a l'identité

$$S_0 = S_1 - n(M - m)^2 \quad (XII.8)$$

Cela est facile à établir, il suffit de développer les expressions de S_0 et S_1 . Pour la première :

$$\begin{aligned} S_0 &= \sum_{i=1}^{i=n} (X_i - m)^2 \\ &= \sum_{i=1}^{i=n} X_i^2 - 2m \sum_{i=1}^{i=n} X_i + n m^2 \\ &= \sum_{i=1}^{i=n} X_i^2 - 2n m M + n m^2 \end{aligned}$$

Pour la seconde :

$$\begin{aligned} S_1 &= \sum_{i=1}^{i=n} (X_i - M)^2 \\ &= \sum_{i=1}^{i=n} X_i^2 - 2M \sum_{i=1}^{i=n} X_i + n M^2 \\ &= \sum_{i=1}^{i=n} X_i^2 - 2n M^2 + n M^2 \\ &= \sum_{i=1}^{i=n} X_i^2 - n M^2 \end{aligned}$$

On voit en soustrayant membre à membre que les termes $\sum X_i^2$ s'annulent et il reste

$$S_0 - S_1 = -2n m M + n m^2 + n M^2 = n(M - m)^2$$

ce qui montre bien (XII.8).

L'espérance mathématique de S_0 est par définition $n \mathbf{Var}(X)$, donc celle de S_1 , $\mathbf{E}(S_1)$, est égale à $n \mathbf{Var}(X) - \mathbf{E}(n[M - m]^2)$.

Il est facile de calculer $\mathbf{E}([M - m]^2)$:

$$\mathbf{E}([M - m]^2) = \mathbf{E}\left(\left[\frac{1}{n} \sum_{i=1}^{i=n} (X_i - m)^2\right]\right) = \frac{1}{n} \mathbf{Var}(X)$$

Par conséquent $\mathbf{E}(S_1) = (n - 1) \mathbf{Var}(X)$, alors que $\mathbf{E}(S_0) = n \mathbf{Var}(X)$. On comprend alors pourquoi on a introduit le coefficient $1/(n - 1)$ dans XII.6 :

d'après la loi des grands nombres appliquée à l'expérience reproductible, si n est grand, les valeurs de S_1 fluctueront autour de leur moyenne $\mathbf{E}(S_1) = (n-1) \mathbf{Var}(X)$, avec des écarts de l'ordre de $\sqrt{n}$. La différence entre $\mathbf{E}(S_1)$ et $\mathbf{E}(S_0)$, égale à $\mathbf{Var}(X)$, est certes faible comparée à $\sqrt{n}$ (amplitude moyenne des fluctuations), mais introduit un décalage systématique ou *biais*. La valeur $\frac{1}{n-1} \mathbf{E}(S_1)$ est une meilleure estimation de $\mathbf{Var}(X)$ que $\frac{1}{n} \mathbf{E}(S_1)$, parce qu'elle est basée sur le centre exact des fluctuations de la variable S_1 .

Il faut bien noter que le choix du coefficient $\frac{1}{n-1}$ repose sur une idée préconçue de la “vraie” variance, à savoir que la “vraie” variance est $\mathbf{Var}(X)$. Cette idée préconçue n'a de sens que par le principe de l'équivalence entre grandeurs théoriques et grandeurs empiriques pour les expériences reproductibles. Le coefficient $\frac{1}{n-1}$ ne sert qu'à ajuster la variance empirique à la variance théorique, de manière à satisfaire au plus juste l'équivalence.

Si on étudie une répartition statistique qui ne provient pas d'une expérience reproductible, par exemple dans l'étude de l'effet d'un médicament sur différentes personnes (cf chapitre **X**), il n'est plus possible d'avoir une idée préconçue de la “vraie” variance. Dans ce cas n'importe quel coefficient autre que $\frac{1}{n-1}$ est tout aussi bon, car la variance empirique (qui seule existe) ne peut servir qu'à comparer les dispersions statistiques sur différents groupes; bien sûr, pour que la comparaison d'études provenant de pays différents soit possible et sensée, il faut une norme internationale (par exemple, que *par convention* tout le monde adopte la somme S_1 des carrés des écarts, ou $\frac{1}{n} S_1$, ou encore $\frac{1}{n-1} S_1$), mais la définition *XII.6*, c'est-à-dire $\frac{1}{n-1} S_1$, n'est alors pas plus juste qu'une autre.

La norme internationale qui a été adoptée est celle de la définition *XII.6*, par analogie avec les expériences reproductibles. Mais en dehors de ces dernières, ce n'est qu'une convention. Il ne faut donc pas en être dupe.

XII. 3. Corrélation statistique.

Nous avons vu au paragraphe 1 que la corrélation était la dépendance linéaire (mesurée par le coefficient de corrélation). Mais cela a été traité pour des variables aléatoires, dont la loi conjointe est connue. De même qu'on a pu, dans la section **XII.1**, dériver la corrélation de deux variables aléatoires de leur covariance et de leurs variances, on doit pouvoir dériver une corrélation empirique de la covariance et de la variance empiriques d'un échantillon. Formellement, cela ne pose aucun problème; mais il faudra en interpréter la signification.

Considérons l'exemple suivant. Dans un groupe de cent personnes

(numérotées de $i = 1$ à $i = 100$ par randomisation), on relève la tension artérielle. On représente celle-ci sur un graphique de la manière suivante: à chaque personne i correspond un point, dont l'abscisse x_i est l'âge, et l'ordonnée y_i la tension (disons la tension systolique pour fixer les idées). On dispose ainsi d'un échantillon de cent points dans le plan $\{x, y\}$, mais ceux-ci ne sont pondérés par aucune loi de probabilité. Les *moyennes d'échantillon* sont alors les nombres $\bar{x} = \frac{1}{100} \sum_{i=1}^{100} x_i$ (âge moyen de l'échantillon) et $\bar{y} = \frac{1}{100} \sum_{i=1}^{100} y_i$ (tension moyenne de l'échantillon). La *covariance d'échantillon* de l'âge et de la tension est la grandeur

$$Q = \frac{1}{99} \sum_{i=1}^{i=100} (x_i - \bar{x}) (y_i - \bar{y})$$

et de même les *variances d'échantillon* sont

$$P = \frac{1}{99} \sum_{i=1}^{i=100} (x_i - \bar{x})^2 \quad (\text{variance de l'âge})$$

$$R = \frac{1}{99} \sum_{i=1}^{i=100} (y_i - \bar{y})^2 \quad (\text{variance de la tension})$$

Rien n'empêche d'introduire un coefficient de corrélation empirique formellement analogue à celui du paragraphe 1 :

$$\rho = \frac{\text{covariance d'échantillon (entre la tension et l'âge)}}{\sqrt{\text{variance de l'âge} \times \text{variance de la tension}}}$$

Sur le graphique on a cent points (dont l'abscisse représente l'âge et l'ordonnée la tension). Il est facile de vérifier que si ρ était égal à ± 1 ces cent points seraient sur une droite: il n'est en effet pas nécessaire de refaire les calculs du paragraphe 1, puisque formellement la situation est la même (au paragraphe 1, les cent points avaient pour poids les probabilités a priori, ici ils ont tous le même poids $1/99$, mais les calculs effectués formellement ne sont pas affectés par la différence de signification). La pente de la droite serait égale à

$$\lambda = \begin{cases} +\sqrt{\frac{\text{variance de la tension}}{\text{variance de l'âge}}} & \text{si } \rho = +1 \\ -\sqrt{\frac{\text{variance de la tension}}{\text{variance de l'âge}}} & \text{si } \rho = -1 \end{cases}$$

Bien entendu il n'arrive jamais dans la réalité que les cent points soient sur une droite (cela signifierait que la tension serait exactement une fonction

linéaire de l'âge !). Mais ce qu'on observe, c'est que la tension est en moyenne plus élevée chez des personnes plus âgées. Sur le graphique, cela devrait avoir l'effet que les points dont l'abscisse est plus grande doivent avoir aussi ("en moyenne") une ordonnée plus grande, ce qui se traduit par un nuage de points allongé et incliné comme ceux qui sont représentés sur la figure 52; un graphique de la tension et de l'âge aurait l'aspect de la figure 52.2 ou 52.3, donc la corrélation est assez faible.

En statistique, la connaissance de la droite de régression joue un rôle capital. Elle sert en effet à corriger des données brutes et à séparer les diverses causes qui influencent un phénomène. Supposons que nous cherchions à déterminer une corrélation entre le tabagisme et l'hypertension. Nous effectuons une étude clinique sur deux groupes de personnes, un groupe de cinquante fumeurs consommant au moins vingt cigarettes par jour, et un groupe témoin de cinquante personnes qui ne fument jamais. À première vue, il peut sembler que, pour que l'étude soit valide, il suffit de choisir ces groupes par randomisation. En effet, il s'agit de montrer que sur l'ensemble de la population (disons la population française pour fixer les idées) la tension artérielle moyenne de l'ensemble des fumeurs est nettement supérieure à la tension moyenne sur les non-fumeurs; comme il est impossible de faire une étude sur tous les fumeurs français, on utilise la méthode du sondage en prélevant des échantillons de cinquante personnes dans ces deux populations (comme on ne s'intéresse pas à connaître la tension artérielle moyenne à trois décimales près, mais seulement de vérifier que l'une est nettement supérieure à l'autre, un échantillon de taille modeste suffit). La randomisation garantira alors que les échantillons sont bien "pris au hasard", conformément à ce qui avait été dit au chapitre X.

Mais il se peut que les fumeurs ne soient pas répartis uniformément selon l'âge. Par exemple, une mode anti-tabac a pu, au cours des années soixante-dix, inciter beaucoup de jeunes à ne jamais fumer, habitude qu'ils garderont ensuite, tandis que chez leurs aînés les non-fumeurs seraient restés rares. Si des échantillons sont choisis par randomisation dans l'ensemble de la population, on aura certes un reflet de la population totale, mais comment savoir si la tension artérielle plus élevée qu'on observera dans le groupe des fumeurs provient du tabagisme et non tout simplement du seul fait que les personnes plus âgées y sont plus nombreuses que dans le groupe des non-fumeurs? Ce problème demeurerait, même si l'étude était effectuée sur la population totale, et ne peut être attribué à la taille modeste de l'échantillon.

Une solution est évidemment de réunir, non pas deux échantillons,

mais quarante répartis selon l'âge, de manière à disposer de vingt couples d'échantillons fumeurs / non-fumeurs, chacun de ces couples étant homogène quant à l'âge. Mais cette solution est beaucoup plus onéreuse. L'étude statistique des corrélations permet de s'en passer.

Il suffit en effet de représenter les résultats concernant les deux échantillons sur deux graphiques, chaque personne correspondant à un point dont l'abscisse est l'âge et l'ordonnée la tension ; on obtient ainsi deux graphiques de cinquante points chacun, l'un représente les non-fumeurs et l'autre les fumeurs. Si, comme il était à craindre, le groupe des fumeurs contient davantage de personnes âgées de plus de quarante ans que le groupe des non-fumeurs, cela se traduira sur le graphique par une inégale répartition de la densité des points : sur le graphique fumeurs, le nuage de points sera plus dense vers la droite, sur le graphique non-fumeurs, il sera plus dense vers la gauche. Mais la droite de régression n'est pas affectée par ces différences de densité. Bien sûr si on prend un autre échantillon, on peut observer une droite de régression un peu différente, et une répartition de densité également différente, mais ce ne sont que des fluctuations par rapport à une valeur moyenne. Ce qui intéresse alors le clinicien n'est pas la répartition, inégale, de la densité selon l'âge, mais la droite de régression : si $y = ax + b$ est la droite de régression des non-fumeurs, et $y = cx + d$ celle des fumeurs, l'influence du tabagisme sur la tension artérielle sera considérée comme établie si **pour tout** x , $cx + d$ est supérieur à $ax + b$. On comprend que, si la corrélation est forte, (c'est-à-dire si le nuage de points ne s'écarte pas beaucoup de la droite), le fait que pour une certaine valeur de x on ait $cx + d > ax + b$ signifie que parmi les personnes d'âge x , la tension artérielle est plus élevée chez les fumeurs que chez les non-fumeurs. Si cette inégalité a lieu pour tout x , cela signifie qu'indépendamment de l'âge, la tension est plus élevée chez les fumeurs que chez les non-fumeurs. Toutefois, pour que cette conclusion soit fondée, certaines conditions doivent être réunies, qui sont les trois suivantes.

1. Si $cx + d - ax - b$, quoique formellement positif, est nettement plus petit que la largeur (mesurée par la variance de $y - ax$) du nuage de points, la signification du résultat est nulle, car la valeur moyenne de $cx + d - ax - b$ pourrait être 0 et la valeur positive trouvée une fluctuation pratiquement aussi fréquente que la valeur 0 elle-même ; en particulier, si le nuage de points est très large (donc si le coefficient de corrélation est petit), il est pratiquement impossible de fonder la conclusion. Si on veut estimer quantitativement la signification du résultat, autrement dit décider si la valeur positive est significative ou au contraire due à des fluctuations normales, on effectuera un test du χ^2 ou apparenté.

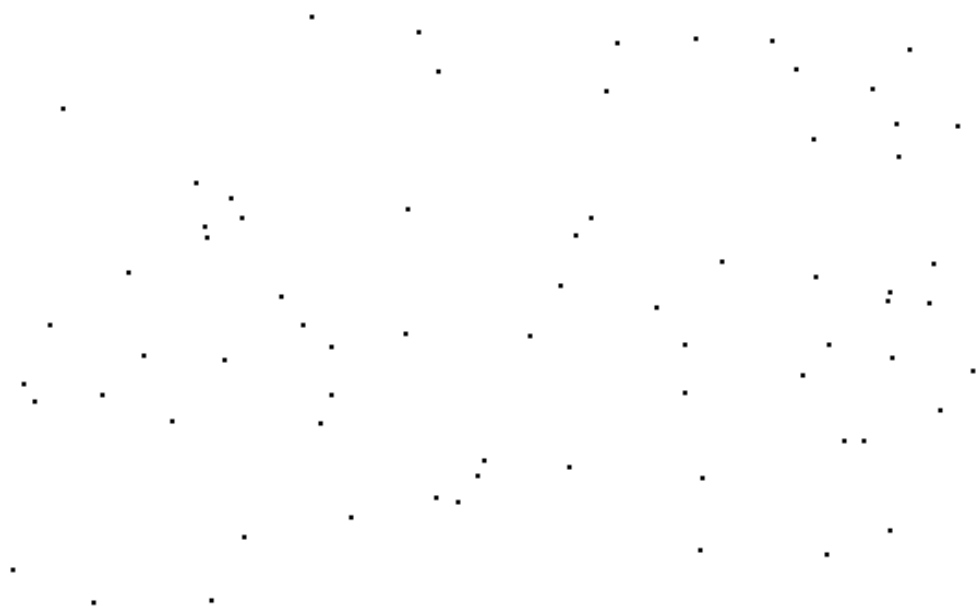


figure 52.1

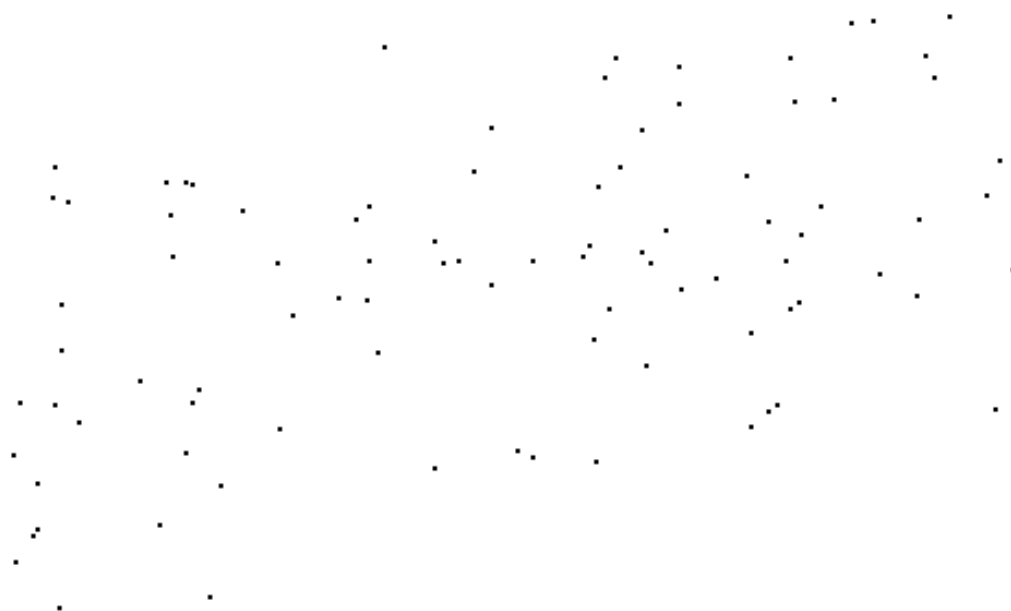


figure 52.2

Nuages de points faiblement corrélés



figure 52.3

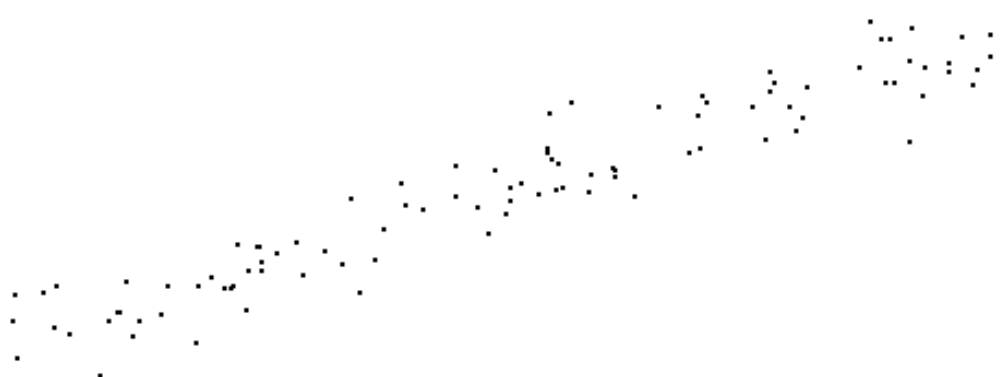


figure 52.4

Ici, la corrélation est plus grande qu'à la page précédente (en fait le coefficient de corrélation ρ dans les différentes figures est tel que d'une figure à la figure suivante $\sqrt{1 - \rho^2}$ est diminué de moitié).

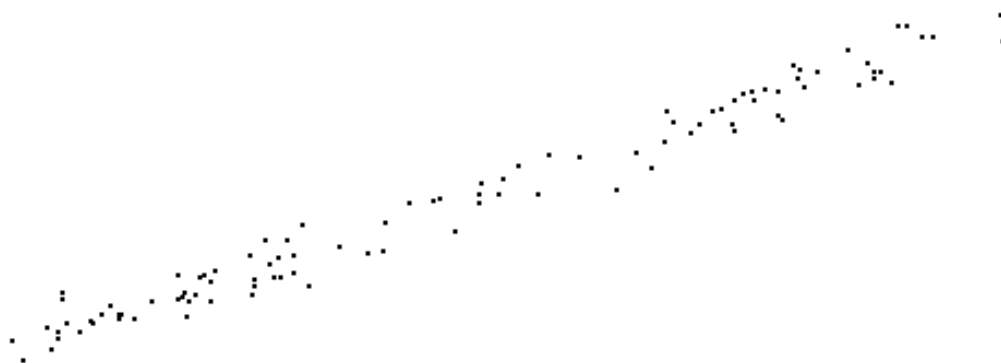


figure 52.5



figure 52.6

Les nuages de points sont maintenant très nettement distribués à proximité de leur droite de régression. On peut dire que celle-ci a une véritable signification.

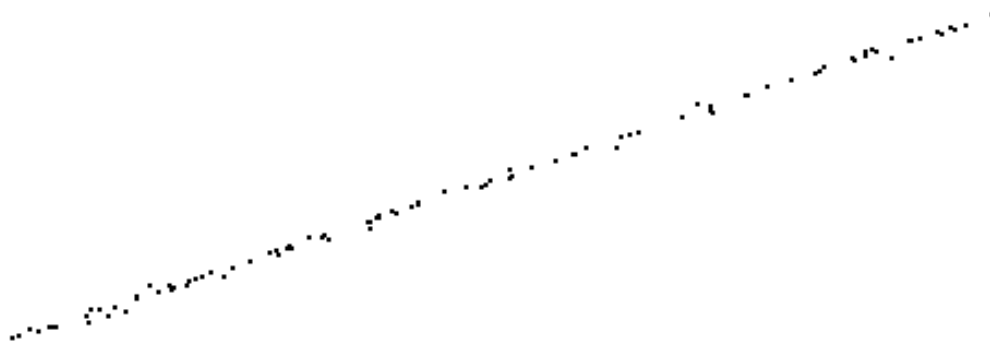


figure 52.7

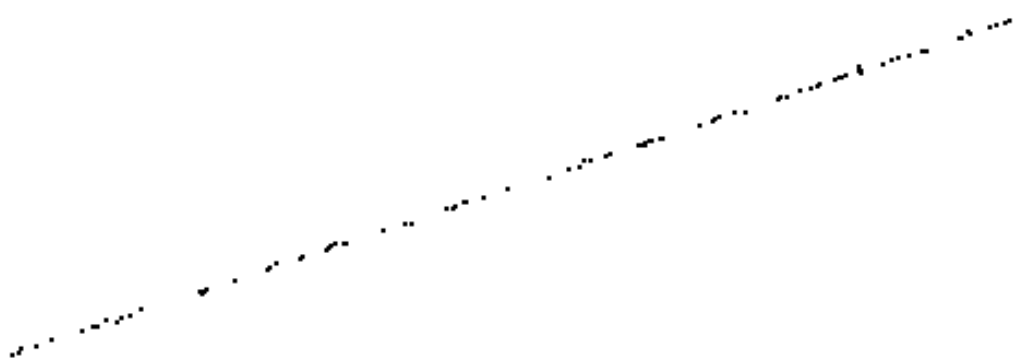


figure 52.8

Enfin, voici l'aspect pris par le nuage de points lorsque la corrélation est vraiment très forte. Sur 52.8, $1 - \rho^2$ est 4000 fois plus petit qu'en 52.1, ce qui veut dire que ρ est pratiquement égal à 1. Dans ce cas on peut dire que l'ordonnée est pratiquement une fonction linéaire de l'abscisse.

2. L'étude de la régression a été rendue nécessaire (au lieu que nous nous contentions d'un simple calcul de moyenne) par le soupçon d'une corrélation entre le tabagisme et l'âge, corrélation qui se superposerait à la corrélation cherchée entre le tabagisme et l'hypertension et fausserait ainsi les observations. Mais qui nous prouve qu'il n'existe pas d'autres corrélations, que nous ne soupçonnons pas, et qui viendraient fausser également les résultats? Il existe par exemple une corrélation entre la tension et le poids (les obèses sont davantage hypertendus), or fumer agit *négativement* sur le poids. Il faudrait alors, en toute rigueur, étudier aussi la régression tension / poids dans les deux groupes, et effectuer une nouvelle comparaison. Il faudrait refaire la même opération pour *toutes* les corrélations possibles (du moins celles qui sont suffisamment fortes et par conséquent non négligeables), et pouvoir s'assurer qu'on n'en ignore aucune.

3. Enfin, toutes ces études de régression sont valables tant qu'il s'agit de dépendances linéaires; la corrélation entre la tension et l'âge est suffisamment nette pour qu'on puisse dire — avec la précision correspondante — que la tension augmente en moyenne linéairement avec l'âge. Mais si la corrélation est nulle, qu'il existe cependant une nette dépendance non linéaire (comme cela a été illustré sur la figure 50), les droites de régression ne nous apprennent plus rien et ne nous permettent plus de corriger les données statistiques. Il faut alors faire appel à la régression non linéaire (voir section suivante). On peut donc avoir calculé soigneusement toutes les corrélations possibles et effectué les corrections correspondantes, mais aboutir à des conclusions fausses simplement parce qu'une dépendance non linéaire très forte aura échappé à toutes ces précautions.

Toutes ces réserves sont nécessaires pour montrer les limites de l'analyse statistique. Dans beaucoup de situations pratiques on rencontre suffisamment peu de corrélations ou de dépendances pour que ces méthodes aient néanmoins une valeur empirique correcte. Mais il est bien clair que tous les abus (pressentis et dénoncés par le bon sens populaire) proviennent du non-respect de ces réserves. Aucune méthode mathématique ne permet d'en faire l'économie: on peut corriger des données par des calculs de régression (linéaire ou non linéaire), mais les résultats ne sont pleinement corrigés que si l'ensemble des dépendances a été pris en compte; il est légitime de négliger les influences faiblement corrélées avec le paramètre étudié, mais il suffit d'avoir ignoré un seul facteur fortement corrélé avec le paramètre étudié pour que la conclusion perde toute validité. Il est vain d'appliquer des méthodes mathématiques, aussi sophistiquées soient-elles: en aucun cas on ne pourra ainsi compenser l'omission des facteurs occultes.

XII. 4. Régression non linéaire.

La recherche de la droite de régression d'une variable Y par rapport à une variable X se ramenait à un problème de moindres carrés : trouver, pour un nuage de points donné (avec ou sans poids) la fonction linéaire $Y = f(X) = aX + b$ qui *minimise* les écarts aléatoires. Il va de soi que cette opération est possible pour n'importe quelle fonction $Y = f(a, b; X)$ dépendant de deux paramètres a et b (ou plus). On aurait pu chercher la *parabole des moindres carrés* pour les nuages de points des figures 52. Mais on comprend bien que cette parabole n'aurait aucune pertinence : c'est flagrant pour la figure 52.8, quoiqu'il puisse y avoir un doute pour la figure 52.1. En effet, que la parabole qui rend minimum la somme des carrés des écarts existe est une chose ; que ce minimum soit petit en est une autre. C'est-à-dire que la parabole des moindres carrés a aussi peu de pertinence dans la figure 52.8 que la droite de régression n'en avait pour la figure 50.

L'idée de la régression *non* linéaire est d'effectuer cette généralisation : au lieu de chercher la droite de régression d'un nuage de points, chercher une courbe de régression. Mais il n'existe évidemment aucune recette permettant de déterminer le bon type de courbe, qu'il n'y aurait ensuite plus qu'à ajuster. Si on a une raison de penser que le nuage de points suit une parabole plutôt que n'importe quel autre type de courbe, on peut poser pour la fonction $Y = f(X) = aX^2 + bX + c$ (tout comme on avait posé $f(X) = aX + b$ pour la régression linéaire), puis déterminer les coefficients a , b , et c de manière à rendre minimum la somme des carrés des écarts. Cherchons cette parabole des moindres carrés.

La somme des carrés des écarts est

$$S = \sum_{i=1}^n p_i [y_i - f(x_i)]^2 = \sum_{i=1}^n p_i [y_i - ax_i^2 - bx_i - c]^2$$

où on a tenu compte du poids ou probabilité p_i de chaque point (x_i, y_i) : s'il s'agit de variables aléatoires X et Y , les p_i sont donnés par la loi conjointe, s'il s'agit d'un nuage empirique, on aura $p_i = 1/n$. En développant le carré et en introduisant les corrélations suivantes, qu'on peut appeler *corrélations non linéaires* :

$$M_{\alpha, \beta}(X, Y) = \sum_{i=1}^n p_i x_i^\alpha y_i^\beta$$

on obtient

$$\begin{aligned} S = & M_{0,2} + a^2 M_{4,0} - 2a M_{2,1} + 2ab M_{3,0} + \\ & + (b^2 + 2ac) M_{2,0} - 2b M_{1,1} - 2c M_{0,1} + 2bc M_{1,0} + c^2 \end{aligned}$$

Cette expression est minimum lorsque ses dérivées partielles par rapport à a , b , et c s'annulent. La condition (nécessaire) de minimum est donc le système de trois équations linéaires à trois inconnues a , b , c suivant :

$$\begin{aligned} a M_{4,0} + b M_{3,0} + c M_{2,0} &= M_{2,1} \\ a M_{3,0} + b M_{2,0} + c M_{1,0} &= M_{1,1} \\ a M_{2,0} + b M_{1,0} + c &= M_{0,1} \end{aligned} \quad (XII.9)$$

Si au lieu d'une parabole, on cherche la courbe de régression non linéaire sous la forme d'un graphe de polynôme de degré Q , soit $Y = P(X) = \sum a_j X^j$, on obtiendra un système de $Q + 1$ équations linéaires à $Q + 1$ inconnues a_j (les coefficients du polynôme) :

$$\sum_{j=0}^Q a_j M_{k+j,0} = M_{k,1} \quad (XII.10)$$

pour $k = 0, 1, 2, \dots, Q$.

Un tel système se résout numériquement par la *méthode du pivot* bien connue.

Rien n'impose d'ailleurs les polynômes et n'importe quelle fonction de la forme $Y = f(X) = \sum a_j \Phi_j(X)$, où les Φ_j sont des fonctions de base, conduiront à un système d'équations linéaires dont les inconnues seront les coefficients a_j . Mais la linéarité des conditions de minimum XII.9 et XII.10 provient évidemment du fait que la fonction f dépend elle-même linéairement des paramètres. Si on cherche la courbe de régression sous la forme $Y = f(X) = a e^{\alpha X}$, qui dépend des deux paramètres a et α (mais non linéairement en α), les conditions de moindre carré seront alors, et pour cause, un système de deux équations à deux inconnues, mais non linéaires.

On comprend que pour les calculs pratiques il vaut mieux chercher les courbes de régression sous une forme qui soit linéaire selon les paramètres.

Toutefois lorsqu'on a de bonnes raisons de chercher une dépendance non linéaire, il existe un procédé algorithmique efficace (du moins si le nombre de paramètres n'est pas trop élevé) connu sous le nom de *méthode de Levenberg-Marquardt*. Ce procédé est présenté dans l'annexe qui suit ce chapitre. L'algorithme de Levenberg-Marquardt est parfaitement robuste quand il n'y a que deux paramètres. Sa robustesse décroît ensuite avec le nombre de paramètres : au delà de six, l'algorithme n'est plus guère fiable, mais il est très rare que la dépendance soit non linéaire selon six paramètres ; presque toujours, on aura une dépendance linéaire par rapport à la plupart des paramètres, en sorte que seuls deux ou trois d'entre eux produiront la dépendance non linéaire. Il suffira alors d'éliminer préalablement les

premiers par la méthode linéaire du pivot, et traiter ensuite les deux ou trois paramètres restants par la méthode de Levenberg-Marquardt.

Le problème général de la dépendance n'est cependant pas résolu pour autant. Car si on peut ajuster les paramètres d'une famille de courbes, on ne sait pas comment choisir cette famille de courbes. Or la méthode de Levenberg-Marquardt ne résoud évidemment pas ce problème là ; elle permet seulement de calculer les paramètres optimaux *pour une famille donnée*.

Si la régression linéaire ne donne aucun résultat, en ce sens que la *droite* de régression correspond à un minimum *qui n'est pas petit* (ce qui se produit sur la figure 50), on peut chercher une *courbe* des moindres carrés qui correspond à un minimum petit ; on peut par exemple augmenter le degré du polynôme jusqu'à ce que le minimum devienne petit. En procédant ainsi, on est assuré du succès, puisqu'on sait à l'avance que pour un polynôme dont le degré est *égal* au nombre de points du nuage, la somme des carrés des écarts sera nulle (polynôme d'interpolation des points). Mais ce procédé, qui aboutira donc *forcément* au succès si le nombre de points du nuage est fini, fournira aussi une courbe de régression dépourvue de la moindre pertinence. La courbe doit représenter une loi, dont le nuage de points serait une expression bruitée ; si on interpole les points du nuage, et qu'on refait une autre série de mesures qui fourniront un autre nuage, la courbe d'interpolation aura changé. Or la *loi* du phénomène observé (si elle existe) est ce qui ne change pas d'une série de mesures à l'autre, le *bruit* est ce qui change.

On peut donc préciser comme suit les conditions dans lesquelles la régression non linéaire est recommandée. La régression *linéaire* est un problème bien posé parce que les fonctions linéaires, c'est-à-dire les rapports de proportionnalité, jouent un rôle privilégié dans l'étude des grandeurs. Mathématiquement la régression *non* linéaire n'est un problème bien posé que si on choisit a priori une famille paramétrée de fonctions. Le problème formulé ainsi : "parmi toutes les fonctions possibles et imaginables, laquelle minimise la somme des carrés des écarts ?" est évidemment absurde. En effet, parmi *toutes* les fonctions possibles et imaginables, il y en a forcément une, et même une infinité, qui rendent exactement nulle la somme des carrés des écarts. Si $P(x)$ est le polynôme d'interpolation des n points, pour lequel $y_i = P(x_i)$, on aura $S = \sum [y_i - P(x_i)]^2 = 0$. À partir de ce polynôme, on peut ensuite fabriquer une infinité de fonctions ayant la même propriété ; soit en effet $Q(x)$ une fonction égale à 1 lorsque x est égal à l'un des x_i , mais à n'importe quoi en dehors des valeurs x_i (rien n'empêche une telle fonction d'être différentiable ou analytique). Alors toute fonction $f(x)$ de la forme $P(x) \cdot Q(x)$ remplira également les conditions requises.

La régression non linéaire est donc une technique qui n'a de sens que lorsqu'on a déjà de bonnes raisons théoriques de postuler un type particulier de dépendance, et qu'on souhaite seulement en ajuster empiriquement les paramètres.

On appellera *modèle de dépendance* un choix particulier d'une famille paramétrée de fonctions; selon le nombre des paramètres, le modèle sera dit à n paramètres.

Mais il ne suffit pas que le problème soit *mathématiquement* bien posé grâce à la présence d'un modèle de dépendance. Il faut aussi que le modèle postulé soit pertinent. Pratiquer la régression non linéaire en utilisant la méthode de Levenberg-Marquardt, ou un logiciel qui la met en œuvre, est une simple technique. Trouver le modèle pertinent est une question d'imagination, mais surtout d'intuition et d'expérience, voire de génie. Car la difficulté n'est pas d'ajuster des paramètres pour minimiser la somme des carrés des écarts, ni de postuler un modèle plus ou moins arbitraire; la difficulté est de séparer ce qui constitue la loi du phénomène (qui reste invariable d'une série de mesures à l'autre), et ce qui constitue le bruit. On peut tester a posteriori un modèle de régression non linéaire par les deux critères suivants :

a) le modèle doit rester valide pour toutes les séries de mesures, c'est-à-dire que le minimum de la somme des carrés des écarts doit être petit pour *toutes* les séries de mesures que l'on effectue;

b) le bruit constitué par les écarts doit également conserver des caractères statistiques constants d'une série de mesures à l'autre. Par exemple si les écarts sont distribués selon une loi gaussienne (cas de loin le plus commun), et si l'écart-type empirique change notablement d'une série de mesures à l'autre, il faudrait conclure que la séparation entre loi et bruit effectuée par le modèle postulé n'était pas correcte.

Le critère b ne peut être négligé au profit du seul critère a . Imaginons par exemple qu'en étudiant la régression linéaire de deux variables X et Y , on constate que les écarts sont élevés pour X petit et deviennent de plus en plus faibles quand X augmente (voir par exemple figure 53.2). Lorsqu'on effectue des mesures il arrive bien plus souvent que les écarts augmentent proportionnellement à X au lieu de diminuer, car il est fréquent que les erreurs soient relatives plutôt qu'absolues. L'inverse est donc surprenant et devrait mettre la puce à l'oreille. Si les fluctuations sont causées (comme les résultats du chapitre **VII** le font soupçonner) par d'innombrables petites perturbations aléatoires indépendantes qui s'accumulent, on comprend difficilement pourquoi ces perturbations s'évanouissent lorsque X devient grand. Il s'est

donc produit un phénomène qui est intéressant par lui-même ⁽¹⁾. Si on veut analyser la situation de manière complète, la seule droite de régression ne suffit pas, il faut inclure une prédiction concernant la variation éventuelle des écarts. On appellera donc *modèle de régression* (linéaire ou non) un modèle de dépendance (famille paramétrée de fonctions) accompagné d'une prédiction concernant le bruit (par exemple que le bruit est gaussien et que son écart-type est proportionnel à X). En pratique les modèles de régression non linéaire sont rarement autres que gaussiens et incluent toujours un écart-type constant ou proportionnel à X , car les phénomènes du type de la figure 53.2 sont des artefacts, comme nous le verrons plus loin.

L'importance du critère b intervient encore sous un autre aspect. Dans des études expérimentales, on ne peut pas donner à la variable X des valeurs arbitrairement grandes ou arbitrairement petites; les valeurs accessibles sont en effet toujours limitées par les possibilités pratiques. C'est pourquoi il n'est pas toujours facile de distinguer une dépendance exponentielle d'une dépendance polynômiale: par exemple $e^X \simeq 1 + X + \frac{1}{2}X^2$; on pourrait aisément départager l'exponentielle du polynôme, même si le nuage est très flou, si on pouvait faire tendre X vers l'infini. Mais si les valeurs mesurables de X sont limitées, on ne le peut pas. Existe-t-il alors un moyen statistique de remarquer la différence? La réponse est oui: supposons par exemple que le bruit soit gaussien et qu'on le constate pour les petites valeurs de X ; supposons en outre que le phénomène soit exponentiel "en réalité", mais qu'on ait choisi un modèle de dépendance de la forme $a + bX + cX^2$, qui après ajustage par moindres carrés donne $a = 1$, $b = 1$, $c = \frac{1}{2}$. Alors les écarts ne seront plus exactement gaussiens pour les grandes valeurs de X , à cause d'une erreur systématique sur la moyenne. Le modèle exponentiel séparerait mieux le signal du bruit que le modèle polynômial.

Ainsi l'analyse statistique du seul bruit permet aussi de détecter si le modèle de dépendance est correct.

Ce qui implique que si l'un des deux critères a et b manque, le modèle de régression doit être revu. Mais il n'existe pas de méthode pour trouver un modèle. La Statistique ne fournit que des critères de vérification a posteriori.

Un point doit encore être précisé. Lorsque les mesures donnent des écarts faibles, c'est-à-dire que le bruit est très petit par rapport au signal, que le nuage de points est concentré au voisinage immédiat d'une courbe qui saute aux yeux, comme sur la figure 52.8, alors les techniques statistiques de régression sont évidemment superflues. Ces techniques deviennent utiles lorsque le bruit est important (ou lorsqu'il varie nettement avec l'une des

⁽¹⁾ Pour l'explication d'un tel phénomène, voir plus loin dans cette section.

variables, comme sur les figures 53 à 57), c'est-à-dire lorsque le nuage est flou, mais qu'on a de bonnes raisons de penser qu'il existe une loi noyée dans ce bruit, et qu'il s'agit de la découvrir. Comme nous avons pu le constater avec les tests statistiques (chapitre **XI**), la Statistique est un outil qui devient inutile quand les fluctuations aléatoires sont faibles. Il est inutile d'appliquer le test du χ^2 pour vérifier l'équilibre d'un dé si on peut le lancer un million de fois et donc obtenir des fluctuations de l'ordre de 0.1% ; ce test ne prend son sens que si, n'ayant jeté le dé que cent fois, les fluctuations sont trop fortes pour donner un résultat flagrant. De même, les méthodes de régression ne sont utiles que si l'amplitude des fluctuations, c'est-à-dire l'épaisseur du nuage, est si grande qu'elle efface la loi du phénomène. Ces méthodes statistiques n'apportent cependant pas de certitude quant à la loi en question ; elles ne doivent pas dispenser d'une étude plus exacte du phénomène (tout comme un alcootest ne doit pas dispenser d'une prise de sang), mais permettent tout au plus d'économiser du temps et des moyens en évitant de retenir pour une étude plus approfondie un modèle déjà incompatible avec les premières données.

Il faut aussi signaler que la recherche d'une *courbe* de régression se ramène le plus souvent à la régression linéaire après un changement de variable adéquat. En pratique ce procédé est infiniment plus courant que la recherche directe des moindres carrés à partir d'un modèle non linéaire : il est tout particulièrement répandu sous la forme du papier logarithmique, analogue au papier millimétré, sauf que les abscisses et les ordonnées sont toutes deux graduées logarithmiquement et non arithmétiquement. Ainsi une dépendance de la forme $Y = aX^\alpha$, c'est-à-dire une fonction puissance, se traduit sur papier logarithmique par une droite de pente α et d'ordonnée à l'origine $\log(a)$, puisque les nouvelles variables $\log(X)$ et $\log(Y)$ sont liées par la dépendance linéaire $\log(Y) = \alpha \log(X) + \log(a)$. Par conséquent, au lieu de chercher une courbe de régression de la forme $Y = aX^\alpha$ en déterminant a et α par les moindres carrés (ce qui est d'autant plus difficile que le paramètre α n'intervient pas linéairement), il est bien plus pratique et plus parlant de chercher la droite de régression de $\log(Y)$ par rapport à $\log(X)$. Mais si on cherche à analyser l'écart lui-même, par exemple pour savoir si l'écart-type est constant en valeur relative ($\Delta X/X$) ou au contraire en valeur absolue (ΔX), il peut être préférable d'employer la méthode directe.

Afin d'illustrer cela, on a donné sur les figures 53 à 57 divers exemples de telles transformations non linéaires. On peut constater que si sur le nuage avant transformation, les écarts sont homogènes, soit en valeur absolue (ΔX), soit en valeur relative ($\frac{\Delta X}{X}$), il n'en est plus du tout de même après

transformation. Il est aisé de donner une description mathématique de ces transformations: supposons que $Y = f(X) + \varepsilon$, où f est une fonction non linéaire et ε une variable aléatoire représentant le bruit. Les points de coordonnées $(x_i, y_i = f(x_i) + \varepsilon_i)$ sont les points du nuage.

Représentons maintenant non plus les points (x_i, y_i) , mais les points $(x_i, g(y_i))$, où $x = g(y)$ est la fonction inverse de $y = f(x)$. Si ε reste assez petit, on peut faire un développement limité de $g(y_i) = g(f(x_i) + \varepsilon_i)$ en puissances de ε_i :

$$\begin{aligned} g(y_i) &\simeq g(f(x_i)) + g'(f(x_i))\varepsilon_i + \frac{1}{2} g''(f(x_i))\varepsilon_i^2 \\ &= x_i + \frac{1}{f'(x_i)} \varepsilon_i - \frac{1}{2} \frac{f''(x_i)}{f'(x_i)^3} \varepsilon_i^2 \end{aligned}$$

On a utilisé les dérivées connues de la fonction inverse: $g'(f(x)) = 1/f'(x)$ et $g''(f(x)) = -f''(x)/f'(x)^3$. On en conclut que la dépendance non linéaire entre Y et X devient une dépendance linéaire entre $g(Y)$ et X , mais la répartition des écarts aléatoires $(1/f'(x))\varepsilon$ (le bruit) subit une distorsion. Si l'écart-type de ε est constant, alors dans la nouvelle représentation il variera en $1/f'(X)$. Si $f'(x)$ est croissante, le nouveau nuage de points sera flou pour les petits x , mais se resserrera au voisinage de la droite de régression $y = x$ (voir figure 53.2). Si $f'(x)$ est décroissante, le nuage s'évasera pour les grandes valeurs de x . Le bruit sera réparti de façon inhomogène le long de la droite de régression.

Pour reprendre une remarque faite plus haut: si dans une série d'observations on constate un tel phénomène, on peut en déduire que les variables X et Y ne sont pas rapportées à la bonne représentation. Une transformation non linéaire adéquate des coordonnées rétablira l'ordre.

Lorsque le bruit apparaît en valeurs relatives, le même développement limité conduit à un bruit de la forme $(f(x)/f'(x))\varepsilon$ après transformation (figures 54.1 et 54.2).

Supposons qu'on recherche par la méthode des moindres carrés sur les paramètres a et α le polynôme $y = ax^\alpha$ de régression du nuage de points de la figure 53.1, puis la droite de régression du nuage de points de la figure 53.2 (ou bien la même chose pour les figures 54.1 et 54.2). Techniquement cela est plus compliqué que si le paramètre α intervenait linéairement, mais ce n'est qu'une affaire de programmation facile et sans importance pour le propos qui suit. Il n'y a aucune raison pour que la droite soit exactement la transformée du polynôme par $g(Y) = \sqrt[3]{Y}$ (toutefois elle en sera assez proche si le bruit ε est petit). C'est-à-dire que le procédé d'optimisation par les moindres carrés n'est pas covariant dans la transformation. Il ne revient

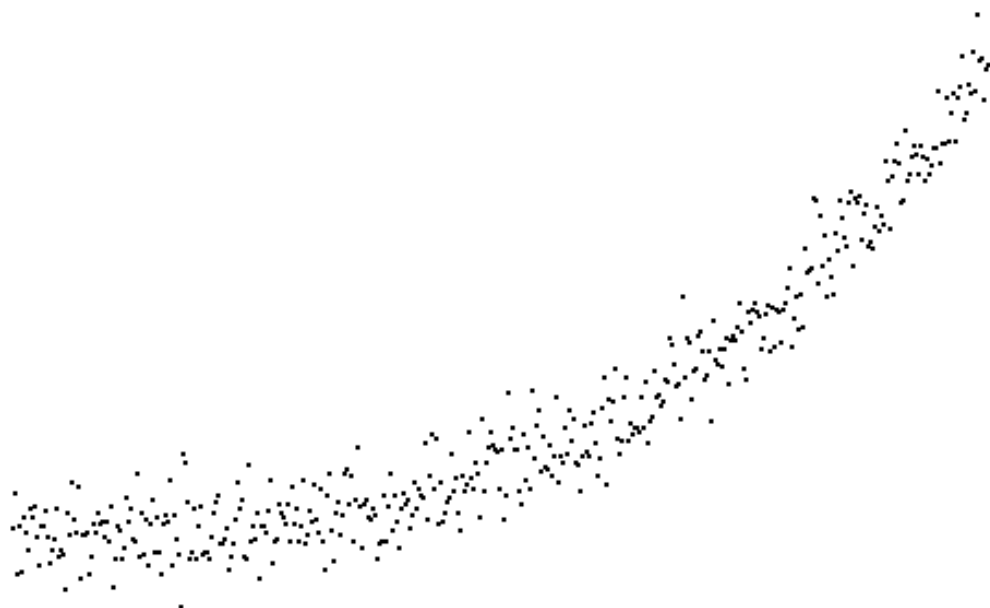


figure 53.1

$$y = ax^3 + \varepsilon$$



figure 53.2

$$y = \sqrt[3]{ax^3 + \varepsilon}$$

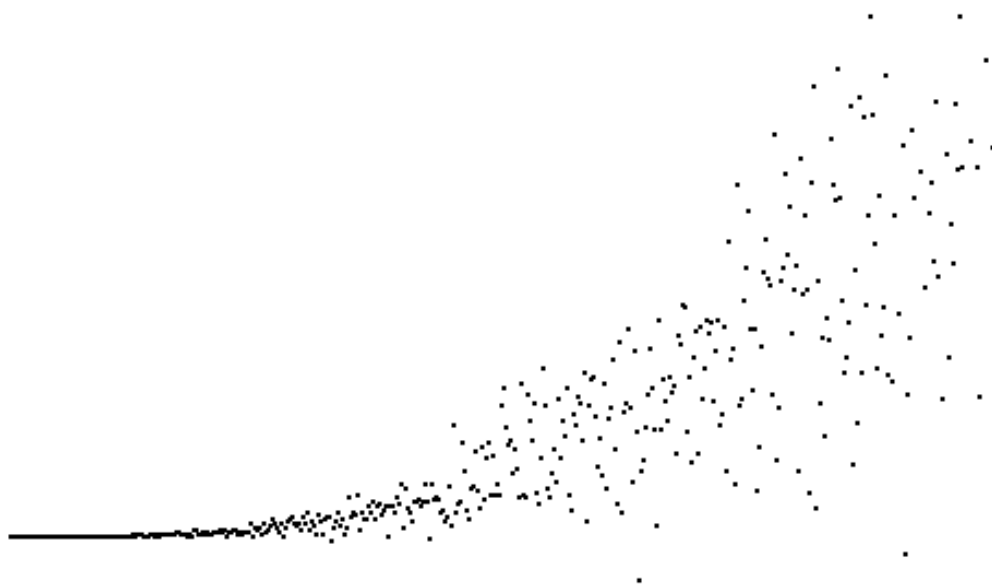


figure 54.1

$$y = ax^3 \cdot (1 + \varepsilon)$$



figure 54.2

$$y = \sqrt[3]{ax^3 \cdot (1 + \varepsilon)}$$

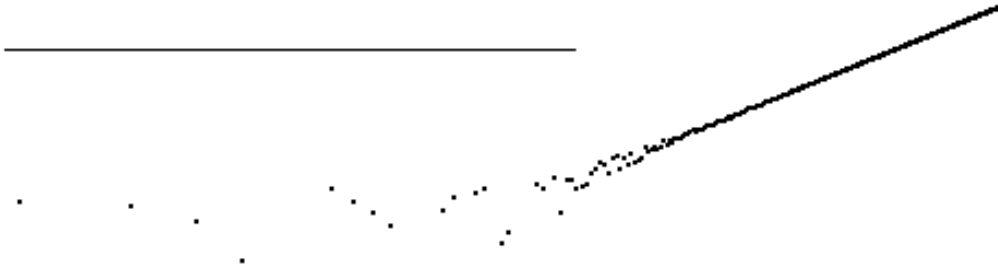


figure 55

$\log y = \log(ax^3 + \varepsilon)$ avec ε extrêmement petit. Le trait horizontal est l'axe des $\log x$ (correspondant à $y = 0$). La pente de la droite de régression est 3, mais l'axe des $\log y$ a été comprimé pour une meilleure mise en page.

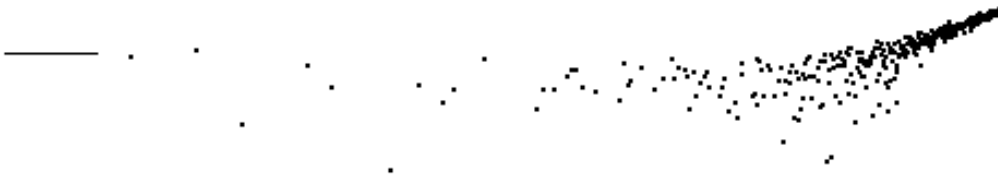


figure 56

$\log y = \log(ax^3 + \varepsilon)$ avec ε modérément petit.

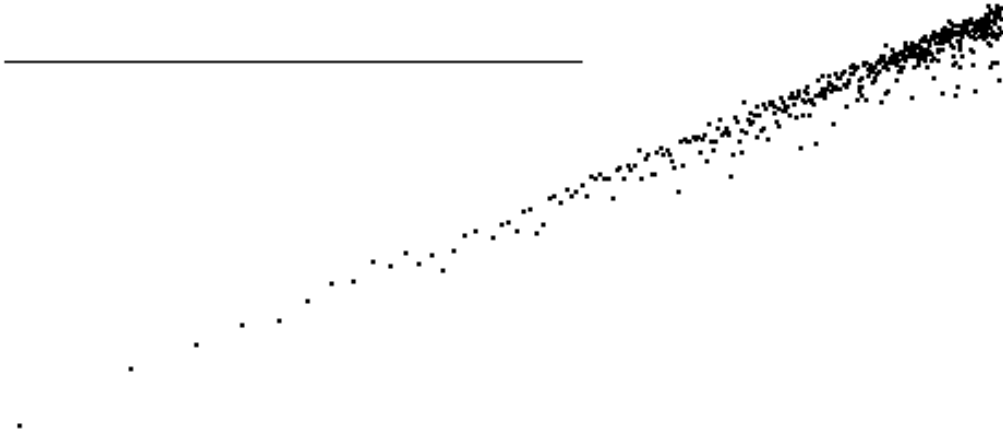


figure 57

$\log y = \log(ax^3 \cdot [1 + \varepsilon])$ avec ε modérément petit. Le trait horizontal est l'axe des $\log x$ (correspondant à $y = 0$).

au même, ni de rechercher la loi du phénomène, ni d'analyser le bruit dans les deux représentations, surtout si le bruit est de forte amplitude.

On ne doit donc pas conclure que dans n'importe quel cas “il suffit de passer à une représentation où la dépendance est linéaire”. Si le bruit est une perturbation d'un signal, on ne peut pas comprendre que sa dispersion varie bizarrement selon la valeur du signal, si ce n'est en admettant que la représentation est artificielle. En comparant par exemple les figures 53.1 et 53.2, on sera conduit à conclure que la répartition du bruit en 53.1 est naturelle, mais que celle de 53.2 ne l'est pas. La représentation dans laquelle la dépendance est linéaire est donc “artificielle”. Cela montre bien que la régression non linéaire n'est pas équivalente à une régression linéaire après transformation.

On peut aussi faire subir aux deux coordonnées à la fois une transformation non linéaire; c'est ce que par exemple on effectue automatiquement lorsqu'on reporte des mesures sur papier logarithmique. Dans ce cas, au lieu de représenter X en abscisse et Y en ordonnée, on représente $\log(X)$ en abscisse et $\log(Y)$ en ordonnée. Cette représentation logarithmique est particulièrement indiquée pour les fonctions puissance $f(x) = aX^\alpha$:

— si $Y = aX^\alpha + \varepsilon$ (écart absolu):

$$\ln(Y) = \alpha \ln(X) + \ln(a) + \ln\left(1 + \frac{\varepsilon}{aX^\alpha}\right)$$

— si $Y = a X^\alpha(1 + \varepsilon)$ (écart relatif) :

$$\ln(Y) = \alpha \ln(X) + \ln(a) + \ln(1 + \varepsilon)$$

Si ε est petit et représente un écart absolu uniformément distribué, on aura dans la représentation logarithmique un bruit en $\varepsilon/a X^\alpha$, qui est donc fortement variable le long de la droite de régression. Si ε est un écart relatif, la transformation donnera un écart absolu égal à ε .

Le résultat est montré sur les figures 55, 56, et 57 pour $\alpha = 3$. La figure 55 doit être comparée à 53.1 : elle représente le même nuage de points, mais sur papier logarithmique. De même 56 doit être comparée à 54.1.

Ces figures illustrent très nettement le phénomène ; si dans la représentation “naturelle”, le bruit est distribué uniformément, alors dans la nouvelle représentation où la dépendance est rendue linéaire, l’inhomogénéité des écarts peut être considérable. Cette idée de représentation “naturelle” est à rapprocher de la discussion à la fin de la section **IX.1**. La répartition du bruit est en effet liée aux invariances naturelles de la Physique (spatiale et temporelle). Prenons pour exemple le cas illustré par les figures 53.1 et 55 où le bruit ε est un écart absolu uniforme, de densité gaussienne. La dépendance de y par rapport à x est alors telle que $y = ax^3 + \varepsilon$. On peut dire que dans cette représentation “naturelle”, l’écart de y par rapport à ax^3 est la somme d’un grand nombre de perturbations stochastiquement indépendantes, puisque c’est généralement ainsi que sont engendrées les fluctuations gaussiennes. Pour que des perturbations puissent s’ajouter, il faut qu’elles représentent des quantités additives, par exemple des déplacements dans l’espace. Notons que des erreurs de mesure se ramènent toujours en dernière instance à des sommes de déplacements dans l’espace ; par exemple les tremblements des aiguilles d’instruments en face d’une graduation sont bien des fluctuations spatiales (les affichages par cristaux liquides rendent cela moins flagrant, mais ce n’est qu’un masque et on sera toujours ramené à l’espace et au temps par une analyse plus poussée). Si le bruit est distribué uniformément, c’est-à-dire si l’écart-type de ε est constant (indépendant de x), cela a une signification concrète : c’est que les nombreux effets indépendants qui contribuent à créer la fluctuation gaussienne ne sont pas influencés par la valeur de x . L’uniformité du bruit est donc le reflet d’une invariance occulte.

En représentant alors les variables x et y sur papier logarithmique, on fait disparaître, ou du moins on occulte davantage cette invariance. On voit bien que dans ce cas la représentation $y = ax^3 + \varepsilon$ est réellement plus “naturelle” que la représentation logarithmique, de même que les repères galiléens sont réellement plus “naturels” que les autres.

En conclusion, le papier logarithmique (ou n'importe quel procédé faisant apparaître artificiellement une dépendance linéaire) est certes un moyen commode pour faciliter la mise en évidence des dépendances, mais il ne faut jamais oublier que la structure du bruit peut elle aussi contenir des informations essentielles. L'étude des corrélations (linéaires) sur le graphique 55 effacerait ces informations; seule l'étude de la régression non linéaire sur le graphique 53.1 permet de les obtenir.

XII. 5. Comment trouver le bon modèle : l'exemple de Planck.

Lorsqu'on est confronté à des résultats de mesures ou d'expériences qui, reportées sur un papier millimétré ou logarithmique se traduisent par un nuage de points, la grande difficulté n'est pas d'appliquer les techniques de Statistique, mais de trouver le bon modèle. Cela est encore plus vrai aujourd'hui (où le travail purement technique de mise en oeuvre des recettes statistiques est entièrement effectué par des logiciels spécialisés) qu'autrefois. Les logiciels de régression non linéaire les plus performants attendent de l'utilisateur qu'il choisisse un *modèle de régression*. Si l'utilisateur n'en propose pas de pertinent, les résultats des calculs seront grossiers et peu significatifs, et la sophistication du logiciel ne compensera pas cette lacune (tout au plus elle servira à mieux tromper les profanes, objectif certes essentiel à notre époque, mais hors de notre sujet). On peut écrire des manuels sur les techniques statistiques, mais enseigner l'intuition, l'imagination, l'astuce, est impossible. Le présent ouvrage ne le prétend pas non plus. Toutefois, cet ouvrage étant voué à la signification et à la compréhension des concepts scientifiques, ne peut laisser de côté ce qui constitue justement *la seule démarche porteuse de sens et de compréhension*, à savoir la création de modèles pertinents. C'est pourquoi nous consacrons une section à présenter un exemple particulièrement fécond : la découverte de la Physique quantique, Berlin, 1899⁽¹⁾.

À la fin du chapitre **II** nous avons rencontré la loi de Planck, qui s'exprime par une fonction du type

$$f(\nu) = \frac{A\nu^3}{e^{\frac{\alpha\nu}{T}} - 1} \quad (XII. 11)$$

où ν est la fréquence et T la température absolue du rayonnement. Les constantes A et α étaient liées à la constante de Planck $\hbar$ (voir **II. 6**). Cette

⁽¹⁾ Nous suivons en gros l'exposé original de Max Planck à la Société allemande de Physique (séance du 19 octobre 1900 : *Über eine Verbesserung der Wienschen Spectralgleichung* Verhandlungen der Deutschen Physikalischen Gesellschaft, Band 2, 1900, pages 237 – 245.

loi de dépendance de l'énergie rayonnée par le corps noir en fonction de la fréquence était inconnue avant Planck. Planck ne l'a pas déduite comme nous de raisonnements probabilistes faisant appel à la Mécanique quantique, c'est au contraire la Mécanique quantique qui a été déduite de la loi. Il l'a obtenue directement à partir de lois empiriques antérieures à la Physique quantique. Dans ces lois classiques il n'y avait évidemment nulle trace de la constante $\hbar$, qui est justement issue de ce travail.

Ce qui était connu avant étaient les deux cas asymptotiques suivants :

a) lorsque la fréquence ν est petite

$$f(\nu) = \frac{8\pi\nu^2}{c^3} k T \quad (XII.12)$$

où c est la vitesse de la lumière et k la constante de Boltzmann. Cette loi portait le nom de Rayleigh–Jeans⁽¹⁾ ; elle avait été obtenue par voie théorique à partir des lois classiques du rayonnement, mais ne se vérifiait expérimentalement que dans l'infra-rouge, c'est-à-dire pour les petites fréquences.

b) lorsque la fréquence ν est grande

$$f(\nu) = A \nu^3 e^{-\frac{\alpha\nu}{T}} \quad (XII.13)$$

appelée loi de Wien, plutôt empirique, et résultant des observations sur les rayonnements de haute fréquence (ultra-violet) ; les constantes A et α étaient empiriques et non (comme pour la loi de Rayleigh–Jeans) exprimées en fonction de constantes déjà connues. On ne savait pas justifier théoriquement la loi de Wien ; c'était une énigme de la Physique.

Peut-on deviner la loi *XII.11* à partir de *XII.12* et *XII.13* ? Si on pose à un mathématicien le problème : "trouver une fonction qui se comporte comme $\frac{1}{x}$ pour x petit et comme e^{-x} pour x grand, il est probable qu'après un temps de réflexion assez court il propose la fonction $1/(e^x - 1)$. Il est donc possible de deviner. Toutefois si on lit les publications de Planck à cette époque charnière de la Physique (1899 – 1900) on constatera qu'il n'a pas procédé ainsi. Il a commencé par chercher quelle expression pour l'entropie S du rayonnement en fonction de la grandeur $U = c^3 f(\nu)/8\pi\nu^2$ impliquaient les relations *XII.12* et *XII.13*, sachant que $\frac{dS}{dU} = \frac{1}{T}$. La grandeur U avait dans ce contexte un sens particulier, qu'il est trop long de

⁽¹⁾ En toute rigueur historique, la loi de Rayleigh–Jeans n'était pas encore connue officiellement au moment où Planck réfléchissait au problème (1899) puisque l'article de Rayleigh où elle fut présentée a paru en juin 1900, mais elle était connue "officieusement". Ce détail historique est toutefois sans intérêt pour notre propos.

développer ici. Mais si on réécrit $XII.12$ et $XII.13$ pour la grandeur U , cela donne respectivement :

$$U = kT \quad (XII.12a.)$$

pour la loi de Rayleigh-Jeans et

$$U = b\nu e^{-\frac{\alpha\nu}{T}} \quad (XII.13a.)$$

où $b = c^3 A / 8\pi$.

On déduit aisément de $XII.12a$ que $\frac{1}{T} = \frac{k}{U}$ d'où

$$\frac{d^2 S}{dU^2} = -\frac{k}{U^2}$$

et de $XII.13a$ que $\frac{1}{T} = -\frac{1}{\alpha\nu} \ln(U/b\nu)$ d'où

$$\frac{d^2 S}{dU^2} = -\frac{1}{\alpha\nu} \cdot \frac{1}{U}$$

Ce détour par l'entropie était nécessaire pour Planck ; sa spécialité était la Thermodynamique et il pensait pouvoir *comprendre* la vraie nature du phénomène par ce biais car (dit-il dans son autobiographie scientifique) "là, il se trouvait en terrain connu". En introduisant la grandeur $R = 1/\frac{d^2 S}{dU^2}$, on obtient dans le cas correspondant à $XII.12a$

$$R = -\frac{1}{k} U^2$$

et dans le cas correspondant à $XII.13a$

$$R = -\alpha\nu U$$

Autrement dit, la loi de Rayleigh-Jeans (pour les petites fréquences) conduisait à une dépendance quadratique entre R et U , tandis que la loi de Wien (pour les hautes fréquences) conduisait à une dépendance linéaire.

Planck proposa alors d'essayer la combinaison la plus simple de ces deux formules, sous la forme

$$R = -\frac{1}{k} U^2 - \alpha\nu U \quad (XII.14)$$

de sorte que si ν est petite, le second terme devient négligeable et le premier prédomine, par contre si ν est grande c'est l'inverse. Il suffit alors de revenir en arrière pour retrouver U en fonction de T :

$$\frac{d^2 S}{dU^2} = \frac{1}{R} = -\frac{1}{\frac{1}{k} U^2 + \alpha\nu U} \quad (XII.15)$$

La primitive de cette fonction de U est $\frac{1}{\alpha\nu} \ln(1 + b\nu/U)$, d'où on déduit que

$$\frac{1}{T} = \frac{dS}{dU} = \frac{1}{\alpha\nu} \ln \left(1 + \frac{b\nu}{U} \right) \quad (XII.16)$$

Ceci est une expression de $1/T$ en fonction de U , qu'on peut inverser pour obtenir une expression de U en fonction de $1/T$, qui est

$$U = \frac{b\nu}{e^{\frac{b\nu}{T}} - 1} \quad (XII.17)$$

On constate que Planck a largement utilisé de la relation thermodynamique $\frac{dS}{dU} = \frac{1}{T}$. En comparant XII.17 avec II.14 on voit que α est égal à $2\pi\hbar/k$ (rappelons que la fréquence ν n'est pas égale à la pulsation ω , mais à $\omega/2\pi$). Les physiciens de 1900 ne pouvaient pas faire cette comparaison. Mais ils savaient que le rayonnement de corps noir est indépendant des matériaux qui constituent la cavité, et que par conséquent les constantes b et α devaient être (comme c ou k) des constantes universelles de la Physique.

C'est ainsi que $\hbar$ apparut pour la première fois (d'où son nom de *constante de Planck*).

Cette histoire est intéressante pour nous ici parce qu'elle montre comment on devine un modèle. Planck n'aurait jamais trouvé l'explication fondamentale du rayonnement du corps noir s'il avait cherché un polynôme qui approche la courbe empirique à toutes les fréquences. Si l'ambition de la Physique consistait à approcher les courbes empiriques par des modèles mathématiques dépourvus d'une signification plus profonde, et ayant pour seule vertu de se rapprocher le plus possible des courbes fournies par l'expérience, ce ne serait plus la Physique et Planck n'aurait pas découvert $\hbar$ (ni Newton la gravitation, etc). La vraie nature de la démarche de Planck décrite ci-dessus est d'avoir attribué a priori aux deux lois empiriques XII.12 et XII.13 un sens plus profond que la simple description des faits expérimentaux. Au départ de sa démarche figurait la conviction intime que les courbes expérimentales sont des apparences, sous-tendues par des causes *intelligibles* qu'on ne peut atteindre que par la pensée (cf. Platon, *La République* Livre **VII**); sa tactique n'a pas consisté à approcher coûte que coûte la courbe expérimentale pour toutes les fréquences, mais à partir du principe (dicté par une intime conviction et non par l'observation) que la *cause intelligible* du phénomène se révélerait mieux dans les deux cas asymptotiques $\nu \rightarrow 0$ et $\nu \rightarrow \infty$ que dans le cas intermédiaire des fréquences moyennes. Le passage par l'entropie S est la meilleure preuve que son approche fut bien celle-là; mais il en témoigne aussi lui-même dans nombre de textes.

Un point essentiel de la démarche (revendiqué par Planck) est également le passage par les relations les plus simples possibles. Deviner la loi générale à partir des deux cas asymptotiques directement sous la forme *XII.12* et *XII.13* eût été possible comme nous l'avons indiqué plus haut. Mais Planck a préféré passer par l'intermédiaire de $R = 1/\frac{d^2S}{dU^2}$ car les relations atteignaient alors le maximum de simplicité (relation linéaire ou quadratique) et surtout, le maximum de **signification**. Le procédé est toujours le même : passer par l'intelligible plutôt que par une description purement empirique.

On peut tirer de cet exemple magistral deux leçons essentielles à retenir lorsqu'on doit chercher un modèle. Ces leçons sont valables même quand on ne nourrit pas l'ambition de révolutionner la Physique :

a) Ne pas chercher simplement à approcher le nuage de points ou la courbe empirique, mais réfléchir d'abord et essayer de trouver à quelle nécessité purement intelligible le modèle doit obéir (symétries ou invariances, analogie avec des situations connues, nécessités logiques). Nous avons déjà fait appel avec insistance à ce type de démarche dans le présent ouvrage, en indiquant que pour résoudre un problème de probabilité "il faut commencer par chercher ce qui est équiprobable").

b) Effectuer sur les variables qui sont en jeu des transformations qui ramènent autant que possible aux types de dépendance les plus simples : proportionnalité ou à la rigueur dépendance quadratique. Ainsi une dépendance du type $Y = aX^\alpha$ devient une régression linéaire entre $\log(Y)$ et $\log(X)$ (papier logarithmique), une dépendance du type $Y = ae^{\alpha X}$ devient une régression linéaire entre $\log(Y)$ et X (papier semilogarithmique).

Bien entendu, pour découvrir la Mécanique quantique, il a fallu pousser l'imagination un peu plus loin, mais le principe reste le même.

XII. 6. Corrélation et causalité.

L'idée essentielle qui se cache derrière ces problèmes de corrélation ou de régression est celle de la *causalité*. Dans le langage courant, y compris les jargons techniques (et même *surtout* dans les jargons techniques) la différence entre la dépendance statistique et la causalité n'est pas clairement définie. C'est pourquoi nous allons maintenant discuter du véritable sens des différentes notions introduites dans ce chapitre, à savoir la corrélation ou la régression. Cette discussion complétera les discussions déjà amorcées au chapitre **X** sur la signification des mesures statistiques. Nous avons alors bien insisté sur la différence entre la mesure d'une probabilité a priori

(cas de l'expérience reproductible) et la mesure d'une proportion dans une population (cas du sondage), les deux pouvant parfois se recouvrir. Ici, nous ne discutons plus les mesures simples, mais la dépendance.

Nous pouvons en effet interpréter une forte corrélation, par exemple dans le cas extrême où le coefficient de corrélation est égal à $+1$ ou -1 , comme une dépendance linéaire. Une telle dépendance n'est pas flagrante dans le cas de la corrélation entre la tension artérielle et l'âge, car cette corrélation est médiocre, et correspond à peu près à ce qu'on peut voir sur les figures 52.2 ou 52.3. La situation serait beaucoup plus claire avec un nuage de points tel que celui de la figure 52.7 ou 52.8. Ce type de figure peut vous rappeler un T.P. d'électricité où il s'agissait de vérifier la loi d'Ohm : supposons qu'on reporte la tension électrique U entre les bornes de la résistance en abscisse et l'intensité I du courant en ordonnée ; on obtiendrait bien des graphiques de ce type.

Les lois de l'électromagnétisme sont souvent enseignées aujourd'hui de façon déductive. Les équations de Maxwell sont postulées comme s'il s'agissait d'une vérité mathématique a priori, dont on déduit tout le reste ; la résistance d'un conducteur n'est alors qu'un effet macroscopique sur un nombre énorme de molécules et d'électrons et on retrouverait la loi d'Ohm, en expliquant "tout" par moyennisation à partir des seules équations fondamentales de Maxwell. Mais le cheminement historique va en sens inverse, et les équations de Maxwell ont été peu à peu *dégagées* par induction à partir d'observations expérimentales, telles que justement cette loi d'Ohm qui a été établie par Ohm en faisant des mesures de tension électrique et d'intensité.

Si on effectue réellement ces mesures, ce qu'on obtient est un tableau de chiffres ou, si on les représente graphiquement, un nuage de points comme celui de la figure 52.8. Si on calcule le coefficient de corrélation du nuage on obtiendra une valeur proche de 1 et une droite de régression dont la pente est appelée la résistance R du conducteur. $U = RI$ est l'équation de la droite de régression.

Toutes les lois de la physique ont été, soit obtenues directement par un tel procédé, soit déduites ou induites mathématiquement à partir de tels procédés. Ce n'est que lorsque la forme mathématique issue de tout ce travail devient vraiment abstraite (par exemple lorsqu'elle prend la forme des équations de Maxwell) qu'on croit pouvoir lui donner une valeur de vérité supérieure, qui serait plus exacte que ce que la variance des mesures expérimentales laisse supposer. On dit alors que les lois abstraites sont les "vraies" lois (par exemple $U = RI$) et que les petits écarts du nuage de points par rapport à cette loi sont "du bruit". Mais cette légende sur

la nature des lois physiques ne doit pas faire oublier qu'au départ, toute l'information a été recueillie dans des graphiques tels que ceux de la figure 52.8. Il n'est pas difficile de comprendre le procédé littéraire qui permet de passer d'un nuage de points comme celui de la figure 52.8 à une loi mathématique comme les équations de Maxwell : la première étape consiste à dire que U est *égal* à RI au lieu de dire que la variance d'échantillon de $U - RI$ est petite.

Mais dans la loi d'Ohm il y a même bien plus que la seule relation mathématique $U = RI$. On dit en effet que la tension électrique appliquée aux bornes est la *cause* du courant.

Cette idée de causalité est tout à fait évidente si on conçoit le courant électrique selon l'électromagnétisme moderne. En effet, le courant électrique est un déplacement d'électrons dans le conducteur, et les électrons ne se déplacent que si on leur applique un champ électrique. Le champ électrique est alors la cause du mouvement des électrons comme le champ de gravité est la cause de la chute des corps. Or on crée un tel champ dans le conducteur en appliquant une tension électrique aux bornes, ce qui veut dire que la tension électrique est la cause du champ et donc aussi la cause du courant électrique.

Mais une simple dépendance ne suffit pas à déterminer ce qui est la cause et ce qui est l'effet. Par exemple, on aurait tout aussi bien pu écrire $I = U/R$, cela n'aurait pas pour autant fait de I la cause et de U l'effet. On voit donc qu'une corrélation statistique, même très forte, ne peut pas à elle seule décider ce qui est la cause ou ce qui est l'effet. La simple constatation d'une corrélation entre le tabagisme et l'hypertension ne suffit pas à certifier que le tabagisme est la cause et l'hypertension l'effet ; il se pourrait par exemple que l'hypertension ait une cause biologique qui favorise chez la même personne le goût pour le tabac.

La causalité est donc une relation plus forte que la dépendance statistique. Si on a observé sur un échantillon de mesures les valeurs de deux paramètres (tels que âge et tension artérielle, ou intensité et tension électrique), et que le nuage de points est distribué le long d'une courbe d'équation $y = f(x)$ de telle sorte que la variance de $y - f(x)$ sur l'échantillon de tous les points est petite, on dira qu'il y a une forte dépendance statistique entre les deux paramètres (si la fonction f est linéaire, on appellera cela une corrélation). Pour qu'en outre on puisse dire que le premier paramètre (celui qui est en abscisse) est la *cause* de l'autre, il faut que la réalisation du premier précède toujours la réalisation du second (en Relativité, la condition est encore plus forte : il faut que le premier soit réalisé avant le second, mais de sorte qu'un photon parti du premier immédiatement après sa réalisation

puisse arriver au second avant sa réalisation).

On peut résumer cela en disant que

$$\text{causalité} = \text{dépendance statistique} + \text{antériorité} \quad (XII.18.)$$

Au delà de cette définition de la causalité, on peut s'interroger sur le problème de savoir si la causalité se réduit entièrement à (XII.18.); si deux paramètres ont entre eux une forte dépendance statistique et un rapport d'antériorité, mais sans qu'on puisse pour autant *comprendre* la relation de nécessité qui les relie, peut-on encore parler de causalité? Cette question a été étudiée au *XVIII^e* siècle par David Hume⁽¹⁾ qui a répondu que la causalité ne peut pas être *plus* que ce qui est exprimé par (XII.18.) Quoique sa réponse ait été critiquée, je dois lui donner entièrement raison. Les critiques qui lui ont été opposées reposent essentiellement sur l'argument que voici : supposons que, chaque fois qu'un événement *A* se produit, un événement *B* se produit peu après, et *B* ne se produit jamais sans que *A* se soit produit juste avant ; on ne peut alors sérieusement tenir *A* pour la cause de *B* que si le rapport de nécessité entre *A* et *B* est clair pour l'entendement. C'est-à-dire que (XII.18.) ne suffit pas, il faut en outre un rapport de nécessité. Cette situation est illustrée sous forme comique dans *La vie criminelle d'Archibald de la Cruz*, un film de Luis Bunuel (1955), dont le personnage principal (Archibald de la Cruz) constate que, chaque fois que dans sa vie il a éprouvé de l'hostilité envers une personne, celle-ci est morte quelques heures ou quelques jours après. Il acquiert ainsi la certitude que, par un mécanisme inconnu, son sentiment de haine est la *cause* de la mort de ses victimes. Lorsque, torturé par les remords, il se livre à la police, personne ne le croit. Or, une situation toute analogue s'est produite lorsque Kepler a découvert les lois du mouvement des planètes. En effet, Kepler a constaté que le temps mis par une planète quelconque pour parcourir une portion de son orbite est proportionnel à l'aire balayée par le segment Soleil – planète (loi des aires). Quoique à cette époque les méthodes statistiques n'étaient pas encore aussi systématiquement quantitatives, ni surtout aussi sophistiquées mathématiquement qu'aujourd'hui, on peut néanmoins, avec un anachronisme certain, mais qui ne touche pas à l'essentiel, imaginer Kepler représentant les résultats de ses mesures et de ses calculs sur un graphique, avec les durées en ordonnée et les aires balayées en abscisse. Les points tracés ont bien la caractéristique de former un nuage concentré au voisinage d'une droite. Il y a donc une dépendance entre la vitesse et la distance au Soleil ; lorsque la planète s'approche du Soleil, elle va plus vite, et la loi des aires permet de calculer cette variation.

⁽¹⁾ David HUME *Enquête sur l'entendement humain* (1748).

Or, constatant cela, Kepler ne s'est pas contenté d'en tirer une loi purement descriptive ; voyant que le Soleil était au foyer de toutes les orbites des planètes, il a estimé que le Soleil ne pouvait pas ne pas être la *cause* du mouvement des planètes : en effet, les autres planètes ne pouvaient pas être tenues pour la cause du mouvement de l'une d'entre elles, car elles étaient interchangeables ; or par définition, la cause n'est pas interchangeable avec l'effet. Seul le Soleil pouvait jouer ce rôle. On voit bien que la simple corrélation constatée ne suffit pas à conclure à une causalité, il a fallu faire appel à un raisonnement a priori.

Mais ce n'est pas tout. La distance entre le Soleil et la planète étant énorme, comment la planète, lorsqu'elle passe à un endroit de son orbite, peut-elle "savoir" que, s'étant par exemple rapprochée du Soleil, il lui convient d'augmenter sa vitesse pour respecter la loi des aires ? D'autre part, la loi des aires est valable pour chaque planète séparément, et ne fait intervenir que le couple planète-Soleil, de sorte que le mouvement d'une planète est indépendant des autres planètes. Il faut dire que du temps de Kepler, on ne connaissait pas encore la gravitation, et à plus forte raison, on ne savait pas que chaque planète subit aussi l'attraction des autres planètes. La gravitation du Soleil étant de beaucoup la plus forte, et la précision des observations étant insuffisante, Kepler ne pouvait pas percevoir l'influence mutuelle des planètes. Si donc une des planètes disparaissait brusquement, cela ne devrait rien changer à la validité de la loi des aires pour celles qui restent. Donc le Soleil (c'est-à-dire la cause) agissait *sans savoir* si les planètes étaient là pour en subir les effets. Comment la corrélation observée est-elle possible s'il n'y a pas une communication entre le Soleil et la planète ? L'étude des textes laissés par Kepler montre qu'il avait été troublé par cette question et qu'il avait cherché une réponse. Son explication a été la suivante. Puisque le Soleil est la cause et que cette cause produit son effet à des centaines de millions de kilomètres, c'est que le Soleil doit émettre constamment un fluide qui se diffuse dans l'espace, et ce fluide, lorsqu'il arrive à la planète, lui donne de la vitesse. Kepler a alors tenté de calculer la loi d'action de ce fluide sur la planète, en partant du principe que, puisque le Soleil agit sans savoir s'il existe ou non des planètes, il doit émettre son fluide indépendamment des planètes, donc isotropiquement (c'est-à-dire uniformément dans toutes les directions). La quantité de fluide présente sur un petit morceau d'orbite est alors facile à calculer, mais ce qu'on trouve n'est pas proportionnel à la vitesse. Kepler, persuadé qu'en vertu d'une causalité nécessaire il *devait* y avoir quelque chose comme un fluide, n'a jamais réussi à faire coïncider cette idée avec les calculs quantitatifs, mais les textes montrent qu'il a beaucoup cherché. C'est Newton qui a trouvé la solution du problème, en disant que ce n'est pas un fluide (dont la densité

est forcément une grandeur scalaire) mais une force (qui est un vecteur) qu'il convenait d'invoquer; et qu'en outre, suivant Galilée, la force agit sur l'accélération et non sur la vitesse.

La moralité de cette histoire est la suivante: l'observation n'a jamais fourni que des corrélations, mais les corrélations brutes ne fournissent aucune *compréhension* du phénomène. Comprendre un phénomène signifie y distinguer des éléments qui sont des causes et d'autres qui sont des effets, et pour cela, l'entendement est obligé d'inventer de toutes pièces des mécanismes abstraits qui ne sont pas observables en tant que tels. Ainsi Archibald de la Cruz ne pouvait pas comprendre la corrélation entre ses fantasmes meurtriers et la mort violente de ses victimes sans faire appel à un fluide, ou à une force, émis par son esprit et se propageant dans l'espace, ou tout autre mécanisme bizarre déclenché dans son cerveau et finissant par produire le déplacement d'objets contondants à proximité de ses victimes. C'est en ce sens que David Hume a raison: si la science, d'une façon ou d'une autre, découvre par l'observation brute une relation telle que (*XII.18.*), mais sans que l'entendement humain puisse y trouver aucune relation de nécessité a priori, la réaction de la postérité ne sera pas de nier que (*XII.18.*) soit une véritable relation de cause à effet, mais de chercher par tous les moyens un artifice conceptuel, tel que les fluides ou les forces, qui *ajoutera* à (*XII.18.*) le caractère de nécessité logique qui lui manquait, afin de rassurer l'entendement. Autrement dit, si les événements *A* et *B* sont liés par (*XII.18.*), l'absence d'une nécessité logique entre eux n'est pas une raison valable pour nier que *A* soit la cause de *B*, car une telle nécessité logique peut toujours être inventée: il suffit d'avoir de l'imagination. En revanche, la dépendance statistique et l'antériorité ne peuvent pas être inventées, car elles doivent se conformer aux observations.

Ainsi les réflexions de Hume dans *Enquête sur l'entendement humain* ne sont pas des spéculations philosophiques que les scientifiques peuvent négliger, mais sont à la base même de l'esprit scientifique et ont une incidence *pratique* sur la méthode: les ignorer mène tout droit à l'erreur. L'insistance de Hume à vouloir évacuer la relation de nécessité n'est pas l'expression d'un positivisme dogmatique, mais d'une exigence de rigueur: définir la causalité comme une simple corrélation avec antériorité, c'est la soumettre totalement et exclusivement au verdict de l'expérience objective; au contraire, la définir comme une corrélation avec relation de nécessité, c'est introduire un élément de subjectivité dans la connaissance. Autrement dit, refuser de reconnaître une corrélation avec antériorité comme causalité au prétexte qu'il y manque une relation de nécessité équivaut à s'interdire de *chercher* une nouvelle relation de nécessité que la science ne possède

pas encore, (donc interdire à Kepler d'inventer son histoire de fluide, à Newton d'inventer les forces, à Faraday d'inventer les champs, etc.) et donc à soumettre la connaissance future à la connaissance présente. De même, voir de la causalité dans une corrélation *sans antériorité établie* au prétexte qu'on y voit une relation de nécessité équivaut à donner à un préjugé la priorité sur l'expérience.

ANNEXE DU CHAPITRE XII.

LA MÉTHODE DE LEVENBERG-MARQUARDT pour le calcul de la régression non linéaire.

Comme on l'a vu à la section **12. 4**, un problème de régression non linéaire consiste à chercher le minimum de la fonction

$$S(\alpha) = \sum_{i=1}^n p_i [f(\alpha, x_i) - y_i]^2 \quad (12 A.1)$$

en faisant varier les paramètres $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_Q)$. Ce minimum étant un point où le gradient de S s'annule, cela revient à trouver les α tels que $\vec{\text{grad}} S = 0$.

Lorsque le modèle de régression non linéaire $f(\alpha, x_i)$ dépend linéairement des paramètres à optimiser, la fonction S est du second degré, donc son gradient est linéaire. L'équation $\vec{\text{grad}} S = 0$ est alors un système linéaire de Q équations à Q inconnues qu'on résout numériquement par la méthode du pivot de Gauss.

N. B. On notera que le terme *linéaire* (ou le terme *non linéaire*) a, dans le présent contexte, deux sens indépendants à ne pas confondre: le modèle de régression *non linéaire* $f(\alpha, x_i)$ peut dépendre *linéairement* ou non des paramètres α : $y = f(\alpha, x)$ est un modèle de régression *non linéaire* si f dépend non linéairement de x , mais il peut dépendre linéairement de α ; ainsi $f(\alpha, x) = \alpha_1 x^2 + \alpha_2 x + \alpha_3$ est un modèle de régression non linéaire, qui dépend linéairement de α .

Lorsque la dépendance en α n'est pas linéaire on doit traiter un système non linéaire de Q équations à Q inconnues et il existe pour cela un algorithme numérique qui a fait ses preuves, connu sous le nom de *Levenberg*

– Marquardt⁽¹⁾. Mon expérience d'enseignement dans une école d'ingénieur m'a appris que les problèmes de moindres carrés sont de loin les plus fréquemment rencontrés par les praticiens. Il me semble donc utile d'inclure ici une description de cette méthode, accompagnée d'explications conformes à l'esprit de cet ouvrage.

Étant donné que la fonction $S(\alpha)$ s'exprime analytiquement à l'aide de fonctions élémentaires (la fonction $f(\alpha, x)$ ne peut offrir un bon modèle de régression que si elle est élémentaire), elle est facile à dériver. On a

$$S_\ell = \frac{\partial S}{\partial \alpha_\ell} = \sum_{i=1}^n 2 p_i [f(\alpha, x_i) - y_i] \frac{\partial f}{\partial \alpha_\ell}(\alpha, x_i) \quad (12 A.2)$$

Le vecteur $\{S_\ell\}_{\ell=1,2,\dots,Q}$ est le gradient déjà mentionné. Les dérivées secondes seront

$$\begin{aligned} S_{k\ell} &= \frac{\partial S_\ell}{\partial \alpha_k} = \frac{\partial S_k}{\partial \alpha_\ell} = \\ &= \sum_{i=1}^n 2 p_i \left\{ [f(\alpha, x_i) - y_i] \frac{\partial^2 f}{\partial \alpha_\ell \partial \alpha_k}(\alpha, x_i) + \frac{\partial f}{\partial \alpha_k}(\alpha, x_i) \cdot \frac{\partial f}{\partial \alpha_\ell}(\alpha, x_i) \right\} \end{aligned} \quad (12 A.3)$$

Ces expressions se prêtent parfaitement bien à la programmation et par conséquent l'algorithme itératif qui conviendra le mieux à la résolution du système d'équations $\vec{\text{grad}} S = 0$ est la méthode de Newton ou méthode de la tangente. Elle consiste à choisir des valeurs initiales $\alpha_1^{(0)}, \alpha_2^{(0)}, \dots, \alpha_Q^{(0)}$ — qu'on notera collectivement $\alpha^{(0)}$ — à partir desquelles on calcule $\alpha^{(1)}$ par

$$\alpha^{(1)} = \alpha^{(0)} - \mathfrak{E}_2^{-1} \circ \mathfrak{E}_1 \quad (12 A.4)$$

où $\mathfrak{E}_1 = \vec{\text{grad}} S(\alpha^{(0)})$ et $\mathfrak{E}_2$ est la matrice des $S_{k\ell}(\alpha^{(0)})$. Après quoi on recommence en remplaçant $\alpha^{(0)}$ par $\alpha^{(1)}$, puis $\alpha^{(1)}$ par $\alpha^{(2)}$, et ainsi de suite. Rappelons que cette méthode de Newton s'interprète géométriquement de la manière suivante : l'équation

$$\beta = \vec{\text{grad}} S(\alpha_1, \alpha_2, \dots, \alpha_Q) \quad (12 A.5)$$

représente une courbe gauche dans l'espace $\mathbb{R}^{2Q}$ des coordonnées $\alpha_1, \alpha_2, \dots, \alpha_Q, \beta_1, \beta_2, \dots, \beta_Q$; la solution du système d'équations $\vec{\text{grad}} S = 0$ (ce qu'on cherche) correspond à l'intersection de cette courbe avec le sous-espace (de dimension Q) $\beta = 0$. La méthode consiste alors à se rapprocher de ce point en suivant la tangente à la courbe : ainsi $\alpha^{(1)}$ est le point d'intersection de la tangente en $\alpha^{(0)}$ avec le sous-espace $\beta = 0$; puis $\alpha^{(2)}$ est le point

⁽¹⁾ D. W. Marquardt J. Soc. Ind. Appl. Math. vol 11 (1963) pp 431 – 441.

d'intersection de la tangente en $\alpha^{(1)}$ avec le sous-espace $\beta = 0$, et ainsi de suite.

La méthode de Newton est recommandée lorsque la dérivée de la fonction (ici $\vec{\text{grad}} S$) est facile à programmer, ce qui est justement le cas. Elle converge extrêmement vite, à condition de prendre $\alpha^{(0)}$ déjà près du point qu'on veut atteindre : l'erreur est alors à chaque itération de l'ordre du carré de l'erreur précédente, ce qui signifie que le nombre de décimales exactes double à chaque itération. Mais si le point de départ $\alpha^{(0)}$ est éloigné du point qu'on veut atteindre, la méthode ne converge plus du tout ou converge trop lentement pour être efficace. Or dans le problème des moindres carrés on ne peut pas deviner a priori un bon point initial.

L'idée essentielle de la méthode de Marquardt est de proposer un moyen simple et algorithmiquement économique pour, dans un premier temps, se rapprocher du minimum (appelons cela la *phase de descente rapide*) ; après quoi, dans un second temps, on peut amorcer avec profit l'itération de Newton (qu'on appellera la *phase critique*).

Le principe est le suivant (la mise en œuvre pratique exigera quelques aménagements semi-empiriques qu'on décrira après). Il s'agit, partant d'un point fixé arbitrairement, de se rapprocher aussi rapidement que possible du minimum (suffisamment pour que l'itération de Newton devienne efficace). Pour cela il est logique de suivre la ligne de plus grande pente sur la surface d'équation

$$\gamma = S(\alpha_1, \alpha_2, \dots, \alpha_Q) \quad (12\ A.6)$$

N. B. L'équation 12 A.5 considérée ci-dessus est vectorielle : c'est un *système* de Q équations, il lui correspond donc une courbe dans l'espace $\mathbb{R}^{2Q}$. Par contre l'équation 12 A.6 est scalaire, il lui correspond donc une (hyper)surface de dimension Q dans l'espace $\mathbb{R}^{Q+1}$ des paramètres $\alpha_1, \alpha_2, \dots, \alpha_Q, \gamma$.

Or la ligne de plus grande pente sur une surface d'équation 12 A.6 (considérée dans le sous-espace $\mathbb{R}^Q$ des coordonnées $\alpha_1, \alpha_2, \dots, \alpha_Q$) a pour vecteur tangent le vecteur $S_1, S_2, \dots, S_Q$. Ce vecteur est centrifuge par rapport au minimum. Donc pour se rapprocher du minimum à partir d'un point $\alpha^{(0)}$ initial situé sur le plan des α , il suffit de prendre la direction $-\vec{\text{grad}} S(\alpha^{(0)})$. Ainsi on choisira le second point $\alpha^{(1)} = \alpha^{(0)} - \lambda \vec{\text{grad}} S(\alpha^{(0)})$ (avec évidemment $\lambda > 0$). Mais il faut choisir λ de manière optimale. Si on prend un λ trop petit il faudra un nombre trop grand d'itérations pour arriver dans le voisinage du minimum ; si on le prend trop grand, on risque de dépasser ce minimum. C'est alors pour déterminer le meilleur choix de ce paramètre λ qu'interviennent plusieurs aménagements empiriques que nous décrivons maintenant.

D'abord les maxima ou minima locaux sont fréquents, surtout si la dimension du modèle, c'est-à-dire le nombre de paramètres α , est grande. Marquardt a donc proposé de placer dans le programme le test simple suivant : on commence avec un λ petit, par exemple 1/100 et on calcule $\alpha^{(1)} = \alpha^{(0)} - \lambda \vec{\text{grad}} S(\alpha^{(0)})$; si $S(\alpha^{(1)}) < S(\alpha^{(0)})$, cela indique qu'on descend la pente; on recommence alors avec un λ beaucoup plus grand (ainsi on évite des maxima locaux).

Par ailleurs il n'est pas indispensable de suivre *exactement* la ligne de plus grande pente; en pratique il suffit de ne pas trop s'en écarter, et sur ce point une approximation même grossière est fortement recommandée si elle économise du temps de calcul. C'est pourquoi Marquardt propose de suivre la direction du vecteur R de composantes $R_\ell = S_\ell / S_{\ell,\ell}$. La bonne raison pour cela est que de cette manière le paramètre λ sera sans dimension. Ce point n'est pas indifférent car le test doit choisir les valeurs petites ou grandes de λ sans connaître les ordres de grandeurs sur le terrain. En introduisant le vecteur R on diminue les risques entraînés par un mauvais choix. Toutefois, d'après la définition 12 A.3 des $S_{k,\ell}$, les éléments diagonaux $S_{\ell,\ell}$ ne sont pas forcément positifs et peuvent parfois s'annuler ou devenir trop petits; c'est pourquoi Marquardt a proposé aussi de remplacer la définition 12 A.3 par

$$S_{k\ell} = \sum_{i=1}^n 2p_i \cdot \frac{\partial f}{\partial \alpha_k}(\alpha, x_i) \cdot \frac{\partial f}{\partial \alpha_\ell}(\alpha, x_i) \quad (12 A.7)$$

qui garantit que les termes diagonaux seront positifs : ils ne peuvent devenir nuls que si $\vec{\text{grad}}_\alpha f$ peut lui-même devenir nul, ce qu'on évitera par le choix du modèle. Cela implique que la méthode de Newton est quelque peu modifiée. Mais les termes négligés (les dérivées secondes de f par rapport aux paramètres α) sont multipliés par les facteurs $f(\alpha, x_i) - y_i$. Or de deux choses l'une :

— ou bien ces facteurs ne deviennent pas tous petits lorsqu'on se rapproche du minimum; alors les termes négligés ne sont pas négligeables et le calcul sera faux;

— ou bien ces facteurs deviennent tous petits lorsqu'on se rapproche du minimum, et alors le calcul sera correct.

Mais si les facteurs $f(\alpha, x_i) - y_i$ ne deviennent pas tous petits lorsqu'on se rapproche du minimum, c'est que le minimum de $S(\alpha)$ n'est pas petit et que donc le modèle est mauvais : il ne permet pas d'approcher les points x_i, y_i . La méthode des moindres carrés n'ayant de toute façon aucun sens dans ce cas, on voit que la proposition de Marquardt est justifiée.

Il existe encore une autre raison, au moins aussi importante que la positivité des $S_{\ell,\ell}$, de remplacer 12 A.3 par 12 A.7. Comme nous venons

de le voir, ce changement ne s'écarte guère de la méthode de Newton si le minimum de $S(\alpha)$ est petit ; par contre si les facteurs $f(x_i) - y_i$ restent appréciables, l'itération ne converge plus vers le minimum puisqu'on ne suit plus la tangente. *Or cela est même un avantage* : en effet, l'un des défauts bien connus de la méthode de Newton est qu'elle converge trop facilement vers des minima locaux (il suffit de l'amorcer avec une valeur initiale α_0 située dans le bassin d'attraction d'un tel minimum local pour qu'elle converge vers celui-ci et non vers celui qu'on veut) ; mais si on applique 12 A.7 au lieu de 12 A.3, la méthode ne convergera justement *que* si on est près du vrai minimum (le minimum absolu). Si on se trouve par hasard au voisinage d'un minimum local où $S(\alpha)$ n'est pas petit, l'itération ne conduira pas à s'en rapprocher ; au contraire, au bout de quelques itérations on sortira de son bassin d'attraction et on recommencera le processus ailleurs. Ainsi seuls les minima effectivement petits feront converger le procédé. En optant pour 12 A.7 au lieu de 12 A.3, Marquardt fait donc d'une pierre deux coups.

Une troisième astuce de Marquardt est encore la suivante. La tactique générale est, comme on l'a vu plus haut, de suivre la pente (ou plutôt le vecteur R) lorsqu'on est loin du minimum, puis de passer à la méthode de la tangente — modifiée comme on vient de voir — lorsqu'on est suffisamment près. Marquardt introduit pour cela la matrice D dont les éléments sont

$$D_{j,k}(\alpha) = \begin{cases} S_{j,j}(\alpha) \cdot (1 + \lambda) & \text{si } j = k \\ S_{j,k}(\alpha) & \text{si } j \neq k \end{cases} \quad (12 A.8)$$

où $S_{j,k}(\alpha)$ est défini par 12 A.7, et propose d'appliquer la formule 12 A.4 avec la matrice D ainsi définie au lieu de la matrice $\mathfrak{S}$:

$$\alpha^{(1)} = \alpha^{(0)} - D(\alpha^{(0)})^{-1} \circ \overrightarrow{\text{grad}} S(\alpha^{(0)}) \quad (12 A.9)$$

Cela se comprend ainsi : lorsque λ est petit, la matrice $D_{j,k}(\alpha^{(0)})$ est pratiquement identique à la matrice $S_{j,k}(\alpha^{(0)})$, donc avec 12 A.9 on applique en fait la méthode de Newton ; si au contraire λ est grand, la matrice $D_{j,k}(\alpha^{(0)})$ est pratiquement identique à la matrice diagonale d'éléments $S_{j,j}(\alpha^{(0)})$, par conséquent $D^{-1} \circ \overrightarrow{\text{grad}} S$ est pratiquement identique au vecteur R introduit ci-dessus. Ainsi on dispose d'une grande souplesse algorithmique pour passer progressivement de la phase de descente rapide à la phase critique : il suffit de jouer sur les valeurs du paramètre λ .

Tous ces aménagements sont des recettes empiriques : choisir λ grand ou petit selon des critères qui ne sont pas absolument fiables, suivre le vecteur R plutôt que la ligne de plus grande pente, remplacer 12 A.3 par 12 A.7, tout cela ne peut pas être justifié par des démonstrations rigoureuses et

formelles et c'est uniquement l'usage pratique qui a tranché. En effet, cette méthode de Marquardt a aujourd'hui convaincu tous les praticiens et elle est implémentée dans les logiciels de calcul (Mathlab, Statistica, etc). Mais il est clair qu'elle ne marche pas à coup sûr. Son succès est dû à ce qu'on n'a pas trouvé mieux.

Parmi les causes de “plantage” possibles, la plus courante est de tourner en rond dans une zone de minima dégénérés; le test qui sert à déterminer λ saute alors éternellement entre un petit λ et un grand. Un programme bien conçu doit donc prévoir un test d'arrêt avec message d'erreur pour sortir d'une telle boucle sans fin; par exemple si λ passe plus de dix fois d'une grande valeur à une petite et vice-versa, arrêter le processus et afficher “Sorry, but it seems I am in a wrong track”. La meilleure solution est que le programme prévienne alors une sortie de secours avec entrée manuelle des valeurs de λ .

On peut donc résumer la méthode comme suit.

a) Créer des routines qui calculent le vecteur $\vec{\text{grad}} S(\alpha)$ donné par 12 A.2 et la matrice D donnée par 12 A.8 en fonction de α et λ , ainsi que son inverse D^{-1} par la méthode du pivot.

b) Choisir une valeur initiale $\alpha^{(0)}$ (en général complètement arbitraire car on n'a aucun critère de choix), calculer $S(\alpha^{(0)})$, et commencer l'itération de 12 A.9 avec un λ moyen mais plutôt petit (de l'ordre de 0.01 ou 0.001), ce qui va donner $S(\alpha^{(1)})$. Le fait de choisir ce premier λ plutôt petit revient à appliquer la méthode de Newton modifiée par 12 A.7.

c) Calculer $S(\alpha^{(1)})$ et le comparer à $S(\alpha^{(0)})$:

— **c1)** si $S(\alpha^{(1)}) \geq S(\alpha^{(0)})$ (ce qui veut dire qu'on n'était certainement pas au voisinage du bon minimum: on devait être au voisinage d'une instabilité, ou d'un minimum local non petit dont on s'est écarté du fait qu'on n'applique pas 12 A.3, mais 12 A.7), prendre un λ beaucoup plus grand, par exemple $\lambda = 1$ ou $\lambda = 0.1$ (ce qui veut dire qu'on suit maintenant plutôt le vecteur R afin de s'écarter nettement de ce point qui n'était visiblement pas bon) et recommencer **b**;

— **c2)** si $S(\alpha^{(1)}) < S(\alpha^{(0)})$ (on peut alors penser qu'on est sur la bonne voie vers le minimum) prendre un λ encore plus petit (disons dix fois, ce qui veut dire qu'on applique maintenant la méthode de Newton modifiée comme on a vu) et recommencer aussi longtemps que $S(\alpha^{(n)})$ diminue, ce qui va donner $\alpha^{(2)}$, $\alpha^{(3)}$, ... Continuer l'itération tant que $\Delta S = S(\alpha^{(n-1)}) - S(\alpha^{(n)})$ est positif et appréciable; revenir à **c1** dès que ΔS devient < 0 ; stopper lorsque ΔS devient petit en restant > 0 , on est alors arrivé à destination (la petitesse ici est aussi question de flair: en général on convient que l'itération peut s'arrêter quand ΔS

devient inférieur à 0.01 ; en réalité, cela dépend fortement du modèle et de l'échantillon).

N. B. Au stade **c2** rien ne prouve encore définitivement qu'on est sur la bonne voie : si on l'est, on s'écartera peu de la tangente à la courbe 12 A.5 et — si le minimum n'est pas dégénéré — l'itération convergera ; sinon, il arrivera tôt ou tard que $S(\alpha^{(n)})$ augmente à nouveau, puisqu'on ne suit pas la tangente.

Les logiciels du commerce sont contraints de choisir forfaitairement les critères tels que le choix de λ ou la petitesse de ΔS . Mais les meilleurs sont programmables (par exemple Matlab) et permettent un ajustement par l'utilisateur. Cette possibilité est essentielle, car ces choix ne peuvent vraiment être rendus optimaux que pour un modèle donné, d'après les exigences de précision et de temps de calcul imposés par les conditions concrètes du problème. Le meilleur ajustement est toujours empirique, mais l'utilisateur ne peut intervenir à bon escient que s'il a compris les principes théoriques exposés ci-dessus.

Il est intéressant de comparer les calculs effectués selon la méthode de Marquardt, lorsque le modèle $y = f(\alpha, x)$ ne dépend pas linéairement des paramètres α , avec les calculs à effectuer lorsque le modèle dépend linéairement des paramètres. Dans ce dernier cas, il y a seulement à résoudre un système linéaire de Q équations à Q inconnues, donc on applique une fois la méthode du pivot. Dans la méthode de Marquardt, la méthode du pivot est réappliquée à chaque itération (cf **a**). Cette partie du programme représente évidemment la quasi totalité du temps de calcul, sauf si la dimension Q est petite. Ainsi, si N itérations en tout sont nécessaires pour parvenir au résultat, le modèle sera N fois plus dispendieux qu'un modèle à dépendance linéaire. Cet argument doit être pris en compte lorsqu'on décide de choisir un modèle de grande dimension. En revanche, un modèle à dépendance non linéaire de dimension $Q = 2$ ou $Q = 3$, est certainement préférable à un modèle à dépendance linéaire de dimension élevée. Par exemple un modèle du type $f(\alpha_1, \alpha_2, x) = x\alpha_1 e^{-\alpha_2 x}$ sera bien meilleur (s'il convient à l'échantillon) qu'un polynôme de degré 10.

On trouvera dans les pages suivantes un programme qui exécute l'algorithme. Le langage PASCAL a été choisi pour sa facilité de lecture, une traduction en C peut se faire rapidement.

Le programme que voici exécute la méthode de Marquardt.

```
program Marquardt ;
```

```
label
```

```
  99 ;
```

```
const
```

```
  N = 12 ;    (nombre de points)
```

```
  stop = 10 ;    (pour limiter le nombre d'itérations.)
```

```
  eps = 1e-5 ;    (précision du calcul.)
```

```
var
```

```
  p, q, u, v, u0, v0, u1, v1, w, z, max, scale : double ;
```

```
  det, lambda, delta, S0, S1 : double ;
```

```
  loi, xx, yy : array[1..N] of double ;
```

```
  i, j, k : integer ;
```

```
function phi (x, y, p : double) : double ;
```

```
  begin
```

```
    phi := exp(x * ln(p) - y * p) ;
```

```
  end ;
```

Cette fonction définit le
modèle de régression choisi.

```
function phix (x, y, p : double) : double ;
```

```
  begin
```

```
    phix := ln(p) * exp(x * ln(p) - y * p) ;
```

```
  end ;
```

Dérivée par rapport à
 x de la fonction `phi`.

```
function phiy (x, y, p : double) : double ;
```

```
  begin
```

```
    phiy := -p * exp(x * ln(p) - y * p) ;
```

```
  end ;
```

Dérivée par rapport à
 y de la fonction `phi`.

Lorsqu'on voudra changer le modèle de régression sans
changer le nombre de paramètres, il suffira de remplacer
ces trois fonctions sans toucher à la suite du programme.
Mais si on veut passer à trois paramètres ou plus, il faudra
modifier aussi les fonctions `SQ`, `Sx`, etc.

function SQ (x, y : double) : double ;

var

s : double ;

l : integer ;

begin

s := 0 ;

for l := 1 **to** N **do**

begin

s := s + loi[l] * sqr(phi(x, y, xx[l]) - yy[l]) ;

end ;

SQ := s ;

end ;

SQ est la somme des carrés des écarts.

function Sx (x, y : double) : double ;

var

s : double ;

l : integer ;

begin

s := 0 ;

for l := 1 **to** N **do**

begin

s := s + loi[l] * (phi(x, y, xx[l]) - yy[l]) * phix(x, y, xx[l]) ;

end ;

Sx := s ;

end ;

Calcul de $\frac{\partial SQ}{\partial x}$.

function Sy (x, y : double) : double ;

var

s : double ;

l : integer ;

begin

s := 0 ;

for l := 1 **to** N **do**

begin

s := s + loi[l] * (phi(x, y, xx[l]) - yy[l]) * phiy(x, y, xx[l]) ;

end ;

Sy := s ;

end ;

Calcul de $\frac{\partial SQ}{\partial y}$.

function Sxx (x, y : double) : double ;

var

 s : double ;

 l : integer ;

begin

 s := 0 ;

for l := 1 **to** N **do**

begin

 s := s + loi[l] * phix(x, y, xx[l]) * phix(x, y, xx[l]) ;

end ;

 Sxx := s * (1 + lambda) ;

end ;

Calcul de $\frac{\partial^2 \text{SQ}}{\partial x^2}$.

function Sxy (x, y : double) : double ;

var

 s : double ;

 l : integer ;

begin

 s := 0 ;

for l := 1 **to** N **do**

begin

 s := s + loi[l] * phix(x, y, xx[l]) * phiy(x, y, xx[l]) ;

end ;

 Sxy := s ;

end ;

Calcul de $\frac{\partial^2 \text{SQ}}{\partial x \partial y}$.

function Syy (x, y : double) : double ;

var

 s : double ;

 l : integer ;

begin

 s := 0 ;

for l := 1 **to** N **do**

begin

 s := s + loi[l] * phiy(x, y, xx[l]) * phiy(x, y, xx[l]) ;

end ;

 Syy := s * (1 + lambda) ;

end ;

Calcul de $\frac{\partial^2 \text{SQ}}{\partial y^2}$.

begin	Début du programme principal.
z := 0 ;	
for i := 1 to N do	Mise en mémoire du nuage de points. xx[i] est l'abscisse du point $N^{\circ} i$, yy[i] son ordonnée. Pour pouvoir tester plus facilement le programme, on a choisi ici un nuage de points situé exactement sur la courbe $y = x^u e^{-vx}$, avec $u = 1.75$ et $v = \sqrt{0.3} \simeq 0.5477$, de sorte que le modèle de régression sera exact.
begin	
w := z ;	
xx[i] := i ;	
q := sqrt(xx[i]) ;	
z := xx[i] * q * sqrt(q) * exp(-sqrt(0.3) * xx[i]) ;	
yy[i] := z ;	
if z > w then	
begin	Recherche du maximum de yy[i] afin de trouver la bonne échelle graphique (scale).
max := z ;	
end ;	
end ;	
scale := 220 / max ;	
for i := 1 to N do	loi[i] est le poids du i -ième point.
begin	Ici il est uniforme.
loi[i] := 1 ;	
end ;	
MoveTo(10, 242) ;	
LineTo(10, 229) ;	
MoveTo(9, 240) ;	
LineTo(490, 240) ;	Représentation graphique du nuage de points.
for i := 1 to N do	
begin	
k := round(scale * yy[i]) ;	
MoveTo(10 + 40 * i, 241 - k) ;	
LineTo(10 + 40 * i, 239 - k) ;	
MoveTo(9 + 40 * i, 240 - k) ;	
LineTo(11 + 40 * i, 240 - k) ;	
end ;	

```

writeln ;
writeln(' enter the initial parameters :') ;
write(' u = ' ) ;
readln(u) ;
write(' v = ' ) ;
readln(v) ;
writeln ;

99 :

lambda := 0.0      delta := 1 ;
i := 0 ;

while ((delta < -eps) or (delta > 0)) and (i <= stop) do
  begin
    MoveTo(10, 240) ;
    for j := 1 to 120 do
      begin
        w := j / 10 ;
        z := scale * phi(u, v, w) ;
        if z >= 500 then
          k := 500
        else if z <= -500 then
          k := -500
        else
          k := round(z) ;
        LineTo(10 + 4 * j, 240 - k) ;
      end ;

    i := i + 1 ;
    S0 := SQ(u, v) ;
    det := Sxx(u, v) * Syy(u, v) - Sxy(u, v) * Sxy(u, v) ;
    u1 := Sx(u, v) * Syy(u, v) - Sy(u, v) * Sxy(u, v) ;
    v1 := -Sx(u, v) * Sxy(u, v) + Sy(u, v) * Sxx(u, v) ;
    u0 := u - u1 / det ;
    v0 := v - v1 / det ;
    S1 := SQ(u0, v0) ;
    delta := S1 - S0 ;
  end ;
end while ;

```

Initialisation manuelle (au clavier) de l'itération.
Cela laisse l'initiative à l'utilisateur.

Début de l'itération.

On représente graphiquement l'évolution du modèle au cours de l'itération. Si tout se passe bien, les courbes obtenues doivent finir par se rapprocher de plus en plus du nuage de points.

Ici, simple précaution pour éviter *overflow*.

Ci-contre on reconnaît l'algorithme de Marquardt.

```
writeln(i : 3, ' lambda = ', lambda : 11) ;
writeln('u = ', u0 : 16 : 14, ' v = ', v0 : 16 : 14) ;
writeln('delta = ', delta : 15) ;
writeln('S0 = ', S0 : 15) ;
writeln ;

if (S0 >= 10000) then
    begin
        writeln ;
        writeln(' I think your initial parameters are highly
        unrealistic. ') ;
        writeln(' The process will probably not converge. ') ;
        writeln(' Could you choose other values ? ') ;
        writeln ;
        writeln(' enter new initial parameters :') ;
        write(' u = ') ;
        readln(u) ;
        write(' v = ') ;
        readln(v) ;
        writeln ;
        goto 99 ;
    end ;

if (delta >= 0) then
    begin
        lambda := 10 * lambda ;
    end
else
    begin
        lambda := lambda / 10 ;
        u := u0 ;
        v := v0 ;
    end ;

writeln ;
write(' ') ;
for j := 1 to 21 do
    write('---') ;
writeln ;
writeln ;
```

On redonne l'initiative à l'utilisateur s'il apparaît
que l'itération diverge trop.

On revient au début de l'itération.

Ci-contre on reconnaît la seconde
partie de l'algorithme de Marquardt.

On trace un pointillé horizontal.

```
if (delta >= -eps) and (delta <= 0) then
    begin
        writeln(' final result : S1 = ', S1 : 15) ;
        writeln(' u = ', u : 16 : 14) ;
        writeln(' v = ', v : 16 : 14) ;

        MoveTo(10, 240) ;
        for j := 1 to 120 do
            begin
                w := j / 10 ;
                z := phi(u, v, w) ;
                k := round(scale * z) ;
                LineTo(10 + 4 * j, 240 - k) ;
            end ;
        end
    else
        ...et si la convergence n'est pas stabilisée, on rend l'initiative à
        l'utilisateur. ...
        begin
            writeln(' I have stopped computing because there is no') ;
            writeln('appreciable progress after ', stop + 1 : 1, '
iterations . ') ;
            writeln(' Try other values for u and v or change the model.') ;
        end ;
    end.
```

À la fin de l'itération, si la convergence est stabilisée, on affiche le résultat final ...

...avec la courbe correspondante ...