Jean-Marie Gogue

Statistique pratique

© Association Française Edwards Deming 2004. Tous droits réservés

Table des matières

La compréhension de cet ouvrage est facilitée par les logiciels de statistique *Norma, Movira, Daisy, Alice* et *Cora*. Leur chargement est libre et gratuit sur le site http://www.fr-deming.org

Introduction

> À long terme, l'efficacité de la statistique dépendra moins de l'existence d'un corps de statisticiens de haut niveau que de l'émergence de toute une génération formée à l'esprit statistique. WALTER SHEWHART

Jusqu'au milieu du vingtième siècle, la statistique pouvait passer pour une discipline réservée à quelques spécialistes rompus au calcul des probabilités, mais les choses ont bien changé. Après la seconde guerre mondiale, un large public a commencé à s'intéresser à l'économie. Les lecteurs des journaux se sont habitués au vocabulaire et aux figures des pages économiques. Or l'économie fait constamment appel aux statistiques. D'ailleurs, pour s'adapter à ce changement, l'Éducation nationale a introduit des éléments de probabilité et de statistique dans ses programmes.

Les études statistiques sont de deux sortes : énumératives et analytiques. Les unes et les autres ont pour but de donner une base rationnelle aux prévisions, aux décisions et aux plans d'actions, mais la différence réside dans la nature des informations recueillies. Une étude énumérative traite des données provenant d'un ensemble fini, invariable au moment de l'étude, par exemple une récolte de blé. Au contraire, une étude analytique traite des données provenant d'un processus qui peut fonctionner indéfiniment, par exemple la culture du blé. Les études énumératives sont les plus connues, parce qu'elles composent souvent des livres et des articles de vulgarisation consacrés à l'économie. En revanche, les études analytiques sont largement répandues dans les entreprises et dans les laboratoires, où elles servent notamment à améliorer des processus de production. Ce livre est consacré essentiellement aux méthodes utilisées dans les études analytiques. Il faut remarquer d'ailleurs que certaines d'entre elles sont utilisées aussi dans les études énumératives.

Quelques précisions s'imposent au sujet des notions de population et d'échantillon. La plupart des études statistiques sont menées à partir d'échantillons. Par exemple, pour évaluer les intentions de vote du corps électoral quelques semaines avant des élections, un institut de sondage travaille avec des échantillons d'un millier de personnes environ, tirées au hasard sur l'ensemble de la population. L'opération est conduite suivant une méthode bien définie. La taille de l'échantillon est calculée en sorte de réaliser un compromis entre la précision de l'estimation et le coût de l'opération. Mais en d'autres circonstances, les statisticiens travaillent avec des échantillons plus petits. Par exemple, dans un

laboratoire de recherches biologiques, les études portent le plus souvent sur quelques dizaines d'individus. Il ne faut pas croire qu'une étude statistique a forcément besoin de grands échantillons pour donner des résultats probants.

Les résultats des sondages étant régulièrement évoqués dans les débats politiques, tout le monde sait ce que ce que ces mots signifient. Mais la notion de population tend à masquer celle de processus. Un échantillon n'est extrait d'une population que dans le cas d'une étude énumérative. Les statistiques démographiques, telles que la proportion d'hommes et de femmes par tranches d'âge dans une région donnée, sont le résultat d'études énumératives sur des populations humaines. Par extension, les professionnels de la statistique emploient ce mot quand ils étudient un ensemble d'objets dont le nombre est défini en un lieu et en un temps donnés, comme par exemple un stock de marchandises dans une chaîne de distribution. Mais le concept de population ne signifie pas grand chose dans le cas d'une étude analytique, puisque le nombre des objets concernés est inconnu et peut varier à chaque instant. Par exemple, quand un chercheur étudie une nouvelle souche de bactérie dans un laboratoire, il n'a pas affaire à une population de bactéries, mais à un processus de reproduction de bactéries. Ce n'est pas sur une population, mais sur un flux continu de données qu'il prélève des échantillons.

Cet ouvrage s'adresse aux statisticiens qui voudraient mettre à jour leurs connaissances ainsi qu'à toute personne désirant se servir de méthodes statistiques pour faire des choix pertinents et améliorer certaines performances. Le principal obstacle à leur diffusion, outre la crainte d'un apprentissage difficile, est l'idée très répandue que l'intuition est souvent préférable à un raisonnement logique, comme si les deux approches étaient incompatibles. Or les ouvrages de statistique ne manquent pas, et certains sont excellents. Mais aucun n'a réussi jusqu'à présent à vaincre cette résistance collective. Pour y parvenir, j'ai associé à ce modeste ouvrage cinq logiciels. Ils dispensent l'opérateur de faire des calculs fastidieux, car les résultats apparaissent à l'écran dès que les données sont enregistrées. J'espère ainsi que beaucoup de lecteurs franchiront enfin le pas.

Il est souhaitable que le lecteur possède les connaissances requises pour le baccalauréat scientifique. Néanmoins, celui qui trouve que son niveau n'est pas suffisant peut facilement rafraîchir ses connaissances grâce à un manuel scolaire. A mon avis, le plus grand mérite du programme de statistique de 1^{e} S est d'ouvrir l'esprit à la notion de variance. En revanche, les manuels scolaires n'apprennent pas à se servir d'une variance expérimentale pour confirmer une hypothèse. A cet égard, il serait utile de montrer aux élèves que certains résultats du calcul statistique, par exemple le coefficient de corrélation, ont un seuil de signification. C'est un raisonnement que l'on rencontre souvent dans ce livre, exemples à l'appui.

1

Les processus

Le soleil, la terre, la pluie, ce livre… toute chose est le résultat d'un processus. L'idée peut sembler des plus banales, mais elle a pris de l'importance dans le monde moderne à mesure qu'on savait mieux agir sur le processus pour le contrôler ou pour en modifier le résultat. Par exemple, dans les fabrications de série, on a commencé par trier les pièces défectueuses. Puis, comprenant qu'il était préférable de ne pas produire de déchets, on a trouvé le moyen d'agir sur les processus de production. Ce fut le miracle japonais. Je me souviens qu'en 1978, dans une usine japonaise d'électronique, les taux de défauts des cartes imprimées étaient 300 fois plus faibles que dans une usine française utilisant les mêmes techniques.

Les méthodes statistiques permettent de contrôler et d'améliorer la plupart des processus présents dans notre société. Voici une liste d'activités où chacun peut trouver un processus qui le concerne :

Administration (ministères, conseils généraux, mairies)
Production industrielle (petite et grande séries)
Recherche (laboratoires publics et privés)
Artisanat (menuiserie, électricité, etc.)
Edition (livres, journaux, jeux vidéo)
Services (banques, assurances, voyages, etc.)
Education (méthodes d'apprentissage)
Sports (méthodes d'entraînement)
Management (gestion du personnel, gestion de projets, etc.)
Finance (bourse, gestion immobilière, etc.)
Commerce (gros, détail, boutiques et grande distribution)
Transport (routier, ferroviaire, maritime et aérien)
Médecine (hôpitaux, cliniques et cabinets privés)
Agriculture (fruits, légumes, vignobles, etc.)
Élevage (petites et grandes exploitations)

Voir un processus à travers un schéma

Pour étudier un processus, il faut le représenter par un diagramme d'événements connu sous le nom de *flugramme*. Utilisant quelques symboles très simples, tels que des rectangles reliés par des traits, il n'est pas fait pour montrer un processus idéal, mais pour rendre compte des faits observés réellement. Une personne qui n'est pas impliquée en permanence dans un processus a toujours

tendance à l'imaginer plus simple qu'il n'est en réalité. C'est le cas de ceux et celles qui dirigent une affaire sans passer sur le terrain un temps suffisant en discutant avec les gens. Un processus est souvent assez compliqué parce qu'il comporte des incidents. Or c'est précisément la connaissance de ces incidents qui permettra d'améliorer les résultats.

Le flugramme est un cadre dans lequel on inscrit les données du processus : quantités, pourcentages, coûts et mesures physiques, à mesure qu'elles se présentent. Il est nécessaire d'interroger tous les acteurs concernés, car chacun ne possède qu'une partie de l'information. Au début de l'étude, on constate que de nombreuses informations sont inconnues ; on les obtiendra progressivement.

Dans cette approche, il ne faut pas pécher par excès de formalisme. On peut commencer tout simplement par écrire la liste des éléments du processus, puis essayer de les disposer dans un ordre logique. C'est ainsi que j'ai vu, par exemple, des enfants de dix ans qui étudiaient leur processus d'apprentissage dans une école primaire de la région parisienne. Ils se sont lancés dans cette étude après qu'on leur eût expliqué pendant une heure seulement la notion de processus. Le maître a beaucoup appris du travail de ses élèves, et les résultats de la classe se sont améliorés de façon spectaculaire. Cet exemple montre l'efficacité de la méthode, même quand elle utilisée hors du cadre traditionnel de la recherche, en dehors des bureaux d'études et des laboratoires.

Deux types de données

A l'issue d'un processus, on trouve beaucoup de données, mais on en retient peu. Par exemple, un homme qui prend le train tous les jours pour aller à son travail rentre dans un processus dans lequel les heures de départ et d'arrivée du train sont des données parmi d'autres. Il ne se souviendra peut-être que des quelques jours où le train était en retard. Le fait n'a pas une grande importance. En revanche, dans la société de transport, il y a un service chargé d'enregistrer toutes les données utiles et nécessaires, soit pour des raisons légales, soit pour la satisfaction des usagers. En s'adressant à ce service, on peut connaître le nombre exact de retards ayant dépassé 10, 20 ou 30 minutes pendant une certaine période. De façon générale, chacun peut trouver sur internet une multitude de données concernant la vie pratique. Pour ne pas se perdre dans cet univers, il est bon de distinguer deux types de données : les mesures et les dénombrements.

Les mesures s'expriment toujours avec une unité. Rappelons brièvement l'existence d'un système international d'unités de mesure dont les bases principales sont le mètre, qui est l'unité de longueur, le kilogramme, qui est l'unité de masse, et la seconde, qui est l'unité de temps. Le système international compte plusieurs dizaines d'unités, dérivées de sept unités de base. Tout le monde sait ou devrait savoir, par exemple, que l'unité de puissance est le watt. En économie, il faut y ajouter les unités monétaires. Les mesures ont comme point commun le fait d'être des *variables continues.*

Les dénombrements sont, par définition, des nombres entiers positifs ou nuls. Ce sont des *variables discrètes*. Beaucoup de données utilisées en économie sont des dénombrements et des nombres qui en sont dérivés par calcul, notamment les pourcentages. Ceux-ci ne sont pas des nombres entiers, mais ils gardent la particularité d'être indépendants des unités de mesure. On passe facilement d'une mesure à un dénombrement. Par exemple, si les heures de départ et d'arrivée des trains sont enregistrés par la SNCF, on peut connaître le nombre et le pourcentage des trains dont le retard a dépassé 30 minutes en 2003 sur la ligne Paris-Chartres.

Les méthodes présentées dans ce livre concernent surtout les variables continues, mais il est utile que le lecteur sache comment la statistique traite les variables discrètes. Les méthodes sont différentes ; elles sont basées en grande partie sur la distribution binomiale.

La distribution binomiale

Certains processus ressemblent à des jeux de hasard. Par exemple, les lois de l'hérédité reposent sur la probabilité qu'un attribut du père ou de la mère passe chez un individu de la génération suivante. Mais bien que chaque résultat soit aléatoire, la proportion d'individus ayant cet attribut dans la population issue du processus est un nombre relativement constant. On comprend donc qu'il soit intéressant, connaissant cette proportion, d'estimer la probabilité de le trouver chez un nouvel individu.

L'étude des probabilités a débuté au XVII^{e} siècle avec Blaise Pascal, le but étant de répondre à la demande d'un joueur qui voulait savoir quelle était sa chance de tirer telle ou telle carte. Partant du principe que la probabilité est la même pour toutes les cartes, la solution du problème reposait sur l'application de deux règles de calcul : l'une pour une association de tirages mutuellement incompatibles, l'autre pour une succession de tirages indépendants. Cette nouvelle algèbre, qui est maintenant nommée *algèbre des événements*, est à l'origine de la distribution binomiale.

Le modèle Quand le résultat d'un processus ne peut prendre que deux valeurs s'excluant mutuellement (nous les désignerons par 0 et 1), il s'agit d'une variable aléatoire discrète. Cette situation a pour modèle le tirage au hasard d'une boule dans une urne contenant un certain nombre de boules blanches et noires, dans une proportion déterminée. Si nous faisons successivement plusieurs tirages en remettant chaque fois la boule dans l'urne après avoir noté sa couleur, nous obtenons une série de variables aléatoires ayant la valeur 0 ou 1. Le nombre de boules noires obtenu dans cette série est une autre variable aléatoire ; elle prendra la valeur 0 s'il n'y a que des boules blanches, 1 s'il y a 1 boule noire, 2 s'il y a 2 boules noires, etc. Nous pouvons répéter l'expérience n fois, avec chaque fois le même nombre de tirages. Nous obtiendrons une série de variables discrètes ayant

des valeurs comprises entre 0 et n. La distribution binomiale est l'ensemble des probabilités affectées à ces expériences. Le calcul donne :

$$P\ (p, k) = C_n^{\ k}\ p^k q^{n-k}$$

Dans cette équation, k est le nombre de boules noires, et p et q sont les proportions de boules blanches et noires (avec par définition $p + q = 1$).

Les coefficients $C_n^{\ k}$ sont les termes successifs du développement du binôme $(p + q)^n$. Ils peuvent être trouvés au moyen du *triangle de Pascal*. Chaque nombre est obtenu en faisant la somme des deux nombres qui sont au dessus de lui.

Pour les valeurs de npq supérieures à 5, la distribution binomiale peut être assimilée à une distribution normale de moyenne np et d'écart-type $\sqrt{\ } npq$. C'est notamment le cas des sondages d'opinion, où le problème du statisticien est d'estimer un pourcentage avec un certain niveau de confiance.

Remarque Quand np est inférieur à 15, même si n est très grand, l'approximation normale n'est pas satisfaisante. Une autre approximation a été proposée par Poisson au XIX^{e} siècle. La principale propriété de la distribution de Poisson est que la variance est égale à la moyenne. Elle a des applications dans de nombreux domaines, notamment dans les télécommunications.

n	Coefficients de la distribution binomiale														
1							1								
2						1		2		1					
3					1		3		3		1				
4				1		4		6		4		1			
5			1		5		10		10		5		1		
6		1		6		15		20		15		6		1	
7	1		7		21		35		35		21		7		1
8	1	8		28		56		70		56		28		8	1

Triangle de Pascal

La notion de stabilité

Quand un processus aboutit à une série de résultats, le premier problème est de les interpréter. Par exemple, quand un commerçant voit son chiffre d'affaires mensuel augmenter notablement trois fois de suite, il sera certainement tenté de prévoir une augmentation régulière des ventes. Or il n'a que quatre valeurs, alors qu'il lui faudrait au moins huit valeurs successives pour faire une prévision rationnelle. Il ne peut pas se fier à sa seule intuition pour faire de telles prévisions. Le seul moyen d'y voir clair est d'utiliser un graphique de contrôle, ce que chacun peut faire facilement avec le logiciel *Movira*.

Par définition, on dit qu'une série de résultats est dans un état stable quand le graphique montre un profil de points qui pourrait être obtenu par tirage au hasard dans un bac de jetons numérotés. Bien entendu, cette définition n'est pas utilisable en pratique. C'est pourquoi les statisticiens ont fixé des critères d'instabilité qui se calculent à partir des valeurs numériques de la série expérimentale. Les principaux critères sont :

Un point est en dehors des limites de contrôle à trois sigma
Huit points successifs sont du même côté de la moyenne
Six points successifs montrent une variation uniforme

Un système stable est un système dont les performances sont prévisibles, à plus ou moins long terme, car les données sont distribuées de façon aléatoire autour de la moyenne. Au contraire un système instable est absolument imprévisible, mathématiquement parlant. Il en résulte que, pour de nombreux processus, la stabilité des résultats doit être recherchée comme un facteur d'économie. D'autre part, nous verrons plus loin qu'un plan d'expériences ne peut se faire valablement que dans une situation stable.

Méthode pour obtenir la stabilité

C'est en 1931 qu'ont été publiés les travaux de Walter Shewhart, chercheur aux *Bell Telephone Laboratories*. Ils sont à l'origine d'une méthode connue en France sous le nom de MSP (Maîtrise Statistique des Processus), et aux États-Unis sous celui de SPC (Statistical Process Control). Shewhart a montré qu'il est possible de déterminer des critères de stabilité uniquement à partir de la série expérimentale. Cette méthode est très répandue actuellement dans l'automobile et dans l'électronique. Il est important d'en connaître le principe.

La MSP se compose de deux parties. La première consiste à porter un jugement sur la stabilité de la caractéristique étudiée. Pour cela, les résultats de mesures sont portés sur un *graphique de contrôle*. La seconde partie consiste à remédier, le cas échéant, aux causes d'instabilité. Le jugement entraîne deux approches complémentaires :

Système instable. Stratégie d'action intensive pour identifier au plus tôt les causes d'instabilité. Les éliminer dans la mesure du possible.

Système stable. Stratégie de veille pour détecter des signes éventuels d'instabilité. On peut aussi déplacer la moyenne et diminuer la variabilité pour des raisons économiques.

Tout l'intérêt de la MSP réside dans le choix d'une stratégie efficace. Des problèmes existent aussi bien dans un système stable que dans un système instable, mais dans le premier cas ils sont imputables au système lui-même, alors que dans le deuxième cas ils sont imputables à des événements particuliers. C'est

pourquoi, face à un problème, celui qui ne connaît pas la méthode risque de faire deux types d'erreurs. Adopter une stratégie d'action intensive dans un système stable, c'est rechercher des causes qui n'existent pas, et parfois même créer de l'instabilité. Adopter une stratégie de veille dans un système instable, c'est négliger des occasions favorables pour une amélioration du système. Certaines personnes ayant une longue expérience savent adopter d'instinct la bonne stratégie, mais dans le doute, il est préférable de porter les résultats sur un graphique et d'observer les critères d'instabilité.

Remarque. Dans certains cas, l'instabilité d'un processus est un signe favorable. Quand un processus est dans une phase d'amélioration, il est évident que les résultats sont instables, puisqu'ils vont suivre une tendance régulière en hausse ou en baisse pendant quelque temps. D'une façon un peu différente, pour un élève qui est en phase d'apprentissage, de même qu'un sportif à l'entraînement, ou une personne accidentée pendant sa rééducation, l'instabilité des performances est bien la preuve qu'une transformation se réalise. Au contraire, la stabilité des performances montre que l'apprentissage, l'entraînement et la rééducation sont terminés, à moins de changer le processus, le système, c'est-à-dire l'entraîneur, les méthodes de travail, le matériel, etc.

Méthodes pour vérifier la normalité

Rappel La distribution normale est entièrement définie par deux paramètres. D'abord, il y a la moyenne, μ , qui situe le centre de la distribution. Ensuite, l'écart-type , σ , qui mesure la dispersion des mesures individuelles. Pour $\mu = 0$ et $\sigma = 1$ la courbe représentant cette fonction est symétrique par rapport à l'axe *oy* , et l'ordonnée à l'origine est $1/\sqrt{2\pi}$. Des tables donnant *y* en fonction de *x* ont été tracées dans ce cas particulier. Elles donnent les valeurs de la distribution normale *réduite*. En écrivant : $x = \mu + \sigma z$, et en remplaçant *x* par *z* , l'équation générale devient :

$$y = 1/\sigma\sqrt{2\pi} \cdot \exp(-z^2/2)$$

C'est l'équation de la distribution normale réduite. Une transformation linéaire permet ainsi de calculer la valeur de *x* en fonction de μ et de σ à partir des tables.

Ce qui intéresse le praticien, ce n'est pas la valeur de *y* mais la probabilité que la variable *z* se trouve entre deux valeurs z_1 et z_2. C'est l'intégrale de *y* . Il existe donc, en plus des tables précédentes, des tables de la distribution normale *cumulée* qui permettent de calculer cette probabilité, avec la même transformation linéaire. Ce sont les plus fréquemment utilisées. On en trouvera un extrait à la fin de cet ouvrage (table A).

Test d'adéquation Dans un grand nombre de calculs statistiques, il faut que les mesures expérimentales aient une distribution normale. Cette condition est fréquemment réalisée en pratique, notamment dans les sciences naturelles. Pour la

vérifier, Karl Pearson a imaginé un test d'adéquation fondé sur la distribution d'une variable aléatoire nommée χ^2 (khi deux). La méthode consiste à comparer les fréquences expérimentales et les fréquences théoriques. On commence par diviser les deux distributions en classes égales. Les premières fréquences étant notés f et les secondes F, la valeur de χ^2 est donnée par la formule :

$$\chi^2 = \Sigma (f_i - F_i)^2 / F_i$$

L'hypothèse de normalité est retenue quand ce résultat est inférieur à un nombre qui se trouve dans une table donnant la limite de χ^2 correspondant à un certain risque d'erreur. Cette table se trouve dans un grand nombre d'ouvrages de statistique. Mais il existe d'autres tests d'adéquation basés sur le même principe. Le test de Kolmogorov-Smirnov est spécialement intéressant parce qu'il se prête bien au calcul sur ordinateur (c'est le test du logiciel *Norma*). Il consiste à diviser les deux distributions en classes égales, puis à comparer les différences des fréquences à une valeur commune à toutes les classes. L'hypothèse de normalité est retenue lorsque, pour chacune des classes :

$$| f_i - F_i | < C / \sqrt{n}$$

Dans cette formule, C est une constante qui dépend du risque d'erreur, et n est l'effectif de l'échantillon. Pour un risque d'erreur de 5%, C = 1,36. Pour pouvoir effectuer ce test, il faut noter que n doit être au moins égal à 25.

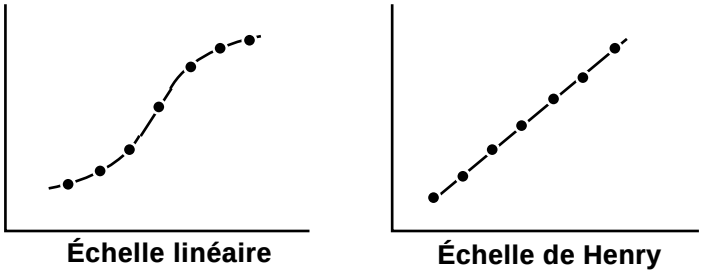
Remarque Ces deux tests ont été conçus pour vérifier l'adéquation d'une distribution expérimentale à toute autre distribution, expérimentale ou théorique. Nous n'avons traité ici que l'adéquation à la distribution normale.

Méthode graphique Il est souvent préférable d'utiliser une méthode graphique connue sous le nom de *méthode de Henry,* du nom d'un ingénieur français du XIX^{e} siècle. Elle est aussi efficace que les tests d'adéquation et donne des informations supplémentaires sur la distribution. Elle est particulièrement utile pour les petits échantillons.

Avant d'aborder cette méthode, rappelons quelle est la relation qui existe entre un histogramme de mesures et une courbe théorique, en l'occurrence celle de la distribution normale. Sur les deux figures, on trouve des surfaces proportionnelles à des fréquences. Imaginons que dans un histogramme expérimental dont on a vérifié la normalité, le nombre de mesures augmente indéfiniment, la largeur des rectangles diminuant de façon correspondante. Le contour de l'histogramme finira par se confondre avec une courbe continue, certainement assez proche de la courbe théorique. Nous pouvons faire le même raisonnement avec les fréquences cumulées, car on peut tracer un graphique de fréquences cumulées représentant des résultats de mesure. Quand le nombre de points augmente, la courbe expérimentale se rapproche de la courbe théorique.

La courbe qui représente l'intégrale de *y* est une courbe en S, l'ordonnée étant comprise entre 0 quand *x* est à - ∞ et 1 quand *x* est à + ∞ , ce qui provient du fait que l'intégrale est une probabilité dont les valeurs extrêmes sont par définition 0 et 1. Un point particulier est le point d'ordonnée 0,5 ; la courbe est symétrique par rapport à ce point. En changeant l'échelle verticale, on peut faire en sorte qu'une distribution normale soit représentée par une droite. C'est le graphique à échelle de Henry. Pour faire le test de normalité avec cette méthode, il faut donc utiliser un papier quadrillé dont l'échelle verticale est particulière. Le lecteur en trouvera un modèle sur la notice "lisez-moi" du logiciel *Cora*.

L'hypothèse de la normalité est admise lorsque tous les points sont alignés. L'idée d'apprécier un alignement par simple coup d'oeil peut sembler discutable à certains théoriciens, mais en pratique cette méthode pose rarement problème.



Distribution normale cumulée

Avantages de la stabilité et de la normalité

La stabilité d'un système et la normalité d'un échantillon issu d'un processus sont deux notions indépendantes. Un système peut être stable sans que les données qui le composent aient une distribution normale. Inversement, un échantillon de données provenant d'un processus instable peut avoir une distribution normale.

L'avantage de la stabilité, nous l'avons dit, est la possibilité de faire des prévisions rationnelles. Par exemple, la méthode *kamban*, une méthode de gestion de production très répandue dans l'industrie automobile, repose entièrement sur la stabilité des processus. Elle présente l'avantage de réduire considérablement le volume des stocks, donc de faire des économies. Mais le nouveau système est tel que la production est fortement ralentie dès que le processus sort de la stabilité. Il faut alors soit reconstituer des stocks, soit retrouver la stabilité au plus tôt.

La stabilité est une condition nécessaire pour une étude expérimentale destinée à améliorer un processus. Une étude faite dans un environnement instable risque fort d'aboutir à des conclusions erronées. Je pourrais citer le cas de la préparation d'un produit cosmétique dans un laboratoire. La température et l'hygrométrie du laboratoire n'étaient pas contrôlées. Par malchance, les propriétés d'une matière

première ont été altérées avant l'expérience par un brusque changement climatique, en sorte que les caractéristiques qui ont été portées dans la gamme de fabrication étaient fausses. L'erreur provenait du fait que le système était instable parce que les facteurs n'étaient pas tous contrôlés. Il aurait fallu sans doute que le laboratoire soit climatisé. De tels exemples sont nombreux.

L'avantage de la normalité tient en grande partie à la possibilité de définir entièrement une distribution expérimentale avec les deux paramètres μ et σ. En pratique, la normalité permet de prévoir avec précision quelle sera, dans un échantillon extrait d'une population donnée, la proportion d'individus dont une caractéristique est comprise entre deux valeurs. D'autre part, la normalité d'un échantillon est nécessaire pour faire un certain nombre de tests statistiques, car les calculs reposent sur cette hypothèse. Par exemple, quand on a vérifié que l'échantillon a une distribution normale, des tests permettent de connaître la probabilité de trouver un résultat expérimental inférieur à une limite fixée. Ce sont les tests *paramétriques*. Quand l'hypothèse de la normalité n'est pas retenue, il reste la possibilité de faire d'autres types de tests, mais ils sont généralement moins précis.

2

Corrélation

Le chapitre précédent montre que pour faire des prévisions ayant de fortes chances de se réaliser, il faut commencer par ordonner les informations disponibles. Ce travail étant fait, la première question qui vient normalement à l'esprit est de connaître les causes des variations, car on estime à juste titre que si une relation de causalité est établie, il sera possible de modifier le cours des événements. Ainsi, dès qu'un accident d'une certaine gravité se produit dans une région, de nombreuses personnes commencent à émettre des hypothèses sur les causes possibles. Mais l'intuition n'est pas toujours bonne conseillère. C'est pour cette raison que le physiologiste anglais Francis Galton, au XIX^{e} siècle, a inventé la *régression*. Ce fut la première méthode d'analyse statistique.

La régression linéaire

Dans ses travaux sur l'hérédité, Galton voulait vérifier l'hypothèse que les particularités des parents : taille, carrure, couleur des yeux, etc. existent chez l'enfant à un moindre degré. Pour cela, il a rassemblé plusieurs centaines de données, des couples de valeurs composés chacun de la taille d'un homme et de celle de son père. Il a porté ces valeurs sur un graphique avec la taille du père en abscisse et celle du fils en ordonnée. Constatant que tous les points obtenus étaient proches d'une droite, et que les fils avaient tendance à être de plus petite taille que les pères, il en a conclu que la taille des fils était en *régression* par rapport à celle des pères. Ce résultat ne concerne évidemment que l'Angleterre du XIX^{e} siècle, mais le terme est resté pour désigner une corrélation entre deux variables dont l'une précède l'autre.

La théorie permettant de calculer les coefficients de la droite de régression est due à Karl Pearson. Son développement sort du cadre de cet ouvrage ; elle ne se limite d'ailleurs pas à la régression linéaire. Par définition, quand une variable aléatoire y dépend d'une variable x, cette relation se nomme la *régression de y selon x*. La première a le nom de variable *expliquée*, et la seconde celui de variable *explicative.* Lorsque la dépendance est linéaire, les variables sont reliées par l'équation :

$$y = \alpha + \beta x + \varepsilon$$

Dans cette équation, α et β sont des constantes, tandis que ε est une variable aléatoire de distribution normale dont la moyenne est nulle et l'écart type indépendant de *x*. Cette variable est responsable de la dispersion.

Dans un problème de régression, on commence toujours par tracer un diagramme de dispersion. Bien que l'informatique en donne les moyens, il est préférable de le faire à la main sur une feuille de papier quadrillé, car c'est un travail rapide qui ne demande pas une grande précision. Par exemple, le conducteur d'une fourgonnette qui effectue des livraisons en région parisienne voudrait connaître le temps moyen du trajet en fonction de la distance à vol d'oiseau mesurée sur une carte. En une semaine, il a fait 23 observations et les a portées sur un diagramme de dispersion.

Distance (km)	Temps (minutes)	Distance (km)	Temps (minutes)
4	19	10	42
5	25	10	40
5	27	10	37
5	30	10	44
7	34	11	44
8	37	11	50
8	31	12	46
8	40	13	51
9	38	13	46
9	44	13	56
9	40	16	60
9	35		

Livraisons en région parisienne
Les distances sont rangées par ordre croissant

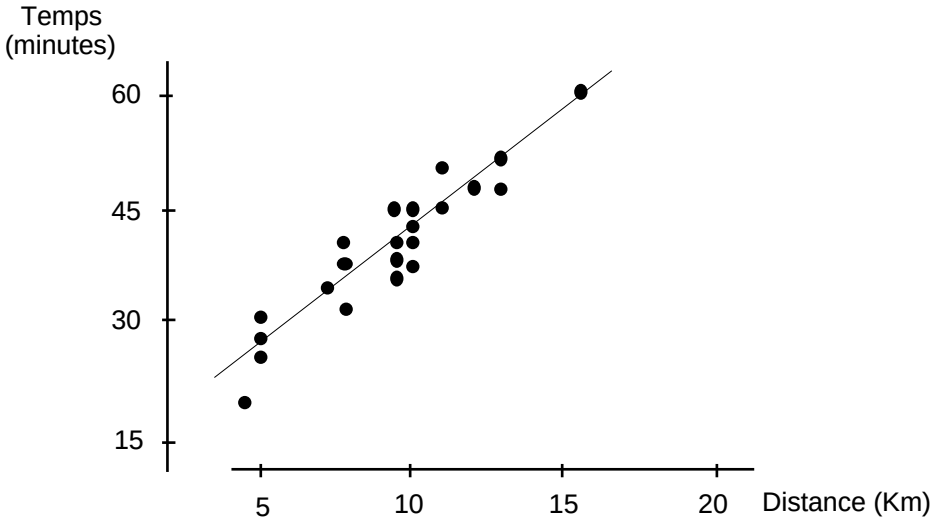
Sur le graphique (page suivante), les résultats forment un nuage de points. Dans le cas présent, il est certain que le temps moyen du trajet augmente en fonction de la distance, mais en général, pour s'assurer qu'il y a bien une relation, il est préférable de faire un test de corrélation. Le logiciel *Cora* donne ici une réponse positive. On peut donc tracer une droite qui passe au plus près de tous les points, ce qui permettra aux personnes qui font les livraisons d'utiliser le diagramme pour prévoir approximativement le temps de trajet. Cette droite est la *régression de l'échantillon de y en x.*

Si nous écrivons l'équation de la droite :

$$y = a + bx$$

la statistique permet de calculer les coefficients a et b correspondant à l'échantillon. Ce calcul peut se faire automatiquement sur une calculette de type Casio Graph 25 qui possède les fonctions statistiques à deux dimensions. On trouve, en valeurs arrondies :

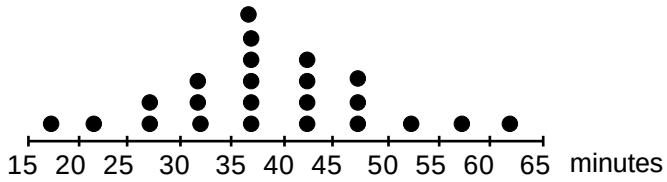
$$y = 11 + 3\,x$$



Régression linéaire du temps de livraison en distance

La condition de normalité

Les calculs précédents ne sont valables que si la distribution de l'échantillon de *y* est normale. Pour le vérifier, nous avons utilisé un papier à échelle de Henry dont le modèle se trouve sur la notice du logiciel *Cora*, disponible sur internet. Les points étant assez bien alignés, nous en avons conclu que cette condition était réalisée. D'ailleurs, celui qui n'a pas ce papier sous la main peut déjà estimer qu'une distribution est normale, avec un peu d'expérience, en regardant simplement l'histogramme des mesures.



Histogramme des temps de livraison

Un exemple fera comprendre l'importance de la condition de normalité. Une boutique de vêtements souhaite mettre en place un système de points de fidélité. Au préalable, le gérant étudie la relation entre la dépense totale d'un client en un mois et le nombre de visites dans le mois. Il fait 14 observations et les porte sur un diagramme de dispersion.

En regardant cette figure, chacun peut dire avec certitude que la dépense moyenne de chaque client augmente avec le nombre de visites. Cette impression est confirmée par *Cora* qui affiche : « l'hypothèse d'une corrélation est retenue ». Mais la distribution des dépenses, sur cet échantillon, n'est pas une distribution normale. On le voit déjà sur l'histogramme des dépenses, et ceci est confirmé par le graphique de Henry (page suivante). Il faut en tirer la conclusion qu'il n'est pas possible, à partir de cet échantillon, d'établir une régression de la dépense en nombre de visites. La méthode est inutilisable. Néanmoins, l'observation du graphique de Henry permet de faire une remarque intéressante.

Visites (nombre)	Dépense (euros)	Visites (nombre)	Dépense (euros)
2	82	7	190
2	105	7	218
2	105	8	173
3	92	8	202
4	108	8	206
4	112	9	238
6	135	9	246

Nombre de visites par mois et dépenses correspondantes
Les nombres de visites sont rangés par ordre croissant

L'échantillon est composé de deux groupes de points séparés par une classe vide (on dit que c'est une distribution *bimodale*). Nous pouvons penser que l'expérimentateur a observé par hasard deux types de clients ayant des habitudes différentes, mais seule une expérience avec un échantillon de plus grande taille permettrait de vérifier cette hypothèse. Supposons un instant que ce soit vrai. Il faudrait alors étudier la relation entre la dépense et le nombre de visites dans chaque catégorie. Le test de normalité conduit à accepter l'hypothèse que chaque groupe a une distribution normale, car nous voyons que les points sont assez bien alignés. En revanche, dans les deux cas, le test de corrélation avec le logiciel *Cora* rejette la possibilité d'une régression. Les enseignements que nous pouvons tirer de cette étude sont donc :

a. L'expérience suggère qu'il y a deux types de clients.
b. On ne constate pas de corrélation entre la dépense et le nombre de visites.
c. La dépense moyenne par client est estimée à :
109 € par mois pour l'un ;
210 € par mois pour l'autre.

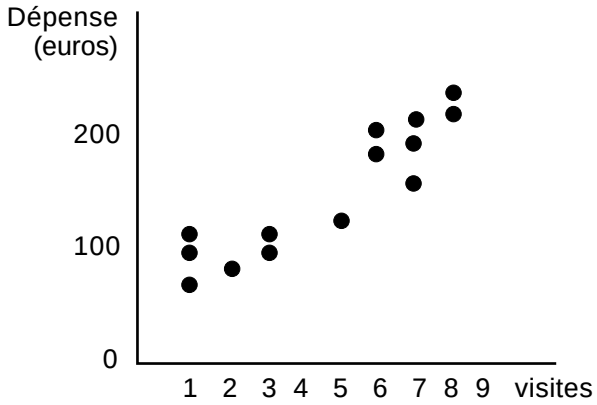
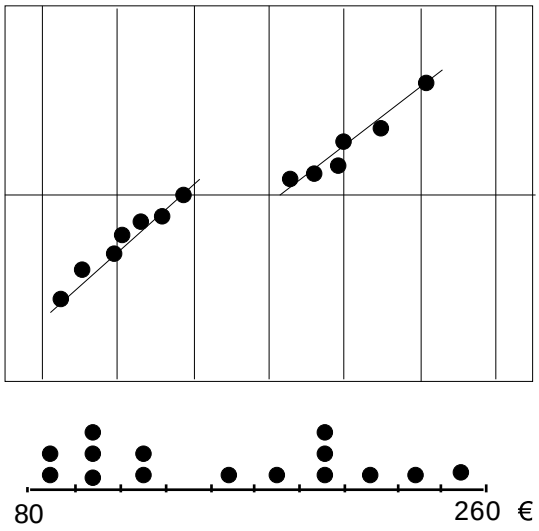


Diagramme de dispersion des dépenses des clients



Histogramme et graphique de Henry des dépenses des clients

Remarque 1. Sur un graphique de Henry, le dernier point de la distribution ne figure pas, car il correspond à la fréquence cumulée de 100%. D'autre part, quand deux points sont superposés sur un tel graphique, il est d'usage de les représenter par des cercles concentriques.

Remarque 2. Le deuxième exemple montre que l'utilisation de méthodes statistiques par une personne qui n'a pas de connaissances suffisantes peut conduire à des conclusions erronées. En particulier, le fait d'oublier ou de négliger la condition de normalité est la source d'un très grand nombre d'erreurs. Pour les éviter, il suffit de prendre l'habitude d'observer soigneusement les histogrammes et les graphiques avant de se lancer dans des calculs.

Test de corrélation paramétrique

L'observation du diagramme de dispersion ne permet pas d'évaluer avec une précision suffisante le degré de finesse de la relation linéaire entre deux variables. On peut avoir l'impression que cette relation existe quand une droite à pente positive est entourée d'un nuage de points allongé, mais il peut arriver malgré tout que les variables soient indépendantes. Le *coefficient de corrélation* permet de conclure avec certitude à l'existence d'une relation. Son expression est :

$$r = \text{Cov}\,(x, y) \;/ \sqrt{\text{Var}\,(x)\ \text{Var}\,(y)}$$

C'est un nombre sans dimension parce que le numérateur et le dénominateur sont formés des mêmes puissances de *x* et de *y*. Ce nombre est compris entre -1 et +1 . Quand il est positif, *x* et *y* varient dans le même sens ; quand il est négatif, ils varient en sens contraire. Le numérateur est appelé *covariance* de *x* et *y* . Au dénominateur, on retrouve les variances de ces deux variables.

On obtient facilement *r* avec une calculette de type Casio Graph 25. Le simple bon sens pourrait nous faire penser que les variables sont étroitement liées quand il est proche de -1 ou +1 et qu'elles sont indépendantes quand il est proche de zéro. Mais les choses ne sont pas aussi simples, car sa valeur dépend de la taille de l'échantillon. Pour trouver la limite de *r* à partir de laquelle l'hypothèse d'une corrélation peut être retenue, il faut se reporter à une table dont l'origine est due à Fisher.

Le calcul des coefficients de corrélation qui se trouvent dans cette table est basé sur la distribution normale à deux dimensions. Pour en comprendre le principe, imaginons une population où deux variables aléatoires *x* et *y* , indépendantes, et chacune de distribution normale, sont attribuées à chaque individu. Concrètement, chaque individu pourrait être une carte sur laquelle sont inscrites une valeur de *x* d'un côté, une valeur de *y* de l'autre. Après avoir tiré un échantillon de cette population, nous pouvons calculer le coefficient de corrélation des *x* et des *y*. C'est une variable aléatoire comprise entre -1 et +1. Pour le statisticien, le problème est de déterminer par un calcul théorique la probabilité que *r* soit inférieur en valeur absolue à une valeur donnée. Et quand nous lui apporterons nos résultats de mesure, après avoir calculé la valeur de *r*, il nous apprendra quelle est la probabilité de trouver une valeur aussi élevée si les deux variables sont indépendantes. On dit que le test de corrélation est le test de l'hypothèse nulle. C'est de cette façon qu'ont été calculées les tables donnant la limite de *r* (en valeur absolue) en fonction de la taille de l'échantillon. La table B à la fin du livre indique cette valeur pour une probabilité de dépassement de 0.05

La limite de *r* varie en sens inverse du nombre de données, c'est-à-dire du nombre de points du diagramme de dispersion. Le lecteur remarquera que l'entrée de la table n'est pas *N*, le nombre de données, mais un nombre marqué *d.l.* (degrés

de liberté). Cette notion sera expliquée au chapitre suivant. Pour l'instant, il suffit de savoir que *d.l.* = *N* - 2. En regardant la table B, on voit par exemple que la limite de *r* est 0,88 pour *N* = 5 et qu'elle tombe à 0,29 pour *N* = 50.

Remarque. Cette méthode fait apparaître pour la première fois la notion de risque d'erreur associé au test d'hypothèse. Les variations aléatoires des données expérimentales peuvent conduire à des erreurs de jugement sans que l'opérateur en soit responsable. L'erreur de première espèce consiste à rejeter l'hypothèse nulle quand elle est vraie. L'erreur de seconde espèce consiste à accepter l'hypothèse nulle quand elle est fausse. Le test ci-dessus limite à 5% le risque d'erreur de seconde espèce.

Test de corrélation non paramétrique

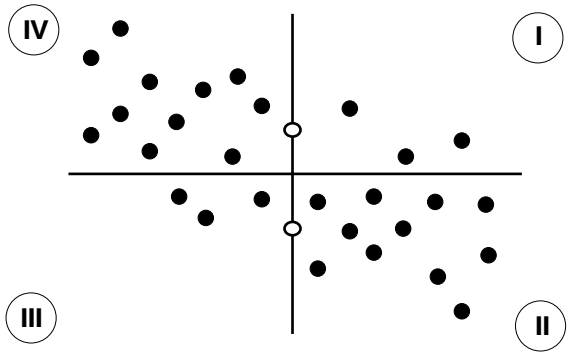
Avant que l'informatique n'apparaisse, les calculs de *a, b* et *r* étaient longs et fastidieux. Pour cette raison, les ingénieurs japonais avaient adopté dès 1950, sur les conseils de Deming, un test de corrélation nommé *test des signes.* Les calculs sont réduits au minimum car il suffit de compter des points sur le diagramme de dispersion pour aboutir à une conclusion. Cette méthode présente aussi l'avantage de ne pas imposer la condition de normalité des distributions (on dit que c'est un test non paramétrique).

Une société de mécanique étudie un traitement électrolytique contre la corrosion. Pour déterminer le temps de traitement, un échantillon de 30 pièces est soumis à des opérations de durée variable entre 12 et 18 minutes, puis placé dans une chambre climatique. On mesure le degré de corrosion au nombre de points de rouille par mètre carré. L'opérateur trace un diagramme de dispersion, avec la durée du traitement en abscisse et le nombre de points de rouille en ordonnée (page suivante). Il divise le diagramme en quatre parties séparées par les médianes horizontale et verticale. Il compte les points dans chaque partie, ceux qui sont sur les médianes n'étant pas comptés. Il additionne les dénombrements des quadrants diamétralement opposées (I et III d'une part, II et IV d'autre part). La somme de ces deux nombres est égale au nombre total de points moins les médianes. Une table déduite de la distribution binomiale indique un seuil de signification en fonction du nombre total de points moins les médianes. Si le plus faible de ces dénombrements lui est inférieur ou égal, l'hypothèse d'une corrélation est retenue (les statisticiens préfèrent dire que l'hypothèse nulle est écartée). La table C à la fin du livre donne les seuils avec un niveau de confiance de 95%, soit une probabilité de dépassement de 0.05.

Dans le cas présent, nous comptons 6 points dans la zone I+III, et 22 points dans la zone II+IV. Étant donné qu'il y a 2 points sur la médiane verticale, *n* = 28. La table C donne une limite de 8 points. L'hypothèse d'une corrélation est donc retenue.

L’efficacité du test des signes est inférieure d’environ 1/3 à celle du test paramétrique. A la différence de ce dernier, le rejet de l’hypothèse nulle n’implique pas que la régression soit linéaire.

Remarque. La meilleure méthode pour tracer la médiane horizontale consiste à déplacer lentement de bas en haut une règle placée horizontalement, et de compter les points à mesure qu’ils sont atteints. Même méthode pour la médiane verticale.



Test des signes pour l'étude d'un traitement anticorrosion

3

Comparaison de deux traitements

Nous avons vu au chapitre précédent comment l'analyse statistique vient confirmer que deux valeurs expérimentales sont liées de manière plus ou moins étroite, ce qui montre la possibilité d'une relation de cause à effet. Tous les auteurs soulignent, naturellement, que la corrélation entre deux valeurs ne signifie pas que la variation de l'une soit provoquée par la variation de l'autre. Cette réserve est importante, mais c'est ainsi que la statistique permet de faire progresser une recherche dans la bonne direction. Quittons ce problème à deux variables pour étudier un problème à une seule variable. Il s'agit de vérifier qu'il existe une différence significative entre deux échantillons issus d'un même processus ayant subi deux traitements différents. On peut remarquer que c'est encore, indirectement, un problème de causalité. En effet, dans le cas d'un test positif, le traitement est la cause de la différence. Comme précédemment, nous verrons que le test statistique comporte un risque d'erreur dont la probabilité est connue.

Propriétés d'un échantillon tiré d'une population normale

Quand un échantillon est tiré au hasard dans une population distribuée normalement, un calcul statistique assez simple permet de faire une *estimation ponctuelle* des paramètres μ et σ de la population. Le résultat pouvant varier d'un échantillon à l'autre, un calcul plus complexe, qui sort du cadre de cet ouvrage, permet de faire une *estimation par intervalle*.

L'écart-type théorique d'une population distribuée normalement, d'effectif N et de moyenne μ, est :

$$\sigma = \sqrt{\Sigma (x_i - \mu)^2 / N}$$

L'estimation ponctuelle de σ au moyen d'un échantillon d'effectif n tiré au hasard devrait utiliser une formule analogue dans laquelle N serait remplacé par n. Mais le calcul montre que si l'on effectue plusieurs tirages, la moyenne de cette estimation est inférieure à σ, ce qui signifie que l'estimateur est entaché d'une erreur systématique. Le seul estimateur *sans biais* de l'écart-type de la population est (la moyenne étant notée par x_b) :

$$s = \sqrt{\Sigma (x_i - x_b)^2 / (n - 1)}$$

Cette formule est déconcertante pour ceux qui commencent à étudier les statistiques. L'explication ne peut pas être donnée complètement ici, car elle est liée à des considérations d'algèbre linéaire. La quantité (n - 1) est appelée *nombre de degrés de liberté* de la fonction statistique s.

D'autre part, la moyenne x. d'un échantillon prélevé au hasard sur une population normale est un estimateur sans biais de la moyenne de la population. Si l'on effectue plusieurs tirages, elle est distribuée normalement. En pratique, une question qui se présente fréquemment est de savoir si un tel échantillon provient d'une population normale de paramètres μ et σ. Pour y répondre, W. S. Gossett a étudié la distribution de

$$t = (x_b - \mu) / s / \sqrt{n}$$

n étant l'effectif de l'échantillon, x_b l'estimation de la moyenne et s celle de l'écart-type.

Des tables de cette distribution, connue sous le nom de *distribution de Student*, ont été publiées. Un extrait de ces tables, le seuil de probabilité étant 5%, se trouve en annexe.

Nombre de degrés de liberté

Si un échantillon comporte n variables indépendantes, nous avons n possibilités de faire varier la moyenne en modifiant l'une de ces variables. C'est dans ce sens que les statisticiens disent que la moyenne comporte n degrés de liberté. Considérons maintenant la somme des carrés des écarts à la moyenne :

$$SCE = \Sigma (x_i - x_b)^2$$

Pour faire varier cette valeur en modifiant l'une des variables, la moyenne restant la même, il ne reste plus que (n - 1) possibilités, parce que le nombre de variables indépendantes n'est plus que (n - 1). C'est dans ce sens que les statisticiens disent que la *SCE* comporte (n - 1) degrés de liberté. La valeur s, qui est l'estimateur sans biais de σ, comporte elle aussi (n - 1) *d.l.* car elle peut s'écrire :

$$s = \sqrt{SCE / (n - 1)}$$

Une propriété importante du nombre de degrés de liberté est son additivité. On peut l'énoncer en disant que le nombre de *d.l.* d'une somme de fonctions est la somme des nombres de *d.l.* individuels. Nous aurons à nous servir de cette propriété plusieurs fois dans la suite du livre.

La distribution de Student

Considérons une population distribuée normalement, d'effectif *N*, de moyenne μ et d'écart-type σ, où 50 échantillons d'effectif *n* ont été tirés au hasard. L'histogramme des moyennes représente une distribution normale de moyenne μ et d'écart-type σ / *n*. La probabilité de trouver une moyenne d'échantillon comprise entre deux valeurs x_1 et x_2 peut être calculée au moyen des paramètres de la population. Mais inversement, si un échantillon d'effectif *n* est tiré au hasard d'une population inconnue, il est plus difficile d'estimer les paramètres de la population.

Si nous connaissions l'écart-type σ de la population, nous pourrions calculer les limites à 95% du paramètre μ avec une table de la distribution normale cumulée

$$x_b. - 1{,}96\ \sigma / \sqrt{n} \leq \mu \leq x_b. + 1{,}96\ \sigma / \sqrt{n}$$

Dans le cas général, σ étant inconnu, on utilise l'écart-type de l'échantillon et la variable de Student. Sur la table D, on voit que cette variable dépend d'un nombre de *d.l.* à déterminer. Dans le cas présent, *d.l.* = *n* - 1.

$$x_b. - t\ s / \sqrt{n} \leq \mu \leq x_b. + t\ s / \sqrt{n}$$

Pour les petits nombres de *d.l.*, la distribution de Student a un sommet plus pointu que la distribution normale et ses queues sont plus hautes. Elle tend vers la distribution normale quand le nombre de *d.l.* augmente.

Comparaison d'échantillons appariés

On dit que deux échantillons sont appariés quand, avec le même effectif, chaque valeur de l'un est couplée avec une valeur de l'autre. Dans cette méthode, le principe est de comparer la série des différences à une série de valeurs nulles. Il s'agit, précisons-le, de comparer la *moyenne.* La première étape consiste donc à calculer les différences entre les résultats pris deux à deux. La seconde consiste à calculer la valeur de *t* pour la série des différences. On cherchera enfin dans la table de Student (table D à la fin du livre) la valeur de *t* qui correspond à *d.l.* = *n* - 1. L'hypothèse nulle est rejetée, ce qui revient à dire que la différence entre les échantillons est significative, si la *valeur absolue* du nombre que l'on a trouvé est supérieure au nombre donnée par la table.

Dans usine de compresseurs, un atelier fabriquait des pièces de haute précision, et l'atelier voisin les assemblait. Les dimensions des pièces étaient mesurées à la sortie du premier atelier, puis à l'entrée du second. Un jour, une contestation s'est élevée entre les deux ateliers sur les résultats de mesure du diamètre des pistons.

Pour tenter de comprendre la situation, le bureau d'études a fait prélever un échantillon de 8 pièces sur lesquelles des numéros d'identification ont été gravés.

Transportées avec soin, les pièces ont été mesurées dans le premier atelier, puis dans le second. Les résultats sont inscrits sur le tableau suivant :

Pièce N°	Atelier A	Atelier B	Différence
1	25,31	25,18	0,13
2	25,20	25,17	0,03
3	25,18	25,14	0,04
4	25,17	25,11	0,06
5	25,09	25,10	-0,01
6	25,08	25,07	0,01
7	25,10	25,05	0,05
8	25,07	25,06	0,01

Diamètres mesurés sur 8 pièces (mm)

Nous laissons au lecteur le soin de vérifier s'il existe une corrélation entre ces deux séries de mesure. Le résultat, à vrai dire, n'a pas grande importance, car le but de cette étude n'est pas de chercher une corrélation, mais de savoir si la différence des moyennes est significative. Il faut commencer par vérifier sur un graphique de Henry que les valeurs de la colonne de droite sont distribuées normalement. Ceci étant, le lecteur se servira d'une calculette scientifique pour calculer la moyenne et l'écart-type des différences (sur la plupart des calculettes, l'estimation de l'écart-type n'est pas notée s mais σ_{n-1}). Nous obtenons :

$x_b = 0{,}04$
$s = 0{,}0431$

D'où : $t = x_b \sqrt{n} / s = 0{,}04 \times \sqrt{8} / 0{,}0431 = 2{,}63$

Pour 7 *d.l.*, la table de Student donne un seuil de signification de 2,36. La valeur absolue du nombre que nous avons trouvé étant supérieure, nous concluons que la différence des moyennes est significative, c'est-à-dire que les méthodes de mesure des ateliers A et B donnent des résultats systématiquement différents. Pour les mettre en harmonie, il faudra apporter une correction de 0,04 mm aux résultats de l'atelier B, mais il faudra également revoir les processus de mesure et d'étalonnage.

On arrive à la même conclusion avec le logiciel *Alice*, sans faire le moindre calcul. Il suffit d'inscrire les huit différences dans la colonne bleue et huit zéros dans la colonne rouge.

Comparaison d'échantillons indépendants

Nous sommes cette fois en présence de deux échantillons indépendants, dont les effectifs peuvent être inégaux, et dont les moyennes $x._1$ et $x._2$ sont les estimations respectives des moyennes des populations dont ils proviennent. On suppose qu'elles sont distribuées normalement. Le test de signification de la différence des moyennes est basé sur la formule :

$$t = [(x_{b1} - x_{b2}) - (\mu_1 - \mu_2)] / s_{1-2}$$

Les paramètres μ_1 et μ_2 sont les moyennes des populations dont proviennent les deux échantillons. La valeur s_{1-2} est l'estimation de l'écart-type de $(x._1 - x._2)$

Il s'agit de tester l'hypothèse nulle, c'est-à-dire l'hypothèse que les deux échantillons proviennent de la même population. La formule précédente devient donc :

$$t = (x_{b1} - x_{b2}) / s_{1-2}$$

Dans cette formule, n_1 et n_2 étant les effectifs des échantillons, SCE_1 et SCE_2 étant les sommes respectives des carrés des écarts à la moyenne :

$$s_{1-2} = \sqrt{} (SCE_1 + SCE_2) (n_1 + n_2) / n_1 n_2 (n_1 + n_2 - 2)$$

$$d.l. = n_1 + n_2 - 2$$

Un exemple Les élèves des classes préparatoires aux grandes écoles passent régulièrement chaque semaine des examens oraux appelés "colles". Un élève voudrait savoir si ses notes du mois d'avril marquent un progrès significatif par rapport au mois de mars. Les notes sont les suivantes :

Mars	Avril
15	12
14	18
10	15
	18
	12

Nous trouvons : $x._1 = 13$ $SCE_1 = 14$ $x._2 = 15$ $SCE_2 = 36$

$(s_{1-2})^2 = (14 + 36) \times (3 + 5) / (3 + 5 - 2) \times 3 \times 5 = 4,44$ $s_{1-2} = 2,11$

$t = -2 / 2,11 = - 0,95$

Il faut comparer cette valeur à celle qui se trouve sur la table de Student pour

$$d.l. = 3 + 5 - 2 = 6$$

La valeur trouvée sur la table est 2,45. La valeur absolue du nombre que nous avons trouvé est donc inférieure au seuil de signification. L'hypothèse nulle ne peut donc pas être rejetée, ce qui signifie en pratique que le test ne permet pas de conclure à une amélioration. Il ne faudrait pas en conclure que l'hypothèse d'une amélioration est exclue, car le résultat du test serait peut-être différent avec de plus grands échantillons et des variances plus faibles. Un test d'hypothèse laisse toujours la possibilité de rejeter l'hypothèse nulle en élargissant la base de données.

Le logiciel *Alice* permet d'arriver au même résultat. En outre, avec ce logiciel, il est facile de faire rapidement de nombreuses simulations en faisant varier les effectifs des séries et les variances. On trouve par exemple une différence significative quand on ajoute 10, 11, 12 aux notes de mars et 16 aux notes d'avril.

Test de comparaison non paramétrique

Le *test de comparaison des rangs* a pour avantage de s'accommoder de distributions quelconques ; il permet ainsi d'éviter le test de normalité. De plus, il ne demande que des calculs simples. Comme précédemment, il faut utiliser une table. Dans ce livre, les effectifs de la table sont limités à 10-15 pour un échantillon et 10-20 pour l'autre. Passons tout de suite à un exemple : dans une usine de semiconducteurs, on veut comparer deux procédés de fabrication différents en comptant le nombre de défauts d'une plaque de silicium. Les échantillons sont de 10 et 12 pièces. La méthode est la suivante :

1. On prépare un tableau composé de deux colonnes de résultats séparées par une colonne vide d'environ 5 cm (page suivante). Les résultats des deux échantillons sont rangés par ordre croissant.

2. On trace des traits reliant les valeurs des deux colonnes dans un ordre croissant.

3. On écrit les rangs en regard des résultats. Si plusieurs résultats sont identiques, on leur affecte un "rang moyen" qui n’est pas forcément un nombre entier. Par exemple, le rang moyen de 7 et 8 sera 7,5.

4. On calcule la somme des rangs du plus petit échantillon. Dans l'exemple que nous avons choisi, c’est l’échantillon B. Le résultat est : 122,5.

5. On compare ce résultat avec la table du test de comparaison des rangs (table E). S’il est extérieur à l’intervalle trouvé sur la table, on peut retenir l'hypothèse

d'une différence des moyennes. Dans le cas contraire, on considère que la différence n'est pas significative. Dans le cas présent, les limites qui correspondent à des effectifs de 10 et 12 sont : 85-145.

Le résultat est dans l'intervalle de la table. Nous en déduisons que la différence entre les deux échantillons n'est pas significative, bien que la moyenne de B soit très supérieure. En l'absence de toute autre information faisant pencher la balance, tel que le coût de l'opération par exemple, il est donc indifférent d'utiliser le procédé A ou le procédé B.

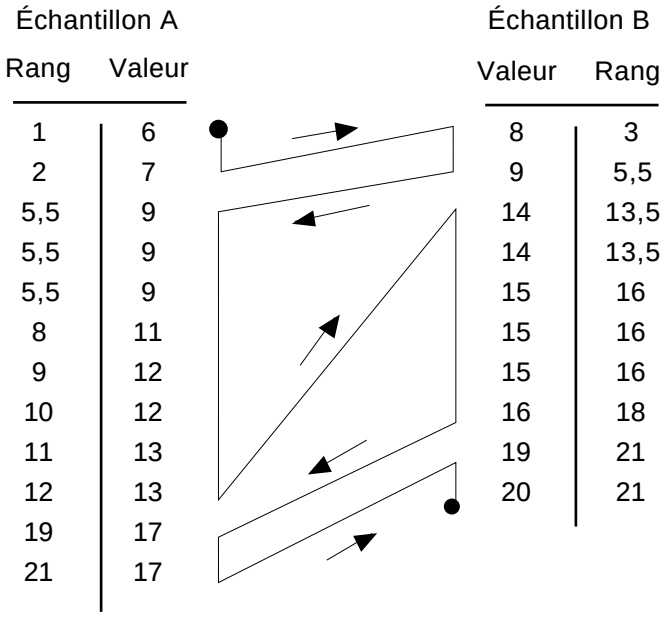
Si les deux échantillons étaient distribués normalement (ce qui n'est pas le cas ici), le test de Student aboutirait à la même conclusion. Mais il faut remarquer que le test de comparaison des rangs a une efficacité plus faible d'environ 30%.

Si les effectifs des échantillons sont supérieurs à ceux qui se trouvent dans la table, on peut calculer les seuils, en fixant à 0.05 la probabilité de dépassement bilatéral, au moyen de la formule suivante, qui donne la limite inférieure :

$$U_{\min} = (n_1 \ n_2 \ + 1) / 2 - 2 \sqrt{ n_1 n_2 \ (n_1 + n_2 \ + 1) / 12}$$

Pour calculer la limite supérieure à partir de la limite inférieure, il suffit d'utiliser la formule qui donne la somme d'une série limitée de nombres entiers. On peut vérifier que :

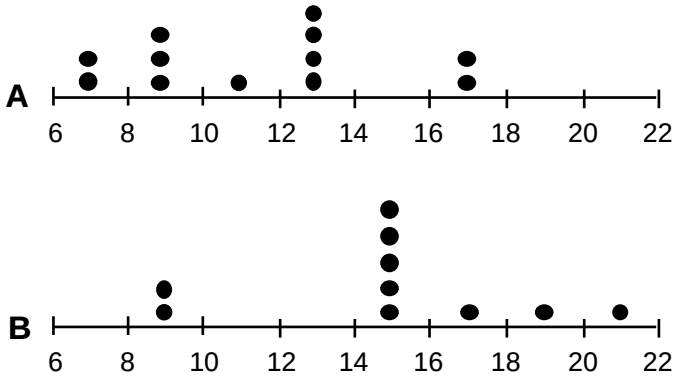
$$U_{\min} + U_{\max} = (n_1 + n_2) (n_1 + n_2 + 1) / 2$$



Somme des rangs de B : 122,5

Schéma du test de comparaison des rangs

Remarque. L'observation des histogrammes des deux échantillons est toujours instructive. Quand les distributions ne sont pas homogènes, comme dans l'exemple précédent, il est plus important de chercher à comprendre pourquoi les points sont répartis irrégulièrement que de savoir uniquement que la différence des moyennes n'est pas significative. C'est une information qui peut conduire à l'amélioration d'un processus de fabrication.



Histogrammes des échantillons A et B

4

Analyse de la variance

Nous avons étudié au chapitre précédent des méthodes permettant de comparer les moyennes de deux échantillons ayant subi deux traitements différents. Nous allons maintenant étendre le problème au cas d'un nombre quelconque de traitements. Il serait possible de comparer les échantillons deux à deux, ce qui nécessiterait 3 tests pour 3 traitements, 6 pour 4 traitements, etc. Les calculs seraient longs. Au début du XX^{e} siècle, le chercheur anglais Ronald Fisher a imaginé une méthode plus rapide permettant de comparer tous les échantillons en une seule fois, et d'identifier ceux qui présentent des différences significatives. Ce progrès décisif en statistique se nomme *analyse de la variance.*

Comparaison de plusieurs échantillons

La comparaison de deux échantillons d'origine différente est une expérience à une entrée et deux traitements. Par définition, une entrée est un facteur de variation. Un traitement est une caractéristique particulière de cette entrée. Les traitements doivent être indépendants. Ainsi, dans le premier exemple du chapitre précédent, la mesure du diamètre était une entrée, et les processus de mesure des ateliers A et B étaient des traitements. Nous allons maintenant étudier le problème général d'une expérience à une entrée et plusieurs traitements, c'est-à-dire que nous allons comparer simultanément plus de deux échantillons. La méthode sera présentée avec quatre traitements, c'est-à-dire avec quatre échantillons.

Une expérience à une entrée et quatre traitements

Dans le cadre d'une étude sur la coagulation du sang, 25 rats ont été soumis à quatre régimes alimentaires différents ABCD. Le but était de savoir si la vitesse de coagulation dépendait du régime alimentaire. Les données du tableau ci-dessous représentent, en secondes, les temps de coagulation. On constate une différence appréciable entre les valeurs des quatre séries, malgré un certain chevauchement. Le lecteur peut comparer graphiquement les échantillons en construisant des histogrammes comme au chapitre précédent.

Avec une calculatrice scientifique en mode SD, nous obtenons la somme des valeurs Σx, la moyenne x_b. et la somme des carrés Σx^2 de chaque échantillon. Ces

nombres figurent en bas du tableau. La somme des carrés des écarts à la moyenne est donnée par la formule :

$$SCE = \Sigma (x^2) - (\Sigma x)^2 / n$$

Une méthode plus simple, donnant le même résultat, consiste à sortir l'écart-type sur la même calculatrice avec la touche marquée σ_n puis à effectuer l'opération $SCE = n (\sigma_n)^2$

On se rappellera que le nombre de *d.l.* de chaque échantillon est égal à (n - 1) et que le nombre total de *d.l.* est la somme des *d.l.* des échantillons.

Régime	A	B	C	D	Total
	62	69	66	61	
	56	65	62	62	
	60	69	64	60	
	63	64	70	63	
	64	68	67	59	
	63	67	63		
	59		63		
Σx	427	402	455	305	1589
x_b	61	67	65	61	254
Σx^2	26 095	26 956	29 623	18 615	101 289
$(\Sigma x)^2 / n$	26 047	26 934	29 575	18 605	100 997
SCE	48	22	48	10	292
d.l.	6	5	6	4	21

Vitesse de coagulation du sang et régime alimentaire

Le problème est de savoir si les différences des moyennes sont significatives. L'analyse de la variance est fondée sur le fait qu'avec plusieurs traitements, il existe trois façons d'estimer la variance. On se place dans l'hypothèse nulle, c'est-à-dire que toutes les données sont supposées provenir d'une même population normale. Dans tous les cas, la variance s'exprime par

$$s^2 = SCE / d.l.$$

Première estimation. En comparant les 25 données à la moyenne générale, toutes classes confondues, nous calculons la *SCE* totale :

$62^2 + 56^2 + 60^2 + ... + 59^2 - 1\,589^2 / 25 = 101\,289 - 100\,997 = 292$

Cette *SCE* ayant 24 *d.l.*, la première estimation de la variance est :

$$292 / 24 = 12{,}17$$

Deuxième estimation. Considérons les moyennes des quatre échantillons : 61, 67, 65 et 61. La *SCE* inter-classes est la somme pondérée des carrés des écarts de ces quatre valeurs à la moyenne générale, soit :

$7x61^2 + 6x67^2 + 7x\ 65^2 + 5x\ 61^2 - 1\,589^2 / 25 = 101\,161 - 100\,997 = 164$

Cette *SCE* ayant 3 *d.l.*, la deuxième estimation de la variance est :

$$164 / 3 = 54{,}67$$

Troisième estimation. La *SCE* intra-classes est simplement la somme des *SCE* des échantillons, soit 48 + 22 + 48 + 10 = 128

Le nombre de *d.l.* est la somme des *d.l.* des échantillons, soit 6 + 5 + 6 + 4 = 21. La troisième estimation de la variance est donc :

$$128 / 21 = 6{,}10$$

Portons ces résultats sur un *tableau d'analyse de la variance*. Il est convenu de nommer *carré moyen* l'estimation de la variance.

source de variation	***SCE***	***d.l.***	carré moyen
Inter-classes	164	3	54,67
Intra-classes	128	21	6,10
Total	292	24	12,17

Analyse de la variance

La principale propriété de ce tableau est l'additivité des *SCE* d'une part, des *d.l.* d'autre part. Cette propriété se démontre mathématiquement. En conséquence, la méthode habituelle d'analyse de la variance consiste à calculer d'abord la *SCE* totale et la *SCE* inter-classes. La *SCE* intra-classes est obtenue ensuite par différence.

Si l'hypothèse nulle était vraie, les carrés moyens inter-classes et intra-classes seraient sensiblement égaux. Le rapport du premier au second est donc un bon critère pour tester l'hypothèse nulle. Historiquement, c'est l'Anglais Ronald Fisher qui a eu l'idée d'étudier la distribution de ce critère pour des échantillons tirés d'une

population normale. George Snedecor, qui en a publié une table quelques années plus tard, l'a nommée pour cette raison *distribution F*. Un extrait, le seuil de probabilité étant 5%, se trouve en annexe.

Dans le cas présent, $F = 54{,}67 / 6{,}10 = 8{,}97$

La limite supérieure de *F* donnée par la table, pour 3 *d.l.* au numérateur (première ligne) et 21 *d.l.* au dénominateur (première colonne), est 3,07. L'hypothèse nulle est donc rejetée, ce qui signifie qu'il y a des différences significatives entre les moyennes des traitements. Le test ne précise rien de plus, mais l'observation des moyennes montre que les régimes alimentaires B et C sont probablement en cause. Pour vérifier ces hypothèses, il suffit de comparer séparément la moyenne de B, puis la moyenne de C, à la moyenne de l'ensemble A et D. Le logiciel *Alice* permet de faire rapidement le test. La conclusion de cette étude est que les régimes B et C donnent une vitesse de coagulation nettement supérieure à celle des régimes A et D, mais qu'il n'y a pas de différence significative entre B et C d'une part, A et D d'autre part.

Une expérience à quatre traitements et cinq blocs

Les expériences à deux entrées sont fréquentes en économie. Par exemple dans une enquête sur les dépenses des ménages, il est habituel de classer les ménages à la fois par nombre de personnes et niveau de revenu. Le nombre d'échantillons est souvent assez élevé, car si toutes les combinaisons sont représentées, c'est le produit du nombre de traitements de l'une par celui de l'autre. Nous allons présenter une expérience agronomique à une entrée avec quatre traitements et cinq répétitions disposées suivant autant de blocs. Elle comporte donc 20 échantillons. On pourrait dire aussi que c'est une expérience à deux entrées, en considérant que les blocs forment une entrée. Mais il faut d'abord comprendre ce qu'est un bloc.

Dans une expérience agronomique, le terrain est divisé en parcelles, nommées "blocs". Chaque traitement est répété sur différents blocs. Les blocs sont disposées de telle façon que les conditions de culture soient aussi uniformes que possible, et les traitements sont répartis au hasard sur les blocs. Ce plan d'expérience est dit « par blocs randomisés ». Le concept s'est étendu à d'autres domaines, en sorte qu'un bloc peut être aussi une provenance de matière première, une tranche horaire de la journée, etc. Le but de la randomisation est d'éviter que des variations de caractéristiques non contrôlées dans les blocs aient une influence sur les conclusions de l'étude.

En d'autres termes, la méthode a pour but d'éliminer une éventuelle perturbation due à un facteur n'ayant pas un intérêt immédiat. Le cas peut se présenter, par exemple, quand on doit faire une série de mesures physiques en utilisant une batterie dont la tension baisse au cours du temps. Le facteur temps sera mis en blocs randomisés pour éliminer l'influence de la batterie.

Le tableau ci-dessous représente une expérience où quatre traitements de semences de soja avec des engrais chimiques (ABCD) sont comparés à un témoin non traité (T). Cent graines ont été semées par parcelle. Les données sont les nombres de plantes qui n'ont pas levé.

Répétitions Traitements	1	2	3	4	5	Total	Moyenne
T	8	10	12	13	11	54	10,8
A	2	6	7	11	5	31	6,2
B	4	10	9	8	10	41	8,2
C	3	5	9	10	6	33	6,6
D	9	7	5	5	3	29	5,8
Total	26	38	42	47	35	188	

Terme correctif $\boldsymbol{C} = 188^2 / 25 = 1\,413{,}76$

SCE totale $8^2 + 2^2 + \ldots + 6^2 + 3^2 - \boldsymbol{C} = 220{,}24$

SCE traitements $(54^2 + 31^2 + \ldots + 29^2) / 5 - \boldsymbol{C} = 83{,}84$

SCE répétitions $(26^2 + 38^2 + \ldots + 35^2) / 5 - \boldsymbol{C} = 49{,}84$

Sources de variation	***d.l.***	***SCE***	carré moyen
Traitements	4	83,84	20,96
Répétitions	4	49,84	12,46
Err. aléatoire	16	86,56	5,41
Total	24	220,24	

Expérience à une entrée avec 5 traitements et 5 blocs randomisés
Les répétitions sont les numéros des blocs

La première étape consiste à calculer le terme correctif $(\Sigma X)^2 / n$ ainsi que la *SCE* totale, la *SCE* traitements et la *SCE* répétitions, comme s'il s'agissait d'une expérience à une entrée. On trouve ensuite par différence la *SCE* de l'erreur aléatoire. Puis on porte sur le tableau les nombres de *d.l.*, qui sont respectivement 4, 4, 16 et 24. On calcule enfin les carrés moyens. Comme dans l'expérience à une entrée, le carré moyen de l'erreur aléatoire sera pris comme dénominateur de *F*.

Le *F* des traitements est 20,96 / 5,41 = 3,87 alors que la table, pour 4 et 16 *d.l.*, nous donne une limite supérieure de 3,01. Il faut en conclure que certains traitements ont une influence significative. Le problème est de les identifier. Or si nous faisons un test de comparaison sur les traitements B et D, qui sont aux deux extrémités de l'échelle des moyennes à l'exclusion du témoin T, nous trouvons que la différence de leurs moyennes n'est pas significative. Donc tous les traitements ont une influence significative sans qu'aucun d'eux ne soit privilégié.

D'autre part, le *F* des répétitions est 12,46 / 5,41 = 2,30. Il est donc inférieur au seuil de signification de la table, mais ceci montre néanmoins que la variance des répétitions est environ quatre fois supérieure à la variance résiduelle. Les différences entre les blocs ne sont certainement pas négligeables.

Une expérience à deux entrées, 3 x 4 traitements

Quand l'une des deux entrées n'est pas simplement faite de répétitions par blocs randomisés, comme dans l'exemple précédent, mais constitue un facteur de variation qui présente autant d'intérêt que le premier, cette organisation prend le nom de *plan factoriel M* ***x*** *N*, ces deux lettres représentant les nombres de traitements dans l'une et l'autre entrée. Nous allons donner ici un exemple de plan factoriel 3 x 4 qui provient de l'industrie textile. Il faut remarquer au préalable que ce plan d'expérience comporte un élément nouveau. En effet, l'interaction des deux facteurs est une possibilité qui n'existait pas dans les expériences précédentes. Les calculs seront donc modifiés pour pouvoir tester l'hypothèse d'une interaction.

Une usine textile se propose de tester plusieurs types de fil et plusieurs types de métiers à tisser dans la mise au point d'un tissu destiné à faire des vestes d'escrime. Plutôt que de conduire des expériences séparées pour choisir le fil d'une part et le métier à tisser d'autre part, le bureau d'études combine les deux expériences en une seule pour faire un plan factoriel 3 x 4 avec 4 mesures sur chaque échantillon. Le résultat comportera donc 3 x 4 x 4 = 48 données. On mesure la résistance à la perforation par impact, qui s'exprime en Newtons, conformément à une norme européenne. Le résultat est présenté sur le tableau ci-dessous :

Métier Fil	A	B	C	D
1	431	482	443	445
	445	450	445	471
	446	488	463	466
	443	472	476	462
2	436	486	444	456
	429	461	435	474
	440	449	431	485
	423	453	440	448
3	428	430	431	430
	421	437	429	436
	435	438	426	431
	423	429	438	433
Moyennes				
1	441,2	473,0	456,5	461,0
2	432,0	462,2	437,5	466,0
3	426,8	433,5	431,0	432,5

On calcule d'abord la *SCE* totale à partir des données brutes. Les moyennes serviront ensuite à calculer la *SCE* inter-classes. Nous recommandons d'utiliser une calculatrice scientifique de type Casio Collège.

Pour calculer la *SCE* totale, il faut entrer en mémoire les 48 données brutes, sortir l'écart-type avec la touche marquée σ_n puis faire l'opération $SCE = n (\sigma_n)^2$. Ensuite, pour calculer la *SCE* inter-classes, il faut entrer en mémoire les 12 moyennes, sortir l'écart-type avec la touche marquée σ_n puis faire l'opération $SCE = 4\ n (\sigma_n)^2$. La somme doit être multipliée par 4 pour que les valeurs obtenues soient homogènes, car chaque moyenne représente 4 données brutes. Pour terminer, la *SCE* erreur aléatoire est obtenue par différence. Les résultats sont :

SCE totale = 16 116 *SCE* inter-classes = 11 800 Err. aléatoire = 4 316

Il ne reste plus qu'à déterminer la *SCE* métier, la *SCE* fil et la *SCE* interaction. Pour cela, on commence par calculer les moyennes des lignes et des colonnes du tableau des moyennes. Ensuite, sur la calculatrice, on entre en mémoire les moyennes des 4 colonnes métier. Puis on fait apparaître l'écart-type pour faire l'opération $SCE = k\ n (\sigma_n)^2$. Le multiplicateur k est le nombre de données brutes qui appartiennent à une colonne. Donc $k = 12$. Le processus est évidemment le même pour la *SCE* fil, avec 3 lignes au lieu de 4 colonnes, et $k = 16$.

	A	B	C	D	moyenne
1	441,2	473,0	456,5	461,0	457,9
2	432,0	462,2	437,5	466,0	449,4
3	426,8	433,5	431,0	432,5	431,0
moyenne	433,3	456,2	441,7	453,2	

SCE métier = *SCE* (33,3 ; 56,2 ; 41,7 ; 53,2) x 12 = 4 027

SCE fil = *SCE* (57,9 ; 49,4 ; 31,0) x 16 = 6 050

On démontre que dans une expérience factorielle à deux entrées, la *SCE* inter-classes est la somme des facteurs et de l'interaction. La SCE interaction se calcule donc par différence :

SCE interaction = 11 800 - 4 027 - 6 050 = 1 723

D'autre part, d'après les calculs précédents, il est clair que les nombres de *d.l.* sont :

- pour la *SCE* totale 48 - 1 = 47
- pour la *SCE* inter-classes 12 - 1 = 11
- pour la *SCE* métier 4 - 1 = 3
- pour la *SCE* fil 3 - 1 = 2

Les autres *d.l.* se déduisent par différence.

On porte ces résultats sur un tableau d'analyse de la variance :

source de variation	***SCE***	***d.l.***	carré moyen
Métier	4 027	3	1 342
Fil	6 050	2	3 025
Interaction	1 723	6	287
Inter-classes	11 800	11	
Err. aléatoire	4 316	36	120
Total	16 116	47	

Résistance du tissu - Plan factoriel 3 x 4
Analyse de la variance

Les valeurs de *F* pour le métier et pour le fil sont très supérieures à la limite donnée par la table. Les différences des moyennes sont donc très significatives. Le tableau des moyennes suggère que le métier B et le fil 1 se détachent nettement en tête des performances. Un test de comparaison utilisant le logiciel *Alice* confirmera cette hypothèse. En revanche, le test de *F* montre que l'interaction n'est pas significative.

Remarque. On pourrait penser que l'observation du tableau des moyennes suivi d'un test de comparaison suffit pour une telle étude, et que, par conséquent, le test de *F* n'était pas nécessaire. Or dans une expérience comportant plus de deux traitements, il peut arriver qu'un test de comparaison retienne l'hypothèse d'une différence significative alors que le test de *F* donne le résultat inverse. Dans ce cas, c'est la conclusion du test de comparaison qui est erronée. Le test de *F* est donc une condition préalable.

Une expérience à trois entrées

Les chercheurs ont parfois besoin de faire des expériences à trois entrées comportant deux familles de blocs randomisés. Chacune des entrées est expérimentée avec le même nombre de modalités, de telle sorte que chaque modalité de chaque entrée soit associée une fois et une seule à chaque modalité des autres entrées.

Par exemple, dans une enquête sur la perception de la publicité par les spectateurs d'une salle de cinéma, on a testé quatre types de publicités : A, B, C, D. L'expérience comporte deux familles de blocs randomisés, l'une étant formée de 4 films et l'autre de 4 salles. Le nombre de personnes interrogées doit être suffisant pour que tous les films et toutes les salles soient représentés. Avec 4 films et 4 salles, nous avons 16 combinaisons. Le plus petit nombre de personnes interrogées est donc 16 (mais on peut aussi prendre un multiple de 16). Ces personnes doivent répondre à un questionnaire en sortant du cinéma. Pour éviter que le résultat soit faussé par l'influence des films et des salles, les combinaisons sont disposées au hasard sur un tableau carré où chaque type de publicité apparaît une seule fois sur une ligne et une seule fois sur une colonne. On pourrait dire que c'est une double randomisation. La figure est un *carré latin.*

Grâce à cet artifice, l'expérience est menée de la même façon qu'une expérience à deux entrées avec une seule famille de blocs randomisés. Le but de l'enquête n'était pas de savoir si telle salle ou tel film était plus favorable qu'une autre salle ou un autre film à l'impact de la publicité sur le spectateur. Néanmoins l'expérience permet de répondre à la question avec le calcul de la *SCE* lignes et de la *SCE* colonnes.

Salles Films	1	2	3	4
I	A	B	C	D
II	C	D	A	B
III	B	C	D	A
IV	D	A	B	C

Carré latin 4 x 4 pour une enquête sur la publicité

Les carrés latins ont d'abord été utilisés en agronomie, et le carré représentait effectivement une répartition de parcelles d'un champ entre les différentes modalités d'un traitement. Comme dans la méthode des blocs randomisés, on introduit le hasard dans l'expérience en tirant au sort le carré latin qui sera utilisé parmi tous les carrés latins de même dimension.

La moindre interaction entre deux traitements fausse les résultats. Par conséquent, les carrés latins ne doivent être utilisés que si l'on est suffisamment sûr de l'absence d'interactions. C'est une condition qui limite sensiblement l'emploi de ce type de plan.

5

Plans factoriels complets à deux niveaux

Le chapitre précédent montre comment on peut comparer des échantillons ayant subi des traitements différents, d'abord sur une seule entrée, puis sur deux entrées. Par exemple, quand nous comparons quatre engrais différents dans une culture de soja, l'entrée unique est le type d'engrais. Puis quand nous comparons quatre métiers à tisser et trois types de fil dans une usine textile, les deux entrées sont le type de métier et le type de fil. Étant donné que le nombre de traitements augmente rapidement avec le nombre d'entrées, on a tendance à limiter ce nombre à deux traitements par entrée. On arrive ainsi à la notion d'expérience factorielle. Le but est toujours d'identifier les différences significatives, mais la limitation du nombre d'essais devient une préoccupation majeure.

La notion d'expérience factorielle

L'analyse de la variance à deux entrées permet de déterminer les effets de deux types de variables, appelés entrées, ou *facteurs*, sur une valeur mesurée, appelée *réponse*. Quand les facteurs sont plus nombreux, par exemple les conditions de croissance d'une plante (hygrométrie, ensoleillement, composition du sol, engrais, etc.), le problème devient plus complexe. C'est pourquoi, vers 1920, Fisher et Yates ont créé pour les besoins de la recherche agronomique une nouvelle approche nommée *expérience factorielle à deux niveaux.* Cette méthode s'est largement répandue dans le monde industriel à partir de 1960, grâce aux efforts de l'ingénieur japonais Genichi Taguchi.

Les ingénieurs d'études des grandes sociétés ont souvent à faire le choix entre plusieurs matériels, plusieurs réglages, plusieurs matières premières, etc. C'est un problème qui se présente invariablement au cours du développement d'un nouveau produit, notamment dans l'électronique ou l'automobile. La méthode permettant de trouver une solution optimale est l'expérience factorielle, l'outil mathématique le plan factoriel. Une expérience factorielle peut comporter jusqu'à 10 facteurs, ou même davantage. Le nombre d'essais se détermine à partir du nombre de facteurs et du nombre de traitements, qui prennent alors le nom de *niveaux.* Il n'est pas difficile de calculer qu'avec 7 facteurs et 2 niveaux pour chaque facteur, une expérience factorielle nécessitera 132 essais pour que toutes les combinaisons de facteurs et de niveaux soient réalisées. On comprend pourquoi, dans ces conditions, le nombre de niveaux est généralement réduit à 2.

Quand toutes les combinaisons de facteurs et de niveaux sont réalisées, nous avons un *plan factoriel complet.* Fisher a démontré qu'il est encore possible, au prix de quelques restrictions que nous étudierons au chapitre suivant, de trouver l'influence des facteurs en n'utilisant pas toutes les combinaisons de facteurs et de niveaux, afin de diminuer le nombre d'essais. C'est le principe du *plan factoriel fractionnaire*.

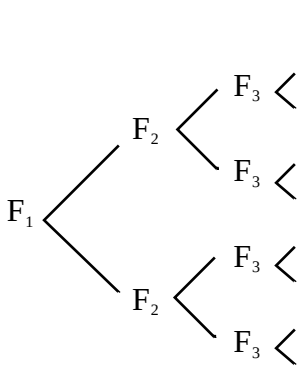
Autrefois, on étudiait un seul facteur à la fois. Mais l'expérience factorielle présente le grand avantage de demander moins d'essais pour obtenir des estimations avec une précision (une variance) déterminée. Pour comprendre cela, prenons le cas d'une expérience factorielle à 2 niveaux et 2 facteurs, avec *n* répétitions. C'est une expérience classique à 2 entrées, qui comporte donc 4 x *n* essais. Soient A et B les facteurs, 1 et 2 les niveaux. On peut démontrer que si les données proviennent d'une population normale de variance σ^2, les variances de l'entrée A et de l'entrée B sont toutes deux égales à σ^2 / n.

Voyons maintenant le cas de la comparaison des niveaux 1 et 2 du facteur A pour le même niveau de B. C'est l'une des 4 comparaisons que l'on peut faire quand on étudie un seul facteur à la fois. On peut démontrer que la variance de la différence est $2 s^2 / n$. Donc il nous faudrait deux fois plus de répétitions que dans l'expérience factorielle pour estimer la différence avec la même précision. En définitive, la supériorité de l'expérience factorielle vient de ce que chaque donnée est utilisée pour estimer l'effet de chaque facteur, alors que dans la méthode "un facteur à la fois", chaque donnée n'apporte d'information que sur l'effet d'un seul facteur.

Un plan factoriel complet à 3 facteurs

Le graphique 5.1 présente l'organisation d'un plan factoriel complet à 3 facteurs et 2 niveaux. Il comporte $2^3 = 8$ essais. Chaque facteur F_1, F_2 ou F_3 peut occuper le niveau 1 ou le niveau 2. Ce sont des symboles arbitraires qui désignent par exemple deux réglages d'une même machine, deux fournisseurs d'une même pièce, etc. Dans la suite du chapitre, nous les remplacerons par les signes (+) et (-). Chaque ligne définit un essai. Par exemple sur la troisième ligne, on voit que F_1 a le niveau 1, tandis que F_2 a le niveau 2 et F_3 le niveau 1.

Quand les essais seront terminés, nous aurons 8 résultats de mesure. Il faudra faire le traitement algébrique de ces données pour calculer les effets principaux et les interactions. On utilise pour cela une *matrice orthogonale* dont nous allons expliquer la construction.



F_1	F_2	F_3
1	1	1
1	1	2
1	2	1
1	2	2
2	1	1
2	1	2
2	2	1
2	2	2

Organisation d'un plan factoriel complet à 3 facteurs

Écrivons les 3 colonnes précédentes dans un ordre différent, en désignant les facteurs par A, B, C, et en remplaçant les symboles 1 et 2 par les symboles (+) et (-). Ajoutons les 4 colonnes d'interactions. La colonne AB s'obtient en multipliant deux à deux les signes des colonnes A et B, suivant les règles habituelles :

(+1) x (+1) = (+1) (+1) x (-1) = (-1) (-1) x (+1) = (-1) (-1) x (-1) = (+1)

Les colonnes AC, BC et ABC s'obtiennent de la même façon. Le résultat final est un tableau à 8 lignes et 7 colonnes. On peut constater que dans toutes les colonnes, le nombre de (+) est égal au nombre de (-). Chaque ligne sera utilisée, comme nous le verrons plus loin, pour calculer une combinaison linéaire des données.

Essais	A	B	C	AB	AC	BC	ABC
1	-	-	-	+	+	+	-
2	+	-	-	-	-	+	+
3	-	+	-	-	+	-	+
4	+	+	-	+	-	-	-
5	-	-	+	+	-	-	+
6	+	-	+	-	+	-	-
7	-	+	+	-	-	+	-
8	+	+	+	+	+	+	+

Matrice orthogonale d'un plan factoriel complet à 3 facteurs

Remarque. La construction d'une matrice orthogonale avec 4 facteurs principaux ABCD se ferait suivant la même méthode. On obtiendrait un tableau avec 16 lignes et 15 colonnes, et ainsi de suite.

Un exemple industriel. Une société de production de matière plastique étudie une réaction chimique dont les facteurs de variation A, B et C sont respectivement la température, le dosage du catalyseur et l'origine de la matière première. On mesure le taux d'impuretés du produit de la réaction, en g/Kg. Pour les facteurs A et B, les niveaux (-) et (+) correspondent aux conditions extrêmes dans lesquelles la réaction se fait encore normalement. Pour le facteur C, ils correspondent aux deux fournisseurs de la société. Le but de l'expérience est de s'approcher des conditions optimales. S'il est établi que les facteurs A ou B ont un effet significatif, il faudra pousser l'étude plus loin en cherchant le meilleur point de réglage. Un calcul de régression sera nécessaire. Si le facteur C a un effet significatif, il faudra certainement abandonner l'un des fournisseurs.

Les résultats de mesure sont :

Essai	1	2	3	4	5	6	7	8
Résultat	42	39	55	54	51	43	51	51

Les coefficients d'une colonne de la matrice orthogonale étant notés $c_1\ c_2\ ...\ c_8$, et les résultats de mesure étant notés $x_1\ \ x_2\ ...\ x_8$, l'effet du facteur ou de l'interaction qui correspond à cette colonne est donné par la formule :

$$E = c_1 x_1 + c_2 x_2 + ... + c_8 x_8$$

Par exemple, calculons l'effet principal du facteur A :

$$E = - 42 + 39 - 55 + 54 - 51 + 43 - 51 + 51 = - 12$$

Nous remarquons que toutes les valeurs affectées du même signe correspondent au même niveau de A. L'effet principal de A est donc la différence entre la somme du niveau (+) et celle du niveau (-). Pour cette raison, il est aussi nommé *contraste* de A. Nous pouvons faire la même remarque pour les facteurs B et C.

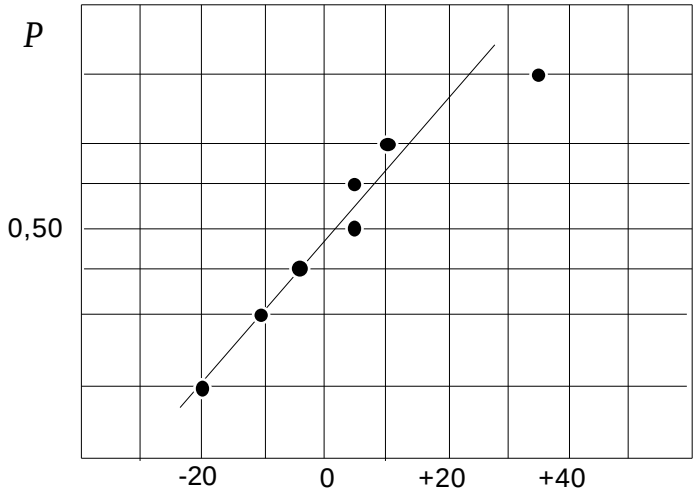
Interprétation des résultats

Le logiciel *Daisy* donne les valeurs suivantes :

Facteur	A	B	C	AB	AC	BC	( )
Effet	-12	36	6	10	-4	-20	6

Le facteur marqué ( ) est en réalité l'interaction ABC. Mais *Daisy* ne donne que les interactions qui seront utiles pour analyser les effets. En revanche, le nombre correspondant à ce facteur sera utilisé plus loin pour la construction du graphique de Henry.

Il ne reste plus qu'à trouver les facteurs et les interactions qui ont un effet significatif. Si nous faisons une analyse de la variance, comme au chapitre précédent, nous voyons que le nombre de *d.l.* d'un effet (effet principal ou interaction) est égal à 1, et qu'il n'y a pas de variations résiduelles. Le calcul de *F* est donc impossible, car nous n'avons pas de carré moyen à mettre au dénominateur. Pour contourner cette difficulté, on pourrait sélectionner des facteurs non prédominants et additionner leurs variances. Mais une méthode plus élégante consiste à utiliser un graphique à échelle de Henry.



Graphique de Henry de l'étude d'une réaction chimique

Rappelons d'abord que la méthode des plans factoriels implique obligatoirement une distribution normale des données. Une expérience factorielle n'est valable que si un test de normalité vient confirmer l'hypothèse que les données sont issues d'une distribution normale. Or, les effets étant des formes linéaires des données, si les données sont distribuées normalement, les effets sont aussi distribués normalement. Par conséquent, dans l'hypothèse où il n'existe aucun effet significatif (effet principal ou interaction), les points représentant les effets sur le graphique seront alignés. En outre, elle passe toujours à proximité du point (0 ; 0,50) car la moyenne des effets est nulle.

L'analyse des effets consiste à porter ces valeurs sur un graphique à échelle de Henry pour trouver les points qui s'écartent assez notablement de la droite normale. L'échelle horizontale représente les taux d'impuretés ; elle est graduée en g/Kg. L'échelle verticale, qui représente la probabilité, est graduée de 0 à 1. Les probabilités cumulées des sept valeurs des effets sont : 0,125 ; 0,25 ; 0,375 ; 0,50 ; 0,875 ; 0,75 ; 0,875. Remarquons au passage que sur cette échelle non linéaire, la probabilité 1 est rejetée à l'infini.

Les six premiers points en partant de la gauche semblent suffisamment alignés pour que l'hypothèse d'une distribution normale soit retenue. Nous traçons donc la

droite qui passe au plus près de ces points. Nous voyons alors que le dernier point (E = 36) s'en écarte assez notablement. La conclusion qui s'impose est que le facteur B (dosage du catalyseur) est probablement le seul facteur ayant un effet significatif. Une étude de régression permettra ensuite de déterminer le dosage optimum avec précision.

Un plan factoriel complet avec répétition

En étudiant l'exemple précédent, le lecteur a compris que si l'on veut utiliser le test F pour trouver les facteurs principaux et les interactions ayant un effet significatif, il faut faire un plan avec au moins une répétition, ce qui double le nombre d'essais nécessaire. Prévoyant cette éventualité, nous n'avons pas traité le cas en entier dans le paragraphe précédent, parce qu'une autre série d'essais identiques eut lieu dans cette entreprise quelques jours plus tard. Les résultats de mesure figurent sur le tableau 5.1.

Essai	1	2	3	4	5	6	7	8
Série I	42	39	55	54	51	43	51	51
Série II	43	46	56	53	51	48	52	45
Différence	-1	-7	-1	1	0	-5	-1	6
Moyenne	42,5	42,5	55,5	53,5	51	45,5	51,5	48

Résultats d'un plan factoriel avec répétition

La *SCE* intra-classes est $[(-1)^2 + (-7)^2 + ... + (6)^2] / 2 = 57$ avec 8 *d.l.*

La *SCE* d'un facteur est $E^2 / 4$, sachant que E est l'effet de celui-ci, avec 1 *d.l.*

Ces nombres sont portés sur un tableau d'analyse de la variance :

source de variation	***SCE***	***d.l.***	***C.M.***	***F***
A	30,25	1	30,25	4,25
B	182,25	1	182,25	25,58
C	1	1	1	0,14
AB	0	1	0	0
AC	12,25	1	12,25	1,72
BC	110,25	1	110,25	15,47
( )	4	1	4	0,56
Err. aléatoire	57	8	7,125	
Total	397	15	26,47	

En cherchant dans la table de *F* le nombre qui correspond à 1 et 8 *d.l.*, nous trouvons 5,32. Ce test conduit à une conclusion légèrement différente de la précédente, car il montre que le facteur B et l'interaction BC ont tous deux des effets significatifs. En conséquence, il faudra non seulement déterminer le dosage optimum du catalyseur, mais aussi changer le dosage du catalyseur quand on change de fournisseur.

Cet exemple montre que le test de *F* aboutit à une conclusion plus précise que le graphique de Henry, mais cet avantage n'est pas gratuit puisqu'il faut doubler le nombre d'essais.

Propriétés d'une matrice orthogonale

Rappel. L'ensemble de *n* variables (x_1 , x_2 , ... x_i , ... x_n) est un vecteur X à *n* dimensions. On peut le transformer en un vecteur Y en le multipliant par une matrice que nous nommerons A :

$$Y = A\,X$$

Par exemple, le vecteur (12 , 2) est le produit du vecteur (7 , 5) par la matrice

$$\begin{pmatrix} +1 & +1 \\ +1 & -1 \end{pmatrix}$$

Car 7 + 5 =12 et 7 - 5 =2

D'après cette définition, le vecteur des effets est le produit du vecteur des réponses par la matrice du plan factoriel. On remarquera que, puisque la matrice comporte *n* lignes et (*n* - 1) colonnes, le vecteur des effets n'a que (*n* - 1) dimensions.

Après ce bref rappel, nous allons étudier la catégorie des matrices constituées exclusivement des coefficients +1 et -1. Pour faciliter le raisonnement, reprenons les symboles du graphique 5.1. En plaçant dans un certain ordre les 3 colonnes principales et les 4 autres, nous obtenons une matrice à 8 colonnes et 7 lignes que les statisticiens désignent par $OA^8(2^7)$. Le graphique 5.4 indique sa composition à partir des 3 colonnes principales, x_1 , x_2 et x_3 . Chaque colonne est affectée d'un *générateur.*

Une propriété évidente de cette matrice est qu'en élevant au carré les éléments d'une colonne quelconque, autrement dit en faisant son produit par elle-même, on obtient la colonne identité, c'est-à-dire une colonne n'ayant que des 1. Nous en déduisons la règle de composition suivante :

> Le produit de deux colonnes s'obtient en juxtaposant leurs générateurs, puis en supprimant les lettres qui sont répétées deux fois.

Exemple le produit de x_1 et de $x_1.x_2$ est $x_1.x_1.x_2$, c'est-à-dire x_2
le produit de $x_1.x_2$ et de $x_1.x_3$ est $x_1.x_2$ $.x_1.x_3$, c'est-à-dire $x_2.x_3$
etc.

Il en résulte que chaque colonne peut être formée à partir d'autres, mais pas de n'importe quelle façon. Par exemple, sur la matrice $OA^8(2^7)$, parmi 35 combinaisons de 3 colonnes, on peut en trouver 28 où les colonnes sont indépendantes, comme par exemple x_1 , x_2 et $x_1.x_3$. Mais dans les 7 autres combinaisons, chaque colonne est le produit des deux autres. Une autre propriété intéressante, que chacun peut facilement vérifier, est que chaque paire de colonnes contient 2 fois les 4 combinaisons d'éléments 2 à 2. Par exemple la paire de colonnes (x_1 ; x_2) du graphique 5.1 contient 2 fois les paires d'éléments (1 ; 1), (1 ; 2), (2 ; 1) et (2 ; 2). Par définition, on dit qu'une matrice ayant cette propriété est une matrice *orthogonale*, par analogie avec un système de vecteurs orthogonaux dans un espace à *n* dimensions.

C'est toujours une matrice orthogonale, quel que soit le nombre de lignes, qui permet de calculer les effets avec un maximum de précision.

Colonne	1	2	3	4	5	6	7
Générateur	x_1	x_2	x_1 x_2	x_3	x_1 x_3	x_2 x_3	x_1 x_2 x_3

Composition de la matrice $OA_8(2^7)$

6
Plans factoriels fractionnaires

Le chapitre précédent montre comment on peut identifier des effets significatifs dans une expérience factorielle à deux niveaux. L'analyse de la variance est toujours à la base de cette méthode, mais la théorie des matrices orthogonales vient s'y ajouter. Comme on pouvait s'y attendre, l'application d'un plan factoriel complet permet d'identifier non seulement des facteurs, mais aussi des interactions. Nous allons voir maintenant qu'il est possible de réduire le nombre d'essais en adoptant l'hypothèse *a priori* que certaines interactions n'ont pas d'effet significatif. La méthode est celle des graphes d'interaction.

Construction d'un plan factoriel fractionnaire

On est souvent contraint de réduire le nombre d'essais pour des raisons économiques. Dans ce cas, il est possible de construire un plan ne comportant pas toutes les combinaisons de niveaux, mais avec lequel on peut calculer un certain nombre d'effets. C'est un plan factoriel fractionnaire.

Pour commencer, nous allons donner une méthode de construction d'un plan fractionnaire en fonction du nombre de facteurs à vérifier et du nombre d'essais que l'on peut accepter d'un point de vue économique. Il nous sera plus commode de reprendre les symboles 1 et 2 utilisés au chapitre précédent. La règle de composition des colonnes est alors :

$$(1) \times (1) = (1) \qquad (1) \times (2) = (2) \qquad (2) \times (2) = (1)$$

Dans une matrice orthogonale, chaque colonne est le produit de deux autres colonnes, ou plus. Par exemple, dans la matrice $OA_8(2^7)$, la colonne 3 est le produit des colonnes 1 et 2, et la colonne 7 est le produit des colonnes 1, 2 et 4. On dit que la colonne 3 contient l'interaction de 1 et 2, et que la colonne 7 contient l'interaction de 1, 2 et 4. De façon générale, on dit que la colonne *i* contient l'interaction d*e j, k, l, ...* quand elle est le produit des colonnes *j, k, l, ...* suivant la règle de composition donnée plus haut. Nous verrons plus loin que cette propriété joue un rôle important dans la construction du plan fractionnaire.

Un plan fractionnaire de 2^n essais se déduit assez facilement de la matrice orthogonale $OA_n 2^{n+1}$. Les graphiques 6.1, 6.2 et 6.3 représentent les matrices qui sont à l'origine des plans de 4, 8 et 16 essais. Le lecteur peut trouver dans un

ouvrage spécialisé les matrices utilisées pour des plans de 32, 64 et même 132 essais, mais en pratique il est assez rare que l'on fasse plus de 16 essais.

	Colonne		
Ligne	1	2	3
1	1	1	1
2	1	2	2
3	2	1	2
4	2	2	1
Générateur	x_1	x_2	x_1 x_2

Matrice $OA_4(2^3)$

	Colonne						
Ligne	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2	1	1	1	2	2	2	2
3	1	2	2	1	1	2	2
4	1	2	2	2	2	1	1
5	2	1	2	1	2	1	2
6	2	1	2	2	1	2	1
7	2	2	1	1	2	2	1
8	2	2	1	2	1	1	2
Générateur	x_1	x_2	x_1 x_2	x_3	x_1 x_3	x_2 x_3	x_1 x_2 x_3

Matrice $OA_8(2^7)$

Prenons un exemple. Un plan complet de 16 essais comporte 4 facteurs principaux A, B, C et D. Supposons que je veuille construire un plan fractionnaire avec 4 facteurs principaux, mais avec 8 essais seulement, comme le montre le graphique 6.4. Dans la matrice $OA_8(2^7)$, je peux choisir les colonnes 1, 2, 4 et 7, dont les générateurs sont x_1, x_2, x_3, et $x_1.x_2.x_3$, mais il est clair que la colonne 7 représentera à la fois le facteur D et l'interaction ABC. On dit que D et ABC sont des *alias.* Au moment d'interpréter les résultats de mesure, je ne saurai pas si l'effet calculé est celui de D ou de son alias. La construction du plan introduit donc un risque de confusion. Si l'effet est significatif, j'opterai pour D, car en pratique il est très rare qu'une interaction d'ordre 3 se produise. En revanche, si j'avais choisi la colonne 3 qui représente à la fois le facteur D et l'interaction AB, la confusion serait plus embarrassante, car il n'est pas rare qu'une interaction d'ordre 2 se produise.

	Colonne														
Ligne	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2
3	1	1	1	2	2	2	2	1	1	1	1	2	2	2	2
4	1	1	1	2	2	2	2	2	2	2	2	1	1	1	1
5	1	2	2	1	1	2	2	1	1	2	2	1	1	2	2
6	1	2	2	1	1	2	2	2	2	1	1	2	2	1	1
7	1	2	2	2	2	1	1	1	1	2	2	2	2	1	1
8	1	2	2	2	2	1	1	2	2	1	1	1	1	2	2
9	2	1	2	1	2	1	2	1	2	2	2	1	2	1	2
10	2	1	2	1	2	1	2	2	1	1	1	2	1	2	1
11	2	1	2	2	1	2	1	1	2	2	2	2	1	2	1
12	2	1	2	2	1	2	1	2	1	1	1	1	2	1	2
13	2	2	1	1	2	2	1	1	2	1	1	1	2	2	1
14	2	2	1	1	2	2	1	2	1	2	2	2	1	1	2
15	2	2	1	2	1	1	2	1	2	1	1	2	1	1	2
16	2	2	1	2	1	1	2	2	1	2	2	1	2	2	1
Générateur	x_1	x_2	x_1 x_2	x_3	x_1 x_3	x_2 x_3	x_1 x_2 x_3	x_4	x_1 x_4	x_2 x_4	x_1 x_2 x_4	x_3 x_4	x_1 x_3 x_4	x_2 x_3 x_4	x_1 x_2 x_3 x_4

Matrice OA $_{16}$ (2 ^{15})

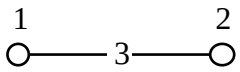
Un plan fractionnaire est moins coûteux que le plan complet ayant le même nombre de facteurs mais il comporte nécessairement des alias. C'est là tout le problème. Quand on prépare un plan fractionnaire, il faut éviter de créer des alias entre certains facteurs principaux et certaines interactions lorsque tous deux risquent d'avoir un effet significatif. On peut utiliser pour cela trois types d'outils : le groupe générateur, le graphe linéaire et le graphe d'interaction. C'est ce dernier que nous allons présenter.

Essai	Facteur			
	A	*B*	*C*	*D*
1	-	-	-	-
2	+	-	-	+
3	-	+	-	+
4	+	+	-	-
5	-	-	+	+
6	+	-	+	-
7	-	+	+	-
8	+	+	+	+
N° de colonne de OA $_8$ (2 ^{7})	4	2	1	7

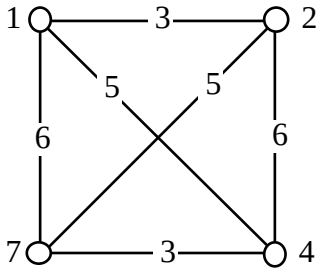
Plan factoriel fractionnaire à 4 facteurs

Graphes d'interaction

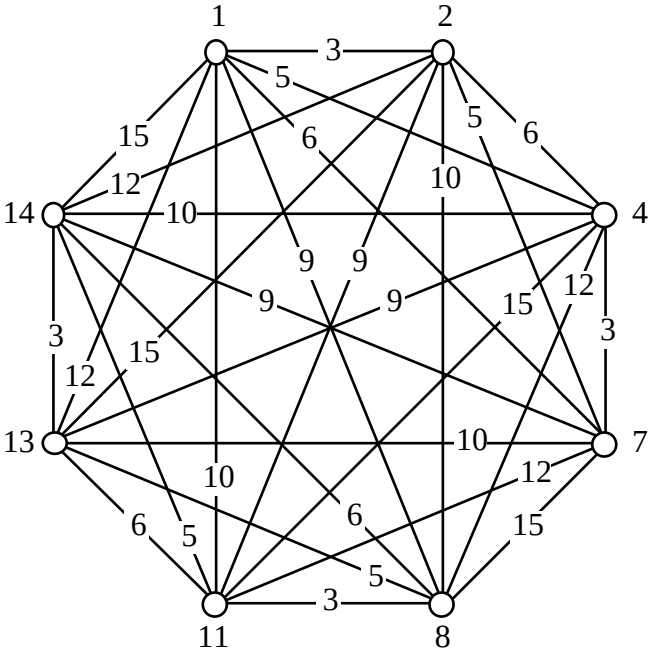
Les 3 colonnes de la matrice $OA_4(2^3)$ forment un groupe d'interaction. Cette relation est représentée ci-dessous. En observant la matrice $OA_8(2^7)$, on distingue 7 groupes d'interaction d'ordre 3. On pourrait les représenter au moyen de graphes linéaires, comme le fait Taguchi, mais il est plus commode de tracer un seul graphe à deux dimensions. Cette figure est le graphe d'interaction $IG_8(2^7)$. Le problème se complique avec la matrice $OA_{16}(2^{15})$ car on ne compte pas moins de 31 interactions d'ordre 3 auxquelles s'ajoutent 6 interactions d'ordre 4. Elles sont représentées sur le graphe d'interaction $IG_{16}(2^{15})$.



Graphe d'interaction IG $_4$ (2 ^{3})



Graphe d'interaction IG $_8$ (2 ^{7})



Graphe d'interaction IG $_{16}$ (2 ^{15})

Pour lire un graphe d'interaction, il faut connaître les trois règles suivantes :

La règle d'identité. Nous avons déjà vu cette règle au chapitre précédent. En faisant le produit d'une colonne par elle-même, on obtient la colonne identité, c'est-à-dire une colonne n'ayant que des 1. D'autre part en faisant le produit d'une colonne quelconque par la colonne identité, on obtient cette colonne.

La règle des lignes. Les trois colonnes représentées par deux nœuds et la ligne qui les joint forment un groupe d'interaction. Le produit de ces 3 colonnes est la colonne identité. Chaque colonne est le produit des 2 autres. Par exemple, suivant cette règle, le graphe $OA_8(2^7)$ permet de voir que les colonnes 3, 4 et 7 forment un groupe d'interaction.

La règle des triangles. Les trois colonnes représentées par les trois côtés d'un triangle forment un groupe d'interaction. Le produit de ces 3 colonnes est la colonne identité. Chaque colonne est le produit des 2 autres. Par exemple, suivant cette règle, le graphe $OA_8(2^7)$ permet de voir que les colonnes 3, 5 et 6 forment un groupe d'interaction.

Les graphes d'interaction sont particulièrement utiles pour préparer un plan factoriel fractionnaire permettant d'estimer les effets de tous les facteurs principaux et de quelques interactions spécifiques, en supposant que les effets des autres interactions sont négligeables.

Résolution d'un plan factoriel fractionnaire

Le plus grand plan factoriel fractionnaire qui peut se déduire d'un plan factoriel complet comportant 2^n essais est un demi-plan comportant $2^{(n-1)}$ essais. Considérons à nouveau le plan fractionnaire comportant 8 essais qui est constitué par les colonnes 1, 2, 4 et 7 de la matrice $OA_8(2^7)$. On voit que c'est la moitié du plan factoriel complet comportant 16 essais qui est constitué par les colonnes 2, 4, 8 et 14 de la matrice $OA_{16}(2^{15})$. De même peut-on voir qu'un plan fractionnaire comportant 16 essais et 5 facteurs est la moitié d'un plan factoriel complet comportant 32 essais, etc.

Nous avons vu qu'un plan $2^{(3-1)}$ présente le sérieux inconvénient de confondre un facteur principal avec une interaction d'ordre 2. Un plan $2^{(4-1)}$ présente seulement l'inconvénient de confondre un facteur principal avec une interaction d'ordre 3 et deux interactions d'ordre 2 entre elles. Un plan $2^{(5-1)}$ présente des inconvénients encore moindres. C'est pour cette raison que la théorie des plans factoriels utilise la notion de *résolution.* Plus la résolution est grande, plus le plan est précis.

Par définition, la résolution d'un plan factoriel à 2 niveaux est le nombre de colonnes contenues dans le plus petit groupe générateur. En pratique, il suffit de

savoir qu'un plan $2^{(3-1)}$ est de résolution III, un plan $2^{(4-1)}$ de résolution IV et un plan $2^{(5-1)}$ de résolution V.

Un plan comportant 4 essais et 3 facteurs

L'installation de tentes en plein air fait partie des prestations demandées pour certaines cérémonies et fêtes, notamment pour des mariages. Un traiteur, après avoir acheté le matériel nécessaire, voudrait déterminer le processus de montage qui prend le moins de temps. Il dispose de 2 types de charpentes, 2 types de bâches et 2 types d'outils. Pour tester ces 3 facteurs, il décide de faire des essais avec un plan factoriel comportant 4 essais seulement, car il voudrait avoir une première estimation le plus tôt possible. Le plan est de résolution III, en sorte que chaque facteur risque d'être confondu avec les interactions des deux autres. Le facteur charpente est A, le facteur bâche est B, et le facteur outils est C. Les résultats bruts, exprimés en minutes, sont :

Essai	1	2	3	4
Résultat	135	150	175	165

Les effets sont :

A 165 - 175 + 150 - 135 = 5
B 165 + 175 - 150 - 135 = 55
C 165 - 175 - 150 + 135 = - 25

L'effet B (bâche) est prédominant. Mais le nombre de données est trop faible pour pouvoir affirmer que ce facteur a une influence significative, et d'autre part il n'est pas exclu que ce soit l'interaction AC, interaction entre l'outil et la charpente, qui se soit manifestée au lieu de B. Le traiteur décide donc de faire 4 nouveaux essais en appliquant un plan factoriel $2^{(4-1)}$. La météo lui offre justement l'occasion de tester un nouveau facteur. Les 4 premiers essais ont eu lieu par un beau temps ensoleillé ; les 4 suivants auront lieu sous une pluie battante. Le facteur météo est D. Les résultats bruts, exprimés en minutes, sont :

Essai	1	2	3	4	5	6	7	8
Résultat	136	152	178	172	145	165	183	172

Les effets sont :

A 172 - 183 + 165 - 145 + 172 - 178 + 152 - 136 = 19
B 172 + 183 - 165 - 145 + 172 + 178 - 152 - 136 = 107
C 172 - 183 - 165 + 145 + 172 - 178 - 152 + 136 = -53
D 172 + 183 + 165 + 145 - 172 - 178 - 152 - 136 = 27
AD 172 - 183 + 165 - 145 - 172 + 178 - 152 + 136 = -1
BD 172 + 183 - 165 - 145 - 172 - 178 + 152 + 136 = 17
CD 172 - 183 - 165 + 145 - 172 + 178 + 152 - 136 = -9

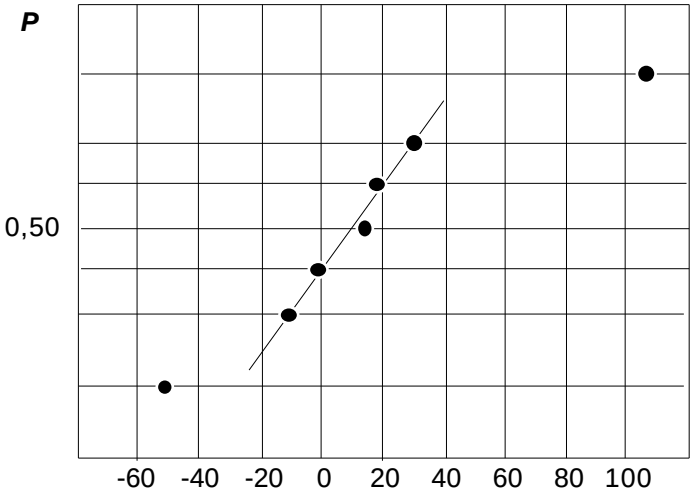
Cette fois, le traiteur possède des données en nombre suffisant pour faire le test du graphique de Henry. Il semble que les facteurs B et C ont un effet significatif. Pour confirmer cette hypothèse, il va faire un test de *F* en calculant une estimation de la variance résiduelle au moyen des facteurs non significatifs.

	A	B	C
1	-	-	+
2	+	-	-
3	-	+	-
4	+	+	+

Premier plan factoriel du traiteur

	A	B	C	D	AD	BD	CD
1	-	-	+	-	+	+	-
2	+	-	-	-	-	+	+
3	-	+	-	-	+	-	+
4	+	+	+	-	-	-	-
5	-	-	+	+	-	-	+
6	+	-	-	+	+	-	-
7	-	+	-	+	-	+	-
8	+	+	+	+	+	+	+

Deuxième plan factoriel du traiteur



Graphique de Henry du traiteur

La méthode du pooling Nous avons vu au chapitre précédent qu'on peut estimer la variance résiduelle en introduisant une répétition dans un plan factoriel complet. Cette méthode conduit à doubler le nombre des essais, alors qu'un plan factoriel fractionnaire tend au contraire à les réduire. Il est possible néanmoins de

faire une estimation de la variance résiduelle sans modifier le plan, mais avec une précision bien moindre, en utilisant la méthode du *pooling.*

Sur le graphique de Henry, on voit que les effets A, D, AD, BD, CD sont pratiquement négligeables. Ils participent au "bruit de fond" comme disent les ingénieurs. Le pooling consiste simplement à calculer le carré moyen des effets de ces facteurs. C'est une estimation de la variance résiduelle affectée de 5 *d.l.*, car l'effectif du pool est 5. On trouve :

$$C.M. = (361 + 729 + 1 + 289 + 81) /5 = 1461$$

Pour 1 et 5 *d.l.*, la table de *F* nous donne une valeur de 6,61. Les valeurs trouvées pour B et C étant respectivement 7,84 et 1,92, il faut retenir l'hypothèse que seule la différence entre les deux types de bâches est significative. En revanche, il est intéressant de constater que le temps de montage est indépendant des conditions climatiques.

Un plan comportant 16 essais et 5 facteurs

Pour la journée nationale de la Croix Rouge, le comité local de Bastillane organise comme d'habitude une quête en ville. Le président du comité voudrait profiter de l'occasion pour vérifier le bien-fondé de quelques innovations. Par exemple, l'année précédente, on avait vu de jeunes quêteurs avec des rollers et des casquettes. Le comité décide de faire un test sur 5 facteurs à la fois :

A : avec ou sans rollers
B : avec ou sans casquette
C : fille ou garçon
D : deux discours différents
E : seul(e) ou accompagné(e) d'un(e) camarade

Le nombre d'essais est fixé à 16. Mais puisque le facteur E consiste à opposer 8 personnes seules à 8 groupes de deux, ce sont 24 jeunes au total qui sont sélectionnés : 12 filles et 12 garçons. La veille de la journée nationale, le comité se réunit pour parler des détails de l'opération. C'est un plan factoriel de *Daisy* qui a été adopté.

Sur ce plan (page suivante), vous reconnaissez les colonnes n° 8, 4, 2 , 14 et 13 du graphique 6.3. Le graphe d'interaction $OA_{16}(2^{15})$ permet de vérifier qu'elles sont indépendantes.

	A	B	C	D	E
1	-	-	-	-	-
2	+	-	-	+	+
3	-	+	-	+	+
4	+	+	-	-	-
5	-	-	+	+	-
6	+	-	+	-	+
7	-	+	+	-	+
8	+	+	+	+	-
9	-	-	-	-	+
10	+	-	-	+	-
11	-	+	-	+	-
12	+	+	-	-	+
13	-	-	+	+	+
14	+	-	+	-	-
15	-	+	+	-	-
16	+	+	+	+	+

Organisation des essais du comité de la Croix Rouge

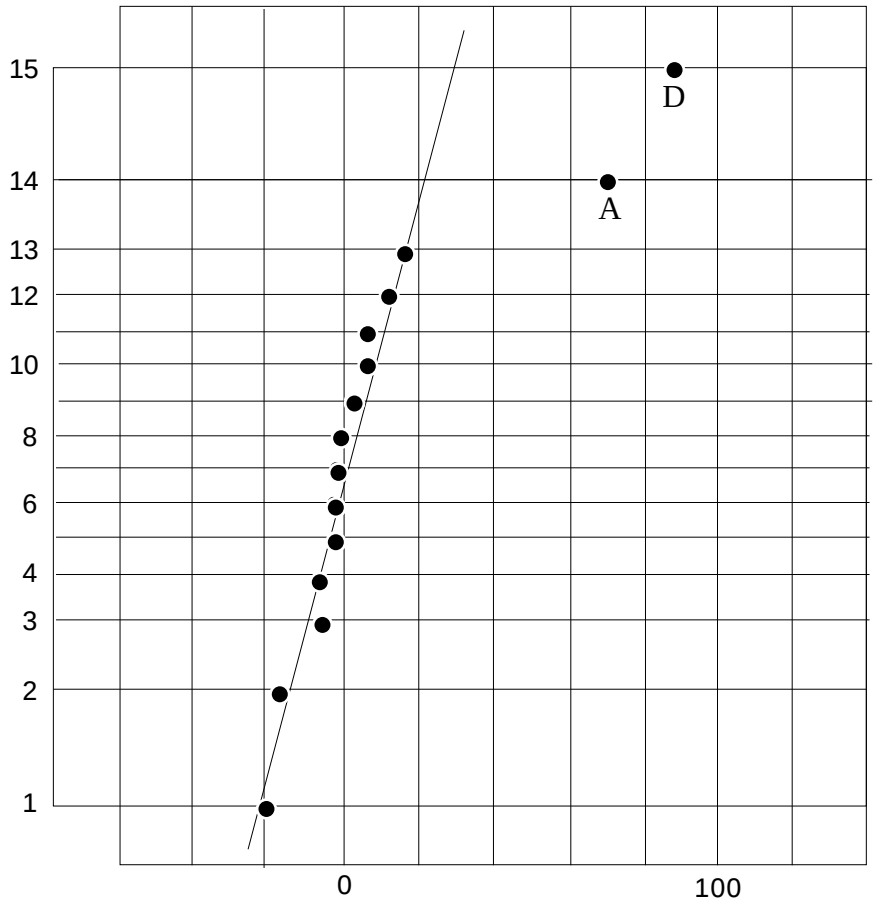
Le soir de la journée nationale, les résultats du test de Bastillane sont les suivants (en Euros) :

Tronc	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Montant	42	62	57	54	55	52	41	60	44	63	58	51	50	53	38	63

Ces nombres sont inscrits sur la page "données" de *Daisy* et les effets apparaissent sur la page "résultats". Les effets principaux sont :

	A	B	C	D	E
Effet	73	1	-19	93	-3

Il n'est pas besoin de faire un test de *F* pour affirmer que les facteurs A et D ont un effet significatif. Chacun peut s'en convaincre en portant sur un graphique de Henry les valeurs données par *Daisy* concernant les 5 facteurs et les 10 interactions. On voit que les points correspondant aux 13 valeurs à partir de la gauche sont alignés, alors que les deux points du haut s'écartent très nettement de la droite. Par conséquent ces 2 facteurs ont une véritable influence, alors que les 3 autres facteurs et les interactions sont négligeables.



Graphique de Henry du comité de la Croix Rouge

Voyant ce résultat, le comité a porté son attention sur les rollers et la présentation orale. La conclusion est double : d'une part les jeunes gens chaussés de rollers ont plus de chance auprès du public, et d'autre part le nouveau discours recommandé par la direction régionale est mieux reçu que le discours traditionnel. Il faut noter, évidemment, que cette conclusion n'est valable que pour la ville de Bastillane.

Arbitrage entre les objectifs et les moyens

Avant de s'engager dans une expérience factorielle, il faut bien réfléchir aux chances de succès. Quand on étudie un seul facteur composé de plusieurs traitements, il s'agit de savoir si certains traitements se distinguent des autres. On n'hésitera pas à répéter les essais pour augmenter le nombre de données, car on sait que l'expérience risque d'être perturbée par des variations n'ayant aucun rapport avec ce facteur. Celui qui a coutume d'observer des variations sur des graphiques de contrôle sait combien l'estimation d'une moyenne est fragile quand elle repose sur un petit nombre de données. Au contraire, quand on prépare un plan d'expérience, sachant que l'expérience factorielle a pour principe d'économiser

les essais, on hésite à les répéter. Or des variations indépendantes des facteurs risquent de se produire, même quand il s'agit d'un processus bien maîtrisé. Par conséquent, plus le nombre d'essais est réduit, plus le risque est grand que des variations intempestives conduisent à des conclusions erronées. La fascination exercée sur les ingénieurs par la théorie des plans factoriels fractionnaires les conduit souvent à négliger ce risque.

Enfin, il faut bien comprendre qu'une expérience factorielle n'a pas pour but de montrer que certains facteurs ont une influence plus grande que d'autres, car on trouvera toujours des différences entre les effets. Le but est d'identifier des facteurs qui permettent d'améliorer un processus.

Conclusion

Notre vie quotidienne est habitée par une question dont la réponse appelle toujours la satisfaction d'un désir ou d'un besoin : « que vais-je faire maintenant ? » Quand deux personnes ne se connaissant pas se trouvent dans la même situation, il peut se faire que leurs décisions soient différentes, d'abord parce qu'elles ont des projets différents, mais aussi en raison de leurs expériences personnelles. Ainsi, à la veille d'un long week-end, l'une peut décider de partir en voyage, l'autre de rester chez elle pour faire des achats en ville et entretenir son jardin. C'est peut-être la crainte des embouteillages qui a inspiré la seconde. L'expérience est à la base de la décision, car c'est le meilleur moyen d'évaluer les chances de réussite d'un projet. Or ce que nous nommons ainsi n'est pas une simple accumulation de faits, mais une synthèse. En effet, nous avons tous l'habitude de réfléchir aux événements passés, tantôt pour les comparer, tantôt pour leur trouver des causes possibles. La plus grande partie des idées qui viennent alors à l'esprit sont des comparaisons et des présomptions de causes. Ceci apparaît de manière flagrante quand les parents discutent avec les enfants au sujet des notes obtenues à l'école, ou bien quand un chef de service réunit ses collaborateurs pour leur parler d'objectifs et de résultats.

Les croyances à des relations de cause à effet sont innombrables. Elles se transmettent souvent de génération en génération, comme beaucoup de dictons populaires. « Quand le chat fait sa toilette, s'il passe une patte derrière l'oreille, c'est qu'il va pleuvoir demain. » Il existe aussi de petites théories, justifiées ou non, que chacun élabore à partir d'observations personnelles. « Un enfant travaille mieux à l'école quand il regarde moins la télévision. » Nous pourrions citer également, parmi ces croyances, des théories économiques qui alimentent régulièrement les débats politiques, et qui mériteraient plutôt le nom de dogmes. Le public est porté à croire que les théories qui lui sont présentées, en économie, en médecine, en biologie, etc. ont toujours été vérifiées par des gens compétents. Quant aux petites théories personnelles, chacun pense évidemment qu'elles sont justifiées par le bon sens. Or le monde est plein d'idées fausses, et le pire est qu'une erreur mille fois répétée par les médias passe pour une vérité. C'est pourquoi il est important d'apprendre à chaque citoyen, en commençant par les élèves des collèges, comment une hypothèse peut être vérifiée expérimentalement au moyen de certains outils mathématiques.

Les idées qui se trouvent dans ce livre datent du début du vingtième siècle, mais elles ne sont pas encore parfaitement entrées dans les mœurs. Pour commencer, il faut admettre que des variations existent entre des choses qu'on voudrait voir identiques, et comprendre que les variations sont nécessaires à la vie. Le premier chapitre montre que toute chose est le résultat d’un processus, et que

cette production est accompagnée de variations, le plus souvent aléatoires. Donc la différence entre plusieurs valeurs observées ne signifie pas forcément que les choses ont changé. Le problème est d'abord de savoir si les variations sont aléatoires ou non. Par exemple, quand on nous annonce qu'un indicateur économique varie dans le bon sens, ce n'est pas forcément une bonne nouvelle ; parfois cela ne veut rien dire du tout. Quelques auteurs américains ont comparé le fonctionnement d'un processus à celui d'un vieux poste de radio. Quand celui-ci ne capte aucune station, le haut parleur fait un bruit caractéristique. Quand il capte une station lointaine, le bruit est toujours présent et couvre presque entièrement le signal. Les ingénieurs disent que le *rapport signal sur bruit* est faible. Avec un processus, la succession des variations aléatoires est l'équivalent du bruit du poste. Une variation ne peut être interprétée comme un signal que lorsqu'elle dépasse le niveau de bruit. C'est ce qui apparaît sur un graphique de contrôle.

La régression linéaire est certainement l'outil statistique le plus utilisé en économie, ce qui explique sa bonne connaissance par le grand public. En présentant au chapitre II la méthode qui permet de tracer une droite de régression, nous avons mis le lecteur en garde contre deux erreurs, malheureusement trop fréquentes, qui peuvent le conduire à des conclusions erronées. En effet, il faut savoir que cette méthode n'est valable que si la variable expliquée a une distribution normale, d'où l'importance du test de normalité. D'autre part, la droite de régression n'a de sens que si le coefficient de corrélation est assez grand, d'où la nécessité de se reporter à la table de Fisher (table B de ce livre) qui indique les limites de ce nombre en fonction de la taille de l'échantillon. C'est une introduction à la notion de risque d'erreur dans un jugement statistique.

Au chapitre 3, le problème de la comparaison de deux échantillons d'objets ayant subi des traitements différents nous fait entrer dans le domaine de la distribution de Student, qui s'étend à tous les chapitres suivants. Le principe est de comparer les moyennes des deux échantillons pour estimer, avec une certaine probabilité, s'ils pourraient être issus d'une même population. L'analyse de la variance, présentée au chapitre 4, n'est autre que l'extension de la méthode au cas de plusieurs échantillons.

Les chapitres 5 et 6 traitent des méthodes qui consistent à combiner dans une même expérience l'étude de plusieurs facteurs, ce qui s'appelle une expérience factorielle. Ces méthodes sont très efficaces, car chaque donnée apporte une information sur tous les facteurs qui interviennent dans l'expérience. Elles permettent de trouver non seulement des facteurs, mais aussi des interactions ayant un effet significatif. On s'aperçoit cependant que leurs performances ont des limites dès que l'on cherche à calculer le rapport des variances qui sert à distinguer les effets significatifs. D'autre part, une variation intempestive portant sur une seule donnée d'entrée peut avoir de graves conséquences sur la conclusion de l'étude. C'est pourquoi il est absolument nécessaire d'assurer au préalable une parfaite stabilité du processus dont proviennent les données, avec une variance aussi réduite que possible. Cette condition n'avait pas été soulignée par Fisher, car

ses méthodes étaient appliquées en agronomie, domaine où les processus sont naturellement stables. Mais les statisticiens lui portent une grande attention chaque fois que des plans factoriels sont appliqués dans l'industrie.

On peut trouver dans certains ouvrages spécialisés beaucoup d'autres méthodes statistiques destinées aux études analytiques. Mais celles qui sont présentées dans ce livre permettent de résoudre la grande majorité des problèmes pratiques.

TABLE A

Distribution normale réduite cumulée

Aire comprise entre la courbe, les axes *ox*, *oy*, et la verticale d'abscisse *x*

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.000	0.004	0.008	0.012	0.016	0.020	0.024	0.028	0.032	0.036
0.1	.040	.044	.048	.052	.056	.060	.064	.067	.071	.075
0.2	.079	.083	.087	.091	.095	.099	.103	.106	.110	.114
0.3	.118	.122	.125	.129	.133	.137	.141	.144	.148	.152
0.4	.155	.159	.163	.167	.170	.174	.177	.181	.184	.188
0.5	.191	.195	.198	.202	.205	.209	.212	.216	.219	.222
0.6	.226	.229	.232	.236	.239	.242	.245	.249	252	255
0.7	.258	.261	.264	.267	.270	.273	.276	.279	.282	.285
0.8	.288	.291	.294	.297	.299	.302	.305	.308	.311	.313
0.9	.316	.319	.321	.324	.326	.329	.331	.334	.336	.339
1.0	.341	.344	.346	.346	.348	.351	.355	.358	.360	.362
1.1	.364	.366	.369	.371	.373	.375	.377	.379	.381	.383
1.2	.385	.387	.389	.391	.392	.394	.396	.398	.400	.401
1.3	.403	.405	.407	.408	.410	.411	.413	.415	.416	.418
1.4	.419	.421	.422	.424	.425	.426	.428	.429	.431	.432
1.5	.433	.434	.436	.437	.438	.439	.441	.442	.443	.444
1.6	.445	.446	.447	.448	.449	.450	.451	.452	.453	.454
1.7	.455	.456	.457	.458	.459	.460	.461	.462	.462	.463
1.8	.464	.465	.466	.466	.467	.468	.469	.469	.470	.471
1.9	.471	.472	.473	.473	.474	.474	.475	.476	.476	.477
2.0	.477	.478	.478	.479	.479	.480	.480	.481	.481	.482
2.1	.482	.483	.483	.483	.484	.484	.485	.485	.485	.486
2.2	.486	.486	.487	.487	.487	.488	.488	.488	.489	.489
2.3	.489	.490	.490	.490	.490	.491	.491	.491	.491	.492
2.4	.492	.492	.492	.492	.493	. 493	.493	.493	.493	.494
2.5	.494	.494	.494	.494	.494	.495	.495	.495	.495	.495
2.6	.495	.495	.496	.496	.496	.496	.496	.496	.496	.496
2.7	.496	.497	.497	.497	.497	.497	.497	.497	.497	.497
2.8	.497	.497	.498	.498	.498	.498	.498	.498	.498	.498
2.9	.498	.498	.498	.498	.498	.498	.498	.498	.499	.499
3.0	.499	.499	.499	.499	.499	.499	.499	.499	.499	.499

Table B

Seuils du coefficient de corrélation

P est la probabilité de dépassement dans l'hypothèse nulle

d.l.	*P* = 0.05	*P* = 0.01	*d.l.*	*P* = 0.05	*P* = 0.01
1	.997	1.000	21	.413	.526
2	.950	.990	22	.404	.515
3	.878	.959	23	.396	.505
4	.811	.917	24	.388	.496
5	.754	.874	25	.381	.487
6	.707	.834	26	.374	.478
7	.666	.798	27	.367	.470
8	.632	.765	28	.361	.463
9	.602	.735	29	.355	.456
10	.576	.708	30	.349	.449
11	.553	.684	35	.325	.418
12	.532	.661	40	.304	.393
13	.514	.641	45	.288	.372
14	.497	.623	50	.273	.354
15	.482	.606	60	.250	.325
16	.468	.590	70	.232	.302
17	.456	.575	80	.217	.283
18	.444	.561	90	.205	.267
19	.433	.549	100	.195	.254
20	.423	.537	150	.159	.208

Table C

Seuils du test des signes

N est le nombre total de points.

S est le seuil de signification avec une probabilité de 0.05

N	*S*	*N*	*S*	*N*	*S*	*N*	*S*
9	1	27	7	45	15	63	23
10	1	28	8	46	15	64	23
11	1	29	8	47	16	65	24
12	2	30	9	48	16	66	24
13	2	31	9	49	17	67	25
14	2	32	9	50	17	68	25
15	3	33	10	51	18	69	25
16	3	34	10	52	18	70	26
17	4	35	11	53	18	71	26
18	4	36	11	54	19	72	27
19	4	37	12	55	19	73	27
20	5	38	12	56	20	74	28
21	5	39	12	57	20	75	28
22	5	40	13	58	21	76	28
23	6	41	13	59	21	77	29
24	6	42	14	60	21	78	29
25	7	43	14	61	22	79	30
26	7	44	15	62	22	80	30

Table D

Distribution de *t*

P est la probabilité de dépassement en valeur absolue

d.l.	*P* = 0.20	*P* = 0.10	*P* = 0.05	*P* = 0.025	*P* = 0.01
1	3.078	6.314	12.706	25.452	63.657
2	1.886	2.920	4.303	6.205	9.925
3	1.638	2.353	3.182	4.176	5.841
4	1.533	2.132	2.776	3.495	4.604
5	1.476	2.015	2.571	3.163	4.032
6	1.440	1.943	2.447	2.969	3.707
7	1.415	1.895	2.365	2.841	3.499
8	1.397	1.860	2.306	2.752	3.355
9	1.383	1.833	2.262	2.685	3.250
10	1.372	1.812	2.228	2.634	3.169
11	1.363	1.796	2.201	2.593	3.106
12	1.356	1.782	2.179	2.560	3.055
13	1.350	1.771	2.160	2.533	3.012
14	1.345	1.761	2.145	2.510	2.977
15	1.341	1.753	2.131	2.490	2.947
16	1.337	1.746	2.120	2.473	2.921
17	1.333	1.740	2.110	2.458	2.898
18	1.330	1.734	2.101	2.445	2.878
19	1.328	1.729	2.093	2.433	2.861
20	1.325	1.725	2.086	2.423	2.845
21	1.323	1.721	2.080	2.414	2.831
22	1.321	1.717	2.074	2.406	2.819
23	1.319	1.714	2.069	2.398	2.807
24	1.318	1.711	2.064	2.391	2.797
25	1.316	1.708	2.060	2.385	2.787
26	1.315	1.706	2.056	2.379	2.779
27	1.314	1.703	2.052	2.373	2.771
28	1.313	1.701	2.048	2.368	2.763
29	1.311	1.699	2.045	2.364	2.756
30	1.310	1.697	2.042	2.360	2.750

Table E

Seuils du test de comparaison des rangs

La probabilité de dépassement bilatéral est de 0.05

	10	11	12	13	14	15
10	78 132					
11	81 139	96 157				
12	85 145	99 165	115 185			
13	88 152	103 172	119 193	137 214		
14	91 159	106 180	123 201	141 223	160 246	
15	94 166	110 187	127 209	145 232	164 256	185 280
16	96 174	114 194	131 217	150 240	169 265	190 290
17	100 180	117 202	135 225	154 249	175 273	195 300
18	103 187	121 209	139 233	159 257	179 283	201 309
19	107 193	124 217	144 240	163 266	184 292	205 320
20	110 200	128 224	148 248	168 274	189 301	211 329

Table F

Distribution de *F*

La probabilité de dépassement est 0.05

Les *d.l.* du numérateur vont de 1 à 9

Les *d.l.* du dénominateur vont de 1 à 30

d.l.	1	2	3	4	5	6	7	8	9
1	161	200	216	225	230	234	237	239	241
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80
13	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.55	2.48	2.43
20	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2,45	2.40
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.47	2.40	2.35
23	4.28	3.42	3.03	2.80	2.64	2.53	2.45	2.38	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.43	2.36	2.30
25	4.24	3.38	2.99	2.76	2.60	2.49	2.41	2.34	2.28
26	4.22	3.37	2.98	2.74	2.39	2.32	2.27	2.22	2.27
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.30	2.25
28	4.20	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.24
29	4.18	3.33	2.93	2.70	2.54	2.43	2.35	2.28	2.22
30	4.17	3.32	2.92	2.69	2.53	2.42	2.34	2.27	2.21