Statistique

Dictionnaire encyclopédique

Springer

Paris
Berlin
Heidelberg
New York
Hong Kong
Londres
Milan
Tokyo

Yadolah Dodge

Statistique

Dictionnaire encyclopédique

Springer

Yadolah Dodge

Professeur honoraire
Université de Neuchâtel
Suisse
yadolah.dodge@unine.ch

ISBN : 978-2-287-72093-2 Springer Paris Berlin Heidelberg New York

SPIN : 12053858

Maquette de couverture : Jean-François Montmarché

Garde-toi bien de la fumée des cœurs blessés
La blessure du cœur se rouvrira sans cesse.
Tant que tu le pourras, n'oppresse pas un cœur
Le soupir d'un seul cœur renversera le monde.

Saadi Shirazi (1184-1271), poète persan

Cet ouvrage est dédié à tous les enfants de par le monde, qui sont les victimes innocentes des atrocités humaines.

Du même auteur

Aux Éditions Springer-Verlag France

Dodge Y (1999) Premiers pas en statistique. Springer-Verlag, Paris.

Dodge Y (2002) Mathématiques de base pour économistes. Springer-Verlag, Paris.

Dodge Y (2005) Optimisation appliquée. Springer-Verlag, Paris.

Autres éditeurs

Arthanari TS and Dodge Y (1981) Mathematical Programming in Statistics. John Wiley and Sons, New York.

Arthanari TS and Dodge Y (1993) Mathematical Programming in Statistics. Classic edition, John Wiley and Sons, New York.

Dodge Y (1985) Analysis of Experiments with Missing Data. John Wiley and Sons, New York.

Dodge Y (1987, 1989, 1996) Mathématiques de base pour économistes. Éditions EDES puis Presses Académiques, Neuchâtel.

Dodge Y (1987) Introduction à la programmation linéaire. Editions EDES, Neuchâtel.

Dodge Y (1987) (Ed.) Statistical Data Analysis based on the L1-norm and Related Methods. North-Holland, Amsterdam.

Dodge Y, Fedorov VV and Wynn HP (1988) (Eds) Optimal Design and Analysis of Experiments. North-Holland, Amsterdam.

Dodge Y (1989) (Ed.) Statistical Data Analysis and Inference. North-Holland, Amsterdam.

Dodge Y, Mehran F et Rousson M (1991) Statistique. Presses Académiques, Neuchâtel.

Dodge Y (1992) (Ed.) L1-Statistical Analysis and Related Methods. North-Holland, Amsterdam.

Dodge Y and Whittaker J (1992) (Eds) Computational Statistics I. Physica Verlag, Heidelberg.

Dodge Y and Whittaker J (1992) (Eds) Computational Statistics II. Physica Verlag, Heidelberg.

Birkes D and Dodge Y (1993) Alternative Methods of Regression. John Wiley and Sons, New York.

Dodge Y (1993) Statistique : Dictionnaire encyclopédique. Dunod, Paris.

Dodge Y (1997) (Ed.) L1-Statistical Procedures and Related Topics. IMS Lecture Notes and Monographs Series, Volume 31, Berkley, U.S.

Dodge Y (1999) Analyse de régression appliquée. En collaboration avec V. Rousson. Dunod, Paris.

Dodge Y and Jureckovà J (2000) Adaptive regression. Springer-Verlag, New York.

Dodge Y (2002) (Ed.) Statistical Data Analysis based on the L1 norm and related methods. Birkhauser, Basel.

Dodge Y (2003) The Oxford Dictionary of Statistical Terms. Oxford University Press, Oxford.

Avant-propos

L'idée de cet ouvrage est née d'un véritable besoin, celui de combler une lacune importante dans la littérature relative à la science statistique. En effet, il n'existait à ce jour aucun dictionnaire de statistique en langue française. Il est vrai qu'il s'agit d'une science relativement jeune, reconnue en tant que telle depuis le début du xx^{e} siècle seulement.

Notre objectif est de donner la définition des concepts les plus couramment utilisés tout en complétant l'information chaque fois que possible, par des historiques, des aspects mathématiques, des domaines et limitations, des exemples, des références et des mots-clés, qui sont donnés avec leur équivalent en anglais (un mot-clé étant une notion qui peut être liée au mot de base en fonction de plusieurs critères). Les mots apparaissant en gras dans le texte indiquent des correspondances qui renvoient à d'autres notions statistiques, permettant d'expliquer et de développer le terme considéré. Il faut toutefois préciser que ce dictionnaire n'a pas la prétention de réunir en un seul volume toutes les notions de statistique existantes à ce jour, cette science étant en constante évolution.

Le premier dictionnaire de cette science à avoir été publié est celui de Sir Maurice Kendall et William Buckland, en 1957. Commandé par l'Institut international de statistique avec l'appui financier de l'UNESCO, il a pour objet la définition des termes statistiques les plus utilisés dans le monde scientifique. En 2003, la sixième édition paraît sous le nom *The Oxford Dictionnary of Statistical Terms*, édité par moi-même.

En 1968, une encyclopédie internationale des statistiques en deux volumes est éditée par William Kruskal et Judith Tanur (rééditée en 1978). Elle comprend 174 articles de différents auteurs classés par ordre alphabétique.

Dès 1982, Samuel Kotz et Norman Johnson, ont publié une encyclopédie des sciences statistiques (*Encyclopedia of Statistical Sciences*) en dix volumes comptabilisant 6 604 pages au total. Or, de l'avis même des éditeurs, ces quelques volumes ne peuvent contenir tous les termes, concepts, notions et méthodes nés des recherches de ce siècle.

En 1998, Peter Armitage et Theodore Colton ont publié une encyclopédie de biostatistique en six volumes (*Encyclopedia of Biostatistics*).

La dernière encyclopédie en date est *The Encyclopedia of Environmetrics* éditée par A.H. El-Shaarawi et W.W. Piergorsch.

Le présent dictionnaire intéressera un vaste public comme celui des sciences économiques et sociales, des sciences humaines ou des sciences de la vie et de la terre. Il est aussi bien destiné aux chercheurs de divers domaines de sciences appliquées qu'aux étudiants envisageant d'approfondir la théorie statistique et ses applications. Certains domaines importants de la statistique moderne n'ont pas été abordés dans cet ouvrage, telles les méthodes statistiques génétiques et la reconnaissance d'image ; mais un dictionnaire exhaustif est impossible à réa-

liser en un seul volume. Nous serions reconnaissants aux lecteurs qui auraient des remarques à formuler, de bien vouloir nous les communiquer afin que nous puissions en tenir compte lors d'une prochaine édition.

Si le lecteur désire approfondir certains sujets, il pourra se reporter à la liste des références à la fin de chaque définition qui, bien que non exhaustive, pourra servir de guide. Il est également conseillé de consulter l'*Encyclopedia of Statistical Sciences*.

Ce manuscrit a été imprimé pour la première fois en 1993 avec le même titre par les éditions Dunod, Paris. Depuis, il a été complété et amélioré.

Pour la première édition, je tiens à remercier tout particulièrement Nicole Rebetez, Sylvie Gonano et Jürg Schmid pour leur étroite collaboration, ainsi que Séverine Pfaff qui a réalisé la saisie de cet ouvrage. Je témoigne ma profonde reconnaissance aux professeurs Gérard Antille, Bernard Fichet, Farhad Mehran et Jean-Pierre Renfer pour avoir apporté les dernières corrections à cet ouvrage, de même qu'à Rachid Maraï pour ses illustrations. Je remercie également mes assistants : Sandrine König, Valentin Rousson, ainsi que les professeurs Ludovic Lebart, Valérie Isham et Joe Whittaker pour leur précieuse collaboration. Enfin, je tiens à remercier le Fonds national suisse de la recherche scientifique (requêtes n° : 20-5648.88 et 20-28989.90) pour sa contribution financière.

Pour cette deuxième édition, je remercie Elisabeth Pasteur, Aline Lykiardopol et Jimmy Brignony pour leur participation à l'élaboration de cet ouvrage. Enfin, sans l'aide précieuse d'Alexandra Fragnière, qui a corrigé et recorrigé avec beaucoup de passion, je n'aurais pas pu présenter cet ouvrage sous sa forme actuelle.

Yadolah Dodge
Professeur de Statistique
Université de Neuchâtel, Suisse
Mars 2004

La statistique : une science omniprésente

INTRODUCTION

La statistique est une discipline qui concerne la quantification des phénomènes et des procédures inférentielles. Elle a trait, en particulier, aux problèmes de mise en œuvre et d'analyse des expériences et échantillons, à l'examen de la nature des erreurs d'observation et des sources de variabilité, et à la représentation simplifiée des grands ensembles de données. La théorie statistique est le cadre qui fournit un certain nombre de procédures que l'on appelle *méthodes statistiques*. Le terme *statistique* est parfois utilisé dans d'autres sens, par exemple, pour se référer non pas à une discipline globale comme décrite ici mais, de manière plus limitée, à un ensemble d'outils statistiques comprenant des formules, des tableaux et des procédures, Dans un sens plus étroit, le terme *statistique* est aussi employé pour désigner un ensemble de données numériques, comme par exemple, les statistiques de chômage en Suisse, et lorsqu'il est au singulier, il peut encore être utilisé pour désigner un paramètre numérique, une estimation calculée à partir d'observations de base. La statistique est une discipline dont les racines remontent loin dans l'histoire. Considérons, par exemple, la notion de moyenne arithmétique qui est une des plus anciennes méthodes utilisée pour combiner des observations afin d'obtenir une valeur approximative unique. Son emploi semble en effet remonter au troisième siècle avant notre ère, époque où les astronomes babyloniens l'utilisaient pour déterminer la position du soleil, de la lune et des planètes. La statistique en tant que discipline est née de l'étude et de la description des états, sous leurs aspects économiques et surtout démographiques. En effet, des formes de recensement se pratiquaient déjà à l'époque des grandes civilisations de l'Antiquité, où l'étendue et la complexité des empires nécessitaient une administration stricte et rigoureuse. Les pratiques dans ce domaine ont bien entendu fortement évolué avec le temps, les méthodes nouvelles se sont développées, et ainsi de la simple collecte des données, la statistique a peu à peu pris un sens nouveau. De plus à la notion de *comptage* ou de *mesure* est venu s'ajouter le problème de la comparaison. Or, la comparaison de mesures nécessite une connaissance de leur exactitude, de leur précision, ainsi qu'une maîtrise des techniques pour formuler l'incertitude rattachée aux valeurs. C'est là que la statistique inférentielle et la théorie des probabilités prennent toute leur importance. Dès le XVII^{e} siècle, de grands savants ont participé au développement de la statistique telle qu'on la connaît

aujourd'hui. Citons, parmi tant d'autres, Jakob Bernoulli au XVII^e siècle ; Pierre Simon de Laplace, Thomas Bayes et Roger Joseph Boskovich au XVIII^e ; Adrien-Marie Legendre, Carl Friedrich Gauss et Siméon-Denis Poisson, au XIX^e ; Karl Pearson, William Sealy Gosset, Ronald Aylmer Fisher, Jerzy Neyman et Egon Sharpe Pearson, Sir David Cox, Jack Kiefer, Bradley Efron, C.R. Rao, Peter Huber et Eric Lehmann au XX^e. Tous ont contribué à faire de la statistique une science en soi. La statistique moderne est beaucoup plus qu'un ensemble de méthodes, d'outils ou de techniques isolées utilisées dans des sciences spécifiques. Il existe une unité des méthodes statistiques, une logique cohérente, une base scientifique, mathématique et probabiliste, solide. La statistique n'est véritablement reconnue comme discipline en tant que telle que depuis le début du XX^e siècle. C'est donc une science relativement jeune, en pleine expansion que ce soit au niveau de la recherche fondamentale ou des applications. Par ailleurs, de nombreuses branches nouvelles, ainsi qu'un grand nombre de spécialisations se sont développées : la théorie de la décision, des séries chronologiques, l'analyse multivariée, l'économétrie, la théorie des jeux, l'inférence non paramétrique, reéchantillonage, envirométrie, biostatistique, statistique génétique, statistique spatiale et simulation (MCMC), modélisation (GLM, régression), modèles graphiques.

RECHERCHE ET STATISTIQUE

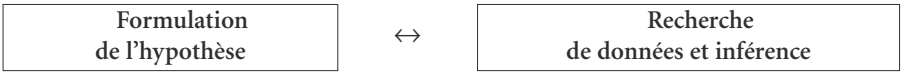
La recherche est une investigation ou une expérimentation critique ayant pour objectif la découverte et l'interprétation correcte de nouveaux faits ou de nouvelles relations entre différents phénomènes. Elle a également pour fonction la vérification de lois, conclusions ou théories acceptées, et à la lumière de faits nouveaux, le développement de nouvelles lois, de nouvelles conclusions et de nouvelles théories.

Comment accédons-nous à la connaissance des phénomènes naturels ? Quels sont les processus mentaux impliqués et comment aborder les investigations ? Comment juger si les données et les résultats obtenus sont exacts ? Ces questions ont en effet troublé la pensée humaine et sont restées longtemps un sujet de discours philosophiques. Cependant, les progrès récents de la logique et de la science statistique nous ont permis de leur apporter des réponses satisfaisantes.

La connaissance d'un phénomène naturel est habituellement définie en des termes et des lois qui permettent de prévoir les évènements de façon précise dans les limites raisonnables du possible. C'est le cas pour la loi de la dynamique de Newton et pour celle de la génétique de Mendel.

Comment ces lois ou théories s'établissent-elles ? Il existe une méthode scientifique. En premier lieu, une loi est formulée en tant qu'hypothèse chargée d'expliquer certains évènements observés, puis, les conséquences de l'hypothèse sont résolues par un raisonnement déductif et vérifiées par d'autres observations provenant d'expériences soigneusement conçues. Si les

données sont en contradiction avec l'hypothèse, cette dernière est écartée et une nouvelle hypothèse est formulée. Dans le cas contraire, elle est provisoirement acceptée et on lui accorde le statut de loi avec des limites spécifiques. Ceci laisse la possibilité de remplacer la loi au cours du temps par une autre loi plus universelle appuyée par un nombre plus important de données. Le cycle logique de la méthode scientifique de la recherche peut être schématisé ainsi :



De ces deux phases, la *recherche de données* est traitée par la statistique, alors que la *formulation de l'hypothèse* naît de la curiosité et de l'observation. Ces deux phases demandent beaucoup d'imagination et des connaissances techniques poussées pour aboutir à un progrès scientifique. Il est à souligner que les lois scientifiques ne progressent ni par le principe de l'autorité, ni par le mythe ou la philosophie moyenâgeuse. **La statistique reste la seule base des connaissances actuelles**.

Grâce à une analyse fine des données destinées à tester des hypothèses fixées, la statistique permet au scientifique d'utiliser au maximum ses capacités créatives et de découvrir des phénomènes nouveaux. Une découverte est, sans doute, le fruit d'une illumination soudaine, Mais ceci arrive seulement à un cerveau capable de trier le bon grain de l'ivraie, de chercher des éléments nouveaux et d'être attentif à l'inattendu pour mettre le tout ensemble, afin de provoquer l'étincelle qui donnera naissance à une découverte. Le travail du statisticien consiste à faire ressortir le maximum d'informations des données pour que les chiffres révèlent ce qu'ils contiennent. Selon son expérience, ses connaissances dans le domaine concerné et sa collaboration avec les experts compétents, le statisticien peut faire un examen plus approfondi des données pour comprendre un phénomène caché.

Cependant, l'acquisition de la connaissance n'est pas toujours liée à des théories ou des lois. Quelle quantité de blé produira-t-on cette saison ? La fumée provoque-t-elle le cancer ? Qui a écrit Hamlet : Shakespeare, Bacon ou Marlowe ? Quelle pourra être la place exacte d'une tumeur dans le cerveau d'un patient ? Comment les langues ont-elles évolué ? Comment mesurer l'effet d'un vaccin contre la polio ? Voici quelques questions ayant été résolues par la statistique et qui ont ainsi contribué au développement des connaissances.

EXEMPLES

Considérons maintenant quelques illustrations d'approches statistiques au service de la recherche de la vérité ou de nouvelles lois.

Diagnostic statistique des maladies

Les médecins font des diagnostics de maladie sur la base de quelques tests et symptômes. Le fait que deux médecins puissent donner des diagnostics différents, à partir de mêmes données et cela même pour des cas simples, montre que les tests sont insuffisants et que chaque médecin utilise sa propre expérience pour établir un diagnostic. Bien sûr, on pourrait effectuer davantage de tests (ce qui augmenterait considérablement les coûts). Mais le problème réel est celui de l'efficacité des tests. Tous les médecins ne sont pas en mesure de comprendre ou de travailler avec une grande masse de données qui, le plus souvent, créent la confusion. Il faut de plus pouvoir relier de petits indices tirés de tests spécifiques pour arriver à la meilleure conclusion. Pour faire face à de telles situations, les statisticiens ont développé ce qu'on appelle l'analyse de variantes multiples par laquelle l'information des mesures individuelles (comme les tests cliniques) peut être combinée de façon efficace pour répondre à une question spécifique (par exemple, quelle est la nature de la maladie ?).

En utilisant les dossiers des patients et l'expérience des meilleurs médecins, on peut construire une fonction discriminante fondée sur divers tests cliniques et ainsi déterminer les tests et valeurs qui peuvent au mieux distinguer les différentes maladies.

Le style littéraire quantifié

La statistique ne se limite pas seulement à la recherche scientifique. À en juger par les articles de la presse scientifiquement « branchée », on est frappé de voir l'accent mis sur les méthodes quantitatives et statistiques, même dans les domaines les plus inattendus comme les arts.

Discutons de quelques explications statistiques dans le domaine de la littérature et de l'étude des langues. L'œuvre de Platon est restée intacte après vingt-deux siècles et ses idées philosophiques tout comme son style, ont été étudiés dans les moindres détails. Malheureusement, personne n'a mentionné ou ne savait l'ordre chronologique correct de ses trente-cinq dialogues, de ses six pièces courtes et de ses treize lettres. Le problème de la chronologie des œuvres de Platon se posait déjà au siècle passé, mais aucun progrès n'apparut. Il y a quelques années, les statisticiens se sont mis à l'œuvre et ont réussi à établir un ordre qui semble assez logique.

La méthode statistique commence par établir pour chaque paire d'œuvres un index de similitude. Dans une étude menée par Lilian I. Boneva (*Mathematics in Archeological and Historical Sciences*, 1971, pp. 174-185), l'index était fondé sur la fréquence dans chaque œuvre des trente-deux descriptions possibles des cinq dernières syllabes d'une phrase.

En se fondant uniquement sur l'hypothèse que les œuvres les plus rapprochées dans le temps se ressemblent d'avantage que celles écrites à une autre période, une méthode a pu être développée et appliquée pour résoudre le problème de leur ordre chronologique.

En étudiant la parenté des langues indo-européennes (le latin, le sanscrit, l'allemand, le persan, le celte, etc.), la linguistique a découvert un ancêtre commun qui a été parlé, on le

pense, il y a quatre mille cinq cent ans. Si un tel ancêtre existe, il y a donc un arbre évolutif des langues qui s'est ramifié à différents moments de l'histoire. Est-il possible de construire cet arbre des langues comme les biologistes l'ont fait avec l'arbre évolutif de la vie? Il s'agit d'un problème très intéressant et scientifiquement important qui relève de ce que l'on appelle la *glottochronologie*. En utilisant une grande quantité d'informations sur les similitudes entre les langues et un raisonnement fort compliqué, les linguistes ont réussi à établir quelques branches principales de l'arbre, mais les relations exactes entre elles et le moment de l'histoire où elles se sont séparées restent à déterminer. Mais, une approche purement statistique de ce problème, tout en utilisant beaucoup moins d'informations, a pu donner des résultats encourageants.

La première étape d'une telle étude est de comparer une catégorie de mots appartenant à des langues différentes et ayant une utilité commune (l'œil, la main, la mère, etc.). Les mots de même sens appartenant à des langues différentes sont marqués d'un « + » s'ils sont analogues, et d'un « - » dans le cas contraire. La comparaison de deux langues devient ainsi une séquence de « + » et de « – ».

En utilisant uniquement cette information, Morris Swadesh (*Proceeding of the American Philosophical Society*, 1952, 96, pp. 452-463) a proposé une méthode pour estimer l'époque de la séparation de chaque paire de langues. Une fois les dates de séparation de chaque paire de langues connues, il est facile de construire un arbre évolutif

La statistique servant à détecter les erreurs

Puisque l'acceptation d'une théorie dépend de sa vérification sur des données observées, le scientifique peut être tenté de *truquer* les données afin de pouvoir imposer ses idées par des théories bricolées. À part le fait que cette pratique est immorale, elle cause aussi beaucoup de tort au progrès de la science. Sans doute, une théorie fausse, sera-t-elle repérée un jour ou l'autre. Il est toutefois possible qu'elle nuise à la société aussi longtemps qu'elle est acceptée. Un exemple récent est celui de la fraude impliquant Cyril Burt, *le père de la psychologie pédagogique de l'Angleterre*, alors qu'il étudiait le quotient d'intelligence (QI). Sa théorie, selon laquelle les différences d'intelligence étaient héréditaires et ne dépendraient pas de facteurs socio-économiques, dirigea la pensée politique en matière d'éducation sur une mauvaise voie. Or, on a montré que l'augmentation du taux de vitamines B et de fer chez les femmes enceintes des groupes socio-économiques désavantagés, affectait positivement les scores moyens de leurs enfants de 5 à 8 points. Il semble évident que des facteurs autres que l'hérédité interviennent aussi dans les résultats des tests.

De telles falsifications peuvent être détectées par le fait que, comme Haldone l'a mentionné, l'homme est un animal méthodique qui ne peut pas imiter le désordre de la nature. En utilisant cette limitation, les statisticiens ont pu inventer des techniques permettant de détecter des données falsifiées.

Quelques grands scientifiques ont falsifié des données dans leur souci de faire avancer des théories qu'ils croyaient irréfutables et indispensables. R.A. Fischer a examiné les données expérimentales à partir desquelles Mendel découvrit les lois fondamentales de l'hérédité. Dans la plupart de ses expériences, les rapports des phénomènes observés étaient presque trop proches de ceux prévus par Mendel. Si on lance, par exemple, une pièce de monnaie 100 fois, il y a des chances d'obtenir 50 fois face, mais si vous prétendez obtenir ce résultat très souvent, on pourrait vous soupçonner de fraude. Ainsi Fisher a montré que la précision des données de Mendel (en accord avec sa théorie) était frappante à un point tel que si l'on calculait la probabilité de ce qu'il a trouvé expérimentalement, on arriverait à un événement qui se produirait 7 fois sur 100 000. Il commente ainsi cette probabilité rarissime : *malgré le fait que l'on ne peut pas donner d'explication satisfaisante, il très probable que Mendel n'ait pas voulu montrer les résultats à ses assistants qui connaissaient aussi bien que lui ce qu'il fallait trouver pour que la théorie soit cohérente.* Cet argument est renforcé par le fait que l'on voit des altérations sur les feuilles de mesures de Mendel, afin que les calculs soient en accord avec ses espérances.

Jusqu'à quel point êtes-vous célèbre ?

Les scènes de lit de mort où une mère mourante reste accrochée à la vie juste assez longtemps pour que son fils, absent de longue date, revienne ne sont que trop fréquentes dans les films. Mais est-ce que cela arrive aussi dans la réalité ? Est-il possible d'ajourner sa mort jusqu'à un évènement important ? On a cru que les gens célèbres pouvaient le faire, dans le sens où ils attachaient une grande importance à leur anniversaire. Une étude menée par David P. Philipps (*Statistics : a Guide to the Unknown*, 1978, pp. 71-85) renforce cette idée avec des faits bien précis. Phillips recueillit les mois de naissance et de mort de 1 251 Américains célèbres, les dates des morts étant classées en trois parties : les mois avant, le mois même et les mois après la date de la personne (voir tableau).

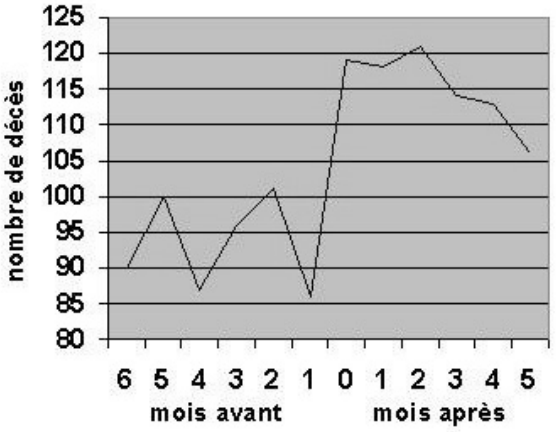
L'échantillon 1 correspond aux personnes très connues énumérées dans Four Hundred Notable Americans et l'échantillon 2 aux personnes remarquables identifiées à partir des trois tomes de *Who was who* pour les années 1951-1960, 1943-1950 et 1897-1947.

Si la date de la mort n'a rien à voir avec celle de la naissance, le nombre attendu de décès pour chacune des trois périodes du tableau devrait être 1 251/12, ce qui fait à peu près 104.

On voit que, pour les deux échantillons, la fréquence observée des décès survenant durant le mois de la naissance est nettement supérieure aux fréquences observées pour les deux autres périodes étudiées ; ces résultats sont illustrés ci-dessous.

Nombre de décès avant, pendant et après le mois de naissance

	Mois Avant						Mois de naissance	Mois Après					Total
	6	5	4	3	2	1		1	2	3	4	5	
Échantillon 1	24	31	20	23	34	16	26	36	37	41	26	34	348
Échantillon 2	66	69	67	73	67	70	93	82	84	73	87	72	903
Ensemble	90	100	87	96	101	86	119	118	121	114	113	106	1 251
1251/12=104	104	104	104	104	104	104	104	104	104	104	104	104	



Or, il est frappant de constater que nombreux sont les individus observés qui meurent au cours du mois de leur naissance, mais il est encore un phénomène plus étonnant : il existe, en effet, une forte diminution de la fréquence des décès précédant les dates de naissance, ce qui pourrait s'expliquer par un ajournement de la mort jusqu'à la date d'anniversaire.

Le rythme journalier

Si l'on vous demande votre taille, vous avez sans doute une réponse toute prête. Quelqu'un vous aura mesuré auparavant et vous l'aura communiquée. Mais vous ne vous êtes peut-être pas demandé pourquoi ce chiffre représente votre taille, et si vous l'aviez fait, on vous aurait

peut-être répondu qu'il a été obtenu en suivant une procédure *ad hoc*. En pratique, cette définition semble satisfaisante. Mais on peut se poser d'autres questions : la caractéristique que nous étudions dépend-elle de l'heure de la mesure ? Par exemple, y a-t-il une différence de taille entre le matin et le soir pour une personne donnée ? Si oui, quelle est l'amplitude de cette différence et quelle en est l'explication physiologique ?

Une enquête statistique simple peut fournir la réponse. Des mesures soigneusement récoltées sur 41 étudiants à Calcutta ont montré une différence de 9,3 mm en moyenne, la mesure du matin étant chaque fois supérieure (C.R. Rao, *Sankhya*, 19, 1957, pp. 96-98). Si, en réalité, la taille d'une personne était constante durant toute la journée, chaque différence de taille observée devrait être attribuée à des erreurs de mesure qui peuvent être soit positives, soit négatives avec une probabilité égale. Dans pareil cas, la probabilité que les 41 différences observées soient positives est de l'ordre de 2-41, ce qui correspond à un évènement qui arrive moins de cinq fois en 1 013 expériences. Ceci révèle que la probabilité que l'hypothèse de la différence de taille en cours de journée soit fausse est extrêmement faible. On peut donc accepter l'idée que nous nous allongeons d'un centimètre pendant que nous dormons et que nous nous rapetissons de la même longueur lorsque nous sommes au travail ! L'explication physiologique plausible est que, pendant la journée, les vertèbres s'écrasent les unes sur les autres, alors qu'une fois couché elles se mettent dans leur position normale.

CONCLUSION

Nous avons vu que la statistique joue un rôle essentiel dans diverses branches de la recherche pure ou appliquée. Son omniprésence a ses racines dans le système scientifique développé par les statisticiens pour la collecte, l'organisation, l'analyse, l'interprétation et la présentation de l'information qui peut être donnée sous forme chiffrée.

Le besoin que la recherche a de travailler avec des méthodes quantitatives a fait de la statistique un outil indispensable, surtout dans les sciences dites *douces* comme l'économie, la psychologie ou la sociologie.

L'importance du développement de l'enseignement et de la recherche dans le domaine des statistiques ne cesse de croître. Pour s'en convaincre, il n'y a qu'à se rappeler qu'en 1992, 2 254 professeurs ordinaires enseignaient cette discipline dans les quelques 289 départements de statistiques que comptent les universités nord-américaines. La même année, les diplômes délivrés par ces départements atteignaient le nombre impressionnant de 1 149 licences, 1 121 postgrades et 392 doctorats.

Aux États-Unis, l'association de statistique dénombre à elle seule 10 000 membres. L'importance des publications en statistique mérite également d'être soulignée, puisqu'au niveau mondial, on dénombre 415 périodiques.

Les avantages des applications statistiques dans l'industrie et dans l'administration sont bien connus. Dans les usines où les méthodes modernes de la statistique ont été utilisées, la production a augmenté de 10 à 100 %, sans recours à des investissements supplémentaires ou à une extension de l'usine. Ainsi, la connaissance statistique est une ressource naturelle. La statistique peut jouer un rôle vital dans les affaires nationales d'un gouvernement. Pour les statisticiens, la connaissance des technologies et celle de leur propre environnement professionnel constitue une exigence indispensable. Ils créent, entre les différentes disciplines, le lien qui permet aux responsables de prendre les décisions nécessaires au progrès. On peut voir que plus un pays est avancé, plus son système de statistique est performant et consulté.

BIBLIOGRAPHIE

Armitage, P. and Colton, T. (1998). *Encyclopedia of Biostatistics*. John Wiley & Sons, New York, 6 volumes.

Dodge, Y. (1993). *Statistique : Dictionnaire encyclopédique.* Dunod, Paris.

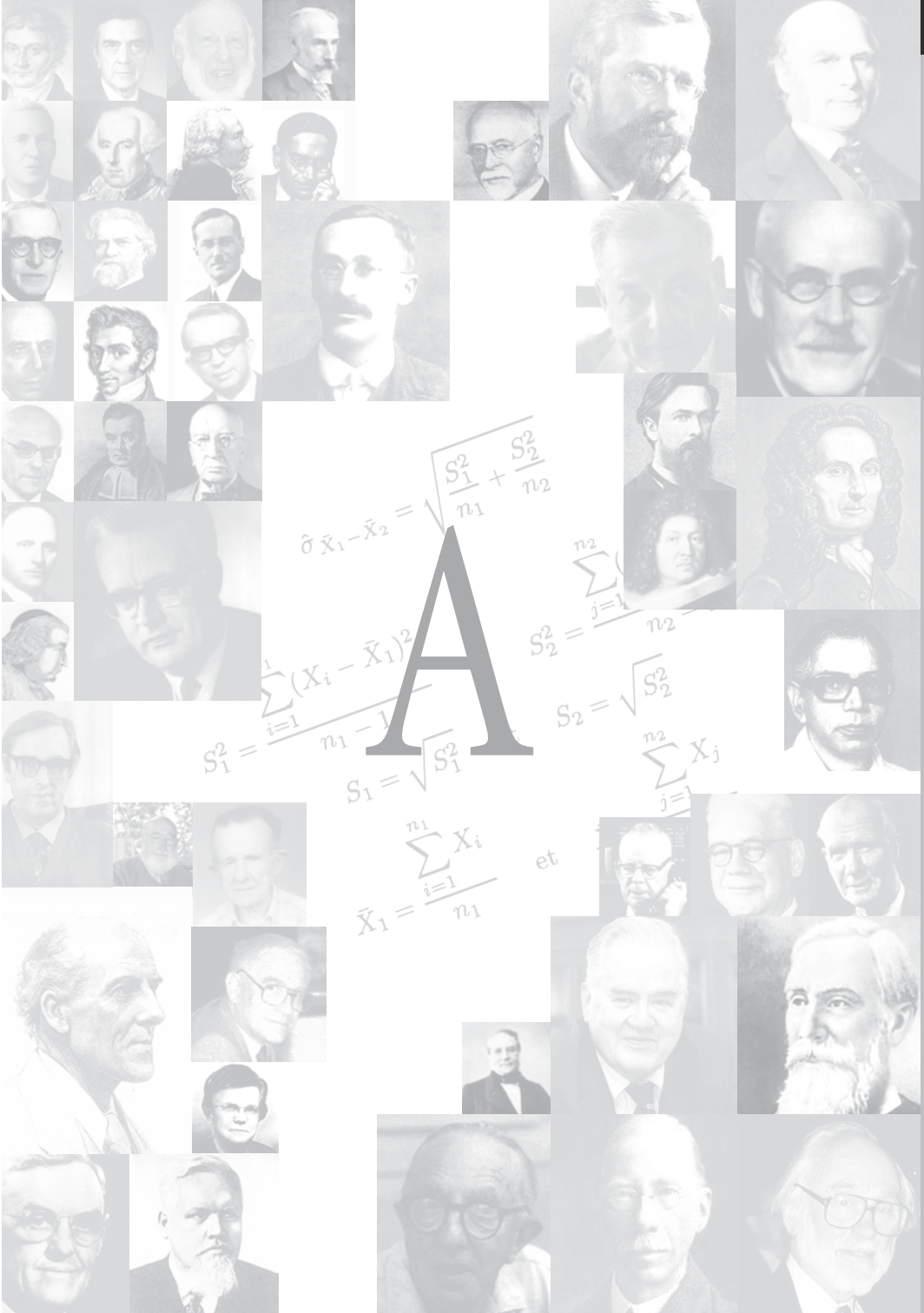
Dodge, Y. (2003). *The Oxford Dictionary of Statistical Terms.* Oxford : Oxford University Press.

El-Shaarawi, A.H., Piegorsch, W.W. (2001). *The Encyclopedia of Environmetrics.* John Wiley & Sons, New York, 4 volumes.

Kendall, M.G. and Bucklard, W.R. (1968). *Dictionary of Statistical Terms.* Longman Scientific & Technical, Essex. Fifth edition, Marriott, F.H.C. (1990).

Kotz, S. and Johnson, N.L. (1982). *Encyclopedia of Statistical Science.* John Wiley & Sons, New York, 10 volumes and update volumes 1 and 2.

Kruskal, W.H. and Tanur, J.M. (1978). *International Encyclopedia of Statistics.* The Free Press, New York.


A
$\hat{\sigma}_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$
$S_1^2 = \frac{\sum_{i=1}^{n_1}(X_i - \bar{X}_1)^2}{n_1 - 1}$
$S_1 = \sqrt{S_1^2}$
$\bar{X}_1 = \frac{\sum_{i=1}^{n_1} X_i}{n_1}$
et
$S_2^2 = \frac{\sum_{j=1}^{n_2}}{n_2}$
$S_2 = \sqrt{S_2^2}$
$\sum_{j=1}^{n_2} X_j$

ADÉQUATION, TEST D'

Goodness of fit test

Voir **test d'adéquation**.

ALGORITHME

Algorithm

Un algorithme est un processus déterminé formé d'une séquence d'étapes bien définies menant à la solution d'un certain type de problèmes.

Ce processus peut être itératif, c'est-à-dire répété plusieurs fois. Il est généralement numérique.

HISTORIQUE

Le terme algorithme vient de la prononciation latine du nom de Abu Ja'far Muhammad ibn Musa Al-Khwarizmi, mathématicien du IX^{e} siècle vivant à Bagdad et précurseur de l'algèbre.

DOMAINES ET LIMITATIONS

Le mot algorithme a pris ces dernières années un autre sens avec l'arrivée des ordinateurs. Il fait alors référence à la description d'un processus dans une forme appropriée à l'exécution d'un programme sur ordinateur.

Le but principal des logiciels statistiques actuels est de dévélopper conjointement les langages de programmation ainsi que le développement des algorithmes statistiques de telle manière que ces algorithmes puissent être présentés sous une forme compréhensible par l'utilisateur. Ceci a pour avantage que l'utilisateur comprend les résultats produits par l'algorithme et qu'il a une plus grande confiance en l'exactitude des solutions. Parmi les revues statistiques traitant des algorithmes, citons notamment le *Journal of Algorithms* édité par l'*Academic Press*, la partie de *Applied Statistics* du *Journal of the Royal Statistical Society-serie C* traitant spécifiquement des algorithmes, *Computational Statistics* édité par *Physica-Verlag* (Heidelberg) et encore *Random Structures and Algorithms* édité par *Wiley* (New York).

EXEMPLE

Nous allons donner ci-dessous un algorithme permettant de calculer la valeur absolue d'un nombre différent de zéro, soit $|x|$.

Processus :

Étape 1. Identifier le signe algébrique du nombre donné.

Étape 2. Si le signe est négatif, passer à l'étape 3. Si le signe est positif, désigner la valeur absolue du nombre donné par le nombre lui-même :

$$|x| = x$$

et arrêter le processus.

Étape 3. Désigner la valeur absolue du nombre donné par l'opposé du nombre :

$$|x| = -x$$

et arrêter le processus.

MOTS CLÉS

Algorithme de Yates *(Yates' algorithm)*
Logiciel statistique *(Statistical software)*

RÉFÉRENCES

Chambers, J.M. (1977). *Computational Methods for Data Analysis.* John Wiley & Sons, New York.

Khwarizmi, Musa ibn Meusba (IX^{e} siècle). Jabr wa-al-muqeabalah. *The algebra of Mohammed ben Musa.* Edited and transl. by Frederic Rosen. Georg Olms Verlag.

Rashed, R. (1985). La naissance de l'algèbre, in *Le Matin des mathématiciens*, Émile Noël éditeur. Belin-Radio France.

ALGORITHME DE YATES

Yates' algorithm

L'**algorithme** de Yates est un processus utilisé pour calculer les **estimateurs** des **effets principaux** et des **interactions** dans une **expérience factorielle**.

HISTORIQUE

L'**algorithme** de Yates a été mis au point par Frank Yates en 1937.

ASPECTS MATHÉMATIQUES

On considère une **expérience factorielle** à n facteurs chacun ayant deux niveaux. On note cette **expérience** 2^n.

Les facteurs sont notés par des lettres majuscules. On emploie une lettre minuscule pour le niveau supérieur du facteur et on omet la lettre pour le niveau inférieur. Lorsque tous les facteurs sont au niveau inférieur, on utilise le symbole (1).

L'**algorithme** se présente sous la forme d'un tableau. Dans la première colonne, on trouve toutes les combinaisons des niveaux des facteurs dans un ordre standard (l'introduction d'une lettre est suivie par les combinaisons avec toutes les combinaisons précédentes). La deuxième colonne donne les observations de toutes les combinaisons.

Pour les n colonnes suivantes, la procédure est la même : la première moitié de la colonne est obtenue en sommant les paires de valeurs successives de la colonne précédente ; la seconde moitié de la colonne est obtenue en soustrayant la première valeur de la deuxième valeur de chaque paire.

La colonne n contient les contrastes des effets principaux et des interactions désignés dans la première colonne. Pour trouver les estimateurs de ces effets, on divise la n-ème colonne par $m \cdot 2^{n-1}$ où m est le nombre de répétitions de l'expérience et n le nombre de facteurs. En ce qui concerne la première ligne, on trouve dans la colonne n la somme de toutes les observations.

DOMAINES ET LIMITATIONS

L'algorithme de Yates peut être appliqué à des **expériences** comprenant des facteurs ayant plus de deux niveaux.

À l'aide de cet **algorithme**, on peut aussi trouver les sommes des carrés correspondant aux effets principaux et interactions de l'**expérience factorielle** 2^n. Elles sont données en élevant la n-ème colonne au carré et en divisant le résultat par $m \cdot 2^n$ (où m est le nombre de répétitions et n le nombre de facteurs).

EXEMPLES

Afin de calculer les estimations des effets principaux et celles des trois interactions, on considère une expérience factorielle 2^3 (c'est-à-dire trois facteurs à deux niveaux) relative à la production agricole de la betterave sucrière. Le premier facteur noté A concerne le type d'engrais (fumier ou non fumier), le second noté B la profondeur du labourage (20 cm ou 30 cm) et le dernier, soit C, la variété de betterave sucrière (variété 1 ou 2). Puisque chaque facteur compte deux niveaux, on peut parler d'une expérience $2 \times 2 \times 2$.

L'interaction entre le **facteur** A et le facteur B, par exemple, est notée par $A \times B$ tandis que l'interaction entre les trois facteurs est désignée par $A \times B \times C$.

Les observations effectuées au nombre de huit indiquent le poids de la récolte, exprimé en kilos, pour un demi-hectare. On a ainsi obtenu :

A		Fumier		Non fumier	
B		20 cm	30 cm	20 cm	30 cm
C	Var. 1	80	96	84	88
	Var. 2	86	96	89	93

En réalisant cette expérience, on cherche à quantifier non seulement l'effet particulier de l'engrais, du labourage et de la variété de betterave (effets simples) mais aussi l'impact sur la récolte des effets combinés (effets d'interaction).

Pour établir le tableau résultant de l'application de l'algorithme de Yates, il faut tout d'abord classer les huit observations selon l'ordre dit de Yates. Pour cela, l'usage du tableau de classement est pratique ; les symboles − et + représentent respectivement les niveaux inférieurs et supérieurs des facteurs. Par exemple, la première ligne indique l'observation "80" correspondant aux niveaux inférieurs pour chacun des trois facteurs (c'est-à-dire A : fumier, B : 20 cm et C : variété 1).

Tableau de classement des observations selon l'ordre de Yates

N° Obs.	Facteurs A	B	C	Observation correspondante
1	−	−	−	80
2	+	−	−	84
3	−	+	−	96
4	+	+	−	88
5	−	−	+	86
6	+	−	+	89
7	−	+	+	96
8	+	+	+	93

On peut désormais construire le tableau de l'algorithme de Yates :

Tableau correspondant à l'algorithme de Yates

Combinaisons	Obs.	Colonnes 1	2	3	Estimations	Effets
(1)	80	164	348	712	89	moyenne
a	84	184	364	−4	−1	A
b	96	175	−4	34	8,5	B
ab	88	189	0	−18	−4,5	$A \times B$
c	86	4	20	16	4	C
ac	89	−8	14	4	1	$A \times C$
bc	96	3	−12	−6	−1,5	$B \times C$
abc	93	−3	−6	6	1,5	$A \times B \times C$

On trouve les termes de la colonne numérotée 2, par exemple, en faisant :

- pour la première moitié de cette colonne (somme des éléments de chacune des paires de la colonne précédente soit 1) :

$$\begin{aligned} 348 &= 164 + 184 \\ 364 &= 175 + 189 \end{aligned}$$

$$\begin{aligned} -4 &= 4 + (-8) \\ 0 &= 3 + (-3) \end{aligned}$$

- pour la deuxième moitié (différence des éléments de chacune des paires, à l'inverse de précédemment) :

$$\begin{aligned} 20 &= 184 - 164 \\ 14 &= 189 - 175 \\ -12 &= -8 - 4 \\ -6 &= -3 - 3 \end{aligned}$$

L'estimation pour l'effet principal du facteur A est égale à -1. En effet, on a divisé la valeur -4 (colonne 3) par $m \cdot 2^{n-1} = 1 \cdot 2^{3-1} = 4$ où m vaut 1 puisque l'expérience n'a été réalisée qu'une seule fois (sans répétition) et n correspond au nombre de facteurs, soit 3 dans notre cas.

MOTS CLÉS

Analyse de variance *(Analysis of variance)*
Contraste *(Contrast)*
Effet principal *(Main effect)*
Expérience factorielle *(Factorial experiment)*
Interaction *(Interaction)*
Plan d'expérience *(Design of experiments)*

RÉFÉRENCE

Yates, F. (1937). The design and analysis of factorial experiments. *Technical Communication of the Commonwealth Bureau of Soils*, 35, Commonwealth Agricultural Bureaux, Farnham Royal.

ANALYSE COMBINATOIRE

Combinatory analysis

D'une façon générale, l'analyse combinatoire se réfère aux techniques permettant de déterminer le nombre d'éléments d'un ensemble particulier sans les compter individuellement.

Les éléments pourraient être les résultats d'une **expérience** scientifique ou les issues différentes d'un **événement** quelconque.

On peut souligner en particulier trois notions liées à l'analyse combinatoire :

— permutations ;

— combinaisons ;

— arrangements.

HISTORIQUE

Les problèmes d'analyse combinatoire ont depuis longtemps intéressé les mathématiciens. En effet, selon Lajos Takács (1982), ces questions remontent à la Grèce antique. Toutefois, les Hindous, les Perses avec Omar Khayyâm, poète et mathématicien, et surtout les Chinois y consacrèrent également leurs travaux. Un livre chinois vieux de 3 000 ans, *I Ching* donna les arrangements possibles d'un ensemble de n éléments, avec $n \leq 6$. En 1303, Shih-chieh Chu publia un ouvrage intitulé *Ssu Yuan Yü Chien* (Précieux miroir des quatre éléments) dont la couverture présente, arrangée en triangle, une **combinaison** de k éléments pris d'un ensemble de taille n, avec $0 \leq k \leq n$. Le problème du **triangle arithmétique** fut traité par plusieurs mathématiciens européens, dont Michael Stifel, Nicolo Fontana Tartaglia et Pierre Hérigone, mais surtout Blaise Pascal qui rédigea en 1654 un *Traité du triangle arithmétique* qui ne sera publié qu'en 1665.

Si l'on peut citer un autre écrit historique sur les combinaisons, publié en 1617, par Erycius Puteanus et intitulé *Erycii Puteani Pretatis Thaumata in Bernardi Bauhusii è Societate Jesu Proteum Parthenium*, ce ne fut qu'avec les travaux de Pierre de Fermat (1601-1665) et de Blaise Pascal (1623-1662) que l'analyse combinatoire prit véritablement toute son importance. En revanche, le terme « analyse combinatoire » fut introduit par Gottfried Wilhelm von Leibniz (1646-1716) en 1666. Dans son ouvrage *Dissertatio de Arte Combinatoria*, il étudia systématiquement les problèmes d'arrangements, de permutations et de combinaisons.

D'autres travaux méritent également d'être cités, comme ceux de John Wallis (1616-1703), rapportés dans *The Doctrine of Permutations and Combinations, being an essential and fundamental part of the Doctrines of Chances*, ou ceux de **Jakob Bernoulli**, d'**Abraham de Moivre**, de Girolamo Cardano (1501-1576), et de Galileo Galilei (Galilée) (1564-1642).

Dans la seconde moitié du XIX^{e} siècle, Arthur Cayley (1829-1895) résolut certains problèmes de cette analyse en utilisant des graphes qu'il désigna sous le nom d'« arbres » ; enfin on ne saurait évoquer ce sujet sans mentionner un ouvrage important, celui de Percy Alexander MacMahon (1854-1929), paru sous le titre *Combinatory Analysis* (1915-1916).

EXEMPLES

Voir **arrangement**, **combinaison** et **permutation.**

MOTS CLÉS

Arrangement *(Arrangement)*
Binôme *(Binomial)*
Combinaison *(Combination)*
Loi binomiale *(Binomial distribution)*
Permutation *(Permutation)*
Triangle arithmétique *(Arithmetic triangle)*

RÉFÉRENCES

MacMahon, P.A. (1915-1916). *Combinatory Analysis.* Vol. I et II. Cambridge University Press, Cambridge.

Stigler, S. (1986). *The History of Statistics, the Measurement of Uncertainty before 1900.* The Belknap Press of Harvard University Press. London, England.

Takács, L. (1982). Combinatorics. In Kotz S. and Johnson N.L. (editors). *Encyclopedia of Statistical Sciences.* Vol 2. John Wiley & Sons, New York.

ANALYSE DE COVARIANCE

Covariance analysis

L'analyse de covariance est une technique d'**estimation** et de test des effets des traitements. Elle permet

de déterminer s'il existe une différence significative entre plusieurs moyennes de traitements en tenant compte des valeurs observées de la **variable** avant le traitement.

L'**analyse de covariance** accroît la précision des comparaisons de traitement puisqu'elle implique l'ajustement de la variable de réponse Y à une variable concomitante X qui correspond aux valeurs observées avant le traitement.

HISTORIQUE

L'analyse de covariance date des années 1930. C'est **Ronald Aylmer Fisher** (1935) qui a développé cette méthode pour la première fois.

Par la suite, d'autres auteurs appliqueront l'analyse de covariance à des problèmes de type agricole ou médical. Nous citerons par exemple Maurice Stephenson Bartlett (1937) qui reprend des études qu'il avait faites sur la culture du coton en Égypte et sur le rendement laitier des vaches en hiver et auxquelles il applique l'analyse de covariance.

C'est Daniel Bertrand DeLury (1948) qui effectue une analyse de covariance pour comparer les effets de différents médicaments (atropine, quinidine, atrophine) sur les muscles des rats.

ASPECTS MATHÉMATIQUES

Considérons une analyse de covariance pour un **plan complètement randomisé** impliquant un seul **facteur**.

Le **modèle** linéaire que nous devons considérer est le suivant :

$$Y_{ij} = \mu + \tau_i + \beta X_{ij} + \varepsilon_{ij}, \qquad \begin{matrix} i = 1, 2, \ldots, t \\ j = 1, 2, \ldots, n_i \end{matrix}$$

où

- Y_{ij} représente la j-ème **observation** recevant le **traitement** i,
- μ la **moyenne** générale commune à tous les traitements,
- τ_i l'effet réel sur l'observation du traitement i,
- X_{ij} la **valeur** de la variable concomitante et
- ε_{ij} l'**erreur** expérimentale de l'observation Y_{ij}.

Calcul des sommes

Notons $\bar{X}_{i.}$ et $\bar{Y}_{i.}$ respectivement les moyennes des valeurs des X et des Y pour le traitement i et $\bar{X}_{..}$ et $\bar{Y}_{..}$ respectivement les moyennes de toutes les valeurs de X et de Y. Pour pouvoir effectuer le **test de Fisher** et calculer le ratio F qui va permettre de déterminer s'il existe une différence significative entre les traitements, nous avons besoin des sommes des carrés et des sommes des produits suivantes :

1. La somme des carrés totale pour X :

$$S_{XX} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{..}\right)^2,$$

2. La somme des carrés totale pour Y (S_{YY}) :
elle se calcule de la même manière que S_{XX} en remplaçant les X par des Y,

3. La somme des produits totale de X et Y :

$$S_{XY} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{..}\right)\left(Y_{ij} - \bar{Y}_{..}\right),$$

4. La somme des carrés des traitements pour X :

$$T_{XX} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} \left(\bar{X}_{i.} - \bar{X}_{..}\right)^2,$$

5. La somme des carrés des traitements pour Y (T_{YY}) :
elle se calcule de la même manière que T_{XX} en remplaçant les X par des Y,

6. La somme des produits des traitements de X et Y :

$$T_{XY} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} \left(\bar{X}_{i.} - \bar{X}_{..}\right)\left(\bar{Y}_{i.} - \bar{Y}_{..}\right),$$

7. La somme des carrés des erreurs pour X :

$$E_{XX} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{i.}\right)^2,$$

8. La somme des carrés des erreurs pour Y : elle se calcule de la même manière que E_{XX} en remplaçant les X par des Y,

9. La somme des produits des erreurs X et Y :

$$E_{XY} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{i.}\right)\left(Y_{ij} - \bar{Y}_{i.}\right).$$

Le calcul de ces sommes correspond à une **analyse de variance** effectuée pour X, pour Y et pour XY.
Les degrés de liberté associés à ces différentes sommes sont les suivants :

1. Pour les sommes totales : $\sum_{i=1}^{t} n_i - 1$,
2. Pour les sommes des traitements : $t - 1$,
3. Pour les sommes des erreurs : $\sum_{i=1}^{t} n_i - t$.

L'ajustement de la **variable** Y à la variable concomitante X donne deux nouvelles sommes de carrés :

1. La somme des carrés totale ajustée :

$$SC_{tot} = S_{YY} - \frac{S_{XY}^2}{S_{XX}},$$

2. La somme des carrés des erreurs ajustée :

$$SC_{err} = E_{YY} - \frac{E_{XY}^2}{E_{XX}}.$$

Les nouveaux degrés de liberté pour ces deux sommes sont :

1. $\sum_{i=1}^{t} n_i - 2$
2. $\sum_{i=1}^{t} n_i - t - 1$ a degré de liberté ayant été soustrait pour l'ajustement.

La troisième somme des carrés ajustée, à savoir la somme des carrés des traitements ajustée est donnée par :

$$SC_{tr} = SC_{tot} - SC_{err}.$$

Elle garde le même nombre de degrés de liberté $t - 1$.

Tableau d'analyse de covariance

Nous avons maintenant tous les éléments pour construire le tableau d'analyse de covariance, les moyennes des carrés étant obtenues par la division des sommes des carrés par les degrés de liberté.

Source de variation	Degrés de liberté	Somme des carrés et des produits $\sum_{i=1}^{t} x_i^2$	$\sum_{i=1}^{t} x_i y_i$	$\sum_{i=1}^{t} y_i^2$
Traitements	t - 1	T_{XX}	T_{XY}	T_{YY}
Erreurs	$\sum_{i=1}^{t} n_i - t$	E_{XX}	E_{XY}	E_{YY}
Total	$\sum_{i=1}^{t} n_i - 1$	S_{XX}	S_{XY}	S_{YY}

Remarque : les nombres dans les colonnes des $\sum_{i=1}^{t} x_i^2$ et des $\sum_{i=1}^{t} y_i^2$ ne peuvent pas être négatifs ; en revanche, les nombres dans la colonne des $\sum_{i=1}^{t} x_i y_i$ peuvent être négatifs.

Ajustement

Source de variation	Degrés de liberté	Somme des carrés	Moyenne des carrés
Traitements	t - 1	SC_{tr}	MC_{tr}
Erreurs	$\sum_{i=1}^{t} n_i - t - 1$	SC_{err}	MC_{err}
Total	$\sum_{i=1}^{t} n_i - 2$	SC_{tot}	

Ratio F Test concernant les traitements

Le ratio F (qui va servir à tester l'**hypothèse nulle** qu'il n'y a pas de différence significative entre les moyennes des traitements une fois la variable Y ajustée) est donné par :

$$F = \frac{MC_{tr}}{MC_{err}}.$$

Le ratio suit une **loi de Fisher** avec $t - 1$ et $\sum_{i=1}^{t} n_i - t - 1$ degrés de liberté.

L'hypothèse nulle :

$$H_0 \quad : \quad \tau_1 = \tau_2 = \ldots = \tau_t$$

sera alors rejetée au **seuil de signification** α si le ratio F est supérieur ou égal à la valeur de la **table de Fisher**, c'est-à-dire si :

$$F \geq F_{t-1,\sum_{i=1}^{t} n_i - t - 1,\alpha}.$$

Il est clair que, pour faire une analyse de covariance, le coefficient β est supposé différent de zéro. Si tel n'était pas le cas, une simple analyse de variance suffirait.

Test concernant la pente β

Nous désirons ainsi savoir s'il y avait une différence significative entre les variables concomitantes avant l'application du traitement.

Pour tester cette **hypothèse**, nous formulons l'hypothèse nulle :

$$H_0 \quad : \quad \beta = 0$$

et l'**hypothèse alternative** :

$$H_1 \quad : \quad \beta \neq 0.$$

Alors le ratio F :

$$F = \frac{E_{XY}^2 / E_{XX}}{MC_{err}}$$

suit une loi de Fisher avec 1 et $\sum_{i=1}^{t} n_i - t - 1$ degrés de liberté. L'hypothèse nulle sera rejetée au seuil de signification α si le ratio F est supérieur ou égal à la valeur de la table de Fisher, c'est-à-dire si :

$$F \geq F_{1,\sum_{i=1}^{t} n_i - t - 1,\alpha}.$$

DOMAINES ET LIMITATIONS

Les hypothèses de base qu'il faut respecter avant de faire une analyse de covariance sont les mêmes que pour l'**analyse de variance** et l'**analyse de régression**. Ce sont des hypothèses de normalité, d'homogénéité (homoscédasticité) de variances et d'indépendance des erreurs du modèle.

En analyse de covariance, comme en analyse de variance, l'**hypothèse nulle** stipule que les échantillons indépendants proviennent de différentes populations dont les moyennes sont identiques.

D'autre part, comme à toute technique statistique, un certain nombre de conditions d'application est associé à l'analyse de covariance :

1. Les distributions des populations doivent être approximativement normales, sinon normales.
2. Les populations d'où sont prélevés les échantillons doivent posséder la même variance σ^2, c'est-à-dire :
   $$\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$
   où k est le nombre de populations à comparer.
3. Les échantillons doivent être choisis aléatoirement et tous les échantillons doivent être indépendants.

   À cela s'ajoute une hypothèse de base spécifique à l'analyse de covariance, à savoir que les traitements effectués ne doivent pas influencer les valeurs de la **variable** concomitante X.

EXEMPLE

Considérons l'**expérience** qui consiste à comparer les effets de trois méthodes de nutrition sur une **population** de porcs.

Les données se présentent sous la forme d'un tableau comprenant trois méthodes de nutrition, chacune soumise à 5 porcs. Les poids initiaux sont notés par la variable concomitante X (en kg), et les gains de poids (après **traitement**) sont notés par Y :

Méthodes de nutrition

1		2		3	
X	Y	X	Y	X	Y
32	167	26	182	36	158
29	172	33	171	34	191
22	132	22	173	37	140
23	158	28	163	37	192
35	169	22	182	32	162

Calculons les moyennes $\bar{X}_{i.}, \bar{Y}_{i.}, \bar{X}_{..}$ et $\bar{Y}_{..}$:

$$\begin{aligned}
\bar{X}_{1.} &= \frac{32+29+22+23+35}{5} = 28,2 \\
\bar{X}_{2.} &= \frac{26+33+22+28+22}{5} = 26,2 \\
\bar{X}_{3.} &= \frac{36+34+37+37+32}{5} = 35,2 \\
\bar{X}_{..} &= \frac{1}{15}(32+...+35+26+...+22 \\
&\quad +36+...+32) \\
&= 29,87 \\
\bar{Y}_{1.} &= \frac{167+172+132+158+169}{5} = 159,6 \\
\bar{Y}_{2.} &= \frac{182+171+173+163+182}{5} = 174,2 \\
\bar{Y}_{3.} &= \frac{158+191+140+192+162}{5} = 168,6 \\
\bar{Y}_{..} &= \frac{1}{15}(167+...+169+182+...+182 \\
&\quad +158+...+162) \\
&= 167,47.
\end{aligned}$$

Nous calculons les différentes sommes des carrés et des produits :

1. La somme des carrés totale pour X :

$$\begin{aligned}
S_{XX} &= \sum_{i=1}^{3}\sum_{j=1}^{5}\left(X_{ij}-\bar{X}_{..}\right)^2 \\
&= (32-29,87)^2+\ldots \\
&\quad +(32-29,87)^2 \\
&= 453,73.
\end{aligned}$$

2. La somme des carrés totale pour Y :

$$\begin{aligned}
S_{YY} &= \sum_{i=1}^{3}\sum_{j=1}^{5}\left(Y_{ij}-\bar{Y}_{..}\right)^2 \\
&= (167-167,47)^2+\ldots \\
&\quad +(162-167,47)^2 \\
&= 3\,885,73.
\end{aligned}$$

3. La somme des produits totale de X et Y :

$$\begin{aligned}
S_{XY} &= \sum_{i=1}^{3}\sum_{j=1}^{5}\left(X_{ij}-\bar{X}_{..}\right)\left(Y_{ij}-\bar{Y}_{..}\right) \\
&= (32-29,87)\,(167-167,47)+ \\
&\quad \ldots+(32-29,87)\,(162-167,47) \\
&= 158,93.
\end{aligned}$$

4. La somme des carrés des traitements pour X :

$$\begin{aligned}
T_{XX} &= \sum_{i=1}^{3}\sum_{j=1}^{5}\left(\bar{X}_{i.}-\bar{X}_{..}\right)^2 \\
&= 5\,(28,2-29,87)^2 \\
&\quad +5\,(26,2-29,87)^2 \\
&\quad +5\,(35,2-29,87)^2 \\
&= 223,33.
\end{aligned}$$

5. La somme des carrés des traitements pour Y :

$$\begin{aligned}
T_{YY} &= \sum_{i=1}^{3}\sum_{j=1}^{5}\left(\bar{Y}_{i.}-\bar{Y}_{..}\right)^2 \\
&= 5\,(159,6-167,47)^2 \\
&\quad +5\,(174,2-167,47)^2 \\
&\quad +5\,(168,6-167,47)^2 \\
&= 542,53.
\end{aligned}$$

6. La somme des produits des traitements de X et Y :

$$\begin{aligned}
T_{XY} &= \sum_{i=1}^{3}\sum_{j=1}^{5}\left(\bar{X}_{i.}-\bar{X}_{..}\right)\left(\bar{Y}_{i.}-\bar{Y}_{..}\right) \\
&= 5\,(28,2-29,87)\,(159,6-167,47) \\
&\quad +\ldots+5\,(35,2-29,87)\,(168,6-167,47) \\
&= -27,67.
\end{aligned}$$

7. La somme des carrés des erreurs pour X :

$$\begin{aligned}
E_{XX} &= \sum_{i=1}^{3}\sum_{j=1}^{5}\left(X_{ij}-\bar{X}_{i.}\right)^2 \\
&= (32-28,2)^2+\ldots+(32-35,2)^2 \\
&= 230,40.
\end{aligned}$$

8. La somme des carrés des erreurs pour Y :

$$E_{YY} = \sum_{i=1}^{3}\sum_{j=1}^{5}\left(Y_{ij}-\bar{Y}_{i.}\right)^2$$

$$= (167 - 159,6)^2 + \ldots + (162 - 168,6)^2$$
$$= 3\,343,20.$$

9. La somme des produits des erreurs X et Y :

$$\begin{aligned} E_{XY} &= \sum_{i=1}^{3}\sum_{j=1}^{5}\left(X_{ij} - \bar{X}_{i.}\right)\left(Y_{ij} - \bar{Y}_{i.}\right) \\ &= (32 - 28,2)\,(167 - 159,6) + \ldots \\ &\quad + (32 - 35,2)\,(162 - 168,6) \\ &= 186,60. \end{aligned}$$

Les degrés de liberté associés à ces différentes sommes sont les suivants :

1. Pour les sommes totales :

$$\sum_{i=1}^{3} n_i - 1 = 15 - 1 = 14.$$

2. Pour les sommes des traitements :

$$t - 1 = 3 - 1 = 2.$$

3. Pour les sommes des erreurs :

$$\sum_{i=1}^{3} n_i - t = 15 - 3 = 12.$$

L'ajustement de la **variable** Y à la variable concomitante X donne deux nouvelles sommes des carrés :

1. La somme des carrés totale ajustée :

$$\begin{aligned} SC_{tot} &= S_{YY} - \frac{S_{XY}^2}{S_{XX}} \\ &= 3\,885,73 - \frac{158,93^2}{453,73} \\ &= 3\,830,06. \end{aligned}$$

2. La somme des carrés des erreurs ajustée :

$$\begin{aligned} SC_{err} &= E_{YY} - \frac{E_{XY}^2}{E_{XX}} \\ &= 3\,343,20 - \frac{186,60^2}{230,40} \\ &= 3\,192,07. \end{aligned}$$

Les nouveaux degrés de liberté pour ces deux sommes sont :

1. $\sum_{i=1}^{3} n_i - 2 = 15 - 2 = 13$
2. $\sum_{i=1}^{3} n_i - t - 1 = 15 - 3 - 1 = 11$

La somme des carrés des traitements ajustée est donnée par :

$$\begin{aligned} SC_{tr} &= SC_{tot} - SC_{err} \\ &= 3\,830,06 - 3\,192,07 \\ &= 637,99. \end{aligned}$$

Elle garde le même nombre de degrés de liberté :

$$t - 1 = 3 - 1 = 2.$$

Nous avons maintenant tous les éléments pour construire le tableau d'analyse de covariance, les moyennes des carrés étant obtenues par la division des sommes des carrés par les degrés de liberté.

Source de variation	Degrés de liberté	Somme des carrés et des produits $\sum_{i=1}^{3} x_i^2$	$\sum_{i=1}^{3} x_i y_i$	$\sum_{i=1}^{3} y_i^2$
Traitements	2	223,33	−27,67	543,53
Erreurs	12	230,40	186,60	3 343,20
Total	14	453,73	158,93	3 885,73

Ajustement

Source de variation	Degrés de liberté	Somme des carrés	Moyenne des carrés
Traitements	2	637,99	318,995
Erreurs	11	3 192,07	290,188
Total	13	3 830,06	

Le ratio F (qui va servir à tester l'**hypothèse nulle** qu'il n'y a pas de différence significative entre les moyennes des traitements une fois la variable Y ajustée) est donné par :

$$F = \frac{MC_{tr}}{MC_{err}} = \frac{318,995}{290,188} = 1,099.$$

Si nous choisissons un **seuil de signification** $\alpha = 0,05$, la valeur de F dans la **table de Fisher** est égale à :

$$F_{2,11,0,05} = 3,98.$$

Comme $F < F_{2,11,0,05}$, nous ne pouvons pas rejeter l'hypothèse nulle et concluons qu'il n'y a pas de différence significative entre les trois méthodes de nutrition, une fois la variable Y ajustée, aux poids initiaux X.

MOTS CLÉS

Analyse de régression *(Regression analysis)*
Analyse de variance *(Analysis of variance)*
Analyse de données manquantes *(Analysis of missing data)*
Plan d'expériences *(Design of experiments)*

RÉFÉRENCES

Bartlett, M.S. (1937). Some Examples of Statistical Methods of Research in Agriculture and Applied Biology. *Journal of the Royal Statistical Society*, (Suppl.), 4, pp. 137-183.

DeLury, D.B. (1948). The Analysis of Covariance. *Biometrics*, 4, pp. 153-170.

Fisher, R.A. (1935). *The Design of Experiments.* Oliver & Boyd, Edinburgh.

Huitema, B.E. (1980). *The analysis of covariance and alternatives.* John Wiley & Sons, New York.

Wildt, A.R. and Ahtola, O. (1978). *Analysis of Covariance.* Sage University Papers series on Quantitative Applications in the Social Sciences (12).

ANALYSE DE DONNÉES

Data analysis

De nombreuses activités scientifiques commencent par un recueil de données, qu'on peut généralement classer dans un tableau à double entrée de grande taille. L'analyse de données vise à permettre à l'utilisateur de ces tableaux d'extraire facilement le maximum d'informations qui lui sont nécessaires.

On peut considérer l'analyse de données au sens large, comme l'essence même des statistiques à laquelle tous les autres aspects sont subordonnés.

HISTORIQUE

Dans la mesure où l'analyse de données regroupe des méthodes très nombreuses et très différentes d'analyse statistique, il est difficile de rapporter un historique précis. On peut cependant mentionner les précurseurs de ces différents aspects :

- la première publication d'analyse de données exploratoire date de 1970-1971, elle est due à **John Wilder Tukey** ; il s'agit d'une première version de son ouvrage paru en 1977 ;
- les principes théoriques de l'**analyse factorielle des correspondances** sont dus à Hermann Otto Hartley (1935) (publiés sous son nom original allemand Hirschfeld) et à **Ronald Aylmer Fisher** (1940). Cependant elle ne fut développée que dans les années 1970 par Jean-Paul Benzécri ;
- les premiers travaux de **classification** furent effectués en biologie et en zoologie. La plus ancienne typologie est due à Galen (129-199), et précède de loin les recherches de Linné (XVIII^{e} siècle). Les méthodes de classifications numériques sont basées sur les idées d'Adanson (XVIII^{e} siècle) et ont été développées entre autres par Joseph Zubin (1938) et Robert L. Thorndike (1953).

DOMAINES ET LIMITATIONS

L'analyse de données regroupe des méthodes très nombreuses et très différentes d'analyse statistique.

L'analyse de données peut se décomposer de la façon suivante :

1. *L'analyse exploratoire de données*, qui consiste comme son nom l'indique à explorer les données, ce qui se résume à :
   - la **représentation graphique** des données ;

- la **transformation**, si nécessaire, des données ;
- la détection d'éventuelles observations aberrantes ;
- l'élaboration d'hypothèses de recherches imprévues au début de l'**expérience** ;
- l'**estimation robuste**.

2. *L'analyse de données initiale* traite :
   - du choix des méthodes statistiques à appliquer aux données.
3. *L'analyse multivariée* des données comprend :
   - le déploiement d'espaces multidimensionnels,
   - la transformation des données pour réduire les dimensions et faciliter l'interprétation ;
   - la recherche de structure.
4. *L'analyse de données* (lorsqu'on parle d'un aspect particulier de cette dernière) inclut :
   - l'**analyse factorielle des correspondances** ;
   - l'analyse des correspondances multiples ;
   - la **classification**.
5. *L'analyse des données confirmatoire* comprend :
   - l'**estimation** de paramètres ;
   - les tests d'hypothèses ;
   - la généralisation et la conclusion.

L'analyse de données est caractérisée par quelques idées de base, à savoir :

- utiliser de façon massive et efficace l'informatique ; (visualisation des résultats avec des moyens d'interprétation clairs et précis) ;
- rester proche des données ;
- réduire au maximum la subjectivité.

MOTS CLÉS

Analyse exploratoire de données *(Exploratory data analysis)*
Analyse factorielle des correspondances *(Correspondence analysis)*
Classification *(Classification)*
Classification automatique *(Cluster analysis)*
Donnée *(Data)*
Statistique *(Statistics)*
Test d'hypothèse *(Hypothesis testing)*
Transformation *(Transformation)*

RÉFÉRENCES

Benzécri, J.-P. (1976). *L'Analyse des données*, Tome 1 : *La Taxinomie* ; Tome 2 : *L'Analyse factorielle des correspondances*. Dunod, Paris.

Benzécri, J.-P. (1976). Histoire et préhistoire de l'analyse des données, *Les Cahiers de l'analyse des données 1*, n$^{os.}$ 1–4. Dunod, Paris.

Everitt, B.S. (1974). *Cluster Analysis*. Halstead Press, London.

Fisher, R.A. (1940). The precision of discriminant Functions. *Annals of Eugenics* (London), 10, pp. 422-429.

Gower, J.C. (1988). *Classification, Geometry and Data Analysis in Classification and Related Methods of Data Analysis*, Bock H.H. (editor). Elsevier Science Publishers B.V. (North-Holland).

Hirschfeld, H.O. (1935). A connection between correlation and contingency, *Proceedings of the Cambridge Philosophical Society*, 31, pp. 520-524.

Thorndike, R.L. (1953). Who belongs in a family ? *Psychometrika*, 18, pp. 267-276.

Tukey, J.W. (1970-1971). *Exploratory data analysis*, limited preliminary edition. Addison-Wesley, Reading, Mass.

Tukey, J.W. (1977). *Exploratory data analysis*. Addison-Wesley, Reading, Mass.

Zubin, J. (1938). A Technique for Measuring Likemindedness. *Journal of Abnormal and Social Psychology*, 33, pp. 508-516.