PUBLICATION DE LA COMMISSION INTER-IREM
ENSEIGNEMENT DE LA STATISTIQUE ET DES PROBABILITES

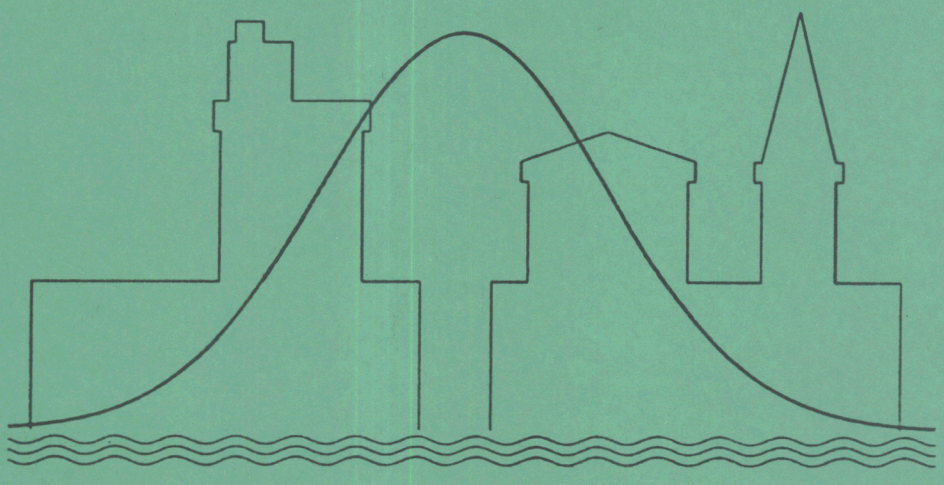
ACTES

DE

L'UNIVERSITE D'ETE

DE

STATISTIQUE



LA ROCHELLE 1 - 5 SEPTEMBRE 1992

édité par PICHARD J.F.
IREM de ROUEN

PUBLICATION DE LA COMMISSION INTER-IREM
ENSEIGNEMENT DE LA STATISTIQUE ET DES PROBABILITES

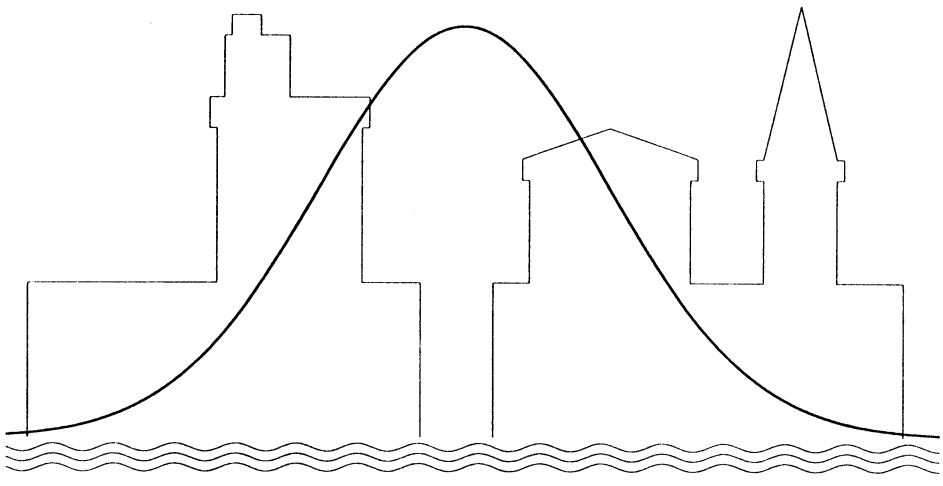
ACTES

DE

L'UNIVERSITE D'ETE

DE

STATISTIQUE



LA ROCHELLE 1 - 5 SEPTEMBRE 1992

édité par PICHARD J.F.
IREM de ROUEN

SOMMAIRE

PRESENTATION

L'Université d'été de Statistique intitulée "Statistique dans les formations technologiques", qui s'est déroulée à l'IUT de La Rochelle du 1 au 5 septembre 1992, a été organisée par la commission inter-IREM "Enseignement de la Statistique et des Probabilités".

La création récente de cette commission (juin 1990) répondait à un besoin ressenti par beaucoup de professeurs de mathématiques, que ce soit en collège ou en lycée : difficulté d'enseigner la statistique alors qu'on n'a pas eu de formation dans cette discipline, voire même en théorie des probabilités, dans son cursus universitaire.

Ce besoin est encore plus fortement ressenti au niveau post-bac où les méthodes de raisonnement de la statistique inférentielle sont très différentes de celles des mathématiques.

Les objectifs préalables que l'on avait fixés pour cette Université d'été étaient :

- accroître les compétences des enseignants en leur donnant des compléments de formation initiale et en indiquant les principales méthodes de la Statistique pouvant être utilisées dans les premières années de formation post-bac : BTS et autres formations technologiques,
- favoriser l'apport de l'informatique en tant qu'outil pour effectuer des traitements statistiques,
- étudier l'utilisation de logiciels statistiques comme support didactique pour l'enseignement de la Statistique,
- confronter des expériences d'enseignement de la Statistique,
- diffuser la recherche en Statistique par l'intervention d'universitaires et établir des liens entre recherche à l'Université et enseignement dans les formations technologiques, lesquels sont actuellement très faibles,
- participer à la formation de formateurs et personnes ressources dans les académies sur le thème de la Statistique.

Ces objectifs - certainement trop ambitieux pour un stage de 5 jours - n'ont peut-être pas été atteints. Néanmoins, cette session a été fructueuse, permettant, outre un apport théorique, de replacer les notions enseignées dans un cadre plus large et d'apporter un éclairage nouveau sur la statistique, selon l'opinion d'un participant.

De plus, la participation de stagiaires de l'Education Nationale et de l'Enseignement Agricole a semblé intéressante à tous, permettant la comparaison de pratiques pédagogiques légèrement différentes.

Cette Université d'été, qui était la première sur le domaine de la Statistique, a permis la discussion des participants sur les difficultés rencontrées pour l'enseignement de la statistique et a mis à jour le besoin d'un développement de la didactique de cette discipline et sa diffusion lors d'universités d'été ultérieures.

En écho et en prolongement de cette session de formation, j'ai rassemblé les textes des exposés et les comptes rendus des ateliers faits à l'Université d'été de Statistique, pour faire la publication de ces Actes.

Selon la demande de nombreux participants qui souhaitaient avoir, dans les meilleurs délais, un document de travail sur tout ce qui s'était déroulé à ce stage et auquel ils n'avaient pu assister, les ateliers se tenant en parallèle, j'ai réuni et publié ces textes, certains dans leur forme initiale, le plus rapidement possible. Aussi ce recueil se présente-t-il dans une mise en page un peu hétéroclite, et les textes eux-mêmes n'ont peut-être pas la meilleure rédaction souhaitable.

Ce document est complété par une bibliographie commentée d'ouvrages de statistique et une liste des brochures récentes sur la statistique produites dans les IREM.

Tel qu'il est, j'espère que ce document sera un outil de travail profitable pour les lecteurs, et leurs critiques ou commentaires sont vivement souhaités.

Je tiens à remercier Madame Lamarche, secrétaire de l'IREM de ROUEN, qui a participé à la lourde tâche de l'organisation matérielle de cette université d'été et a assuré son suivi sur place à La Rochelle.

Le responsable pédagogique de l'université d'été,
responsable de la commission
Pichard Jean-François.

UNIVERSITE D'ETE DE STATISTIQUE - LA ROCHELLE

1 au 5 septembre 1992

Programme

mardi 1 sept.

8h30-9h30 : accueil
9h30-11h : conférence présentation, Piednoir (I.G.)
11h-11h15 : pause
11h15-12h30 : exposé : Quelques points d'histoire de la Statistique, Pichard (U. de Rouen)

14h-16h : conférence : Méthodes d'estimation, Henry (U. de Besançon)
16h-16h15 : pause
16h15-18h15 : atelier
18h30-19h30 : atelier libre : utilisation de logiciels

mercredi 2 sept.

8h45-10h45 : conférence : Tests d'hypothèse, Beninel (IUT Niort)
10h45-11h : pause
11h-12h30 : exposé : Sur la genèse d'un test, Gras (U. de Rennes)

14h-16h : atelier
16h-16h15 : pause
16h30-18h30 : atelier
18h30-19h30 : atelier libre : utilisation de logiciels

jeudi 3 sept.

8h30-10h30 : conférence : Méthodes bayésiennes, Cellier (U. de Rouen)
10h30-10h45 : pause
10h45-12h30 : conférence : Méthodes linéaires et régression, Foucart (U. d'Orléans)

a.m. : découverte du pays

vendredi 4 sept.

8h45-10h45 : conférence : Analyse de variance, Courcoux (ENITIAA-Nantes)
10h45-11h : pause
11h-12h30 : exposé : Méthodes linéaires et régression II, Foucart (U. d'Orléans)

14h-16h : atelier
16h-16h15 : pause
16h30-18h30 : atelier
18h30-19h30 : atelier libre : utilisation de logiciels

samedi 5 sept.

8h30-10h30 : atelier
10h30-10h45 : pause
10h45-12h15 : évaluation-bilan de l'université d'été

Liste des ateliers

mardi 1 sept.

tranche de 16h15 à 18h15

. Henry : estimation et intervalles de confiance (niveau 2)
. Labroue, Saint-Pierre : T.P. sur échantillonnage, estimation (niveau 1)
. Présentation de logiciels statistiques

mercredi 2 sept.

tranche de 14h à 16h

. Gras : test d'hypothèse, traitement d'exemples sur logiciel
. Fredon : régression, aspects déterministes (niveau 1)
. Mérigot : analyse de sujets d'examen de BTS (niveau 1)

tranche de 16h30 à 18h30

. Beninel : test d'hypothèse (niveau 2)
. Janvier : présentation d'un didacticiel sur l'ajustement linéaire

jeudi 5 sept. : pas d'atelier

vendredi 4 sept.

tranche de 14h à 16h

. Cellier : exemple d'application de méthodes bayésiennes (niveau 2)
. Courcoux : analyse de la variance (niveau 2)

tranche de 16h30 à 18h30

. Foucart : régression (niveau 2)
. Mérigot : analyse de sujets d'examen de BTS (niveau 1)

samedi 5 sept.

. Fredon : tests non paramétriques : une introduction (niveau 1)
. Saint-Pierre : application de la régression : loi de Weibull (niveau 1)
. Pavy : exemple d'utilisation d'un logiciel de statistique : STAT-ITCF (niveau 1)

Ateliers libres de 18h30 à 19h30 : Présentation de logiciel(s) de statistique

Pour chaque tranche horaire de 2h, il y a 2 ou 3 ateliers en parallèle.

La difficulté des ateliers est classée en 2 niveaux :
niveau 1 : notions de base ou exemple de T.P. pour les élèves
niveau 2 : approfondissement

METHODES ET MODELES STATISTIQUES

PIEDNOIR Jean-Louis
I.G.

LA STATISTIQUE EN SECTION DE TECHNICIEN SUPERIEUR

Un programme de statistique est enseigné dans certaines sections de techniciens supérieurs. De plus en plus, les représentants des professions dans les organes consultatifs de l'éducation nationale demandent une initiation aux méthodes statistiques qui sont utilisées de façon croissante dans des secteurs toujours plus nombreux de la vie économique. Les progrès techniques facilitent l'enregistrement de données et ont abaissé considérablement le coût et la difficulté de leur traitement. Le développement des conduites de qualité, les besoins de connaissances précises dans le secteur tertiaire poussent également à un usage accru des techniques existantes.

La statistique met souvent le professeur du second degré mal à l'aise. Il y a peu de temps qu'elle s'est infiltrée dans les programmes, y compris ceux des collèges. Dans les cursus de mathématiques des universités, elle est peu enseignée voire absente, sauf en option. Aussi l'initiation aux méthodes et aux modèles de la statistique est en général embryonnaire. Les morceaux de statistique enseignés dans les lycées en section B et D par exemple, ou en BTS, apparaissent de deux types. Premier type : des choses élémentaires, triviales, que l'on peut apprendre tout seul, qui se prolongent par des définitions d'indicateurs telle la moyenne, la médiane, qui paraissent arbitraires. Deuxième type : des procédures faisant appel aux modèles probabilistes et donc complexes ; on est alors réduit à enseigner des recettes. De toute façon, dans les deux cas les exercices prévus pour le contrôle des connaissances sont débiles.

Les élèves, par contre, apprécient souvent un enseignement de statistique. Il s'agit pour eux d'une branche nouvelle des mathématiques et ceux qui ont eu des difficultés dans la discipline ont l'impression de pouvoir réussir sur un domaine inexploré pour eux. De plus l'aspect nécessairement pluridisciplinaire de la statistique crée un intérêt supplémentaire. On choisit, pour illustrer un cours de statistique, des exemples issus d'autres disciplines. Cette démarche montre l'aspect opérationnel des mathématiques.

En classe de technicien supérieur, il ne s'agit pas de transformer des étudiants en statisticien mais de montrer la pertinence des méthodes statistiques pour atteindre une certaine connaissance sur des populations données. A partir de quelques situations courantes, on apprend une démarche, on sait interpréter un résultat. Cela permettra au futur technicien de pouvoir dialoguer avec des spécialistes quand il faut traiter des cas plus complexes.

L'objectif poursuivi par l'exposé ci-dessous est, à partir d'exemples variés sur lesquels des méthodes statistiques ont été réellement appliquées, de faire l'analyse des situations pour lesquelles la statistique peut être employée, puis de présenter quelques modèles statistiques. Bref, on observe les moeurs de la tribu des statisticiens pour rentrer après dans un mode de pensée. Ensuite, on fera quelques remarques sur la validité des méthodes statistiques pour terminer sur quelques considérations de nature plus pédagogique.

QUELQUES SITUATION CONCRETES

Exemple 1 – On se propose d'étudier le comportement des candidats dans les différents baccalauréats généraux en 1989. Pour cela, on recueille pour chaque candidat l'ensemble des notes qu'il a obtenu aux différentes épreuves. Que peut-on tirer de cette masse de données pour tenir un discours sur les caractéristiques de chaque baccalauréat.

Exemple 2 – Un historien veut étudier le vocabulaire des hommes politiques au début du siècle et voir comment l'appartenance politique influe sur le vocabulaire employé. Pour cela, il sélectionne un certain nombre de mots et pour chaque homme politique il compte dans ses discours le nombre de fois où ces mots sont employés et note son appartenance politique. Comment représenter les rapports réciproques entre l'étiquette politique et la fréquence des mots employés.

Exemple 3 – Sur un territoire donné, un archéologue dispose d'un lot de haches néolithiques. Elles sont probablement de type et d'origine diverses. Pour chacune d'entre elles, il note des caractéristiques de formes, de composition chimique de la pierre utilisée, les principales dimensions. A partir de ces données est-il possible de conclure à une certaine homogénéité de la population ou au contraire peut-on détecter des sous-populations qui pourraient ensuite orienter les recherches à venir ?

Exemple 4 – Un commerçant reçoit une caisse de cartouches à vendre. Il veut en connaître la qualité. Pour cela il en prélève au hasard quelques unes et les expérimente. Au vu des résultats obtenus sur cet échantillon va-t-il accepter ce lot ou au contraire le renvoyer à son fabriquant ?

Exemple 5 – Une machine automatique fabrique des pièces qui sont censées être toutes identiques. De temps à autre un opérateur en prélève quelques-unes, les mesure. Va-t-il déclarer, après avoir procédé à ces essais, que la machine fonctionne bien ?

Exemple 6 – Un lots d'équipements est supposé représentatif de toute la population des équipements de la même espèce. Pour chacun d'eux on mesure la durée de vie. Que peut-on dire, du point de vue de la fiabilité, sur les équipements en question ?

Exemple 7 – Un gaz est formé de molécules. Pour chacune d'entre elles on peut déterminer à un instant donné la position, le vecteur vitesse, on peut aussi connaître le nombre de chocs entre deux instants donnés. A partir de ces mesures est-il possible de caractériser le gaz en question ?

Exemple 8 – Un radar envoie des signaux électromagnétiques, il reçoit des échos ; certains sont ceux de la cible recherchée, si elle existe, les autres sont considérés comme parasites. Quelle procédure mettre en route pour avoir la meilleures détection possible sur l'écran sans faire apparaître les signaux parasites appelés bruit ?

Exemple 9 – Au vu d'un diagnostic classique, un médecin est capable d'assigner une probabilité pour que le malade étudié ait telle ou telle maladie. Il fait ensuite procéder à des analyses. Comment modifie-t-il les probabilités données précédemment pour tenir compte des résultats des analyses.

ANALYSE DES SITUATIONS ETUDIEES

On remarque qu'aucune des situations précédentes n'est justiciable d'un schéma de causalité simple. On effectue à chaque fois plusieurs mesures et elles sont toutes différentes même si les facteurs contrôlés sont constants. La statistique se nourrit de cette variabilité. Dès qu'elle existe on peut songer à utiliser des méthodes statistiques.

Dans les exemples étudiés on peut identifier une population, appelée $\mathcal{C}$, et des individus. Sur chaque individu on effectue une mesure. A partir de ces mesures, on cherche à dire quelque chose sur la population $\mathcal{C}$; ou en d'autres termes à effectuer une mesure sur la population. Là est la démarche fondamentale de la statistique. On peut commencer à la formaliser selon le schéma ci-dessous.

LA DEMARCHE STATISTIQUE

Soit $\mathcal{C}$ une population, $I_n = \{1,2,....,n\}$ un ensemble d'individus de $\mathcal{C}$. On a $I_n \subset \mathcal{C}$ ou $I_n = \mathcal{C}$.
On appellera X l'ensemble dans lequel les individus prennent leurs mesures et x la mesure ; $x : I_n \rightarrow X$.

Sur $\mathcal{C}$ on cherche une mesure. Soit A l'ensemble dans lequel cette mesure est prise. On cherche donc $y_{\mathcal{C}} \in A$, $y_{\mathcal{C}}$ sera calculé à partir des mesures sur les individus. Chercher une procédure statistique, c'est trouver une application $g_n : X^n \rightarrow A$.

L'art de la statistique, c'est donc de trouver la ou les bonnes applications g_n. Pour cela il faudra évidemment introduire d'autres éléments.

Au niveau du langage on parlera aussi de collectif au lieu de population, d'éléments du collectif au lieu d'individus. Cela est mieux adapté à certaines situations. On peut remarquer que cette structure est très courante dans les sciences. En sociologie on a une population et des individus. En biologie un être vivant est composé de cellules. En chimie un corps pur est fait de molécules, etc...

La démarche statistique existe dans la vie courante. D'une façon intuitive, tout commerçant fait de la statistique comme Monsieur JOURDAIN faisait de la prose, sans le savoir. C'est bien à partir de l'observation des ventes dans les semaines précédentes que celui-ci procède, produit par produit, à ses commandes. Cet exemple sera repris par la suite.

On peut, sur les exemples précédents, déterminer les ensembles I_n, X et A et donc l'application x.

Exemple 1 : Prenons cinq épreuves écrites, $E = \{A,B,C,D,E\}$ l'ensemble des baccalauréats généraux, $X = E \times \mathbf{R}^5$, A sera l'ensemble des discours possibles (cf. étude sur le baccalauréat).

Exemple 2 : Soit I avec Card I = p l'ensemble des formations politiques, si k est le nombre de mots choisis, $X = I \times \mathbf{N}^k$. Une proximité entre une formation et un mot peut être exprimé comme le nombre de répétitions d'un mot dans une formation politique donnée. Alors $A = (\mathbf{R}^+)^{pk}$.

Exemple 3 : On posera J_1 l'ensemble des caractères qualitatifs de forme, J_2 l'ensemble des caractères qualitatifs géologiques et on suppose que l'on procède à mesure, donc $X = J_1 \times J_2 \times \mathbf{R}^k$.

On cherche des sous-populations, A est donc l'ensemble des partitions de I_n.

Exemple 4 : Si la cartouche fonctionne on notera 1, sinon 0. p sera la proportion de bonnes cartouches dans le lot global. $X = \{0,1\}^n$, $A = [0,1]$.

etc...

LE COLLECTIF ET LES MESURES

Le collectif n'est pas n'importe quelle collection d'individus. Il doit être relatif à une réalité cohérente ainsi que l'illustre le contre-exemple suivant. On veut étudier l'efficacité d'un médicament sur une collection d'individus composée d'hommes et de femmes. On notera H l'ensemble des hommes, F celui des femmes, T l'ensemble des individus traités, $\overline{T}$ celui des individus non traités, G l'ensemble des individus guéris, $\overline{G}$ l'ensemble des individus qui restent malades. Sur la population globale, on note les effectifs suivants :

	G	$\overline{G}$
T	120	50
$\overline{T}$	200	100

Au vu de ces résultats le médicament paraît efficace. Effectuons le même travail sur les sous-populations hommes, femmes.

Hommes

	G	$\overline{G}$
T	8	10
$\overline{T}$	60	60

Femmes

	G	$\overline{G}$
T	112	40
$\overline{T}$	140	40

L'examen de ces résultats montre à l'évidence que sur chacune des sous-populations le médicament est dangereux. On voit immédiatement que les hommes et les femmes réagissent de façon très différente à la dite maladie et que la proportion de personnes ayant subi le traitement est très différente d'une sous-population à l'autre. Cet exemple montre a contrario que la définition d'un collectif et de ses différents sous-collectifs nécessite une réflexion a priori. Le "piège" précédent est un exemple de ce qui peut arriver si les définitions ne sont pas précises et les procédures (voir plus loin) non rigoureusement observées.

Simpson fournit dans son paradoxe un exemple analogue. Il s'agit d'étudier si, dans sa sélection, une université favorise les hommes au détriment des femmes. Les résultats globaux tendraient à le montrer.

Mais il existe une "variable cachée" : le département. En effet, on s'aperçoit que dans chaque département la sélection favorise les femmes. Dans le tableau ci-dessous le numérateur donne pour chaque sexe le nombre de candidats sélectionnés, le dénominateur le nombre total de candidats.

	Hommes	Femmes
Département A	512/825=0,621	89/108=0,824
Département B	22/373=0,059	24/341=0,070
Total	534/1198=0,446	113/449=0,252

La mesure sur le collectif déduite des mesures sur les individus est aussi appelée résumé statistique. D'un point de vue naïf, on peut dire que le résumé extrait l'information interne dans les mesures faites sur les individus et qui intéresse la population. Il en résulte que le résumé doit être peu sensible à la présence ou à l'absence d'un individu particulier donné. Il doit y avoir une certaine stabilité du résumé par rapport aux variations inter-individuelles. En effet, un collectif est un entité distincte des individus qui le composent, il est d'un autre ordre. Le commerçant évoqué plus haut le sait bien intuitivement. Une seule mauvaise journée ne l'amène pas à modifier ses commandes.

Pour trouver le bon résumé statistique, il faudrait, cela paraît évident :

1/ Savoir ce que l'on cherche sur le collectif,

2/ Savoir ce que l'on veut faire de la mesure sur le collectif,

3/ Trouver, compte tenu du but fixé, les mesures pertinentes sur le collectif.

Il existe des cas où ces conditions sont remplies. On sait dans l'exemple 7, à partir de mesures sur les molécules d'un gaz dans diverses conditions, démontrer la loi de Mariotte : pression × volume = constante × température, loi physique qui peut aussi s'observer directement. Mais cet exemple est limite. Il n'est pas possible en général de vérifier la validité de la procédure statistique directement. S'il en était toujours ainsi, la statistique serait une quasi curiosité scientifique.

Souvent, le point 1 précédent ne peut être explicité complètement ; c'est le cas de l'exemple 1. Alors, on applique à ces situations des procédures statistiques connues, on examine les résultats. C'est l'intelligibilité de ces résultats par rapport à la situation de départ qui valide la procédure. La validation est alors a posteriori et non a priori par le modèle de départ.

En résumé, on peut dire que la statistique veut passer des éléments au collectif en éliminant les aspects individuels sans éliminer les individus.

L'ELABORATION DES RESUMES

Pour élaborer des résumés, il est nécessaire d'avoir sur le phénomène étudié des idées a priori, une connaissance sur la façon dont les données ont été recueillies. Sur les exemples précédents on va illustrer le propos.

Dans l'exemple 1 : le collectif des candidats au baccalauréat est divisé en 5 sous-populations suivant les baccalauréats. Pour la description statistique des résultats obtenus, on fera l'hypothèse que la mesure sur chaque individu est un vecteur de l'espace euclidien à 5 dimensions.

Dans l'exemple 2 : on appellera q_{ij} la fréquence d'emploi du mot j par les individus ayant l'appartenance i ($\sum q_{ij} = 1, \forall i$) et r_{ij} la fréquence des individus d'appartenance i pour l'emploi du mot j ($\sum r_{ij} = 1, \forall j$).

Alors, les vecteurs $(q_{i1}, q_{i2},...,q_{ik})$, $i=\{1,2,...,p\}$ appartiennent au simplexe δ_k , les vecteurs $(r_{1j}, r_{2j},...,r_{pj})$, $j=\{1,2,...,k\}$ au simplexe δ_p. On suppose que la distance dite du χ^2 sur ces simplexes est adéquate pour mesurer l'écart entre deux vecteurs.

Dans l'exemple 3 : on construit sur $J_1 \times J_2 \times \mathbf{R}^k$ une distance, ou au moins un indice de dissimilarité, et on s'efforcera de mettre dans la même classe les individus proches les uns des autres.

Dans l'exemple 4 : on supposera que le procédé d'obtention des individus de l'échantillon testé est le suivant : il y a tirage au hasard des cartouches essayées et chaque cartouche a la même probabilité de figurer dans l'échantillon.

Dans l'exemple 5 : on supposera que la mesure faite sur chaque pièce est le résultat d'une variable aléatoire gaussienne de moyenne μ et d'écart-type σ. Toutes les variables aléatoires sont indépendantes.

L'exemple 6 est justifiable d'une modélisation analogue ; on supposera que la variable aléatoire durée de vie est celle d'un système qui ne vieillit pas, donc suit une loi exponentielle de paramètre λ.

Dans l'exemple 8 : on prélève un échantillon de n observations, réalisations de n variables aléatoires indépendantes admettant une densité g(.) qui est la loi des échos suspects. On a : $\forall\, x \in \mathbf{R},\ g(x) = f(x - \Delta)$.

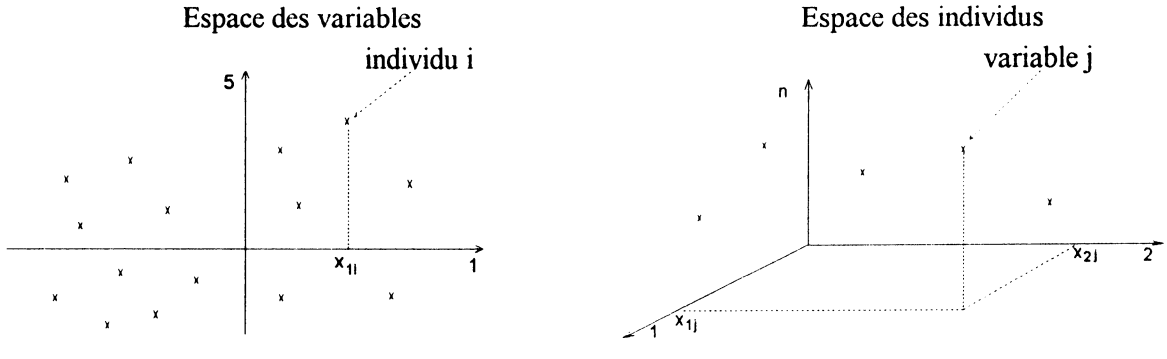
Si $\Delta = 0$ l'écho suspect n'est que du bruit, si $\Delta > 0$ il y a cible.

C'est à partir de ces hypothèses précises que l'on pourra élaborer des résumés statistiques. Leur pertinence dépend bien entendu de l'adéquation du modèle à la réalité.

On voit qu'il existe dans les exemples précédents deux grands types de modèles. Dans les exemples 1, 2, 3, la population étudiée forme tout le collectif et le modèle a priori est formalisé sous une forme géométrique, topologique ou algébrique. Les techniques mises en oeuvre forment la statistique descriptive dite aussi analyse des données. Dans les exemples 4, 5, 8, la population étudiée est un échantillon supposé extrait d'un collectif plus grand, réel ou mythique, souvent supposé infini et décrivable par une mesure de probabilité. Faire une mesure sur le collectif, c'est alors dire des choses sur la probabilité inconnue, en d'autres termes, mesurer cette probabilité. Les techniques mises en oeuvre forment la statistique inductive. C'est le jugement sur échantillon.

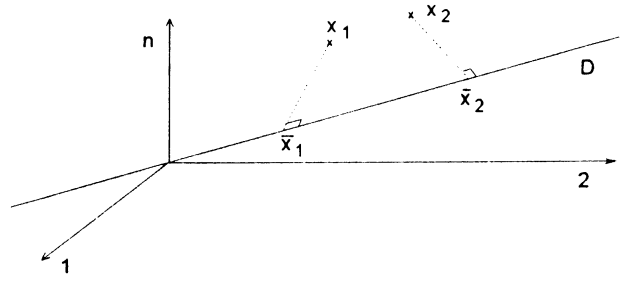
EXEMPLE DE STATISTIQUE DESCRIPTIVE

Il est hors de question de faire un panorama de l'ensemble des méthodes relevant de l'analyse des données. On se contentera des méthodes relevant du cadre euclidien. Prenons l'exemple 1. Chacun des n candidats d'un bac donné (C par exemple) est décrit par un vecteur à 5 composantes. On peut faire deux représentation duales : représenter les n individus dans l'espace E_5 des variables par un nuage de n points, ou représenter les 5 variables dans l'espace E_n des individus. Cela peut-être illustré par le schéma suivant.



A partir de ces représentations, il est possible de déterminer pour chaque variable des indicateurs de centralité, de dispersion, de chercher l'intensité de la liaison linéaire entre deux variables, d'avoir des résumés fidèles plus simples.

INDICATEUR DE CENTRALITE, DE DISPERSION, DE LIAISON LINEAIRE.



Prenons la représentation dans l'espace des individus. Il y a un cas où l'indicateur de centralité pour un ensemble de valeurs prises par une variable donnée s'impose de lui-même. C'est bien entendu le cas où toutes les observations sont identiques. On a $x_{11} = x_{12} = ... = x_{1n}$. L'ensemble des variables ayant cette propriété est la droite D n-sectrice des axes dont les coefficients directeurs sont tous égaux à 1. Si la variable X_1 n'est pas de ce type et si la description euclidienne est adéquate, on prendra comme indicateur de centralité celui de la variable qui a toutes ses coordonnées identiques et qui est la plus proche de X_1. Il suffit donc de projeter X_1 sur D. Il est facile de voir que le point $\bar{X}_1$ à toutes ses coordonnées égales à

$$\bar{x}_1 = \frac{1}{n} \sum x_{1i}$$

L'indicateur de dispersion suit aussi immédiatement du modèle euclidien : c'est la distance entre le point-variable X_1 et la droite D. On a alors

$$d^2(X_1, \bar{X}_1) = \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2$$

Pour pouvoir comparer des dispersions quand les tailles des populations diffèrent, on normalise par le facteur taille.

On a alors : $\sigma^2_{X_1} = \frac{1}{n} \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2$

S'il existe une liaison affine entre deux variables : $\forall$ i, $x_{1i} = a.x_{2i} + b$, les deux vecteurs $X_1 - \bar{X}_1$ et $X_2 - \bar{X}_2$ ont des supports parallèles. Si la liaison entre ces deux variables est proche d'une liaison affine, l'angle de ces deux variables sera proche de 0 (a>0) ou de l'angle plat (a<0). Il en résulte que l'intensité de la liaison linéaire peut être mesurer par l'angle entre les deux vecteurs ou par une de ses lignes trigonométriques.

On pose : $\rho = \cos (X_1 - \bar{X}_1, X_2 - \bar{X}_2)$. On a : $\rho = \frac{\sum (x_{1i} - x_1).(x_{2i} - x_2)}{n.\sigma_{x_1}.\sigma_{x_2}}$

ANALYSE EN COMPOSANTES PRINCIPALES

Il n'est pas facile de se représenter un espace à 5 dimensions. On va alors chercher une représentation du nuage des individus dans l'espace des variables, dans un espace à une ou deux dimensions et qui soit aussi fidèle que possible. Le cadre euclidien va permettre de répondre à la question. On va chercher le sous-espace affine à 1, 2 ou k dimensions le plus proche du nuage de points. Soit P_i le point représentatif de l'individu i, F_k un sous-espace affine à k dimensions, P'_i projection orthogonale de P_i sur F_k.

On cherche F_k tel que : $d^2(P_i, P'_i)$ soit minimum, d étant la distance euclidienne.

Un tel sous-espace passe par le point G dont les coordonnées sont les moyennes des différentes variables. Le plan minimum passe par la droite minimum. Une fois déterminé le sous-espace, on projette les points P_i sur F_k en P'_i. Le nuage des P'_i est alors le nuage le plus proche du nuage des P_i dans un espace affine à 1,2 ou k dimensions.

Le rapport : $R^2 = \frac{\sum GP'^2_i}{\sum GP^2_i}$ donne la qualité de la représentation ; il est dit rapport de l'inertie expliquée par la représentation à l'inertie totale du nuage.

Le langage utilisé rappelle la mécanique d'un système de points matériels. Si chaque point P_i est muni de la masse 1/n, on peut interpréter G comme le centre de gravité du système, F_1 comme étant l'axe principal d'inertie portant l'inertie maximum, de même (F_1, F_2) est le plan principal d'inertie portant la plus grande inertie. Chercher les composantes principales c'est déterminer les axes principaux de l'ellipsoïde d'inertie.

Une fois déterminé, ces différents axes d'inertie sont souvent interprétables en terme de caractéristiques du collectif. Ainsi dans l'exemple 1, le premier axe d'inertie est interprétable comme

représentant la performance globale des individus. Les candidats refusés ont sur cet axe une coordonnée petite, ceux qui ont la mention bien, une grande coordonnée.

Le deuxième axe oppose, au bac C, la réussite en mathématiques et celle en physique. Les candidats ayant une bonne note en maths et une mauvaise en physique ont une forte coordonnée sur ce deuxième axe. La conclusion est inversée pour ceux ayant une bonne note en physique et une mauvaise en mathématiques.

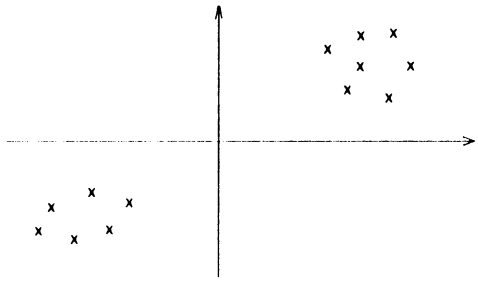
PERTINENCE DE LA REPRESENTATION

Tous les résumés présentés précédemment ne sont pertinents que si la représentation euclidienne est valide. Ainsi, par exemple, si les points représentatifs du nuage de points des individus ont, pour une variable, la présentation suivante :

———————X———————————————————XXXXXXXXXX

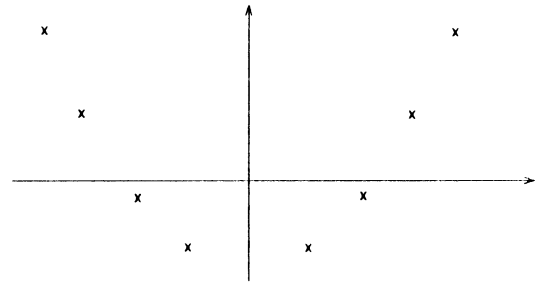
la moyenne ne peut être un bon indicateur de centralité. En effet, la présence ou l'absence de la plus petite observation influe considérablement sur la moyenne. Il n'y a plus stabilité du résumé par rapport aux variations individuelles. Le cadre euclidien n'est plus adéquat. Les fonctionnaires royaux du 18ème siècle qui voulaient résumer la capacité productive des provinces éliminaient la plus petite et la plus grande observation et faisait la moyenne tronquée. On élimine aussi des valeurs dites aberrantes.

De même, quand on étudie la liaison entre deux variables, un nuage de points tel que celui ci-dessous :



conduit à un coefficient de corrélation élevé qui ne traduit pas une liaison linéaire positive entre les deux variables. On a affaire à des données qui peuvent s'interpréter comme un mélange de deux populations distinctes.

Dans le cas suivant, on a une liaison fonctionnelle, mais un coefficient de corrélation nul. L'absence de liaison linéaire ne veut pas dire absence de liaison. L'indicateur choisi pour détecter la liaison est dans ce cas inadéquat.



STATISTIQUE INDUCTIVE

1. LE MODELE STATISTIQUE

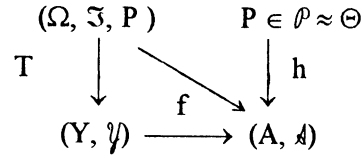
Faire de la statistique inductive, c'est définir sur le collectif étudié une mesure de probabilité, observer un échantillon de ce collectif et, à partir de cette observation, mesurer au moins partiellement la probabilité en question. On peut décrire mathématiquement cette suite d'opérations.

On appelle Ω l'ensemble de tous les échantillons a priori possibles ; Ω est l'espace fondamental. C'est un espace probabilisé, il est donc muni d'une tribu d'événements $\mathfrak{I} \subset \wp(\Omega)$ et d'une probabilité P. P étant inconnu, on a $P \in \mathcal{P}$ ensemble des lois de probabilité a priori possibles. $\mathcal{P}$ peut en général être mis en bijection avec un ensemble Θ appelé ensemble des paramètres. On veut connaître des choses sur Θ. Cela peut se formaliser en introduisant l'ensemble A des actions à mener et une application $h : \Theta \rightarrow A$. Pour la suite on supposera A probabilisable, donc muni d'une tribu d'événements $\mathcal{A}$.

Faire de la statistique c'est associer à un échantillon observé une action à mener. On cherche donc une application S, mesurable: $S : (\Omega, \mathfrak{I}) \rightarrow (A, \mathcal{A})$. S est aussi appelé une stratégie. S en général ne peut

déterminé directement mais à partir d'un résumé intermédiaire. Un résumé des données est une application mesurable $T : (\Omega, \mathfrak{I}) \to (Y, \mathcal{Y})$. T est appelée une statistique ($S = f_o T$).

On peut maintenant faire le schéma global de la statistique inductive.



Dans le cas fréquent où l'échantillon est interprétable comme la réalisation de n variables aléatoires indépendantes, on a alors :

$\Omega = X^n, \quad \mathfrak{I} = \mathcal{X}^{\otimes n}, \quad P = \Pi^{\otimes n}$

où $(X, \mathcal{X}, \Pi)$ est l'espace probabilisé formalisant une expérience. $(\Omega, \mathfrak{I}, P)$ est alors un espace produit.

Illustration du modèle

On reprend les exemples de l'introduction pour illustrer la formalisation proposée.

Exemple 4 - Pour le lot de cartouches supposé illimité, on pose : $X = \{0,1\}$, $\mathcal{X} = \wp(X)$. Soit p la proportion de bonnes cartouches. On a donc $\Omega = \{0, 1\}^n$ et si $k(\omega)$ est le nombre de bonnes cartouches dans l'échantillon, on écrit $P(\{\omega\}) = p^{k(\omega)}.(1-p)^{n-k(\omega)}$. On a $\Theta = [0, 1]$.

Si on veut connaître cette proportion, on pose : $A = [0, 1]$, et h est l'application identique.

Si on veut seulement savoir si $p \leq p_0$ ou $p > p_0$, on écrit $A = \{0, 1\}$ et $h(p) = 0$ si $p \leq p_0$, $h(p) = 1$ si $p > p_0$.

Exemple 5 : On s'intéresse au diamètre des pièces produites, on suppose que celui-ci suit une loi de Gauss de moyenne μ et d'écart type σ. La machine est considérée comme bien réglée si $\mu = \mu_0$. On a alors :

$X = \mathbf{R}, \quad \Omega = \mathbf{R}^n, \quad \Theta = \mathbf{R} \times \mathbf{R}^+, \quad A = \{0,1\}$

$$\frac{dP}{d\lambda_n}(x_1,\ldots,x_n) = \frac{1}{(2\pi)^{n/2}\,\sigma^n}\, e^{-\frac{1}{2}\sum\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

, où λ_n est la mesure de Lebesgue de $\mathbf{R}^n$.

$h(\mu, \sigma) = 0$ si $\mu = \mu_0$, $h(\mu, \sigma) = 1$ sinon.

Exemple 6 : la durée de vie d'un équipement est un nombre réel positif donc : $X = \mathbf{R}^+$, $\Omega = (\mathbf{R}^+)^n$.

Si cet équipement ne vieillit pas , on montre simplement que la loi est exponentielle, c'est-à-dire que l'on peut écrire :

$$P\left(\prod_{i=1}^{n} [x_i, +\infty[\right) = \prod_{i=1}^{n} \exp(-\lambda . x_i), \quad \lambda \in \mathbf{R}^+$$

Donc $\Theta = \mathbf{R}^+$, $1/\lambda$ est la durée moyenne de vie que l'on désire connaître. Il en résulte que $A = \mathbf{R}^+$ et h est l'application identique.

Exemple 8 : Dans les problèmes de détection - radar, échos de bruit et échos de cible peuvent être considérés comme des réalisations de variables aléatoires indépendantes de même loi pour chacun des deux types. La loi du bruit change suivant les conditions atmosphériques, celle de la cible se déduit de celle du bruit par une simple translation. On est amené alors à procéder comme suit. On prélève un échantillon de m observations de bruit et un échantillon de n observations de ce qui peut être une cible. Soit f la densité de probabilité du bruit et Δ l'intensité de l'écho cible. On a $\Omega = \mathbf{R}^{n+m}$, $\Theta = \mathbf{R}^+ \times \mathcal{J}$, où $\mathcal{J}$ est l'ensemble des densités de probabilité continues.

$$\frac{dP_\theta}{d\lambda}(x_1,\ldots,x_m,x_{m+1},\ldots,x_{m+n}) = \prod_{i=1}^{m} f(x_i) \cdot \prod_{i=m+1}^{m+n} f(x_i - \Delta) \quad A = \mathbf{R}^+, \quad h(\Delta, f) = \Delta.$$

CHOIX D'UNE STRATEGIE STATISTIQUE

Pour choisir une stratégie statistique il est nécessaire de se donner des critères de choix. On peut se restreindre à certaines classes d'applications, se donner un préordre sur l'ensemble des stratégies. Pour cela, il faut enrichir le modèle précédent :

Stratégie convergente : on connaît en probabilité la loi des grands nombres. Soit un événement E de probabilité $p > 0$, n répétitions indépendantes, f_n la fréquence empirique d'apparition de E. On peut démontrer : $\forall \varepsilon > 0$, $P_p(\{|f_n - p| > \varepsilon\}) \to 0$ quand $n \to \infty$, P_p étant la probabilité quand p est "l'état de la nature".

Ce théorème permet de dire que quand on réalise un grand nombre d'expériences indépendantes, on approche de très près la vraie loi de probabilité. A noter que pour exprimer la proximité entre probabilité et fréquence, on est obligé d'employer le concept de probabilité. D'où la difficulté de définir ce dernier comme limite de la fréquence empirique. Quoi qu'il en soit de cette discussion épistémologique, il est naturel de concevoir la probabilité comme limite de la fréquence empirique. Ce point de vue est celui des statisticiens fréquentistes.

Dans ces conditions, il est naturel d'exiger qu'une stratégie statistique soit d'abord convergente. Pour formaliser cela il faut plonger la situation particulière dans une suite de situations. Un problème statistique δ c'est la donnée de : $\forall n$, $\delta_n = (X^n, \Theta, A, h ; P_{n,\theta})$.

Supposons que A soit un espace métrique muni d'une distance d.

La stratégie $S_n : X^n \to A$ sera dite convergente si :

$\forall \varepsilon > 0, \forall \theta \in \Theta$, $P_{n,\theta}(\{\omega / d[S_n(\omega), h(\theta)] > \varepsilon\}) \to 0$, quand $n \to \infty$,

En terme intuitif cela se dit : la stratégie choisie approche la valeur cherchée quand n est grand.

Dans l'exemple des cartouches, si on prend comme stratégie $S_n(\omega) = k(\omega)/n$ (fréquence empirique des bonnes cartouches), on a bien, si p est la proportion de bonnes cartouches :

$\forall \varepsilon > 0$, $P_p(\{|S_n - p| > \varepsilon\}) \to 0$ quand $n \to \infty$.

C'est la loi des grands nombres. La distance sur [0,1] choisie est la distance classique sur **R**.

Fonction de coût : Pour introduire un préordre sur l'espace des stratégies on introduit une mesure de l'erreur faite en remplaçant ce que l'on cherche h (0) par la décision statistique s (ω).

Pour cela on introduit une fonction de coût : $L : \Theta \times A \to \mathbf{R}^+$

Le coût de la décision statistique est donc : $L(\theta, S_n(\omega))$ si θ est l'état de la nature, et ω le constat expérimental, S_n la stratégie choisie. Ce nombre dépend de ω donc du hasard, c'est le résultat d'une loterie. Depuis Pascal, pour comparer entre elles les loteries on compare leurs espérances mathématiques. On définit ainsi la fonction de risque de la stratégie S_n par :

$$R(\theta, S_n) = E_\theta(S_n) = \int_\Omega L(\theta, S_n(\omega)) dP_\theta(\omega)$$

Une stratégie S_n sera préférée à la stratégies S_n^* ($S_n \succ S_n^*$) si et seulement si

$\forall \theta \in \Theta$, $R(\theta, S_n) \leq R(\theta, S_n^*)$.

On a un préordre partiel sur l'espace des stratégies.

Prenons l'exemple 4 du lot de cartouches pour illustrer le calcul de la fonction de risque. On a $\Theta = A = [0, 1]$. Il est d'usage de prendre la fonction de coût quadratique $L(\theta, a) = (\theta - a)^2$.

La stratégie S_n qui consiste à estimer la proportion p par la fréquence empirique f_n a comme fonction de risque : $R(p, f_n) = E_p((f_n - p)^2) = Var(f_n) = p.(1-p) / n$.

La stratégie triviale qui consiste à décider p=1/2 sans regarder l'échantillon a pour fonction de risque $(p-1/2)^2$. On voit immédiatement que ces deux stratégies ne sont pas comparables.

Le préordre précédent ne permet donc pas de choisir entre ces deux stratégies. C'est pour cela que l'on est amener à restreindre l'ensemble des stratégies. Il est bien évident que la stratégie triviale n'est pas convergente, par exemple. On va introduire le cas de l'estimation qui est un autre concept.

CAS DE L'ESTIMATION

On dit que l'on a un problème d'estimation quand l'ensemble A est un espace vectoriel normé. Dans les cas classiques : $A = \mathbf{R}^k$. On dit alors que S_n estime $h(\theta)$, S_n est l'estimateur, $S_n(\omega)$ l'estimation déduite de l'échantillon.

Un critère possible de sélection d'une stratégie est le suivant. On veut qu'en moyenne les estimations différentes qui seraient possibles, si on pouvait répéter l'épreuve statistique, coïncident avec la chose à estimer. Cela peut se traduire par : $E_\theta(S_n) = h(\theta)$.

L'estimateur S_n est alors dit sans biais. Dans l'exemple des cartouches, on a bien : $E_p(F_n) = p$.

Dans le cas où $A = \mathbf{R}$, introduire la fonction de coût quadratique et se restreindre à des estimateurs sans biais revient à comparer les variances des estimateurs. Cette variance est la fonction de risque de l'estimateur sans biais.

Estimer un paramètre, c'est attribuer à ce paramètre une valeur que l'on espère proche de la valeur réelle. On est alors tenté d'indiquer la précision de cette approximation et donc de donner une borne supérieure à l'erreur commise. Mais comme on est dans un modèle probabiliste, celui-ci sera formalisé en termes probabilistes.

ESTIMATION PAR INTERVALLE

On se contentera d'étudier le cas où $A = \mathbf{R}$. Pour définir la précision on veut trouver un intervalle qui, avec une forte probabilité, recouvre la valeur cherchée. Pour cela, on se fixe un risque α et on cherchera un intervalle I tel que $P_\theta(\{\omega / I(\omega) \ni h(\theta)\}) = 1 - \alpha$.

Si $I = [a,b]$ cela peut s'écrire : $P_\theta(\{\omega / a(\omega) \leq h(\theta) \leq b(\omega)\}) = 1 - \alpha$.

Attention, c'est bien a et b qui sont aléatoires, $h(\theta)$ est inconnu mais certain. Il ne faut pas confondre aléatoire et inconnu. Cette situation d'estimation peut-être illustrée par l'exemple 6 sur les durées de vie. On veut estimer $1/\lambda$, durée moyenne de vie.

On appelle $X_1,...,X_n$ les variables aléatoires durées de vie des équipements à observer. On propose comme estimation la moyenne des durées de vie observées. On posera :

$\widehat{1/\lambda} = \frac{1}{n}(X_1 + X_2 + ... + X_n)$. On a : $E(\widehat{1/\lambda}) = 1/\lambda$ et $Var(\widehat{1/\lambda}) = 1/\lambda\sqrt{n}$.

On montre que $n\lambda\ \widehat{1/\lambda}$ suit la loi Γ_n, ce qui permet de construire l'intervalle de confiance.
Fixons $\alpha = 0{,}10$, on cherche a_n et b_n tels que $\Gamma_n(a_n) = 0{,}05$ et $1 - \Gamma_n(b_n) = 0{,}05$.

où : $\Gamma_{n+1}(x) = \int_0^x e^{-t}\frac{t^n}{n!}dt$

On a $P_\lambda(a_n < n\lambda.\widehat{1/\lambda} < b_n) = 0{,}90$ et $P_\lambda(n/b_n.\widehat{1/\lambda} < 1/\lambda < n/a_n.\widehat{1/\lambda}) = 0{,}90$.

On dit qu'avec une confiance de 90% la durée moyenne de vie est comprise entre $n/a_n\ \widehat{1/\lambda}(\omega)$ et $n/b_n\ \widehat{1/\lambda}(\omega)$.

La confiance n'est pas la probabilité. Ce terme rappelle que la procédure ayant abouti au constat fait devait réussir avec une probabilité 0,9.

PROBLEMATIQUE DES TESTS

Dans l'exemple 5 la décision à prendre est dichotomique. Doit-on oui ou non faire régler la machine. Dans un premier temps, il en est de même dans l'exemple 8 : y a-t-il oui ou non écho, si oui quelle est son intensité ? Quand l'ensemble des décisions à prendre ne comporte que deux éléments on peut écrire :

$A=\{0,1\}$.

L'ensemble Θ qui représente tous les "états de la nature" a priori possibles peut alors être dichotomisé $\Theta = \Theta_0 \cup \Theta_1$ avec $\Theta_0 \cap \Theta_1 = \varnothing$ et $\Theta_0 = h^{-1}(\{0\})$.

Décider 0 c'est dire que le paramètre θ qui régit le phénomène est un élément de Θ_0. Ainsi dans l'exemple 5, cela revient à affirmer que la moyenne du diamètre des pièces est μ_0. Quand décide-t-on 0 ? Quand la règle est l'application S, on pose : $C = S^{-1}(\{1\})$ qui est appelée région critique du test.

Si on examine les erreurs on se trouve devant le cas de figure suivant :

Décision du statisticien — Etat de nature

	Θ_0	Θ_1
0	Bien	Erreur 2
1	Erreur 1	Bien

Pour pouvoir introduire un préordre partiel sur les règles de décisions, il faut mesurer les erreurs.

L'erreur 1 peut être représentée par la fonction : $P_\theta(C)$, $\theta \in \Theta_0$. $P_\theta(C)$ est la probabilité de décider 1 quand l'état de la nature est θ, ici $\theta \in \Theta_0$.

L'erreur 2 peut être représentée par la fonction : $1 - P_\theta(C)$, $\theta \in \Theta_1$. $1 - P_\theta(C)$ est la probabilité de décider 0 quand l'état de la nature est θ, $\theta \in \Theta_1$.

Jusqu'à présent, les deux cas possibles jouent des rôles parfaitement symétriques. Mais dans les problèmes de décision dichotomique, il n'en est pas en général ainsi. L'une des hypothèses, disons $\theta \in \Theta_0$ est privilégiée. Elle correspond par exemple à une théorie scientifique et on ne l'abandonnera que si on a de très sérieuses raisons pour cela. Ou bien, comme dans l'exemple 5, rejeter l'hypothèse que la machine est bien réglée conduit à tout un processus qui coûte cher. On ne le mettra en route que si l'hypothèse choisie a priori est peu plausible au vu des résultats.

On traduit mathématiquement ce choix de l'hypothèse nulle de la façon suivante. On va fixer a priori une borne à l'erreur 1. On choisira une règle de décision S, donc une région C telle que

$$\sup_{\theta \in \Theta} P_\theta(C) \leq \alpha$$

Une fois l'erreur 1 choisie, un test sera meilleur qu'un autre si l'erreur 2 est plus petite pour le premier que pour le second et ce quel que soit $\theta \in \Theta_1$. On appelle puissance du test C la fonction : $\beta_C(\theta) = P_\theta(C)$.

Le test C_1 sera préféré à C_2 (C_1 â C_2) si, et seulement si $\forall\, \theta \in \Theta_1$, $\beta_{C_1}(\theta) \geq \beta_{C_2}(\theta)$.

Le choix de α dépend, lui, de la confiance que l'on a, a priori, dans l'hypothèse Θ_0, α sera d'autant plus petit que l'on a une confiance plus grande dans l'hypothèse Θ_0, ou bien que l'abandon de l'hypothèse Θ_0 conduit à des complications diverses et coûteuses. Ainsi dans les problèmes tels que ceux de l'exemple 5, on prend souvent α=0,01. Les psychologues, les biologistes choisissent traditionnellement α=0,05 ou 0,10.

Dans l'exemple 8 de détection de signaux radar, le nombre α a une interprétation concrète. Pendant un temps de surveillance donné, le nombre de fois où il apparaît un point sur l'écran alors qu'il n'y a pas de cible est proportionnel à α, ce que le radaristes appellent taux de fausse-alarme. De même la puissance du test est la probabilité de faire apparaître un point sur l'écran quand il y a cible. Accepter trop de faux échos, c'est prendre le risque de mobiliser pour rien des moyens importants et trop souvent. On limite donc ce taux et on prend souvent $\alpha = 10^{-5}$ ou 10^{-6}. Le réglage de ce seuil est d'ailleurs possible, sur les appareils existants, par l'utilisateur.

Un test sera dit sans biais s'il est toujours meilleur que la stratégie triviale qui consiste à décider sans regarder les observations. On a alors $\beta_C(\theta) \geq \alpha$, $\forall\, \theta \in \Theta_1$ (prendre comme stratégie triviale : je décide $\theta \in \Theta_1$ avec la probabilité α quel que soit l'échantillon).

CHOIX D'UNE PROCEDURE

Qu'il s'agisse de test, d'estimation ou d'autres problèmes statistiques, il faut ensuite mettre au point des procédures présentant un certain nombre de qualités prescrites a priori et ayant, au vu des critères développés ci-dessus, de bonnes performances.

C'est l'objet d'une partie de la statistique mathématique de faire ce travail. Il a été commencé au début du siècle par K. Pearson ; il s'est développé entre les deux guerres pour les modèles gaussiens avec Sir Ronald Fisher et son équipe, en particulier. Il se continue depuis la guerre en cherchant toujours plus de généralité ou en travaillant sur des modèles dont jusqu'à présent la complexité avait interdit la mise au point ou l'étude de procédures adéquates.

LA STATISTIQUE BAYESIENNE

Dans ce qui précède, on a présenté la probabilité comme une caractéristique objective du phénomène étudié dont on pouvait s'approcher en observant un grand nombre de répétitions. Dans l'optique Bayésienne une opinion sur les choses est une interprétation possible de la probabilité. On peut démontrer d'ailleurs qu'un système cohérent d'opinions peut se traduire par une distribution de probabilité.

Reprenons le schéma statistique. On a vu que l'ensemble $\mathcal{P}$ des lois de probabilité a priori possibles pouvaient être mis en correspondance biunivoque avec un ensemble Θ appelé ensemble des paramètres. On supposera que Θ est un espace probabilisable. On va munir cet espace d'une loi de probabilité Q qui exprime l'opinion a priori du statisticien (ou d'un spécialiste) sur les états de la nature. Il existe des cas où cette opinion a priori est le résultat d'une construction, reflète l'expérience du praticien. L'exemple 9 du diagnostic médical illustre cette situation. L'examen clinique permet au médecin de donner une probabilité a priori à chaque maladie possible.

On peut maintenant utiliser cette probabilité a priori. On supposera que toutes les probabilités admettent des densités par rapport aux mesures σ-finies μ sur $(\Omega,\mathfrak{I})$ et ν sur $(\Theta,\mathfrak{R})$; $q(\theta)$ est la densité de probabilité de Q par rapport à ν, $p(\omega,\theta)$ est la densité de probabilité de P_θ par rapport à μ.

Appliquons le théorème de Bayes. Dans le cas fini celui-ci peut s'énoncer ainsi : soit $(A_1,..,A_n)$ une partition de Ω, B un événement. En posant P(A/B) la probabilité de A sachant B, on peut facilement démontrer :

$$P(A_i/B) = P(B/A_i).P(A_i) \,/\, \Sigma\, P(B/A_i)\, P(A_i)$$

$P(A_i/B)$ s'interprète comme la probabilité de la "cause" A_i quand l'événement B se produit et $P(B/A_i)$ la probabilité de B quand "A_i agit" ; $P(A_i)$ la probabilité a priori de la cause A_i.

On appelle $q(\theta,\omega)$ la densité de la loi de probabilité a posteriori Q_ω sur Θ par rapport à la mesure ν. Le théorème de Bayes s'écrit :

$$q(\theta,\omega) = \frac{p(\omega,\theta).q(\theta)}{\int_\Theta p(\omega,\theta).q(\theta).\nu(d\theta)}$$

C'est à partir de cette densité a posteriori que s'élaborent les stratégies statistiques dites alors bayésiennes.

Quand on peut définir une fonction de coût $L : \Theta\times A \to \mathbf{R}$, $L(\theta,a)$ est le coût de l'action a quand l'état du système est θ. On introduit alors des risques a priori et a posteriori de l'action a.

Risque a priori $\tau_Q(a) = \int_\Theta L(\theta,a).q(\theta).\nu(d\theta)$ Risque a posteriori $\tau_Q(a,\omega) = \int_\Theta L(\theta,a).q(\theta,\omega).\nu(d\theta)$

La stratégie optimale quand l'opinion a priori est Q est celle qui minimise le risque a posteriori.

On peut écrire sous réserve d'existence $\tau_Q(S(\omega),\omega) = \min_{a\in A} \tau_Q(a,\omega)$.

Le point de vue Bayésien permet donc de trouver des stratégies optimales au sens bayésien. On montre que si $\forall\, \theta \in \Theta$, $q(\theta) > 0$, les stratégies bayésiennes sont convergentes au sens précédent.

QUELQUES REMARQUES POUR FINIR

L'exposé ci-dessus a voulu illustrer la démarche statistique par la modélisation de quelques situations particulières suivie des méthodes de traitement les plus usitées. Mais la statistique intervient également en amont. Etant donné un problème, comment mener les expériences pour avoir les renseignements souhaités, pour optimiser le nombre d'expériences suivant leur coût. En aval, il existe aussi des techniques destinées à

des conclusions dépend de l'adéquation du modèle à la réalité décrite et du mode de recueil des informations.

Du point de vue de l'enseignement, la statistique est une discipline très riche. A un certain niveau on y fait des mathématiques relativement sophistiquées. A un niveau plus modeste, on peut montrer un champ d'application des mathématiques très vaste. En outre, elle oblige à un travail interdisciplinaire au plein sens du terme.

Les données traitées sont toujours issues d'une autre discipline. La mise au point du modèle représentatif, l'interprétation des résultats supposent que le statisticien entre dans la problématique du spécialiste demandeur et que ce dernier soit capable de comprendre ce qu'est un phénomène aléatoire, une probabilité, une confiance. Sans cette compréhension au moins intuitive, il risque de faire des erreurs graves au niveau des conclusions qu'il tire pour sa discipline.

Si on veut intéresser les élèves à la statistique, il faut avant tout faire ressortir la démarche, son intérêt scientifique et pour cela utiliser de nombreux exemples. Le traitement numérique concret des données peut maintenant être considérablement allégé grâce à l'utilisation des logiciels ad'hoc, voire de calculatrices. On a donc en plus une application intéressante de l'informatique. Il est toujours intéressant d'aller jusqu'au bout des calculs afin de pouvoir faire l'interprétation des résultats...

La pratique montre que pédagogiquement l'étude de la statistique permet d'intéresser à la mathématique des élèves, des étudiants, qui jusque là, étaient rétifs à cette discipline. Son caractère fortement interdisciplinaire permet ce déblocage.

La statistique voit actuellement son champ d'application s'étendre considérablement. La demande d'enseignement ne peut donc que croître au niveau des formations à finalité professionnelle. De plus, pouvoir interpréter des données existantes est un acte que tout citoyen voulant se construire une opinion doit faire. Il n'y a qu'à voir le nombre de bêtises qu'écrivent les journalistes à propos des sondages ou de la correction des variations saisonnières, pour se convaincre de l'importance de la statistique aussi dans les mathématiques pour tous.

BREVE BIBLIOGRAPHIE

Cette bibliographie ne prétend pas à l'exhaustivité, il existe des ouvrages remarquables et adaptés aux besoins de ceux qui s'intéressent à l'enseignement de la statistique jusqu'au niveau Bac+2 et qui n'y figurent pas. Elle veut seulement guider le premier choix de ceux qui découvrent cette discipline.

– W. PETER Countring for something, statistical principal and personalities. Ed. Sprinper – Verlag
Ouvrage simple, donne une présentation historique simple de la statistique.

– A.P.M.E.P. Analyse de données (2 tomes)
Une présentation simple, pédagogique, et intéressante des méthodes de description de la statistique.

– ROZENGARD Probabilité et statistique en recherche sociale. Ed. Dunod.
Initiation aux probabilités et aux statistiques de niveau moyen. On utilise les mathématiques, mais de nombreux résultats sont admis.

– FOUCART, BENSABERT, GARNIER. Méthodes pratiques et de la statistique. Ed. Dunod.
Exposé très pratique des méthodes statistiques, mais les idées sous jacentes ne sont pas toujours très bien mises en valeur.

– J.J. DROESBEKE Eléments de statistique, collection SMA, Ed. Ellipses.
Bon compromis entre la statistique appliquée et le modèle mathématique, intéressant, complet.

– A. MONFORT Cours de probabilité. Ed. Economica.
Pour ceux qui veulent plonger dans le calcul des probabilités et lire un exposé complet de bon niveau écrit en utilisant le langage de la théorie de la mesure (le seul adapté).

– Ph. TASSI Méthodes statistiques. Ed. Economica.
Exposé classique de la statistique inférentielle. Nécessité de connaître un peu le calcul des probabilités.

– ANDERSON, SCLOVE. Statistical analysis data, Ed. The Scientific press.
Exposé des principales méthodes de statistiques descriptive et inférentielle avec de nombreux exemples, lecture agréable.

– Th WONACOTT et R. WONACOTT. Statistique (économie gestion, sciences humaines). Ed. Economica.
Présentation claire des méthodes statistiques (surtout pour la statistique inférentielle) ; livre à lire, de très nombreux exemples, une mine d'exercices possibles, mais les applications sont surtout orientées vers les sciences humaines et économiques.

– SARDADI, VINCZE. Mathematical methods of statistic quality control. Ed. Académic Press.
Initiation aux probabilités, à la statistique inférentielle avec des applications au contrôle de qualité essentiellement et un peu à la fiabilité, beaucoup d'exemples.

–INTER IREM TECHNIQUE. Fiabilité.
Toutes les procédures statistiques au programme des BTS maintenance et de quelques autres illustrées par des exemples issus des examens.

QUELQUES POINTS D'HISTOIRE DE LA STATISTIQUE

Les théorèmes limites du calcul des probabilités, fondement de la Statistique

Pichard Jean François IREM de ROUEN

Préambule

Le but de cet exposé est surtout d'ordre culturel et épistémologique. Dans les classes de S.T.S., l'ampleur du programme de mathématiques ne permet pas de s'étendre sur une approche historique du calcul des probabilités et de la Statistique. Cependant, en introduction aux techniques statistiques étudiées, on peut indiquer les problèmes, les savants ayant donné une contribution significative pour la résolution de ces problèmes et poser ainsi quelques jalons chronologiques concernant l'élaboration de la théorie des probabilités et de la Statistique.

Introduction

La théorie des probabilités et la Statistique étant des sciences relativement jeunes (milieu du 17e siècle), leur développement s'est surtout fait à partir de problèmes externes : les jeux, la démographie et la sociométrie - l'arithmétique politique -, la physique statistique, la biométrie... L'apport de problèmes internes est assez récent (surtout à partir du 19e siècle) et reste faible par rapport aux mathématiques ; en effet, la Statistique est essentiellement une science appliquée où les incitations externes sont fortes et beaucoup de problèmes viennent des autres disciplines.

L'idée principale développée ici concerne des résultats qui font le lien entre la théorie des probabilités et la statistique ; ils ont été énoncés (et pour partie démontrés) au 18e et au début du 19e siècle. Les deux théorèmes limites - la loi des grands nombres et le théorème central-limit - sont le fondement mathématique de la statistique inférentielle.

Les développements plus récents de la statistique, en particulier l'apport de l'école anglaise de biométrie à la fin du 19e et la première moitié du 20e siècle, peuvent être trouver dans les ouvrages donnés en bibliographie (partie histoire).

La mise en place des premiers concepts

Les jeux de hasard (dés, cartes) et les questions économiques en avenir incertain (espérances sur un héritage, bourse, rentes viagères,...) ont joué un rôle moteur dans la naissance et le développement du calcul des probabilités et de la statistique ([1]).

C'est le cas en particulier de la correspondance de 1654 entre B. Pascal et P. Fermat, qui porte surtout sur le problème des partis (partage d'un enjeu ou d'un héritage à venir) ([2]). La connaissance des sujets traités entre Fermat et Pascal a incité C. Huygens à écrire le premier ouvrage à être publié en cette matière ([3]). On peut dire, de même, que ce sont des questions posées par des joueurs qui ont amenés J. Bernoulli, P. de Montmort ([4]) et A. de Moivre au calcul des probabilités.

---

[1]) voir par exemple Hacking I. : *The Emergence of Probability*, Cambridge University Press, 1975.

[2]) voir par exemple Pascal B. : Traité du triangle arithmétique, in *Oeuvres complètes*, Gallimard, Coll. La Pléiade, Paris, 1954, où il traite le problème des partis par une méthode de récurrence.

[3]) Huygens, Christiaan : De Ratiociniis in Ludo Aleae, in Frans van Schooten, *Exercitationum Mathematicarum*, Elsevirii, 1657, où il introduit la notion d'espérance mathématique.

[4]) Montmort, Pierre Remond de : *Essay d'Analyse sur les jeux de hazard*, Paris, 1708, 2e éd. 1713. C'est le deuxième ouvrage publié sur le sujet après celui de Huygens.

Un autre thème d'importance commence en Angleterre à la même époque ([5]) sous le nom d'arithmétique politique ([6]). A partir de tables de mortalité, on calcule l'espérance de vie à différents âges pour obtenir les annuités de rentes viagères ([7]) ; on assiste ainsi au début de la théorie des assurances, avec une mesure raisonnée du risque.

La loi des grands nombres

Le premier théorème limite du calcul des probabilités - la loi faible des grands nombres ([8]) - est établi par Jacques Bernoulli dans la quatrième partie de son ouvrage *Ars Conjectandi* [3], publié après sa mort. Il y montre sa proposition principale :

"... j'appellerai *fertiles* les cas dans lesquels un événement peut se produire, et *stériles* les cas dans lesquels le même événement ne peut se produire : de même, j'appellerai expériences *fertiles* celles pour lesquelles on constate qu'un des cas fertiles peut survenir, et *stériles* celles pour lesquelles on observe qu'un des cas stériles se produit. Soit donc le nombre de cas fertiles au nombre de cas stériles précisément ou approximativement dans le rapport r/s et qu'il soit en conséquence, au nombre de tous dans le rapport $r/r+s$ ou r/t, rapport qu'encadrent les limites $r+1/t$ & $r-1/t$. Il faut montrer que l'on peut concevoir des expériences en un nombre tel qu'il soit plus vraisemblable d'autant de fois que l'on veut (soit c) que le nombre des observations tombe à l'intérieur de ces limites plutôt qu'en dehors, c'est-à-dire que le nombre des observations fertiles soient au nombre de toutes les observations dans un rapport ni plus grand que $r+1/t$, ni plus petit que $r-1/t$." en utilisant pour cela des méthodes combinatoires : comparaison des coefficients du binôme dont l'exposant est très élevé.

Le programme ambitieux mis en titre de cette quatrième partie : "L'usage et l'application de la doctrine précédente aux affaires civiles, morales et économiques", qui est l'objectif de l'arithmétique politique, ne sera pas traité par Bernoulli, mais va être une ligne de conduite pour certains de ses successeurs, avec cependant une éclipse pendant la 2e moitié du 19e siècle.

Au 18e siècle, l'arithmétique politique se propose, en s'appuyant sur le calcul des probabilités, de mettre en évidence une régularité des phénomènes démographiques, sociaux et politiques pour fonder des lois ([9]) et faire de la prévision ([10]).

Par exemple, avant l'instauration des recensements (1801 en France), pour obtenir la population d'un pays on dénombrait sur un échantillon les foyers (les feux), ensuite on a utilisé les naissances de l'année, puis on multipliait par un coefficient convenable pour avoir une valeur approchée de cette population. Cette démarche est celle de l'estimation, mais le problème de la représentativité se pose aussi ([11]), ainsi que celui de l'influence d'autres facteurs, et donnera lieu à de nombreuses discussions.

A. de Moivre, après avoir publié un mémoire *De Mensura Sortis* en 1711 ([12]), développe ses recherches dans son traité *The Doctrine of Chances* [9]. En particulier, il y incorpore un résultat de 1733 ([13]) sur une approximation de la probabilité que la somme d'un grand nombre de jets de dés soit dans un intervalle donné, en utilisant un développement en série de logarithme hyperbolique pour les coefficients binomiaux (appelé maintenant formule de Stirling pour la factorielle).

---

5) Graunt, John : *Natural and Political Observations ... made upon the Bills of Mortality*, London, 1662.

6) une définition en est donnée dans l'*Encyclopédie méthodique*, mathématiques, par MM d'Alembert, l'abbé Bossut, de la Lande, le Marquis de Condorcet, &c, Paris, 1784. Réédition du Bicentenaire, ACL-éditions, Paris, 1987.

7) voir par exemple de Moivre : Annuities upon the Lives, in [9].

8) la dénomination serait due à Poisson.

9) dans le même sens que loi en physique.

10) sur ce sujet, on peut lire avec profit l'*Essai philosophique*.. [7a] de Laplace.

11) voir l'article de Bru B. : Estimations laplaciennes dans [17].

12) traduit en anglais par Hald A. dans : *International Statistical Review*, 1984, vol. 52 n°3, pp. 229-262.

13) de Moivre A. : A Method of approximating the Sum of the Terms of the Binomial $(a+b)^n$ expanded into a Series, in [9].

Cette approximation de Moivre améliore la loi des grands nombres de J. Bernoulli ([14]).

Laplace reprend ce résultat dans son traité *Théorie analytique des probabilités* [7d], Livre II, chapitre iii : "Des lois de la probabilité qui résultent de la multiplication indéfinie des événements" et montre, en utilisant des approximations établies (?, ce sont souvent des affirmations sans justification, fondées sur des analogies) dans son Livre I "Du calcul des fonctions génératrices", que la probabilité que le rapport des arrivées d'un événement au nombre total de coups soit renfermé dans des limites fixées peut être approchée à l'aide de l'intégrale de la densité de la loi normale ([15]).

De Moivre obtient aussi par la même méthode une approximation dans le cas où la probabilité de l'événement favorable devient très petite ; mais la caractérisation de la loi limite obtenue est due à Poisson ([10]) et elle porte son nom.

Méthode de Bayes-Laplace

Le problème d'évaluation d'une probabilité inconnue, dans le cas binomial, est apportée par la loi des grands nombres de J. Bernoulli. Le problème inverse : déterminer la distribution de la possibilité d'un événement à partir d'une série d'observations reçoit une première réponse par Bayes ([16]) qui suppose que la possibilité a priori est uniforme entre 0 et 1. Indépendamment, Laplace montre un théorème connu sous le nom de théorème de Bayes (ou Bayes-Laplace) dans son mémoire sur la probabilité des causes ([7b]) en utilisant un principe de vraisemblance maximum, énoncé par Lambert en 1760. Comme application de cette méthode, on trouve les recherches sur la probabilité des jugements pour lesquelles on peut mentionner Condorcet ([17]), Laplace ([18]) et Poisson [10].

De la théorie des erreurs au théorème central-limit

Au cours des observations astronomiques effectuées depuis l'Antiquité pour essayer de déterminer les lois qui régissent les astres - fixité ou périodicité -, on avait remarqué que les mesures faites sur un objet céleste n'étaient pas toujours les mêmes. De plus, lors de la mesure du méridien faite au 18e, malgré tout le soin apporté, on s'est aperçu qu'il y avait encore des discordances, des erreurs d'observation, écart entre la vraie valeur inconnue et les résultats d'observation.

Pour obtenir une valeur approchée de la constante inconnue, diverses méthodes étaient utilisées sans justification. Simpson, en 1757, applique le calcul des probabilités à ce problème en considérant les erreurs de mesure comme des pertes au cours d'un jeu de hasard (contre la Nature), propose quelques lois de distribution - appelées loi des erreurs - et étudie la distribution de la moyenne arithmétique de n observations indépendantes.

Ce problème est du même type que la recherche du rapport entre les naissances des garçons et des filles à laquelle vont participer D. Bernoulli, J.L. Lagrange et P. Laplace ([19]).

---

[14]) Laplace, dans son *Essai* ([7a], p.199), écrit : "Il ne se contente pas de faire voir, comme Bernoulli, que le rapport des événements qui doivent arriver approche sans cesse de celui de leurs possibilités respectives, il donne de plus une expression élégante et simple de la probabilité que la différence de ces deux rapports est contenue dans des limites données."

[15]) Il semble que c'est D. Bernoulli qui étudia le premier (vers 1794) la distribution continue appelée maintenant loi normale (d'après K. Pearson, 1893 ou Poincaré) ou de Laplace-Gauss.

[16]) Bayes, Thomas : An essay towards solving a Problem in the Doctrine of Chances, *Phil. Trans. of the Royal Society*, Londres, 1763, reproduit dans [15] pp. 131-153 et traduit dans [1].

[17]) Condorcet, Caritat de : *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*, 1785, in *Sur les élections*, Fayard, 1986.

[18]) Laplace met dans l'avertissement de la 2e édition de *Théorie Analytique* : "La théorie de la probabilité des témoignages, omise dans la première édition, est ici présentée avec le développement qu'exige son importance". Il y consacre une partie de l'introduction (l'*Essai*, p.118), le chapitre xi : "De la probabilité des témoignages", ainsi qu'une partie du *Premier Supplément* "De la probabilité des jugements".

[19]) voir en particulier l'article de Bru op. cit.

A propos de ce problème, on peut voir chez Laplace [7c] le premier test d'hypothèse lorsqu'il fait la comparaison de ces rapports observés à Londres et à Paris ([20]).

Cette théorie des erreurs conduit d'une part à la recherche d'estimateurs d'une valeur centrale, par exemple chez Daniel Bernoulli (voir [2]) et Laplace ([7c]), et d'autre part à la détermination de la loi des erreurs pour laquelle plusieurs distributions furent proposées, et l'énonciation du théorème central-limit ([21]) par Laplace (1810), en particulier dans [7d] livre II, chapitre iv : "De la probabilité des erreurs des résultats moyens d'un grand nombre d'observations, et des résultats moyens les plus avantageux".

La "démonstration" du théorème central-limit par Laplace a paru fort obscure à ses successeurs. Son argumentation est basée sur les résultats indiqués dans son Livre I : "Du calcul des fonctions génératrices", et en particulier de la Seconde Partie "Théorie des approximations des formules qui sont fonctions de grands nombres". Il fait des extensions par analogie du cas fini au cas infini, que ce soit lors du passage des équations aux différences finies à des équations différentielles que pour les intégrations. Il utilise des développements en série et les intègre ou les différencie sans se préoccuper de la convergence. Il admet que la transformation d'une suite à sa fonction génératrice et inversement, se conserve pour les opérateurs sur ces fonctions. Il considère, par analogie, que les intégrales de fonctions d'une variable complexe ont les mêmes propriétés que celles des fonctions d'une variable réelle, par exemple $\int dt.f(t).c^{-\alpha t\sqrt{-1}}$, où "c est le nombre dont le logarithme hyperbolique est l'unité". Il utilise des extensions qui ont paru mal fondées. Au n°18, après avoir montré le résultat si les erreurs peuvent prendre des valeurs entières uniformément entre -n et +n, il écrit : "Supposons généralement que la probabilité de chaque erreur positive ou négative soit exprimée par $\varphi(x/n)$, x et n étant des nombres infinis."

Le problème des erreurs d'observation en astronomie, où l'on a plusieurs paramètres inconnus, va amener à la méthode des moindres carrés par Legendre A.M., Laplace P.S. ([7d], II n°20 à 23) et Gauss C.F. ; pour Gauss et Laplace, cette méthode est basée sur la nature probabiliste des erreurs ([22]). La méthode des moindres écarts, développée auparavant par Laplace ([7c] et [7d], II n°23), rend nécessaire l'utilisation du milieu de probabilité (la médiane) et de la distribution dite 1ère loi de Laplace.

Pendant la première moitié du 19e siècle, les travaux se poursuivent sur l'application du calcul des probabilités à des phénomènes sociaux : Laplace ([7d], chapitre XI : De la probabilité des témoignages), Poisson (op. cit.), Quetelet A. (*Essai de Physique Sociale*, Paris, 1835). Cependant, en raison de la difficulté pour attribuer une probabilité, comme Laplace l'avait noté ([23]), les controverses sur les bases scientifiques de ces études et les critiques ([24]) portées contre la notion d' "homme moyen" introduite par Quetelet jettent le discrédit sur ce genre d'étude, tout au moins en France.

---

[20]) Dans l'*Essai* ([7a]), Laplace écrit : (p.88) "Mais avant que d'en rechercher les causes, il est nécessaire, pour ne point s'égarer dans de vaines spéculations, de s'assurer qu'ils sont indiqués avec une probabilité qui ne permet point de les regarder comme des anomalies dues au hasard." et (p.116), "Le calcul des probabilités peut faire apprécier les avantages et les inconvénients des méthodes employées dans les sciences conjecturales. Ainsi, pour reconnaître le meilleur des traitements en usage dans la guérison d'une maladie, il suffit d'éprouver chacun d'eux sur un même nombre de malades, en rendant toutes les circonstances parfaitement semblables... le calcul fera connaître la probabilité correspondante de son avantage et du rapport suivant lequel il est supérieur aux autres."

[21]) appellation de Polya de 1920, que je conserve telle que, la francisation de cette expression en "théorème de la limite centrale, ou centrée" n'ayant aucun sens. Laplace présente ce résultat dans son Essai [7a, p.91].

[22]) Laplace écrit dans [7d], II chapitre iv, n°23 : "La méthode des moindres carrés des erreurs devient nécessaire, lorsqu'il s'agit de prendre un milieu entre plusieurs résultats donnés, chacun, par l'ensemble d'un grand nombre d'observations de divers genres." Il obtient aussi la loi normale comme loi des erreurs "qui donne constamment la règle des milieux arithmétiques."

[23]) ([7a, p 38]) : "DEUXIEME PRINCIPE : Mais cela suppose les divers cas également possibles. S'ils ne le sont pas, on déterminera d'abord leurs possibilités respectives, dont la juste appréciation est un des points les plus délicats de la théorie des hasards."

[24]) tout d'abord par Cournot, A. : *Exposition de la théorie des chances et des probabilités*, Paris, 1843, et repris par Bertrand, J. : *Calcul des probabilités*, Hachette, Paris, 1889.

L'époque moderne : fin du 19e, 20e

Pendant la deuxième moitié du 19e siècle, les mathématiciens qui s'intéressent au calcul des probabilités se garderont de toute application aux phénomènes humains et sociaux. Ils vont partir des résultats de leurs prédécesseurs (en particulier Laplace, [7d]) pour en donner des démonstrations avec des conditions moins restrictives. Dès cette époque, les deux théorèmes limites vont devenir des problèmes internes pour la théorie des probabilités.

Bienaymé ([25]) démontre une inégalité - dite aussi de Tchebychev - avec laquelle il obtient "un théorème sur les probabilités qui contient comme cas particuliers le théorème de Bernoulli et la loi des grands nombres". C'est une loi faible des grands nombres pour une suite de variables aléatoires dont les variances sont majorées par un même nombre. Dans un autre mémoire ([26]), il montre un théorème central-limit pour une suite de v.a. dont "les espérances mathématiques de toutes leurs puissances ne dépassent pas une limite finie quelconque".

Dans la recherche de conditions plus larges pour la loi des grands nombres et le théorème central-limit, on peut noter l'apport de l'école russe de la fin du 19e siècle avec Tchebychev P.L. ([27]), Markov A.) qui donne la première démonstration rigoureuse du théorème central-limit en 1898, Liapounov A. (1901), Lindeberg J.W. (1922), jusqu'à Khintchine A. ([28]).

Dans cette recherche sur la loi des grands nombres, on va utiliser une autre sorte de convergence : la convergence presque sûre. La démonstration est basée sur un résultat appelé lemme de Borel-Cantelli (voir [4]).

On peut trouver différentes conditions pour lesquelles une suite de variables aléatoires vérifie la loi des grands nombres (faible ou forte) et le théorème central-limit par exemple dans Feller [5] ou Lévy [8] ; la démonstration du théorème central-limit par la méthode des fonctions caractéristiques est due à P. Lévy, qui justifie en les précisant les affirmations de Laplace.

Vers l'axiomatisation de la théorie des probabilités

Le calcul des probabilités va bénéficier des progrès effectués dans la théorie de la mesure des ensembles linéaires, puis quelconques, et de la théorie de l'intégration faits par : Baire M., Borel E., Lebesgue H. ([29]) et une généralisation par Fréchet M. ([30]).

L'extension de l'intégrale à un espace abstrait (Radon, Fréchet) va permettre l'axiomatisation de la théorie des probabilités qui est réalisée par A.N. Kolmogorov ([31]) ; la notion fondamentale est l'espace probabilisé, c'est-à-dire mesuré avec une mesure totale égale à 1. On peut donc utiliser tous les résultats démontrés en théorie de la mesure dans le cas particulier d'une mesure finie. Le calcul des probabilités sort du rôle de calculs sur les jeux pour devenir une théorie mathématique à part entière.

---

[25]) Bienaymé, I.J. : Des valeurs moyennes, *Comptes Rendus de l'Académie des Sciences* (C.R.A.S.), Paris, entre 1850 et 60 (?)

[26]) Bienaymé, I.J. : Sur deux théorèmes relatifs aux probabilités, idem

[27]) Tchebychev donne en 1887 un énoncé clair et une démonstration (incomplète) du théorème central-limit.

[28]) Khintchine, A. : Sur la loi des grands nombres, *C.R.A.S.*, Paris, 1929.

[29]) Lebesgue, H. : *Leçons sur l'intégration et la recherche des fonctions primitives*, Gauthier-Villars, Paris, 1905.

[30]) Fréchet, M. : Sur l'intégrale d'une fonctionnelle étendue à un ensemble abstrait, *Bull. Soc. Math. Fr.*, 1915.

[31]) Kolmogorov, A.N. : *Foundations of the Theory of Probability*, Chelsea Press, New-York, 1950 ; 1e éd. en allemand, Springer, 1933.

Bibliographie

Ouvrages originaux (ou traduction)

[1] Bayes, Thomas : *Essai en vue de résoudre un problème de la doctrine des chances*, 1763 ; traduit par J.P. Cléro, Cahiers d'Histoire et de Philosophie des Sciences, n° 18, 1988.

[2] Bernoulli, Daniel : *Dijudicatio maxime probabilis plurium observationum discrepantium..*, Acta Acad. Sc. Petrop., 1777 ; traduit dans [14] : The most probable choice between several discrepant observations..

[3] Bernoulli, Jacques : *Ars Conjectandi*, 1713 ; 4e partie traduite par N. Meusnier dans *Jacques Bernoulli & l'ars conjectandi*, IREM de Rouen, 1987.

[4] Borel, Emile : Oeuvres en 4 vol., CNRS, Paris, 1972.
- *Les probabilités dénombrables et leurs applications arithmétiques*, R. C. Circolo Mat. Palerma, 1909 ; Oeuvres t.2, Ed. du CNRS, Paris, 1972.
--- : *Valeur pratique et philosophie des probabilités*, Gauthier-Villars, Paris, 2e éd. 1952.

[5] Feller, William : *An introduction to probability theory and its applications*, Wiley, New-York, tome I, 1e éd. 1950, 2e éd. 1957.

[6] Fréchet, Maurice : *Généralités sur les probabilités. Eléments aléatoires*, 2e éd., Gauthier-Villars, Paris, 1950.

[7] Laplace, Pierre Simon de : a- *Essai philosophique sur les probabilités*, préface de R. Thom, postface de B. Bru, Bourgois, Paris, 1986 ([32])
--- : Oeuvres complètes, Gauthier-Villars, Paris, de 1878 à 1886
b- *Mémoire sur la probabilité des causes par les événements*, 1774, O.C. t.8.
c- *Mémoire sur les probabilités*, 1781, O.C. t.9.
d- *Théorie analytique des probabilités*, 3e éd. 1820, O.C. t.7.

[8] Lévy, Pierre : *Théorie de l'addition des variables aléatoires*, Gauthier-Villars, Paris, 1937.

[9] Moivre de, Abraham : *The Doctrine of Chances*, 1e éd., 1718 ; 3e éd., 1756 réimprimée par Chelsea, New-York, 1967.

[10] Poisson Denis : *Recherches sur la probabilité des jugements en matière criminelle et en matière civile*, Paris, 1837.

Ouvrages sur l'histoire de la théorie des probabilités et de la statistique

[10] Benzécri, Jean-Pierre : *Histoire et préhistoire de l'analyse des données*, Dunod, Paris, 1982.

[11] Dieudonné, Jean et alii : *Abrégé d'histoire des mathématiques, 1700-1900*, t. 2, Hermann, Paris, 1978.

[12] Droesbeke, J.-J. et Tassi, Ph. : *Histoire de la Statistique*, P.U.F., Coll. "Que sais-je?", Paris, 1990.

[13] Feldman J., Lagneau G., Matalon B. éds : *Moyenne, Milieu, Centre ; histoires et usages*, Ed. de l'EHESS, Paris, 1991.

[14] Gillispie C.C. : *Dictionary of Scientific Biography*, C. Scribner's Sons, New-York, 16 vol. de 1970 à 1980.

[15] Pearson, Egon S. and Kendall, Maurice G. eds : *Studies in the history of Statistics and Probability*, vol.1, Griffin & Co, London, 1970 ([33]).

[16] Kendall Maurice G. and Plackett Robin L. eds : *Studies...*, vol.2, Griffin & Co, 1977.

[17] Mairesse, Jacques éd. : *Estimation et sondages*, cinq contributions à l'histoire de la statistique, Economica, Paris, 1988.

[18] Stigler, Stephen M. : *The History of Statistics*, Belknap Harvard, 1986.

[19] Todhunter, Isaac : *A history of the mathematical theory of probability*, Cambridge, 1865 ; réédition Chelsea, New-York, 1965.

---

[32]) L'*Essai philosophique* sert d'introduction à la *Théorie Analytique* à partir de la 2e éd. de 1812. Les références à l'*Essai* sont données pour cette dernière édition qui est plus facile à trouver que les Oeuvres Complètes.

[33]) Cet ouvrage, ainsi que le suivant, est en grande partie une compilation d'articles de Biometrika, revue créée par Karl Pearson.

METHODES EN STATISTIQUE : ESTIMATION

HENRY Michel
IREM de BESANÇON

I - LE PROBLEME DE L'ESTIMATION

A) LES ORIGINES ; CONCEPTIONS DE JACQUES BERNOULLI.

B) L'ESPRIT FREQUENTISTE DES PROGRAMMES DU SECOND DEGRE

C) PROBLEME D'AJUSTEMENT D'UNE LOI

II - MODELE DE LA STATISTIQUE INFERENTIELLE

A) MODELE PROBABILISTE

B) MODELE DE LA STATISTIQUE

C) MODELE POUR L'ECHANTILLONNAGE

III - ECHANTILLONNAGE ET STATISTIQUES

A) LOI D'UN ECHANTILLON ET FONCTION DE VRAISEMBLANCE

B) RESUME STATISTIQUE ET LOI D'UNE STATISTIQUE

C) STATISTIQUES USUELLES $\overline{X}$ ET S^2
CAS PARTICULIER D'UN ECHANTILLON GAUSSIEN

IV - ESTIMATEURS

A) CADRE DE L'ESTIMATION

B) QUALITES D'UN ESTIMATEUR : **biais, convergence, efficacité, limite théorique des performances d'un estimateur**

C) RECHERCHE D'UN ESTIMATEUR : METHODE DU MAXIMUM DE VRAISEMBLANCE

V - ESTIMATION PAR INTERVALLE

A) POSITION DU PROBLEME

B) PRINCIPE DE LA DETERMINATION D'UN INTERVALLE DE CONFIANCE

C) MISE EN PRATIQUE DE LA RECHERCHE DE L'INTERVALLE DE CONFIANCE

D) FORMULATION DE LA CONCLUSION

I - LE PROBLEME DE L'ESTIMATION

A) LES ORIGINES ; CONCEPTIONS DE JACQUES BERNOULLI.

Dans une approche naïve du Calcul des Probabilités, on cherche à évaluer les "chances" que l'on a d'observer un événement dont l'apparition est conditionnée par le hasard.

Dans une démarche scientifique, il s'agit de discerner les variables pertinentes dans des situations concrètes permettant de dégager un modèle suffisamment simple, dans lequel il est possible d'utiliser les outils mathématiques.

Cette intention de calculer des probabilités suppose une prise de parti épistémologique qui n'a été ni simple ni précoce dans l'histoire des sciences. Il a fallu accepter l'idée que les phénomènes aléatoires sont des phénomènes objectifs qui peuvent être quantifiés indépendamment de la sensibilité et des préjugés de l'observateur (position objectiviste).

Dans ce cadre, le calcul des probabilités s'accommode-t-il du déterminisme ?

Dans la préface à l'"*Essai philosophique sur les probabilités*" de Laplace, René Thom indique [1, p.23] : "*En science, l'aléatoire pur, c'est le processus markovien, où toute trace du passé s'abolit dans la genèse du nouveau coup... l'aléatoire pur exige un fait sans cause, c'est-à-dire un commencement absolu.*"

Ainsi, dans la dyade déterminisme - hasard, une position philosophique interprétant les phénomènes naturels comme purement aléatoires, ou soumis à la volonté divine imprévisible, conduit nécessairement l'observateur à exclure tout modèle mathématique et le réduit à l'attribution de "probabilités subjectives" (selon les termes repris par Poincaré [2, p.195]) qui ne peuvent, en l'absence de préférences, qu'être réduites à l'équiprobabilité pour les différentes éventualités qui sont susceptibles de se présenter : l'âne de Buridan aurait une chance sur deux de choisir d'abord le seau d'eau - s'il ne veut pas mourir de faim et de soif.

Dans son livre "*Au hasard*", Ivar Ekeland nous livre de belles pages sur l'approche moderne du hasard, évoquant l'évolution historique des conceptions qui conduisent au traitement scientifique [4, p.63] : "*Si la réalité ultime est décrite par le calcul des probabilités, le monde sera soumis aux lois de la statistique.*" Et, [p.80] : "*l'incertitude est une des données fondamentales de l'histoire humaine et de notre vie quotidienne. En permanence, il nous faut prendre des décisions dans un contexte que nous apprécions mal*".

Cette nécessité de prendre des décisions avait conduit Pascal, dans sa correspondance avec Fermat, à retenir comme concept de probabilité ce qui est a priori calculable par logique, symétries et dénombrements, dans le degré d'incertitude que l'on a sur l'apparition de tel ou tel résultat d'une expérience aléatoire clairement décrite et dont les événements élémentaires sont des cas équiprobables (problème des Partis : [6, p.217]). D'où la formule donnant la probabilité d'un événement composé :

$$\frac{\text{nombre de cas favorables}}{\text{nombre de cas possibles}}$$

qui semble exclure toute subjectivité, mais qui ne s'applique qu'aux expériences aléatoires bien caractéristiques (dés, urnes...) ayant un petit nombre d'issues que l'on peut, selon les termes de Laplace [1, p.35] : "*réduire à un certain nombre de cas également possibles*".

La modélisation (par exemple en termes d'urnes) de telles expériences est alors élémentaire et le calcul des probabilités est essentiellement du ressort mathématique. D'où le terme de "géométrie du hasard" introduit par Pascal, qu'il faut comprendre dans le sens où la géométrie est la "mathématique du réel" qui, du point de vue de Platon, s'opposerait à l'arithmétique opérant dans le domaine des Idées.

Mais cette conception est insuffisante pour s'attaquer aux vrais problèmes et est marquée d'un vice épistémologique à la base, comme le fait remarquer Poincaré [2, p.192] : "*On est donc réduit à compléter cette définition [de Pascal] en disant* : «... au nombre total des cas possibles, pourvu que ces cas soient également probables». *Nous voilà donc réduits à définir le probable par le probable.*"

Comment, en réalité, relier le "degré d'incertitude" (terme de Laplace) que nous avons sur un événement à la probabilité objective de cet événement, que seule la compilation statistique dans un grand nombre de réalisations peut révéler a posteriori (Poincaré) ?

Cette question, qu'en termes actuels nous appelons "problème d'estimation des probabilités", était déjà posée en toute clarté par Jacques Bernoulli dans son *Ars Conjectandi*, publié en 1713 après sa mort ([5, p.40]). Il souligne les conditions restrictives (jeux de hasard,...) auxquelles s'applique la définition de Pascal et montre que la complexité des phénomènes naturels suppose une autre manière de concevoir la probabilité qui, pour ne pas être subjective, doit provenir d'une étude des fréquences des événements issus de nombreuses expériences ou situations aléatoires identiques. Cette assimilation fréquence - probabilité est justifiée par la loi des grands nombres qui, d'une perception intuitive commune, passe au statut de résultat mathématiquement établi : c'est l'objet de l'*Ars Conjectandi*.

Il est saisissant de voir en quels termes contemporains Bernoulli introduit cette problématique [5, p.42,44] :

« On en est ainsi venu à ce point que pour former selon les règles des conjonctures sur n'importe quelle chose il est seulement requis d'une part que les nombres de cas soient soigneusement déterminés, et d'autre part que soit défini combien les uns peuvent arriver plus facilement que les autres. Mais c'est ici enfin que surgit une difficulté, nous semble-t-il : cela peut se voir à peine dans quelques très rares cas et ne se produit presque pas en dehors des jeux de hasard que leurs premiers inventeurs ont pris soin d'organiser en vue de se ménager l'équité, de telle sorte que fussent assurés et connus les nombres de cas qui doivent entraîner le gain ou la perte, et de telle sorte que tous ces cas puissent arriver avec une égale facilité. En effet lorsqu'il s'agit de tous les autres résultats, dépendant pour la plupart soit de l'oeuvre de nature soit de l'arbitre des hommes, cela n'a pas du tout lieu. Ainsi, par exemple, les nombres de cas sont connus lorsqu'il s'agit des dés, car pour chacun des dés les cas sont manifestement aussi nombreux que les bases, et ils sont tous également enclins à échoir. ...
Mais qui donc parmi les mortels définira par exemple le nombre de maladies,... ?
Qui encore recensera les cas innombrables des changement auquel l'air est soumis chaque jour, en sorte qu'on puisse à partir de là conjecturer ce que sera son état après un mois ?...
En outre qui aurait sur la nature de l'esprit humain, ou sur l'admirable fabrique de notre corps une vue suffisante pour oser déterminer dans les jeux, qui dépendent en totalité ou en partie de la finesse de celui-là ou de l'agilité de celui-ci, les cas qui peuvent donner la victoire ou l'échec. ...

Mais à la vérité ici s'offre à nous un autre chemin pour obtenir ce que nous cherchons. Ce qu'il n'est pas donné d'obtenir a priori l'est du moins *a posteriori*, c'est-à-dire qu'il sera possible de l'extraire en observant l'issue de nombreux exemples semblables ; car on doit présumer que, par la suite, chaque fait peut arriver et ne pas arriver dans le même nombre de cas qu'il avait été constaté auparavant, dans un état de choses semblables, qu'il arrivait ou n'arrivait pas. »

B) L'ESPRIT FREQUENTISTE DES PROGRAMMES DU SECOND DEGRE

(cf. les programmes de 1ère et de terminales de 1991 et l'article de REPERES-IREM [9])

A partir de ces considérations épistémologiques, nous pouvons analyser la démarche de ces programmes qui s'affirment plus "fréquentiste" que les précédents dans l'approche de la notion de probabilité.

Modélisation d'une expérience aléatoire

Ces programmes inscrivent d'abord l'introduction des probabilités dans la continuité des apprentissages en statistique, particulièrement lors de l'étude des séries statistiques réalisée en seconde.

Mais, et ceci est fondamental, une probabilité ne peut avoir de sens que si elle concerne un événement associé à une expérience aléatoire.

En l'absence d'expérience aléatoire, pas de probabilité !

Le concept de probabilité suppose donc acquis ceux d'expérience aléatoire et d'événement. C'est d'ailleurs souligné en premier lieu dans le programme : "*l'objectif est d'entraîner les élèves à décrire quelques expériences aléatoires simples*". Ce concept se dégagera alors de cette activité de description portant sur différents exemples dans des cadres variés.

Si, ensuite, on veut développer un travail mathématique, la notion d'expérience aléatoire doit pouvoir être modélisée. Nous savons que le langage et les opérations sur les ensembles le permettent effectivement. Nous ferons les trois hypothèses suivantes :

a- Une telle expérience met en jeu un phénomène aléatoire : son issue ne peut être prévue à l'avance, elle est le "fruit du hasard". Ainsi, les conditions de l'expérience, telles qu'elles sont décrites, ne déterminent pas l'un des résultats possibles de manière absolue.

b- Pour représenter une expérience aléatoire, on considère donc un ensemble de résultats possibles, bien identifiés.

c- Enfin, une expérience aléatoire doit être reproductible dans les mêmes conditions (au moins par la pensée).

Ainsi, les faits historiques ne peuvent être considérés comme résultant d'expériences aléatoires et ne sauraient être probabilisés.

A la notion d'expérience aléatoire est donc liée celle de hasard, avec les difficultés épistémologiques ou philosophiques que nous connaissons et qui justifient une longue dissertation en introduction de l'ouvrage de Henri Poincaré "*Calcul des probabilités*" publié en 1912. En voici un extrait :

Le déterminisme Laplacien pose que "*le mot hasard est tout simplement un synonyme d'ignorance, qu'est-ce que cela veut dire ? ... Il faut donc bien que le hasard soit autre chose que le nom que nous donnons à notre ignorance, que parmi les phénomènes dont nous ignorons les causes, nous devions distinguer les phénomènes fortuits, sur lesquels le calcul des probabilités nous renseignera provisoirement, et ceux qui ne sont pas fortuits et sur lesquels nous ne pouvons rien dire, tant que nous n'aurons pas déterminé les lois qui les régissent*" [3, p.3].

Cette remarque montre que pour comprendre les difficultés conceptuelles de la notion de probabilité, nous nous heurterons à des obstacles épistémologiques et l'on n'évitera pas un travail de fond avec les élèves qui abordent l'aléatoire pour la première fois en mathématiques

Ces obstacles nous sont révélés par les hésitations historiques des Bernoulli, D'Alembert, Laplace, Poincaré. Il est bon, alors, que les enseignants de mathématiques aient fait le point sur leurs propres conceptions, les aient confrontées à celles des mathématiciens du passé pour ne pas être pris au dépourvu par tel ou tel comportement d'élève qui rencontre à cette étape des difficultés de compréhension.

Approche fréquentiste de la probabilité

C'est ici un objectif particulier du programme que d'associer la notion de probabilité d'un événement à la fréquence stabilisée de cet événement au cours d'un grand nombre d'expériences identiques. Le programme de première ajoute : "*Pour introduire la notion de probabilité, on s'appuiera sur l'étude des séries statistiques obtenues par répétition d'une expérience aléatoire, en soulignant les propriétés des fréquences et la relative stabilité de la fréquence d'un événement donné lorsque cette expérience est répétée un grand nombre de fois.*"

Les spécialistes y verront une allusion à la loi des grands nombres, tout en pesant les difficultés épistémologiques qui lui sont inhérentes et les difficultés didactiques d'un énoncé précis à ce niveau.

La probabilité est ainsi liée à la notion de fréquence, laquelle prend du sens par la description de multiples exemples de phénomènes aléatoires. C'est pour cela que nous parlons de conception fréquentiste, en opposition à la conception qui postule l'équiprobabilité des événements élémentaires, posée a priori pour des raisons de symétrie et que nous appellerons "*l'approche pascalienne*", en référence à l'auteur de la formule de définition de la probabilité dans la "*géométrie du hasard*".

La probabilité est donc conçue dans les programmes du secondaire comme une valeur numérique injectée dans un modèle de l'expérience sensible, modèle dégagé de l'activité qui, en statistique, est centrée sur le recueil et l'organisation de données.

Remarquons que cette démarche est cohérente avec les objectifs des programmes de collège et de seconde en vigueur, privilégiant l'activité des élèves, l'observation et la formulation de conjectures, avant que le modèle mathématique soit institutionnalisé. S'opposant à une présentation formelle des mathématiques, ces programmes mettent en avant l'étude d'outils de description d'une réalité concrète.

Ces questions ayant été clarifiées, on est placé devant un problème d'estimation : dans quelle mesure les 48,5 % de "pile" observés justifient-ils le 0,5 attribué à la probabilité du "pile" ? Problème pas très difficile au niveau BTS, mais exclu en classe de première sur le plan conceptuel même. Au fait, pourquoi 0,5 et non 0,49 ? (la pièce est peut-être déséquilibrée).

Ainsi la fréquence observée sur un grand nombre d'expériences ne doit pas être confondue avec la notion mathématique de probabilité. Celle-ci est conçue comme donnée numérique du modèle, estimée à partir de l'observation de la stabilité de cette fréquence.

Dans sa cohérence, le programme propose "*l'observation de la stabilité approximative de la fréquence f_n d'un événement donné lorsque l'expérience est répétée un grand nombre n de fois*".

Il ne s'agit donc pas d'établir une estimation de la probabilité p limite. D'ailleurs, dans ce cas, la meilleure estimation est la fréquence observée pour le plus grand nombre d'expériences et l'on n'a que faire de sa stabilisation.

De plus, et le programme le précise, cette stabilisation ne peut être que relative. En effet, en théorie, la convergence de f_n vers p n'est pas monotone :
si $p = ½$, il y a une chance sur 2 pour qu'au prochain tirage f_n s'éloigne de p.

Il convient cependant d'habituer les élèves à cette stabilité relative, de leur faire apprécier sur des exemples le nombre n d'expériences nécessaires pour l'observer et de leur donner confiance dans l'approche fréquentiste. Celle-ci, en effet, correspond mieux à la pratique sociale et à l'usage actuel en statistique dans l'induction suivante : si, dans un échantillon assez vaste pris au hasard dans une population, j'observe une proportion p d'éléments de catégorie A, le choix au hasard d'un autre élément de la population donnera A avec une probabilité que je peux prendre égale à p.

Cette pratique imagine la répétition de l'expérience une fois de plus et suppose la stabilité de la fréquence observée dès que l'échantillon préalable est assez grand ([1]).

Le programme propose donc de telles observations, concrètement, en poursuivant l'objectif de privilégier l'activité des élèves comme support à leur conceptualisation.

c) PROBLEME D'AJUSTEMENT D'UNE LOI

La démarche empirique qui consiste à "estimer" une probabilité par la fréquence observée de l'événement reste très limitée dans ses résultats pratiques.

D'une part, elle ne dit pas quel est le degré d'approximation qu'on obtient ainsi, ni avec quelle précision on peut introduire les probabilités dans des calculs complexes, où des incertitudes sur les données initiales peuvent provoquer de grandes variations sur les résultats.

D'autre part, elle limite le concept de probabilité ou son domaine d'application aux expériences aléatoires effectivement répétables un grand nombre de fois et exclut de son champ les situations complexes non reproductibles, économiques par exemple.

---

1) voir cependant Emile BOREL : *Valeur pratique et philosophie des probabilités*, Gauthier-Villars, Paris, 2ème éd. 1952.

Le développement de la théorie des probabilités a montré l'importance de la notion de variable aléatoire et, pour le développement des modèles, de celle de loi.

Ainsi, plutôt que de déterminer a posteriori toutes les probabilités relatives aux événements issus d'une expérience aléatoire, il est plus avantageux de décrire cette expérience en introduisant les variables pertinentes et de déterminer les lois de ces variables qui permettent ensuite tous les calculs de probabilités souhaités.

Dans la pratique, la loi est un outil théorique sensé décrire au mieux la répartition des valeurs possibles de la variable aléatoire. Le problème est donc de choisir, parmi un ensemble catalogué de lois théoriques bien connues, celle qui satisfait le mieux aux conditions de l'expérience aléatoire. C'est ce qu'on appelle un problème d'ajustement.

En réalité, suivant l'expérience aléatoire et le choix des variables la décrivant, il y a certains types de lois qui viennent s'imposer. Les connaissances et le savoir-faire du probabiliste - statisticien lui permettront de faire le bon choix.

Par exemple, s'il pleut et si, sur un territoire limité, je m'intéresse à la quantité d'eau reçue par les éléments de surface du sol, il sera naturel de suggérer une loi uniforme en dimension 2.

Par contre si j'arrose mon gazon avec un jet fixe muni d'une pomme de dispersion, je penserai plutôt à une loi normale que j'essaierai de contrôler expérimentalement .

Ainsi, l'outil mathématique puissant que donne le modèle de Kolmogorov, dans lequel la notion de loi est centrale, nous conduit-il à estimer non plus la probabilité de chaque événement, mais les paramètres qui déterminent numériquement les lois des variables en cause.

II - MODELES DE LA STATISTIQUE INFERENTIELLE

(Pour cet enseignement à des sections de techniciens supérieurs, on pourra se reporter à [8] et [10])

A) MODELE PROBABILISTE

En statistique inférentielle, on s'intéresse à des expériences aléatoires particulières et à un problème particulier :
Etant donnée une population $\mathcal{P}$ que l'on désire étudier, l'expérience consiste en un prélèvement (non exhaustif, i.e. avec remise, en théorie) "au hasard" de n éléments de $\mathcal{P}$ (l'échantillon au sens commun).

Le problème est d'inférer, à partir de l'observation de cet échantillon, les propriétés de $\mathcal{P}$.

On va supposer pour la suite que ces propriétés sont interprétées par les valeurs d'un caractère χ (éventuellement multidimensionnel) que nous prendrons quantitatif (pour nous limiter).

Le modèle probabiliste décrivant le prélèvement au hasard d'un élément de $\mathcal{P}$ introduit un ensemble Ω (avec les notations habituelles) représentant cette population :

$$(\Omega, \mathcal{T}, P) \xrightarrow{X} (\mathbf{R}^d, \mathcal{B}, P_X)$$

avec la signification des lettres :
- Ω ensemble des éléments de la population, • $\mathcal{T}$ tribu sur Ω, • P distribution de la probabilité sur Ω
- X v.a. représentant le caractère χ, à valeurs dans $\mathbf{R}^d$, • $\mathcal{B}$ boréliens de $\mathbf{R}^d$,
- P_X mesure image de P par X décrite généralement par l'un des moyens suivants :
  - probabilités élémentaires discrètes
  - fonction de répartition
  - densité
  - fonction caractéristique

B) MODELE DE LA STATISTIQUE

Le modèle statistique que nous allons utiliser doit décrire le prélèvement de l'échantillon.

Plaçons nous directement dans l'espace image de $\mathcal{P}$ par le caractère χ.

Soit E_0 l'ensemble des valeurs possibles de χ. On introduit X_0 la v.a.(2) représentant le tirage au hasard d'un élément de $\mathcal{P}$ et l'application à cet élément de χ (observation de la valeur du caractère). X_0 dans ce modèle est alors l'identité de E_0, mais son introduction assure la prise en compte du caractère aléatoire des valeurs observées.

Les valeurs de X_0 sont réparties suivant une loi de probabilité P_{X_0}. X_0 sera appelée variable parente (de l'échantillon) et P_{X_0} la loi parente.

En réalité, c'est P_{X_0} que l'on désire connaître. Il y a divers degrés d'incertitude sur cette loi qui déterminent la répartition des valeurs de χ et par conséquent les éléments caractéristiques de $\mathcal{P}$ comme les valeurs moyenne, écart-type, etc...

Suivant la connaissance de $\mathcal{P}$ et de χ que l'on a, P_{X_0} sera à rechercher parmi une famille plus ou moins grande de lois possibles, chacune déterminée numériquement par différents paramètres.

On indique cette indétermination par l'introduction d'un paramètre $\theta \in \Theta$ qui peut être numérique, éventuellement multidimensionnel. Il caractérise donc le couple $(\mathcal{P}, \chi)$.

P_{X_0} dépend de θ, on la note alors $P_{X_0,\theta}$ et l'ensemble des lois possibles par $(P_{X_0,\theta})$, $\theta \in \Theta$.

D'où le modèle représentant le problème de l'inférence : $\left(E_0, \mathcal{B}_0, (P_{X_0,\theta})\ \theta \in \Theta\right)$.

C) MODELE POUR L'ECHANTILLONNAGE

L'expérience consiste à observer (et traiter) n réalisations indépendantes de X_0, prélèvements au hasard avec remise, c'est-à-dire n observations de χ.

On notera X_i la i-ème opération de ce type :

X_i est une v.a. définie sur E_0 (identité) de même loi que X_0.

La réplique n fois de la même expérience aléatoire dans les mêmes conditions est interprétée dans le modèle par l'hypothèse que les X_i sont indépendantes.

Soit alors $E = E_0^n$ l'ensemble des n-uples d'éléments de E_0 et $X = (X_1,...,X_n)$.

X est une v.a. définie sur E ; on l'appelle le n-échantillon de X_0. X est caractérisé par le fait que les X_i sont de même loi (que X_0) et indépendantes.

La loi de X est entièrement déterminée par $P_{X_0,\theta}$ et par l'hypothèse d'indépendance. On la désigne par $P_{X,\theta}$. Pour les spécialistes, c'est le produit tensoriel de n mesures identiques à celle qui représente $P_{X_0,\theta}$. Notamment, si $P_{X_0,\theta}$ est représentée par une densité $f_{X_0}(x_0,\theta)$, alors X aura pour densité :

$$f_X(x,\theta) = \prod_{i=1}^{n} f_{X_0}(x_i,\theta)$$

Le modèle de l'échantillonnage est ainsi : $(E, \mathcal{B}, (P_{X,\theta})\ \theta \in \Theta)$.

L'estimation consiste donc à déterminer θ à partir d'une observation de X, qui est un n-uple de valeurs $x=(x_1,...,x_n) \in E$ obtenu par prélèvement de n éléments de $\mathcal{P}$. On le note avec des lettres minuscules.

La théorie de l'estimation peut être insérée dans un modèle plus vaste : le modèle de la décision qui s'applique aussi bien en théorie des tests d'hypothèses ou en théorie des jeux. Sa présentation est un peu lourde et n'est pas nécessaire ici.

---

2) v.a. : abréviation de variable aléatoire.

III - ECHANTILLONNAGE ET STATISTIQUE

A) LOI D'UN ECHANTILLON ET FONCTION DE VRAISEMBLANCE

Dans le cadre de la théorie de la mesure, $P_{X,\theta}$ est considérée comme une mesure sur E. Dans la pratique courante, il y a deux situations bien différentes :
- X est discrète et $P_{X,\theta}$ est alors donnée par la liste des probabilités élémentaires,
- X est continue et, le plus souvent, $P_{X,\theta}$ est donnée par sa densité par rapport à la mesure de Lebesgue (uniforme) sur E, muni de la tribu des boréliens.

Donnons des exemples :

* n-échantillon de Bernoulli

$$X_0 = \begin{cases} 1 & \text{avec la probabilité } p \\ 0 & \text{avec la proba } q = 1-p \end{cases}, \quad p \text{ inconnu}$$

alors $P_{X_0,p}$ est déterminée par les valeurs $p^{x_0}.q^{1-x_0}$ des probabilités associées aux valeurs x_0 de X_0.
$X = (X_1,...,X_n)$ de loi $P_{X,p}$ donne $f_X(x,p) = p^{\Sigma x_i}.q^{n-\Sigma x_i}$ pour la probabilité associée à la valeur $x = (x_1,...,x_n)$ de l'échantillon.

* n-échantillon de Poisson :

X_0 à valeurs dans **N** avec $P(X_0 = x_0) = \dfrac{\lambda^{x_0}.e^{-\lambda}}{x_0!}$, λ inconnu

La loi de X est donnée par les probabilités $f_X(x,\lambda) = \dfrac{\lambda^{\Sigma x_i}.e^{-n\lambda}}{\prod_{i=1}^{n}(x_i!)}$

* n-échantillon normal :

La v.a. X_0 de loi $\mathcal{N}(m, \sigma^2)$ a pour densité $f_{X_0}(x_0,\theta) = \dfrac{1}{\sqrt{2\pi}.\sigma}.e^{-\frac{1}{2}\left(\frac{x_0-m}{\sigma}\right)^2}$, $\theta = (m,\sigma)$ inconnu

La loi de l'échantillon X : $P_{X,\theta}$ est donnée par la densité : $f_X(x,\theta) = \dfrac{1}{(2\pi)^{n/2}\sigma^n}.e^{-\frac{1}{2}\Sigma\left(\frac{x_i-m}{\sigma}\right)^2}$

par rapport à la mesure de Lebesgue sur $\mathbf{R}^n$.

Aussi bien dans le cas discret que continu, les fonctions $f_X(x,\theta)$ associées à la loi $P_{X,\theta}$ sont appelées fonctions de vraisemblance du modèle.

B) RESUME STATISTIQUE ET LOI D'UNE STATISTIQUE

Dans la pratique, n peut être grand et la succession des valeurs observées de l'échantillon est lourde et peu lisible. Comme en statistique descriptive, on résume l'ensemble des valeurs observées par des paramètres caractérisant leur position et leur dispersion (statistiques d'ordre, de position : médiane, moyenne ; de dispersion : variance, écart-type ou autres résumés plus compliqués).

Cela revient à résumer l'information contenue dans X par un ou plusieurs nombres significatifs, calculés à partir des valeurs observées.

Pour décrire cela, on complète le modèle de l'échantillonnage par l'introduction d'une application $T : E \to \mathbf{R}$ (ou $\mathbf{R}^2$,...) qui, composée avec X, est alors une variable aléatoire :

$$E \xrightarrow[\text{aléatoire}]{X} E \xrightarrow[\text{statistique}]{T} \mathbf{R}$$ dont la loi sur **R** est $P_{T,\theta}$.

Les valeurs observées t de T sont en effet le fruit du tirage au hasard de l'échantillon.

D'où le modèle : $(\mathcal{P}^n, \mathcal{I}, P_\theta) \xrightarrow{X} (E, \mathcal{B}, P_{X,\theta}) \xrightarrow{T} (\mathbf{R}, \mathcal{B}_R, P_{T,\theta})$

modèle abstrait tirage de l'échantillon	modèle de la statistique	modèle de l'inférence

Ainsi l'observation t de T donnera des renseignements sur θ (i.e. la loi de X_0), à condition de connaître $P_{T,\theta}$ à partir de $P_{X,\theta}$ et T (mesure image). C'est tout le problème de la statistique inférentielle : savoir calculer $P_{T,\theta}$.

T est appelée en général une statistique. Dans le cadre de l'estimation [$E = E_o^n$ et $X = (X_1,...,X_n)$], on dit aussi que T est un estimateur de θ.

c) Statistiques usuelles $\bar{X}$ et S^2

Dans l'étude des séries statistiques issues d'un caractère quantitatif, on a vu le rôle important joué par la moyenne $m = {}^1/_N \Sigma x_i$ de la population comme paramètre de position, et par la variance $\sigma^2 = {}^1/_N \Sigma(x_i - m)^2$ comme paramètre de dispersion (où N est la taille de la population).

En probabilités, on constate que lorsque l'on connaît le type de loi d'une v.a., dans les cas les plus fréquents, le paramètre qui détermine complètement cette loi est connu dès que l'on connaît son espérance mathématique (Bernoulli, binomiale, Poisson, Pascal, hypergéométrique, exponentielle, Γ) ou en outre son écart-type (uniforme, normale).

Pour connaître la loi parente de X_0, si elle est de ces types, il suffit donc de déterminer des valeurs assez précises pour $m = E(X_0)$ et $\sigma^2 = Var(X_0)$.

A partir d'un échantillon X, il est naturel de regarder sur X les valeurs observées des moyenne et variance empiriques pour en inférer les valeurs (ou un encadrement) de ces paramètres pour la population. La loi des grands nombres nous garantira une bonne probabilité pour qu'avec n assez grand, on ne soit pas trop loin des bons résultats.

C'est ainsi que se pose le problème de l'estimation. D'où l'intérêt d'introduire les statistiques T(X) suivantes, définies sur l'échantillon :

$$\bar{X} = \frac{1}{n}\sum X_i \qquad\qquad \Sigma^2 = \frac{1}{n}\sum (X_i - m)^2$$

La 2ème est souvent peu utile, car m est généralement inconnu ; on ne peut donc calculer la valeur observée. On cherche alors à remplacer m par une valeur approchée.

Or $\bar{X}$ est considéré (on le verra plus loin) comme un "bon" estimateur de m. En prenant la valeur observée $\bar{x}$ à la place de celle de m, on obtient la statistique :

$$S'^2 = \frac{1}{n}\sum (X_i - \bar{X})^2 \quad \text{qui en réalité est moins "bonne" que} \quad S^2 = \frac{1}{n-1}\sum (X_i - \bar{X})^2$$

Ces deux statistiques $\bar{X}$ et S^2 font l'objet de nombreux résultats intéressants. On ne va pas les démontrer ici, on se reportera à la bibliographie [7, p.266].

Pour $\bar{X}$

* $E(\bar{X}) = m, \quad Var(\bar{X}) = \sigma^2/n$
* $\bar{X} \to m$, en proba (loi faible des grands nombres)
* $\bar{X} \to m$, p.s. (loi forte)
* $\dfrac{\bar{X} - m}{\sigma/\sqrt{n}} \xrightarrow{\text{loi}} U \in \mathcal{N}(0,1)$ (théorème central-limit)

On utilisera l'approximation en loi $\bar{X} \approx \mathcal{N}(m, \sigma^2/n)$ dès que n>30 ([3]). Cette dernière propriété permet de calculer les probabilités pour que $\bar{X}$ s'écarte de m de plus d'une valeur donnée. Cela permet donc d'établir les fourchettes d'encadrement pour m avec leur fiabilité, à la base de la détermination des intervalles de confiance.

Pour S^2

* $E(S^2) = \sigma^2$ (raison pour la préférer à S'^2, car $E(S'^2) = \sigma^2 - \sigma^2/n$)

* $\mathrm{Var}(S^2) = \frac{\mu_4 - \sigma^4}{n} + \frac{2\sigma^4}{n(n-1)}$, μ_4 moment centré d'ordre 4 de la loi de X_0

* $\mathrm{cov}(\bar{X}, S^2) = \frac{\mu_3}{n}$ (non corrélées si $\mu_3 = 0$, de toute façon asymptotiquement non corrélées)

* $S^2 \xrightarrow{p.s.} \sigma^2$ (loi forte des grandes nombres)

$\frac{S^2 - \sigma^2}{\sqrt{\mu_4 - \sigma^4}}\sqrt{n} \xrightarrow{\text{loi}} U \in \mathcal{N}(0,1)$ (application du théorème central-limit et autres théorèmes de convergence)

Dans la pratique, avec une erreur uniforme sur les probabilités calculées inférieure à 10^{-3} ([4]), on fait les approximations suivantes :

dès que n > 30, $\bar{X} \approx \mathcal{N}(m, \sigma^2/n)$

dès que n > 50, $S^2 \approx \mathcal{N}(\sigma^2, \frac{\mu_4 - \sigma^4}{n})$.

Cas particulier d'un échantillon gaussien

Si $X_0 \sim \mathcal{N}(m, \sigma^2)$, alors :

* $\bar{X} \sim \mathcal{N}(m, \sigma^2/n)$ exactement pour tout n
* $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ (loi du χ^2 à n-1 degrés de liberté)
* $\bar{X}$ et S^2 sont indépendantes
* $\frac{\bar{X} - m}{S} \sim T_{n-1}$ (T_{n-1} de Student à n-1 degrés de liberté)

Cette dernière statistique présente l'intérêt de ne pas faire intervenir σ.

Mais pour n > 50, $T_{n-1} \sim \mathcal{N}(0,1)$, d'où son utilisation essentiellement pour les "petits" échantillons.

IV- ESTIMATEURS

A) CADRE DE L'ESTIMATION

L'estimation consiste à donner des valeurs approchées aux paramètres d'une population (m, σ,...) à l'aide de l'observation d'un n-échantillon d'une variable parente X_0.

On s'intéresse de préférence aux valeurs d'un caractère représenté par X_0 : sa moyenne m, sa variance σ^2, la proportion p d'objets d'un certain type A dans la population, et on introduit les statistiques

---

3) si la loi de X_0 n'est pas trop dissymétrique, n>50 ou plus pour une loi fortement dissymétrique.

4) loi assez régulière.

$\bar{X}$, S^2 et F (où F est la fréquence du caractère dans l'échantillon) ayant les propriétés d'estimateurs convergents :

$$\bar{X} \xrightarrow{p.s.} m, \quad S^2 \xrightarrow{p.s.} \sigma^2, \quad F \xrightarrow{p.s.} p$$

On a déjà étudié $\bar{X}$ et S^2 ; remarquons que F est un cas particulier de $\bar{X}$:
F est la fréquence empirique de la réalisation par X_0 d'une qualité donnée : situation de Bernoulli,

$$X_o = \begin{cases} 1 & \text{si qualité obtenue} \\ 0 & \text{sinon} \end{cases}, \quad \text{(avec la probabilité p)}$$

Alors $X = (X_1,...,X_n)$ est une "bernoullade" (5) et $F = {}^1/_n \Sigma X_i$ désigne la fréquence du type A dans l'échantillon. Donc $F = \bar{X}$. La loi de F est connue : $n.F \sim B(n,p)$. Le paramètre à estimer est p :

On a $E(X_0) = p$, $var(X_0) = p.q$, $E(F) = p$, $var(F) = \frac{pq}{n}$

la loi de X est donnée par les probabilités des valeurs $x = (x_1,...,x_n)$: $p^{\Sigma x_i}.q^{n-\Sigma x_i}$
la loi de F est : $P(F = f) = P(\Sigma X_i = nf) = \Sigma_x P(X=(x_1,...,x_n) / \Sigma x_i = nf)$

or les x qui vérifient $\Sigma x_i = nf$ sont au nombre de C_n^{nf} parmi les 2^n valeurs possibles,

d'où $P(F = f) = C_n^{nf} p^{nf}.q^{n(1-f)}$

$f_F(f,p) = C_n^{nf} p^{nf}.q^{n(1-f)}$ est la fonction de vraisemblance du modèle qui permet d'estimer une proportion p.

B) QUALITES D'UN ESTIMATEUR

Si T est susceptible d'estimer θ, on espère que la valeur t observée sera "proche de θ".

Il y a deux manières d'être proche :

1- <u>en moyenne</u> : si on réalise de nombreux échantillonnages, les valeurs de t seront réparties autour de θ. Cela se traduit dans le modèle par la propriété : $E_\theta(T) = \theta$; E_θ parce que cette espérance est calculée sur la base de la loi de T qui dépend de θ.

Cette propriété est celle d'un estimateur <u>sans biais</u> (e.s.b., c'est le cas des précédents).

Par exemple $E_\sigma(S'^2) = \sigma^2 - \sigma^2/n$. Le biais de S'^2 est σ^2/n, ce n'est pas rédhibitoire ; par une homothétie, on peut toujours rendre sans biais un estimateur, à condition de savoir calculer son biais.

2- <u>asymptotiquement</u> : plus n est grand (plus on paye cher l'échantillonnage), plus on se rapproche de θ. En probabilité, c'est le sens le plus opératoire, presque sûrement, c'est le plus fort.

Cela se traduit par : $T_n \xrightarrow[n\to\infty]{} \theta$, en proba.

C'est la propriété d'être <u>convergent</u> pour T_n, et pour n assez grand, de minimiser la probabilité : $P_\theta(|T_n - \theta| > \varepsilon)$. Le n assez grand et la valeur de cette probabilité supposent de connaître la loi de T_n.

Les estimateurs précédents sont convergents ; si T_n est sans biais et si $var(T_n) \to 0$, alors T_n est convergent (inégalité de Bienaymé-Tchebychev).

3- <u>Efficacité, limite théorique des performances d'un estimateur</u>

Pour minimiser la probabilité précédente, on cherche les estimateurs les plus précis, qui sont donc les moins dispersés possibles autour de la valeur θ, pour un n donné, ceci en moyenne. Cela sera traduit par la recherche d'un minimum pour l'espérance $E_\theta(|T - \theta|^2)$ (= $var_\theta(T)$ si T est un e.s.b. de θ en dim. 1).

Cette espérance est appelée risque quadratique $R(\theta,T)$ de l'estimateur T.

Lorsque T réalise ce minimum, il est dit "admissible" (il n'est pas forcément sans biais).

Un "bon" estimateur sera alors un e.s.b. convergent de variance minimale.

---

5) c'est-à-dire un n-échantillon de Bernoulli !

Cette variance minimale ne peut être aussi petite que l'on veut, il y a une borne inférieure donnée par un théorème important de la statistique, le théorème de Cramer-Rao-Fréchet-Darmois :

Si $f_X(x,\theta)$ est la fonction de vraisemblance du modèle (existence supposée), avec certaines hypothèses de régularité, en posant :

$$I_n(\theta) = E_\theta\left[\left(\frac{\partial \ln f_X(x,\theta)}{\partial \theta}\right)^2\right]$$ si cette espérance existe (quantité d'information de Fisher),

alors, si T est un e.s.b. de θ défini sur X dont la variance existe, on a :

$$Var_\theta(T) \geq \frac{1}{I_n(\theta)}$$ (borne inférieure de Rao-Cramer).

Si T réalise l'égalité, c'est l'e.s.b. le meilleur, il est dit efficace. C'est le cas de $\bar{X}$, F, et de Σ^2 quand m est connu.

C) RECHERCHE D'UN ESTIMATEUR : METHODE DU MAXIMUM DE VRAISEMBLANCE

On peut penser que pour les exemples que nous avons pris, nous avons les meilleurs estimateurs.

Dans le cas général, il y a des critères pour vérifier qu'un tel estimateur existe et que l'on a le meilleur estimateur (Th. de Rao-Blackwell et Lehmann-Scheffé dans le cas de statistiques exhaustives, c'est-à-dire conservant l'information contenue dans X, et calculable quand le modèle est de type exponentiel) (Saporta [7, p.291]).

Empiriquement, il y a une méthode qui conduit le plus souvent à l'expression d'un bon estimateur : la méthode du maximum de vraisemblance.

Elle consiste à prendre comme estimation de θ la valeur $\hat{\theta}$ qui rend maximale la fonction de vraisemblance $f_X(x,\theta)$.

Cela revient à supposer que l'échantillon observé était le plus "probable" puisque f_X désigne une densité de probabilité. C'est explicite dans le cas discret où f_X est la valeur des probabilités élémentaires, c'est intuitif dans le cas continu où, pour un encadrement donné ε de cette valeur de θ, on a le maximum de la probabilité de s'y trouver lorsque $\hat{\theta}$ réalise le maximum de f_X.

Cela procède d'une hypothèse empirique : l'événement observé est le plus probable (on a plus de chance de le voir que les autres), ce qui est une hypothèse hardie...

Avec cette méthode, la détermination de $\hat{\theta}$ est du domaine mathématique, à partir du problème :

$$\forall\, x \in E,\ f_X(x, \hat{\theta}) = \sup_{\theta \in \Theta} f_X(x, \theta),$$

A chaque $x \in E$ on obtient (peut-être) une valeur pour $\hat{\theta}$ qui est alors une fonction de x, donc une statistique définie sur E , l'estimateur du maximum de vraisemblance ; pratiquement :

- ou on trouve $\hat{\theta}(x)$ facilement, vu la tête de f_X
- ou on a recours à l'analyse et aux régularités de f_X.

Mais en échantillonnage, $f_X(x,\theta) = \prod_{i=1}^{n} f_{X_o}(x_i,\theta)$; or maximiser un produit de termes positifs revient à maximiser son logarithme ; ceci revient à poser que $\hat{\theta}$ satisfait

$\frac{d}{d\theta} \ln f_X(x,\theta) = 0$, en vérifiant qu'on a bien un maximum ; cela s'écrit : $\sum_{i=1}^{n} \frac{d}{d\theta} \ln f_{X_o}(x_i,\theta) = 0$

[équation de vraisemblance] qui est une fonction implicite de θ en x dont la solution est $\hat{\theta}(x)$.

exemple - loi de Bernoulli : estimation de p.

$X_o \sim \mathcal{B}(1,p)$, $f_X(x,p) = p^{\Sigma x_i}.(1-p)^{n-\Sigma x_i}$

équation de vraisemblance : $\frac{d}{dp}\ln f_X(x,p) = \frac{\Sigma x_i}{p} - \frac{n-\Sigma x_i}{1-p} = 0$

solution : $np - \Sigma x_i = 0$, $p = \frac{\Sigma x_i}{n}$, avec $\frac{d^2}{dp^2}\ln f_X(x,p) = -\frac{\Sigma x_i}{p^2} - \frac{n-\Sigma x_i}{(1-p)^2} < 0$.

On a bien un maximum sur]0, 1[, d'où l'estimateur du maximum de vraisemblance : $\hat{p} = \bar{X}$.

exemple - loi normale. Dans ce cas, θ est à plusieurs dimensions : $\theta = (m,\sigma^2)$, on obtient par le même principe un système d'équations aux dérivées partielles. Traitons l'exemple :

$X_o \sim \mathcal{N}(m, \sigma^2)$, $f_X(x,\theta) = \frac{1}{(2\pi)^{n/2}\sigma^n}.e^{-\frac{1}{2}\Sigma\left(\frac{x_i-m}{\sigma}\right)^2}$, avec un extremum local pour f_X si

$$\frac{\partial}{\partial m}\ln f_X(x,\theta) = 0 \quad \text{et} \quad \frac{\partial}{\partial(\sigma^2)}\ln f_X(x,\theta) = 0$$

La première équation donne : $\Sigma\frac{x_i - m}{\sigma^2} = 0$, d'où $m = \frac{\Sigma x_i}{n}$ et $\hat{m} = \bar{X}$.

La deuxième donne : $-\frac{n}{2}\times\frac{1}{\sigma^2} - \frac{1}{2}\Sigma\frac{(x_i-m)^2}{\sigma^4} = 0$, d'où $\sigma^2 = 1/n\ \Sigma(x_i - \bar{x})^2$, et $\hat{\sigma^2} = S'^2$ (biaisé).

Cela, si on vérifie qu'on a bien un maximum (local), condition donnée pour une fonction de 2 variables par $rt - s^2 > 0$. Ici, le calcul des dérivées partielles secondes avec m et σ^2 et l'expression donne bien

$$rt - s^2 = -\frac{n^2}{2\sigma^6} + \frac{n^2}{\sigma^6} > 0$$

V- ESTIMATION PAR INTERVALLE

A) POSITION DU PROBLEME

Fournir une valeur numérique pour un paramètre à estimer (estimation ponctuelle) n'est pas satisfaisant en pratique, où il faut pouvoir quantifier les risques pris en retenant cette valeur. En effet, cela ne fournit ni la marge d'erreur : la précision de la donnée numérique, ni la probabilité que l'échantillon prélevé conduise effectivement à une "bonne" valeur : la fiabilité du procédé (on pourrait être très malchanceux avec l'échantillon !).

D'où le problème de l'estimation par intervalle ainsi posé pour le paramètre θ en dimension 1 :

Trouver un intervalle [a(X) ; b(X)] dont les bornes sont aléatoires puisque déterminées par l'échantillon X et tel que $P_\theta(\theta \in [a(X) ; b(X)]) \geq 1 - \alpha$.

* l'échantillon étant alors prélevé, son observation x fournit l'intervalle réel [a(x) ; b(x)] appelé "fourchette" pour l'estimation de θ, sans qu'on puisse en toute certitude dire si la valeur réelle θ_0 du paramètre θ pour la population $\mathcal{P}$ est entre a(x) et b(x).
* $1 - \alpha$ est appelé le niveau de confiance (par ex. 0,95 ou 95%), indice de fiabilité du résultat et α est le risque (de tomber à côté !).

* la probabilité P_θ qu'on cherchera à rendre la plus voisine de 1- α (pour des raisons d'économie), est calculable si on connaît la loi de X, ou plutôt de T statistique servant à déterminer les bornes a(X) et b(X).
* l'intervalle est d'autant plus étroit que T est moins dispersée, d'où l'intérêt d'avoir des estimateurs sans biais le plus efficaces possibles, de loi connue et convergents "rapidement" pour minimiser la taille n de l'échantillon nécessaire pour réaliser le niveau 1- α.
* élargir l'intervalle de confiance (perdre en précision), c'est augmenter P_θ et donc se permettre d'atteindre le niveau de confiance, c'est en fait diminuer le risque α (gagner en fiabilité). On sera toujours à la recherche du compromis entre précision et fiabilité.
* la valeur de α, donnée a priori (à moins que ce soit n), est déterminée par les coûts que représentent les fausses informations (cas où θ_0 n'est pas dans la fourchette). Ces coûts sont calculables sur un grand nombre d'estimations liées à une pratique industrielle et commerciale. Sans les contraintes financières, la détermination d'un intervalle de confiance n'a d'intérêt que qualitatif lorsque des valeurs standards de α sont données (en médecine ou sciences humaines par exemple), on prend alors souvent $\alpha = 0{,}05$.

B) PRINCIPE DE LA DETERMINATION D'UN INTERVALLE DE CONFIANCE

La solution du problème précédent où on doit déterminer a(X) et b(X) n'est pas unique : 1 relation pour 2 inconnues. Les conditions du problème restreignent cette indétermination.

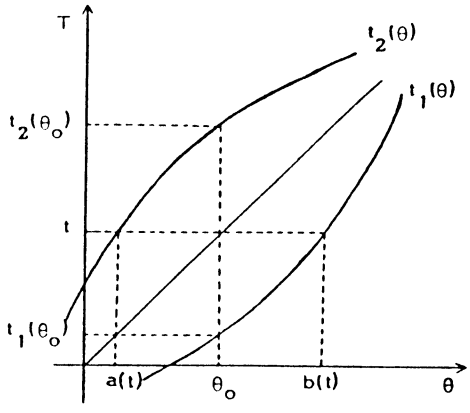
On peut chercher • un maximum de confiance : $P_\theta(\theta \le M(X)) = 1 - \alpha$

• ou un minimum : $P_\theta(m(X) \le \theta) = 1 - \alpha$

• ou penser qu'il n'y a pas de raison pour que le risque soit plus porté à droite qu'à gauche de l'intervalle et équilibrer ce dernier par les conditions : $P_\theta(\theta \le a(X)) = \alpha/2$ et $P_\theta(\theta \ge b(X)) = \alpha/2$.

Plaçons nous dans cette situation.

Si on a un "bon" estimateur T de θ et si, pour chaque valeur de θ, on trouve deux réels t_1 et t_2 tels que $P_\theta(t_1 < T < t_2) = 1 - \alpha$, on peut penser que la valeur θ_0 de θ est proche de la valeur observée t de T et considérer que si $t \in [t_1, t_2]$ cette valeur est possible pour θ_0. Cela se traduit par le graphique



d'où la détermination
$a(t) = t_2^{-1}(t)$
$b(t) = t_1^{-1}(t)$
si c'est possible

alors $t_1(\theta_0) < T(X) < t_2(\theta_0)$
équivaut à
$t_2^{-1}(T) < \theta_0 < t_1^{-1}(T)$
de probabilité 1- α

En théorie, il faudrait admettre que les fonctions $t_1(\theta)$ et $t_2(\theta)$ (non uniques) sont croissantes et inversibles localement. On préfère résoudre directement ce problème d'inversion d'inégalités dans chaque cas, comme nous allons le voir sur un exemple.

choix d'une statistique de décision.

En réalité, l'estimateur T ne contient pas explicitement le paramètre θ qui intervient dans le calcul des bornes t_1 et t_2 par l'intermédiaire de la loi P_θ, inconnue.

On transforme alors cette statistique en une statistique contenant θ mais dont la loi n'en dépend plus et est bien connue, permettant de calculer la probabilité demandée. Cette dernière statistique sera la variable de confiance.

Exemple : estimation de la moyenne d'une loi normale avec écart-type connu.

$X_0 \sim \mathcal{N}(m, \sigma^2)$; $\bar{X} \sim \mathcal{N}(m, \sigma^2/n)$ est le meilleur estimateur de m, mais sa loi contient m.

On prend alors $\dfrac{X - m}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ qui sera la variable de confiance adéquate.

C) MISE EN PRATIQUE DE LA RECHERCHE DE L'INTERVALLE DE CONFIANCE

Il faudra l'adapter à chaque cas d'espèce, d'où un ensemble de résultats particuliers donnés dans les manuels [8, p.149] ou [7, p.304].

Reprenons l'exemple précédent ; on part de la condition

$P_m(a(X) < m < b(X)) = 1 - \alpha$ équivalente par transformations affines à une condition du type :

$P_m(-u_{\alpha/2} < U < u_{\alpha/2}) = 1 - \alpha$, où $u_{\alpha/2} > 0$ est le fractile d'ordre $\alpha/2$ de la loi normale centrée réduite, défini par $P[U > u_{\alpha/2}] = \alpha/2$ (lu dans la table : $\alpha = 0{,}05$ donne $u_{\alpha/2} = 1{,}96$).

On a bien $-u_{\alpha/2} < U < u_{\alpha/2} \Leftrightarrow -u_{\alpha/2}\,\sigma/\sqrt{n} < \bar{X} - m < u_{\alpha/2}\,\sigma/\sqrt{n}$

$\Leftrightarrow \bar{X} - u_{\alpha/2}\,\sigma/\sqrt{n} < m < \bar{X} + u_{\alpha/2}\,\sigma/\sqrt{n}$

d'où l'intervalle de confiance équilibré :

$P[\bar{X} - u_{\alpha/2}\,\sigma/\sqrt{n} < m < \bar{X} + u_{\alpha/2}\,\sigma/\sqrt{n}] = 1 - \alpha$

et la fourchette obtenue avec l'observation $\bar{x}$ de $\bar{X}$: $(\bar{x} - u_{\alpha/2}\,\sigma/\sqrt{n}\ ;\ \bar{x} + u_{\alpha/2}\,\sigma/\sqrt{n})$,

avec les remarques d'usage :

- La fiabilité $1 - \alpha$ influe sur la valeur $u_{\alpha/2}$, ici 1,96 si $\alpha = 0{,}05$; 1,65 si $\alpha = 0{,}1$, donc peu influente devant les variations de n (l'écart de la fourchette dû à un choix de fiabilité raisonnable peut varier du simple au double) ;

- l'écart de la fourchette est proportionnel à σ, écart-type de la population, qui trouve ici son sens concret. On retrouve le fait que la fourchette est d'autant resserrée que la dispersion de X_0 est faible ;

- l'écart (la précision) est inversement proportionnel à $\sqrt{n}$: il faut multiplier par 4 le nombre d'observations pour réduire de moitié l'intervalle de confiance. Ce sera le cas général de l'estimation d'une moyenne du fait que $\mathrm{Var}(\bar{X}) = \sigma^2/n$;

- numériquement, si X_0 est une mesure en cm autour de 20 cm et $\sigma = 1$ cm (les 2/3 des mesures tombent entre 19 et 21 cm), avec $\alpha = 0{,}05$ et $n = 49$, l'intervalle observé est pour m : [19,7 ; 20,3].

Dans cet exemple, avec l'hypothèse forte $X_0 \sim \mathcal{N}(m, \sigma^2)$, on voit qu'un échantillon modeste de taille 50, donne un intervalle appréciable de longueur 0,6 pour les données numériques choisies.

Dans l'estimation d'une proportion où l'on ne connaît pas la loi de la variable qui conduit à attribuer telle ou telle qualité au caractère, on verra que la taille de l'échantillon doit être bien plus grande pour estimer avec une bonne fiabilité cette proportion (n = 1000 dans les sondages, cf. l'atelier).

D) FORMULATION DE LA CONCLUSION

Revenons à l'exemple.

Mathématiquement, la réponse est : l'intervalle de confiance pour estimer m au niveau de confiance $1 - \alpha$ est : $[\bar{X} - u_{\alpha/2}\,\sigma/\sqrt{n}\ ;\ \bar{X} + u_{\alpha/2}\,\sigma/\sqrt{n}]$.

Cette formulation ne peut satisfaire l'utilisateur qui doit prendre des décisions. De plus il aimerait avoir de bonnes valeurs numériques justifiées par l'échantillon qu'il a payé assez cher.

On retombe sur un problème épistémologique :

- ou on ne fera ce prélèvement qu'une seule fois, et mon penchant fréquentiste a de la difficulté à donner du sens opératoire à la notion de probabilité (que θ soit effectivement dans l'intervalle retenu), car du point de vue numérique, θ est, ou n'est pas, dans l'intervalle observé ; il n'y a plus, après cette observation, d'expérience aléatoire, donc de probabilité ;

- ou il fait partie d'un contrôle régulier, répété un assez grand nombre de fois, et je préfère formuler le résultat en les termes suivants :
au niveau de confiance $1 - \alpha = 0{,}95$ (par exemple), ma méthode mathématique conduit à un résultat

$(m \in [\ \bar{x} - u_{\alpha/2}\ \sigma/\sqrt{n}\ ;\ \bar{x} + u_{\alpha/2}\ \sigma/\sqrt{n}\ ])$ vrai en moyenne 95 fois sur 100 (mais je ne sais pas quand je me trompe). Vous pourrez donc me faire relativement confiance.

Pour vous, utilisateur, vous pouvez savoir que si vous utilisez un grand nombre de fois une telle estimation de m (tous les mois par exemple) dans votre contrôle de production, sachez que 5 fois sur 100 en moyenne l'intervalle observé (la fourchette) ne contiendra pas la vraie valeur de m. Vous pourrez alors évaluer les coûts que cela représente pour vous, prendre vos dispositions, et revoir peut-être le niveau de confiance que vous m'avez donné, en fonction de la taille n de l'échantillon que vous acceptez de payer.

BIBLIOGRAPHIE

[1] LAPLACE Pierre Simon : *Essai philosophique sur les probabilités* (1825) Editions Ch. BOURGEOIS, 1986.

[2] POINCARE Henri : *La Science et l'hypothèse* (1902) Editions CHAMPS-FLAMMARION, 1968.

[3] POINCARE Henri : *Calcul des probabilités* (1912), réédition J. GABAY, 1987.

[4] EKELAND Ivar : *Au hasard*, Editions du SEUIL, 1990.

[5] BERNOULLI Jacques : *Ars Conjectandi* (1713), traduction de N. Meusnier, Brochure IREM de Rouen, 1987.

[6] INTER-IREM : *Histoire et épistémologie : Maths au fil des âges* Editions GAUTHIER-VILLARS, 1987.

[7] SAPORTA Gilbert : *Probabilités, Analyse des données et Statistique* Editions TECHNIP, 1990.

[8] BIGOT Bernard et VERLANT Bernard : *Mathématiques, statistiques et probabilités*, cours de BTS Editions FOUCHER, 1990.

[9] HENRY Michel et Annie : L'enseignement des probabilités dans le nouveau programme de Première, in *Repères-IREM*, n°6, 1992.

[10] IREM de BESANÇON : *L'enseignement des statistiques et des probabilités en STS* Brochure IREM de Besançon, 1990.

Atelier

Pratique de la recherche d'un intervalle de confiance pour l'estimation d'une proportion p. Application aux sondages

---

Modelisation

Dans la population P, une proportion p des individus possède le caractère observé. On cherche un intervalle de confiance pour p.
On fait un tirage avec remise d'un n-échantillon (ou sans remise si la population est assez vaste et si l'échantillonnage ne modifie pas p).

Soit F la proportion (ou fréquence) trouvée dans le n-échantillon des individus ayant le caractère examiné. F est un estimateur sans biais, efficace de p (exercice 33).

On a $nF \sim \mathcal{B}(n, p)$.

L'expression de cette loi est trop compliquée pour donner un intervalle de confiance explicite bâti sur cette variable.

a) Petits échantillons :

On utilise une table de la loi binômiale pour déterminer pour différents p les valeurs $k_1(p)$ et $k_2(p)$ telles que :

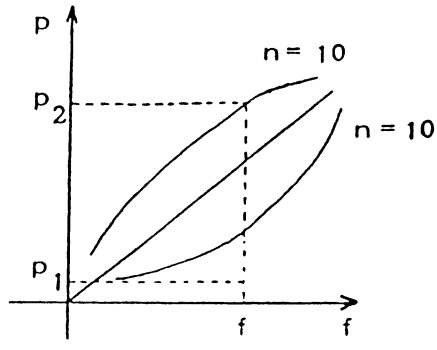
$$P\left(k_1 < nF < k_2\right) = \sum_{k=k_1(p)}^{k_2(p)} C_n^k \, p^k (1-p)^{n-k} = 1-\alpha$$

(avec par exemple $\sum_{k=0}^{k_1} C_n^k \, p^k (1-p)^{n-k} = \frac{\alpha}{2}$).

L'intervalle de confiance pour p (ou plutôt une de ses réalisations) sera déterminé par l'ensemble des p tels que si nf est l'observation issue de l'échantillon, on a $k_1(p) < nf < k_2(p)$ selon la présentation théorique du début.

Pour éviter de nombreux calculs fastidieux, on utilise des abaques de la loi binômiale construites à cet effet.

Une telle abaque est un réseau de courbes. Chaque courbe corresponc à une taille d'échantillon. Elle donne les bornes p_1, p_2 de l'intervalle de confiance pour p en fonction de l'observation f, selon le schéma suivant :



b) Grands échantillons (tirage avec remise ou ne modifiant pas sensiblement p) :

Une nouvelle application du théorème central limite montre que l'on a l'approximation :

$$nF \approx \eta(np, \sqrt{np(1-p)})$$

d'où

$$F \approx \eta\left(p, \sqrt{\frac{p(1-p)}{n}}\right) \text{ et } \frac{F-p}{\sqrt{\frac{p(1-p)}{n}}} \simeq U \sim \eta(0,1)$$

On obtient, comme pour le 1er paragraphe, l'intervalle de confiance

$$(*) \qquad F - u_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} < p < F + u_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

mais p figure dans les bornes et il faut résoudre en p cette double inégalité. On a trois solutions :

α - Utiliser un agrandissement de l'intervalle de confiance par la majoration $p(1-p) < \frac{1}{4}$ (car $0 < p < 1$) d'où l'intervalle

$$\boxed{F - \frac{u_{\alpha/2}}{2\sqrt{n}} < p < F + \frac{u_{\alpha/2}}{2\sqrt{n}}}$$

qui suppose en fait que p est voisin de $\frac{1}{2}$ (mais on sait que pour que l'approximation normale soit valable, il ne faut pas que p soit trop proche de 0 ou de 1). On a alors un intervalle de confiance de niveau supérieur à $1-\alpha$, mais on ne connaît pas le niveau exact.

β - Faire une résolution graphique de (*)

Les bornes de l'encadrement (*) de la variable de confiance sont

$$f = p \pm u_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

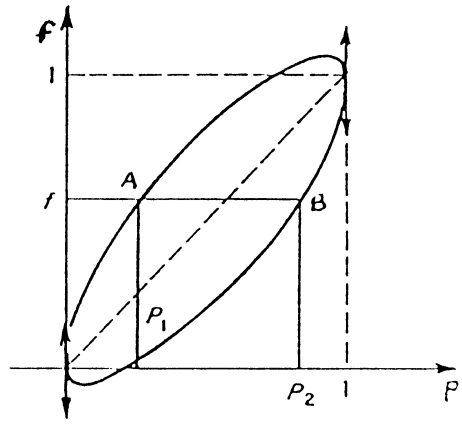
d'où

$$(f-p)^2 = u^2_{\frac{\alpha}{2}} \frac{p(1-p)}{n}$$

d'où l'équation

$$f^2 + p^2\left(1 + \frac{u^2_{\alpha/2}}{n}\right) - 2pf - \frac{u^2_{\alpha/2}\, p}{n} = 0$$

d'une ellipse passant par l'origine et le point (1, 1)



Les points intérieurs à l'ellipse vérifient les inégalités *.

Pour chaque observation f de F, on obtient donc (comme dans le cas des petits échantillons) un couple $p_1(f)$ et $p_2(f)$ et l'intervalle [A, B] représente aussi bien l'évènement $F - u_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < p < F + u_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$ que $p_1(F) < p < p_2(F)$ de probabilité $1-\alpha$.

Lorsque n (grand) varie, on obtient aussi une abaque constituée d'ellipses dont l'utilisation est analogue à celle des petits échantillons.

Lorsque $n \to +\infty$, les ellipses se rétrécissent et tendent vers la diagonale.

γ - Utiliser F comme estimation ponctuelle de p, on obtient alors l'intervalle de confiance

$$\boxed{F - u_{\frac{\alpha}{2}}\sqrt{\frac{F(1-F)}{n}} < p < F + u_{\frac{\alpha}{2}}\sqrt{\frac{F(1-F)}{n}}}$$

de niveau $1-\alpha$.

[Ceci est justifié par le calcul suivant : résoudre en p l'équation de l'ellipse donnée

$$p = \frac{\left(2f + \frac{u^2_{\alpha/2}}{n}\right) \pm \sqrt{\frac{u^4_{\alpha/2}}{n^2} + 4f\,\frac{u^2_{\alpha/2}}{n} - 4f^2\,\frac{u^2_{\alpha/2}}{n}}}{2\left(1 + \frac{u^2_{\alpha/2}}{n}\right)}$$

Comme n est grand, on prend plutôt un équivalent, ce qui donne

$$\frac{2f + \frac{u^2_{\alpha/2}}{n}}{2\left(1 + \frac{u^2_{\alpha/2}}{n}\right)} \sim f$$

et

$$\frac{\sqrt{\frac{u^4_{\alpha/2}}{n^2} + 4f\,\frac{u^2_{\alpha/2}}{n} - 4f^2\,\frac{u^2_{\alpha/2}}{n}}}{2\left(1 + \frac{u^2_{\alpha/2}}{n}\right)} \sim \frac{u_{\alpha/2}}{\sqrt{n}}\sqrt{f(1-f)}$$

d'où $p \sim f \pm u_{\alpha/2}\sqrt{\frac{f(1-f)}{n}}$.

Exercice

1) Un encadrement de confiance de p est exigé à $\pm$ 0,01 ("fourchette" à 2 % pour un sondage par exemple) avec un niveau $1-\alpha = 95$ %.
On sait que la valeur f observée sera voisine de 0,5 (cas le plus défavorable où $\sqrt{f(1-f)}$ atteint son maximum $\frac{1}{4}$).
Déterminer la taille de l'échantillon nécessaire (Rep. : 9600).

2) Avec un échantillon de 1000 personnes et une fourchette à 2 %, quel est le niveau de confiance du sondage ? (Rep. : 0,46).

c) Grands échantillons exhaustifs (taille de la population = N) :

F est encore un estimateur sans biais de p, on a $\text{var } F = \frac{p(1-p)}{n}\,\frac{N-n}{N-1}$.

On est alors ramené à l'intervalle de confiance précédent, de la forme

$$F - u_{\frac{\alpha}{2}}\sqrt{\frac{F(1-F)(N-n)}{n(N-1)}} < p < F + u_{\frac{\alpha}{2}}\sqrt{\frac{F(1-F)(N-n)}{n(N-1)}}$$

Exercice

Construire un intervalle de confiance à 10 % pour le paramètre p de Bernoulli si une observation d'un échantillon de taille 50 donne $\Sigma x_i = 15$. (Utiliser l'approximation normale et comparer les résultats dans les deux cas où p(1-p) est remplacé par $\frac{1}{4}$ ou par son estimation tirée de l'échantillon).

Exercice

1) Lancer une pièce 50 fois et compter le nombre de piles. Déterminer alors un intervalle de confiance pour p à 90 %.

2) Lancer un dé 15 fois et compter le nombre de 6. Déterminer un intervalle de confiance pour la probabilité d'obtenir 6 avec ce dé (utiliser une abaque). Votre dé est-il pipé ?

Exercice

A la veille d'une consultation électorale, on a interrogé 100 électeurs pris au hasard. 64 d'entre eux se sont déclarés favorables au candidat z. Entre quelles limites, au moment du sondage, au niveau 0, 95, la proportion du corps électoral favorable à z se situe-t-elle ?

Exercice

Un médecin désire estimer la proportion des cas qui seront guéris par un nouveau traitement.

a) A combien de patients doit-il appliquer le traitement avant de pouvoir conclure, s'il veut que son estimateur ait un écart type de 0, 005 et s'il pense que le traitement guérira à peu près 75 % des malades.

b) Lors de l'expérience pratiquée en a), quelle est approximativement la probabilité pour que l'estimateur excède 0, 8 alors que, en réalité, le traitement ne soigne que 60 % des cas ?

Voici quelques éléments bibliographiques pour les applications aux sondages:

HENRY Michel: Eléments sur les sondages, cours de statistiques, Maîtrise SMI, Besançon, 1983.

KOSMANEK Edith: Sondages stratifiés, article de "Quadrature n°6, Septembre 1990.

GENET, PUPION et REPUSSARD: Probabilités, statistiques et sondages, cours et exercices corrigés, Vuibert 1974.

SAPORTA Gilbert: Probabilités, analyse des données et statistiques, éditions Technip, 1992.

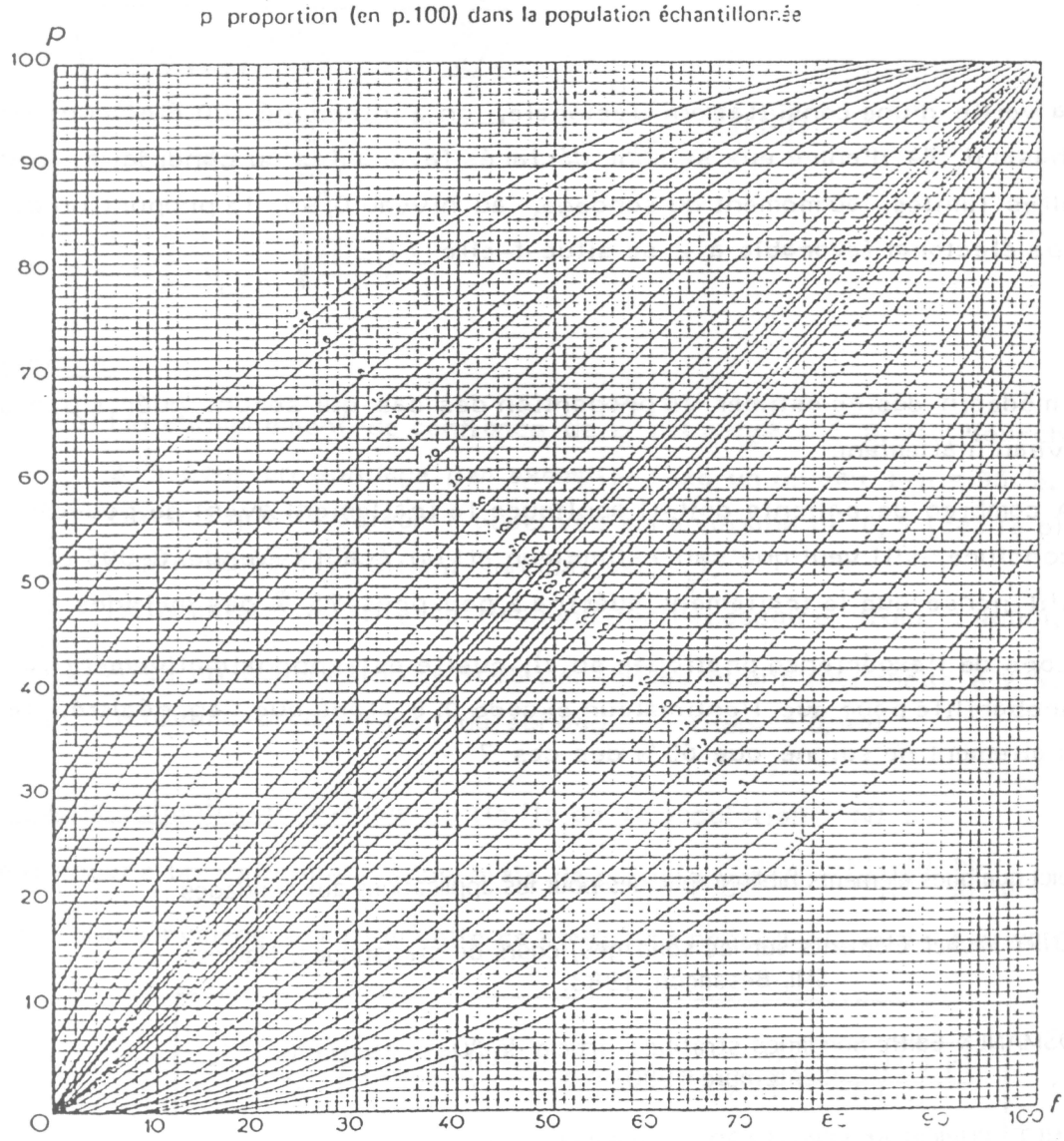
DROESBEKE Jean-Jacques et TASSI Philippe: Histoire de la statistique, col. "que sais-je", PUF 1990.

Cet abaque est extraite du Manuel de Gilbert SAPORTA: Probabilités, analyse des données et statistiques, éditions Technip, 1992.

Table 3 bis

Abaque donnant en fonction de f
l'intervalle de confiance à 0,95 ($p_{0,025}$ à $p_{0,975}$)

f fréquence observée (en p.100) sur un échantillon d'effectif n
p proportion (en p.100) dans la population échantillonnée



1

TESTS D'HYPOTHESES

F. BENINEL
I.U.T. - DPT STID
Centre Du Guesclin
79000 NIORT

Nombreuses sont les problématiques de sciences expérimentales conduisant à la réalisation d'un test d'hypothèses.

En général, la réalisation d'un test d'hypothèses consiste à partir de l'observation d'échantillon(s) convenablement choisi(s), et au moyen d'un mécanisme de décision à choisir entre hypothèses antagonistes ; ces dernières portant sur des caractéristiques inconnues des populations statistiques étudiées.

Φ_1- EXEMPLES INTRODUCTIFS

EXEMPLE 1. CONTROLE STATISTIQUE DE FABRICATION

Une entreprise industrielle produit des bouteilles d'eau minérale dont la teneur en sodium annoncée est de 44mg/l.

Ne pouvant mesurer la teneur en sodium de toutes les bouteilles d'eau produites et désirant être en règle avec la législation du point de vue conformité aux normes, l'entreprise a "à choisir" pour chaque lot (ensemble des bouteilles fabriquées en 1 journée : population statistique), entre les 2 hypothèses suivantes :

H_0 : le lot est conforme à la norme
(en matière de teneur en sodium)

H_1 : le lot n'est pas conforme.

En désignant par TS_m la teneur moyenne en sodium dans le lot, la personne chargée du contrôle statistique de la fabrication, reformule le test comme suit :

H_0 : $TS_m = 44$
H_1 : $TS_m \neq 44$

TS_m est un paramètre inconnu, la réalisation de ce test nécessite la mesure de la teneur en sodium sur un échantillon de bouteilles (une partie du lot).

La validité du mécanisme permettant d'opter pour l'une ou l'autre des 2 hypothèses, dépend du nombre de bouteilles analysées (taille d'échantillons) et de la façon dont ces bouteilles ont été prélevées (Echantillonnage).

Le raisonnement (de type inductif) présidant à la réalisation du test, peut conduire aux éventualités résumées par le tableau ci-après :

2

<**1>

DECISION \ VERITE	H_0	H_1
H_0	Bonne décision	Lot commercialisé alors qu'il est non conforme
H_1	Lot non commercialisé alors qu'il est conforme	Bonne décision

EXEMPLE 2 : COMPARAISON DE 2 VARIETES DE MAïS DU POINT DE VUE DU RENDEMENT

Une entreprise de commercialisation de semences désire comparer les performances des variétés (V_A, V_B) de maïs, quant au rendement.

Un plan d'expérimentation est donc mis en place, celui-ci consistant à semer chacune des deux variétés sur un certain nombre de parcelles de 1 ha (n_A parcelles élémentaires pour V_A ; n_B parcelles pour V_B :Echantillons).

Désignons par X_1,X_2 designent les variables rendement de la variété V_A et rendement de la variété V_B.

Les deux variétés sont semées dans des conditions similaires et fournissent des résultats du type suivant :

Données expérimentales

	Variété V_A	Variété V_B
	$x_{1\ 1}$	$x_{2\ 1}$
	$x_{1\ 2}$	$x_{2\ 2}$
Rendement de V_A obtenu sur la parcelle n° i <-------	$x_{1\ i}$	$x_{2\ j}$ --> Rendement de V_B obtenu sur la parcelle n° j
	$x_{1\ n_A}$	$x_{2\ n_B}$

Les résultats obtenus "suggèrent", que globalement, la variété V_B donnerait un meilleur rendement que la variété V_A. Ainsi se propose-t-on de vérifier cela, par la réalisation du test:

$H_0 : R_A = R_B$ (rendements identiques)
$H_1 : R_A < R_B$ (V_B a un meilleur rendement que V_A)

ici R_A et R_B désignent les rendements moyens (ceux que l'on obtiendrait en semant les 2 variétés sur toutes les parcelles imaginables : population) des variétés V_A et V_B.

3

A l'issue du test, on est conduit à se trouver dans une des 4 éventualités, données par le tableau suivant :

<**2>

DECISION \ VERITE	H_0	H_1
AH_0	Bonne décision	On considère les 2 variétés à mêmes rendements alors que V_B est meilleur. Erreur du type II
AH_1 (RH_0)	On considère V_B à meilleur rendt alors que les rendements sont identiques Erreur du type I	Bonne décision

A travers ces deux exemples, il apparaît que la décision que l'on prend à l'issue d'un test est entachée de risques d'erreurs qu'il s'agira de minimiser.

Ces risques d'erreurs sont dûs au hasard de l'échantillonnage.

La qualité statistique d'un test exige un choix rigoureux des échantillons. [Les échantillons doivent être représentatifs des populations étudiées].

Φ_2-PRINCIPES GENERAUX ET DEFINITIONS

La réalisation d'un test nécessite la définition des hypothèses :

H_0 (Hypothèse nulle)

H_1 (Hypothèse alternative)

Une et une seule de ces hypothèses est vraie. Par ailleurs ces hypothèses ne sont pas interchangeables. En effet, il est plus facile de réaliser le test (T_1) ci-dessous que le test (T_2).

$$(T_1)\quad \begin{cases} H_0 : TS_m = 0,45 \\ H_1 : TS_m \neq 0,45 \end{cases} \qquad (T_2)\quad \begin{cases} H_0 : TS_m \neq 0,45 \\ H_1 : TS_m = 0,45 \end{cases}$$

L'hypothèse H_0 joue un rôle privilégié car, et nous le verrons plus loin, influant sur le mécanisme de décision.

Comme nous l'avons vu en (**1) et (**2), quatres situations mettant en évidence le caractère aléatoire de la décision, sont à envisager :

RH_0/H_0 : on considère H_0 fausse alors qu'elle est vraie. Il s'agit là d'une erreur dite de 1ère espèce

RH_1/H_1 : on rejette H_1 alors que H_1 est vraie, cette erreur est appelée erreur de 2ème espèce

AH_0/H_0 : on conserve H_0 (comme hypothèse plausible) alors qu'elle est vraie. Bonne décision

AH_1/H_1 : on conserve H_1 alors qu'elle est vraie. Bonne décision

RISQUES, TABLEAU DE VERITE

Le risque de commettre l'erreur de 1ère espèce (ou risque de 1ère espèce) est donné par :

$$\alpha = \text{Proba}\ (RH_0/H_0)$$

Ce risque est fixé, au préalable, par le statisticien ; Selon les domaines d'application des tests, ce risque est choisi petit ou très petit (5 % , 1 % , 1 %.)

Le risque de commettre l'erreur de 2ème espèce (risque de 2ème espèce), s'il est souhaité aussi petit que possible, dépend de la valeur α, de la dimension des données expérimentales (taille d'échantillons). Ce risque est donné par :

$$\beta = \text{Proba}\ (RH_1/H_1) = P(AH_0/H_1)$$

En général, ce risque est difficile à calculer avec exactitude.

Relativement à l'exemple du test de l'exemple1 , on obtient une valeur de β chaque fois qu'on fixe une valeur de TS_m # 0,45 (H_1 vraie).

Le risque $\beta = \beta\ (TS_m)$ est d'autant petit que TS_m s'éloigne de 0,45 (valeur de TS_m sous H_0).

Il découle de ce qui précède :

$$P(AH_0/H_0) = 1 - \alpha$$

$$P(AH_1/H_1) = 1 - \beta \text{ (puissance du test)}$$

d'où le tableau récapitulatif suivant :

Tableau de vérité

DECISION \ VERITE	H_0	H_1
AH_0 (RH_1)	$1 - \alpha$	β
AH_1 (RH_0)	α	$1 - \beta$

α : RISQUE DE 1re Espèce

$1 - \beta$: Puissance du test

Φ_3-EXEMPLE DE REALISATION D'UN TEST

Une machine produit des boulons de 1 cm de diamètre.

La variabilité du processus de production est telle que malgré les entretiens dont fait l'objet la machine, le diamètre des boulons ne peut être considéré comme une constante égale à la norme (D_0 = 1 cm)

La recherche de la maîtrise de la fabrication conduit à supposer que le Diamètre D est une variable aléatoire, distribuée selon la loi normale de moyenne m_D et d'écart type σ_D (on note N (m_D, σ_D)).

Ainsi pour bien contrôler le bon fonctionnement de la machine, on décide de prélever chaque 2 heures, 10 boulons, d'en mesurer le diamètre, pour réaliser le test:

H_0 : Bon fonctionnement de la machine ($m_D = D_0$)

H_1 : Fonctionnement défaillant ($m_D \neq D_0$)

(α = 5 %)

ELABORATION DU MECANISME DE DECISION

Notons D_1, D_2, ..., D_{10}, les variables aléatoires d'échantillonnage (D_i : variable aléatoire mesurant le diamètre du ième boulon prélevé $1 \leq i \leq 10$)

Le prélèvement des boulons se faisant avec remise (Echantillon simple) et sachant que les boulons, faisant partie d'un lot, ont la même chance de faire partie de l'échantillon (Echantillon aléatoire).

6

Nous pouvons considérer $D_1, D_2, \ldots, D_{10}$ comme des V.A. indépendantes et de même loi que D ($D_i \rightsquigarrow N(m_D, \sigma_D)$) (**3)

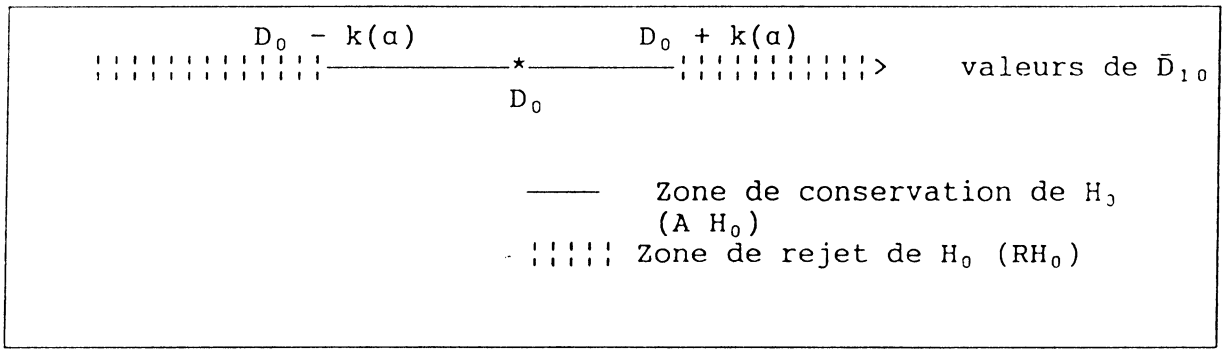
Statistique de test :

Etant donné un échantillon de 10 boulons, il nous paraît clair que H_0 est d'autant invraisemblable que

$\bar{D}_{10} = \frac{1}{10} \sum_{i=1}^{10} D_i$ s'écarte de D_0.

Il convient donc de rejeter H_0 lorsque

$|\bar{D}_{10} - D_0| > k(\alpha)$ [$k(\alpha)$ un nombre dépendant de α, qu'il s'agira de déterminer]



un écart entre $\bar{D}_{10}$ et D_0 n'excédant pas $k(\alpha)$ est attribué au hasard de l'échantillonnage.

($\bar{D}_{10}$ et appelé statistique de test.)

En raison de (**3) on a :

$$\bar{D}_{10} \rightsquigarrow N(m_D, \sigma_D/\sqrt{10})$$

DETERMINONS $k(\alpha)$

α = Proba (RH_0/H_0 vraie)

= Proba ($|\bar{D}_{10} - D_0| > k\alpha / m_D = D_0$)

$$= \text{Proba} \left(\frac{|\bar{D}_{10} - D_0|}{\sigma_D / \sqrt{10}} > \frac{k\alpha}{\sigma_D / \sqrt{10}} \Big/ m_D = D_0 \right)$$

Ayant $m_D = D_0$ (ie. H_0 vraie), cela nous donne :

$$\bar{D}_{10} \rightsquigarrow N(D_0, \sigma_D / \sqrt{10})$$

et par suite:

$$\frac{\bar{D}_{10} - D_0}{\sigma_D / \sqrt{10}} \rightsquigarrow N(0,1)$$

il s'ensuit :

$$k_c = U_{1-\alpha/2}\ \sigma_D/\sqrt{10}$$

$U_{1-\alpha/2}$: fractile de la loi Normale centreé réduite d'ordre $1-\alpha/2$

LA REGLE DE DECISION PEUT SE FORMULER AINSI :

Ayant un échantillon de 10 boulons, on rejettera l'hypothèse H_0 avec un risque de se tromper égal à α, lorsque :

$$|\bar{D}_{10} - D_0| > U_{1-\alpha/2}\ \sigma_D/\sqrt{10}$$

APPLICATION

Examinons ce que donnent deux contrôles différents quant au test :

H_0 : Bon fonctionnement de la machine.

H_1 : Fonctionnement défaillant.

Ce test est réalisé avec un risque de première espèce fixé à 5% .

L'écart type σ_D est supposé connu et égal à 0,01 cm.

CONTROLE 1	CONTROLE 2
ECHANTILLON : 1,01 1,02 0,99 0,97 0,98 0,97 0,98 1,00 1,03 1,02	1,07 1,10 0,94 1,05 1,12 1,03 1,04 0,98 0,97 0,93
CALCUL DE $\bar{D}_{10}$	
$\bar{D}_{10} = 0,997$ d'où	$\bar{D}_{10} = 1,033$
$\lvert\bar{D}_{10} - D_0\rvert = (0,997 - 1) = 0,003$	$\lvert\bar{D}_{10} - D_0\rvert = 0,033$
CALCUL DU SEUIL:	
$u_{1-\alpha/2} = u_{0,975} = 1,96$	
$k_{(\alpha)} = u_{1-\alpha/2}\ \sigma_D/\sqrt{10}$	IDEM
$= 1,96\ (0,01/\sqrt{10})$	
$= 0,0062$	

CONCLUSION

On a $\lvert\bar{D}_{10} - D_0\rvert < k(\alpha)$;	$\lvert\bar{D}_{10} - D_0\rvert > k(\alpha)$
On conserve H_0; le risque de se tromper est égal à ß	On rejette H0; le le risque de se tromper est égal à 5 %
LE FONCTIONNEMENT DE LA MACHINE PEUT ETRE CONSIDERE COMME NORMAL	LE FONCTIONNEMENT DE LA MACHINE PEUT ETRE CONSIDERE COMME DEFAILLANT

PUISSANCE DU TEST

Posons $D = \{ D_1 / D_1 \neq D_0 \}$

D désigne l'ensemble des valeurs que peut prendre m_0 lorsque l'hypothèse H_1 est vraie.

Ou encore, D est l'ensemble des valeurs de réglage de la machine autres que (D_0 = 1 cm).

La fonction puissance du test est donnée par :

$$g : D \longrightarrow [0,1]$$

$$D_1 \longrightarrow g(D_1) = P(AH_1 / m_0 = D_1)$$

$$= 1 - \beta(D_1)$$

Ainsi à chaque valeur D_1 (pour un risque de 1ère espèce et une taille d'échantillon, fixés) correspond une valeur de la puissance.

Ecrivons de façon plus explicite, relativement à notre exemple, $\beta(D_1)$:

$$\beta(D_1) = P(RH_1 / H_1)$$

$$= P(\lvert\bar{D}_{10} - D_0\rvert < u_{1-\alpha/2}\, \sigma_0/\sqrt{10} / m_0 = D_1)$$

$$= P(D_0 - u_{1-\alpha/2}\, \sigma_0/\sqrt{10} < \bar{D}_{10} < D_0 + u_{1-\alpha/2}\, \sigma_0/\sqrt{10})$$

avec $m_0 = D_1$

$$= P\{(D_0-D_1)/(\sigma_0/\sqrt{10}-u_{1-\alpha/2} < U < (D_0-D_1)/(\sigma_0/\sqrt{10}+u_{1-\alpha/2}\}$$

avec $U = (\bar{D}_{10} - D_1) / \sigma_0/\sqrt{10} \rightsquigarrow N(0, 1)$

et par suite:

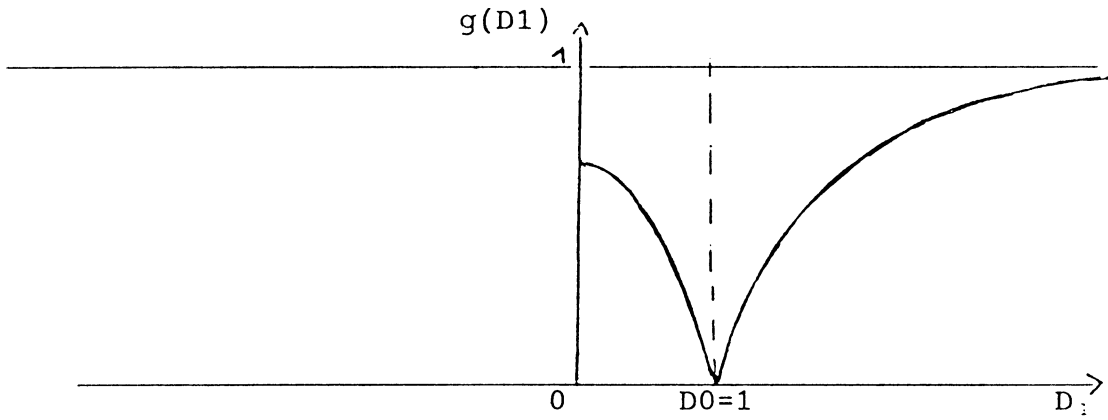
$$\beta(D_1) = \pi\{(D_0-D_1)/\sigma_0/\sqrt{10}+u_{1-\alpha/2}\} - \pi\{(D_0-D_1)/\sigma_0/\sqrt{10}-u_{1-\alpha/2}\}$$

avec π désignant la fonction de répartition de la loi normale centrée réduite.

Calculons la valeur de la fonction puissance pour quelques valeurs de D_1:

D1	0.7	0.8	0.9	0.95	0.99	0.999
g(D1)	1	1	1	1	0.885	0.061

Figure1: Allure de la courbe de la fonction puissance



Comme on peut le voir sur le graphique ci-dessus, si la vraie valeur m_0 a un écart D_0 de l'ordre de 1/1000, le test est peu puissant, ne serait-il pas plus judicieux dans ce cas de réaliser le test à "PILE ou FACE " ! ?

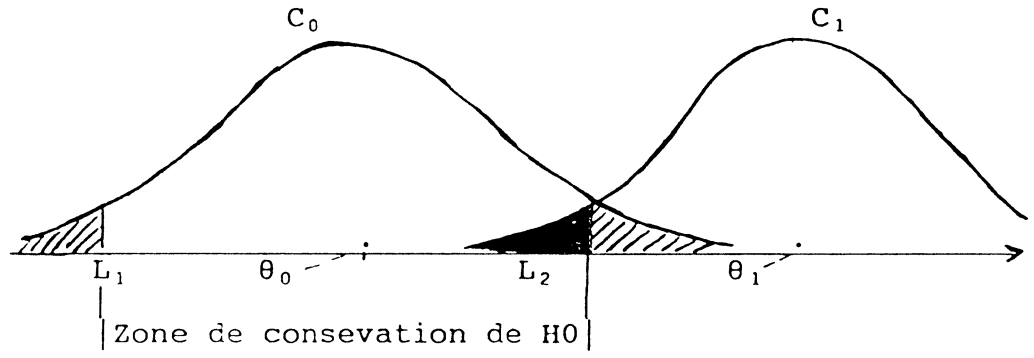
L'erreur de deuxième espèce, consisterait à considérer que la machine est réglée pour produire des boulons de 1 cm de diamètre alors qu'elle est réglée pour produire des boulons de 0.999 cm (par exemple).

Cette erreur si elle est probable (ß =0.994) n'a pas d'incidence négative sur la fabrication des boulons;En effet un boulon de 0.999cm de diamètre assure " sans peine " la fonction d'un boulon de 1cm.

D'une façon générale, le test est d'autant puissant que la valeur de θ_1 s'éloigne de θ_0 , pour le test: $H_0:\theta=\theta_0$ - $H_1:\theta=\theta_1$ (graphique ci après).

Figure2: Visualisation graphique des risques.

C_0 :Courbe de la fonction densité de T sous H_0
C_1 :Courbe de la fonction densité de T sous H_1



α = Aire hachurée. ß = Aire en noir.

Ainsi donc,nous pouvons résumer la démarche d'un test comme suit:

1) Définition des hypothèses H_0 et H_1 et la donnée du risque de première espèce α

2) Détermination de la statistique de test et étude de sa loi de probabilité sous H_0;

3) Détermination de la région critique (ou régle de décision);

4) Etude de la puissance 1-ß ;

5) Calcul de la valeur expérimentale de la statistique de test (ou variable de décision);

6) Conclusion : rejet ou conservation de H_0.

Remarque relative au risque α

On peut envisager deux façons de procéder:

La première façon consiste à fixer α et à voir si l'on rejette H_0 ou pas.
La deuxième façon consiste à déterminer α,dans le cas ou l'on décide de rejeter H_0.

Examinons ce que donnent ces deux démarches sur l'exemple (paragraphe Φ_3-controle2 sur les boulons).

α1) En prenant α = 5%,on obtient comme seuil de rejet de H_0, k(α) = 0,0062 .

α2) On se pose la question suivante:"En rejettant H_0,dans la configuration de nos données,quel rique α encourt-on".

Dans ce cas on prend $|\bar{D}_{10} - D_0| = 0,033$ comme seuil de rejet de H0.
Dans ce cas on pose:

$k(\alpha)=0,033$

Cela donne:

$U_{1-\alpha/2} = 10,48$

Et par suite:

$\alpha = \alpha_0 < 10^{-5}$

En résumé $\alpha = 5\%$ est le risque maximum quand on rejette H_0. $\alpha=\alpha_0$ est le risque exacte quant on rejette H_0.

Nous pouvons classifier les tests que l'on retrouve dans la littérature statistique, en distinguant les catégories suivantes:

C1- TESTS DE COMPARAISONS

(Moyennes, écart-types, proportions, distributions de probabilité, ...)

C2- TESTS DE CONFORMITE

(Moyenne, écart-types, proportions, ...)

C3- TESTS D'INDEPENDANCE ENTRE VARIABLES

(test du Chi-2, test de corrélation, ...)

C4- TESTS D'AJUSTEMENT D'UNE DISTRIBUTION EXPERIMENTALE A UNE DISTRIBUTION THEORIQUE

(test du Chi-2, test de Kolmogorov-Smirnov, ...)

Sur le plan méthodologique la réalisation des tests cités ci-dessus,se fait selon deux approches différentes:

(*) Approche paramétrique.

(**)Approche non paramétrique.

Pour voir ce qui différencie ces deux approches intéréssons nous au test relatif à la comparaison du rendement des deux variétés de maïs V_A et V_B (cf:Exemple introductif n° 2).

Approche paramétrique du test:

$$H0: E(X_1)=E(X_2)$$

$$H1: E(X_1) < E(X_2)$$

Approche non paramétrique du test:

$$H0: P[rg(X_1)>rg(X_2)] = P[rg(X_1)<rg(X_2)]=1/2$$

$$H1: P[rg(X_1)>rg(X_2)] = P[rg(X_1)<rg(X_2)]$$

(E:symbole espérance P: probabilité et rg:le rang)

L'approche paramétrique suppose connue la distribution des variables aléatoires dont on étudie les paramètres.

L'approche non paramétrique ne fait aucune supposition quant à la distributions des variables aléatoires étudiées.

Ces deux approches ne sont pas en concurrence;c'est la nature des donnéesqui rend telle ou telle autre approche efficiente.

En ce qui concerne le rendement (exemple 2),divers connaissances nous permettent de considérer cette variable comme distribuée selon la loi normale.

S'il s'agissait de notes attribuées à des vins quant à leur bon goût,le caractère particulièrement subjectif de la notation nous amène à considérer comme plus fiable le classement des vins,que les notes attribuées.

Φ_4-METHODOLOGIE DE CONSTRUCTION D'UN TEST

Par choix d'un test,il est entendu dans le cas présent le choix de la régle de décision ,une fois les hypothèses posées.

Dans ce qui suit,on se place dans le cas général:

$$H0: \quad \theta \in \theta_0$$

$$H1: \quad \theta \in \theta_1$$

θ: paramètre d'une variable aléatoire X.

θ_0:ensemble des valeurs du paramètre θ sous H0.

θ_1:ensemble des valeurs du paramètre θ sous H1.

Pour réaliser le test posé précédement,on dispose d'échantillons de taille n,du type $(x_1,x_2,...,x_n)$,de réalisations de la variable X.

Posons D={d0,d1},l'ensemble des décisions possibles à l'issue du test.

d0:Décider H0.

d1:Décider H1.

Remarque

Il faut dire que pour certaines procédures,l'ensemble des décisions peut comporter plus que deux décisions.

A titre d'exemple on peut avoir comme décisions possibles:

d0:décider H0.

d1:décider H1.

d2:s'abstenir(ou réaliser d'autres observations pour mieux décider).

Un test peut être considéré comme une fonction:

$\Phi: X^n \longrightarrow D$

(X^n:ensemble des échantillons de valeurs de X,de taille n)

On appellera région critique,l'ensemble:

$$W=\{ (x_1,x_2,...,x_n)\in X^n \ / \ \Phi(x_1,x_2,...,x_n)=d1\}$$

$$=\Phi^{-1}(d1)$$

On appellera région de"conservation" de H0,l'ensemble:

$$\overline{W}=\{ (x_1,x_2,...,x_n)\in X^n \ / \ \Phi(x_1,x_2,...,x_n)=d0\}$$

$$=\Phi^{-1}(d0)$$

Le mécanisme de décision est élaboré dés lors que l'on connait W (ou, ce qui est équivalent,la fonction Φ).

Supposons maintenant que pour un même test l'on ait deux régions critiques différentes W et W'(ou encore deux fonctions test Φ et Φ').

La règle de décision se basant sur W comme région critique sera dite meilleure que celle se basant sur W' si:

(i) $\alpha(\Phi) \leq \alpha(\Phi')$.

(ii) $\beta(\Phi) \leq \beta(\Phi')$.

Avec $\alpha(\Phi)=P((X_1,X_2,\ldots,X_n)\in W / H_0 \text{ vraie})$

$\beta(\Phi)=P((X_1,X_2,\ldots,X_n)\in \overline{W} / H_1 \text{ vraie})$

$X_1,X_2,\ldots,X_n$:variables aléatoires d'échantillonnage indépendantes et de même loi que X.

En d'autres termes,la meilleure règle de décision sera celle minimisant à la fois l'erreur de première espèce et celle de deuxième èspece.

En réalité,on ne peut minimiser simultanément les deux risques(α et β) (voir figure 2).

La démarche proposée par NEYMAN-PEARSON consiste pour α fixé,à construire une règle de décision (ou fonction test Φ^*),minimisant le risque de deuxième éspèce β.

càd: $\beta(\Phi^*)=\min \beta(\Phi)$.

METHODOLOGIE DE NEYMAN-PEARSON

Fonction de vraisemblance (rappel):

Soit $(x_1,x_2,\ldots,x_n)$ un échantillon indépendant d'une variable aléatoire X.

La loi de X dépend du paramètre θ.

notons $L(x_1,x_2,\ldots,x_n,\theta)$,la vraisemblance de l'echantillon $(x_1,x_2,\ldots,x_n)$,on a si:

X est une V.A discrète:

$$L(x_1,x_2,\ldots,x_n,\theta)=P[X=x_1]\ P[X=x_2]\ldots P[X=x_n]$$

X est une V.A continue:

$$L(x_1,x_2,\ldots,x_n,\theta)=f(x_1)f(x_2)\ldots f(x_n)$$

(ou f désigne la foncion densité de X)

Considérons le test:

H0: $\theta=\theta_0$

H1: $\theta=\theta_1$ (θ_0 différent de θ_1)

Théorème de Neyman-Pearson:

Pour tout $\alpha\in[0,1]$, il éxiste un test pur de niveau α, de puissance maximum, défini par la région critique W:

$$W=\{(x_1,x_2,\ldots,x_n)\in X^n / \frac{L(x_1,x_2,\ldots,x_n,\theta_0)}{L(x_1,x_2,\ldots,x_n,\theta_1)} \leq k(\alpha) < 1\}$$

Cela signifie que lorsque pour un échantillon $(x_1,x_2,\ldots,x_n)$ donné la vraisemblance sous H_0 est plus faible que celle sous H_1, on rejette H_0.
En d'autres termes H_1 est des deux hypothèses, la plus vraisemblable.

Exemple d'application:

Considérons le test de conformité relatif à la moyenne d'une population gaussienne. [$N(m,1)$]

H_0: $m = m_0 = 2$ [$m = m_0$]

H_1: $m < m_0$ [$m = m_1 < m_0$]

(on prendra comme risque de 1ère éspèce $\alpha = 5\%$)

Pour réaliser ce test, on a prélevé l'échantillon $x_1,x_2,\ldots,x_{16}$ de moyenne $\overline{x} = 2,6$.

Forme de la région critique:

On a $L(m)=(x_1,x_2,\ldots,x_{16},m) = \pi(1/\sqrt{2\pi})\ \exp(-1/2(x_i-m)^2)$

cela nous donne l'équivalence entre les relations (1) et (2) suivantes:

(1) $L(m_0)/L(m_1) \leq k(\alpha)$

(2) $-1/2(m_1-m_0)\Sigma[2xi-(m_0+m_1)] \leq k'(\alpha)$ avec $k'=Ln(k)$.

On déduit de ce qui précède(cf. relation (2)):

$\overline{x} = 1/16\ \Sigma x_i \leq k''(\alpha)$ (k":un nombre indépendant des valeurs x_i)

La région critique est selon Neyman-Pearson de la forme:

$W=\{(x_1,x_2,\ldots,x_{16})/\ \overline{x} \leq k''(\alpha)\}$

Explicitons davantage cette région critique:

$\alpha = P[(X_1,X_2,\ldots,X_{16})\in W\ /\ m=m_0]$

$= P[\ \overline{X} \leq k''\ /\ m=m_0]$

En utisant le fait que sous H_0 $\quad \overline{X} \rightsquigarrow N(m_0 , 1/\sqrt{16})$ on a:

$$k''(\alpha) = m_0 + U_\alpha$$

Et par suite:

$$W=\{(x_1,x_2,\ldots,x_{16})/\ \overline{x} \leq m_0+(1/4)\ U_\alpha\}$$

Relativement à notre exemple on est conduit à rejetter H0 puisque:

$$k''(\alpha) = k''(0,05) = 1,59$$

et

$$\overline{x} = 2,6$$

OUVRAGES CONSEILLES:

1-C.R.RAO: Linear statistical inference and its applications,Wiley,1973.

2-P.TASSI: Méthodes statistiques,Economica,1989.

3-G.SAPORTA: Probabilités,analyse des données et statistiques,Technip,1990.

4-P.DAGNELIE: Théorie et méthodes statistiques,vol1,2, presses agronomiques de Gembloux,1973.

EN DEÇA ET AU-DELA DE LA GENESE D'UN TEST

GRAS Régis
IRMAR ([1])

INTRODUCTION

L'objectif premier poursuivi à travers cet exposé consiste en la reconstruction, au sens épistémologique du terme, d'un concept statistique. Je souhaiterais vous restituer, à partir de la problématique originale, la modélisation et la formalisation choisies, puis les développements théoriques consécutifs, et enfin quelques applications obtenues. J'essaierai de montrer en quoi le choix opportun d'un modèle accroit la fécondité de son extension théorique et la variété de ses applications. Le temps me manquera certainement pour vous faire part des errances vécues mais je voudrais qu'elles ne soient pas ignorées, la construction théorique ne suivant jamais une trajectoire complètement prévisible, sans point d'arrêt et sans remise en cause.

En 1978, la problématique de ma recherche s'inscrit dans le cadre de l'évaluation nationale d'une expérience pédagogique à l'origine du courant réformateur des programmes du 1er cycle. Plus spécifiquement, apparaît la nécessité de hiérarchiser des performances d'élèves en fonction des niveaux de complexité a priori des questions qui leur sont posées. Pour y parvenir, il me faut donner un sens à des énoncés du type : "la question **a** est plus complexe que la question **b**", énoncé qui se ramène à celui-ci : "tout élève qui réussit **a** réussit également **b**". C'est cette situation de quasi-implication ou, en langage ensembliste de quasi-inclusion, qu'il faut tester.

Or, étant donné la taille de l'échantillon d'élèves et le nombre d'items à hiérarchiser, je ne dispose à ce moment que d'outils statistiques classiques mettant à l'épreuve une hypothèse nulle, différente de celle en question, ou encore de méthodes d'analyse de données ne traitant la multidimensionnalité qu'à travers la similarité ou la distance mutuelle, approches essentiellement symétriques.

Je vous fais grâce des nombreux tâtonnements qui m'ont fait croiser par hasard (et m'en rendre compte après coup, hasard que rencontrera également plus récemment A. Bodin) la même réponse que celle donnée par un chercheur britannique, J. Loevinger, en 1947. Mais cet indice, dont je parlerai plus loin, n'est pas une échelle de probabilité et je l'ai rejetée.

Une réponse non quantifiée à la même question est fournie par certains psychologues à partir de l'examen du tableau de croisement des variables **a** et **b**. Pour décider, par exemple, que a implique b, ils estiment "à vue" la petitesse de l'effectif de l'ensemble de ceux qui satisfont **a** sans satisfaire **b**. Mais ce jugement est subjectif.

La première brique théorique va donc consister en une substitution de l'indice de J. Loevinger par un indice fondé sur les probabilités. Les briques suivantes sont le fruit des thèses de A. LARHER (1991) et de celles de S. Ag. Almouloud, A. Totohasina et H. Ratsimba-Rajohn (thèses à soutenir fin 1992).

---

[1]) Institut de Recherche Mathématique de Rennes, Campus de Beaulieu - 35042 RENNES CEDEX.

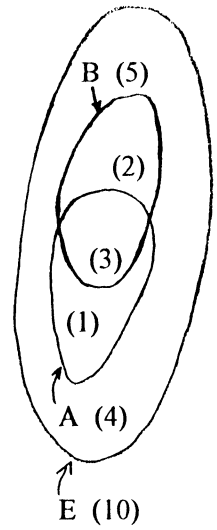
§ 1 - IMPLICATION ENTRE VARIABLES BINAIRES, TEST, EXTENSION.

1.1. Modélisation.

Dans le cas binaire, la situation générique est la suivante. Croisant une population E et un ensemble de variables V et du fait de l'observation exceptionnelle de l'implication stricte de la variable **a** sur la variable **b**, on veut donner un sens statistique à une implication non stricte : **a** $\Rightarrow$ **b**. En termes ensemblistes, A et B représentant les sous-populations respectives où les variables **a** et **b** prennent la valeur 1 (ou "vrai"), il y a équivalence à mesurer l'inclusion non stricte de A dans B.

Par exemple, si **a** est l'attribut "cheveux blonds" et **b** l'attribut "yeux bleus" dans une population E d'étudiants, les données pour étudier si **a** $\Rightarrow$ **b**, dans le cas où Card E = 10, peuvent se présenter de 3 façons différentes :

V / Sujets	**a**	**b**
1	0	0
2	0	1
3	1	1
4	1	0
5	0	0
6	1	1
7	1	1
8	0	0
9	0	1
10	0	0
Total	4	5



a / **b**	1	0	Marges
1	3	1	4
0	2	4	6
Marges	5	5	10

S'inspirant de la méthode de I.C. LERMAN [81] pour définir la similarité, R. GRAS [79] axiomatise la notion d'implication statistique de la façon suivante :

Soient X et Y deux parties aléatoires quelconques de E, choisies indépendamment (absence de lien a priori) et de mêmes cardinaux respectifs que A et B. Soient $\bar{Y}$ et $\bar{B}$ les complémentaires respectifs de Y et de B dans E.

Axiome 1.

> (**a** $\Rightarrow$ **b**) *est admissible au niveau de confiance 1 - α si et seulement si* $Pr\,[card(X \cap \bar{Y}) \leq card(A \cap \bar{B})] \leq \alpha$.

Intuitivement et qualitativement, ceci signifie que l'implication **a** $\Rightarrow$ **b** sera admissible à l'issue d'une expérience si le nombre d'individus de E la contredisant dans l'expérience est "*invraisemblablement petit*" par rapport au nombre d'individus attendu dans une hypothèse d'absence de lien.
Par exemple si Card E = 100, Card A = 35, Card B = 50, alors Card ($A \cap \bar{B}$) = 3 est "*invraisemblablement petit*" pour une absence de lien entre **a** et **b**.

La modélisation probabiliste que nous retenons de façon privilégiée est décrite par un processus de tirage aléatoire en 3 étapes (cf. I.C. LERMAN, R. GRAS, H. ROSTAM [81]). Notons n_a, n_b, $n_{\bar{b}}$, $n_{a\wedge\bar{b}}$, $n_{av\bar{b}}$, les cardinaux respectifs de A, B, $\bar{B}$, $A \cap \bar{B}$, $A \cup \bar{B}$:

. on considère le référentiel E comme la réalisation d'un référentiel aléatoire $\mathcal{E}$ dont le cardinal $\mathcal{N}$ serait une variable aléatoire de Poisson de paramètre le cardinal n de E observé, hypothèse compatible avec les situations les plus fréquemment rencontrées et qui ne nuit pas à la généralité de la modélisation :

$$\Pr[\mathcal{N} = m] = \frac{n^m}{m!} e^{-n}$$

. le choix aléatoire d'une partie quelconque (par exemple X) de cardinal aléatoire K pour une distribution uniforme de probabilité sur les éléments de $\mathcal{E}$ et égale à $\frac{d}{m}$ (dans le cas de X, $d = n_a$) est alors de type binomial :

$$\Pr[K = k / \mathcal{N} = m] = \binom{m}{k} \left(\frac{d}{m}\right)^k \left(1 - \frac{d}{m}\right)^{m-k} \qquad \text{(pour } k \leq m\text{)}$$

. X et $\bar{Y}$ étant deux parties quelconques choisies de façon indépendante parmi les parties ayant respectivement pour cardinaux n_a et $n_{\bar{b}}$, la probabilité qu'un élément de $\mathcal{E}$ appartienne à $X \cap \bar{Y}$ est :

$$p(a).p(\bar{b}), \text{ où } p(a) = \frac{n_a}{m} \text{ et } p(\bar{b}) = \frac{n_{\bar{b}}}{m}$$

Proposition 1. La variable aléatoire Card $(X \cap \bar{Y})$ suit la loi de Poisson de paramètre $n.p(a).p(\bar{b})$.

La loi de probabilité conditionnelle du cardinal de $X \cap \bar{Y}$ est binomiale de paramètres m et $\pi = p(a).p(\bar{b})$

$$\Pr[\text{Card}\,(X \cap \bar{Y}) = s / \mathcal{N} = m] = \binom{m}{s} \pi^s (1 - \pi)^{m-s} \quad \text{pour } s \leq n_{a\wedge\bar{b}} \text{ et } m \geq n_{av\bar{b}} .$$

Par suite, en faisant varier le conditionnement de $\mathcal{N}$, on obtient :

$$\Pr[\text{Card}(X \cap \bar{Y}) = s] = \sum_{m \geq s} \Pr[\text{Card}(X \cap \bar{Y}) = s / \mathcal{N} = m] \times \Pr[\mathcal{N} = m] = \frac{(n\pi)^s}{s!} e^{-n\pi} .$$

La variable aléatoire Card $(X \cap \bar{Y})$ suit donc la loi de Poisson de paramètre $n.\pi = n.p(a).p(\bar{b})$ (de moyenne et de variance $n.\pi$). D'autres modélisations conduiraient à une loi hypergéométrique ou à une loi binomiale.

Corollaire. Comme I.C. LERMAN, nous réduisons et centrons cette variable de Poisson en la variable :

$$Q(a,\bar{b}) = \frac{\text{Card}\,(X \cap \bar{Y}) - n\,p(a)\,p(\bar{b})}{\sqrt{n\,p(a)\,p(\bar{b})}} = \frac{\text{Card}\,(X \cap \bar{Y}) - \frac{n_a\,n_{\bar{b}}}{n}}{\sqrt{\frac{n_a\,n_{\bar{b}}}{n}}} .$$

Dans l'expérience, la valeur observée de $Q(a,\bar{b})$ est $q(a,\bar{b}) = \dfrac{n_{a\wedge\bar{b}} - \frac{n_a\,n_{\bar{b}}}{n}}{\sqrt{\frac{n_a\,n_{\bar{b}}}{n}}}$, indicateur de la non-implication de **a** sur **b**.

Dans les cas légitimant convenablement l'approximation ($\frac{n_a\, n_{\bar{b}}}{n} \geq 3$), la variable $Q(a,\bar{b})$ suit la loi normale centrée réduite. L'intensité d'implication, qualité de l'admissibilité de **a** $\Rightarrow$ **b**, pour $n_a \leq n_b$, est alors définie à partir de l'indice $q(a,\bar{b})$ par :

Définition 1.

$$\varphi(a,\bar{b}) = 1 - Pr[Q(a,\bar{b}) \leq q(a,\bar{b})] = \frac{1}{\sqrt{2\pi}} \int_{q(a,\bar{b})}^{\infty} e^{-\frac{t^2}{2}}\, dt$$

Axiome 1'. L'axiome 1 devient :

> L'implication **a** $\Rightarrow$ **b** sera admissible au niveau de confiance $1 - \alpha$, si et seulement si
> $$\varphi(a,\bar{b}) = 1 - Pr\,[Q(a,\bar{b}) \leq q(a,\bar{b})] \geq 1 - \alpha.$$

Notons que si l'approximation gaussienne n'est pas valide, il est loisible de revenir aux origines poissonniennes de la variable Card $(X \cap \bar{Y})$ et de considérer :

$$\varphi(a,\bar{b}) = 1 - Pr\,[Card\,(X \cap \bar{Y}) \leq n_{a\wedge\bar{b}}\,].$$

Exemple

a \ b	1	0	Marges
1	5	1	6
0	10	84	94
Marges	15	85	100

Considérons les données ci-contre dans lesquelles :

Card $(A \cap \bar{B}) = n_{a\wedge\bar{b}} = 1$.

Alors $q(a,\bar{b}) = -1{,}816$

et $\varphi(a,\bar{b}) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-1{,}816}^{\infty} e^{-\frac{t^2}{2}}\, dt = 0{,}9653.$

On dira dans ce cas que (**a** $\Rightarrow$ **b**) est admissible au niveau de confiance 96,5 %.

Remarquons deux valeurs particulières :

$\varphi(a,\bar{b}) \geq 0{,}95 \Leftrightarrow q(a,\bar{b}) \leq -1{,}65$

et $\varphi(a,\bar{b}) \geq 0{,}5 \Leftrightarrow q(a,\bar{b}) \leq 0$.

On notera que l'approche ci-dessus peut s'exprimer en termes de test d'hypothèse. Cependant, nous ne retenons pas cette voie qui, du fait de sa visée de prise de décision, limiterait les considérations qui vont suivre. Mais cette possibilité marque bien la différence avec l'approche de J. Loevinger (LOEVINGER [1947] qui définissait la quasi-implication de **a** sur **b** par l'indice qui prend ses valeurs sur $]-\infty,1]$:

$H(a,b) = 1 - \frac{n.n_{a\wedge b}}{n_a\, n_{\bar{b}}}$. Si $H(a,b)$ est "proche" de 1, l'implication est "presque" satisfaite.

Mais, comme on le voit, cet indice présente l'inconvénient, ne se référant pas à une échelle de probabilité, de ne pas fournir de seuil de vraisemblance et d'être invariant dans toute dilatation de E, A, B et $A \cap \bar{B}$. On a d'ailleurs la relation suivante entre $H(a,b)$ et $q(a,\bar{b})$:

$$\frac{q(a,\bar{b})}{H(a,b)} = -\sqrt{\frac{n_a\, n_{\bar{b}}}{n}}\,.$$

Cette limitation apparaît dans l'approche de J. Pearl (PEARL [1988]), de S. Acid et als (ACID [1991]) et de A. Gammerman, Z. Luo (GAMMERMAN A. et LUO Z. [1991]). Chez ces derniers chercheurs, c'est l'écart entre la distribution conjointe entre **a** et **b** (et non **a** et $\bar{\mathbf{b}}$) et la distribution produit qui tient lieu de critère comparatif.

Dans la recherche sur l'apprentissage de bases de connaissances de J.G. Ganascia (GANASCIA J.G. [1991]), où "*l'incertitude*" sur l'implication **a** $\Rightarrow$ **b** est évaluée par l'indice : 2 Pr[b/a] - 1 et étendue à des classes de variables, la simplicité de l'approche de la quasi-implication se paie selon ces deux mêmes inconvénients. De plus, cet indice ne sépare pas, numériquement, deux implications dont l'une serait triviale et l'autre hautement informative.

1.2. Test d'hypothèse

L'indice d'implication étant défini entre deux variables binaires, il est possible de considérer cet indice comme la valeur d'une statistique du cadre inférentiel. En effet, examinons, relativement aux 2 variables **a** et **b**, l'hypothèse nulle suivante :

$\mathbf{H_0}$: les variables **a** et **b** sont indépendantes

Sous cette hypothèse, l'effectif des individus respectant **a** et non **b** doit être $\frac{n_a\, n_{\bar{b}}}{n}$. En considérant comme hypothèse alternative $\mathbf{H_1}$: la variable **a** implique la variable **b**, le test va consister à estimer jusqu'à un certain seuil (par exemple à 5 %) la faiblesse de l'effectif observé, eu égard, d'une part à $\mathbf{H_0}$, d'autre part aux variations dues aux aléas de l'échantillonnage. Ainsi nous accepterons l'hypothèse $\mathbf{H_1}$ si l'indice $q(a,\bar{b})$ associé est inférieur à la valeur-limite définie par la loi de Poisson ou par la loi de Laplace-Gauss (par exemple, pour 5 %, $q(a,\bar{b}) = -1{,}65$).

En fait, on retiendra, préférentiellement, le cas échéant, l'une des 2 hypothèses alternatives :

"la variable **a** implique la variable **b**" ou "la variable **b** implique la variable **a**".

Le critère décisif tiendra à l'ordre entre les nombres $n_{a\wedge\bar{b}}$ et $n_{\bar{a}\wedge b}$ comme il a été vu dans le cas de l'indice d'implication, sachant que si ces 2 nombres sont petits, on acceptera simultanément les 2 hypothèses.

1.3. Quelques propriétés du modèle

A. LARHER dans sa thèse [1991] étudie et démontre différentes propriétés de q et φ. Les plus importantes, auxquelles nous adjoignons les propositions 5 et 6, sont les suivantes :

Proposition 2. Si, n_a étant fixé et A inclus dans B, n_b tend vers n (B croît vers E), alors $\varphi(a,\bar{b})$ tend vers 0,5. Un prolongement par continuité nous permet donc de définir : si B = E alors $\varphi(a,\bar{b}) = 0{,}5$.

Proposition 3. Pour toute variable a :

$$0{,}95 \le \varphi(a,\bar{a}) \le 1 \quad \Leftrightarrow \quad n_a \in \left[\frac{n - \sqrt{n(n-11)}}{2} \; ; \; \frac{n + \sqrt{n(n-11)}}{2}\right].$$

Or, pour n assez grand, l'intervalle ci-dessus est voisin de [0,n], ce qui permet d'affirmer que l'implication statistique **a** $\Rightarrow$ **a** a un sens, tout en réservant une limite de confiance à la stabilité du caractère reproductible de la variable **a**.

Proposition 4. *La relation $\mathcal{R}$ sur V^2 définie par :*

$$\forall (a,b) \in V^2 \qquad a \mathcal{R} b \quad \text{dès que} \qquad \varphi(a,\bar{b}) \geq 0{,}95 \text{ et que } n_a \text{ vérifie (1)}$$

est réflexive, mais ni symétrique, ni antisymétrique, ni transitive.

Pour lui associer un graphe valué, sans cycle et transitif, on en prendra la restriction $\mathcal{R}'$ aux variables vérifiant la condition : si $a \mathcal{R}' b$ et $b \mathcal{R}' c$, alors l'arc (a,c) appartient au graphe seulement si $\varphi(a,\bar{c}) \geq 0{,}5$. $\mathcal{R}'$ définit alors un préordre partiel et permet une représentation claire de la relation d'implication statistique (cf. GRAS [1979]). On décrit dans [LERMAN I.C., GRAS R., ROSTAM H. 1981] l'algorithme qui permet de le construire. Ce seuil de 0,5 permet, en outre, la satisfaction d'un objectif d'accroissement informationnel. En effet, si I est l'incertitude associée aux variables a et b, il permet de vérifier $I(a|b) \leq I(a)$.

Remarque 1. Notre approche diffère également de celle de S. AMARGER et als (AMARGER S., DUBOIS D. et PRADE H. [1991]) qui, à partir d'une certaine inférence (une probabilité conditionnelle telle que $p(b|a)$ appartient à un intervalle, sans être parfaitement connue), induisent transitivement, de proche en proche, des probabilités conditionnelles sur un graphe incomplet, et cela sans la contrainte d'un seuil. Notre problématique, à l'opposé, vise l'analyse d'un tableau donné, sans ambition inductive a priori, mais en imposant un seuil qui autorise la fermeture transitive du graphe implicatif.

Proposition 5. *Comparons le coefficient de corrélation $\rho(a,b)$ et l'indice $q(a,\bar{b})$.*

$$\text{Supposons } q(a,\bar{b}) \neq 0. \text{ Alors : } \frac{\rho(a,b)}{q(a,\bar{b})} = -\sqrt{\frac{n}{n_b\, n_{\bar{a}}}}$$

$$\text{En effet, } q(a,\bar{b}) = \frac{n_{a\wedge\bar{b}} - \dfrac{n_a\, n_{\bar{b}}}{n}}{\sqrt{\dfrac{n_a\, n_{\bar{b}}}{n}}} = \frac{n_a\, n_b - n\, n_{a\wedge b}}{\sqrt{n\, n_a\, n_{\bar{b}}}}$$

et $\rho(a,b) = \dfrac{n.n_{a\wedge b} - n_a\, n_b}{\sqrt{n_a\, n_b\, n_{\bar{a}}\, n_{\bar{b}}}}$ d'où la relation entre $\rho(a,b)$ et $q(a,\bar{b})$.

Ainsi $q(a,\bar{b}) = 0 \Leftrightarrow \rho(a,b) = 0$ et $\rho(a,b) \geq 0 \Leftrightarrow \varphi(a,\bar{b}) \geq 0{,}5$. Ceci signifie que implication et corrélation linéaire vont plutôt *"dans le même sens"*. Cependant, on peut observer une croissance de l'implication en même temps qu'une décroissance de la corrélation. Ce qui montre bien, qu'outre la dépendance aux effectifs n, $n_{\bar{a}}$ et n_b, le rapport $\frac{\rho}{q}$ indique la non-coïncidence des deux concepts.

La double situation suivante l'illustre :

b_1 / a_1	1	0	Marges
1	61	1	62
0	57	81	138
Marges	118	82	200

$\rho_1(a_1,b_1) = 0{,}537$

$q_1(a_1,\bar{b}_1) = -4{,}843$

b_2 / a_2	1	0	Marges
1	69	5	74
0	48	78	126
Marges	117	83	200

$\rho_2(a_2,b_2) = 0{,}540$

$q_2(a_2,\bar{b}_2) = -4{,}639$

Par suite :

. d'une part, $\mathbf{a_1}$ et $\mathbf{b_1}$ sont moins corrélées que $\mathbf{a_2}$ et $\mathbf{b_2}$,

. d'autre part, l'intensité d'implication de $\mathbf{a_1}$ sur $\mathbf{b_1}$ est plus forte que celle de $\mathbf{a_2}$ sur $\mathbf{b_2}$ puisque $q_1 < q_2$.

Proposition 6. Considérons le χ^2 d'indépendance des variables **a** et **b** ; alors

$$\frac{\chi^2}{q^2(a;\bar{b})} = \frac{n^2}{n_b\, n_{\bar{a}}}$$

La démonstration est aisée, en particulier si l'on remarque que $\chi^2 = n\,\rho^2$.

On constate ainsi que les deux concepts χ^2 et q^2 ne se superposent pas. $q(a,\bar{b})$ est la racine carrée de la contribution à χ^2 de la case $(a,\bar{b})$ du tableau de croisement des 2 variables **a** et **b**. La seule considération, non relative, de χ^2 et des effectifs des 4 cases de ce tableau, comme le font souvent les psychologues, ne peut donc pas rendre compte précisément de l'implication.

Remarque 2. Etudions la sensibilité de q aux faibles variations d'effectif. Supposons pour cela que, par exemple, $n_{a\wedge\bar{b}}$ devienne $n'_{a\wedge\bar{b}} = n_{a\wedge\bar{b}} + k$, où $k \in \mathbf{Z}$, sans que varient n_a et $n_{\bar{b}}$. Alors :

$$q'(a,\bar{b}) = q(a,\bar{b}) + \frac{k}{\sqrt{\frac{n_a\, n_{\bar{b}}}{n}}}.$$

L'effet de l'erreur de mesure k est atténué en $\sqrt{\frac{n}{n_a\, n_{\bar{b}}}}$, ce qui permet dans la plupart des cas de maintenir la validité d'une implication au même seuil ou à un seuil voisin.

1.4. Extension à des variables fréquentielles.

Les problématiques didactiques nous conduisent à des situations où les données issues des faits observés ne sont plus binaires.

a) C'est le cas des faits dont la valeur de vérité n'est pas nécessairement 0 (faux) ou 1 (vrai), mais toute valeur modale de [0,1], quantification d'un "peut-être", "sans doute", rarement, etc. A ce type de variable modale, nous pouvons rattacher les variables ordinales représentant un ordre préférentiel parmi les énoncés proposés. Par exemple, Marc Bailleul, dans sa recherche sur les représentations des mathématiques et de leur enseignement par les professeurs, propose 10 termes associables à des

conceptions personnelles de ceux-ci. Il leur demende d'en choisir 3, jugés représentatifs et de leur associer les valeurs 1, 2 ou 3 par ordre d'adéquation décroissant. La variable ordinale correspondante pourrait donc prendre les valeurs 1, 2, 3 ou 4 auxquelles nous substituons les valeurs modales respectives 1, 2/3, 1/3 et 0 (cas de non choix).

b) C'est aussi le cas de faits répétés (un type d'erreur par exemple) rencontrable dans un questionnaire ou au cours d'une observation. A l'effectif observé, on associera la fréquence du fait en considérant le nombre maximum de fois où il peut se répéter. Par exemple, S.Ag Almouloud a identifié une vingtaine de conduites observables au cours d'une démonstration de géométrie, chacune pouvant apparaître plusieurs fois dans la rédaction de la démonstration.

A l'égard de ces 2 types de variables, nous procédons comme pour les variables binaires.

Soient $(\mathbf{a},\mathbf{b}) \in V^2$ et $i \in E$, α (resp. β) la valeur maximum de **a** (resp. **b**) sur E. α_i (resp. β_i) la **valeur observée** de **a** (resp. **b**) sur i. Posons :

$$\alpha'_i = \frac{\alpha_i}{\alpha},\quad \beta'_i = \frac{\beta_i}{\beta},\quad \nu_a = \sum_{i\in E} \alpha'_i,\quad \nu_b = \sum_{i\in E} \beta'_i,\quad \nu_{a\wedge\bar{b}} = \sum_{i\in E} \alpha'_i\,(1-\beta'_i) \text{ et } \nu_{\bar{b}} = n - \nu_b .$$

Donc posons, comme pour les variables binaires : $q(a,\bar{b}) = \dfrac{\nu_{a\wedge\bar{b}} - \dfrac{\nu_a\,\nu_{\bar{b}}}{n}}{\sqrt{\dfrac{\nu_a\,\nu_{\bar{b}}}{n}}}$

qui sera, pour $\nu_a \le \nu_b$, l'indicateur retenu pour l'implication statistique de **a** sur **b**.

§ 2 - IMPLICATION ENTRE CLASSES DE VARIABLES.

Elle ne prend véritablement son sens qu'à condition qu'à l'intérieur de chaque classe de variables dont on examine la relation avec d'autres, existe une certaine "**cohésion**" entre les variables qui la constituent, ceci afin que le "*flux*" implicatif d'une classe $\mathcal{A}$ sur une classe $\mathcal{B}$ soit nourri d'un "flux" interne à $\mathcal{A}$ et alimente un "*flux*" interne à $\mathcal{B}$. Cette cohésion, généralement nourrie de cohérence sémantique ou, dans le cas de la didactique, de conditions psychologiques, cognitives, situationnelles, etc., doit se traduire ici par une mesure (quantitative). On pourrait penser qu'un ensemble d'indices de similarité assez élevés entre les éléments de la classe serait un bon indicateur de cohésion. Nous ne retenons pas cette approche qui ne rendrait compte que d'une cohésion de profils symétriquement comparables, ne restituant pas une dynamique interne orientée (donc non symétrique). Or, nous disposons avec les intensités d'implication entre variables d'un instrument de mesure d'un emboîtement de deux parties d'une population E. Par exemple, si dans la classe à 3 éléments *a*, *b*, *c*, on observe : $\varphi(a,\bar{b}) = 0{,}97$, $\varphi(b,\bar{c}) = 0{,}95$, $\varphi(a,\bar{c}) = 0{,}92$, on pourra dire que la classe orientée de *a* vers *c* admet une bonne cohésion.

Ce ne serait pas le cas si $\varphi(a,\bar{b}) = 0{,}82$, $\varphi(b,\bar{c}) = 0{,}38 < \varphi(c,\bar{b}) = 0{,}70$ et $\varphi(a,\bar{c}) = 0{,}48 > \varphi(c,\bar{a})$.

C'est donc cette voie que nous choisissons pour une cohésion implicative donc orientée, comme peut l'être une filiation procédurale ou une genèse. Nous verrons ensuite quel indicateur permettrait de rendre compte d'une extension aux classes, de la notion d'implication.

2.1. Cohésion implicative.

Afin d'en améliorer l'intuition, nous la définirons progressivement pour 2, puis 3 et r éléments de classe.

La cohésion, se voulant indicateur d'ordre implicatif au sein d'une classe de variables, s'oppose en cela au"**désordre**" dont rend compte l'entropie d'une expérience aléatoire. Rappelons au sujet de celle-ci que, X étant une variable aléatoire prenant ses valeurs dans $S = \{m_1,m_2,..,m_k\}$ muni de la loi

$\{p_1,p_2,...,p_k\}$, l'entropie est l'espérance mathématique de la variable I(X) prenant les valeurs $I(m_1)$, $I(m_2),...,I(m_k)$; $I(m_j)$ est l'incertitude sur $\{m_j\}$ ou information apportée par la réalisation de $\{m_j\}$. Ainsi :

$$\mathcal{E}[I(X)] = \sum_{j=1}^{k} - p_j \log_2 p_j$$

est **l'entropie de l'expérience.**

2.1.1. Cas de 2 éléments : classe (a,b).

Supposons $n_a \leq n_b$. Nous allons définir la cohésion du couple (a,b).

Soit χ la variable aléatoire indicatrice de l'évènement $[Q(a,\bar{b}) \geq q(a,\bar{b})]$. Alors :

$$\Pr(\chi = 1) = \varphi(a,\bar{b}) = p$$

et $$\Pr(\chi = 0) = 1 - \varphi(a,\bar{b}) = 1 - p.$$

L'entropie ou incertitude de cette expérience est alors :

$$\mathcal{E}[I(\chi)] = -p \log_2 p - (1-p)\log_2(1-p)$$.

Par exemple :

. si $\varphi(a,\bar{b}) = p = 0{,}95$, alors :

$$\mathcal{E}[I(X)] = \frac{-0{,}95 \ell n\, 0{,}95 - 0{,}05 \ell n\, 0{,}05}{\ell n\, 2} = 0{,}286.$$

. si $\varphi(a,\bar{b}) = 1$, alors $\mathcal{E}[I(\chi)] = 0$ en convenant que $0\ \ell n\ 0 = 0$

. si $\varphi(a,\bar{b}) = 0{,}5$, alors $\mathcal{E}[I(\chi)] = 1$ (entropie maximale).

Mais si $\varphi'(a,\bar{b}) = 1 - \varphi(a,\bar{b})$, alors $\mathcal{E}[I(\chi)] = \mathcal{E}[I(1-\chi)]$.

Précisément, en posant : $\mathcal{E} = f(p) = -p \log_2 p - (1-p) \log_2 (1-p)$, on a : $f(1-p) = f(p)$.

L'entropie est donc symétrique par rapport à $p = 0{,}5$.

De plus, $\frac{df}{dp} = \log_2 \frac{1-p}{p}$ pour $p \in]0,1[$

donc $\mathcal{E}$ croît de 0 à 1 sur $]0;0{,}5]$ et décroît de 1 à 0 sur $[0{,}5;1[$.

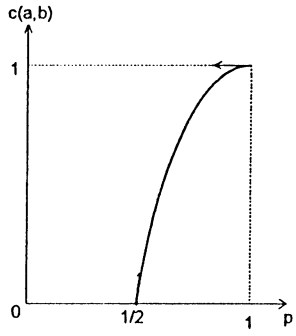
Aussi, la propriété de symétrie de $\mathcal{E}$ allant à l'encontre de la dissymétrie de la quasi-implication, nous retiendrons finalement, comme indicateur de cohésion, l'application c définie sur $V \times V$ par :

* **Définition 2.** *La cohésion du couple de variables (a,b) tel que $n_a \leq n_b$ est le nombre c(a,b) où :*

> *. si* $\varphi(a,\bar{b}) = p \geq 0{,}5$, $c(a,b) = [1 - [p \log_2 p + (1-p)\log_2(1-p)]^2]^{1/2} = \sqrt{1-\mathcal{E}^2}$
> *. et si* $\varphi(a,\bar{b}) = p < 0{,}5$, $c(a,b) = 0$ *(absence de cohésion).*

La fonction "**carré de l'entropie**" est choisie pour des raisons de renforcement du contraste sur [0,1]. Nous prenons la racine carrée de son complément à 1 pour donner à la cohésion la dimension de l'entropie et pour accroître sa valeur numérique (en effet pour $x \in [0,1]$, $\sqrt{1-x^2} \geq 1-x$).

* **Etude aux limites.**



Posons : $c(a,b) = g(\mathcal{E}) = g[f(p)]$

$$\frac{dg}{d\mathcal{E}} = -\frac{\mathcal{E}}{\sqrt{1-\mathcal{E}^2}} \qquad \text{et} \qquad \frac{dc}{dp} = \frac{-\mathcal{E}}{\sqrt{1-\mathcal{E}^2}} \times \log_2 \frac{1-p}{p}.$$

$c(a,b)$ croît donc de 0 à 1 quand $p = \varphi(a,\bar{b})$ croît de 0,5 à 1. La fonction c de p est continue en $p = \frac{1}{2}$.

Nous prolongeons par continuité en prenant :

$c(a,b) = 1$ lorsque $p = 1$ (c'est-à-dire lorsque l'implication $a \Rightarrow b$ est stricte).

* **Remarque 3.** Rappelons un résultat démontré dans la thèse d'A. LARHER :

si $n_a \leq n_b$ et si $q(a,\bar{b}) \leq 0$ alors l'intensité d'implication de a sur b est supérieure à celle de b sur a. Ce qui signifie que :

$$n_a \leq n_b \text{ et } q(a,\bar{b}) \leq 0 \Rightarrow \varphi(a,\bar{b}) = \max(\varphi(a,\bar{b}), \varphi(b,\bar{a})).$$

Appelant **classe** (a,b) le couple (a,b) tel que $n_a \leq n_b$, la cohésion de la classe (a,b) est définie sans équivoque à partir de la plus grande des intensités relatives à $(a \Rightarrow b)$ et $(b \Rightarrow a)$.

D'où la :

* **Définition 2'.**

> *La cohésion de la classe (a,b) est le nombre c(a,b) tel que :*
> *. si* $p = max\ [\varphi(a,\bar{b}), \varphi(b,\bar{a})] \geq 0{,}5$ *et* $\mathcal{E} = -p \log_2 p - (1-p)\log_2(1-p)$
> $c(a,b) = \sqrt{1-\mathcal{E}^2}$
> *. si* $p = 1$ $c(a,b) = 1$
> *. si* $p \leq 0{,}5$ $c(a,b) = 0$.

* **Remarque 4.** Lorsque $b = a$ et que $n_a \in [\frac{n-\sqrt{n(n-11)}}{2} ; \frac{n+\sqrt{n(n-11)}}{2}]$, nous avons vu que (cf. p. 7) $0{,}95 \leq \varphi(a,\bar{a}) \leq 1$. La cohésion de la classe (a,a) a donc un sens et, en général, puisque $\lim_{p\to 1^-} c = 1$, cette cohésion est très voisine de 1. Par définition, nous poserons donc :

$$\forall a \in V \qquad c(a,a) = 1.$$

2.1.2. Cas de 3 éléments a, b et c.

Six valeurs d'intensité correspondent a priori à l'ensemble $\dot{A} = \{a,b,c\}$:

$$\varphi(a,\bar{b}),\ \varphi(a,\bar{c}),\ \varphi(b,\bar{a}),\ \varphi(b,\bar{c}),\ \varphi(c,\bar{a}) \text{ et } \varphi(c,\bar{b}).$$

Nous souhaitons que l'indice de cohésion implicative contienne l'information révélée par les relations implicatives binaires entre tous les éléments de l'ensemble $\dot{A}$. Mais, en même temps, pour conserver la dynamique dissymétrique de l'implication, seule la relation la plus puissante entre deux éléments quelconques reste pertinente par rapport à notre objectif. Par suite, parmi toutes les associations 3 à 3 ne faisant intervenir qu'une fois chaque couple d'éléments de $\{a,b,c\}$ et restituant au mieux la puissance de certaines implications, nous retenons les valeurs :

$$\max[\varphi(a,\bar{b}),\ \varphi(b,\bar{a})],\ \max[\varphi(a,\bar{c}),\ \varphi(c,\bar{a})] \text{ et } \max[\varphi(b,\bar{c}),\ \varphi(c,\bar{b})].$$

Comme précédemment, dans le cas où ils sont supérieurs ou égaux à 0,5, les maxima obtenus sont compatibles avec l'ordre des effectifs n_a, n_b et n_c. Par exemple, si $n_a \leq n_b \leq n_c$, les trois maxima sont : $\varphi(a,\bar{b})$, $\varphi(a,\bar{c})$ et $\varphi(b,\bar{c})$.

Définition 3. *Le couple $\mathcal{A} = ((a,b),c)$ sera encore appelé* **classe** *et sa cohésion implicative sera définie ainsi :*

$$C(\mathcal{A}) = [c(a,b) \times c(b,c) \times c(a,c)]^{1/3}$$

moyenne géométrique des cohésions des classes à 2 éléments

La préférence accordée à la moyenne géométrique plutôt qu'à la moyenne arithmétique tient à notre volonté, d'une part d'obtenir une cohésion nulle pour une classe dès que la cohésion d'un de ses couples est nulle, c'est-à-dire dès que les implications mutuelles sont inférieures ou égales à 0,5, d'autre part de "**ramener**" $C(\mathcal{A})$ au voisinage de 1 lorsque les cohésions des couples sont assez fortes.

2.1.3. Cas de r éléments $a_1, a_2,...,a_r$.

Nous opérons comme ci-dessus, c'est-à-dire en retenant les maxima des intensités d'implication entre 2 éléments quelconques de l'ensemble $\dot{A} = \{a_1,a_2,...,a_r\}$. A ces maxima sont associés les cohésions implicatives des couples et l'ordre induit sur $\dot{A}$ par les effectifs n_{a_1}, n_{a_2} ,..., n_{a_r}. Par exemple, si $n_{a_1} \leq n_{a_2} \leq ... \leq n_{a_r}$, nous appellerons <u>classe</u> le couple $\mathcal{A} = (((a_1,a_2),a_3),...,a_r)$, et comme il y a $\frac{r(r-1)}{2}$ paires, sa cohésion implicative sera :

Définition 4.

$$C(\mathcal{A}) = \Big[\prod_{\substack{i \in \{1,...r-1\} \\ j \in \{2,...,r\} \\ j>i}} c(a_i,a_j) \Big]^{\frac{2}{r(r-1)}}$$ *moyenne géométrique des cohésions de classes à 2 éléments*

2.2. Implication entre classes.

Nous souhaitons que l'implication entre deux classes se constitue à partir des informations suivantes :

- les cohésions respectives des 2 classes,

- une intensité d'implication extrémale des éléments d'une classe sur les éléments de l'autre,

- les cardinaux respectifs des 2 classes.

Chacune de ces informations crédite l'indice que nous retiendrons si :

- l'indice croît avec les cohésions de chaque classe et s'annule lorsque la cohésion de l'une d'entre elles est nulle,

- l'indice croît avec la liaison extrémale (minimale si l'on vise un degré d'exigence élevé, maximale si l'on recherche une souplesse réaliste),

- l'indice décroît avec les cardinaux des classes, eu égard à la prise en compte d'une liaison maximale.

Posons : . $\dot{A}$ et $\dot{B}$ deux parties disjointes : $\dot{A} = \{a_1,...,a_r\}$ et $\dot{B} = \{b_1,...,b_s\}$,

. $\mathcal{A}$ et $\mathcal{B}$ les classes qui leur sont respectivement associées ,

. $C(\mathcal{A})$ et $C(\mathcal{B})$ leurs cohésions respectives.

Conformément aux lois de probabilité des sup. de variables aléatoires, a priori uniformément distribuées, nous définissons l'indice d'implication $\psi(\mathcal{A},\mathcal{B})$ de la classe $\mathcal{A}$ vers la classe $\mathcal{B}$ par :

Définition 5.

$$\psi(\mathcal{A},\mathcal{B})=\left[\sup_{\substack{i\in\{1,\ldots,r\}\\ j\in\{1,\ldots,s\}}} \varphi(a_i,\bar{b}_j)\right]^{rs} \times [C(\mathcal{A})\times C(\mathcal{B})]^{\frac{1}{2}}$$

L'expression $[C(\mathcal{A})\ C(\mathcal{B})]^{\frac{1}{2}}$ représente la cohésion moyenne (géométrique) de $\mathcal{A}$ *et* $\mathcal{B}$; elle intègre les informations de cohésivité des 2 classes en jeu ; de plus, cette expression est telle que si $C(\mathcal{A})$ et $C(\mathcal{B})$ sont simultanément multipliées par k, alors $\psi(\mathcal{A},\mathcal{B})$ est multipliée par k.

Dans sa thèse, A.Totohasina étudie, par des simulations, la loi de variation de ψ(A,B) afin de munir cette implication d'un "seuil de crédibilité" comparable à celui des variables seules et fonction des effectifs de variables constituant respectivement A et B.

2.3. Agrégations successives des classes.

Selon l'objectif classique des méthodes de classification hiérarchique ascendante, nous avons construit un algorithme d'agrégations successives des classes basé sur le critère de leur cohésion maximale. S. Ag Almouloud a réuni dans un même logiciel, CHIC (Classification Hiérarchique Implicative et Cohésitive), le traitement des analyses de similarité et d'implication entre variables de nature différentes, ainsi que la construction graphique des hiérarchies orientées correspondantes. H. Ratsimba-Rajohn considère, dans sa thèse en cours, les niveaux significatifs de la hiérarchie, moments cruciaux de formation d'une typologie orientée par l'implication de classes. Par ailleurs, il s'intéresse à la contribution des individus et de leurs descripteurs par rapport à cette typologie.

L'agrégation en classes, comme le graphe implicatif, conduit à la mise en évidence de phénomènes orientés, assimilables quelquefois à des conceptions dont on restituerait la genèse. C'est le cas de la recherche d'A. Totohasina qui, sur l'acquisition par les élèves de Terminale de la notion de probabilité conditionnelle, a mis en évidence, entre autres, 2 conceptions : l'une causaliste, l'aute chronologiste, avec leur forme initiale et achevée.

§ 3 – QUELQUES EXEMPLES TRAITES

3.1. Questionnaire sur de courtes démonstrations de géométrie.

3.1.1 Présentation du questionnaire

FAITS

1 (EF) et (CD) sont symétriques par rapport au point I
2 [MN] est le symétrique de [PR] par rapport au point I
3 (AB) et (CD) sont symétriques par rapport au point 0
4 (MN) // (PR)
5 (CD) // (EF)
6 (AB) // (CD)
7 (AB) // (EF)
8 MN = PF
9 CD = EF
10 AB = CD
11 AB = EF

THEOREMES

1 La symétrie centrale conserve les longueurs.
2 Si (D) // (D') et (D') // (D") alors (D) // (D").
3 Le symétrique d'une droite (D) par rapport à un point est une droite (D') parallèle à (D).
4 Si deux droites sont symétriques par rapport à un point alors elles sont parallèles.
5 Deux segments symétriques par rapport à un point ont même longueur.
6 La symétrie centrale conserve les directions.

En fait, chaque question se présente schématiquement ainsi :

Hypothèse : fait n° p **Théorème** : n° q **Conclusion** : fait n° ?

Schématiquement, l'ensemble questions-réponses peut être présenté ainsi :

	HYPOTHESES	THEOREME	? CONCLUSION à trouver
Q1 { Hypothèse : 1 / Théorème : 3 → / Conclusion : 5	(EF) et (CD) symétriques par rapport à I	Le symétrique de (D) par rapport à un point est (D') // (D)	(EF) // (CD)
Q2 { Hypothèse : 3 / Théorème : 4 → / Conclusion : 6	(AB) et (CD) symétriques par rapport à 0	Si 2 droites sont symétriques par rapport à un point, alors elles sont parallèles	(AB) // (CD)
Q3 { Hypothèse : 2 / Théorème : 5 → / Conclusion : 8	[MN] est symétrique de [PR] par rapport à I	2 segments symétriques par rapport à un point ont même longueur	MN = PR
Q4 { Hypothèse : 3 / Théorème : 6 → / Conclusion : 6	(AB) et (CD) symétriques par rapport à 0	La symétrie centrale conserve les directions	(AB) // (CD)
Q5 { Hypothèse : 6 et 5 / Théorème : 2 → / Conclusion : 7	(AB) // (CD) et (CD) // (EF)	Si (D) // (D') et (D') // (D"), alors (D) // (D")	(AB) // (EF)
Q6 { Hypothèse : 2 / Théorème : 1 → / Conclusion : 8	[MN] est symétrique de [PR] par rapport à I	La symétrie centrale conserve les longueurs	MN = PR

Exemple. 3-6-10 (Q_4).

Hypothèse : (AB) et (CD) sont symétriques par rapport à 0.

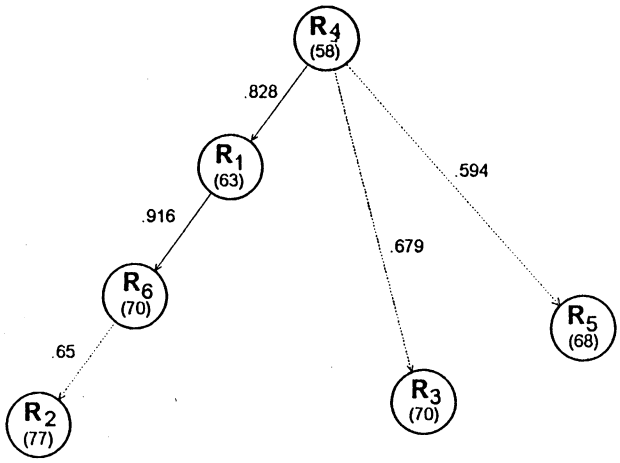
Théorème : la symétrie centrale conserve les directions.

Conclusion donnée par l'élève : AB = CD.

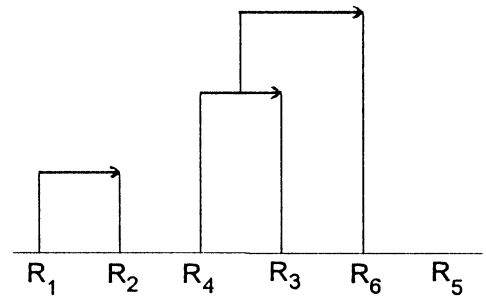
Ce questionnaire a été proposé à l'aide du logiciel "Premier Pas" à 80 élèves de 4ème. Compte tenu des 31 comportements de réponse des élèves, nous disposons d'un tableau 80*31 dont les régularités, les traits pertinents, en particulier les relations implicatives entre les modalités, sont impossibles à analyser "à la main".

3.2. Graphe implicatif et hiérarchie implicative entre les réussites.

Le graphe implicatif des réussites du "6 questions" traduit une relation de préordre partiel dans cet ensemble. Il est donc orienté et valué. R_4, réussite à la question (Q_4) la plus complexe du questionnaire (et la moins bien réussie), en est la source. R_2, R_3, R_5 en sont les puits :



Ci-contre, nous présentons le graphe implicatif entre les 6 modalités de réussite.



hiérarchie cohésitive

Les valeurs indiquées sont les intensités des implications.
Sur le même chemin, les réussites se placent dans l'ordre croissant de leurs effectifs.

R_2 , réussite à la question (Q_2) dont le théorème est exprimé en "si .... alors", formulation facilitant, semble-t-il, le succès.

R_3 , relative à la conservation des longueurs dans la symétrie centrale. La longueur est une notion plus familière à l'élève que celle de direction.

R_5 , seule question relative à la transitivité du parallélisme.

Notons la séparation de R_5 des autres réussites, ce qui ne se produit pas en analyse des similarités où l'on force celles-ci, même lorsqu'elles sont faibles. Je renvoie à la thèse d'A. Larher pour lire l'analyse fouillée des graphes de réussites et de toutes les modalités de réponse, de même que les hiérarchies implicatives associées.

3-2 Enquête sur les représentations des enseignants relativement à leur pratique

3-2-1 Présentation de l'enquête menée par Marc Bailleul

L'objectif de ce questionnaire est de faire apparaître certains aspects des représentations que se font les enseignants de mathématiques de l'enseignement de leur discipline, de plusieurs points de vue.

En effet, cette représentation peut être différente, nous dit M. Bailleul, selon qu'il s'agit :

- de leur propre enseignement des mathématiques,
- de leur idéal de l'enseignement des mathématiques,
- de l'enseignement des mathématiques tel qu'ils le perçoivent autour d'eux
- de l'enseignement des mathématiques tel que l'institution l'attend d'eux (à leur avis).

Pour cela, il leur est demandé de choisir, pour chacun des 4 points de vue, dans chacune des 3 colonnes proposées, 3 mots exactement qu'ils numéroteront de 1 (celui auquel ils accordent le plus d'importance) à 3.

MA PERCEPTION DE MON ENSEIGNEMENT DES MATHEMATIQUES.

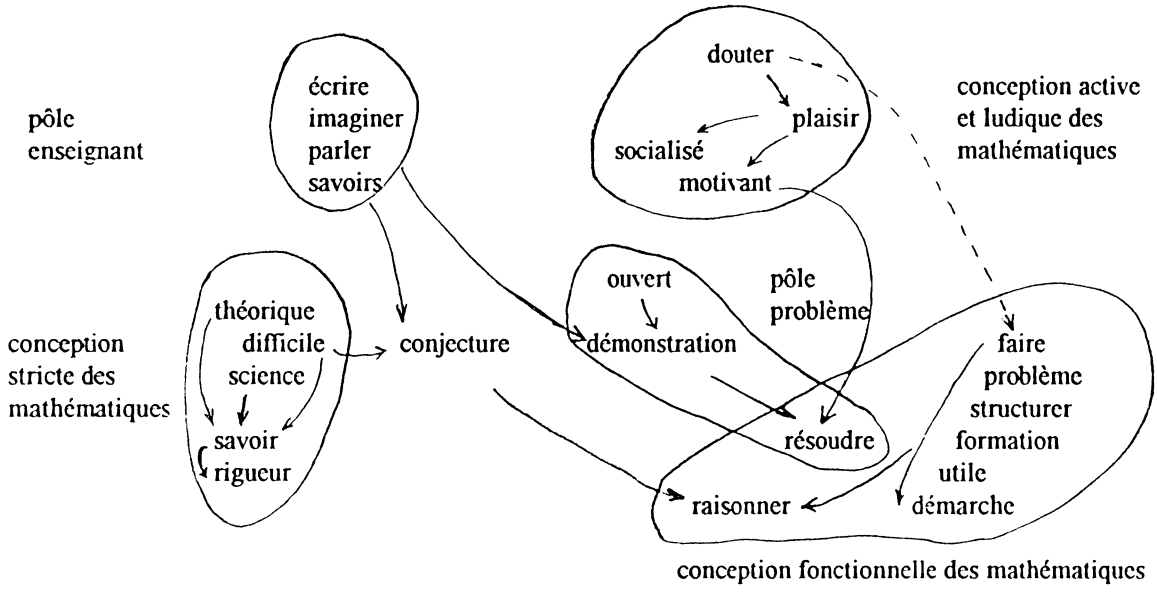
théorique		faire		rigueur	
symbolique		parler		démonstration	
concret		écrire		conjecture	
motivant		raisonner		science	
lassant		structurer		savoirs	
socialisé		savoir		jeu	
individuel		imaginer		plaisir	
difficile		douter		problème	
utile		appliquer		démarche	
ouvert		résoudre		formation	

Disposant des réponses d'une trentaine d'enseignants, nous avons procédé à une analyse des variables modales définies par les choix hiérarchisés des enseignants et obtenues à partir du point de vue "mon enseignement".

Bien entendu, il serait intéressant, par rapport aux mêmes expressions, d'analyser et comparer les représentations selon les 3 autres points de vue.

3-2-2 Analyse implicative

Schématiquement, nous obtenons l'arbre suivant :



CONCLUSION.

La méthode présentée ici apparaît donc en plein développement théorique et, compte tenu de ses premières applications, semble un complément statistique très riche, dépassant de trèsl oin l'intérêt d'un simple test statistique. C'est en effet un complément original des autres méthodes d'analyse de données. Ses extensions actuelles et la programmation des algorithmes de calcul autorisent maintenent son usage au même titre que les autres méthodes, en leur ajoutant son approche dissymétrique riche d'informations en didactique comme en intelligence artificielle.

REFERENCES

[ACID S., de CAMPOS L.M., GONZALEZ A., MOLINA R., PEREZ de la BLANCA N. 1991], Learning with Castle - in Symbolic and quantitative approaches to uncertainty (R. KRUSE, P. SIEGEL), Springer-Verlag, 99-106.

[AMARGE S., DUBOIS D., PRADE H. 1991], Imprecise quantifiers and conditional probabilities - in Symbolic and quantitative approaches to uncertainty (R. KRUSE, P. SIEGEL), Springer-Verlag, 33-37.

[DIDAY E. 1991] - Towards a statistical theory of intentions for knowledge analysis, rapport de recherche 1494, INRIA Rocquencourt.

[GAMMERMAN A., LUO Z. 1991] - Constructing Causal Trees from a medical database, Technical Report TR 91 002, Dep^{t} of Computer Sci., Heriot-Watt Univ., Edimburgh.

[GANASCIA J.G. 1991] - CHARADE : Apprentissages de bases de connaissances dans "Induction symbolique - numérique à partir de données", Ed. KODRATOFF et DIDAY, CEPADUES, 1991.

[GRAS R., 1979] - Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques, Thèse d'Etat, Université de Rennes I, octobre 1979.

[GRAS R. et LARHER A., 1989] - La quasi-implication : une méthode d'analyse de relations non symétriques entre attributs et entre classes d'attributs, Public. interne I.R.M.A.R., Rennes, 1989.

[GUIGUES J.L. et DUQUESNE V. 1986] - Familles minimales d'implications informatives résultant d'un tableau de données binaires, Mathématiques et Sciences Humaines n° 95, p. 5-18, 1986.

[LARHER A., 1991] - Implication statistique et applications à l'analyse de démarches de preuve mathématique, Thèse de l'Université de Rennes I, février 1991.

[LERMAN I.C., GRAS R., ROSTAM H., 1981] - Elaboration et évaluation d'un indice d'implication pour des données binaires, I et II, Mathématiques et Sciences Humaines n° 74, p 5-35 et n° 75, p 5-47, 1981.

[LERMAN I.C., 1981] - Classification et analyse ordinale des données, Dunod, 1981.

[LOEVINGER J. 1947] - A systematic approach to the construction and evaluation of tests of ability, Psychological Monograhs, 61, n° 4.

[PEARL J. 1988] - Probabilistic Reasoning in intelligent systems, San Mateo, CA, Morgan Kaufmann.

L'ANALYSE STATISTIQUE BAYÉSIENNE

DOMINIQUE CELLIER (*)

...Toi, qui de l'univers en marche ne sait rien
Tu es bâti de vent : par suite tu n'es rien.
Ta vie est comme un pont jeté entre deux vides :
Tu n'as pas de limite, au milieu tu n'es rien...
Omar Khayyâm

- Introduction -

La difficulté de cet exposé réside dans le fait qu'il n'est pas évident de présenter en si peu de temps une théorie qui, d'une part, est un véritable champ de bataille de polémiques et qui, d'autre part, est peu développée en France tant au niveau de l'enseignement de la statistique qu'au niveau des applications pratiques.

Les raisons de cet état de fait seraient longues à exposer ici. Cependant on peut penser que "l'objectivité" des esprits dits "cartésiens" ne peut être que choquée, ignorante parfois ou méprisante vis à vis d'une théorie qui fait "apparemment" appel à la "subjectivité" : l'analyse bayésienne prend fondamentalement en compte les "a priori", l'apprentissage et l'expérience acquise de l'expérimentateur. Certains aspects élémentaires, intuitifs d'un point de vue mathématique font préférer souvent à l'analyse statistique bayésienne des théories plus complexes mathématiquement.

L'analyse statistique bayésienne est essentiellement un principe de dualité et une démarche cohérente d'inversion.

Le travail d'un statisticien consiste en dernière analyse à "*remonter des effets aux causes*". On observe les effets d'un phénomène et on cherche, sur la base de cette observation, à faire une déduction (une inférence) sur les causes qui provoquent ces effets :

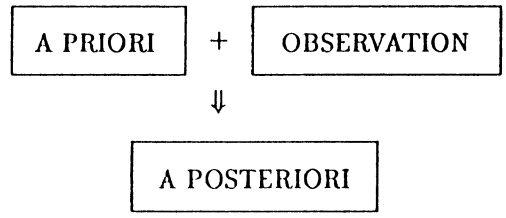
- estimer des paramètres inconnus,
- émettre, accepter ou rejeter des hypothèses,
- prédire des observations futures ...etc...

La statistique inférentielle est donc une démarche d'"*inversion*".

EFFETS $\hookrightarrow$? CAUSES

(*) Université de Rouen - Laboratoire Analyse et Modèles Stochastiques - U.R.A. C.N.R.S. 1378.

La démarche bayésienne est sans doute celle qui s'inscrit de façon la plus cohérente dans cette problématique d'inversion : elle met en œuvre effectivement cette dernière qui consiste à remonter des effets (les observations) aux causes (les paramètres).



L'a posteriori consiste en une réactualisation de la connaissance ou du degré d'ignorance du ou des paramètres.

- I - Généralités sur la démarche bayésienne -

...Quand vous avez éliminé l'impossible
Ce qui reste, même improbable,
Doit être la vérité...
Conan Doyle

I.1 - Statistique inférentielle et théorie des probabilités.

Quel rapport existe-t-il entre la Statistique qui repose sur l'observation de phénomènes concrets et la théorie des probabilités qui traite des propriétés de certaines structures modélisant des phénomènes où le hasard intervient ?

1 - Les données observées sont imprécises, entachées d'erreurs. Le modèle probabiliste permet de les présenter comme des variables aléatoires (l'aléa provenant de la déviation entre vraies valeurs et valeurs observées).

2 - On constate parfois que la répartition statistique d'une variable au sein d'une population est voisine de modèles mathématiques proposés par le calcul des probabilités.

3 - On est souvent amené à modéliser des situations très complexes (nombre important de paramètres, paramètres cachés...). Le modèle proposé est alors simplifié grâce au calcul des probabilités.

4 - Enfin, le lien le plus important réside dans le fait que les échantillons d'individus observés sont la plupart du temps tirés au hasard dans la population, ceci pour en assurer la "représentativité". Chaque individu a alors une certaine probabilité d'appartenir à l'échantillon. Les caractéristiques observées deviennent, grâce au "tirage au sort", des variables aléatoires dont le calcul des probabilités permet d'étudier les propriétés, le comportement etc...

I.2 - Modèle statistique et vraisemblance.

Compte tenu de ce qui précède, on comprend bien le rôle important joué par la modélisation en statistique inférentielle. Un modèle statistique consiste en l'observation d'une variable aléatoire X, de loi de probabilité P_θ. Le modèle peut alors s'écrire

$$\left(\mathcal{X}, (P_t)_{\theta\in\Theta}\right)$$

où

- $\mathcal{X}$ est l'espace des observations (les effets)
- P_θ est la loi de l'observation qui dépend d'un paramètre inconnu $\theta \in \Theta$ (représentant les causes).

L'espace des observations peut être discret. Par exemple X peut prendre ses valeurs dans $\{0, 1, 2, \cdots, n\}$ et avoir pour loi une loi Binomiale de paramètre (n, θ), $\theta \in [0, 1]$ inconnu. Dans ce cas on a

$$\forall k \in \{0, 1, 2, \cdots, n\} \qquad P_\theta(X = k) = C_n^k \theta^k (1-\theta)^{n-k} = l(\theta, k)$$

La fonction $l(\theta, \cdot)$ est appelée la *vraisemblance*.

L'espace des observations peut être continu. Par exemple X peut prendre ses valeurs dans $\mathbf{R}$ et avoir pour loi une loi de Laplace-Gauss $\mathcal{N}(m, \sigma^2)$, $\theta = (m, \sigma^2)$ inconnu. Dans ce cas on a

$$\forall [a.b] \subset \mathbf{R} \qquad P_\theta(X \in [a, b]) = \int_a^b \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{1}{2\sigma^2}(x-m)^2\right) dx$$
$$= \int_a^b l(\theta, x) dx$$

La fonction $l(\theta, \cdot)$ est encore appelée la *vraisemblance*.

Une fois le modèle statistique construit, on cherche à établir, sur la base de l'observation et de sa vraisemblance, une inférence sur le paramètre θ inconnu.

I.3 - Le théorème de Bayes.

De manière générale, cette démarche d'inversion qui consiste à remonter des effets aux causes est décrite par le *Théorème de Bayes* : si A et E sont deux évènements tels que $P(E) \neq 0$, on a

$$P(A \mid E) = \frac{P(E \mid A) \cdot P(A)}{P(E \mid A) \cdot P(A) + P(E \mid \bar{A}) \cdot P(\bar{A})}$$

Ce théorème est une simple conséquence de la définition de la probabilité conditionnelle

$$P(A \mid E) = \frac{P(A \cap E)}{P(E)}$$

Ce théorème a constitué un saut conceptuel majeur dans l'histoire de la théorie des probabilités et de la statistique : c'est la première formule d'inversion des probabilités. En termes d'apprentissage, il décrit l'actualisation de la vraisemblance de A après que E ait été observé. En termes statistiques, il actualise l'information sur le paramètre inconnu θ (les causes) au vu de l'observation x (l'effet).

Ce théorème fondamental est à la base de la "*statistique bayésienne*" : il met sur un pied d'égalité causes et effets, tous deux pouvant être probabilisés.

- Exemple.

Dans une usine, deux machines M_1 et M_2 fabriquent des boulons de même type. M_1 sort en moyenne 0,3% de boulons défectueux et M_2 en sort 0,8% .

On mélange dans une caisse 250 boulons provenant de M_1 et 750 de M_2. On tire au hasard 1 boulon dans la caisse, on constate qu'il est défectueux. *Quelle est la probabilité qu'il ait été fabriqué par* M_1 *?*

Lorsqu'on tire un boulon au hasard, *a priori* la probabilité qu'il provienne

de M_1 est 0,25

de M_2 est 0,75 .

Si on observe qu'il est défectueux (évènement D), on calcule les probabilités conditionnelles $P(M_1 \mid D)$ et $P(M_2 \mid D)$ pour répondre à la question

$$P(M_1 \mid D) = \frac{P(D \mid M_1) \cdot P(M_1)}{P(D \mid M_1) \cdot P(M_1) + P(D \mid M_2) \cdot P(M_2)}$$
$$= \frac{0,003 \cdot 0,25}{0,003 \cdot 0,25 + 0,008 \cdot 0,75}$$
$$\simeq 0,11$$

Evidemment on a $P(M_2 \mid D) \simeq 0,89$.

$P(M_1 \mid D)$ et $P(M_2 \mid D)$ sont les probabilités *a posteriori* sachant que le boulon observé est défectueux.

- version continue du théorème de Bayes.

Supposons que l'observation X à valeurs réelles est de vraisemblance $l(\theta, \cdot)$, $\theta \in \mathbf{R}$ inconnu. Supposons que θ ait une loi Π de densité $\pi(\cdot)$, c'est-à-dire

$$\forall [a,b] \subset \mathbf{R} \qquad \Pi([a,b]) = \int_a^b \pi(t)dt$$

Alors, a posteriori, la loi conditionnelle $\Pi(\cdot \mid X = x)$ de θ sachant qu'on a observé $X = x$ a une densité $\pi(\cdot \mid x)$ donnée par

$$\pi(\theta \mid x) = \frac{l(\theta, x) \cdot \pi(\theta)}{\int_{-\infty}^{+\infty} l(t, x) \cdot \pi(t)dt}$$

Le numérateur de l'expression est la densité de la *loi conjointe* du couple (θ, X), le dénominateur est la densité, notée ρ, de la *loi prédictive* de X. Alors

$$\forall [a,b] \subset \mathbf{R} \qquad \Pi(\theta \in [a,b] \mid X = x) = \int_a^b \pi(t \mid x)dt$$

I.4 - Modèle statistique bayésien.

La problématique de l'analyse statistique bayésienne consiste à introduire une loi de probabilité Π sur l'espace des paramètres : idée révolutionnaire qui continue à diviser les statisticiens.

On passe alors de la notion de *paramètre inconnu* à la notion de *paramètre aléatoire*. Cette loi de probabilité sur l'espace des paramètres est appelée *loi a priori*.

La première question qui vient à l'esprit est évidemment : que représente en fait cette loi a priori ?

- Dans certains cas, le paramètre inconnu est réellement aléatoire ou peut être perçu comme tel.
- Mais dans la majorité des cas c'est impossible et c'est là que réside l'argument principal des adversaires de l'approche bayésienne.

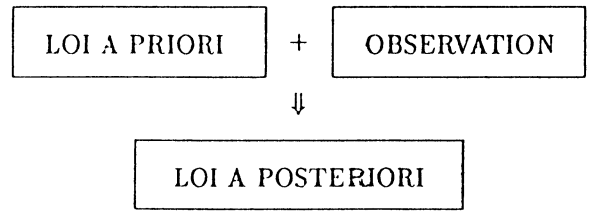
Prenons un exemple : La vitesse c de la lumière est en fait à jamais inconnue du fait de la limitation des appareils de mesure. Il est alors légitime de considérer c comme une variable aléatoire uniforme dans l'intervalle $[c_0 - \epsilon, c_0 + \epsilon]$ où

- ϵ représente la précision maximale actuelle des appareils de mesure
- et c_0 la mesure usuellement retenue.

L'importance de la loi a priori réside dans le fait qu'elle représente un moyen efficace de résumer l'information a priori disponible sur le paramètre inconnu ainsi que l'incertitude sur la valeur de cette information.

I.5 - Loi a priori, loi a posteriori.

L'essentiel des méthodes de l'analyse statistique bayésienne consiste, sur la base de la loi a priori et de l'observation effectuée, à déterminer la loi du paramètre conditionnellement à l'observation : la *loi a posteriori* qui actualise l'information sur le paramètre



Toute la statistique bayésienne repose sur cette loi a posteriori. Les difficultés essentielles proviennent, d'une part, de la détermination et du choix de la loi a priori et, d'autre part, du calcul explicite de la loi a posteriori.

Illustrons cela sur un exemple.

I.6 - Exemple

On observe une variable aléatoire réelle X dont la loi est une loi de Laplace-Gauss $\mathcal{N}(\theta, 1)$ où
- la moyenne $\theta \in \mathbf{R}$ est inconnue
- et la variance est connue et égale à 1.

On cherche, sur la base d'une observation, à estimer la paramètre θ.
Supposons que l'on choisisse comme loi a priori pour θ une loi de Laplace-Gauss $\mathcal{N}(\mu, \tau^2)$.
Déterminons la loi a posteriori. En vertu de la version continue du théorème de Bayes on a

$$\pi(\theta \mid x) = \frac{\exp\left(-\frac{1}{2}(x-\theta)^2\right) \cdot \exp\left(-\frac{1}{2\tau^2}(\theta-\mu)^2\right)}{\int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2}(x-\theta)^2\right) \cdot \exp\left(-\frac{1}{2\tau^2}(\theta-\mu)^2\right) d\theta}$$

Le calcul du dénominateur donne

$$\frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{1+\tau^2}} \cdot \exp\left(-\frac{1}{2(1+\tau^2)}(x-\mu)^2\right)$$

On reconnaît la loi de Laplace-Gauss $\mathcal{N}(\mu, 1+\tau^2)$: c'est la *loi prédictive* de X.
Il vient alors

$$\pi(\theta \mid x) = \frac{1}{\sqrt{2\pi}} \cdot \sqrt{\frac{1+\tau^2}{\tau^2}} \cdot \exp\left(-\frac{1+\tau^2}{2\tau^2}\left[\theta - \frac{\tau^2}{1+\tau^2}\left(x + \frac{\mu}{\tau^2}\right)\right]^2\right)$$

On reconnaît de nouveau une loi de Laplace-Gauss. La *loi a posteriori* est donc la loi

$$\mathcal{N}\left(\frac{\tau^2}{1+\tau^2}\left(x + \frac{\mu}{\tau^2}\right), \frac{\tau^2}{1+\tau^2}\right)$$

Il y a eu réactualisation des paramètres de la loi a priori :

- la moyenne a priori μ devient la moyenne a posteriori

$$\frac{\tau^2}{1+\tau^2}\left(x + \frac{\mu}{\tau^2}\right)$$

- la variance a priori τ^2 devient la variance a posteriori

$$\frac{\tau^2}{1+\tau^2}$$

- II - Estimation de la moyenne d'une loi de Laplace-Gauss. -

...Estimer ne coûte rien.
Estimer incorrectement coûte cher..
Vieux proverbe chinois

2.1 - Estimation.

Nous allons traiter un exemple d'analyse statistique bayésienne : le problème de l'estimation de la moyenne inconnue d'une loi de Laplace-Gauss. Reprenons pour cela le cadre de l'exemple précédent.

On observe une variable aléatoire réelle X de loi $\mathcal{N}(\theta, 1)$. La moyenne θ est inconnue et la variance est connue (ici égale à 1).

Le modèle statistique associé à une observation est donc

$$\left(\mathbf{R}, (\mathcal{N}(\theta, 1))_{\theta \in \mathbf{R}}\right)$$

La vraisemblance $l(\theta, \cdot)$ est égale à

$$\forall x \in \mathbf{R} \qquad l(\theta, x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\theta)^2\right)$$

Le problème est donc d'estimer le paramètre θ inconnu sur la base d'une observation. Un *estimateur* ϕ de θ est une application de $\mathbf{R}$ dans $\mathbf{R}$. Pour tout $x \in \mathbf{R}$, $\phi(x)$ est une *estimation* de θ.

L'estimateur usuel, noté $\hat{\phi}$, est l'estimateur des moindres carrés et du maximum de vraisemblance dans ce cas. Il est défini par

$$\forall x \in \mathbf{R} \qquad \hat{\phi}(x) = x$$

Intuitivement, il est naturel, sur la base d'une seule observation, d'estimer la moyenne inconnue θ par cette observation elle-même.

Cependant, on peut imaginer d'autres estimateurs possibles de θ. Il est donc indispensable d'utiliser un estimateur "le meilleur possible" et pour cela, on doit disposer de critères pour juger de la performance d'un estimateur et pour pouvoir comparer les estimateurs entre eux.

2.2 - Coût quadratique, risque quadratique.

Estimer θ par $\phi(x)$ entraîne une erreur inévitable que l'on peut mesurer par $(\phi(x) - \theta)^2$: le coût quadratique (ou erreur quadratique) encouru si on estime θ par $\phi(x)$.

Pour un estimateur ϕ de θ, on définit alors le *risque quadratique* de ϕ, noté R_ϕ par

$$\forall \theta \in \mathbf{R} \qquad R_\phi(\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (\phi(x) - \theta)^2 \exp\left(-\frac{1}{2}(x-\theta)^2\right) dx$$

c'est-à-dire le coût moyen encouru si on utilise l'estimateur ϕ. Il s'agit d'une fonction de $\mathbf{R}$ dans $\bar{\mathbf{R}}_+$.

Par exemple, pour l'estimateur usuel $\hat{\phi}$ on a

$$\forall \theta \in \mathbf{R} \qquad R_{\hat{\phi}}(\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (x-\theta)^2 \exp\left(-\frac{1}{2}(x-\theta)^2\right) dx = 1 \ .$$

Le risque de l'estimateur usuel est donc constant en θ.

2.3 - Comparaison des estimateurs.

On peut utiliser le critère du risque quadratique pour comparer les estimateurs entre eux.
Si ϕ et ψ sont deux estimateurs de θ, on dira que ϕ est meilleur (préférable) que ψ, on note $\phi \prec \psi$, si

$$\forall \theta \in \mathbf{R} \qquad R_\phi(\theta) \leq R_\psi(\theta)$$

On remarque immédiatement que deux estimateurs ne sont pas toujours comparables.

Si un estimateur n'est pas améliorable, on dit qu'il est *admissible*. Le problème de l'admissibilité est très compliqué (existence, unicité, calcul...) et demeure aujourd'hui un secteur important de la recherche en statistique.

Appliquons maintenant l'analyse statistique bayésienne pour proposer d'autres estimateurs dans le cas particulier que nous étudions.

2.4 - Loi a priori, loi a posteriori.

Comme dans l'exemple 1.6, nous choisissons comme loi a priori sur θ une loi de Laplace-Gauss $\mathcal{N}(\mu, \tau^2)$. Nous introduisons donc deux nouveaux paramètres

* μ : on pense a priori que θ est plutôt voisin de μ
* τ^2 : la variance, paramètre d'échelle mesurant la dispersion autour de μ, est ici un paramètre nuisible.

Comme dans l'exemple 1.6, on calcule la loi a posteriori $\Pi(\cdot \mid X = x)$

$$\Pi(\cdot \mid X = x) = \mathcal{N}\Big(\frac{\tau^2}{1+\tau^2}\big(x + \frac{\mu}{\tau^2}\big), \frac{\tau^2}{1+\tau^2}\Big)$$

Il y a réactualisation de notre information sur θ :

* la moyenne a posteriori est un barycentre de x et de μ
* la variance a posteriori vérifie la propriété suivante si on définit la précision comme l'inverse de la variance :

Précision a posteriori = Précision a priori + Précision de l'observation

2.5 - Risque bayésien, estimateur bayésien.

Soit ϕ un estimateur de θ. On peut calculer la moyenne de la fonction de risque R_ϕ de ϕ relativement à la loi a priori puisque θ est considéré comme aléatoire. On définit alors le *risque bayésien* de ϕ relativement à la loi a priori Π par

$$R_\Pi(\phi) = \frac{1}{\sqrt{2\pi}\tau} \int_{-\infty}^{+\infty} R_\phi(t) \cdot \exp\big(-\frac{1}{2\tau^2}(t-\mu)^2\big)\, dt$$

Par exemple, on a $R_\Pi(\hat{\phi}) = 1$.

Le risque bayésien permet de définir un nouveau critère de comparaison des estimateurs :

$$\phi \prec\prec \psi \qquad \Longleftrightarrow \qquad R_\Pi(\phi) \leq R_\Pi(\psi)$$

Deux remarques s'imposent.

1 - Comme R_Π est un nombre, deux estimateurs sont toujours comparables au sens du risque bayésien.

2 - Si $\phi \prec \psi$ alors $\phi \prec\prec \psi$. La réciproque est fausse.

Un estimateur ϕ^* est dit *bayésien* pour Π si, pour tout estimateur ϕ on a $\phi^* \prec\prec \phi$.

Plusieurs problèmes se posent alors : existence d'un estimateur bayésien, unicité, calcul.

2.6 - Coût moyen a posteriori.

Par interversion de l'ordre d'intégration, on montre que

$$\begin{aligned} R_\Pi(\phi) &= \int_{-\infty}^{+\infty} \Big(\int_{-\infty}^{+\infty} (\phi(x) - \theta)^2 l(\theta, x) dx\Big) \pi(\theta) d\theta \\ &= \int_{-\infty}^{+\infty} \Big(\int_{-\infty}^{+\infty} (\phi(x) - \theta)^2 \pi(\theta \mid x) d\theta\Big) \rho(x) dx \end{aligned}$$

L'intégrale $\int_{-\infty}^{+\infty} (\phi(x) - \theta)^2 \pi(\theta \mid x) d\theta$ représente le *coût moyen a posteriori*.

Le calcul précédent présente deux intérêts.

- Un intérêt pratique : si on minimise ce coût moyen a posteriori, alors on minimise le risque bayésien. Cela fournit donc une méthode de calcul de l'estimateur bayésien.
- Un intérêt sur le fond : on moyenne l'erreur quadratique sur toutes les valeurs possibles de θ ayant observé $X = x$. C'est plus naturel que de faire la moyenne sur toutes les valeurs possibles de X alors qu'on a observé une seule valeur de X.

2.7 - Existence et calcul de l'estimateur bayésien.

L'intégrale $\int_{-\infty}^{+\infty}(a-\theta)^2\pi(\theta \mid x)d\theta$ est minimisée pour

$$a = \int_{-\infty}^{+\infty} \theta\pi(\theta \mid x)d\theta$$

expression qui représente la moyenne de la loi a posteriori. Ainsi, *l'estimateur bayésien est la moyenne de la loi a posteriori.*

Dans notre exemple, l'estimateur bayésien ϕ_Π^* est donc défini par

$$\forall x \in \mathbf{R} \qquad \phi_\Pi^*(x) = \frac{\tau^2}{1+\tau^2}\left(x + \frac{\mu}{\tau^2}\right) = \mu + \left(1 - \frac{1}{1+\tau^2}\right)(x-\mu)$$

2.8 - Remarques.

On peut calculer le risque de l'estimateur bayésien ϕ_Π^* ainsi déterminé

$$\forall \theta \in \mathbf{R} \qquad R_{\phi_\Pi^*}(\theta) = \left(1 - \frac{1}{1+\tau^2}\right)^2 + \left(\frac{1}{1+\tau^2}\right)^2(\theta-\mu)^2$$

Quant au risque bayésien de ϕ_Π^*, il vérifie

$$R_\Pi(\phi_\Pi^*) = 1 - \frac{1}{1+\tau^2} < 1 = R_\Pi(\hat{\phi})$$

Pour conclure, on peut faire un certain nombre de remarques

1 - On a bien $\phi_\Pi^* \prec\prec \hat{\phi}$.

2 - Mais ϕ_Π^* et $\hat{\phi}$ ne sont pas comparables au sens du risque quadratique.

3 - ϕ_Π^* est meilleur que $\hat{\phi}$ (au sens du risque quadratique) au voisinage de μ.

4 - ϕ_Π^* est mauvais pour des grandes valeurs de θ puisque son risque quadratique converge vers l'infini lorsque θ tend vers l'infini.

5 - Si on fait tendre τ^2 vers l'infini alors ϕ_Π^* converge vers $\hat{\phi}$ et il y a aussi convergence du risque quadratique et du risque bayésien de ϕ_Π^* vers les risques respectifs de $\hat{\phi}$.

Le paramètre τ^2 apparaît bien comme un paramètre nuisible. On aimerait s'en débarrasser dans l'expression de l'estimateur bayésien. C'est une autre histoire qui sera abordée dans l'atelier correspondant.

Atelier :

ESTIMATION BAYÉSIENNE
- EFFET STEIN -

DOMINIQUE CELLIER (*)

- Le paradoxe de Stein(**) -

La meilleure estimation de la probabilité qu'un évènement se réalise est généralement identifiée à la moyenne arithmétique des résultats obtenus antérieurs. Le paradoxe de Stein définit les circonstances où existent des estimateurs meilleurs que cette moyenne.

Pour illustrer ce paradoxe, utilisons l'exemple développé par Efron et Morris dans leur article. On analyse le taux de réussite dans le renvoi de la balle avec la batte de 18 joueurs de baseball au cours de la saison 1970. Ces taux de réussite sont reproduits dans le tableau 1 page 2.

Pour le joueur numéro i, $1 \leq i \leq 18$, on note

θ_i : le taux de réussite du joueur pour l'année 1970,

y_i : le taux de réussite de ce même joueur à l'issu des 45 premiers essais de la même saison.

A l'issue des 45 premiers essais, on a donc observé $y = (y_1, y_2, \cdots, y_{18})$.

Si on nous avait demandé à ce moment précis d'estimer le taux de réussite $\theta = (\theta_1, \theta_2, \cdots, \theta_{18})$ sur l'ensemble de la saison, qu'aurions-nous prédit ? Probablement

$$\hat{\theta} = (y_1, y_2, \cdots, y_{18})$$

Car, traditionnellement, l'estimation fondée sur une observation est la valeur observée elle-même.

Le résultat de Stein est paradoxal en ce sens qu'il dément cette loi élémentaire de la théorie statistique : *si nous avons 3 joueurs ou plus, il existe une estimation "meilleure", c'est-à-dire avec plus de précision.*

Soit $\mu = \frac{1}{18} \sum_{i=1}^{18} y_i$ la moyenne des valeurs observées (ici $\mu = 0,265$). La phase essentielle de la méthode de Stein consiste à *rapprocher* chaque valeur observée y_i de μ de la façon suivante

$$\forall i \ , \ 1 \leq i \leq 18 \qquad\qquad \theta_i^* = \mu + c(y_i - \mu)$$

où c est une constante de rapprochement calculée à partir de l'observation (ici $c = 0,212$) .

On choisit alors l'estimateur $\theta^* = (\theta_1^*, \theta_2^*, \cdots, \theta_{18}^*)$ de θ défini par

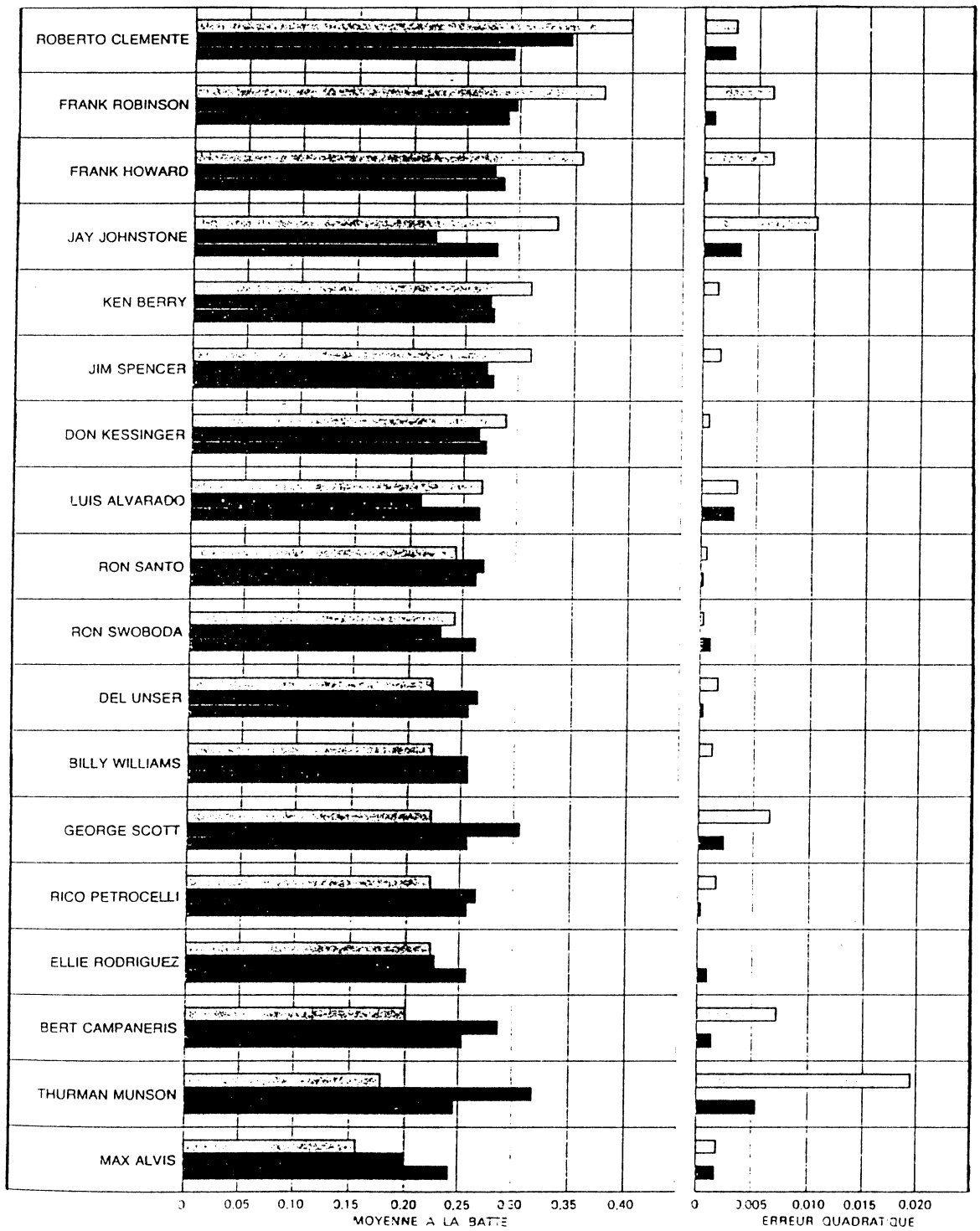
$$\forall i \ , \ 1 \leq i \leq 18 \qquad\qquad \theta_i^* = 0,265 + 0,212 \cdot (y_i - 0,265)$$

(voir tableau 2 page 3).

(*) Université de Rouen - Laboratoire Analyse et Modèles Stochastiques - U.R.A. C.N.R.S. 1378.

(**) D'après Bradley Efron et Carl Morris : Le paradoxe de Stein - Pour la Science, 1979, n° 1, p28-37.

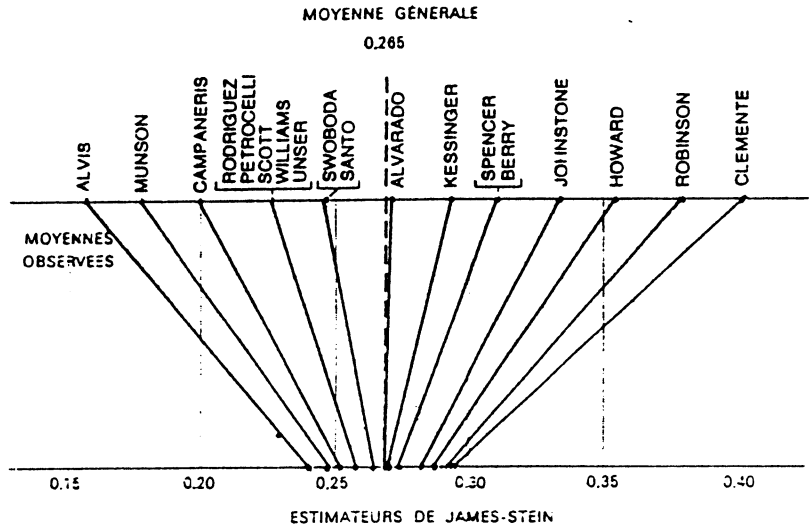
TABLEAU 1(*)



1. UNE ESTIMATION DES CAPACITÉS A LA BATTE de 18 joueurs de baseball est obtenue de façon plus précise par la méthode de James-Stein qu'en prenant les moyennes individuelles. Les moyennes employées comme estimateurs ont été calculées après que chaque joueur ait réalisé 45 essais pendant la saison 1970. La capacité réelle d'un joueur à la batte est une quantité inobservable, mais elle est approchée de près par la moyenne de ses performances sur une longue période. Ici, la capacité réelle est représentée par la moyenne à la batte obtenue pendant le reste de la saison de 1970. Pour l'estimation de la capacité à la batte de 16 joueurs sur 18, la moyenne arithmétique individuelle constitue une estimation moins bonne que l'estimation de James-Stein. L'ensemble des estimateurs de James-Stein a une erreur quadratique totale associée inférieure.

(*) Bradley Efron et Carl Morris : Le paradoxe de Stein - Pour la Science, 1979, n° 1, p29.

TABLEAU 2(*)



2. LES ESTIMATEURS DE JAMES STEIN pour les 18 joueurs de baseball ont été calculés en « rapprochant » les moyennes individuelles à la batte d'une « moyenne des moyennes individuelles » (moyenne globale). Dans ce cas, la moyenne globale vaut 0,265 et chacune des moyennes voit diminuer d'environ 80 % sa distance à cette valeur. Ainsi, le théorème sur lequel est basé la méthode de Stein affirme que les capacités réelles à la batte sont plus étroitement regroupées que les moyennes préliminaires n'auraient d'abord semblé le suggérer.

Nous disposons donc de deux estimateurs $\hat{\theta}$ et θ^* de θ. Lequel est le meilleur ?

Pour le joueur numéro i, les erreurs de prévision dans l'utilisation de ces estimateurs sont $(\hat{\theta}_i - \theta_i)$ et $(\theta_i^* - \theta_i)$.

On compare alors les deux estimateurs $\hat{\theta}$ et θ^* à l'aide de l'erreur quadratique globale

$$e(\hat{\theta}) = \sum_{i=1}^{18} (\hat{\theta}_i - \theta_i)^2 = 0,077$$

$$e(\theta^*) = \sum_{i=1}^{18} (\theta_i^* - \theta_i)^2 = 0,022$$

Ainsi l'*estimateur de Stein* θ^* est 3,5 fois plus précis au sens de l'erreur quadratique globale, il est par ailleurs meilleur pour 16 des 18 joueurs (voir tableau 1 page 2).

C'est un véritable défi au bon sens : pourquoi la réussite ou l'insuccès d'un joueur devrait influencer notre estimation d'un autre joueur ?

Plus choquant encore !

Supposons qu'on choisisse un échantillon de 45 voitures à Paris. Notons y_{19} la proportion de voitures étrangères dans l'échantillon et θ_{19} la proportion de voitures étrangères à Paris.

On peut injecter cette observation dans l'exemple précédent. L'observation est maintenant y', la moyenne globale devient μ' et la constante c c'.

On obtient un nouvel estimateur de $\theta' = (\theta_1, \theta_2, \cdots, \theta_{18}, \theta_{19})$

$$\theta' = \mu' + c'(y' - \mu')$$

qui est encore meilleur que l'estimateur usuel au sens de l'erreur quadratique globale.

Pour comprendre ce paradoxe appliquons l'analyse bayésienne dans le cas général suivant.

(*) Bradley Efron et Carl Morris : Le paradoxe de Stein - Pour la Science, 1979, n° 1, p30.

D.CELLIER : Estimation bayésienne - effet Stein

- Estimation bayésienne empirique -

On observe n variables aléatoires réelles indépendantes $(X_1, X_2, \cdots, X_n)$. On suppose que pour tout i, $1 \leq i \leq n$, la loi de X_i est une loi de Laplace-Gauss $\mathcal{N}(\theta_i, 1)$. Le paramètre $\theta = (\theta_1, \theta_2, \cdots, \theta_n) \in \mathbf{R}^n$ est inconnu et on désire l'estimer.

Un estimateur de θ est donc une application Φ de $\mathbf{R}^n$ dans $\mathbf{R}^n$

$$\forall (x_1, x_2, \cdots, x_n) \in \mathbf{R}^n \quad \Phi(x_1, x_2, \cdots, x_n) = \big(\phi_1(x_1, x_2, \cdots, x_n), \phi_2(x_1, x_2, \cdots, x_n), \cdots, \phi_n(x_1, x_2, \cdots, x_n)\big)$$

où $\phi_i(x_1, x_2, \cdots, x_n)$ estime θ_i.

Par exemple, l'estimateur usuel est $\hat{\Phi}$ défini par

$$\forall (x_1, x_2, \cdots, x_n) \in \mathbf{R}^n \qquad \Phi(x_1, x_2, \cdots, x_n) = (x_1, x_2, \cdots, x_n)$$

Les variables observées étant indépendantes, la vraisemblance $L(\theta, \cdot)$ du modèle statistique est définie par

$$\forall (x_1, x_2, \cdots, x_n) \in \mathbf{R}^n \quad L(\theta, (x_1, x_2, \cdots, x_n)) = \frac{1}{(2\pi)^{\frac{n}{2}}} \cdot \exp\Big(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta_i)^2\Big)$$

Si on utilise le critère du risque quadratique, le risque R_Φ de tout estimateur Φ de θ est défini par

$$\forall \theta \in \mathbf{R}^n \qquad R_\Phi(\theta) = \sum_{i=1}^{n} \int_{\mathbf{R}^n} \big(\phi_i(x_1, x_2, \cdots, x_n) - \theta_i\big)^2 L(\theta, (x_1, x_2, \cdots, x_n)) dx_1 dx_2 \cdots dx_n$$

On vérifie facilement que le risque de l'estimateur usuel $\hat{\Phi}$ est constant et égal à n.

Estimation bayésienne.

Choisissons comme loi a priori pour θ_i une loi de Laplace-Gauss $\mathcal{N}(\mu_i, \tau^2)$ et supposons que les θ_i sont indépendants. Un calcul simple montre que la loi a posteriori de θ_i est une loi de Laplace-Gauss

$$\Pi(\cdot \mid (x_1, x_2, \cdots, x_n)) = \mathcal{N}\Big(\frac{\tau^2}{1+\tau^2}(x_i + \frac{\mu_i}{\tau^2}), \frac{\tau^2}{1+\tau^2}\Big)$$

L'estimateur bayésien $\Phi^* = (\phi_1^*, \phi_2^*, \cdots, \phi_n^*)$ de θ est la moyenne de la loi a posteriori et est donc défini par

$$\forall i \in \{1, 2, \cdots, n\} \quad \forall (x_1, x_2, \cdots, x_n) \in \mathbf{R}^n \quad \phi_i^*(x_1 . x_2, \cdots, x_n) = \frac{\tau^2}{1+\tau^2} \cdot (x_i + \frac{\mu_i}{\tau^2}) = \mu_i + \big(1 - \frac{1}{1+\tau^2}\big)(x_i - \mu_i)$$

On montre que le risque quadratique de cet estimateur est défini par

$$\forall \theta \in \mathbf{R}^n \qquad R_{\Phi^*}(\theta) = n\big(1 - \frac{1}{1+\tau^2}\big)^2 + \frac{1}{(1+\tau^2)^2}\sum_{i=1}^{n}(\theta_i - \mu_i)^2$$

Nous pouvons faire alors les mêmes remarques que dans l'exposé précédent concernant la comparaison de $\hat{\Phi}$ et Φ^*

1 - Φ^* est meilleur que $\hat{\Phi}$ au sens du risque bayésien ($\Phi^* \prec\prec \hat{\Phi}$).
2 - Ces deux estimateurs ne sont pas comparables au sens du risque quadratique.
3 - Φ^* est meilleur que $\hat{\Phi}$ (au sens du risque quadratique) au voisinage de $\mu = (\mu_1, \mu_2, \cdots, \mu_n)$.
4 - Φ^* est mauvais pour des grandes valeurs de $\|\theta\|$ puisque son risque quadratique converge vers l'infini lorsque $\|\theta\|$ tend vers l'infini.
5 - Si on fait tendre τ^2 vers l'infini, alors Φ^* converge vers $\hat{\Phi}$ et il y a aussi convergence du risque quadratique et du risque bayésien de Φ^* vers ceux de $\hat{\Phi}$.

Le paramètre τ^2 apparaît encore comme un paramètre nuisible dont on aimerait se débarrasser. Mais comment ?

Estimation bayésienne empirique.

Si on utilise la loi prédictive de l'observation, on montre que relativement à cette loi, la variable aléatoire

$$\frac{n-2}{\sum_{i=1}^{n}(x_i-\mu_i)^2}$$

est de moyenne $1/(1+\tau^2)$, c'est-à-dire que cette variable aléatoire est un *estimateur sans biais* de la quantité $1/(1+\tau^2)$.

On peut donc remplacer cette dernière dans l'expression de l'estimateur bayésien Φ^* par une estimation sans biais. On obtient ainsi un *estimateur bayésien empirique* Φ^s défini par

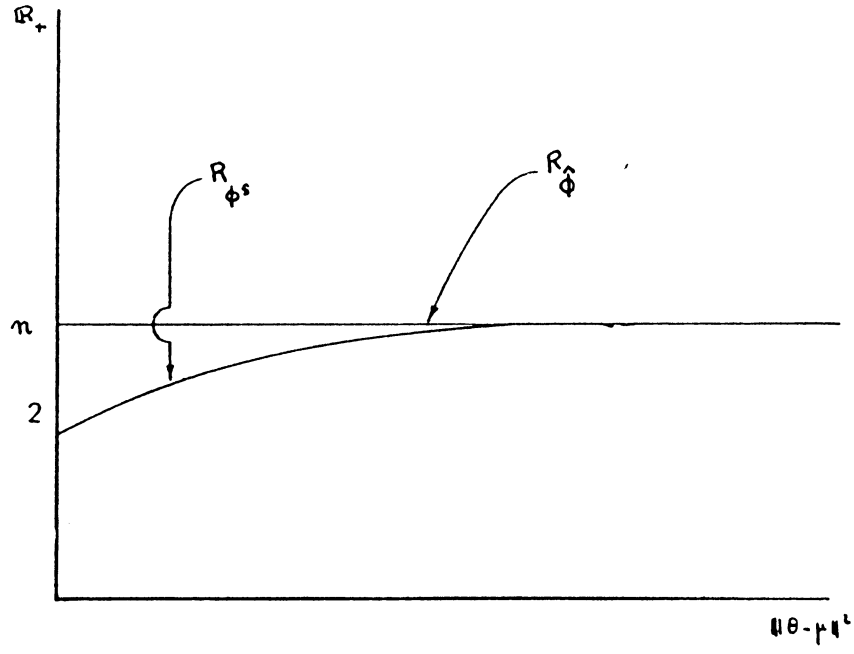
$$\forall i \in \{1,2,\cdots,n\} \quad \forall (x_1,x_2,\cdots,x_n) \in \mathbb{R}^n \quad \phi_i^s(x_1,x_2,\cdots,x_n) = \mu_i + \left(1 - \frac{n-2}{\sum_{i=1}^{n}(x_i-\mu_i)^2}\right)(x_i-\mu_i)$$

Cet estimateur porte le nom d'*estimateur de Stein.*

On reconnaît évidemment la forme de l'estimateur construit dans l'exemple introductif d'Efron et Morris dans lequel

$$\mu = \frac{1}{18}\sum_{j=1}^{18} y_i = 0,265 \qquad \text{et} \qquad c = 1 - \frac{16}{\sum_{j=1}^{18}(y_j-\mu)^2} = 0,212$$

Pour terminer, un calcul plus compliqué permet de vérifier que cet estimateur de Stein a un risque quadratique strictement inférieur à celui de l'estimateur usuel dès que $n \geq 3$.



Nous renvoyons aux deux tableaux 3 et 4 de simulations donnés en annexe pour analyer et juger des performances réciproques des différents types d'estimateurs introduits et de l'intérêt des estimateurs de Stein.

TABLEAU 3

Lois simulées	N(0,1)	N(5,1)	N(10,1)	N(15,1)	N(20,1)
	0.11	5.74	9.22	14.93	18.82
	-0.58	5.72	9.97	14.71	18.63
	1.00	6.98	10.26	14.21	19.11
	-1.62	5.08	9.41	15.41	20.56
	-0.92	4.70	10.68	14.33	22.54
	-0.07	4.13	9.67	15.03	18.91
	0.63	5.48	7.74	15.25	20.86
	0.43	4.59	8.96	14.60	20.16
	-0.08	4.97	10.98	15.31	20.75
	-1.61	4.83	10.16	16.41	20.02
	0.36	4.88	11.30	14.14	21.44
	-0.05	4.41	9.84	14.55	18.63
	-0.53	6.15	8.84	15.55	20.58
	-1.47	4.04	11.52	14.33	20.36
	0.04	5.40	12.08	15.44	20.61
	1.32	6.36	10.02	15.00	19.75
	-0.56	6.20	10.29	14.58	20.81
	0.01	5.60	10.55	13.34	20.25
	0.73	3.79	10.98	13.49	19.88
	-0.28	3.31	11.59	14.66	21.09
Estimateur usuel	-0.16	5.12	10.20	14.76	20.19
Erreur quadratique	0.025	0.014	0.041	0.056	0.035
Loi a priori N(0,1)					
Estimateur bayésien	-0.08	2.56	5.10	7.38	10.09
Erreur quadratique	0.006	5.959	23.994	58.039	98.126
Loi a priori N(0,10)					
Estimateur bayésien	-0.16	5.07	10.10	14.62	19.99
Erreur quadratique	0.024	0.004	0.010	0.147	0.000
Loi a priori N(0,5)					
Estimateur Bayésien	-0.15	4.92	9.81	14.20	19.41
Erreur quadratique	0.023	0.006	0.036	0.647	0.346

D.CELLIER : Estimation bayésienne - effet Stein

TABLEAU 4

Lois simulées	N(0,1)	N(10,1)	N(20,1)	N(30,1)	N(40,1)	N(50,1)	N(60,1)	N(70,1)	N(80,1)	N(90,1)
Valeur du paramètre	0	10	20	30	40	50	60	70	80	90
Valeurs observées	-0.26	10.51	19.21	30.25	39.00	50.86	59.01	72.09	81.15	89.54
ESTIMATEUR USUEL	-0.26	10.51	19.21	30.25	39.00	50.86	59.01	72.09	81.15	89.54
Erreurs quadratiques	0.07	0.26	0.63	0.06	1.00	0.73	0.98	4.35	1.33	0.21
Erreur quadratique globale	9.62									
ESTIMATEUR DE STEIN										
Moyenne des observations	45.13									
(Observ. - Moyen.)**2	2060.25	1199.01	672.33	221.68	37.63	32.73	192.48	726.35	1297.30	1971.87
Rétrécissement	0.00095									
Estimateur de Stein	-0.21	10.54	19.23	30.26	39.01	50.85	58.99	72.06	81.12	89.50
Erreurs quadratiques	0.05	0.29	0.59	0.07	0.99	0.72	1.01	4.24	1.25	0.25
Erreur quadratique globale	9.46									
Performance	+	-	+	-	+	+	-	+	+	-
ESTIMATEUR DE BAYES										
Loi a priori N(45.13;1)										
Estimateur bayésien	22.44	27.82	32.17	37.69	42.07	47.99	52.07	58.61	63.14	67.34
Erreur quadratique	503.53	317.59	148.11	59.14	4.27	4.02	62.86	129.73	284.14	513.59
Erreur quadratique globale	2027.00									
Performance	-	-	-	-	-	-	-	-	-	-
ESTIMATEUR DE BAYES										
Loi a priori N(45.13;10)										
Estimateur bayésien	0.19	10.85	19.46	30.39	39.06	50.80	58.87	71.82	80.80	89.10
Erreur quadratique	0.04	0.72	0.29	0.15	0.88	0.64	1.28	3.31	0.63	0.81
Erreur quadratique globale	8.75									
Performance	+	-	+	-	+	+	-	+	+	-

REGRESSION LINEAIRE MULTIPLE

Méthodes, applications, programmes

Thierry Foucart
Département de Mathématiques
Université d'Orléans

INTRODUCTION

La régression linéaire est la méthode statistique vraisemblablement la plus utilisée par les praticiens de toutes disciplines: la recherche d'une liaison entre deux ou plusieurs caractères est une démarche très courante en médecine, en psychologie, en physique, en économie etc...

Cette recherche correspond en premier lieu à la vérification d'un modèle construit pour représenter une certaine réalité: c'est le cas des sciences exactes, physique, chimie par exemple, où le modèle doit représenter très fidèlement les phénomènes étudiés. C'est aussi le cas des sciences humaines, économie, sociologie, psychologie ..., où l'on se contente par contre d'une approximation jugée suffisante compte tenu du contexte.

Inversement, les liens existant entre différents paramètres peuvent être inconnus: il s'agit alors de déterminer les liaisons qui existent et d'en déduire un modèle approprié. Cette démarche, plus difficile que la précédente, est souvent employée sans suffisamment de précautions: une liaison mise en évidence par la statistique est en fait une question posée au praticien: pourquoi cette liaison existe-t-elle? Il faut apporter une explication déterministe à un phénomène détecté par des méthodes basées sur l'analyse du hasard.

Ces deux approches de la régression multilinéaire demandent une bonne connaissance théorique de la méthode, un logiciel commode et suffisamment puissant et une grande expérience dans le traitement des données statistiques.

La plupart des ouvrages généraux de statistique disponibles actuellement contiennent un chapitre au moins consacré à la régression: on ne rencontre donc guère de difficultés à se mettre au courant de la méthode au plan théorique. Nous donnons à la fin de cet article une courte bibliographie commentée.

On pourra se procurer pour un prix dérisoire le logiciel de régression multilinéaire LORELI (LOgiciel de REgression LInéaire), qui permet d'appliquer la plupart des méthodes classiques de régression linéaire, auprès de l'IREM d'Orléans.

Nous donnons dans le texte qui suit des applications détaillées des méthodes à des données réelles publiées dans l'ouvrage de Saporta et figurant en annexe. Tous les résultats numériques ont été établis à l'aide du logiciel LORELI.

Dans le texte qui suit, on ne discute que des méthodes de base de la régression linéaire; les lecteurs intéressés par les méthodes plus complexes telles que la régression sur composantes principales et la ridge régression, pourront consulter les ouvrages donnés en bibliographie en particulier ceux de Tomassone (en français) et de Weisberg (en anglais).

Chapitre 1

REGRESSION ET MODELE LINEAIRES

1. REGRESSION ET MODELE LINEAIRES. DEFINITIONS.

La régression et le modèle linéaires sont deux méthodes statistiques souvent confondues parce que les procédures de calcul sont les mêmes.

La régression concerne un couple de variables aléatoires (X,Y). On considère un échantillon $(X_i,Y_i)i=1,...,n$ de ce couple: les v.a. X_i et Y_i vérifient donc les propriétés suivantes:

— X_i et X_j sont indépendantes, de même que Y_i et Y_j, X_i et Y_j $(i \neq j)$.

— (X_i, Y_i) est un couple aléatoire de même loi que (X,Y).

Supposons que X et Y soient liées par la relation:

$$Y = f(X) + \varepsilon$$

où les v.a. ε et X sont indépendantes. On en déduit que cette relation est vérifiée pour tous les couples (X_i, Y_i):

$$\forall i=1,...,n \quad Y_i = f(X_i) + \varepsilon_i$$

Dans la formule précédente, les suites (Y_i), (X_i), (ε_i) constituent des échantillons des v.a. Y, X, et ε.

4 *Régression et modèle linéaires* *Chap. 1*

Un modèle s'applique lorsque la variable X n'est plus aléatoire mais contrôlée par l'utilisateur, ce qui est souvent le cas en physique par exemple; il consiste à poser:

$$\forall i=1,\ldots,n \quad Y_i = f(x_i) + \varepsilon_i$$

Dans ce modèle, les v.a. Y_i ne constituent pas un échantillon d'une v.a. puisqu'elles n'ont pas la même loi. L'hypothèse fondamentale du modèle linéaire est que, par contre, les v.a. ε_i constituent un échantillon d'une v.a. ε: elles sont indépendantes et de même loi.

On raisonne toujours en régression conditionnellement aux valeurs observées x_i, ce que l'on note X = x: il n'y a plus aucune différence formelle avec le modèle, et cela explique pourquoi on utilise parfois la terminologie de la régression en étudiant un modèle.

On cherche donc à ajuster une fonction f représentant la liaison entre les deux variables et à donner une estimation de ses paramètres. Lorsque la fonction f peut s'écrire sous la forme d'une fonction linéaire de ses paramètres à estimer, on parle de régression ou de modèle linéaire.

Les v. a. ε_i, qui constituent un échantillon d'une v.a. ε, suivent donc toutes la même loi de probabilité, que l'on suppose être dans la plupart des cas la loi normale centrée de variance σ^2. Il existe des modèles plus généraux, dans lesquels les ε_i ne sont pas indépendants ou n'ont pas la même variance, mais nous nous limiterons au cas le plus simple qui est d'ailleurs le plus employé.

L'hypothèse de normalité de la v.a. ε se justifie par le fait que ce terme d'erreur représente dans le modèle toute l'information non prise en compte; en physique par exemple, toutes les erreurs de mesure. Le cumul de ces erreurs justifie alors le choix de la loi normale.

Définitions:

— la variable Y est appelée variable expliquée (ou dépendante).

— la variable X èst appelée variable explicative (ou indépendante).

— la variable ε est appelée variable résiduelle.

— la variance σ^2 de la v.a. ε est appelée variance résiduelle.

Les représentations graphiques des couples (x_i, y_i) donnent des indications sur la nature de la fonction f à calculer; lorsqu'elle n'est pas linéaire, on peut procéder à une transformation des variables, par le logarithme par exemple; on peut penser à une fonction exponentielle de la fréquence cardiaque pour ajuster la résistance pulmonaire (fig. 1.1).

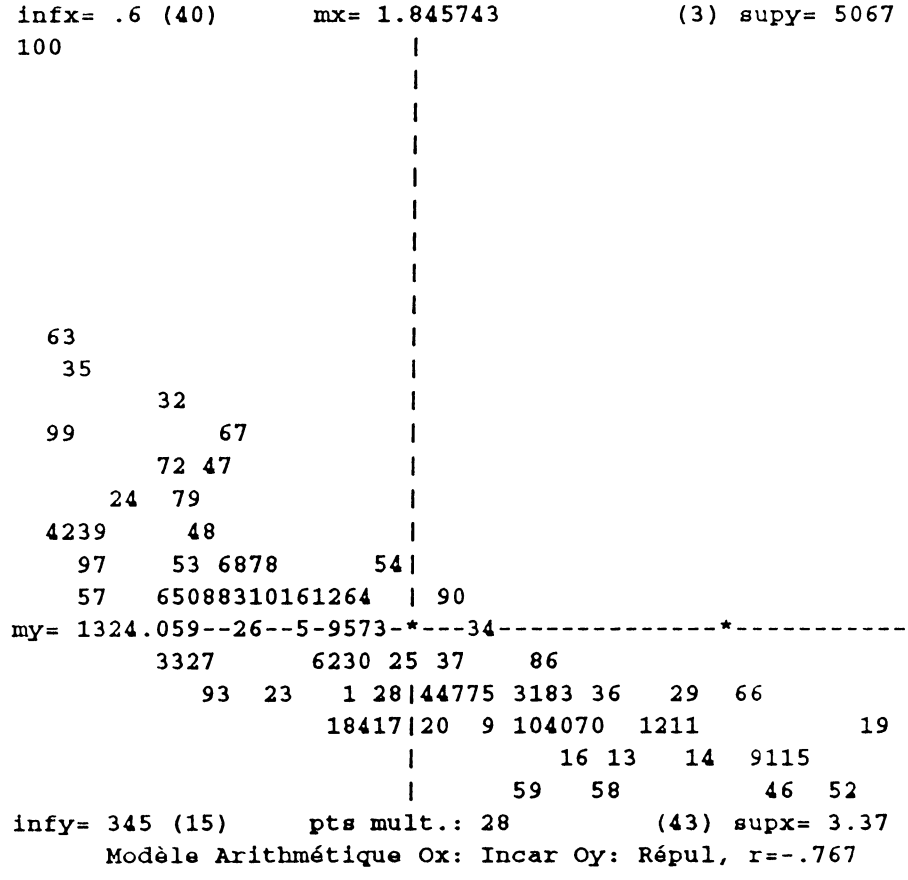


Fig.1.1: Index cardiaque x résistance pulmonaire
(repère arithmétique)

2. AJUSTEMENT D'UNE DROITE.

La courbe la plus simple à ajuster au nuage de points est la droite. Pour ajuster une fonction exponentielle, une simple transformation des données suffit pour se ramener au cas linéaire, alors que la régression polynomiale sera introduite comme une régression multilinéaire (chapitre 2).

Le modèle prend alors la forme, sur les valeurs observées:

$$\forall i=1,\ldots,n \quad y_i = a\,x_i + b + e_i$$

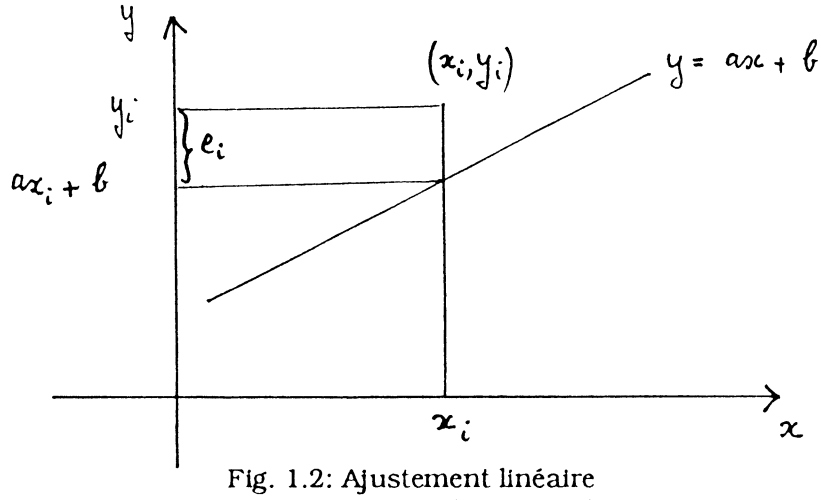


Fig. 1.2: Ajustement linéaire
(critère des moindres carrés)

La variable expliquée étant toujours placée en ordonnée, l'erreur que l'on commet en remplaçant y_i par $a\,x_i + b$ est $e_i = y_i - (a\,x_i + b)$.

On cherche à minimiser ces erreurs: le critère des moindres carrés consiste à minimiser la fonction S(a,b):

$$S(a,b) = \sum_{i=1}^{n} [y_i - (a\,x_i + b)]^2$$

Les résultats sont connus et figurent dans tous les livres; mais nous présentons la démonstration de façon à introduire le calcul matriciel que l'on

utilise en régression multiple.

On sait que le minimum de la fonction S(a,b), s'il existe, est obtenu pour les valeurs a et b qui annulent les dérivées partielles premières. Nous admettrons la condition suffisante, qui fait intervenir des dérivées partielles secondes, et calculons a et b tels que:

$$\partial S/\partial a = 0 \qquad \partial S/\partial b = 0$$

On a:

$$\partial S/\partial a = -2 \sum_{i=1}^{n} [(y_i - (a\, x_i + b)]\, x_i = 0$$

$$\partial S/\partial b = -2 \sum_{i=1}^{n} [(y_i - (a\, x_i + b)] = 0$$

On trouve le système des "équations normales":

$$\sum_{i=1}^{n} x_i^2\, a + \sum_{i=1}^{n} x_i\, b = \sum_{i=1}^{n} x_i\, y_i$$

$$\sum_{i=1}^{n} x_i\, a + n\, b = \sum_{i=1}^{n} y_i$$

Ce système se met sous la forme matricielle suivante:

$$M \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} x_i\, y_i \\ \sum_{i=1}^{n} y_i \end{bmatrix} \quad \text{avec} \quad M = \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{bmatrix}$$

La résolution mathématique de ce système ne pose pas de problème. On trouve:

$$\boxed{a = \mathrm{cov}(x,y)/s^2(x) \qquad b = \bar{y} - a\,\bar{x}}$$

Ce qui nous intéresse ici, c'est la résolution matricielle, qui consiste à

inverser la matrice M (si elle est inversible). On obtient:

$$\begin{bmatrix} a \\ b \end{bmatrix} = M^{-1} \begin{bmatrix} \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i \end{bmatrix}$$

Cette résolution matricielle du système des équations normales sera utilisée pour calculer les coefficients de régression b_j, j=0,...,p du modèle de régression multilinéaire (Cf. chap. 2).

3. ETUDE DES ESTIMATEURS ET DES RESIDUS.

Le modèle linéaire s'exprime de la façon suivante:

$$\forall i=1,\ldots,n \quad Y_i = \alpha x_i + \beta + \varepsilon_i$$

Définition: les coefficients α et β sont appelés coefficients de régression.

Nous avons calculé dans le paragraphe précédent les coefficients a et b de la droite la plus proche des points observés au sens des moindres carrés. Les valeurs que l'on obtient sont en fait des estimations des vrais coefficients, puisqu'ils dépendent de l'échantillon; nous les appellerons aussi coefficients de régression, en précisant qu'ils sont estimés en cas d'ambiguïté.

Théorème: les estimateurs A et B des coefficients α et β sont des estimateurs linéaires efficaces (sans biais et de variance minimale) indépendants de la v.a. ε. Si la variable résiduelle est gaussienne, ils sont gaussiens.

Nous ne démontrerons pas ce théorème ni les formules des variances des estimateurs A et B conditionnellement à x que nous donnons ci-dessous:

$$V(A) = \frac{\sigma^2}{n\, s^2(x)} \qquad V(B) = \frac{\sigma^2}{n} + \frac{\sigma^2\, \bar{x}^2}{n\, s^2(x)} \qquad cov(A,B) = -\frac{\sigma^2\, \bar{x}}{n\, s^2(x)}$$

Ces formules montrent que les estimateurs sont convergents (leurs variances tendent vers 0 lorsque n tend vers l'infini) et que la variance de B est une fonction croissante de $\bar{x}$. On peut montrer que le coefficient de corrélation entre A et B tend vers -1 lorsque $\bar{x}$ tend vers l'infini.

La réalisation de la v.a. ε_i est, d'après le modèle, égale à $y_i - (\alpha x_i + \beta)$. Ne connaissant pas les vraies valeurs des coefficients de régression α et β, on ne peut la calculer: la variable résiduelle n'est pas observable. On peut toutefois calculer les valeurs des erreurs commises après estimation des coefficients de régression:

$$e_i = y_i - (a\, x_i + b)$$

Définition: les termes $e_i = y_i - (a\, x_i + b)$ sont appelés résidus.

Ces résidus, qui ne constituent pas comme nous le verrons ultérieurement un échantillon d'une v.a., possèdent un certain nombre de propriétés que nous retrouverons en régression multiple et que nous admettrons:

— la série des résidus est de moyenne nulle:

$$1/n \sum_{i=1}^{n} e_i = 0$$

— elle est non corrélée avec la variable explicative:

$$1/n \sum_{i=1}^{n} e_i\, x_i = 0$$

— sa variance est égale à:

$$s^2(e) = 1/n \sum_{i=1}^{n} e_i^2 = s^2(y)\ (1-r^2)$$

où r est le coefficient de corrélation empirique:

$$r = \frac{cov\ (x,y)}{s(x)\ s(y)}$$

On trouve ici une autre interprétation du coefficient de corrélation: une valeur proche de ±1 montre que la variance de la série des résidus est faible par rapport à la variance observée de la variable expliquée. Ces propriétés ont pour conséquence:

Théorème: la quantité $s^2 = n\ s^2(e) / (n-2)$ est une estimation sans biais de la variance résiduelle σ^2.

L'estimation de la variance résiduelle est utilisée pour effectuer des prédictions de la variable expliquée en fonction de la variable explicative.

Le modèle étant donné par:

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

où ε suit la loi normale centrée, il est facile de montrer que l'espérance E(Y/X=x) de Y pour X=x, est égale à $\alpha\ x + \beta$.

La première prédiction que l'on effectue est donc celle de l'espérance conditionnelle de Y sachant X=x. On peut aussi vouloir prédire une valeur particulière: la formule est la même, seule la variance de la prévision diffère.

Théorème: L'estimateur de E(Y/X=x) est Y':

$$Y' = A\ x + B$$

dont la variance est égale à:

$$V(Y') = \frac{\sigma^2}{n} \left[1 + \frac{(x - \bar{x})^2}{s^2(x)} \right]$$

La variance de la prévision d'une valeur est obtenue en ajoutant le terme σ^2 à la variance précédente; pour estimer ces variances, il suffit de remplacer la variance résiduelle σ^2 par son estimateur sans biais s^2.

On s'efforce, dans la pratique, d'aboutir à des résidus gaussiens. En effet, dès que cette hypothèse est acceptée, on peut effectuer des tests et des estimations par intervalle de confiance sur les coefficients de régression, sur le coefficient de détermination et sur les prévisions (on trouvera des exemples numériques dans le paragraphe 4 de ce chapitre):

— Pour donner une estimation du coefficient de régression α par intervalle de confiance et tester l'égalité à une valeur spécifiée, on utilise la statistique $T = (A - a)/ S_A$, où S_A^2 est l'estimateur sans biais de la variance de A, qui suit la loi de Student de degré de liberté n-2.

— Le test de Fisher Snedecor peut être appliqué pour tester la nullité du coefficient de corrélation théorique: sous cette hypothèse, la statistique $F = (n-2)R^2/(1-R^2)$ suit la loi de Snedecor de degrés de liberté 1 et n-2.

— On peut donner des intervalles de confiance aux prévisions de la variable expliquée, en utilisant la statistique $(Y'-y)/S_{Y'}$ qui suit la loi de Student de degré de liberté n-2 si y est l'espérance de Y' et $S_{Y'}^2$ l'estimateur sans biais de sa variance.

Remarque: cas des données groupées.

Nous avons présenté dans les paragraphes précédents l'ajustement d'un nuage de points par une droite dans le cas de données individuelles. Tous les résultats que nous avons donnés peuvent être appliqués au cas de données groupées sous la forme d'un tableau de corrélation: il suffit d'introduire dans les formules un terme correspondant à l'effectif $n_{k,l}$ des valeurs observées dans la classe $C_k \times D_l$ et identifiées au couple (c_k, d_l) constitué des centres.

4. EXEMPLE NUMERIQUE.

Nous effectuons dans ce paragraphe la régression de la résistance pulmonaire (Répul) par l'index cardiaque (Incar) sur l'échantillon constitué des 101 malades observés. Nous n'appliquons ici que les définitions et résultats présentés précédemment.

Nous avons observé dans le paragraphe 1.1 que la liaison entre l'index cardiaque et la résistance pulmonaire présente un caractère exponentiel qui disparait si l'on considère le logarithme de la résistance pulmonaire au lieu de la variable initiale.

En outre, cette transformation atténue considérablement la particularité de l'unité statistique n° 100, que nous conserverons donc dans les données.

Notre première opération consiste donc à définir le modèle:

$$Y_i = \alpha\, x_i + \beta + \varepsilon$$

où Y est le logarithme népérien de la résistance pulmonaire. Pour simplifier, nous continuerons à la noter "Répul".

Les paramètres statistiques observés sur les données sont les suivants:

Variables	Moyennes	écarts-types	Variances
Incar	1.8457	.655747	.4300047
Répul	7.0499	.529936	.2808321

Le coefficient de corrélation entre la variable expliquée et la variable explicative est égal à -0.839, plus élevé en valeur absolue que lorsque l'on considère comme variable expliquée la résistance pulmonaire proprement dite (-0.767), ce qui justifie a posteriori le choix du logarithme.

L'analyse de variance nous donne la variance résiduelle estimée et le carré du coefficient de corrélation:

$$s^2 = 0.0846 \qquad r^2 = 0.7044 \qquad F(1\ ,99) = 235.942$$

La régression donne de bons résultats: le coefficient de corrélation est significativement non nul puisque la valeur observée de F appartient à la région critique [6.90, + ∝ [définie pour un risque de première espèce α égal à 0.01, et la variance résiduelle estimée est très inférieure à la variance de la variable expliquée.

Les coefficients de régression sont:

estimation	écart-type	t de Student	Interv de conf.
a = -0.67827	0.04416	-15.360	[-0.76482,-0.59172]
b = 8.30185	0.08649	95.982	[8.13233, 8.47137]

Etudions les résidus.

On peut rechercher les points aberrants parmi les données en examinant ce que l'on appelle la représentation linéaire des résidus (fig. 1.3).

Cette représentation montre que, sur les 101 résidus observés, seuls 4 sortent de l'intervalle ± 2 fois l'écart-type représenté par les deux demi-droites supérieure et inférieure: il s'agit des résidus e_{18} = -0.6581, e_{59} = -0.8847, e_{71} = -0.7044 et e_{100}= 0.6356.

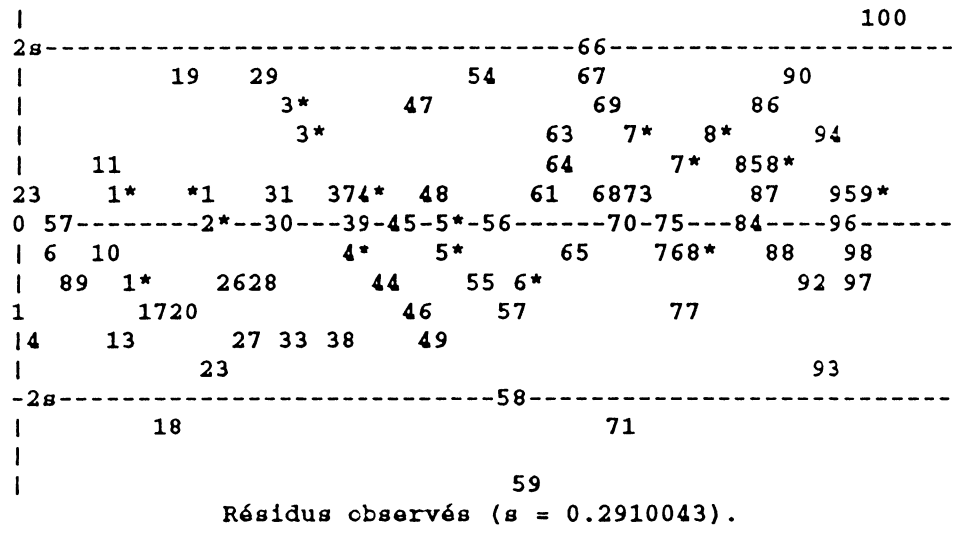


Fig. 1.3: Représentation linéaire des résidus
(Régression de Répul par Incar)

Dans le cas de la loi normale, il y a en moyenne 5% de valeurs à l'extérieur de l'intervalle ± 2 fois l'écart-type; la série des résidus observés ne présente donc rien d'extraordinaire. Il est toutefois intéressant d'examiner les unités statistiques correspondantes. On peut noter une grande différence entre les valeurs observées de certaines variables sur les unités statistiques en question et les moyennes calculées sur la totalité des observations (pour juger de ces différences, on la compare à l'écart-type; c'est pourquoi nous avons fait figurer simultanément la moyenne et l'écart-type des variables en-dessous des valeurs observées).

*	frcar	incar	insys	prdia	papul	pvent	répul	prono
18*	86	1.7	19.8	10	14	10.5	659	2
59*	75	2.32	30.9	8	10	6	345	2
71*	100	2.31	23.1	8	12	1	416	2
100*	116	0.60	5.2	33	38	10	5067	1
moy.	91.9	1.86	20.97	19.1	25.9	9.5	1286.6	
e-t	16.2	0.65	8.67	5.64	7.22	4.34	638.8	

C'est particulièrement vraie pour la pression artérielle pulmonaire (papul) et surtout la pression diastolique (prdia), pour laquelle les valeurs observées sortent de l'intervalle moyenne ± 2 x écart-type suivant le signe des résidus. On peut donc imaginer que la variable explicative Incar devrait être complétée par l'une de ces deux variables pour mieux reconstruire la résistance pulmonaire.

L'histogramme est donné en figure 1.4; les résidus paraissent un peu dissymétriques mais le test d'ajustement du χ^2 permet d'accepter l'hypothèse de normalité (l'écart-type étant le seul paramètre estimé de la loi normale ajustée, le degré de liberté est égal à 5):

Classes	Effectifs	Probabilité	condition (np_i)
1	12	0.0948	9.57
2	8	0.1393	14.07
3	24	0.2106	21.27
4	26	0.2280	23.02
5	14	0.1767	17.85
6	12	0.0980	9.90
7	5	0.0526	5.32

Résidus observés Test du Chi-2: 5.2587 Ddl: 5 Prob. cr.: 0.3851

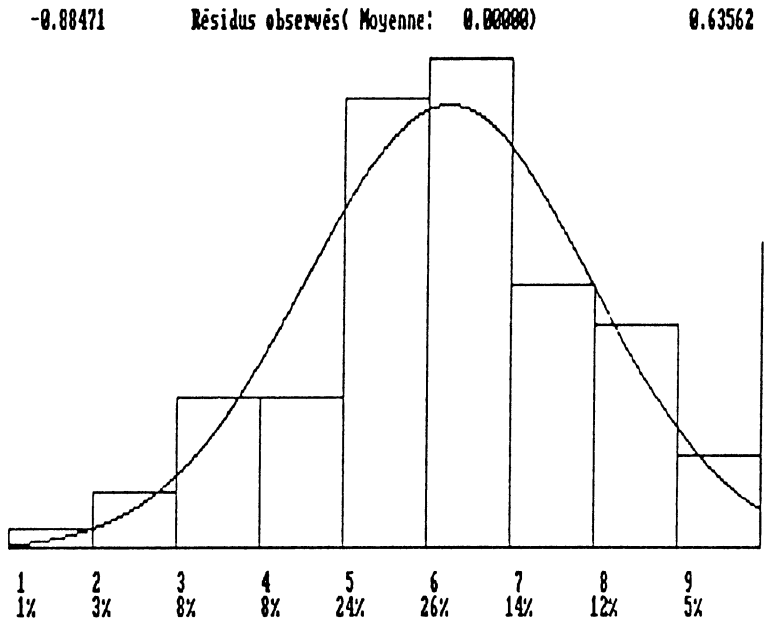


Fig. 1.4: Histogramme des résidus observés

On peut donc accepter l'hypothèse de normalité des résidus, et donc de la variable résiduelle (le nombre d'observations est suffisamment élevé pour que cette approximation soit justifiée). Les tests de Student et de Fisher peuvent donc être utilisés et montrent évidemment que l'on ne peut pas accepter l'hypothèse d'un coefficient de régression α nul ou d'un coefficient de corrélation ρ nul. Les intervalles de confiance sur α et β (p. 13) sont calculés pour un risque de première espèce égal à 0.05.

Chapitre 2

INTRODUCTION
A LA REGRESSION MULTILINEAIRE

1. ESTIMATION DES COEFFICIENTS DE REGRESSION.

1.1 Modèle multilinéaire.

Nous avons étudié dans le chapitre précédent le modèle linéaire simple: une variable expliquée, notée Y, et une variable explicative notée X. Le modèle multilinéaire consiste à introduire plusieurs variables explicatives que nous noterons X_1, X_2, ..., X_p. La distinction entre modèle linéaire et régression linéaire que nous avons précisée reste valable dans le cas multilinéaire: en régression, les variables X_j, j=1,...,p sont des variables aléatoires, dans le modèle multilinéaire, ce sont des variables contrôlées.

Le modèle de régression s'écrit donc:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_p X_p + \varepsilon$$

où ε est la variable résiduelle, indépendante des v.a. explicatives X_1, X_2, ..., X_p d'espérance nulle et de variance σ^2. Nous disposons d'un échantillon de (X_1, X_2,..., X_p, Y) de taille n.

En raisonnant conditionnellement aux valeurs observées x_i, on peut considérer l'espérance conditionnelle de Y:

$$E(Y / X_j=x_j \ \forall j=1,p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Le modèle multilinéaire est un système de n équations dont chacune correspond à une suite de valeurs des variables contrôlées $X_1,\ldots, X_p$:

$$Y(1) = \beta_0 + \beta_1 x_1(1) + \beta_2 x_2(1) +\ldots + \beta_p x_p(1) + \varepsilon(1)$$
$$Y(2) = \beta_0 + \beta_1 x_1(2) + \beta_2 x_2(2) +\ldots + \beta_p x_p(2) + \varepsilon(2)$$
$$\ldots\ldots\ldots$$
$$Y(i) = \beta_0 + \beta_1 x_1(i) + \beta_2 x_2(i) +\ldots + \beta_p x_p(i) + \varepsilon(i)$$
$$\ldots\ldots\ldots$$
$$Y(n) = \beta_0 + \beta_1 x_1(n) + \beta_2 x_2(n) +\ldots + \beta_p x_p(n) + \varepsilon(n)$$

Dans ce modèle nous faisons l'hypothèse que les v.a. $\varepsilon(i)$ constituent un échantillon indépendant d'une v.a. ε d'espérance nulle et de variance σ^2.

Dès l'instant que l'on raisonne en régression conditionnellement aux observations $x_j(i)$ des v.a. X_j, c'est-à-dire en les supposant fixées, le modèle de régression est équivalent au modèle linéaire: c'est ainsi que nous allons raisonner jusqu'à nouvel ordre. L'expression X=x signifiera "pour $X_j = x_j$, $\forall\, j = 1,\ldots,p$ ".

Il est fréquent d'utiliser la notation matricielle pour exprimer le modèle linéaire. Pour cela, on note:

— X la matrice à n lignes numérotées de 1 à n et p+1 colonnes numérotées j de 0 à p dont le terme général $x_j(i)$ est l'observation de la j^e variable explicative sur l'unité statistique de rang i. La ligne x(i) est définie par la suite $(x_j(i))$ j=0,p et la colonne x_j par la suite $(x_j(i),i=1,n$. La variable X_0 prend par définition la valeur 1 pour toute unité statistique i: le coefficient β_0 est alors le coefficient de régression de cette variable.

— Y le vecteur colonne à n lignes dont le terme y(i) est la variable expliquée sur l'unité statistique de rang i.

— ε le vecteur colonne à n lignes dont le terme $\varepsilon(i)$ est la variable résiduelle de rang i.

— β le vecteur colonne à p+1 lignes dont le terme général est le coefficient de régression β_j.

Le terme aléatoire est ε; la matrice X est connue et le vecteur β est certain mais inconnu. Le modèle linéaire s'écrit alors de la façon suivante:

$$Y = X\beta + \varepsilon$$

Il n'existe mathématiquement qu'une contrainte dans le choix des variables explicatives: elles ne doivent pas être linéairement liées, c'est-à-dire qu'il ne faut pas que l'une des variables se déduise des autres par combinaison linéaire. Nous verrons pourquoi dans le paragraphe suivant.

Rien n'empêche donc, a priori, de choisir comme variables explicatives des fonctions d'autres variables explicatives, pourvu que ce ne soit pas des fonctions linéaires. Par exemple, on peut introduire en X_1 une variable X, en X_2 son carré X^2, en X_3 son cube X^3 etc... Ce modèle est dit polynomial en X; il est souvent utilisé lorsque la variable X est le temps (Cf chapitre précédent). C'est encore un modèle linéaire par rapport aux coefficients β_j.

Nous discuterons ultérieurement du choix raisonné des variables explicatives, qui est un problème ardu de la régression.

1.2 Estimation des coefficients de régression.

Pour estimer les coefficients de régression, on procède comme dans le chapitre précédent; le critère des moindres carrés s'écrit:

$$S(b_0, b_1, b_2, \ldots, b_p) = \sum_{i=1}^{n} [y(i) - \sum_{j=0}^{p} b_j x_j(i)]^2$$

en posant $x_0(i)=1$ $\forall i=1,\ldots,n$.

Cette fonction admet un minimum au point où toutes les dérivées partielles sont nulles:

$$\forall j=0,\ldots,p \quad \partial S/\partial b_j = 0$$

On a:

$$\forall j=0,\ldots,p \qquad \partial S/\partial b_j = -2 \sum_{i=1}^{n} [(y(i) - \sum_{j=0}^{p} b_j\, x_j(i)]\, x_j(i) = 0$$

On retrouve le système des "équations normales", à p+1 inconnues et p+1 équations:

$$\begin{aligned}
&b_0 \sum_{i=1}^{n} x_0(i)^2 + b_1 \sum_{i=1}^{n} x_0(i)x_1(i) + \ldots + b_p \sum_{i=1}^{n} x_0(i)x_p(i) = \sum_{i=1}^{n} x_0(i)y(i)\\
&\qquad\qquad\qquad\qquad \ldots\ldots\ldots\\
&b_0 \sum_{i=1}^{n} x_0(i)x_1(i) + b_1 \sum_{i=1}^{n} x_1(i)^2 + \ldots + b_p \sum_{i=1}^{n} x_1(i)x_p(i) = \sum_{i=1}^{n} x_1(i)y(i)\\
&\qquad\qquad\qquad\qquad \ldots\ldots\ldots\\
&b_0 \sum_{i=1}^{n} x_0(i)x_p(i) + b_1 \sum_{i=1}^{n} x_0(i)x_p(i) + \ldots + b_p \sum_{i=1}^{n} x_p(i)^2 = \sum_{i=1}^{n} x_p(i)y(i)
\end{aligned}$$

La matrice M du système linéaire ci-dessus possède p+1 lignes et p+1 colonnes et a pour terme général $\Sigma\, x_k(i)\, x_l(i)$; elle est égale au produit matriciel $X^t X$, X^t étant la matrice transposée de X. Ce système se met sous la forme matricielle très simple suivante:

$$M B = X^t Y$$

Ces notations sont analogues aux notations utilisées dans le paragraphe précédent.

Pour calculer les coefficients de régression $b_0, b_1, \ldots, b_p$, il suffit donc de calculer la matrice inverse de M, si elle existe: on retrouve exactement la même démarche qu'en régression simple. Si la matrice M n'est pas inversible, cela signifie que les variables explicatives sont liées: l'une au moins peut être reconstruite exactement à l'aide des autres par une combinaison linéaire. Il est indispensable d'éliminer ces variables de l'ensemble des variables explicatives.

Notons en outre que les programmes ne détectent pas toujours les matrices non inversibles.

A partir de maintenant, nous supposons que la matrice M est inversible. En notant M^{-1} la matrice inverse, on a:

$$B = M^{-1} X^t Y$$

Les termes du vecteur B ainsi calculé sont les estimations des coefficients de régression β_j pour j=0,...,p. On en déduit:

— la valeur de la variable estimée en chaque point:

$$y_e(i) = \sum_{j=0}^{p} b_j \, x_j(i)$$

On note Y_e le vecteur colonne défini par la suite $(y_e(i))i=1,...n$.

— les résidus observés:

$$e(i) = y(i) - y_e(i).$$

Ces résidus ne sont pas les observations de la variable ε puisqu'ils sont calculés à l'aide des coefficients de régression estimés b_j. On note E le vecteur colonne défini par la suite (e(i))i=1,n.

Les notations précédentes permettent d'écrire les relations:

$$Y = X\beta + \varepsilon = X\,B + E$$

2. PROJECTION ORTHOGONALE ET VARIABLES REDUITES.

2.1 Opérateur de projection orthogonale.

Nous allons maintenant étudier l'application de $\mathbf{R}^n$ dans $\mathbf{R}^n$ qui à Y fait correspondre $Y_e = X\,B$ de façon à en proposer une interprétation géométrique.

Le vecteur Y, constitué des n observations y(i), appartient en effet à $\mathbf{R}^n$ et la variable estimée $Y_e = X\,B$, qui appartient aussi à $\mathbf{R}^n$, s'exprime de la façon suivante:

$$Y_e = X\,M^{-1}\,X^t\,Y$$

$$= X\,[X^t\,X\,]^{-1}\,X^t\,Y$$

Nous définissons ainsi l'application P qui à Y fait correspondre Y_e définie par la formule ci-dessus. Cherchons l'image Y_{ee} de Y_e par cette application:

$$Y_{ee} = X\,[X^t\,X\,]^{-1}\,X^t\,Y_e$$

$$= X\,[X^t\,X\,]^{-1}\,X^t\,X\,[X^t\,X\,]^{-1}\,X^t\,Y$$

$$= X\,[X^t\,X\,]^{-1}\,X^t\,Y$$

$$= Y_e$$

Nous venons de montrer une propriété caractéristique d'un projecteur, qui est l'idempotence:

$$P\,[\,P(Y)\,] = P(Y) = Y_e$$

Calculons le produit scalaire $[Y - Y_e]\,.\,Y_e$ défini dans $\mathbf{R^n}$ comme la somme des produits des termes de même rang; ce produit scalaire s'exprime à l'aide des produits matriciels ci-dessous:

$$[Y - Y_e]^t\,Y_e = Y^t\,Y_e - Y_e{}^t\,Y_e$$

$$= Y^t\,Y_e - [X\,M^{-1}\,X^t\,Y]^t\,[\,X\,M^{-1}\,X^t\,Y]$$

$$= Y^t\,Y_e - Y^t\,X\,[M^{-1}]^t\,X^t\,X\,M^{-1}\,X^t\,Y$$

$$= Y^t\,Y_e - Y^t\,X\,[M^{-1}]^t\,X^t\,Y$$

$$= Y^t\,Y_e - Y^t\,Y_e \qquad ([M^{-1}]^t = M^{-1})$$

$$= 0$$

Cette propriété caractérise les opérateurs de projection orthogonale définis sur l'espace euclidien $\mathbf{R^n}$.

La représentation géométrique (fig. 2.1) montre l'orthogonalité du sous-espace F engendré par les variables X_j et de $Y\text{-}Y_e$; cette dernière n'est autre que la variable E définie par les résidus e(i). On généralise ainsi la propriété que nous avons donnée sans démonstration dans le chapitre 1 à l'ensemble des variables explicatives:

Théorème: la série des résidus (e(i))i=1,n est centrée et non-corrélée aux variables explicatives X_j pour j=1 à p.

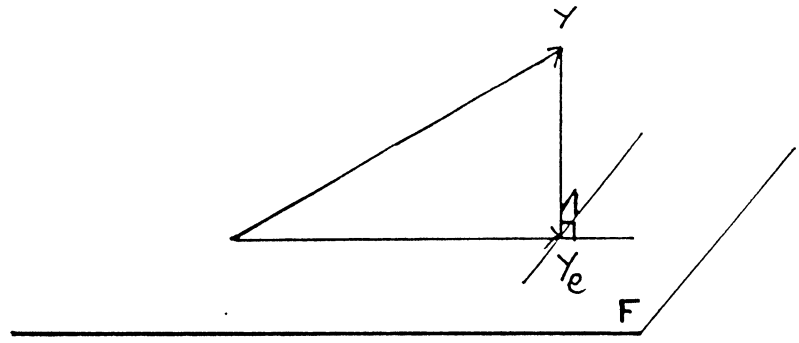


Fig. 2.1: Interprétation géométrique
de la régression multilinéaire

2.2 Variables centrées réduites (régression multilinéaire).

Pour simplifier les notations nous avons introduit la variable x_0 constante et égale à 1. On peut réécrire les équations normales en tenant compte de cette propriété:

La première équation s'écrit :

$$n\, b_0 + b_1 \sum_{i=1}^{n} x_1(i) + \ldots + b_p \sum_{i=1}^{n} x_p(i) = \sum_{i=1}^{n} y(i)$$

Soit, en divisant par n:

$$b_0 + b_1 \bar{x}_1 + \ldots + b_p \bar{x}_p = \bar{y}$$

On peut diviser les autres équations normales par n puis leur soustraire l'équation ci-dessus: on obtient alors:

$$\forall j=1,\ldots,p \quad b_1\, s(x_j)^2 + \ldots + b_p\, cov\,(x_j, x_p) = cov\,(x_j, y)$$

où $s(x_j)^2$ et $cov(x_j, x_k)$ sont les variances et covariances des séries statistiques $x_j(i)$ et $x_k(i)$, pour i=1,n.

Le système des équations normales peut donc être exprimé à l'aide de la matrice de covariances de $(X_1, X_2, ..., X_p)$, le terme constant b_0 étant déduit des autres coefficients a_j par la relation:

$$b_0 = \bar{y} - (b_1\bar{x}_1 + ... + b_p \bar{x}_p)$$

Cette propriété du coefficient constant permet de supposer que les variables mises en jeu dans le modèle sont centrées. La matrice $X^t X$, de dimension p puisqu'il n'y a plus de terme constant, est égale alors à n fois la matrice des covariances entre les variables explicatives.

En outre, il est souvent commode, dans la pratique, d'étudier les coefficients de régression calculés sur les variables centrées réduites. La relation entre ces derniers et les coefficients de régression sur les variables initiales est immédiate; le modèle de régression est, sur les variables initiales:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_p X_p + \varepsilon$$

En centrant les variables, on fait disparaître la constante. Notons Y' et X_j' les variables centrées réduites, σ_j^2 les variances des variables explicatives X_j et σ_y^2 la variance de la variable expliquée; il vient:

$$Y' = \beta_1 \sigma_1 X_1' / \sigma_Y + \beta_2 X_2' \sigma_2 / \sigma_Y + ... + \beta_p X_p' \sigma_p / \sigma_Y + \varepsilon / \sigma_Y$$

Soit:

$$Y' = \beta_1' X_1' + \beta_2' X_2' + \beta_3' X_3' + ... + \beta_p' X_p' + \varepsilon'$$

La variance de la variable résiduelle est divisée par σ_Y^2 et les coefficients de régression β_j' se déduisent des coefficients de régression β_j par la formule:

$$\beta_j' = \beta_j \sigma_j / \sigma_Y$$

Les coefficients de régression β'_j sont relatifs à des variables centrées réduites: on peut donc interpréter directement leur taille et les comparer entre eux.

Toutes ces propriétés sont évidemment vraies pour les estimations.

Dans les représentations graphiques, nous serons amenés à supposer que les variables sont centrées et réduites. La figure 2.2 donne ainsi l'interprétation géométrique dans $\mathbf{R^n}$ de la régression d'une variable centrée Y par deux variables centrées X_1 et X_2; la variable constante X_0 égale à 1 ne figure pas sur ce schéma puisqu'elle appartient à l'orthogonal du sous-espace qui est représenté et qui est engendré par les variables X_1 et X_2 considérées.

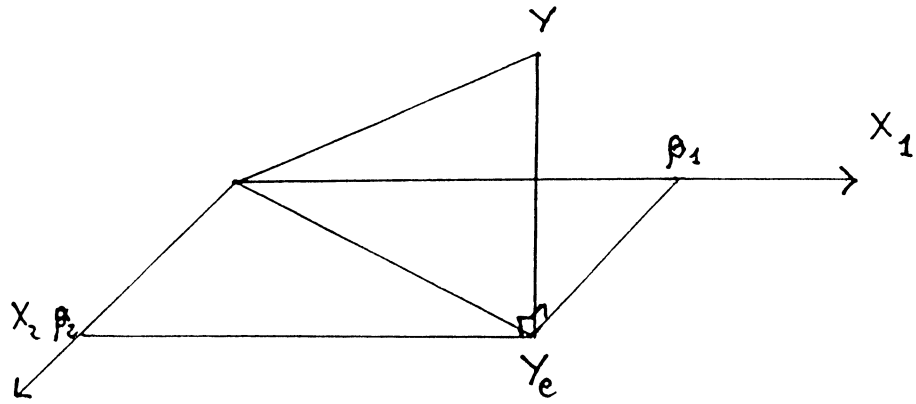


Fig. 2.2: Régression de Y par X_1 et X_2
(variables centrées)

3. PROPRIETES DES ESTIMATEURS ET DES RESIDUS

3.1 Propriétés des estimateurs des coefficients de régression.

Nous avons calculé dans le paragraphe précédent un vecteur $b = (b_0, b_1, \ldots, b_j, \ldots, b_p)$ qui minimise la somme des carrés des erreurs. Ce vecteur b est une estimation du vecteur de régression $\beta = (\beta_0, \beta_1, \ldots, \beta_j, \ldots, \beta_p)$ dont l'existence est supposée par le modèle linéaire.

Théorème: le vecteur B est l'estimateur linéaire efficace du vecteur de régression β et sa matrice de covariance est donnée par:

$$V_B = \sigma^2 M^{-1}$$

Un estimateur efficace est un estimateur de variance minimale dans la classe des estimateurs sans biais.

Nous admettrons ce théorème dont la démonstration est donnée dans tous les ouvrages classiques (Saporta, 1990 par exemple).

L'estimateur de V_B est obtenu en remplaçant la variance résiduelle σ^2 par son estimateur sans biais s^2 (Cf paragr. suivant).

Conséquence: la v.a. B_j est un estimateur sans biais du coefficient de régression β_j et sa variance σ_j^2 est égale au terme diagonal de la matrice V_B.

Remarques:

— les estimateurs des coefficients de régression ne sont donc pas indépendants; nous verrons ultérieurement qu'ils peuvent être fortement corrélés, comme nous l'avons vu dans le chapitre précédent à propos du coefficient directeur de la droite de régression et de son terme constant;

— on montre en outre que, si la variable résiduelle ε est gaussienne, B est l'estimateur du maximum de vraisemblance et est lui-même gaussien; les estimateurs B_j de chaque coefficient de régression β_j sont alors gaussiens et on peut en donner des intervalles de confiance ou effectuer des tests de Student: pour tester l'égalité d'un coefficient de régression β_j à une valeur spécifiée β_j^0, on utilise la statistique T_j définie par:

$$T_j = (B_j - \beta_j^0) / S_j$$

où S_j est l'estimateur sans biais de l'écart-type de B_j, qui suit la loi de Student à de degré de liberté n-2 si la variable résiduelle est gaussienne (paragr. 1.2);

— les estimateurs B_j et B_k n'étant pas indépendants, tester l'égalité du coefficient β_j à β_j^0 puis de β_k à β_k^0 n'est pas équivalent à tester l'égalité du couple (β_j, β_k) à (β_j^0, β_k^0).

3.2 Etude des résidus. Coefficient de corrélation multiple. Prévision.

L'étude de la variable résiduelle rencontre les mêmes difficultés en régression multilinéaire qu'en régression linéaire: les coefficients de régression β_j n'étant pas connus, la variable ε n'est pas observable.

On ne dispose que des résidus e(i) en chaque point; nous avons vu que ces résidus sont centrés et non corrélés aux variables explicatives.

Soit $s^2(e)$ la variance observée des résidus:

$$s^2(e) = 1/n \sum_{i=1}^{n} [y(i) - \sum_{j=0}^{p} b_j x_j(i)]^2 = 1/n \sum_{i=1}^{n} e^2(i)$$

Cette variance nous donne une estimation sans biais $s^2 = n\, s^2(e)/ (n-p-1)$ de la variance résiduelle. C'est la réalisation de la variable $S^2 = SCR/(n - p -1)$, où SCR est la somme des carrés des résidus:

$$SCR = \sum_{i=1}^{n} [\, Y(i) - Y_e(i) \,]^2$$

$Y_e(i)$ est ici l'estimateur ci-dessous:

$$Y_e(i) = \sum_{j=0}^{p} B_j x_j(i)$$

Théorème: la statistique SCR/σ^2 suit la loi du χ^2 de degré de liberté n-p-1. Elle est non corrélée aux v.a. B_j et aux v.a. $Y_e(i)$ (indépendante si la variable résiduelle est gaussienne).

Le degré de liberté s'explique par l'orthogonalité aux p+1 variables explicatives.

Ce théorème nous permet d'estimer la variance résiduelle par intervalle de confiance.

Définition: on appelle coefficient de corrélation multiple le coefficient de corrélation entre la variable expliquée Y et la variable estimée Y_e. Le coefficient de détermination est le carré du coefficient de corrélation multiple. Ces coefficients sont notés R et R^2.

Théorème: le coefficient de détermination vérifie l'égalité ci-dessous:

$$s^2(e) = (1 - R^2)\ s^2(y)$$

Théorème: si tous les coefficients de régression β_j, j=1,...,p sont nuls, la statistique:

$$F = \frac{(n - p - 1)}{p} \frac{R^2}{1 - R^2}$$

suit la loi de Snedecor de degrés de liberté p et n - p - 1.

On rejettera donc l'hypothèse que tous les coefficients de régression (sauf le terme constant) sont nuls, ou, ce qui revient au même, que la valeur théorique de R^2 est égale à 0, pour les valeurs de F appartenant à la région critique $[F_\alpha, +\infty[$, où F_α dépend du risque de première espèce α et des degrés de liberté.

Nous admettons bien sûr toutes les propriétés précédentes dont des exemples sont donnés dans le paragraphe 2.

La figure 2.3.1 donne la représentation des variables non centrées, la figure 2.3.2 celle des variables centrées. La seconde peut être considérée comme la projection orthogonale de la première sur le sous-espace de $\mathbf{R}^n$ orthogonal au vecteur constant égal à 1.

La figure 2.3.2 permet d'interpréter géométriquement la variance des résidus et le coefficient de corrélation multiple:

— le coefficient de corrélation multiple est le cosinus de l'angle θ formé par les vecteurs Y et Y_e;

— la variance des résidus est le carré de la norme du vecteur Y - Y_e divisé par n;

— les coefficients de régression b_1 et b_2 sont les coordonnées de Y_e sur les axes représentant les variables explicatives.

— le terme constant, qui est le coefficient de régression sur la variable X_0, est invisible sur la figure 2.3.2. Par contre on peut le voir sur la figure 2.3.1.

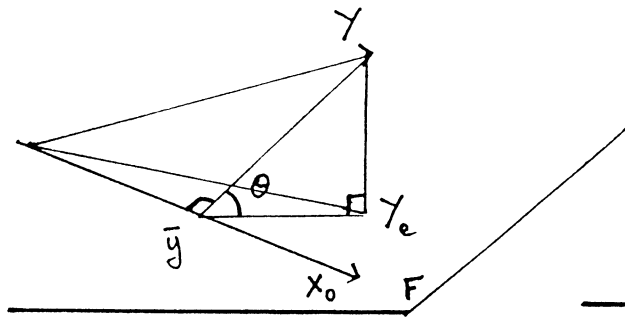


Fig. 2.3.1: Variance des résidus et corrélation multiple (variables non centrées)

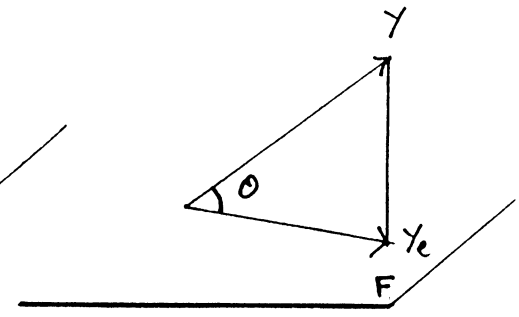


Fig. 2.3.2: Variance des résidus et corrélation multiple (variables centrées)

Pour prédire la variable expliquée en fonction des variables explicatives, on procède comme en régression simple: on peut prédire soit l'espérance conditionnelle, soit la valeur d'un point, en utilisant les observations des variables explicatives. La prédiction ponctuelle est la même, seule change la variance de l'estimateur. Pour estimer E(Y/X=x), on utilise les estimateurs des coefficients de régression:

Théorème: la statistique Y' définie par:

$$Y' = \sum_{j=1}^{p} B_j \, x_j$$

est un estimateur efficace de l'espérance conditionnelle E(Y/X=x):

$$E(Y/X=x) = \sum_{j=1}^{p} \beta_j \, x_j$$

et sa variance est égale à:

$$V(Y') = \sigma^2 \, [x]^t \, M^{-1} \, [x]$$

où $[x]^t$ est le vecteur ligne $[x_0,....,x_j....x_p]$.

La variance de la prévision d'une observation est obtenue en ajoutant simplement σ^2 à la variance précédente.

Pour obtenir des estimateurs, il suffit de remplacer la variance résiduelle σ^2 par son estimateur sans biais s^2.

Pour obtenir une estimation par intervalle de confiance, on utilise la statistique $(Y' - y)/ S_{Y'}$, où $S_{Y'}^2$ est l'estimateur sans biais de la variance V(Y'). Cette statistique suit en effet le loi de Student de degré de liberté n - p-1.

4. EXEMPLE NUMERIQUE.

Nous avons suggéré dans le chapitre 2 de compléter l'index cardiaque par la pression artérielle pulmonaire ou la pression diastolique pour mieux expliquer la résistance pulmonaire des malades étudiés. Nous donnons ci-dessous les résultats numériques obtenus en introduisant la pression diastolique.

Pourquoi avoir choisi la pression diastolique ? Pourquoi pas les deux variables envisagées? En examinant leur coefficient de corrélation (0.928), on peut se dire qu'introduire les deux variables est inutile car les informations qu'elles apportent sur les données sont presque les mêmes. Nous verrons ulté-

rieurement comment en juger de manière satisfaisante (par l'étude des coefficients de corrélation partielle).

Le modèle de régression que nous considérons est donc le suivant:

$$\text{Répul} = \beta_0 + \beta_1 \text{ Incar} + \beta_2 \text{ Prdia} + \varepsilon$$

Les paramètres statistiques des variables explicatives considérées sont les suivants:

VARIABLES	MOYENNES	ECARTS-TYPES	VARIANCES
Incar	1.846	.65575	.43000
Prdia	19.259	5.78051	33.41429
Répul	7.049929	.529936	.28083

Les variances des variables explicatives sont très différentes l'une de l'autre; pour interpréter les coefficients de régression, il sera utile d'examiner leurs estimations sur les variables réduites.

MATRICE DES CORRELATIONS ENTRE LES VARIABLES

	Incar	Prdia	Répul
Incar	1.000		
Prdia	-0.361	1.000	
Répul	-0.839	0.761	1.000

La matrice ci-dessus fait apparaître de fortes corrélations en valeur absolue entre la variable expliquée et les variables explicatives. Le coefficient de détermination (supérieur aux carrés des coefficients de corrélation), sera élevé; le coefficient de corrélation entre les variables explicatives est relativement faible; il semble donc qu'il y ait complémentarité entre ces deux variables explicatives (ce raisonnement n'est valable que pour deux variables).

Dans le tableau ci-dessous, on peut lire que le coefficient de détermination est égal à 0.946; la valeur du F que l'on en déduit est égale à 864.502, sa

probabilité critique P(F > 864.502) est numériquement nulle: cela montre qu'elle appartient à la région critique quel que soit le risque de première espèce choisi: l'hypothèse de nullité des coefficients de régression est rejetée.

La variance résiduelle estimée est ici égale à 0.0155, nettement plus faible que dans la régression à l'aide de la seule variable Incar (0.0846); l'écart-type résiduel estimé est égal à 0.125 contre 0.291 précédemment. L'introduction de la pression diastolique améliore donc considérablement la reconstruction de la résistance pulmonaire.

ANALYSE DE VARIANCE

	ddl	Somme des Carrés	Variance Estimée	Pourcentage de var.tot
Tot	100	283.6404D-01	283.6404D-03	1
Exp	2	268.4260D-01	268.1155D-03	946.3603D-03
Res	98	152.1440D-02	155.2490D-04	0.0536

Corrélation multiple 0.9728 Détermination 0.9464
F(2 , 98) 864.502 Probabilité critique 0.0000

Comme nous l'avons suggéré précédemment, nous examinons tout d'abord les coefficients de régression calculés sur les variables centrées réduites: ils sont de taille relativement proche en valeur absolue l'un de l'autre, ce qui signifie qu'aucune des variables n'a une importance prédominante dans la régression. Ils sont du signe du coefficient de corrélation correspondant, ce qui paraît naturel (cette propriété n'est pas toujours vérifiée).

COEFFICIENTS DE REGRESSION

N°	Estimation	Ecart-type	Estimation (var. réd.)	T de Student
Incar	-0.52458	0.020271	-0.6491	-25.878
Prdia	0.04835	0.002300	0.5274	21.024
Cst	7.08705	0.068631	0	103.264

Le t de Student est très élevé, supérieur à 20 en valeur absolue et a donc une probabilité critique numériquement nulle (P(|T|>21.024)=0). Nous avons une propriété supplémentaire ici par rapport au test du F sur le coefficient de

détermination: on rejette l'hypothèse de nullité de chaque coefficient.

Les intervalles de confiance des coefficients de régression sont définis comme l'ensemble des valeurs β_j tels que la valeur observée t de la statistique $T = (B_j - \beta_j)/s_j$ soit comprise entre -1.96 et 1.96 pour un risque de première espèce $\alpha = 0.05$ (la loi de Student de degré de liberté 98 est confondue avec la loi normale centrée réduite). On en déduit:

$$b_0 \in [\ 6.9525\ ,\ 7.2216\ ]$$
$$b_1 \in [-0.5643\ ,-0.4849\ ]$$
$$b_2 \in [\ 0.04384,\ 0.05286\ ]$$

Examinons maintenant les résidus obtenus: souvenons-nous que c'est par l'étude des résidus que nous avons été amenés à introduire la variable pression diastolique en seconde variable explicative: la résistance pulmonaire des unités statistiques 18, 59, 71 et 100 était particulièrement mal reconstruite par l'index cardiaque.

Les résidus concernant ces malades sont égaux, dans l'ordre précédent, à -0.188, -0.413 et -0.231 et 0.163 au lieu de -0.658, -0.885 et -0.704 et 0.636. On retrouve ici l'amélioration constatée sur un plan général quand on compare les variances résiduelles estimées. Seul le second (n° 59) reste anormal par rapport à l'écart-type résiduel estimé (0.125) puisqu'il est supérieur à 3 x 0.125 en valeur absolue: en examinant les données (p. 39) on remarque une forte valeur de l'index systolique sur l'unité statistique 59, contrairement aux autres unités 18 et 71: c'est peut-être l'explication de la particularité de cette observation.

La figure 2.4 et le test d'ajustement du χ^2 confirment la normalité des résidus.

La figure 2.5 donne la représentation linéaire des résidus. On y retrouve le résidu n° 59, mais trois résidus supérieurs à 2 fois l'écart-type apparaissent: il s'agit des n° 19, 63, 67. ils restent dans des limites acceptables (inférieurs à 3 fois l'écart-type).

Le nombre de valeurs à l'extérieur de l'intervalle $\pm$ 2 x s n'est pas contradictoire avec l'hypothèse que la variable résiduelle suive la loi normale; cette

hypothèse n'est donc pas contredite par les observations effectuées; les tests de Student, de Fisher et les intervalles de confiance sont justifiés.

Régression de la résistance pulmonaire
par l'index cardiaque et la pression diastolique

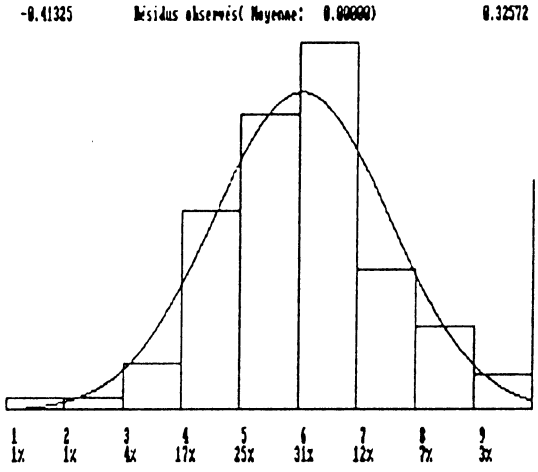


Fig 2.4 Histogramme des résidus observés.
Ajustement par la loi normale.

TEST DU χ^2

Cl.	Eff.	Probab.	np_i
1	6	0.0869	8.78
2	17	0.1579	15.94
3	25	0.2464	24.89
4	31	0.2500	25.25
5	12	0.1647	16.64
6	7	0.0705	7.12
7	3	0.0236	2.38

Résidus observés
Test du Chi-2: 3.7168 Ddl: 5
Prob. cr.: 0.5909

```
|                                    6367
|          19
2s-------------------------------------------------------
|                                                  89
|     11    2*          41             6*          88     100
|      12         29          47              72   81   91  97
|       15         30     42        56   65 70       84       99
| 6   1*1*              39 45      5* 64        **82 86
0 58----17-22--------**---4*5*---------69-7*-------9*93-------
|3  9           27  3*    *3  5155  61                    98
2           20 26 3*35               62       7579       9*
|             23 28                58              8387   95 101
14 7                          49             73      85
|         18       32            54
-2s----------25-----------------------71-------------------
|
|
|
|                                    59
```

Fig. 2.5: Représentation linéaire des résidus observés (s=0.124599)

Chapitre 3

CORRELATIONS PARTIELLES
REGRESSIONS PAS A PAS

1. NOTION DE CORRELATION PARTIELLE.

Dans l'exemple numérique étudié, nous avons constaté que l'introduction de la pression diastolique pour compléter l'index cardiaque améliore considérablement la régression: la variance résiduelle estimée passe de 0.0812 à 0.0154 et le coefficient de détermination augmente de 0.7044 à 0.9464. La variable X = Prdia apporte donc une information sur la variable expliquée Y = Répul étrangère à celle qui est donnée par la première variable explicative X_1 = Incar. Cette information supplémentaire se mesure par le coefficient de corrélation partielle.

1.1 Coefficient de corrélation partielle.

Définition: on appelle coefficient de corrélation partielle de X et Y conditionnellement à X_1 le coefficient de corrélation $R(X,Y/X_1)$ entre $Z_1 = Y - (\beta_1 X_1 + \beta_0)$, et $Z_2 = X - (\alpha_1 X_1 + \alpha_0)$, où β_1, β_0 et α_1, α_0 sont les coefficients des régressions de Y et de X par X_1.

Les variables Z_1 et Z_2 sont toutes deux non corrélées à X_1 puisqu'elles sont définies par les variables résiduelles des régressions de Y et X par X_1. L'estimateur de ce coefficient $R(X,Y/X_1)$ est l'estimateur empirique, calculé sur

les valeurs observées, que nous appellerons aussi coefficient de corrélation partielle et que nous noterons de la même façon.

L'interprétation géométrique est donnée par la figure 3.1, dans laquelle nous avons supposé que les variables sont centrées et réduites.

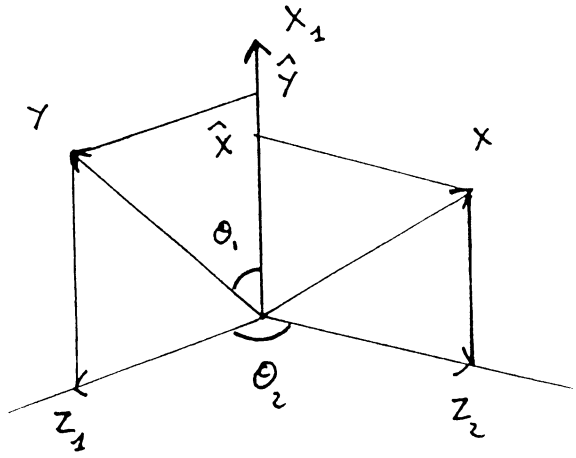


Fig. 3.1: Coefficient de corrélation partielle
de Y et X conditionnellement à X_1

Ce coefficient de corrélation partielle se généralise facilement à des groupes de variables: si X_1, X_2, ..., X_q sont les variables explicatives considérées, le coefficient de corrélation partielle de X et de Y conditionnellement à X_1, X_2, ..., X_q est le coefficient de corrélation des résidus obtenus par la régression de Y par les variables X_j et des résidus obtenus par la régression de X par les variables X_j. On le note $R(Y,X/X_1, X_2,..., X_q) = R(Y,X/X_{.q})$

Théorème: si la variable résiduelle ε est gaussienne et si le coefficient de corrélation partielle théorique est nul, la statistique $F_{1,n-q-2}$ définie par:

$$F_{1,n-q-2} = (n - q - 2) \frac{R(Y,X/X_{.q})^2}{1 - R(Y,X/X_{.q})^2}$$

suit la loi de Snedecor de degrés de liberté 1 , n - q - 2.

Dans notre exemple, ce coefficient de corrélation partielle est estimé à R(Prdia, Répul/Incar) = 0.905 et $F_{1,98}$ = 418.41. Une table nous donne comme région critique du test sur $F_{1,98}$ [6.90, + ∝ [pour un risque de première espèce égal à 0.01 et des degrés de liberté égaux à 1 et 98: il est clair que l'on rejette l'hypothèse de nullité du coefficient de corrélation partielle théorique et que la diminution de la variance résiduelle due à l'introduction de la variable Prdia comme variable explicative ne peut être un effet du hasard.

L'étude des autres coefficients de corrélation partielle montre que l'on aurait pu introduire à la place de la pression diastolique la variable pression artérielle pulmonaire (Papul) dont le coefficient de corrélation partielle est 0.936, supérieur au précédent (0.905):

```
Corrélations partielles avec Répul

Variable explicative considérée: Incar
R2 = 0.70443        F(1,99) = 235.9418        Prob. crit. = 0.0000

         Frcar  Insys  Prdia  Papul  Pvent
Répul    0.357 -0.354  0.905  0.936  0.156

Coef. de corr. 0.3540:
           Valeur du F  14.04 Probabilité critique  0.0004

Coef. de corr. 0.1560:
           Valeur du F   2.44 Probabilité critique  0.1170
```

Seul le coefficient de corrélation partielle de la résistance pulmonaire (Répul) et de la pression ventriculaire (Pvent) n'est pas significatif (r = 0.156, F = 2.44 P(F > 2.44) = 0.117) et ne témoigne pas d'une liaison non aléatoire entre les deux variables compte tenu de l'information apportée par l'index cardiaque.

Nous retrouvons ici ce que nous avions pressenti et expliqué dans le chapitre 2: l'information apportée sur la résistance pulmonaire par l'index cardiaque peut être complétée par la pression diastolique ou la pression articulaire pulmonaire.

A l'aide de la figure 3.1, on montre que le carré du coefficient de corrélation partielle mesure la diminution relative de la variance résiduelle observée. Plus précisément, on montre que:

$$R^2(Y,X/X_1) = \frac{s_q(e)^2 - s_{q+1}(e)^2}{s_q(e)^2}$$

où $s_q(e)^2$ et $s_{q-1}(e)^2$ sont les variances empiriques des résidus dans les régressions de Y par $X_1, ..., X_q$ et par $X_1, ..., X_q$, X respectivement.

L'introduction d'une variable parmi l'ensemble des variables explicatives a donc pour effet de diminuer la variance des résidus, ou, ce qui revient au même, d'augmenter le coefficient de détermination. Cette propriété n'est pas nécessairement vérifiée par l'estimateur sans biais de la variance résiduelle qui dépend du nombre de variables explicatives.

En conclusion, l'examen des coefficients de corrélation partielle nous permet donc de mieux choisir les variables explicatives supplémentaires en évitant les redondances d'information et en mettant en évidence la complémentarité des variables: l'augmentation du coefficient de détermination est d'autant plus importante que le coefficient de corrélation partielle de la variable introduite dans l'ensemble des variables explicatives est élevé.

Ainsi, suivant que l'on ajoute la pression articulaire pulmonaire ou la pression diastolique, le coefficient de détermination atteint 0.94636 ou 0.96356.

1.2 Coefficients de détermination partielle.

La formule précédente permet de généraliser la notion de coefficient de corrélation partielle à un ensemble de variables $Z_1, Z_2, ..., Z_r$, en mesurant la diminution de la variance des résidus quand on introduit les Z_j comme variables explicatives de la régression: la statistique $R^2(Y,Z_1,Z_2,...,Z_r/X_1,...,X_q)^2$, que l'on note $R^2(Y,Z_{.r}/X_{.q})^2$, analogue à un coefficient de détermination est appelée coefficient de détermination partielle:

$$R^2(Y,Z_{.r}/X_{.q}) = \frac{s_q(e)^2 - s_{q+r}(e)^2}{s_q(e)^2}$$

où $s_{q+r}(e)^2$ est la variance des résidus dans la régression de Y par les variables X_j, j = 1,q et Z_k, k = 1,r.

Théorème: si la variable résiduelle ε est gaussienne et si le coefficient de détermination partielle théorique est nul, la statistique $F_{r,n-q-r-1}$ définie par:

$$F_{r,n-q-r-1} = \frac{(n - q - r - 1)}{r} \frac{R(Y,Z_{.r}/X_{.q})^2}{1 - R(Y,Z_{.r}/X_{.q})^2}$$

suit la loi de Snedecor de degrés de liberté r , n - q - r - 1.

Par exemple, en ajoutant simultanément la pression diastolique et la pression artérielle pulmonaire à l'index cardiaque, les sommes des carrés des résidus passent de 8.3837 à 0.94642, les statistiques $R^2(Y,Z_{.r}/X_{.q})$ et $F_{r,q}$ prennent les valeurs 0.88711 et 179.16. La probabilité critique $P(F_{r,q} > 179.16)$ est numériquement nulle: l'information apportée par les deux variables est hautement significative. Cette conclusion n'est d'ailleurs pas étonnante puisque l'on a déjà refusé l'hypothèse de nullité du coefficient de régression de la pression diastolique.

2. ANALYSE DES CORRELATIONS PARTIELLES

2.1 Choix raisonné des prédicteurs.

Le choix raisonné des prédicteurs est effectué par l'utilisateur, en fonction des données qu'il étudie et de ses objectifs.

L'avantage, par rapport aux procédures automatiques que nous présentons dans le paragraphe suivant, est de toujours contrôler le système en cours de constitution et de pouvoir tenir compte simultanément de plusieurs critères, en particulier de la colinéarité entre les variables explicatives (il est préférable en effet que le coefficient de détermination de chaque variable explicative dans la régression par les autres soit faible) et de la variance résiduelle estimée.

La démarche consiste à étudier, après chaque introduction d'un prédicteur, les coefficients de corrélation partielle des variables restantes avec la variable expliquée (pour augmenter le coefficient de détermination de la

régression), la variance résiduelle (pour diminuer la variance des estimations), les coefficients de détermination entre les variables explicatives et le système des prédicteurs déjà introduits (pour éviter les colinéarités).

Nous avons appliqué cette démarche aux données sur l'infarctus:

— on calcule les coefficients de corrélation des variables explicatives avec la variable expliquée: le premier prédicteur est celui pour lequel ce coefficient est, en valeur absolue, le plus élevé en étant significativement non nul bien que cette restriction ne soit pas toujours justifiée comme nous l'indiquons en 2.2 .

CORRELATIONS ENTRE LES VARIABLES

	Frcar	Incar	Insys	Prdia	Papul	Pvent
Répul	0.287	-0.839	-0.833	0.761	0.716	0.318

Nous introduisons ainsi pour expliquer la résistance pulmonaire l'index cardiaque Incar dont le coefficient de corrélation (-0.839) est très largement significatif. Cette décision est uniquement basée sur les résultats numériques: en tenant compte de considérations médicales ou autres, on aurait pu choisir l'index systolique Insys dont le coefficient de corrélation avec la résistance pulmonaire est presque égal au précédent (-0.833).

— on calcule les coefficients de corrélation partielle des variables explicatives restantes et de la variable expliquée conditionnellement au prédicteur précédemment introduit, et leur coefficient de détermination avec le système de prédicteurs.

Les résultats numériques ci-dessous font apparaître deux coefficients de corrélation partielle hautement significatifs (leur probabilité critique est numériquement nulle): les variables concernées sont la pression diastolique et la pression artérielle pulmonaire (Prdia et Papul).

Les coefficients de détermination avec le système de prédicteurs considéré (limité ici à l'index cardiaque) sont tous deux voisins de 0: il est à peu près équi-

valent, au plan statistique, de choisir l'une ou l'autre de ces variables: on peut choisir par exemple celle qui est la plus facile ou la moins chère à observer.

```
                Variables explicatives considérées:
Incar
R2 = 0.70443          F(1,99) = 235.9418          Prob. crit. = 0.0000
Variance résiduelle estimée = 8.4683D-02
-------------------------------------------------------------------
Coefficients de détermination de chaque var. par rapport aux var.
explicatives:
Frcar:1 (R2 = 0.013)|Insys:3 (R2 = 0.787)|Prdia:1 (R2 = 0.013)|
Papul:5 (R2 = 0.073)|Pvent:6 (R2 = 0.080)|
-------------------------------------------------------------------
                      Corrélations partielles
          Frcar  Insys  Prdia  Papul  Pvent
Répul     0.357 -0.354  0.905  0.936  0.156

Coef. de corr. 0.9050 Valeur du F 443.51 Prob. critique  0.0000
```

Nous avons choisi la pression diastolique Prdia dont le coefficient de détermination avec l'index cardiaque est le plus faible.

— on recalcule les coefficients de corrélation partielle et les coefficients de détermination.

```
                Variables explicatives considérées:
Incar  Prdia
R2 = 0.94626          F(2,98) = 864.5018          Prob. crit. = 0.0000
Variance résiduelle estimée = 1.5525D-02
-------------------------------------------------------------------
Coefficients de détermination de chaque var. par rapport aux var.
explicatives:
Frcar:1 (R2 = 0.160)|Insys:3 (R2 = 0.817)|Papul:5 (R2= 0.866)|
Pvent:6 (R2 = 0.118)|
-------------------------------------------------------------------
                      Corrélations partielles
          Frcar  Insys  Papul  Pvent
Répul     0.019 -0.029  0.615 -0.069

Coef. de corr. 0.6150 Valeur du F  59.00 Prob. critique  0.0000
```

Le coefficient de détermination de la pression artérielle avec les prédicteurs déjà introduits est élevé (0.866); on montre que cela peut augmenter

la variance des estimateurs des coefficients de régression, ce qui serait gênant, mais on diminuera (peut-être) la variance résiduelle estimée. Le coefficient de corrélation partielle, largement significatif, ne laisse guère le choix.

— les coefficients de corrélation partielle après introduction de la pression artérielle pulmonaire sont les suivants:

```
                Variables explicatives considérées:
Incar  Prdia  Papul
R2 = 0.96663        F(3,97) = 936.69          Prob. crit. = 0.0000
Variance résiduelle estimée = 9.7569D-03
------------------------------------------------------------------
Coefficients de détermination de chaque var. par rapport aux var.
explicatives:
Frcar:1 (R2 = 0.160)|Insys:3 (R2 = 0.818)|Pvent1:6(R2 = 0.118)|
------------------------------------------------------------------
                     Corrélations partielles
         Frcar  Insys  Pvent
Répul    0.029  0.018 -0.073
```

Aucun de ces coefficients n'est significatif: on peut considérer que le système est complet.

Cette démarche donne finalement des résultats satisfaisants. Mais le système obtenu ne possède pas de propriété d'optimalité; en particulier, il dépend de l'ordre d'introduction des prédicteurs.

2.2 Algorithmes de sélection d'un système de prédicteurs.

Il peut être intéressant d'automatiser la démarche précédente: on parle alors d'algorithme de construction d'un ensemble de prédicteurs. Il existe trois algorithmes principaux:

— Le premier est le même que dans le paragraphe précédent, mais le choix de la variable explicative à introduire éventuellement est uniquement basé sur des considérations numériques; en général, l'algorithme ne tient pas compte du coefficient de détermination des variables explicatives avec le système de pré-

dicteurs et se limite à introduire la variable dont le coefficient de corrélation partielle est le plus élevé, tant que ce coefficient est significatif pour un risque de première espèce fixé.

On aurait ainsi introduit en deuxième variable explicative la pression artérielle pulmonaire au lieu de la pression diastolique suivant leur coefficient de corrélation partielle avec la résistance pulmonaire (0.936 contre 0.905). Le système final aurait été le même que précédemment, la pression diastolique étant introduite en troisième variable explicative (coefficient de corrélation partielle égale à 0.291, $P(F > 8.94)=0.0036$).

Certains programmes vérifient toutefois que la variable à introduire n'est pas colinéaire aux prédicteurs déjà considérés pour éviter des difficultés d'ordre numérique.

Cet algorithme définit ce que l'on appelle régression ascendante.

```
/// Introduction de la variable Incar
          F  235.931 Probabilité critique  0.00000
/// Introduction de la variable Papul
          F  696.582 Probabilité critique  0.00000
/// Introduction de la variable Prdia
          F    8.944 Probabilité critique  0.00363
```

— Le deuxième algorithme consiste à procéder de façon inverse: au départ, toutes les variables explicatives disponibles sont introduites, et l'on élimine au fur et à mesure celle dont le coefficient de corrélation partielle avec la variable expliquée conditionnellement aux autres prédicteurs est le plus faible, tout en étant non significatif.

Cette régression, dite descendante, est parfois efficace et donne en général un autre système de prédicteurs:

```
/// Elimination de la variable Pvent
          F    0.471 Probabilité critique  0.50145
/// Elimination de la variable Insys
          F    0.670 Probabilité critique  0.42015
/// Elimination de la variable Prcar
          F    0.079 Probabilité critique  0.77594
```

Sur ces données, on obtient le même système de prédicteurs, mais il s'agit d'un cas particulier

— Dans le troisième algorithme, chaque introduction d'une variable explicative dans le système de prédicteur est suivie d'une procédure d'élimination analogue à la précédente. Il faut donc définir deux risques: le premier concerne l'introduction et le second, l'élimination. Le second risque doit être supérieur au premier pour assurer la convergence de l'algorithme. Cet algorithme permet l'introduction forcée de variables explicatives au départ: la régression est effectuée suivant ces variables, on complète le système de prédicteurs par une procédure d'introduction, et, en cas d'introduction, on effectue une procédure d'élimination. On recommence ensuite une procédure d'introduction et ainsi de suite. C'est la régression Stepwise.

Nous avons forcé l'introduction de la fréquence cardiaque et de la pression ventriculaire; les résultats sont les suivants, les risques pour l'introduction et l'élimination étant fixés à 0.05:

```
/// Introduction de la variable Incar
          F  218.352 Probabilité critique  0.00000
/// Introduction de la variable Papul
          F  555.559 Probabilité critique  0.00000
/// Elimination de la variable Pvent
          F    0.068 Probabilité critique  0.79093
/// Elimination de la variable Frcar
          F    0.542 Probabilité critique  0.46974
/// Introduction de la variable Prdia
          F    8.944 Probabilité critique  0.00363
```

L'introduction forcée de variables explicatives, possible aussi en régression ascendante, est parfois indispensable pour faire démarrer les algorithmes ascendants: il peut se produire en effet que tous les coefficients de corrélation soient non significatifs, alors qu'un modèle comportant plus d'une variable donne un coefficient de détermination élevé.

A. Bensaber et B. Bleuse-Trillon (1989) en donnent un exemple sur les données d'Anderson (1971) relatives à la consommation de viande aux Etats-

Unis: l'ajustement d'un polynôme en fonction du temps t donne de bons résultats, alors qu'aucun des coefficients de corrélation de t, t^2, t^3, t^4 et t^5 avec la variable expliquée n'est significatif (matrice ci-dessous). Un algorithme ascendant ne donne donc pas de modèle.

	t	t^2	t^3	t^4	t^5	Cons.
t	1.000					
t^2	0.971	1.000				
t^3	0.921	0.986	1.000			
t^4	0.872	0.959	0.992	1.000		
t^5	0.827	0.929	0.975	0.995	1.000	
Cons	-0.327	-0.228	-0.117	-0.020	0.061	1.000

Consommation de viande aux Etats Unis
Corrélations avec la variable t = temps, t^2, t^3, t^4, t^5

Il est préférable de procéder à une régression descendante: on cherche ici à ajuster un polynôme en t de degré minimum: à partir de l'ajustement du polynôme de degré 5, on teste l'égalité à 0 des coefficients de t^k, pour k variant de 5 à 1, et l'on s'arrête au premier rejet; on obtient ainsi un modèle polynomial de degré 3 dont le coefficient de détermination R^2= 0.6708 est hautement significatif (F=12.906, P[F>12.906]=0.0001).

Les algorithmes pas à pas présentent d'autres inconvénients que le précédent. Ils sont tout d'abord construits sur des critères uniquement numériques, l'utilisateur n'ayant pas d'autre choix que d'en attendre les résultats. La différence entre deux coefficients de corrélation partielle n'est pas prise en compte dans le choix de la variable à introduire ou à éliminer, le rang est à lui seul déterminant.

Le critère numérique est en outre critiquable: le risque de première espèce que l'on choisit est calculé sur la loi de chaque coefficient de corrélation partielle (loi de Snedecor), et non sur la loi du plus grand ou du plus petit d'entre eux; la différence est importante, et est discutée dans Draper et Smith (1981, p.311).

BIBLIOGRAPHIE

Anderson T.W. (1971): *The Statistical Analysis of Time series*, Wiley, New York. (application du modèle linéaire pour l'étude des séries chronologiques; introduction aux polynômes orthogonaux pour le modèle polynomial).

Anderson T.W. (1958): *An Introduction to Multivariate Statistical Analysis*, J. Wiley & Sons, New York (la régression dans le modèle gaussien).

Bensaber A., Bleuse-Trillon B. (1989): *Pratique des chroniques et de la prévision à court terme.* Masson, Paris (plus accessible que l'Anderson, avec des exemples numériques)

Cailliez F. et Pagès J.P. (1976): *Introduction à l'analyse des données*, SMASH, 9 rue Duban, 75016 Paris (présentation de la régression et de l'analyse des données à base de l'algèbre linéaire et de la dualité entre un espace euclidien et son dual).

Draper N.R. et Smith H. (1981): *Applied Regression Analysis*, J. Wiley & Sons, New York. (ouvrage fondamental en régression).

Foucart T. et Lafaye J.Y. (1983): *Régression linéaire sur Micro-ordinateur*, Masson, Paris. (présentation géométrique de la régression accompagnée de programmes).

Foucart T. (1985): *Analyse Factorielle. Programmation sur Micro-ordinateur.* Masson, Paris. (présentation simple de l'analyse des données).

Foucart T. (1991): *Introduction aux tests statistiques. Enseignement assisté par ordinateur.* Technip, Paris (tests du F, de Student avec des exemples et des simulations).

Bibliographie

46

Malinvaud E.: *Méthodes Statistiques de l'Econométrie*, Dunod, Paris, 1964. (ouvrage de base en économétrie.)

Mardia K.V., Kent J.T., Bibby J.M.: *Multivariate Analysis*, Academic Press, Londres 1979. (des remarques intéressantes sur le choix des variables explicatives).

Saporta G. (1990): *Probabilités, analyse des données et statistique*, Technip, Paris (ouvrage général sur la statistique; l'exposé sur la régression est complet et bien fait).

Tomassone R., Lesquoy E. et Millier C. (1983): *La régression. Nouveaux regards sur une ancienne méthode statistique*, Masson, Paris. (destiné aux utilisateurs de la régression; bourré de commentaires judicieux, mais des erreurs dans les exemples, surtout à la fin du livre, peut-être corrigés dans la nouvelle édition).

Weisberg S. (1980): *Applied Linear Regression*, Wiley, New-York. (seul défaut: en anglais).

ANNEXE: DONNEES ETUDIEES
(Saporta, 1991)

	frcar	incar	insys	prdia	papul	pvent	répul	prono
1*	90	1.71	19	16	19.5	16	912	2
2*	90	1.68	18.7	24	31	14	1476	1
3*	120	1.4	11.7	23	29	8	1657	1
4*	82	1.79	21.8	14	17.5	10	782	2
5*	80	1.58	19.7	21	28	18.5	1418	1
6*	80	1.13	14.1	18	23.5	9	1664	1
7*	94	2.04	21.7	23	27	10	1059	2
8*	80	1.19	14.9	16	21	16.5	1412	2
9*	78	2.16	27.7	15	20.5	11.5	759	2
10*	100	2.28	22.8	16	23	4	807	2
11*	90	2.79	31	16	25	8	717	2
12*	86	2.7	31.4	15	23	9.5	681	2
13*	80	2.61	32.6	8	15	1	460	2
14*	61	2.84	47.3	11	17	12	479	2
15*	99	3.12	31.8	15	20	11	513	2
16*	92	2.47	26.8	12	19	11	615	2
17*	96	1.88	19.6	12	19	3	809	2
18*	86	1.7	19.8	10	14	10.5	659	2
19*	125	3.37	26.9	18	28	6	665	2
20*	80	2.01	25	15	20	6	796	2
21*	82	3.15	38.4	13	20	6	508	2
22*	110	1.66	15.1	23	31	6.5	1494	1
23*	80	1.5	18.7	13	17	12	907	1
24*	118	1.03	8.7	19	27	10	2097	1
25*	95	1.89	19.9	25	27	20	1143	1
26*	80	1.45	18.1	19	23	15	1269	1
27*	85	1.3	15.1	13	18	10	1108	1
28*	105	1.84	17.5	18	22	10	957	1
29*	122	2.79	22.9	25	36	10	1032	2
30*	81	1.77	21.9	18	27	11	1220	2
31*	118	2.31	19.6	22	27	10	935	2
32*	87	1.2	13.8	34	41	20	2733	1
33*	65	1.19	18.3	15	18	13	1210	1
34*	84	2.15	25.6	27	37	10	1377	2
35*	103	0.91	8.8	30	33.5	10	2945	1
36*	75	2.54	33.9	24	31	16	976	2
37*	90	2.08	23.1	20	28	6	1077	2
38*	90	1.93	21.4	11	18	10	746	2
39*	90	0.95	10.6	20	24	6	2021	1
40*	65	2.38	36.6	16	22	12	739	2
41*	95	0.99	10.4	20	27.5	8	2222	1
42*	95	0.85	8.9	19	22	15.5	2071	1
43*	86	2.05	23.8	21	28	10	1093	2
44*	82	2.02	24.6	16	22	14	871	2
45*	70	1.44	20.6	19	26.5	11	1472	1
46*	92	3.06	33.3	10	15	6	392	2
47*	94	1.31	13.9	26	40	15	2443	1
48*	79	1.29	16.3	24	31	10	1922	1
49*	67	1.47	21.9	15	18	16	980	2

Données Infarctus du myocarde (début)

	frcar	incar	insys	prdia	papul	pvent	répul	prono
50*	75	1.21	16.1	19	24	4	1587	1
51*	80	2.41	30.9	19	24	7	797	2
52*	61	3.28	54	12	16	7	390	2
53*	110	1.24	11.3	22	27.5	11	1774	1
54*	116	1.85	15.9	33	42	13	1816	1
55*	75	2	26.7	16	22	5	880	2
56*	92	1.97	21.4	18	27	3	1096	1
57*	110	0.96	8.80	15	19	16	1583	2
58*	95	2.56	26.9	8	13	3	406	2
59*	75	2.32	30.9	8	10	6	345	2
60*	80	2.65	33.1	13	19	9	574	2
61*	102	1.60	15.7	24	31	16	1550	1
62*	86	1.67	19.4	18	23	8.5	1102	2
63*	60	0.82	13.7	22	32	13	3122	1
64*	100	1.76	17.6	23	33	2	1500	2
65*	80	3.28	41	12	17	2	415	2
66*	108	2.96	27.4	24	35	6.5	946	2
67*	92	1.37	14.8	25	46	11	2686	1
68*	100	1.38	13.8	20	31	11	1797	1
69*	80	2.85	35.6	25	32	7	898	2
70*	87	2.51	28.8	16	24	20	765	1
71*	100	2.31	23.1	8	12	1	416	2
72*	120	1.18	9.9	25	36	8	2441	1
73*	115	1.83	15.9	25	30	8	1311	1
74*	101	2.55	25.2	23.2	30.5	9	957	2
75*	92	2.17	23.5	19	24	3	885	2
76*	87	1.42	16.1	20	26	10	1465	1
77*	80	1.59	19.9	13	20.5	4	1031	2
78*	88	1.47	16.7	23	32.5	10	1769	1
79*	104	1.23	11.8	27	33	11	2146	1
80*	90	1.45	16.1	17	24	8.5	1324	2
81*	67	0.85	12.7	26	33	11	3106	1
82*	87	2.37	27.2	15	22	10	743	2
83*	108	2.40	22.2	26	31	4	1033	2
84*	120	1.91	15.9	18	27	15	1131	1
85*	108	1.50	13.9	28	43	16	1813	1
86*	86	2.36	27.4	24	34	8	1153	2
87*	112	1.56	13.9	24	29	4	1487	1
88*	80	1.34	17	16	25	16	1493	1
89*	95	1.65	17.4	20	33	7	1600	1
90*	90	2.04	22.7	28	41	10	1608	1
91*	90	3.03	33.6	17	23.5	7	620	2
92*	94	1.21	12.9	17	22	3	1455	1
93*	51	1.34	26.3	11	17	6	1015	1
94*	110	1.17	10.6	29	35	10.5	2393	1
95*	96	1.74	18.1	24	29	6	1333	1
96*	132	1.31	9.9	23	28	12	1710	1
97*	135	0.95	7	15	20	7	1684	1
98*	105	1.92	18.3	18	24	3	1000	1
99*	99	0.83	8.4	23	27	8	2602	1
100*	116	0.60	5.2	33	38	10	5067	1
101*	112	1.54	13.8	25	31	8	1610	1

Données Infarctus du myocarde (fin)

L'ANALYSE DE LA VARIANCE

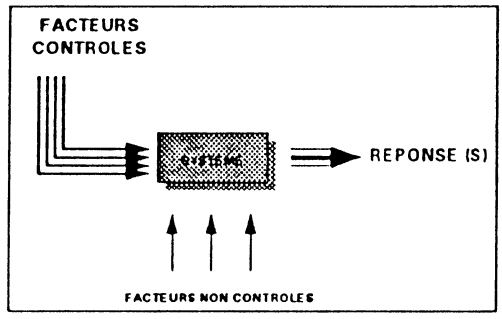
Ph. COURCOUX
ENITIAA
Domaine de la Géraudière
44072 NANTES

PLAN

"L'analyse de la variance consiste à analyser des données qui dépendent de plusieurs types d'effets opérant simultanément, afin de quantifier ces effets, et éventuellement, d'en évaluer l'importance "

M.SCHEFFE.

On utilise l'analyse de variance pour analyser un système expérimental sous la dépendance d'un ou plusieurs facteurs controlés. L'objet de l'analyse est de comparer la variabilité due aux facteurs à la variabilité résiduelle (ou bruit ou erreur expérimentale).



En amont du traitement, on a donc un plan expérimental (plan d'expériences) qui prévoit le nombre d'expériences à faire pour chaque niveau (ou modalité) du ou des facteurs dont on souhaite évaluer l'effet.

Le plan est dit équilibré si il y a même nombre d'observations par niveau de facteur.

I - ANALYSE DE LA VARIANCE A 1 FACTEUR.

I.1 - GENERALITES.

C'est le cas le plus simple rencontré lorsqu'il n'y a qu'un facteur agissant sur les résultats. Le schéma est alors le suivant :

	niveaux du facteur			
	G_1	G_2	.....G_j.....	G_I
observations	x_{11} x_{12} x_{1n1}	x_{21} x_{22} x_{2n2}	....	x_{I1} x_{I2} x_{InI}
Nb d'observations	n_1	n_2	....	n_I
Moyennes	$\bar{x}_1$	$\bar{x}_2$	....	$\bar{x}_I$

La moyenne empirique générale peut être calculée,

soit par : $\bar{x} = \frac{1}{n} \sum_i \sum_j x_{ij}$ avec $n = \sum_i n_i$

ou par : $\bar{x} = \frac{1}{n} \sum_i n_i \bar{x}_i$

Exemple

Considérons une étude portant sur l'apréciation sensorielle de la texture de trois viandes. Il a été prévu une séance de dégustation par viande et 5 dégustateurs différents ont été convoqués à chaque séance. On ne considérera que le caractère fibreux de la viande. Pour ce caractère, les échantillons ont été notés en utilisant une échelle en 15 points.

Comme il n'a pas été possible de réunir les mêmes juges à chacune des séances, l'influence du facteur "Juge" sur les valeurs données ne pourra être étudiée.
Les 15 évaluations seront considérées comme provenant de 15 dégustateurs différents, l'attribution d'un juge à un échantillon se faisant aléatoirement. On parle d'un plan en randomisation totale.

Les valeurs collectées sont les suivantes :

notes pour le caractère fibreux

viande		
A	B	C
3	10	13
5	8	11
6	5	7
3	7	11
3	5	8
$\bar{x}_1 = 4$	$\bar{x}_2 = 7$	$\bar{x}_3 = 10$

$\bar{x} = 7$

I.2 - MODELE.

Le modèle de l'analyse de la variance à un facteur peut s'écrire :

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

où *dans l'exemple des viandes*

X_{ij} est la variable à expliquer ..... *Texture fibreuse d'une viande*
α_i est l'effet du i^{ieme} niveau du facteur..... *caractère fibreux de la viande i*
μ est l'effet moyen général..... *caractère fibreux potentiel*

et ε_{ij} est la variable aléatoire résiduelle .....*due à l'ensemble des causes qui déterminent la note fibreuse d'une viande autre que la nature de la viande (âge de l'animal, condition de conservation, appréciation du juge.....)*

Hypothèses . On suppose que, pour tout couple (i,j) :

Les ε_{ij} sont indépendants
$E(\varepsilon_{ij}) = 0$
$Var(\varepsilon_{ij}) = \sigma^2$
$\varepsilon_{ij} \sim \mathcal{N}(0,\sigma^2)$

I.3 - ESTIMATION DES PARAMETRES DU MODELE (μ,α).

En choisissant de travailler par rapport à l'effet moyen général, c'est-à-dire en posant comme contrainte :

$$\sum_i \alpha_i = 0 \qquad \text{dans le cas d'un plan équilibré}$$

ou $$\sum_i (n_i \alpha_i) = 0 \qquad \text{dans le cas d'un plan non équilibré,}$$

on estime les paramètres μ et α_i (i=1,..,I) du modèle par : $\mu = \bar{x}$ et $a_i = \bar{x}_i - \bar{x}$ pour tout i

I.4 - DECOMPOSITION DE LA VARIABILITE.

Décomposons l'élément x_{ij} en deux termes :

$$x_{ij} = \bar{x}_{.} + (x_{ij} - \bar{x}_i)$$

c'est-à-dire en moyenne de la $i^{ième}$ colonne (parfois écrite $x_{i.}$) plus écart de la valeur individuelle à la moyenne de la colonne.

La moyenne de la colonne i peut de la même manière être décomposée en deux termes :

$$\bar{x}_i = \bar{x} + (\bar{x}_i - \bar{\bar{x}})$$

soit en moyenne globale $\bar{\bar{x}}$ (parfois écrite $x_{..}$) plus la différence de la moyenne de colonne à la moyenne globale

On obtient ainsi :

$$x_{ij} = \bar{\bar{x}} + (\bar{x}_i - \bar{\bar{x}}) + (x_{ij} - \bar{x}_i)$$

Ce développement effectué pour chaque case du tableau de données conduit à l'écriture matricielle suivante :

$$\begin{bmatrix} 3 & 10 & 13 \\ 5 & 8 & 11 \\ 6 & 5 & 7 \\ 3 & 7 & 11 \\ 3 & 5 & 8 \end{bmatrix} = \begin{bmatrix} 7 & 7 & 7 \\ 7 & 7 & 7 \\ 7 & 7 & 7 \\ 7 & 7 & 7 \\ 7 & 7 & 7 \end{bmatrix} + \begin{bmatrix} -3 & 0 & 3 \\ -3 & 0 & 3 \\ -3 & 0 & 3 \\ -3 & 0 & 3 \\ -3 & 0 & 3 \end{bmatrix} + \begin{bmatrix} -1 & 3 & 3 \\ 1 & 1 & 1 \\ 2 & -2 & -3 \\ -1 & 0 & 1 \\ -1 & -2 & -2 \end{bmatrix}$$

$$X = M + B + W$$

Données = Moyennes globales + Ecarts Inter-colonnes + Ecarts Intra-colonne

On a, matriciellement : $(X - M) = B + W$

ou algébriquement : $(x_{ij} - \bar{x}) = (\bar{x}_i - \bar{x}) + (x_{ij} - \bar{x}_i)$ (1)

les termes de cette équation étant les estimations de : $(X_{ij} - \mu) = \overset{V}{\alpha_i} + \overset{V}{\varepsilon_{ij}}$

En élevant au carré et en sommant, pour toutes les observations, les termes de l'équation (1) on a :

$$\sum_i \sum_j (x_{ij} - \bar{x})^2 = \sum_i n_i (\bar{x}_i - \bar{x})^2 + \sum_i \sum_j (x_{ij} - \bar{x}_i)^2$$

Explicitement cette égalité se traduit par :

Somme des carrés des écarts totaux	=	Somme des carrés des écarts inter-niveaux	+	Somme des carrés des écarts intra-niveau

Dans la suite, les termes de cette égalité seront notés respectivement

SCE_{Totale} SCE_{Inter} et SCE_{Intra}.

Vérification sur l'exemple :

$SCE_{Totale} = 140$
$SCE_{Inter} = 90$
$SCE_{Intra} = 50$

I.5 - TEST D'HYPOTHESE.

On souhaite tester les hypothèses :

H_0 : Il n'y a pas d'effet "Produit",
i.e.: les moyennes pour les différents produits (niveaux du facteur) sont égales.
soit $\alpha_1 = \alpha_2 = \ldots\ldots = \alpha_I$

contre H_1 : Il y a un effet dû au produit,
ou deux moyennes au moins sont différentes.

Le principe du test est le suivant :

Si les différences entre produits (entre niveaux du facteur ou entre colonnes du tableau de données) sont grandes par rapport aux écarts intra-produit, alors on concluera qu'il y a un effet différentiel en fonction du produit.

Il s'agit donc de comparer la variablité inter-niveaux à la variabilité intra-niveau du facteur, en faisant intervenir les quantités SCE_{Inter} et SCE_{Intra}.

Reprenons la décomposition :

$$SCE_{Totale} = SCE_{Inter} + SCE_{Intra}$$

équation de l'analyse de la variance

A chacune de ces quantités SCE, on va associer le nombre de degrés de liberté (ddl) adéquat :

source de variation	SCE	ddl
Inter-niveaux	SCE Inter	I - 1
Intra-niveau	SCE Intra	n - I
Totale	SCE Totale	n - 1

On définit le carré moyen inter-groupes par :

$$CM_{Inter} = \frac{SCE_{Inter}}{I-1}$$

De même, on définit le carré moyen intra-groupes par :

$$CM_{Intra} = \frac{SCE_{Intra}}{n-1}$$

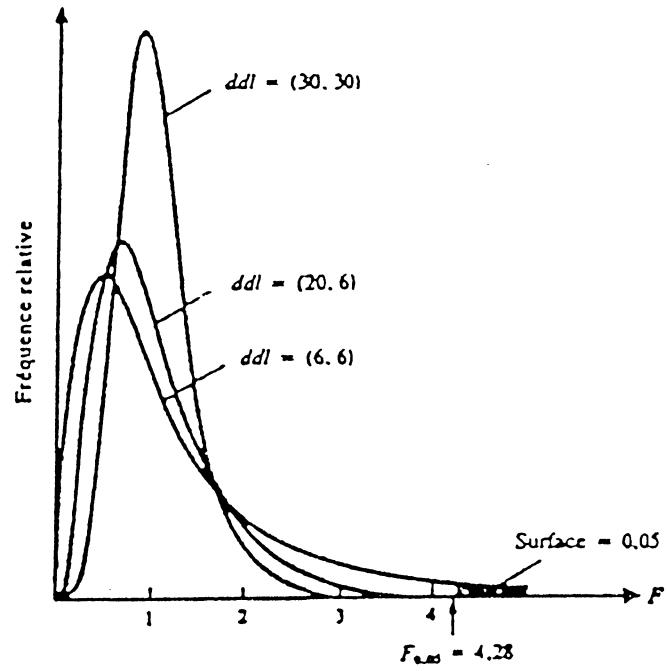
Pour tester H_0 contre H_1, on évalue la quantité :

$$F = \frac{CM_{Inter}}{CM_{Intra}}$$

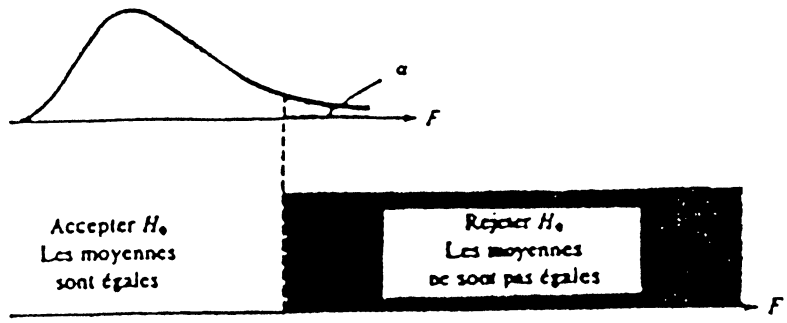
Si l'hypothèse H_0 est vraie , la valeur F est voisine de 1. Plus on s'éloigne de cette hypothèse, en faveur de l'hypothèse H_1, plus le rapport F augmente. Il reste à déterminer à partir de quelle valeur observée de F on rejetera H_0. Cette prise de décision est fondée sur la connaissance de la loi de F sous H_0 .

Dans la mesure où les résidus du modèle d'analyse de la variance suivent une loi normale et si H_0 est vraie, on sait que F est l'observation d'une variable qui suit une loi de Fisher ayant (I-1) ddl au numérateur et (n-I) ddl au dénominateur (notée $F_{(I-1,n-I)}$).

L'allure de la fonction de densité de la distribution de Fisher est la suivante :



Pour un niveau de signification donné α (risque de 1er espèce), on adopte la stratégie de décision suivante:



Exemple

TABLEAU D'ANALYSE DE LA VARIANCE

source de variation	SCE	ddl	CM	F
Type de viande	90	2	45	10,80
Residuelle (Intra-produit)	50	12	4,17	(p<0,01)
Totale	140	14		

EXEMPLE

Les teneurs en tannins de 4 barils de vin californien (Cabernet Sauvignon) provenant de producteurs différents (A, B, C, D) sont comparés. Ces teneurs sont évaluées sur une échelle de 0 à 30 (30 représentant la plus grande teneur).

Les scores sont les suivants :

vin A	vin B	vin C	vin D
8	9	8	1
6	7	5	2
5	6	6	1
7	8	6	0
6	8	7	0
7	7	7	2
7	8	7	0
8	8	5	6
5	7	8	
6	9	59	
7	77		
7			
79			
$n = 12$	$n = 10$	$n = 9$	$n = 7$
$X = 6.58$	$X = 7.7$	$X = 6.56$	$X = 0.86$

$T = 221$
$N = 38$

D'après M. O'MAHONY - *Sensory Evaluation of Food*

Ces observations permettent de construire le tableau d'analyse de la variance suivant ;

Source	SCE	dll	CM	F
Totale	253.71	37		
Entre les vins	219.61	3	73.2	73.2***
Erreur	34.10	34	1.0	

Pour 3 et 34 degrés de liberté, les tables de Fisher donnent :

Prob (F > 2.88) = 0.05
Prob (F > 4.42) = 0.01
Prob (F > 7.05) = 0.001

La valeur F calculée est de 73.2. On en déduit qu'il existe une différence de teneur moyenne en tannin entre barils très significative.

Un test LSD, de comparaisons multiples, au niveau de signification de 0.001, est utilisé afin de répondre à la question : Lequel (ou lesquels) de ces vins diffère(nt) des autres ?

$$LSD = t \sqrt{MS_E \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

	Vin			
	D	C	A	B
Moyenne	0.86			
		6.56	6.58	7.7
n	7	9	12	10

Les vins A, B et C ne sont pas significativement différents les uns des autres quant à leur teneur en tannins, alors que le vin D obtient des scores significativement plus petits que les autres ($\alpha = 0.001$).

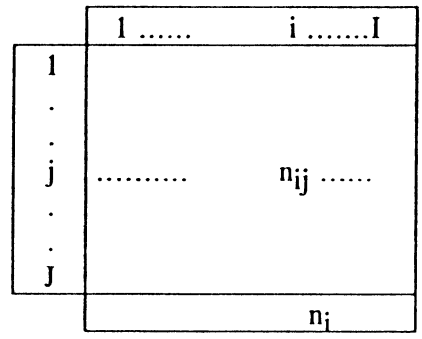
II - ANALYSE DE LA VARIANCE A 2 FACTEURS CROISES.

Dans l'exemple traité en ANOVA à un facteur, portant sur la texture de plusieurs viandes, les séries de mesures en colonne étaient considérées comme étant indépendantes. Ce dispositif complètement randomisé présente un inconvénient majeur: les conditions expérimentales d'évaluation sont différentes d'un type de viande à un autre (juges différents, séances différentes).

Il aurait été préférable de faire intervenir cinq juges, au cours d'une même séance, en leur présentant dans un ordre aléatoire un échantillon de chaque viande. Les séries de mesures sont dans ce cas liées et l'effet "Juge" peut être évalué.

Le plan d'expérience considéré est un dispositif à deux facteurs croisés, complet et sans répétition. Nous considérerons dans un deuxième temps le cas de dispositifs avec répétitions.

D'une manière générale, l'analyse à deux facteurs croisés est mise en oeuvre dans les situations se présentant comme suit :



A et B étant deux facteurs respectivement à I et J niveaux.
n_{ij} étant le nombre d'observations pour le niveau i de A et le niveau j de B

Vocabulaire

* On dit que le dispositif factoriel est croisé lorsque les 2 facteurs jouent le même rôle. Si l'un est subordonné à l'autre, le modèle est dit hiérarchisé.

* Un modèle est dit complet quand il y a au moins une observation par couple (i,j) ($n_{ij} \geq 1$).

* Si le nombre d'observations $n_{ij} = 1$, pour tout couple (i,j), le dispositif est dit sans répétition. si $n_{ij} > 1$, il s'agit d'un dispositif avec répétitions. Seul un plan d'expérience avec répétitions permet de tester l'interaction entre les facteurs.

* Lorsque, pour tout couple (i,j), $n_{ij} = r$, constant, le dispositif est dit équilibré.

II.1 - DISPOSITIF COMPLET SANS REPETITION.

Reprenons la comparaison du caractère fibreux de trois viandes A, B et C et supposons que les mêmes juges aient évalué une série de 3 échantillons . Les valeurs collectées se présenteraient ainsi :

notes pour le caractère fibreux

		viande			
		A	B	C	
	1	3	10	13	$\bar{x}_{.1} = 8.7$
	2	5	8	11	$\bar{x}_{.2} = 8$
juges	3	6	5	7	$\bar{x}_{.3} = 6$
	4	3	7	11	$\bar{x}_{.4} = 7$
	5	3	5	8	$\bar{x}_{.5} = 5.3$
		$\bar{x}_1 = 4$	$\bar{x}_2 = 7$	$\bar{x}_3 = 10$	$\bar{x} = 7$

Si l'effet des facteurs est additif, le modèle est le suivant :

$$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

où *dans l'exemple de la viande*

X est la variable à expliquer .....*Texture fibreuse d'une viande*
α_i est l'effet du i^{ieme} niveau du facteur A..*caractère fibreux de la viande i*
β_j est l'effet du j^{ieme} niveau du facteur B...*effet lié au juge j*
μ est l'effet moyen général.....*caractère fibreux potentiel*

et ε_{ij} est la variable aléatoire résiduelle .....*dûe à l'ensemble des causes qui déterminent la note fibreuse d'une viande autre que la nature de la viande et l' appréciation du juge.*

Avec les hypothèses :

Pour tout couple (i,j), ε_{ij} est distribuée selon une loi normale de moyenne nulle et de variance σ^2

Les variables aléatoires ε_{ij} sont indépendantes 2 à 2.

L'équation de décomposition de la variance s'écrit dans ce cas :

$$\sum_i \sum_j (x_{ij} - \bar{x}_{..})^2 = \sum_i \sum_j (\bar{x}_{i.} - \bar{x}_{..})^2 + \sum_i \sum_j (\bar{x}_{.j} - \bar{x}_{..})^2 + \sum_i \sum_j (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..})^2$$

ou encore

$$\sum_i \sum_j (x_{ij} - \bar{x}_{..})^2 = J \sum_i (\bar{x}_{i.} - \bar{x}_{..})^2 + I \sum_j (\bar{x}_{.j} - \bar{x}_{..})^2 + \sum_i \sum_j (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..})^2$$

soit

$$SCE_T = SCE_A + SCE_B + SCE_R$$

T comme Totale — R comme Résiduelle
A comme liée au facteur A — B comme liée au facteur B

D'où le tableau d'analyse de la variance à 2 facteurs croisés sans répétition :

source de variation	SCE	ddl
Effet principal de A	$SCE_A = J\sum_i (\bar{x}_{i.} - \bar{x}_{..})^2$	I - 1
Effet principal de B	$SCE_B = I\sum_j (\bar{x}_{.j} - \bar{x}_{..})^2$	J - 1
Résiduelle	$SCE_R = SCE_T - (SCE_A + SCE_B)$	(I-1)(J-1)
Totale	$SCE_T = \sum_i \sum_j (\bar{x}_{ij} - \bar{x}_{..})^2$	n - 1

Exemple :

source de variation	SCE	ddl	CM	F
Type de viande	90.00	2	45.00	13.21
Effet Juge	22.73	4	5.68	1.66
Residuelle	27.27	8	3.41	
Totale	140	14		

1er Test H_0 : Absence d'effet du facteur A, c'est-à-dire $\alpha_1 = \alpha_2 = ... = \alpha_I = 0$

contre H_1 : Au moins 2 termes α_i diffèrent

La statistique du test $F = \dfrac{SCE_A / I-1}{SCE_R / (I-1)(J-1)}$ est distribuée, sous H_0, comme une loi de Fisher à (I-1) et (I-1)(J-1) degrés de liberté.

Exemple : $F_{obs} = 13.21 \quad >> \quad F_{\alpha,2,8} = 4.46$ avec $\alpha = 5\%$.

On rejette donc H_0 avec un risque de 5% , c'est-à-dire on conclue à l'existence d'un effet "Produit"

2ièmeTest H_0 : Absence d'effet du facteur B, c'est-à-dire $\beta_1 = \beta_2 = ... = \beta_J = 0$

contre H_1 : Au moins 2 termes β_j diffèrent

La statistique du test $F = \dfrac{SCE_B / J-1}{SCE_R / (I-1)(J-1)}$ est distribuée, sous H_0, comme une loi de Fisher à (J-1) et (I-1)(J-1) degrés de liberté.

Exemple : $F_{obs} = 1.66 < F_{\alpha,4,8} = 3.84$ avec $\alpha = 5\%$.

On ne rejette donc pas H_0, c'est-à-dire qu'aucun effet "Juge " n'a pu être décelé.

EXEMPLE

On considère les données collectées par I. LEGUERINEL pour la caractérisation sensorielle de différents cidres.

La question ici est de savoir, d'une part, si les 10 cidres concernés par l'étude sont significativement différents en ce qui concerne leur caractéristiques sensorielles ou au contraire comparables, et d'autre part si les juges sont homogènes lorsqu'ils notent un même produit. Il s'agit donc de tester l'existence de deux effets : l'effet *produit* et l'effet *juge*.

L'analyse de la variance étant une méthode unidimensionnelle, ce test ne pourra être conduit que pour chacune des variables, prises séparément.

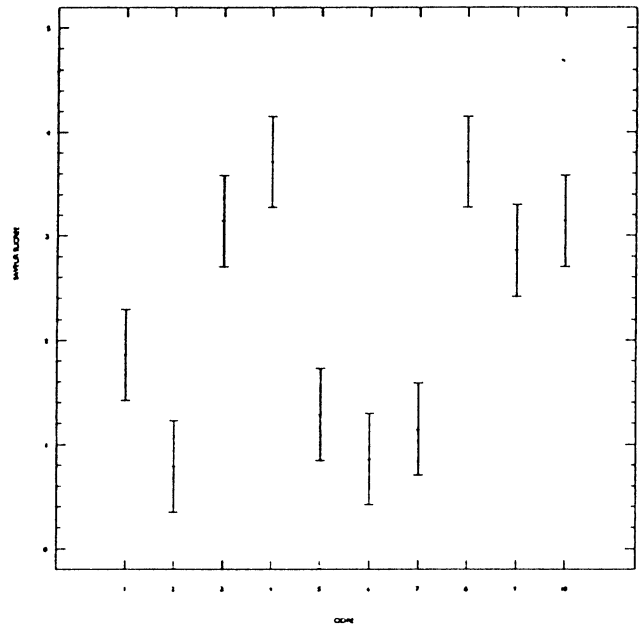
Considérons comme variable à expliquer la :

SAVEUR SUCREE

Tableau d'analyse de la variance. Modèle additif à deux facteurs (juge, produit)

Source de variation	SCE	ddl	CM	Fobs	Niv. signif.
Liée aux cidres	88.52	9	9.83	26.70	0.00
Liée aux juges	1.85	6	0.31	0.90	0.50
Résiduelle	18.51	54	0.34		
Totale	108.88	69			

Intervalles de confiance (LSD) à 95% autour des moyennes par niveau du facteur

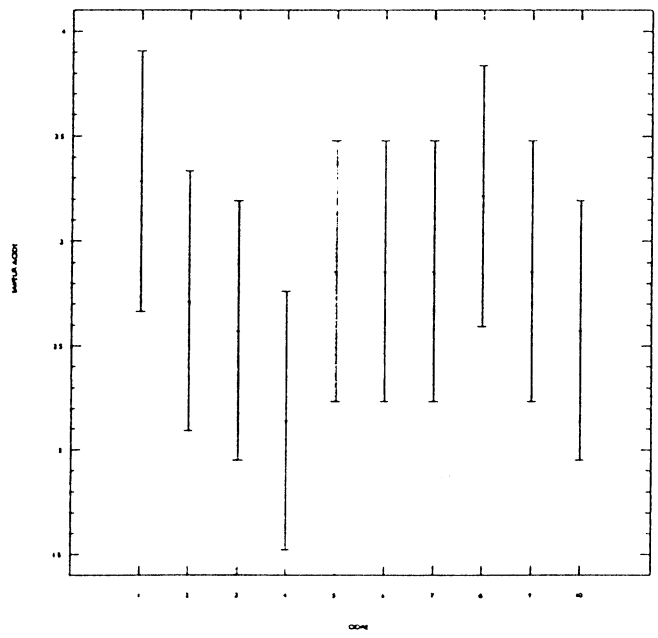


SAVEUR ACIDE

Tableau d'analyse de la variance. Modèle additif à deux facteurs (juge, produit)

Source de variation	SCE	ddl	CM	Fobs	Niv. signif.
Liée aux cidres	6.75	9	0.75	1.77	0.10
Liée aux juges	17.62	6	2.94	6.93	0.00
Résiduelle	22.88	54	0.42		
Totale	47.25	69			

Intervalles de confiance (LSD) à 95% autour des moyennes par niveau du facteur.



Remarque :

Envisageons le cas où, pour savoir si les cidres ont des saveurs acide différentes, on ait omis le facteur "juge".

Le modèle est alors un modèle d'analyse de la variance à un facteur (les cidres). Le tableau de cette analyse est le suivant :

Source de variation	**SCE**	**ddl**	**CM**	**Fobs**	**Niv. signif.**
"Cidre" (inter)	6.75	9	0.75	1.11	0.37
Résiduelle (intra)	40.50	60	0.68		
Totale	47.25	69			

II.2 - DISPOSITIF COMPLET AVEC REPETITIONS.

Considérons un plan d'expérience équilibré, avec plusieurs répétitions par croisement de deux facteurs A et B.

Ces répétitions rendent possible une évaluation de l'interaction des deux facteurs.

Plan à 2 facteurs croisés SANS répétition

	A1	A2	A3
B1	x	x	x
B2	x	x	x
B3	x	x	x
B4	x	x	x
B5	x	x	x

Plan à 2 facteurs croisés équilibré AVEC répétitions

	A1	A2	A3
B1	xxx	xxx	xxx
B2	xxx	xxx	xxx
B3	xxx	xxx	xxx
B4	xxx	xxx	xxx
B5	xxx	xxx	xxx

Le tableau de données se présente alors sous la forme suivante :

	1		i		I	moy.
1						$\bar{x}_{.1.}$
	x_{1j1} x_{1j2} x_{1jk} $\bar{x}_{1j.}$		x_{ij1} x_{ij2} x_{ijk} $\bar{x}_{ij.}$		x_{Ij1} x_{Ij2} x_{Ijk} $\bar{x}_{Ij.}$	$\bar{x}_{.j}$
J						$\bar{x}_{.J}$
moy.	$\bar{x}_{1..}$		$\bar{x}_{i..}$		$\bar{x}_{I..}$	$\bar{x}_{...}$

Le modèle d'analyse de la variance associé est :

$$X_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

où
- μ ..... est l'effet moyen général
- α_i $(i=1,\ldots,I)$ est l'effet principal du facteur A,
- β_j $(j=1,\ldots,J)$ est l'effet principal du facteur B,
- $(\alpha\beta)_{ij}$ ..... est l'interaction des niveaux i et j des 2 facteurs,
- ε_{ijk} ..... est la variable aléatoire résiduelle attachée à la k^{ieme} $(k=1,\ldots,K)$ observation pour la combinaison i,j des facteurs.

On suppose que

les variables ε_{ijk} sont indépendantes,

et pour tout (i,j,k) ε_{ijk} suit une loi normale de moyenne nulle et de variance σ^2.

En procédant comme pour l'analyse de la variance à un critère de classification, l'équation de l'analyse de la variance peut être obtenue en élevant au carré les deux membres du modèle, et en sommant pour toutes les valeurs observées :

$$\sum_i \sum_j \sum_k (x_{ijk} - \bar{x}_{...})^2 = JK \sum_i (\bar{x}_{i..} - \bar{x}_{...})^2 + IK \sum_j (\bar{x}_{.j.} - \bar{x}_{...})^2 + K \sum_i \sum_j (x_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x}_{...})^2$$
$$+ \sum_i \sum_j \sum_k (x_{ijk} - \bar{x}_{ij.})^2$$

La somme totale des carrés des écarts (SCE_T) se divise donc ici en quatre composantes que l'on peut désigner respectivement par :

SCE_A liée au premier facteur (par exemple le produit),
SCE_B liée au second facteur (par exemple le juge),
SCE_{AB} liée à l'interaction des deux facteurs et
SCE_R qui est une somme résiduelle.

$$SCE_T = SCE_A + SCE_B + SCE_{AB} + SCE_R$$

A ces différentes sommes des carrés des écarts sont attachés des nombres de degrés de liberté, qui vérifient la relation :

$$IJK-1 = (I-1) + (J-1) + (I-1)(J-1) + (IJK-IJ)$$

où $IJK = n$ le nombre total d'observations.

Le tableau de l'analyse de la variance peut enfin être dressé :

source de variation	SCE	ddl
Effet principal de A	$SCE_A = JK\sum_i (\bar{x}_{i..} - \bar{x}_{...})^2$	I - 1
Effet principal de B	$SCE_B = IK\sum_j (\bar{x}_{.j.} - \bar{x}_{...})^2$	J - 1
Interaction A X B	SCE_{AB} (par différence)	(I-1)(J-1)
Résiduelle	$SCE_R = \sum_i \sum_j \sum_k (\bar{x}_{ijk} - \bar{x}_{ij.})^2$	IJK - IJ
Totale	$SCE_T = \sum_i \sum_j \sum_k (\bar{x}_{ijk} - \bar{x}_{...})^2$	n - 1

La stratégie des tests d'hypothèses adoptée sera la suivante :

a- *Test de l'absence d'interaction* Ho : $(\alpha\beta)_{ij}=0$ pour tout i et j

La statistique du test $F = \dfrac{SCE_{AB}/(I-1)(J-1)}{SCE_R/(IJK-IJ)}$ est distribuée, sous H_0, comme une loi de Fisher à (I-1)(J-1) et (IJK - IJ) degrés de liberté.

b- *Test d'absence d'effet du facteur A* $H_0 : \alpha_i=0$ pour tout i

La statistique du test $F \quad \dfrac{SCE_A/(I \quad 1)}{SCE_R/(IJK \quad IJ)}$ est distribuée, sous H_0, comme une loi de Fisher à (I-1) et (IJK - IJ) degrés de liberté.

c - *Test d'absence d'effet du facteur B* $H_0 : \beta_j=0$ pour tout j

La statistique du test $F = \dfrac{SCE_B/(J-1)}{SCE_R/(IJK-IJ)}$ est distribuée, sous H_0, comme une loi de Fisher à (J-1) et (IJK - IJ) degrés de liberté.

EXEMPLE

La tenue à la fonte de 4 marques de savon est étudiée à l'aide d'un testeur mécanique. Deux techniciens ont réalisé 3 mesures de suite pour chaque marque.

Les résultats, exprimés en pourcentage, et les totaux marginaux sont fournis dans le tableau suivant :

	Marques de savon				
Techniciens	A	B	C	D	Totaux
1	60.6	73.9	66.1	76.7	822.5
	59.0	74.2	64.4	75.6	
	59.2	73.8	65.0	74.0	
2	59.9	73.9	63.5	73.8	817.9
	61.6	74.0	64.9	73.0	
	62.0	72.3	63.0	76.0	
Totaux	362.3	442.1	386.9	449.1	1640.4
Moyennes	60.4	73.7	64.5	74.9	
Sous-totaux pour les marques suivant le technicien					
1	178.8	221.9	195.5	226.3	
2	183.5	220.2	191.4	222.8	

D'après M. GACULA, J. C. JAGBIR
Statistical Methods in Food and Consumer Research

Les résultats de l'analyse de la variance sont les suivants :

Source	DF	SS	MS	*F* ratio
Total	23	921.54		
Savon	3	894.68	298.23	160.33
Technicien	1	0.88	0.88	0.47
Savon x technicien	3	8.12	2.71	1.46
Erreur	16	17.86	1.86	

Pour un modèle à effets fixes, l'hypothèse nulle selon laquelle les différentes marques sont équivalentes est testée en formant le rapport : "MS Savons / MS erreur". La valeur F calculée conduit au rejet de cette hypothèse nulle.

III - TESTS DE COMPARAISONS MULTIPLES

Utilisation :

Tests paramétriques qui interviennent après un test de comparaison globale dans le cas où l'hypothèse nulle a été rejetée; c'est-à-dire lorsqu'on a conclu à une différence significative entre modalités d'un facteur, au risque de première espèce α

Différents tests :

Il existe de nombreux tests de comparaisons multiples. Sans être exhaustif on peut citer :

* le test de SCHEFFE
* le test HSD(1) de TUKEY ("Studentized Range Test")
* le test de NEWMAN-KEULS
* le test de DUNCAN
* le test LSD(2) de FISHER
* le test de DUNNETT

Ces tests sont ordonnés ici par ordre de puissance croissante, ou selon un autre critère, du plus conservatif(3) au moins conservatif.

On sait que le problème de la comparaison de deux moyennes est résolu par le test "t" de Student: une méthode peut donc consister à comparer les moyennes deux à deux à l'aide de ce test. Mais cela nécessite, pour p moyennes, C_p^2 comparaisons.

Un certain nombre de méthodes permettent d'effectuer ces C_p^2 comparaisons comme par exemple la méthode de la plus petite différence significative (LSD de FISHER), peu correcte mais largement utilisée. Les tests de DUNCAN et de NEWMAN-KEULS sont "similaires au test LSD" mais n'en présentent pas les inconvénients" (GOUET J.P., 1974). Ces inconvénients sont de deux ordres. D'une part, il n'y a pas effectivement indépendance des échantillons comparés, puisque le résultat d'une comparaison est lié aux comparaisons entre moyennes intermédiaires. D'autre part,si on compare p moyennes 2 à 2 au niveau nominal α de 5%, le seuil au niveau global n'est plus nécessairement de 5%, mais augmente avec le nombre de moyennes à comparer.

Dans certains cas, il se peut que quelques comparaisons de moyennes soient intéressantes et d'autres inutiles. Ces situations justifient l'utilisation de tests plus spécifiques comme le test de DUNNETT.

Principe général :

On calcule pour chaque comparaison une valeur seuil r (ou range), fonction, entre autres, du carré moyen des erreurs de l'analyse de la variance. Si la différence entre les moyennes comparées est inférieur à ce seuil r, la différence est déclarée non significative. Traditionnellement, on prend le même niveau de signification α pour les tests de comparaisons multiples que celui pris pour l'ANOVA.

Les tests les plus puissants auront un "r" petit, les plus conservatifs un "r" élevé.

Certains tests de comparaisons multiples font intervenir des seuils r différents suivant le nombres de moyennes situées entre les deux moyennes à comparer. Parmi ces tests dits à plages multiples on trouve les tests de DUNCAN et de NEWMAN-KEULS.

---

(1) HSD : Honesty Significant Difference.
(2) LSD : Least Significant Difference.
(3) "conservative" en anglais : propriété à déceler une différence au risque que celle-ci n'existe pas.

Les différentes expressions de calcul des seuils r sont les suivantes :

LSD	$t\sqrt{2\,CM_E / n}$	t -> table de Student n : nb d'obs.par échantillon CM_E : carré moyen résiduel
SCHEFFE	$S\sqrt{2\,CM_E / n}$	S -> $\sqrt{(p-1)F}$ et F -> table de Fisher
DUNNETT (*)	$D\sqrt{2\,CM_E / n}$	D -> table de Dunnett
NEWMAN-KEULS (*)	$Q\sqrt{2\,CM_E / n}$	Q -> table du "Range Studentized"
DUNCAN (*)	$Q_D\sqrt{2\,CM_E / n}$	Q_D-> table du "Range Studentized" adaptée
HSD de TUKEY (*)	$Q_{max}\sqrt{2\,CM_E / n}$	Q_{max} -> table du "Range Studentized"

(*)test ne s'appliquant que dans le cas de dispositifs équilibrés.

<u>Choix d'un test de comparaisons multiples</u> :

Le choix d'un test n'est pas forcément aisé. Le tableau ci-dessous fournit quelques éléments pour faire ce choix

DUNNETT	pour comparer toutes les moyennes à l'une d'elles. comparaison à un contrôle
SHEFFE	comparaisons simples ou complexes.
HSD de TUKEY	risque α final maintenu au niveau du risque fixé a priori
NEWMAN-KEULS	comparaisons simples. "range" variables. ajustement des niveaux
DUNCAN	nominaux pour un seuil global α choisi.
LSD	comparaisons simples, dans le cas où il y a très peu de comparaisons à faire

Régression, aspects déterministes

Fredon Daniel
IREM de Limoges

L'atelier a pour but d'étudier quelques aspects de la régression simple dans sa première approche déterministe.

I. Choix des variables

La recherche d'une droite de régression $y = a + bx$ suppose a priori que les caractères statistiques X et Y ont des statuts différents. Y est la variable à expliquer et X la variable potentiellement explicative (ou susceptible d'expliquer Y). Il peut aussi arriver que X soit plus facile à mesurer et que la recherche d'un modèle ait pour but d'obtenir Y par le calcul.
Le calcul simultané des droites de régression de Y par rapport à X et de X par rapport à Y n'a donc guère de sens ni sur le plan des applications concrètes, ni sur le plan des aspects plus développés de la régression.

II. Ajustement affine par la méthode des moindres carrés : décomposition de la variance

Après obtention de la droite de régression de Y par rapport à X, cherchons à décomposer $V(Y)$.

$$V(Y) = \frac{1}{n}\sum_{i=1}^{k} n_i\,(y_i - \overline{y})^2$$

$$= \frac{1}{n}\sum_{i=1}^{k} n_i\,[(y_i - a - bx_i) + (bx_i - b\overline{x})]^2 \qquad \text{car} \qquad \overline{y} = a + b\overline{x}$$

On a : $\dfrac{1}{n}\sum_{i=1}^{k} n_i\,(bx_i - b\overline{x})^2 = \dfrac{b^2}{n}\sum_{i=1}^{k} n_i\,(x_i - \overline{x})^2 = b^2 V(X) = V(a + bX)$

et $\dfrac{2}{n}\sum_{i=1}^{k} n_i(y_i - a - bx_i)\,(bx_i - b\overline{x}) = \dfrac{2b}{n}\sum_{i=1}^{k} n_i\,[y_i - \overline{y} + b(\overline{x} - x_i)]\,[x_i - \overline{x}]$

$$= 2b\,[Cov(X,Y) - bV(X)] = 0.$$

Pour les valeurs de a et de b correspondant à la droite de régression, on a donc :

$$V(Y) = V(a + bX) + \frac{1}{n}\sum_{i=1}^{k} n_i(y_i - a - bx_i)^2$$

égalité que l'on interprète par :
variance de Y = variance expliquée (par l'ajustement affine) + variance résiduelle

On constate donc que :

$$\frac{\text{variance expliquée}}{\text{variance totale}} = \frac{V(a+bX)}{V(Y)} = b^2\frac{V(X)}{V(Y)}$$
$$= \frac{[Cov(X,Y)]^2}{V(X)\,V(Y)} \quad \text{car} \quad b = \frac{Cov(X,Y)}{V(X)}$$
$$= r^2$$

r^2 apparaît donc comme mesure de la qualité de l'ajustement affine.

III. Ajustement linéaire

Le modèle affine, même après avoir appliqué des transformations aux variables, n'est pas le seul modèle possible. Dans le but de donner un autre exemple, considérons le modèle linéaire $Y = aX$ et conservons le choix classique de mesure de la distance entre les points expérimentaux et une courbe de la famille par la somme des carrés des écarts verticaux. Cette variante peut faire l'objet d'un problème en STS. En voici un énoncé possible.

Soit $(x_1; y_1), \ldots, (x_n; y_n)$ une série statistique à deux dimensions. On se propose d'ajuster sur ces données une relation linéaire $y = ax$.

1. Déterminez a tel que $S(a) = \sum_{i=1}^{n}(y_i - ax_i)^2$ soit minimum.

2. a étant égal à la valeur obtenue dans la première question, vérifiez qu'après ajustement on a :

$$\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right) = \left(\sum_{i=1}^{n} x_i y_i\right)^2 + \left(\sum_{i=1}^{n}(y_i - ax_i)^2\right)\left(\sum_{i=1}^{n} x_i^2\right)$$

3. Pour mesurer la qualité de l'ajustement linéaire, on calcule :

$$d = \frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2}{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right)}$$

Vérifiez que $0 \leq d \leq 1$ et que $(d = 1) \Longleftrightarrow$ (tous les points sont alignés).

4. Les rayons γ pur émis par une substance radioactive sont en partie absorbés par les écrans de plomb. Si N désigne le nombre d'atomes radioactifs exprimé dans une unité qui correspond à la mesure au compteur Geiger, on a la loi théorique : $N = N_0 e^{-ax}$ où N_0 correspond à l'absence d'écran, x désigne l'épaisseur des écrans et a une constante.

On a obtenu les mesures ci-dessous où n est le nombre d'écrans de 2 mm d'épaisseur.

n	0	1	2	3	4	5	6	7
N	8 623	7 333	6 273	5 347	4 679	4 031	3 384	2 956

Déterminez la valeur de la constante a en m^{-1} en utilisant un ajustement linéaire. Calculez la valeur du coefficient d.

Pour une solution, voir page 32
F.Couty, J.Debord, D.Fredon
Probabilités et statistiques pour les biologistes
collection Flash U A.Colin

IV. Régression orthogonale

Si on choisit le modèle affine alors que Y et X jouent des rôles symétriques, au lieu de retenir comme mesure de la distance entre une droite et les points expérimentaux la somme des carrés des écarts verticaux, il est plus logique de considérer la somme des carrés des distances des points à la droite. C'est la régression orthogonale.
Comme la droite obtenue est le premier axe factoriel de l'analyse en composantes principales des données, ce thème peut aussi servir d'introduction à l'analyse des données.
Voici un énoncé possible sous forme de problème; seules les deux premières questions sont faciles pour un élève de STS.

Soit (X,Y) une série double représentée dans le plan rapporté à un repère $(O;\vec{i};\vec{j})$ par les points $M_k(x_k,y_k)$, $k \in \{1,2,\ldots,n\}$. On considère le point moyen $G(\bar{x},\bar{y})$ et le changement de variables : $x' = x - \bar{x}$, $y' = y - \bar{y}$.
Etant donné une droite Δ dont une équation dans le repère $(G;\vec{i};\vec{j})$ est :

$$x' \cos\alpha + y' \sin\alpha - \rho = 0,$$

on note H_k la projection orthogonale de M_k sur Δ.
On se propose de déterminer Δ de manière à minimiser $S = \sum_{k=1}^{n} H_k M_k{}^2$.

1. Montrez que $S = \sum_{k=1}^{n} (x'_k \cos\alpha + y'_k \sin\alpha - \rho)^2$.

2. Montrez que, s'il existe une droite Δ qui minimise S, cette droite passe par G.

3. Déterminez le coefficient directeur de Δ qui rend S minimum. Précisez dans quel cas la solution est unique. La droite Δ ajustée est appelée *droite de régression orthogonale.*

4. On suppose que la droite Δ de régression orthogonale est unique. Soit D et D' les droites de régression de y par rapport à x et de x par rapport à y. Etudiez la position relative des droites D, D' et Δ.

5. Soit $T = \sum_{k=1}^{n} GM_k^2$ et $S' = \sum_{k=1}^{n} GH_k^2$. Montrez que le critère de l'ajustement revient à maximiser S'.

On pose $q = \frac{S'}{T}$. Montrez que q est un indice de qualité de l'ajustement compris entre $\frac{1}{2}$ et 1.

Quelle est la signification des valeurs $q = 1$ et $q = \frac{1}{2}$?

6. On donne les notes d'un groupe de 8 étudiants dans deux disciplines : mathématiques (note x sur 20) et informatique (note y sur 20).

x	12	15	8	6	10	12	11	9
y	10	16	11	4	9	14	17	8

Ajustez sur ces données la droite Δ de régression orthogonale et calculez le coefficient q de qualité de l'ajustement. Interprétez l'ajustement de Δ.

Pour une solution, voir page 73

C.Raffin
Statistiques et probabilités DEUG A 2ème année
collection Flash U A.Colin

Daniel FREDON
IREM de LIMOGES

Tests non paramétriques

Il s'agit de présenter quelques tests non paramétriques comme réponses à des problèmes non réglés par les tests paramétriques quand leurs hypothèses de fonctionnement ne sont pas satisfaites.

I. Test de Mann et Whitney

• Utilisation

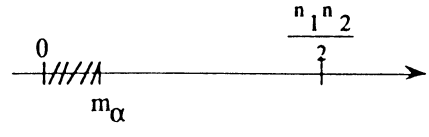
Il permet de comparer les moyennes de deux échantillons indépendants dans le cas où l'on ne sait rien des populations et où les tailles des échantillons sont insuffisantes pour appliquer un test de Student.
On dispose de deux échantillons. indépendants et non-exhaustifs, E_1 et E_2, de tailles respectives n_1 et n_2. On veut comparer les moyennes expérimentales, c'est-à-dire tester l'hypothèse nulle $(H_0) : \mu_1 = \mu_2$.

• Mise en place du test

On classe par *ordre croissant* l'ensemble des valeurs des deux échantillons. Pour distinguer les valeurs qui proviennent de E_1 et de E_2, on peut utiliser divers moyens; il est commode d'utiliser deux couleurs si on le peut.
Pour tout élément x_i de E_1, on compte le nombre d'éléments de E_2 situés après x_i (en comptant pour 0,5 tout élément de E_2 ex-æquo avec x_i).
On note u_1 la somme de toute les valeurs ainsi associées à tous les éléments de E_1.
On définit de même u_2 en permutant les rôles de E_1 et de E_2. Puis on pose $u = \min(u_1, u_2)$.
On vérifie que $u_1 + u_2 = n_1 n_2$.

• Règle de décision

Soit $\mathcal{U}$ la variable aléatoire qui prend la valeur u à l'issue de l'expérience aléatoire.
Les tables jointes donnent, en fonction de n_1, n_2 et α la valeur m_α telle que, sous (H_0), $P(\mathcal{U} \leq m_\alpha) = \alpha$ dans les cas $\alpha = 0{,}05$ et $\alpha = 0{,}01$. On rejette l'hypothèse nulle si $u \leq m_\alpha$.



II. Test de Wilcoxon

• Utilisation

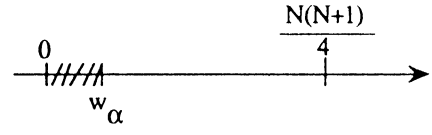
Il permet de comparer les moyennes de deux échantillons appariés dans le cas où l'on ne sait rien des populations et où la taille des échantillons n'est pas suffisante pour appliquer un test de Student.
On veut comparer les moyennes expérimentales, c'est-à-dire tester l'hypothèse nulle $(H_0) : \mu_1 = \mu_2$.

• Mise en place du test

On calcule les différences entre les valeurs appariées. On supprime les différences nulles et on note N le nombre de différences non nulles.
On classe ces différences par ordre croissant des valeurs absolues.
On affecte à chaque différence son rang dans ce classement. s'il y a des ex-æquo, on attribue à chacun un rang égal à la moyenne des rangs qu'ils occupent.
On calcule w_+ la somme des rangs des différences positives et w_- somme des rangs des différences négatives. On vérifie que $w_+ + w_- = \dfrac{N\,(N+1)}{2}$.
On note $w = \min\,(w_+, w_-)$.

• Règle de décision

Soit W la variable aléatoire qui prend la valeur w à l'issue de l'expérience aléatoire.
Si $N \leq 25$, la table jointe donne, en fonction de N, et α la valeur w_α telle que, sous (H_0), $P(W \leq w_\alpha) = \alpha$ dans les cas $\alpha = 0{,}05$ et $\alpha = 0{,}01$. On rejette l'hypothèse nulle si $w \leq w_\alpha$.



Daniel FREDON
IREM de LIMOGES

bibliographie :

F.Couty, J.Debord, D.Fredon
Probabilités et statistiques pour les biologistes (chap. 17)
collection Flash U A.Colin

TABLE 7

Test de Mann et Whitney ($\alpha = 0{,}05$)

La table donne la valeur m_α telle que $P(U \leqslant m_\alpha) = \alpha = 0{,}05$ pour deux échantillons d'effectifs n_1 et n_2 avec $n_1 \leqslant n_2$.

n_1 \ n_2	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	–	–	–	–	0	0	0	0	1	1	1	1	1	2	2	2	2
3	–	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	14
5		2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6			5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7				8	10	12	14	16	18	20	22	24	26	28	30	32	34
8					13	15	17	19	22	24	26	29	31	34	36	38	41
9						17	20	23	26	28	31	34	37	39	42	45	48
10							23	26	29	33	36	39	42	45	48	52	55
11								30	33	37	40	44	47	51	55	58	62
12									37	41	45	49	53	57	61	65	69
13										45	50	54	59	63	67	72	76
14											55	59	64	69	74	78	83
15												64	70	75	80	85	90
16													75	81	86	92	98
17														87	93	99	105
18															99	106	112
19																113	119
20																	127

TABLE 8

Test de Mann et Whitney ($\alpha = 0{,}01$)

La table donne la valeur m_α telle que $P(U \leqslant m_\alpha) = \alpha = 0{,}01$ pour deux échantillons d'effectifs n_1 et n_2 avec $n_1 \leqslant n_2$.

n_1 \ n_2	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0	0
3	–	–	–	–	–	0	0	0	1	1	1	2	2	2	2	3	3
4	–	–	0	0	1	1	2	2	3	3	4	5	5	6	6	7	8
5		0	1	1	2	3	4	5	6	7	7	8	9	10	11	12	13
6			2	3	4	5	6	7	9	10	11	12	13	15	16	17	18
7				4	6	7	9	10	12	13	15	16	18	19	21	22	24
8					7	9	11	13	15	17	18	20	22	24	26	28	30
9						11	13	16	18	20	22	24	27	29	31	33	36
10							16	18	21	24	26	29	31	34	37	39	42
11								21	24	27	30	33	36	39	42	45	48
12									27	31	34	37	41	44	47	51	54
13										34	38	42	45	49	53	57	60
14											42	46	50	54	58	63	67
15												51	55	60	64	68	73
16													60	65	70	74	79
17														70	75	81	86
18															81	87	92
19																93	99
20																	105

TABLE 9

Test de Wilcoxon

La table donne la valeur w_α telle que $P(W \leqslant w_\alpha) = \alpha$, dans les cas $\alpha = 0{,}05$ et $\alpha = 0{,}01$.

α \ N	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
0,05	2	4	6	8	11	14	17	21	25	30	35	40	46	52	59	66	73	81	89
0,01	–	0	2	3	5	7	10	13	16	20	23	28	32	38	43	49	55	61	68

Extrait de F.Couty, J.Debord, D.Fredon
Probabilités et statistiques pour les biologistes
Collection Flash U, A.Colin.

DIDACTICIEL DES TECHNIQUES DE LA STATISTIQUE
- Module Ajustement linéaire -

Michel JANVIER
E.R.E.S.
Université Montpellier II
Sciences et Techniques du Languedoc
Place Eugène Bataillon
34095 MONTPELLIER CEDEX 5
France

Eric VERDOIRE
Conservatoire National des Arts et Métiers
(C.N.A.M.)
Centre Régional du Languedoc-Roussillon
11, avenue Professeur Grasset
34000 MONTPELLIER
France

Présentation

Ce didacticiel fait partie d'un ensemble qui couvrira tout le programme de la demi-valeur du premier cycle du CNAM.

Le public visé comprend :

* en formation professionnelle, tous les auditeurs CNAM,
* en formation pemanente, toutes les personnes utilisant les statistiques dans leur activité professionnelle,
* en formation initiale, de nombreuses catégories d'étudiants dont le cursus comporte une initiation à la statistique : sciences économiques, médecine, pharmacie, biologie, technologie, gestion, écoles de commerce, lycées, classes de S.T.S. .

Contextes d'utilisation

Le didacticiel est susceptible d'être utilisé :

* en remplacement de l'enseignement traditionnel (utilisation à distance),
* comme soutien ou approfondissement des connaissances (utilisation en auto-formation à domicile ou en centre ressource),
* pour illustrer un enseignement présentiel en exploitant les possibilités de simulation ou de traitements des données disponibles dans le didacticiel ainsi que les nombreux exemples et fonctions de données inclus dans le didacticiel (utilisation en salle de cours).

Stratégie didactique

La stratégie didactique choisie tend à faciliter l'appropriation des concepts par la présentation de nombreuses situations-problèmes. L'élaboration des situations s'appuie sur une classification préalable des relations de base intervenant dans le champ conceptuel considéré et l'analyse des classes de problèmes que l'on peut générer à partir de ces relations. La résolution des situations-problèmes

présentées permet d'introduire l'utilisation de procédures de traitement. Pour faciliter l'assimilation des concepts, le cours ne se limite pas à une présentation traditionnelle dans le symbolisme mathématique. Il est imagé par de nombreuses représentations graphiques et simulations. Des questions sont posées en langue naturelle et les réponses de l'élève, attendues sous cette forme, sont analysées par le système.

Un grapheur intégré permet la construction et l'analyse de graphiques. La progression dans le cours impose la lecture et l'utilisation de formules mathématiques. L'élève doit aussi se familiariser avec la pratique du calcul numérique à l'aide d'une calculette intégrée. Ainsi, les concepts sont présentés sous des formes symboliques aussi variées que possible.

Pour prendre en compte la diversité des voies d'accès à un problème, ainsi que les diversités cognitives des apprenants, des changements de registre ont été utilisés pour faire évoluer les conceptions. Par exemple, passage du registre géométrique à celui de l'ajustement d'un modèle à partir d'un échantillon, dans le module sur l'ajustement linéaire.

Contenu

Le module "ajustement linéaire" a pour but de présenter deux aspects fondamentaux d'une analyse de données :

* il introduit la notion de modèle, qui conduit naturellement à la partie inférentielle, complémentaire de la partie exploratoire d'une analyse statistique,
* par l'analyse des résidus on aborde la problématique du diagnostic et du contrôle de la qualité du modèle.

Le modèle de la régression simple met en relation, par l'intermédiaire d'une fonction affine, la variable expliquée Y et la variable explicative X.

$$Y = aX + b$$

Mais à cette fonction affine qui permet de modéliser le comportement de Y en fonction de X s'ajoute un terme d'erreur traduisant notamment le fait que la fonction retenue ne permet pas de prendre en compte tous les éléments explicatifs de Y. Ainsi, pour un ensemble de n observations relatives aux variables Y et X :

$$yi = axi + b + ri \qquad i \in [1, n] \text{ et } i \in \mathbb{N}$$

La méthode des moindres carrés consiste à minimiser la somme des carrés des résidus r_i :

$$\sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - ax_i + b_i)^2$$

Elle permet de déterminer les coefficients a et b, à partir des n observations.

Une décomposition de la variance totale, en variance expliquée par le modèle et variance résiduelle, permet d'introduire le coefficient de déterminaton R^2. La qualité d'un ajustement peut être, en partie jugée par la valeur de R^2, mais il est aussi nécessaire d'étudier le graphique des résidus. Ce graphique est obtenu facilement grâce au grapheur intégré au didacticiel.

Lorsque les relations entre les variables ne sont pas affines, différentes transformations permettent de se ramener au cas affine, et on peut juger de la qualité du nouveau modèle, au moyen des instruments introduits.

Mardi 1/09/92 - Atelier 16h à 18 h

LABROUE, SAINT-PIERRE :

Echantillonnage, Estimation (niveau 1)

Au cours de cet atelier, nous avons d'abord présenté une activité préparatoire destinée aux élèves de section de technicien supérieur : cette activité consiste à faire prélever aux élèves (de préférence chez eux) des échantillons aléatoires, avec remise, de taille 30 dans une population de 100 éléments, (pages 2 et 3) .
Les élèves déterminent, pour chaque échantillon prélevé, soit la moyenne, soit le pourcentage des éléments de cet échantillon ayant une propriété donnée.
Les résultats reportés dans un tableau fournissent, sans formalisme inutile à ce niveau de formation, des données utiles pour introduire le cours sur l'échantillonnage, mais également pour celui concernant l'estimation par intervalle de confiance.

Des sujets de brevets de technicien supérieur se rapportant à ces deux parties ont été analysés et traités, nous avons passé pas mal de temps à discuter sur le contenu et la rédaction de ces sujets. D'autres rédactions ont été proposées, (pages 4, 5 et 6).

Nous avons été limités par le temps, peu d'exercices sur l'estimation ont été abordés, aussi, avons nous distribué aux intéressés des propositions de rédactions de sujets de brevets de technicien supérieur des dernières années avec leurs corrigés (20 pages).

ECHANTILLONNAGE : Travaux pratiques

D'après TP 1 et 2 page 135 (livre de Mr Verlant)

On considère la population constituée des 100 éléments suivants:

classe	0	1	2	3	4	5	6	7	8	9	10
effectif	1	3	7	12	17	20	17	12	7	3	1

1° Calculer la moyenne m et l'écart type σ de cette population. Calculer le pourcentage p d'éléments de la population correspondant à un nombre supérieur ou égal à 6.

2° Prélever au hasard et avec remise, 10 échantillons de 30 éléments de la population. Dans chacun de ces échantillons, calculer les moyennes respectives $\bar{x}_1, .. \bar{x}_n$, ainsi que les pourcentages f_1, f_2, ..f_n d'éléments correspondant à un nombre supérieur ou égal à 6.

3° Calculer la moyenne μ et l'écart type s de la nouvelle série statistique constituée des 10 moyennes ainsi que la moyenne $\bar{f}$ et l'écart type s' de la série statistique constituée des 10 pourcentages f_i.

4° a) Comparer m et μ d'une part, et d'autre part $\frac{\sigma}{\sqrt{30}}$ et s

b) Comparer p et $\bar{f}$ d'une part, et d'autre part $\sqrt{\frac{p(1-p)}{30}}$ et s' : leurs valeurs numériques sont-elles proches ?

5° a) Pour chaque échantillon, déterminer les bornes de l'intervalle :

$\left[\bar{x}_i - 1{,}96 \frac{\sigma}{\sqrt{30}} ; \bar{x}_i + 1{,}96 \frac{\sigma}{\sqrt{30}} \right]$ et placer ces dix intervalles sur 10 axes parallèles en alignant verticalement les abscisses communes.

Quel pourcentage de ces intervalles contient la moyenne m de la population ?

b) Même question avec les intervalles $\left[f_i - 1{,}96 \sqrt{\frac{p(1-p)}{30}} ; f_i + 1{,}96 \sqrt{\frac{p(1-p)}{30}} \right]$.

Quel pourcentage de ces intervalles contient la valeur de p calculée au 1° ?

Placer les résultats dans ce tableau :

échantillon	1	2	3	4	5	6	7	8	9	10
$\bar{x}_i$										
f_i										

0 1 1 1 2 2 2 2

2 2 2 3 3 3 3 3

3 3 3 3 3 3 3 4

4 4 4 4 4 4 4 4

4 4 4 4 4 4 4 4

5 5 5 5 5 5 5 5

5 5 5 5 5 5 5 5

5 5 5 5 6 6 6 6

6 6 6 6 6 6 6 6

6 6 6 6 6 7 7 7

7 7 7 7 7 7 7 7

7 8 8 8 8 8 8 8

9 9 9 10

ECHANTILLONNAGE

Exercice 1 MICROTECHNIQUE 89 *(extrait)*

Une machine fabrique des pièces de forme circulaire en grande série. A chaque pièce, tirée au hasard, on associe son diamètre exprimé en millimètres ; on définit ainsi une variable aléatoire X. On suppose que X suit une loi normale ; on désigne par m sa moyenne et par σ son écart type.

On suppose connue la moyenne : m = 150. On a constaté que 8 % des pièces de la production ont un diamètre supérieur à 150,3 mm.

1° Quelle est la loi suivie par la variable $T = \dfrac{X - 150}{\sigma}$? Etablir que $\sigma = 0{,}21$.

2° Calculer le pourcentage de pièces de cette fabrication dont le diamètre est compris entre 149,79 et 150,42.

3° Soit $\overline{X}$ la variable aléatoire qui à chaque échantillon de 400 pièces associe la moyenne des diamètres des pièces de cet échantillon. $\overline{X}$ suit la loi normale de moyenne m et d'écart type

$s = \dfrac{\sigma}{\sqrt{400}} = 0{,}0105.$

Déterminer h pour que $P(m - h \leq \overline{X} \leq m + h) = 0{,}95$.

Exercice 2 MECANIQUE ET AUTOMATISMES INDUSTRIELS 88 *(extrait)*

Une machine fabrique des pièces en grande série. La variable aléatoire X qui, associe à chaque pièce tirée au hasard sa longueur suit la loi normale de moyenne m = 28,20 mm et d'écart type $\sigma = 0{,}027$ mm

On admet que la variable aléatoire $\overline{X}$ prenant pour valeurs les moyennes des échantillons de même taille n suit la loi normale de moyenne m et d'écart type $\dfrac{\sigma}{\sqrt{n}}$.

1° Une pièce est " bonne " si sa longueur appartient à l'intervalle [28,15 ; 28,27] .
Calculer le poucentage de pièces " bonnes " dans la fabrication.

2° On prélève un échantillon (non exhaustif) dans cette production.
Quelle doit être la taille de l'échantillon pour que la moyenne des longueurs des pièces prélevées appartienne à l'intervalle [28,195 ; 28,205] avec une probabilité de 0,95 ?

Exercices non corrigés :

I)

Dans une population donnée, la proportion p de fumeurs est 0,36. Quelle est la probabilité qu'en prélevant avec remise un échantillon simple de 400 personnes dans cette population la proportion de fumeurs dans l'échantillon soit plus grande ou égale à 0,40 ?

II)

Au cours d'une consultation électorale, un candidat a recueilli 46 % des suffrages.

1° Quelle est la probabilité qu'un groupe de 200 personnes, prises au hasard, lui ait donné la majorité ?

2° Même question avec un groupe de 1000 personnes.

Exercice 4 BTS MAINTENANCE 84

Une machine automatique fabrique des pièces. On suppose que la variable aléatoire X qui à chaque pièce associe son poids x exprimé en grammes suit la loi normale de moyenne 0,90 et d'écart type 0,06.

Une deuxième machine met les pièces fabriquées en boîtes, à raison de 100 pièces par boite ; on prélève au hasard un certain nombre de boites et pour chaque boîte on mesure la moyenne $\bar{x}$ des poids des pièces de cette boîte. On désigne par $\overline{X}$ la variable aléatoire qui à chaque boîte associe la moyenne des poids $\bar{x}$.

1° Quelle est la loi suivie par $\overline{X}$?

2° Déterminer $P(|\overline{X} - 0,9| \leq 0,01)$ c'est à dire $P(0,9 - 0,01 \leq \overline{X} \leq 0,9 + 0,01)$

3° Déterminer un intervalle centré en 0,9 tel que $\overline{X}$ prenne une valeur dans cet intervalle avec la probabilité 0,95.

Exercice 5 BTS MAI 86

Une machine fabrique des disques pleins en grande série. On suppose que la variable aléatoire X qui, à chaque disque tiré au hasard, associe son diamètre suit la loi normale $N(\mu,\sigma)$ où $\mu = 12,8$ mm et $\sigma = 2,1$ mm.

1° Quelle loi suit la variable aléatoire $\overline{X}$ qui, à tout échantillon aléatoire non exhaustif de taille $n = 49$, associe la moyenne des diamètres des disques de cet échantillon ?

2° Déterminer un intervalle centré en 12,8 tel que la variable aléatoire $\overline{X}$ prenne ses valeurs dans cet intervalle avec la probabilité 0,95.

3° On se propose de prélever un échantillon aléatoire non exhaustif de taille n. Déterminer n pour que la moyenne des diamètres des disques prélevés ne s'écarte pas de 12,8 de plus de 0,2 mm, avec une probabilité de 0,95.

Exercice 6 BTS PRODUCTIONS ANIMALES 87

On admet que dans un élevage de poulets fermiers agés de 3 mois, la variable aléatoire prenant pour valeur le poids d'un poulet suit la loi normale de moyenne 1 325 g et d'écart type 175 g.
On prélève au hasard de manière non exhaustive 16 poulets de cet élevage.

1° Déterminer la loi de probabilité de la variable aléatoire $\overline{X}$ qui, associe à chaque échantillon de 16 poulets le poids moyen des poulets de cet échantillon.

2° Calculer la probabilité que le poids total d'un échantillon de 16 poulets :

a) dépasse 22 kg ; [c'est à dire $P(\overline{X} > \frac{22000}{16})$]

b) soit compris entre 20 kg et 22 kg ; [c'est à dire $P(\frac{2000}{16} < \overline{X} < \frac{22000}{16})$]

c) soit inférieur à 20 kg . [c'est à dire $P(\overline{X} < \frac{20000}{16})$]

ESTIMATION

Exercice 1 CHIMISTE 91 (*extrait*)

On a contrôlé le dosage d'un produit dans un mélange à la sortie d'une chaine de conditionnement. On a prélèvé de manière aléatoire, un échantillon de 100 lots de 5 kilogrammes de mélange analysés, on a obtenu les résultats suivants où P_i représente la masse du produit exprimée en grammes et n_i l'effectif correspondant :

P_i	142	144	146	148	150	152	154	156	158	160
n_i	1	5	6	21	32	22	7	4	1	1

1° Calculer la moyenne et l'écart type des masses du produit dans cet échantillon.

2° A partir des résultats obtenus pour cet échantillon, donner une estimation ponctuelle de la moyenne m et de l'écart type σ de la masse du produit de la population.

3° On suppose que la variable aléatoire qui, à tout échantillon de 100 lots associe la moyenne des masses du produit, suit la loi normale $\mathcal{N}(m, \frac{\sigma}{\sqrt{100}})$ et on prend pour σ l'estimation ponctuelle obtenue au 2° .Déterminer un intervalle de confiance de la moyenne m de la population avec le coefficient de confiance 95 % .

4° Même question avec le coefficient de confiance 99 %, puis avec le coefficient de confiance 90 %

Exercice 2 CONSTRUCTION NAVALE 85

On a mesuré les longueurs en mm d'un échantillon de 100 tiges d'acier, tirées au hasard, à la sortie d'une machine automatique :

longueurs	effectif
[132 ; 134 [	2
[134 ; 136 [	5
[136 ; 138 [	13
[138 ; 140 [	24
[140 ; 142 [	19
[142 ; 144 [	14
[144 ; 146 [	10
[146 ; 148 [	8
[148 ; 150 [	3
[150 ; 152 [	2

1° Calculer la moyenne et l'écart-type des longueurs des tiges dans cet échantillon.

2° A partir des résultats obtenus pour cet échantillon, proposer une estimation ponctuelle de la moyenne μ et de l'écart type σ de la longueur de toutes les tiges sorties de la machine.

3° Donner un intervalle de confiance à 99 % de la longueur moyenne μ des tiges de toute la production de la machine.

4° Quelle doit être la taille d'un échantillon extrait de la population pour que la moyenne des longueurs des tiges de la production soit estimée à 10^{-1} près, avec le coefficient de confiance 95 %

Samedi 5/09/92 - Atelier 8h30 à 10h30

SAINT-PIERRE

Application de la régression : loi de Weibull.

L'atelier était destiné à montrer à l'aide d'un TP (construction du papier de Weibull) une application de la régression linéaire au niveau BTS.
Les auditeurs ne connaissant pas la loi de Weibull, il a fallu, au préalable, donner les principales définitions et caractéristiques de la fiabilité et en particulier de la loi de Weibull (pages 8 et 9).

Le TP (pages 10 et 11) a été traité partiellement par les auditeurs, la fin présentant moins d'intéret.
Le groupe a préféré traiter les deux exercices d'un TP (pages12 et 13).
Ce TP a permis de comprendre l'utilisation du papier de Weibull pour déterminer les paramètres η puis β et γ dans les deux cas nul ou non nul.

Nous avons pu, au cours de cet atelier, contaster que pour comprendre ces domaines concrets et trouver des applications dans l'industrie, la démarche des professeurs était la même dans des spécialités pourtant très différentes (agronomie, maintenance etc..).

LA FIABILITE

1. Fonctions de défaillance F(t) et de fiabilité R(t) :

- Modèle mathématique:
La variable aléatoire T qui mesure la durée de vie sans défaillance ou le temps de bon fonctionnement avant défaillance (TBF) est une variable aléatoire de type continu prenant ses valeurs dans R^+, de densité de probabilité f .
Sa fonction de répartition F est appelée fonction de défaillance.
F(t) est la probabilité d'avoir vu le matériel en panne avant une durée t de fonctionnement.

$$F(t) = P(T \leq t)$$

F(t) est appelée fonction de défaillance.

La fiabilité R(t) est la probabilité que le matériel fonctionne sans panne jusqu'au temps t.

$$R(t) = 1 - F(t) = P(T > t)$$

R(t) est appelée fonction de survie ou fonction de fiabilité du matériel.

f étant la densité de probabilité et F la fonction de répartition, on a, si F est dérivable :

$$F'(t) = f(t) \qquad F(t) = \int_0^t f(x)dx$$

- Aspect statistique :

$$R(t) = \frac{N(t)}{N(0)} = \frac{\text{nombre de survivants}}{\text{nombre de matériel initial}}$$

La fonction de fiabilité R(t) est alors la limite d'une fréquence.
La fonction de défaillance est alors :
On a également $F(t) = \dfrac{N(0) - N(t)}{N(0)} = \dfrac{\text{nombre de défaillants}}{\text{nombre de matériel initial}}$

2. Taux d'avarie λ(t) :

- Aspect statistique :

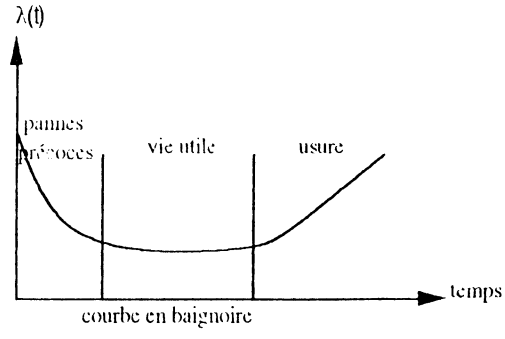
$$\lambda(t) = \frac{f(t)}{R(t)} = \frac{f(t)}{1 - F(t)}$$

- Modèle mathématique :

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{F(t + \Delta t) - F(t)}{R(t)}$$

$$\lambda(t) = \frac{F'(t)}{R(t)} = \frac{f(t)}{R(t)}$$

λ(t) est appelé le taux d'avarie instantané ou taux de défaillance instantané.
On constate expérimentalement , que pour la plupart des matériels, la courbe représentative du taux d'avarie instantané a la forme suivante :



courbe en baignoire

3. Relation entre R(t) et λ(t)

On a vu que $\lambda(t) = \frac{f(t)}{1 - F(t)} = \frac{F'(t)}{1 - F(t)}$ donc $\lambda(t)dt = \frac{F'(t)}{1 - F(t)}dt$

$$\int_0^t \frac{F'(x)}{1 - F(x)}dx = -[\ln(1 - F(x))dx]_0^t$$

$$\int_0^t \lambda(x)dx = -[\ln R(x)]_0^t, \qquad \int_0^t \lambda(x)dx = -\ln R(t)$$ et finalement :

$$\boxed{R(t) = e^{-\int_0^t \lambda(x)dx}} \qquad \boxed{F(t) = 1 - e^{-\int_0^t \lambda(x)dx}}$$

Le choix du modèle mathématique pour la fonction λ que l'on sait, devoir répondre à certaines conditions nous permet, par un calcul intégral de déterminer la fiabilité R(t) du matériel considéré.

LOI DE WEIBULL

DEFINITION

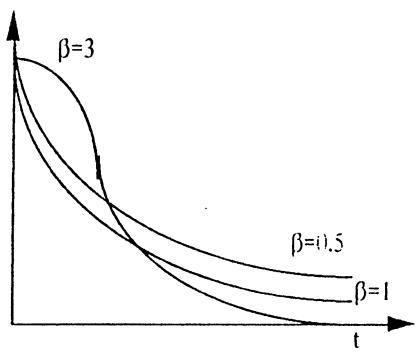
Pour couvrir tous les cas où le taux d'avarie λ(t) est variable et afin de faciliter le calcul intégral, Weibull a choisi pour λ(t) une fonction puissance :

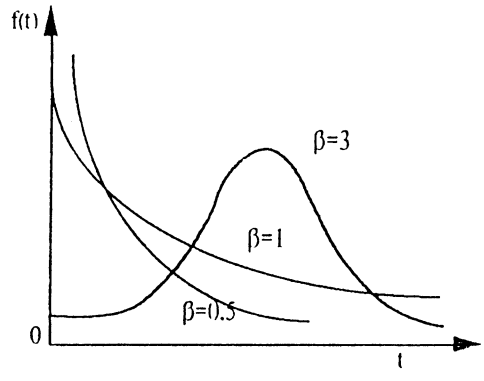
$$\boxed{\lambda(t) = \frac{\beta}{\eta}\left(\frac{t - \gamma}{\eta}\right)^{\beta - 1}}$$

avec $t - \gamma > 0$; $\beta > 0$; $\eta > 1$

$R(t) = e^{-\int_0^t \lambda(x)dx}$ donc en intégrant la fonction puissance $\lambda(t) = \frac{\beta}{\eta}\left(\frac{t - \gamma}{\eta}\right)^{\beta - 1}$ on obtient :

$$\boxed{R(t) = e^{-\left(\frac{t - \gamma}{\eta}\right)^{\beta}} \qquad F(t) = 1 - e^{-\left(\frac{t - \gamma}{\eta}\right)^{\beta}} \qquad f(t) = \frac{\beta}{\eta}\left(\frac{t - \gamma}{\eta}\right)^{\beta - 1} e^{-\left(\frac{t - \gamma}{\eta}\right)^{\beta}}}$$





UN TP POUR CONSTRUIRE DU PAPIER DE WEIBULL

Le but de ce TP est de reconstituer le papier de Weibull.

On rappelle que dans le modèle de Weibull. lorsque $\gamma = 0$, la probabilité de défaillance F(t) est donnée par la relation : $F(t) = 1 - e^{-[t/\eta]^{\beta}}$

1°) Démontrer que $\ln\left(\frac{1}{1 - F(t)}\right) = \left(\frac{t}{\eta}\right)^{\beta}$.

Démontrer en posant $X = \ln t$ et $Y = \ln\left[\ln\left(\frac{1}{1 - F(t)}\right)\right]$ que l'on obtient l'équation de droite :

$$Y = \beta X - \beta \ln \eta$$

2°) Compléter les tableaux suivants :

t	0,1	0,3	e^{-1}	0,5	1	3	5	8	10	20	50
ln t	- 2,3				0				2,3		
4 × ln t	- 9,2				0				9,2		

F(t) en %	0,1 %	1 %	10 %	20	30	50	60	70	80	90	99 %
$\ln\left(\ln\frac{1}{1-F(t)}\right)$	- 6,9						- 0,08				
Y × 2 cm	- 13,8										

3°) Sur l'axe "X_1" du papier millimétré ci- joint la valeur de t placée en H est 0,1. Pour t = 0,1 on a ln t ≈ - 2,3 cm, d'où avec l'échelle choisie 4 × ln 0,1 ≈ - 9,2 cm.

L'axe des ordonnées "Y" aura sa position déterminée par le point I correspondant à ln t = 0 donc à t = 1 donc tel que $\overline{IH}$ = - 9,2 cm.

Placer sur cet axe "Y" les valeurs de t données dans le premier tableau.

4° Sur l'axe "Y_1" la valeur de F(t) placée en H est 0,001 ou 0.1 %. Pour F(t) = 0,001 on a $\ln\left[\ln\left(\frac{1}{1 - F(t)}\right)\right]$ ≈ - 6,9 cm d'où avec l'unité choisie - 13,8 cm.

L'axe des abscisses "X" aura sa position déterminée par le point K correspondant à $\ln\left[\ln\left(\frac{1}{1 - F(t)}\right)\right] = 0$ donc tel que $\overline{KH}$ = - 13,8 cm. Placer alors la valeur de F(t) correspondant au point K. puis les autres valeurs de F(t) données dans le deuxième tableau.

5°) Pour obtenir le coefficient directeur de la droite on trace un nouvel axe des ordonnées "Y_2" parallèle à "Y_1" et "Y" tel que son abscisse soit (- 1) , sa position est donc telle que ln t = - 1 ou $t = \frac{1}{e}$. Construire cet axe. Le module de cet axe est l'unité choisie sur l'axe des ordonnées, son origine est sur l'axe "X" et son vecteur unitaire est opposé à celui de l'axe "Y". Graduer cet axe.

Tracer la droite d'équation Y = 2 X - 2 ln 10, en déduire F(t) et R(t) en fonction de t.

ln t
t
axe X1
1
1
H
0,1
0,1
axe Y1
K
F(t) =
F(t) en %
99%

TRAVAUX DIRIGES

Exercice 1

Une usine utilise 19 machines de même modèle. L'étude du bon fonctionnement en heures, avant la première panne de chacune de ces 19 machines, a permis d'obtenir l'historique suivant :

TBF en heures	]0, 250]	]250, 450]	]450, 600]	]600. 800]	]800, 1100]	]1100, 1400]
Nombre de pannes	1	2	2	3	4	4

Trois autres machines ont fonctionné correctement au moins jusqu'à la date :" 1400 heures ".

1° Déterminer à l'aide du papier de Weibull les paramètres de la loi de Weibull ajustant cette distribution. Donner l'expression de R(t).

2° Calculer la MTBF et la Médiane de cette série statistique.

3° Déterminer graphiquement puis par le calcul la périodicité d'un entretien systématique basé sur une fiabilité de 0,9.

4° Déterminer graphiquement et par le calcul, la probabilité qu'une machine de ce type fonctionne plus de 2000 heures sans panne.

Exercice 2

Une usine utilise 19 machines de même modèle. L'étude du bon fonctionnement en heures, avant la première panne de chacune de ces 19 machines, a permis d'obtenir l'historique suivant :

TBF en heures	]1000, 1250]	]1250, 1450]	]1450, 1600]	]1600, 1800]	]1800, 2100]	]2100, 2400]
Nombre de pannes	1	2	2	3	4	4

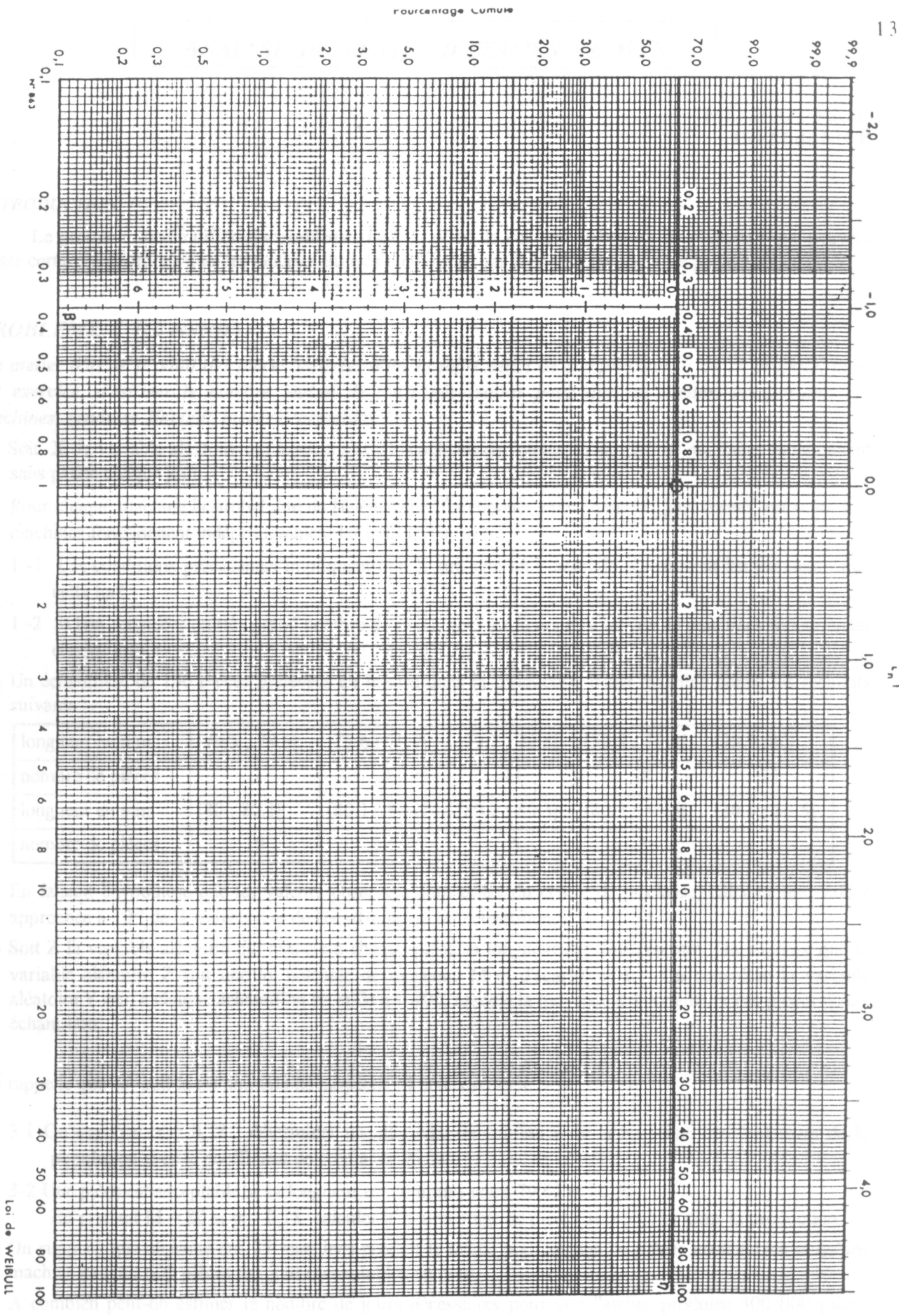
Trois autres machines ont fonctionné correctement au moins jusqu'à la date : " 2400 heures ".

1° Déterminer à l'aide du papier de Weibull les paramètres de la loi de Weibull ajustant cette distribution. Donner l'expression de R(t).

2° Calculer la MTBF et la Médiane de cette série statistique.

3° Déterminer graphiquement puis par le calcul la périodicité d'un entretien systématique basé sur une fiabilité de 0,9.

4° Déterminer graphiquement et par le calcul, la probabilité qu'une machine de ce type fonctionne plus de 2000 heures sans panne.


Pourcentage Cumulé
Loi de WEIBULL

ANALYSE DE TEXTES D'EXAMEN DE BTS

MERIGOT Michel
IREM de NICE

INTRODUCTION

Le groupe a étudié différents problèmes tant du point de vue des difficultés de rédaction que peuvent poser certaines questions que des modèles utilisés qui ne sont pas à développer dans les problèmes de BTS.

PROBLEME I (10 POINTS)

Un atelier d'usine produit des pièces utilisées dans la fabrication de compresseurs ; on se propose, dans cet exercice, d'estimer le nombre moyen de jours nécessaires pour que cet atelier, comportant 25 machines, produise 300 000 pièces acceptables.

1 - Soit X la variable aléatoire qui associe aux 25 machines de l'atelier le nombre de machines fonctionnant sans panne chaque jour.

Pour chaque machine la probabilité de l'événement "fonctionner sans panne un jour donné" est 0.96. Les machines fonctionnent indépendamment les unes des autres.

1 -1 Les conditions précédentes impliquent que X suit une loi binomiale. Quels sont les paramètres de cette loi?

1 -2 Si l'on considère un très grand nombre de jours, quel est le nombre moyen de machines fonctionnant chaque jour ?

2 - Un échantillon de 100 pièces prélevées au hasard dans la production d'une machine donne les résultats suivants :

longueur en mm	[79.5 ; 79.6[	[79.6 ; 79.7[	[79.7 ; 79.8[	[79.8 ; 79.9[	[79.9 ;80[
nombre de pièces	2	4	11	18	23
longueur en mm	[80 ; 80.1[	[80.1 ; 80.2[	[80.2 ; 80.3[	[80.3 ; 80.4[	[80.4 ; 80.5[
nombre de pièces	19	14	5	3	1

En faisant l'hypothèse que les valeurs observées sont celles du centre de la classe, calculer une valeur approchée à 10^{-2} de la moyenne et de l'écart-type de cet échantillon.

3 - Soit Z la variable aléatoire qui associe à chaque pièce sa longueur exprimée en mm. On suppose que la variable aléatoire Z suit une loi normale de moyenne M et d'écart-type σ. On considère la variable aléatoire Y qui, à chaque échantillon de taille n = 100, associe la moyenne des longueurs des pièces de cet échantillon.

On rappelle que Y suit la loi normale de moyenne M et d'écart type $\frac{\sigma}{\sqrt{n}}$.

3.1 On suppose σ = 0,18 , déterminer un intervalle de confiance de la moyenne des longueurs de la population avec le coefficient de confiance 95 %.

3-2 Une pièce est acceptable si sa longueur est comprise entre 79,65 et 80,35.
En prenant M = 80 et σ = 0,18, calculer le pourcentage de pièces acceptables dans la production.

4 - On suppose que chacune des 25 machines produit le même pourcentage de pièces acceptables et qu'une machine produit 600 pièces par jour (acceptables ou non).

A combien peut-on estimer le nombre de jours nécessaires pour que l'atelier produise 300 000 pièces acceptables. (On utilisera les questions (1-2 et 3-2).

REMARQUES SUR LE PROBLEME 1

1- La première question ne pose aucune difficulté, par contre la question 1-2 a amené une longue discussion. On a retenu deux possibilités de rédaction :

a) Sur n jours (n éventuellement précisé), le nombre moyen de machines en fonctionnement est une variable aléatoire N. Quelle est son espérance ? (On supposera que les événements "fonctionner sans panne un jour donné" sont indépendants) (*).

b) Si l'on veut garder l'approche fréquentiste, on peut demander d'étudier la loi de N.

Soit X_i la variable aléatoire "nombre de machines fonctionnant sans panne le jour i"

$X_i \sim \mathcal{B}(25 ; 0{,}96)$

$Y = X_1 + \ldots + X_n$

Les variables sont indépendantes : $Y \sim \mathcal{B}(n\times25 ; 0{,}96)$ et $N = \frac{Y}{n}$. N prend des valeurs de la forme k/n et les probabilités sont celles d'une loi binomiale.

$$E(N) = \frac{n\times25\times0{,}96}{n} = 24 \quad \text{et} \quad \text{var}(N) = \frac{n\times25\times0{,}96\times0{,}04}{n^2} = \frac{0{,}96}{n}$$

Quand n est grand, la variance de N est très petite et la loi est très concentrée autour de 24.

2- $\bar{x}$ nous est inconnue, mais la valeur obtenue en prenant pour valeur observée le centre de la classe est une bonne approximation de $\bar{x}$.

3-1 Même si l'on utilise l'écart-type calculé, on peut déterminer un intervalle de confiance de la moyenne car la variable $\frac{Y - M}{\sigma/\sqrt{n}}$ est bien représentée par une loi normale réduite.

4- On retombe sur les difficultés de la question 1-2 avec en plus la combinaison d'une loi binomiale et d'une loi normale.

PROBLEME 2

Au cours des élections présidentielles du 10 mai 1981, 51.75% des électeurs ont voté pour Mitterrand et 48.25 % pour Giscard d'Estaing.

a) Si un sondage de 1000 électeurs choisis au hasard dans la population française avait été réalisé, quelle aurait été la probabilité de prévoir à tort la défaite de Mitterrand ?

b) Si le risque de se tromper dans la question précédente doit être réduit à 1%, quelle doit être la taille de l'échantillon ?

REMARQUES SUR LE PROBLEME 2

2 interprétations ont été données aux questions :

- Soit, on cherche la probabilité pour qu'un sondage propose un résultat défavorable
- Soit, on attache à chaque sondage un intervalle de confiance et, pour un risque donné, on recherche la probabilité que la borne supérieure de l'intervalle soit encore inférieure à 0.50

*) N.d.E. Il faut de plus supposer qu'une machine tombant en panne est réparée pour le lendemain.

UTILISATION D'UN LOGICIEL STATISTIQUE : STATITCF

PAVY Jacques

PLAN

1 – INTRODUCTION

L'étude personnelle de ce logiciel est étroitement liée à des nécessités pédagogiques :

- ouverture d'une formation BTS Productions Végétales au LEGTA Le Robillard (14) en 1989–90 et demande d'étudiants, pour leur rapport de stage, d'interpréter l'analyse de variance et des comparaisons multiples de moyennes d'essais agronomiques en blocs, voire en Split–Plot (hors programme),
- puis mise en place, dans le programme officiel de la rénovation de ce BTS à la rentrée 91–92, d'un module particulier : le D4–2 "expérimentation" (annexe 1).

Le logiciel STATITCF a été créé par le service statistique de l'Institut Technique des Céréales et des Fourrages (ITCF). Il a été conçu dans le souci de coller au plus près des besoins de leurs collègues ingénieurs agronomes pour traiter les résultats d'essais agronomiques.

Cet atelier utilise la version n° 4 (1987–88). Une version plus récente est parue en novembre 1991.

2 – PRESENTATION D'UN ESSAI

Essai variétal, piloté par l'ENSA Rennes, sur la féverole précoce avec 5 variétés et 3 blocs, donc 15 parcelles de 2 m x 7 m.

- Alfred : variété régulière
- Alscott : variété ancienne 1965–70
- Blandine : variété à gros grains (coût élevé au semis)
- Diana : petits grains
- Skladia : variété inconnue.

L'objectif en sélection est d'obtenir des variétés à petits grains avec un poids 1000 grains plus faible, qu'on veut compenser pour le rendement par une variété ayant plus de grains/pied/gousse.

En septembre 1991, les gousses arrivant à maturité, l'essai a été réutilisé en essai pédagogique pour la classe BTA option QP3 "expérimentation".

1ère séance (1 h)

- chaque élève reçoit une gerbe de 20 pieds, échantillon prélevé sur une même parcelle ; il note le n° de parcelle
- il fait les mesures suivantes :
  - . la hauteur de la gerbe
  - . le nombre de tiges fertiles par pied (une tige est dite fertile si elle a au moins une gousse)
  - . le poids total de la gerbe

. il enlève toutes les gousses, écosse et compte le nombre total de grains par gerbe
. enfin il pèse ces grains par gerbe ;

d'où la feuille de résultats sur les 15 parcelles (cf. p.4).

2ème séance : visite sur le terrain de l'essai (1 h)

- chaque élève reçoit un plan (cf. annexe)
- le collègue décrit l'essai :
  . 4 rangs par parcelle ; dose de semis : 50 grains/m², soit 50 gr x 14 m²=700 grains/parcelle
  . chaque variété est affectée, par tirage au sort, à une seule parcelle dans chaque bloc
  . par bloc, il y a donc autant de parcelles que de variétés
  . les parcelles sont côte à côte, sans séparation par un rang de bordure.

3ème séance : séance informatique animée par le collègue informaticien du lycée (2 h)

découverte du clavier et rudiments du DOS.

4ème séance : saisie de l'essai sur STATITCF, 3 professeurs présents (2 h 30)

présentation d'une fiche élève.

5ème séance (2 h)

création des variables à calculer avec intervention du collègue de phyto pour expliquer les objectifs de la sélection.

6ème séance (2 h)

petite recherche de relations entre variables (matrice de corrélation, régression) en parallèle avec la progression du cours de statistiques sur variable à 2 dimensions.

L'atelier reprend comme support les résultats élèves de cet essai et a pour objectifs :

- initier aux difficultés particulières de saisie d'un essai sur STATITCF en vue d'une analyse de variance dans le module G,
- commenter et interpréter l'analyse de variance avec le module H.

3 – LE MODULE G

L'analyse de variance ne se fera pas à partir des données recueillies, mais d'indicateurs construits par l'agronome : – rendement, – aptitude à maximiser la matière végétale en grains, – poids de 1000 grains, afin de comparer les variétés sur ces critères.

Plan de travail :
- descriptif de l'essai → fichier
- plan de l'essai à saisir
- saisie des données recueillies
- construction des variables.

On entre dans STATITCF en tapant "STATITCF" puis Entrée ; le menu général est alors affiché. En tapant G, on sélectionne le module "Gestion de données pour une analyse de variance", d'où le nouvel écran de menu.

Remarque : se mettre en majuscule, sinon STATITCF proteste !

A . DESCRIPTIF DE L'ESSAI

L'objectif de l'étude est de comparer les variétés. On intègre le facteur "terrain" dans le dispositif afin d'augmenter la part expliquée de la variabilité totale, donc de diminuer la partie non maîtrisée – dite résiduelle – de celle-ci.

Phillipeau, [1, p.77] : "*ce facteur n'est pas à proprement parler étudié, il est seulement pris en compte de façon que son influence dans la variation des résultats soit éliminée.*"

Goupy, [2, p.128] : "*... le blocking étant l'art de regrouper les expériences pour éliminer l'influence du facteur qui peut être génant.*"

Facteur étudié, facteur contrôlé, parcelles

. 5 variétés . 3 blocs	Plan d'expérience avec contrôle d'hétérogénéité en blocs complets	Plan à 2 facteurs sans répétition

. un bloc est un terrain homogène

. chaque variété se retrouve donc une fois par bloc, au hasard.

1er facteur : variété, 5 niveaux → facteur étudié.

2ème facteur : hétérogénéité, 3 niveaux (3 blocs) → facteur contrôlé.

Le statut de ces 2 facteurs n'est pas symétrique.

B . SAISIE DU PLAN

Dans le menu du module G, on choisit l'option B : "Création du fichier d'un essai déjà mis en place".

Le nombre de facteur étudié est 1 (la variété). On entre le nom du facteur :

Intitulé (16 caractères (1))	intitulé réduit (5 caractères)
VARIETE FEVEROL	VA FE

Le facteur a 5 niveaux ; pour chacun, on tape le nom, puis le code (sur 3 caractères).

Pour le facteur contrôlé, on choisit le dispositif BLOC avec 3 blocs.

A la question "TEMOIN ADJACENT ?", on répond : N(on).

On donne le nom du fichier : FE_UETE et le logiciel crée 3 variables :

V_1 : VA FE avec 5 valeurs

V_2 : BLOC avec 3 valeurs No DE BLOC

V_3 : PARCE avec 15 valeurs No DE PARCELLE

ENREGISTREREZ-VOUS LE PLAN DE L'ESSAI (O/N) ? : réponse O

PAR PARCELLE, COMBIEN PREVOYEZ-VOUS DE VARIABLES A INTRODUIRE (max=50) ? : réponse 5 ; ce sont les 5 variables mesurées sur chaque gerbe.

VARIABLE	LIBELLE	LIBELLE (16 car.)	NOMBRE DE DECIMALES
V_4	N TIG	N TIGES FERTILES	0
V_5	H GE	HAUTEUR GERBE	1
V_6	PT GE	POIDS TOTAL GERB	0
V_7	N GR	NOMBRE GRAINS	0
V_8	PT GR	POIDS TOTAL GRAI	0

PAR PARCELLE, COMBIEN PREVOYEZ-VOUS DE VARIABLES A CALCULER (max=55) ? : réponse 3 ; elles définiront des critères de sélection.

V_9	RT GR	RENDT GRAIN Q/HA	1
V_{10}	I RG	INDICE RDT GRAIN	1
V_{11}	PMG	POIDS 1000 GRAINS	0

Bilan : le logiciel enregistre le fichier qui contient 11 variables et 15 individus parcelles.

---

1) Il n'y a pas de contrôle de longueur des entrées ; si l'intitulé est trop long, il faut tout réécrire correctement.

Saisie de l'essai

Le logiciel demande la description du champ. Il contient 15 parcelles par ligne et 1 par colonne.

Pas de difficulté pour le bloc 1 ; la saisie est assez délicate pour les blocs 2 et 3. Pour la variété et le bloc marqués en bas d'écran, déplacer le curseur jusqu'à la position voulue et valider par Entrée ◄┘ .

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	ALF B1	ASC B1	BLA B1	DIA B1	SKL B1	BLA B2	ASC B2	ALF B2	SKL B2	DIA B2	SKL B3	BLA B3	ALF B3	ASC B3	DIA B3
2															

```
BLOC No  3
SKL
 →↑↓ = déplacement curseur  RETURN = enregistrement  C = corrections  F=fin
```

Avant de taper Fin avec F, contrôler et corriger.

C . SAISIE DES DONNEES MESUREES

L'écran est revenu au menu du module G. Sélectionner l'option F : "Saisie des données mesurées". Le logiciel rappelle les caractéristiques du fichier FE_UETE.

Il faut saisir les 5 variables V_4, V_5, V_6, V_7, V_8, soit A sur un tableau de données (lignes: parcelles, colonnes: variables), soit B sur le plan de l'essai. On choisit A, puis C pour une saisie par ligne en commençant par l'angle A ; on obtient alors après saisie :

PARCELLE			SAISIE				
VA FE	BLOC	PARCE	N TIG	H GE	PT GE	N GR	PT GR
ALF	B1	101	21	140.0	926	884	435
ASC	B1	102	21	136.0	1031	967	443
BLA	B1	103	27	121.0	827	723	401
DIA	B1	104	21	139.5	810	1057	344
SKL	B1	105	26	133.0	1297	1353	598
BLA	B2	106	27	117.0	786	679	345
ASC	B2	107	22	139.0	1153	1279	528
ALF	B2	108	24	142.5	1070	875	491
SKL	B2	109	26	140.0	1222	1410	562
DIA	B2	110	24	140.0	926	1140	416
SKL	B3	111	28	130.0	858	857	374
BLA	B3	112	25	136.0	728	960	315
ALF	B3	113	23	136.0	1090	1055	471
ASC	B3	114	25	143.0	1281	1441	621
DIA	B3	115	25	140.0	1134	1083	543

```
ALFRED                BLOC 3
 →↓↑ = Déplacement d'une donnée  PgUp,PgDn = Déplacement de 10 lignes F=Fin
```

Remarque : cette phase est délicate, donc il faut vérifier la saisie en utilisant les touches de déplacement.

D . DEFINITION DES VARIABLES A CALCULER

L'écran est revenu au menu du module G. On sélectionne l'option H : "Elaboration de variables".

Le logiciel rappelle les caractéristiques du fichier FE_UETE. En répondant O(ui) à la question : Avez-vous une variable à calculer, le logiciel affiche la liste des variables déclarées à calculer.

On demande le calcul de la variable V_9 = RD GR en Q/Ha. Il y a 50 pieds par m². La gerbe contient 20 pieds, d'où :

$$V_9 = \frac{\text{poids grains x 50}}{20} \times \frac{10\,000\ m^2}{100\,000\ g} \qquad \text{soit } V_9 = 0{,}25 \times V_8$$

L'écran affiche les codes possibles de transformations, puis

```
***** NO ET NOMS DES VARIABLES *****

 1. VA FE    2. BLOC     3. PARCE    4. N TIG    5. H GE     6. PT GE
 7. N GR     8. PT GR    9. RT GR   10. I RG    11. PMG
-----------------------------------------------------------------------
INDIQUEZ LES TRANSFORMATIONS  (MAX= 0 ). ENTREZ LE CODE 'Z' POUR FINIR
 NO.  CODE  OPERATION      X1  X2 ----A----   ----B----  -NOM-
-----------------------------------------------------------------------
 12   J    A + B*(X1)     8        0          0.25        [on valide par ◄┘ ]
 13  [on tape Z]
TRANSFORMATIONS OK (O,N):
```

La variable V_{10} = I RG mesure la capacité de la variété à transformer la matière végétale en grains.

$$V_{10} = \frac{\text{PT GR}}{\text{PT GE}} = \frac{V_8}{V_6}$$

On a donc comme codage

```
12   H    X1 / X2      8   6
```

Remarque : cette transformation peut bloquer la machine s'il y a 0 dans les mémoires pour V_6 (division par 0). Il faut d'abord saisir les données, puis définir les variables à calculer.

La variable V_{11} = PMG est le poids mille grains.

$$V_{10} = \frac{\text{PT GR}}{\text{N GR}} = \frac{V_8}{V_7} \times 1000$$

Il faut faire 2 transformations : H, puis J

```
12   H    X1 / X2      8   7
13   J    A + B*(X1)   12       0        1000
14  [Z]
```

Lister – éditer les résultats

On sort du module G en tapant Z pour revenir au menu général. On tape A pour passer au module A : "Gestion des données", puis on sélectionne l'option A : "Liste des données". On peut faire la sortie sur écran ou sur imprimante.

N.d.E. : La fin du compte rendu n'est pas parvenu.

Bibliographie

[1] Phillipeau G. : *Théorie des plans d'expérience*, Service des études statistiques, éd. de l'ITCF, 91720 Boigneville, 1985.

[2] Goupy J. : *La méthode des plans d'expérience*, Dunod, 1988.

ATELIER DE L'UNIVERSITE D'ÉTÉ DE STATISTIQUES DE LA ROCHELLE : LE TRAITEMENT D'UN PROBLÈME DANS LE RAPPORT DE STAGE BTS

HUBERT RAYMONDAUD
CFPPA NANCY-PIXERECOURT
54220 MALZEVILLE 83-21-65-22
1 - 5 Septembre 1992

I - INTRODUCTION

Le but de l'atelier est de faire apparaître quelques unes des difficultés rencontrées par les enseignants et les stagiaires (les acteurs de la formation initiale pourront remplacer "stagiaire" par "élève") lors du traitement d'un problème pour le rapport de stage et de proposer une méthodologie d'analyse.

Après avoir fait une analyse "brute" du problème proposé, nous nous efforcerons d'en tirer une méthodologie ayant pour but de guider les enseignants (et/ou tuteurs) et les stagiaires dans leur travail de traitement d'un problème, ou plutôt, dans la prise en compte d'une problématique, car bien souvent le temps manque pour faire un traitement complet.

Il va sans dire que cette démarche ne prétend pas à un quelconque absolu (elle ne fournira certainement pas une recette pour tous les problèmes rencontrés), mais propose seulement quelques pistes de recherche.

Dans la présentation qui va suivre, on peut parfois identifier le groupe à l'enseignant (et/ou tuteur) qui assure le suivi du stagiaire dans le traitement du problème et l'intervenant au stagiaire demandant conseil.

II - PRÉSENTATION DU PROBLEME

Le problème est présenté au groupe comme l'a fait le stagiaire lors de son entretien : par les documents 1 et 2 c'est-à-dire l'ensemble des données disponibles et un bref compte-rendu des objectifs et des conditions d'expérimentation.

Cette présentation sommaire amène une série de questions de la part du groupe, qui désire se faire préciser les conditions et les objectifs de l'étude, questions dont il est intéressant d'observer le déroulement chronologique.

2-1. Interrogations sur les conditions et les objectifs de l'étude

Questions posées par le tuteur (groupe) au stagiaire (intervenant) :

a) La première remarque porte sur le caractère incomplet des informations fournies : comment identifier les types de vaccin (cerveau et culture cellulaire), le plan expérimental n'est pas clair, il faut le préciser.

Le groupe veut ensuite avoir plus de détails sur les conditions expérimentales (le protocole) : quels sont les chiens vaccinés, à partir de quelle quantité d'anticorps un

chien est-il immunisé, connaît-on le résultat des vaccinations quant à l'immunisation contre la rage.

b) Puis le groupe s'interroge sur les objectifs précis de l'étude, par exemple pourquoi a-t-on pris des chiens de deux origines (rurale ; urbaine), comment sont-ils choisis ; quelles comparaisons veut-on faire et pourquoi ?

2-2. Quelques indications complémentaires

Réponses du stagiaire au tuteur :

a) Ce sont des chiens "sains" (non atteints de la rage) qui sont vaccinés, par des sérums préparés à partir de cerveaux de chiens enragés ou de cultures cellulaires de tissus contaminés (ce sont les deux types de vaccins ayant donné lieu aux deux essais). Rabisin et Rabigen se rapportent à des sérums type culture cellulaire (2ème essai), les autres sont des sérums issus de cerveaux (1er essai).

On ne connaît pas la relation exacte entre quantité d'anticorps titrée et immunité, mais ce que l'on sait, c'est que plus il y a d'anticorps, moins on a de chance de développer la maladie quand on est contaminé, d'où la recherche du titrage le plus élevé.

Les chiens vaccinés ont ensuite été soumis à contamination, mais on ne connaît pas encore les résultats (le stage s'est terminé avant). De toute façon, ça n'intervient pas dans le problème posé.

b_1) Les races de chiens vivant en ville sont différentes de celles vivant en milieu rural, c'est pour cela que l'on a prélevé 2 échantillons simples au hasard en zone urbaine et en zone rurale pour le 1er essai. Pour le 2ème essai, on ne connaît pas l'origine des chiens mais on sait que l'échantillonnage est aléatoire et simple dans une zone géographique connue.

* Pour le 1er essai : deux échantillons prélevés aléatoirement, en milieu rural et urbain respectivement, puis répartis ensuite entre 1 ml et 5 ml de façon aléatoire ; le déséquilibre observé étant dû à la mortalité.
* Pour le 2ème essai : un échantillon de 70 chiens réparti de façon aléatoire en 2 x 35, la mortalité a mené aux effectifs de 32 et 34.

b_2) Les objectifs de l'étude sont de déterminer les combinaisons (type ; doses ; formulations) donnant les meilleures réponses sur les populations rurales et urbaines.

2-3. Remarques (de l'intervenant)

Le stagiaire possède tous les éléments nécessaires au traitement, il a cependant des difficultés à poursuivre. Pourquoi ?
Mettons en lumière les principales raisons.

a) Les objectifs de l'étude n'ont pas été "traduits" en méthodes statistiques, même si à un moment donné on parle de comparaisons.

b) Le stagiaire a du mal à **identifier les composants des protocoles** avec le vocabulaire statistique : "facteurs", "modalités", et non sans raison, car il est en présence de deux expérimentations différentes statistiquement parlant (les deux "essais"), même si pour le vétérinaire ça n'est qu'un seul et même "ensemble".

Notons ici l'importance du vocabulaire, dont la maîtrise sera capitale dans la mise en œuvre des méthodes (cf. l'atelier de J. PAVY).

c) Les méthodes du programme de base de statistique inférentielle (de BTS), si tant est qu'il les ait déjà abordées, ne concernent que des comparaisons 2 à 2, ou ne font intervenir qu'un seul facteur ; sur des structures de données très simples. Or, ici, les structures (notion qu'il sera nécessaire de définir plus précisément) ne sont plus triviales, ce qui explique qu'il ne les identifie pas clairement.

d) Par contre, à côté de ces éléments posant problème, l'identification de la réponse à l'expérimentation est suffisamment claire : on ne s'intéresse qu'au titre dosé chez les chiens vaccinés, et dont on connaît bien les caractéristiques techniques, il s'agit de la variable UI.

e) D'une façon plus générale, le cas concret abordé dans cet atelier met en évidence une lacune importante : l'absence de méthodologie du traitement. Certaines méthodes sont acquises (parfois réduites à de simples calculs mécaniques), mais le stagiaire ne sait pas identifier, dans son cas concret, les situations nécessitant telle ou telle méthode.

Et pour cause, l'enseignement classique est basé sur le contrat implicite que le modèle (souvent aussi implicite) pris comme hypothèse, est toujours présent et évident. Dans la majorité des exercices d'école, les structures de données sont simples et évidentes.

On aura remarqué que nous parlons déjà de traitement *(sensu lato)*, dans la partie intitulée "présentation du problème". Ça n'est pas fortuit, car nous ne réduisons pas le "traitement" à la simple application des méthodes statistiques (ou même pire, à la seule mécanique calculatoire).

Le traitement (*sensu lato*) pourrait se décomposer en plusieurs phases :

- la présentation (qui n'est d'ailleurs pas indépendante des méthodes qui seront appliquées par la suite),
- la mise en forme,
- le choix des méthodes statistiques (modèles),
- leur mise en œuvre,
- l'analyse des résultats,
- la rédaction,

le tout suivant une méthodologie faisant intervenir plusieurs disciplines (techniques ; informatique ; statistiques).

Il serait intéressant de mener une réflexion concrète sur la nécessité de la mise au point et de l'enseignement d'une telle méthodologie (nécessairement pluridisciplinaire, rejoignant en cela, sur le fond, les propos de la conférence introductive de Mr PIEDNOIR).

III - LES PROPOSITIONS DE TRAITEMENT (sensu lato)

3-1.Les propositions du groupe (stagiaires-élèves)

A ce point de la présentation le groupe a estimé avoir suffisamment d'éléments pour proposer quelques méthodes statistiques : analyse de la variance, comparaisons de moyennes.

L'intervenant a posé la question de la mise en œuvre de ces méthodes, question restée en suspens. Cet embarras illustre bien quand et comment se pose le problème de la méthodologie.

A ce point de "l'analyse" du problème, les propositions de méthodes arrivent trop tôt. En effet, la présentation faite précédemment est notoirement insuffisante :

- les différents **objectifs** n'ont pas été suffisamment **détaillés en termes "opérationnels"** (nous verrons plus loin),
- les différentes **structures de données**, en relation avec les objectifs et les outils, n'ont pas été **établies**.

Il manque une étape capitale **entre la présentation et la mise en œuvre**, c'est celle que nous avions appelé plus haut la "**mise en forme**".

Il aurait été intéressant, dans un premier temps, de poursuivre l'atelier sous cette forme "informelle", afin de mettre à jour les autres points "clés" du traitement en identifiant ainsi quelques difficultés, pour ensuite reprendre l'ensemble des éléments, dans le cadre plus formel d'une méthodologie (stratégie de traitement). Les contraintes de durée de l'atelier nous ont obligé à proposer de suite une stratégie.

3-2.Propositions pour une stratégie de traitement

Le tuteur suggère de travailler sur les points suivants qui peuvent constituer les différentes étapes d'une stratégie de traitement (qui n'est certainement ni unique, ni la meilleure) :

RESTAURATION DU PROTOCOLE	**PRÉSENTATION**	1.Identifier les différentes "expérimentations" (au sens statistique du terme), les facteurs et les modalités qui y interviennent respectivement (on peut appeler cela le protocole). 2.Proposer un schéma récapitulatif clair.
	MISE EN FORME	3.Proposer une première structure (tableau) pour les données, en précisant qu'elle ne sera pas forcément la meilleure pour les traitements ultérieurs. 4.Traduire les "buts" (objectifs) de l'étude en objectifs opérationnels précis, et puis tenir compte des informations dont on dispose pour préciser les objectifs réalisables en réalité.

5.Déterminer les méthodes statistiques (modèles) applicables pour atteindre les objectifs précités.
6.Mettre en œuvre ces méthodes, ce qui nécessitera aussi, très certainement, de restructurer les données.
7.Analyser les résultats ; valider le modèle et éventuellement choisir de nouvelles méthodes (modèles).
8.Rédiger le rapport.

On notera que cette stratégie diffère sensiblement de celle proposée sur les documents généraux présentés à La Rochelle, qui suivait rigoureusement la démarche scientifique expérimentale, soit l'ordre : objectifs - modèle - données - traitements - analyses - éventuel retour au modèle - conclusions.

En effet, dans notre pratique, il n'est pas rare que le seul élément objectif dont nous disposions au départ soit les données. Les étapes 1, 2 et 3 ont donc pour but de "restaurer" le protocole, base sur laquelle travaillera le stagiaire.

3-3.Application au cas concret

Nous proposons une ébauche, élaborée à partir du cas concret abordé précédemment, en mettant en lumière quelques points clés qui sont à l'origine des principales difficultés des stagiaires.

ETAPES 1, 2 ET 3

Le document 2 structure les données et permet de visualiser le protocole expérimental, en y intégrant quelques informations récoltées dans la présentation du problème.

On y remarque les 2 essais, c'est-à-dire, les deux expérimentations (au sens statistique), les dispositifs déséquilibrés , les deux facteurs du premier essai : origine (2 modalités ou niveaux) et dose (2 modalités ou niveaux) ; le facteur unique du 2ème essai qui est la formulation (2 modalités).

Il est aussi très important de remarquer que cette structure ne pourra pas être utilisée par la suite pour les traitements, ou autrement dit, que les fichiers utilisés pour les traitements n'auront pas cette forme. Celle-ci sera déterminée une fois choisis les méthodes statistiques et les logiciels mis en œuvre.

Restauration du protocole

L'expérimentateur a réalisé deux expériences :

- l'une permettant d'étudier l'influence des deux facteurs, origine des chiens vaccinés et dose de vaccin ayant chacun deux modalités : Rural ; urbain et 1 ml ; 5 ml, sur la réponse UI.
- l'autre permettant d'étudier l'effet du facteur formulation, à deux modalités, rabigen et rabisin.
- l'unique réponse analysée est le taux d'immunisation mesurée par un dosage d'anticorps exprimé en UI.
- les répétitions n'étant pas identiques, nous avons un dispositif déséquilibré.
- l'échantillonnage est aléatoire et simple (cf. 2-2 b_1).
- les autres détails opératoires ont été précisés auparavant (cf. 2-2).

Les deux essais mènent classiquement aux deux plans expérimentaux :

- 1er essai : plan à deux facteurs fixes avec répétitions déséquilibrées.
- 2ème essai : comparaison à partir de deux échantillons.

On remarquera que le choix de ces plans revient très rigoureusement au choix d'un modèle pour le traitement qui suivra (modèle linéaire de l'analyse de la variance), et donc d'une méthode de traitement : l'analyse de la variance.

ETAPE 4

Nous précisons, à nouveau, que cette étape devrait normalement intervenir en premier dans la démarche de résolution d'un problème.

Nous allons présenter successivement la demande des vétérinaires puis les objectifs réalisables avec les données et les informations disponibles.

- La demande

Les objectifs des vétérinaires sont :

1- déterminer le "type" de vaccin induisant la meilleure immunisation. En termes statistiques : existe-t-il une différence significative entre l'immunisation induite par les deux "types" de vaccins et si oui quel est le meilleur ?
2- déterminer la meilleure dose, ou : existe-t-il une différence significative entre les deux doses et si oui quelle est la meilleure ?
3- déterminer la meilleure formulation, ou : existe-t-il une différence significative entre les deux formulations et si oui quelle est la meilleure ?
4- les réactions sont-elles différentes selon l'origine des chiens ? ou : existe-t-il une interaction significative entre l'origine des chiens et les autres facteurs ?

- **Les objectifs réalisables avec des informations disponibles** ou de la nécessité de faire la différence entre les problèmes que l'on se pose, les problèmes qui se posent, et ceux auxquels on peut répondre, afin d'éviter le risque de 3ième espèce.

Ce sont les objectifs 2 et 3 avec les précisions suivantes :

* Les résultats de l'objectif 2 ne seront valides que pour les vaccins de type "cerveau"; les résultats de l'objectif 3 ne seront valides que pour les vaccins de type "culture cellulaire" et pour les deux uniquement dans les régions sur lesquelles a porté l'échantillonnage.
* L'objectif 4 n'est réalisable que dans le premier essai, puisque l'on ne connaît pas l'origine des chiens dans le deuxième essai.

Par contre, l'objectif 1 n'est pas réalisable de façon simple, il sera donc mis de côté.

Les conditions d'échantillonnage sont toujours très importantes pour la validité des résultats. L'échantillonnage aléatoire et simple est celui le plus utilisé dans les méthodes classiques. Bien souvent, cette condition n'est pas respectée, ou bien les conditions d'échantillonnage sont mal connues.

C'est ici qu'il convient de dire avec force qu'un habillage mathématique, aussi rigoureux soit-il, ne donnera jamais de sens à une mesure inepte.

Les tests non paramétriques basés sur les permutations n'ont pas besoin de l'échantillonnage aléatoire dans la population, mais simplement de la repartition aléatoire des modalités des facteurs entre les individus. Bien sûr les conclusions n'auront pas le même degré de généralité.

Mais n'est-il pas préférable d'obtenir des conclusions moins générales plutôt que des conclusions fausses ?

ETAPE 5

Cette étape va consister à déterminer les méthodes statistiques les mieux à même d'atteindre les objectifs fixés à l'étape 4. Notons que le choix d'une méthode correspond en toute rigueur, toujours, au choix d'un modèle mathématique, et qu'il est important de l'avoir présent à l'esprit pour en faire bon usage.

La tendance des stagiaires est de se précipiter (et le mot n'est pas trop fort) sur l'inférence statistique (les tests), en négligeant complètement les aspects description et validation .

Il nous semble capital de toujours **effectuer une description complète et attentive** que l'on peut appeler **exploration des données**, avant de mettre en œuvre d'autres méthodes.

Tukey, dans l'E.D.A., nous propose toute une panoplie d'outils graphiques simples qui complètent les quelques méthodes classiques. Ces outils étant peu mathématisés, ils peuvent être enseignés même à partir des classes de B.T.A.

- ✶ On détaillera d'abord très précisément les descriptions à effectuer : que décrit-on et comment?

  En effet, se pose non seulement le problème des outils de description, mais aussi et surtout le problème des différents groupes à décrire. Il faut les identifier, regrouper les données puis les décrire. Or, là aussi, les stagiaires sont peu formés à ce travail d'identification des structures et de gestion des données (les exercices d'école sur la description portent en général sur un seul groupe).
- ✶ Une fois la description effectuée et après avoir éventuellement réalisé les opérations qu'elle a pu suggérer (correction de données, transformation de variables ...) viennent les méthodes de l'inférence statistique.

Application :

A partir du cas concret nous servant d'exemple, il faut identifier les 6 groupes (les 6 colonnes du document 8) : 4 groupes pour le premier essai, et 2 groupes pour le deuxième essai. Nous distinguons groupes et échantillons : les 4 groupes du 1er essai sont issus de 2 échantillons indépendants, alors que les 2 groupes du 2ème essai sont issus d'un seul échantillon. La notion de structure des données et de sa maîtrise prennent ici toute leur importance.

- ✶ Les descriptions faites sur les 4 groupes du premier essai figurent sur les documents 4 et 5. Nous y avons présenté quelques outils de l'analyse exploratoire : le branche et feuille ; l'histogramme ; la boite à pattes.
- ✶ Il nous semble instructif de commenter les avantages respectifs des différents outils graphiques, tant dans le fond (que représentent-ils?) que dans la forme (que **voit**-on?).

Quant aux méthodes d'inférence que nous pouvons proposer, ce sont **l'analyse de la variance à 2 facteurs fixes** pour le premier essai (objectifs 2 et 4) et **la comparaison à partir de 2 groupes** pour le deuxième essai.

- ✶ Est-il utile de rappeler que l'analyse de la variance est basée sur les hypothèses du modèle linéaire, de la normalité des résidus, de leur indépendance et de l'homogénéité de leur variance ; autant d'**hypothèses qu'il sera impératif de valider dans le courant de l'analyse**.
- ✶ Les documents 6 et 7 illustrent certaines de ces méthodes que nous commenterons plus loin, et au risque de nous redire, nous précisons que nous ne proposons ici que quelques indications ne pouvant en aucun cas faire figure de traitement complet.

ETAPE 6

L'étape de la mise en œuvre des analyses est, de loin, celle qui pose le plus de problèmes aux stagiaires, à tel point que l'on en voit souvent certains réaliser tous les calculs à la calculette. Et pour cause, c'est un travail qu'ils n'ont que rarement effectué dans leur scolarité (nous parlons essentiellement des B.T.S.).

Les difficultés sont de plusieurs ordres :

- ✶ Il faut déterminer la nouvelle structure des données qui permettra leur exploitation par une chaîne de traitement. Cette structure va dépendre des méthodes statistiques utilisées, et de la chaîne de traitement utilisée. Seule l'expérience permet de trouver rapidement la meilleure structure.
- ✶ Il faut ensuite créer le fichier informatique des données, donc mettre en œuvre un logiciel pour ce faire, sachant qu'il n'est pas du tout judicieux (parfois impossible) de faire la saisie à partir des modules de saisie des chaînes de traitement statistique.
- ✶ Il faut mettre en œuvre une chaîne de traitement, dont on connait la validité des algorithmes utilisés.

Sur ce vaste sujet, à peine abordé, on peut lire l'intéressant article de la revue "Statistiquement vôtre" n°1 de Sept. 92. Est-il utile, encore, de signaler, qu'ici informatique et statistiques doivent étroitement coopérer ? Encore la pluridisciplinarité !

Ce que l'on peut dire, en simplifiant beaucoup, c'est que le choix d'une chaîne de traitement s'opère, non pas tellement sur les méthodes statistiques elles-mêmes, car maintenant presque tous les logiciels proposent "l'artillerie" (!!!) classique au grand complet, mais surtout sur :

- la **souplesse** (convivialité) et les **possibilités** qu'elle offre dans **la gestion des données,** essentiellement dans la sélection des "groupes", sous forme de requêtes,
- la **souplesse** dans l'utilisation des méthodes, sous forme de procédures indépendantes, que l'on peut enchaîner,
- la **qualité des graphiques réalisés**, et la qualité de présentation,
- **"l'intelligence" des sorties numériques**.

L'atelier de J.PAVY nous a bien montré les difficultés de cette mise en œuvre et la nécessité de former les stagiaires à une méthodologie d'utilisation de ces outils.

Application :

Il est intéressant de noter la différence entre les 2 structures de données présentées dans les documents 2 et 3. Seule, celle du document 3 permet l'exploitation par une chaîne statistique. **Le passage de l'une à l'autre n'a rien de trivial.** C'est d'ailleurs une des principales difficultés rencontrées par les stagiaires, qui se situe, contrairement à ce que l'on dit souvent, beaucoup moins au niveau du "mode opératoire" des logiciels, que dans l'identification des structures des données, induites par le protocole expérimental.

ETAPE 7

Nous ne ferons que quelques commentaires sur l'analyse des résultats présentés dans les documents 4 à 7 ; notre propos portant moins sur la méthode que sur la méthodologie. Nous n'avons présenté qu'une partie des analyses faites sur l'essai n° 1.

Il est intéressant de comparer les différents outils de description de la distribution de la variable UI dans les 4 groupes. Les 2 outils les plus "parlants", à notre avis, étant le branche et feuille et les boites à pattes.

La forte dissymétrie des distributions et la présence de valeurs fortement éloignées de chaque groupe, nous mènent à proposer l'utilisation d'une transformation. **L'échelle des transformations de Tukey** nous permet de choisir la transformation logarithme décimal (on obtient alors la variable UILOG) dont le résultat **"visuel"** est satisfaisant (doc. 5).

Afin de montrer le bien fondé et le résultat de la transformation, nous avons effectué 2 analyses sur les variables, non transformée UI et transformée UILOG (doc. 6 et 7).

L'analyse (ANOVA) de la variable non transformée (UI) n'est pas validée, puisque ni la normalité, ni l'homogénéité des variances des résidus ne sont respectées ! L'analyse de UILOG est validée, et l'effet dose presque significatif (P~0,06), alors qu'il ne semblait pas l'être du tout d'après l'analyse de UI (P~0,1).

Notons au passage que la seule analyse du tableau d'analyse de la variance est insuffisante. À la limite, il vaut mieux faire une bonne description graphique, que de se limiter à cette seule analyse.

ETAPE 8

Cette étape n'a pas du tout été abordée dans l'atelier, faute de temps. Nous proposons 3 parties pour la réalisation du rapport :

- ✶ Elaboration du plan, qui doit comprendre l'introduction dans laquelle est présenté le problème à traiter, ses tenants et ses aboutissants, la présentation du protocole, avec matériel et méthodes, l'analyse des résultats, les discussions et la conclusion.
- ✶ Elaboration de la synthèse des analyses et des discussions dans laquelle les résultats doivent être présentés et commentés brièvement et clairement (voir problèmes de rédaction avec l'enseignement de français).
- ✶ Mise en page du rapport qui nécessite de bien positionner les commentaires, les sorties chiffrées et les graphiques, de façon à ce que la lecture en soit facilitée.

Les possibilités actuelles de rediriger toutes les sorties des chaînes de traitement vers des fichiers textes ou graphiques que l'on peut ensuite reprendre, par exemple, sous ***Windows*** et ses logiciels, permettent de faire des présentations claires et rapides.

Les documents 5 à 7 sont des documents bruts, directement sortis de la chaîne de traitement, mais qui permettent déjà d'élaborer un début de présentation, en regroupant les graphiques et les sorties chiffrées. L'ensemble devrait être repris et intégré à une synthèse rédigée.

IV - EN GUISE DE CONCLUSION

Nous avons essayé de montrer que les méthodes statistiques ne sont qu'une partie d'un tout, la démarche de traitement d'un problème.

La profusion des méthodes, des moyens de calculs et des logiciels rendent encore plus nécessaire l'apprentissage d'une méthodologie : contrairement à une croyance répandue, l'apprentissage de méthodes supplémentaires ne remplace pas celle d'une stratégie de leur utilisation.

Il ne faudrait jamais oublier, qu'une réponse, même exacte, à une question partielle ou mal posée, peut constituer une tromperie, ou conduire à des erreurs grossières.

Cette démarche est fondamentalement pluridisciplinaire car doivent y intervenir les "techniciens", les informaticiens et les statisticiens, et si possible une personne ressource sachant faire le lien (communication !!!) entre tous les protagonistes.

Chaque traitement étant un cas particulier, seule l'expérience de la pratique des traitements permet d'acquérir le recul nécessaire à une synthèse et un enseignement pratique et efficace.

Enfin, il existe actuellement, pour se former à la méthodologie et aux méthodes, des propositions de la branche enseignement du Ministère de l'Agriculture, parmi lesquelles :

- un didacticiel interactif sur les statistiques, recouvrant le programme BTS (agricole), en licence mixte : VIDEOSTAT,

Je remercie J.F. Pichard ; R. Tomassonne ; H. Joannes ; A. Mélan d'avoir bien voulu lire le document et pour leurs remarques judicieuses.

DOC.N° 1

PROBLEME 1 : RAGE 1 - AUBERT##

Vétérinaire sans frontière a effectué, au cours d'une mission au Népal, des essais de vaccination des chiens contre la rage, afin de comparer différents vaccins.

Des chiens "sains" (non atteints de la rage) sont vaccinés, par des sérums préparés à partir de cerveaux de chiens enragés ou de cultures cellulaires de tissus contaminés (ce sont les deux types de vaccins ayant donné lieu aux deux essais). Rabisin et Rabigen se rapportent à des sérums type culture cellulaire (2ème essai), les autres sont des sérums issus de cerveaux (1er essai).

Les races de chiens vivant en ville sont différentes de celles vivant en milieu rural, c'est pour cela que l'on a prélevé 2 échantillons simples au hasard en zone urbaine et en zone rurale pour le 1er essai. Pour le 2ème essai, on ne connaît pas l'origine des chiens mais on sait que l'échantillonnage est aléatoire et simple dans une zone géographique connue.

- Pour le 1er essai : deux échantillons prélevés aléatoirement, en milieu rural et urbain respectivement, puis répartis ensuite entre 1 ml et 5 ml de façon aléatoire ; le déséquilibre observé étant dû à la mortalité.
- Pour le 2ème essai : un échantillon de 70 chiens réparti de façon aléatoire en 2 x 35, la mortalité a mené aux effectifs de 32 et 34.

Les chiens vaccinés ont ensuite été soumis à contamination.

Les mesures d'efficacité sont effectuées 18 mois après la vaccination, en dosant par la méthode RFFIT, un anticorps qui assure la protection contre le virus. L'unité est l'UI (unité internationale).

Normalement, la quantité d'anticorps produite, augmente avec la quantité de vaccin introduite, dans certaines limites biensûr.

Les résultats obtenus figurent sur le document n° 2.

Le but des essais est de comparer les deux formulations, les deux doses, les deux types, et de déterminer s'il existe une relation entre origine des chiens et efficacité de la vaccination.

DOC.N° 2

DOCUMENT N° 2 : RÉSULTATS DE L'EXPÉRIMENTATION RAGE 1

Résultats de 2 essais de vaccination de chiens contre la rage (origine des données : Véterinaires sans Frontière et Centre National d'Etude sur la Rage-Pixérécourt).

Mesures 18 mois après vaccination en **UI/ml** (Unité Internationale), dosage par la méthode RFFIT.

Type de vaccin (type de tissu ayant servi à la fabrication du sérum)	CERVEAU (1er essai)				CULTURE CELLULAIRE (2ième essai)	
Origine des chiens ayant été vaccinés	RURAL		URBAIN			
Dose puis formulation	1 ml	5 ml	1 ml	5 ml	rabigen	rabisin
Effectif testé	22	15	15	14	32	34

MESURES

UI	UI	UI	UI	UI	UI
0.20	0,12	0,08	0,14	0,12	0,30
0.22	0.13	0,13	0,14	0,21	0,37
0.27	0.13	0,21	0,23	0,26	0,38
0.30	0.20	0,37	0,23	0,29	0,40
0.35	0.26	0,37	0,23	0,30	0,59
0.36	0.31	0,37	0,37	0,37	0,60
0.48	0.32	0,40	0,37	0,42	0,73
0.50	0.37	0,44	0,45	0,48	0,80
0.53	0.50	0.53	0,54	0,51	0,80
0.54	0.51	0,70	0,65	0,56	0,91
0.56	0.70	0,98	0,68	0,58	0,94
0.61	0.74	1,05	0,70	0,67	1,05
0.68	0.94	1,94	0,70	0,68	1,10
0.70	1.17	2,54	1,05	0,70	1,10
0.70	3.16	6,96		0,79	1,17
0.80				0,91	1,24
0.87				1,10	1,29
1,18				1,20	1,30
1.35				1,45	1,40
1.70				1,70	1,80
1.70				1,86	1,83
2.00				1,90	1,90
				2,00	2,10
				2,06	2,20
				2,12	2,25
				2,48	2,40
				2,60	2,68
				3,45	3,14
				6,17	3,26
				11,88	4,04
				12,60	5,35
				33,40	8,98
					9,44
					17,50

DOC.3

File AUBERT11 12/ 2/92

row	LIEU	VACCIN	DOSE	UI	UILOG
1	RURAL	CERVEAU	1ml	0.61	-0.215
2	RURAL	CERVEAU	1ml	0.20	-0.699
3	RURAL	CERVEAU	1ml	0.27	-0.569
4	RURAL	CERVEAU	1ml	0.70	-0.155
5	RURAL	CERVEAU	1ml	0.35	-0.456
6	RURAL	CERVEAU	1ml	1.70	0.230
7	RURAL	CERVEAU	1ml	1.70	0.230
8	RURAL	CERVEAU	1ml	1.35	0.130
9	RURAL	CERVEAU	1ml	0.87	-0.060
10	RURAL	CERVEAU	1ml	0.68	-0.167
11	RURAL	CERVEAU	1ml	1.18	0.072
12	RURAL	CERVEAU	1ml	0.36	-0.444
13	RURAL	CERVEAU	1ml	0.70	-0.155
14	RURAL	CERVEAU	1ml	0.53	-0.276
15	RURAL	CERVEAU	1ml	0.48	-0.319
16	RURAL	CERVEAU	1ml	0.22	-0.658
17	RURAL	CERVEAU	1ml	0.80	-0.097
18	RURAL	CERVEAU	1ml	0.56	-0.252
19	RURAL	CERVEAU	1ml	2.00	0.301
20	RURAL	CERVEAU	1ml	0.54	-0.268
21	RURAL	CERVEAU	1ml	0.50	-0.301
22	RURAL	CERVEAU	1ml	0.30	-0.523
23	URBAN	CERVEAU	1ml	2.54	0.405
24	URBAN	CERVEAU	1ml	0.21	-0.678
25	URBAN	CERVEAU	1ml	0.08	-1.097
26	URBAN	CERVEAU	1ml	0.37	-0.432
27	URBAN	CERVEAU	1ml	1.94	0.288
28	URBAN	CERVEAU	1ml	0.37	-0.432
29	URBAN	CERVEAU	1ml	0.98	-0.009
30	URBAN	CERVEAU	1ml	0.44	-0.357
31	URBAN	CERVEAU	1ml	0.40	-0.398
32	URBAN	CERVEAU	1ml	6.96	0.843
33	URBAN	CERVEAU	1ml	1.05	0.021
34	URBAN	CERVEAU	1ml	0.53	-0.276
35	URBAN	CERVEAU	1ml	0.13	-0.886
36	URBAN	CERVEAU	1ml	0.37	-0.432
37	URBAN	CERVEAU	1ml	0.70	-0.155
38	RURAL	CERVEAU	5ml	0.37	-0.432
39	RURAL	CERVEAU	5ml	0.70	-0.155
40	RURAL	CERVEAU	5ml	0.94	-0.027
41	RURAL	CERVEAU	5ml	0.20	-0.699
42	RURAL	CERVEAU	5ml	0.31	-0.509
43	RURAL	CERVEAU	5ml	0.50	-0.301
44	RURAL	CERVEAU	5ml	0.32	-0.495
45	RURAL	CERVEAU	5ml	0.26	-0.585
46	RURAL	CERVEAU	5ml	0.74	-0.131
47	RURAL	CERVEAU	5ml	0.13	-0.886
48	RURAL	CERVEAU	5ml	0.51	-0.292
49	RURAL	CERVEAU	5ml	0.13	-0.886
50	RURAL	CERVEAU	5ml	1.17	0.068
51	RURAL	CERVEAU	5ml	3.16	0.500
52	RURAL	CERVEAU	5ml	0.12	-0.921
53	URBAN	CERVEAU	5ml	0.23	-0.638
54	URBAN	CERVEAU	5ml	0.45	-0.347
55	URBAN	CERVEAU	5ml	0.68	-0.167
56	URBAN	CERVEAU	5ml	0.23	-0.638
57	URBAN	CERVEAU	5ml	0.23	-0.638
58	URBAN	CERVEAU	5ml	0.70	-0.155
59	URBAN	CERVEAU	5ml	0.70	-0.155
60	URBAN	CERVEAU	5ml	0.54	-0.268
61	URBAN	CERVEAU	5ml	0.14	-0.854
62	URBAN	CERVEAU	5ml	1.05	0.021
63	URBAN	CERVEAU	5ml	0.37	-0.432
64	URBAN	CERVEAU	5ml	0.14	-0.854
65	URBAN	CERVEAU	5ml	0.37	-0.432
66	URBAN	CERVEAU	5ml	0.65	-0.187

DOC. 4

AUBERT 11 **REPRESENTATION EN BRANCHE ET FEUILLE**

Branche-et-feuille pour la variable UI ;
Modalités : LIEU = 'RURAL' et DOSE = '1ml' ;
Unité = 0.01 ; 1|2 représente 0.12 ;

3	2	027
6	3	056
7	4	8
[11]	5	0346
[11]	6	18
9	7	00
7	8	07
5	9	
5	10	
5	11	8
4	12	
4	13	5
HI		170,170,200

Branche-et-feuille pour la variable UI ;
Modalités : LIEU = 'RURAL' et DOSE = '5ml' ;
Unité = 0.01 ; 1|2 représente 0.12 ;

3	1	233
[5]	2	06
(3)	3	127
[7]	4	
7	5	01
5	6	
5	7	04
3	8	
3	9	4
2	10	
2	11	7
HI		316

Branche-et-feuille pour la variable UI ;
Modalités : LIEU = 'URBAN' et DOSE = '1ml' ;
Unité = 0.1 ; 1|2 représente 1.2 ;

2	0*	01
[6]	0T	2333
(3)	0F	445
[6]	0S	7
5	0o	9
4	1*	0
3	1T	
3	1F	
3	1S	
3	1o	9
HI		25,69

Branche-et-feuille pour la variable UI ;
Modalités : LIEU = 'URBAN' et DOSE = '5ml' ;
Unité = 0.01 ; 1|2 représente 0.12 ;

2	1	44
5	2	333
[7]	3	77
[7]	4	5
6	5	4
5	6	58
3	7	00
1	8	
1	9	
1	10	5

LEGENDE :

Colonne des unités où *, T,F,S,o symbolisent les différentes classes nous avons ici des classes d'étendue 0,2

Effectifs cumulés croissants et décroissants jusqu'à la classe médiane dont l'effectif est (3).

La somme des effectifs encadrés donne l'effectif de la série.

2	0 *	01
[6]	0 T	2333
(3)	0 F	445
[6]	0 S	7
5	0 o	9
4	1 *	0
3	1 T	
3	1 F	
3	1 S	
3	1 o	9
2	HI	25,69

L'unité est 0.1 et est l'unité de chaque "feuille".

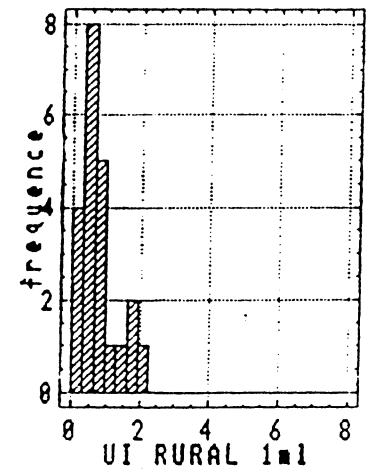
0* | 01 représente donc les mesures 0,08 et 0,13 arrondies à 0,0 et 0,1

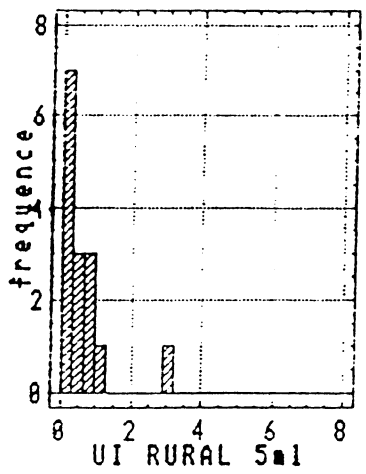
Chaque feuille est un chiffre représentant une mesure (observation).

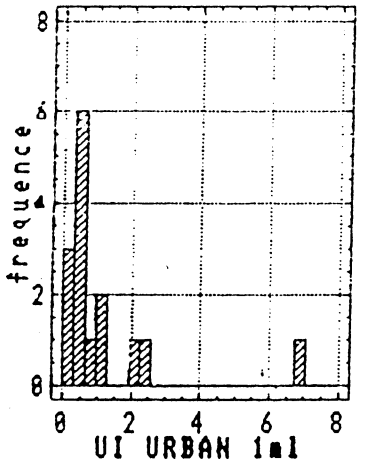
Mesures supérieures à Q3+1,5 fois l'intervalle interquartille

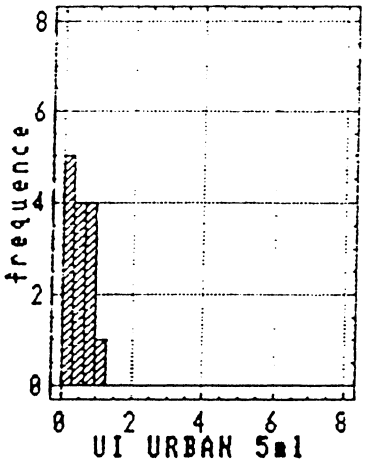
HI|25,69 représentent les mesures 2,54 et 6,96 arrondies à 2,5 et 6,9

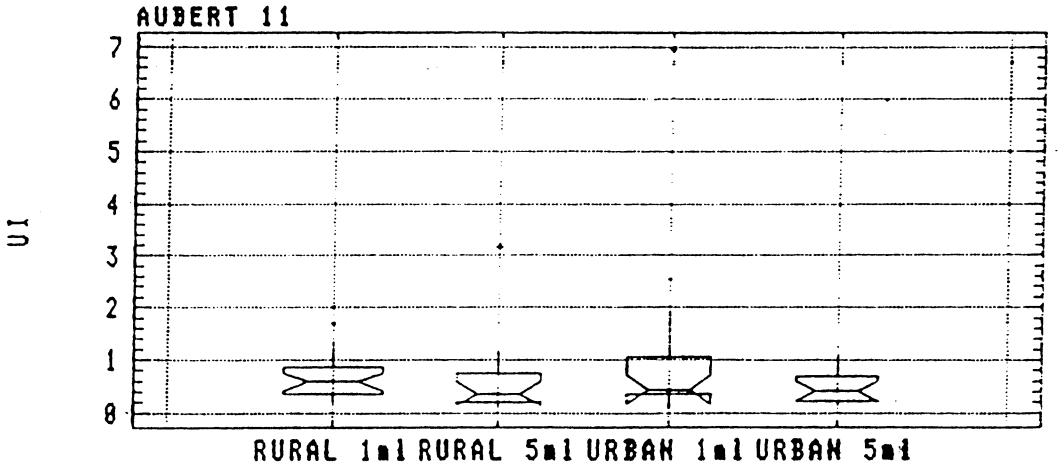
Doc. 5

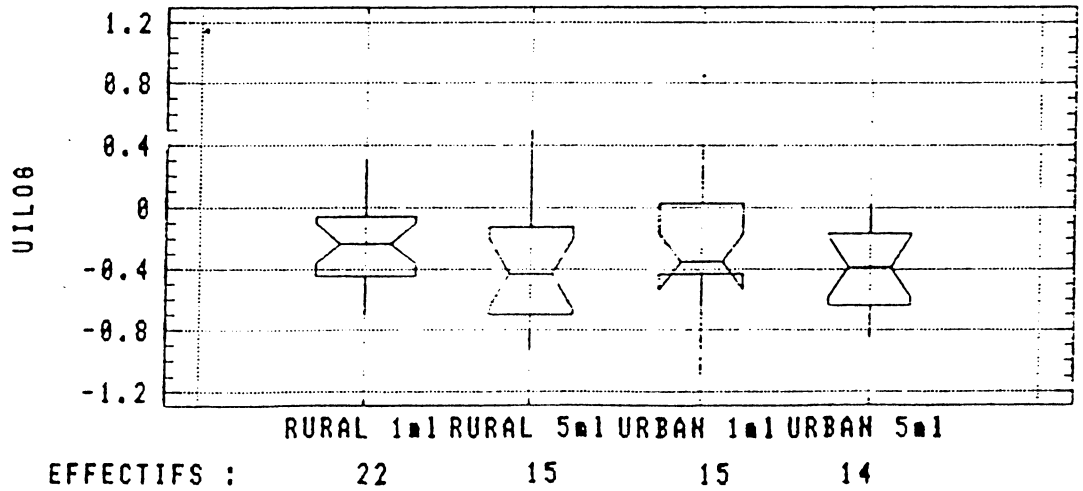






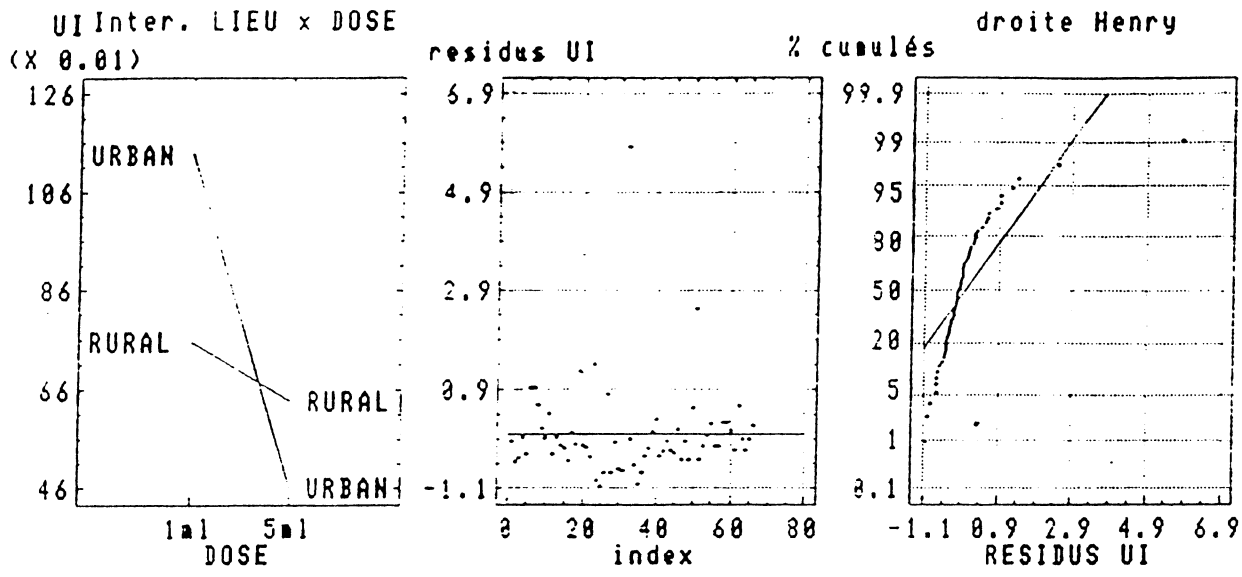


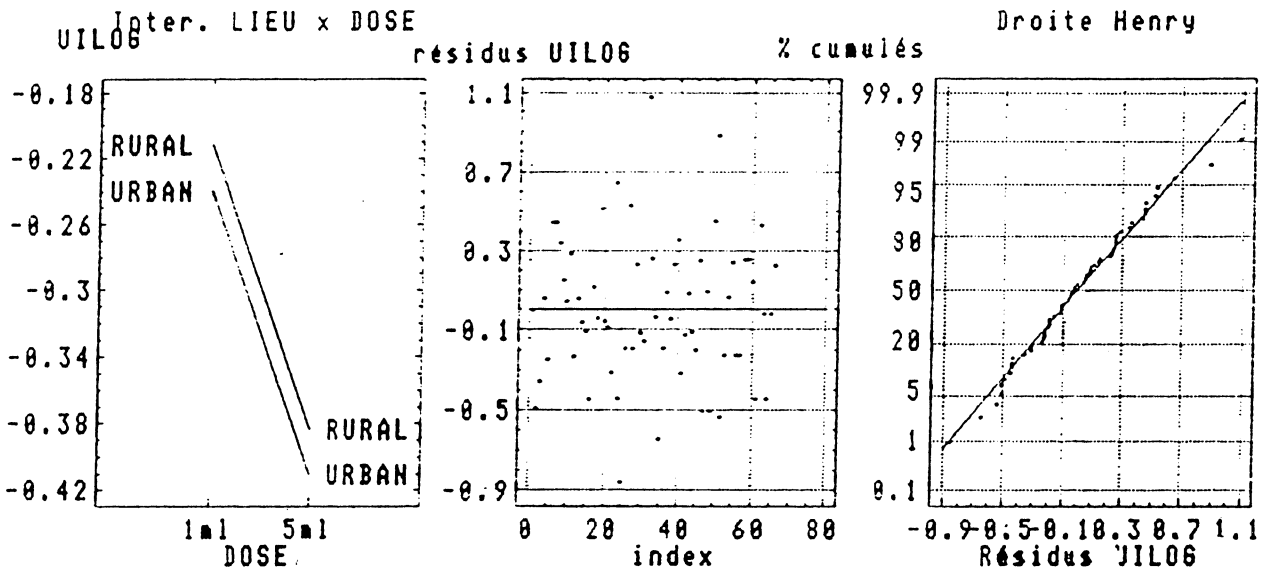




EFFECTIFS : 22 15 15 14

DOC. 6





Analysis of Variance for UI - Type III Sums of Squares

Source of variation	Sum of Squares	d.f.	Mean square	F-ratio	Sig. level
MAIN EFFECTS					
A:LIEU	0.1745367	1	0.1745367	0.188	0.6707
B:DOSE	2.5091332	1	2.5091332	2.702	0.1053
INTERACTION					
AB	1.2440697	1	1.2440697	1.340	0.2515
RESIDUAL	57.565365	62	0.9284736		
TOTAL (CORREC.)	61.168032	65			

0 missing values have been excluded.
All F-ratios are based on the residual mean square error.

Tests for Homogeneity of Variances : LIEU Cochran's C test: 0.804035 P = 0.000136321 Bartlett's test: B = 1.27125 P(15.1187) = 0.000100961 Hartley's test: 4.10296	**Tests for Homogeneity of Variances : DOSE** Cochran's C test: 0.805168 P = 0.000126664 Bartlett's test: B = 1.24341 P(13.7237) = 0.000211864 Hartley's test: 4.13264
Tests for Homogeneity of Variances : LIEU x DOSE Cochran's C test: 0.768937 P = 1.81247E-9 Bartlett's test: B = 2.21889 P(48.0724) = 2.05526E-10 Hartley's test: 42.6525	**Test de normalité des résidus :** Estimated KOLMOGOROV statistic DPLUS = 0.218703 Estimated KOLMOGOROV statistic DMINUS = 0.161769 Approximate significance level = 0.00362258

DOC. 7.

Analysis of Variance for UILOG - Type III Sums of Squares

Source of variation	Sum of Squares	d.f.	Mean square	F-ratio	Sig. level
MAIN EFFECTS					
A:LIEU	0.0121859	1	0.0121859	0.090	0.7684
B:DOSE	0.4696007	1	0.4696007	3.469	0.0673
INTERACTION					
AB	7.79684E-6	1	7.79684E-6	0.000	0.9941
RESIDUAL	8.3937951	62	0.1353838		
TOTAL (CORREC.)	8.8961768	65			

0 missing values have been excluded.
All F-ratios are based on the residual mean square error.

Table of Least Squares Means for UILOG

Level	Count	Average	Stnd. Error	95 Percent Confidence for mean	
GRAND MEAN	66	-0.3111155	0.0460131	-0.4031152	-0.2191158
A:LIEU					
RURAL	37	-0.2973108	0.0616023	-0.4204800	-0.1741415
URBAN	29	-0.3249202	0.0683664	-0.4616137	-0.1882267
B:DOSE					
1ml	37	-0.2254191	0.0616023	-0.3485884	-0.1022499
5ml	29	-0.3968118	0.0683664	-0.5335053	-0.2601184
AB					
RURAL 1ml	22	-0.2112652	0.0784462	-0.3681126	-0.0544179
RURAL 5ml	15	-0.3833563	0.0950031	-0.5733078	-0.1934048
URBAN 1ml	15	-0.2395730	0.0950031	-0.4295246	-0.0496215
URBAN 5ml	14	-0.4102674	0.0983375	-0.6068859	-0.2136488

Multiple range analysis for UILOG by DOSE

Method : 95 Percent Newman-Keuls

Level	Count	LS Mean	Homogeneous Groups
5ml	29	-0.3968118	X
1ml	37	-0.2254191	X

contrast	difference (* denotes a statistically significant difference)
1ml - 5ml	0.17139

Tests for Homogeneity of Variances : LIEU
Cochran's C test: 0.596463 P = 0.274319
Bartlett's test: B = 1.01916 P(1.19538) = 0.274248
Hartley's test: 1.47809

Tests for Homogeneity of Variances : DOSE
Cochran's C test: 0.557119 P = 0.520022
Bartlett's test: B = 1.00643 P(0.40345) = 0.525313
Hartley's test: 1.25794

Tests for Homogeneity of Variances : LIEU x DOSE
Cochran's C test: 0.441362 P = 0.0476843
Bartlett's test: B = 1.12857 P(7.29555) = 0.0630511
Hartley's test: 3.23911

Test de normalité des résidus UILOG :
Estimated KOLMOGOROV statistic DPLUS = 0.05419
Estimated KOLMOGOROV statistic DMINUS = 0.0510
Estimated overall statistic DN = 0.0541944
Approximate significance level = 0.990197

PRÉSENTATION DE VIDÉOSTAT :
DIDACTICIEL INTERACTIF D'INTRODUCTION AUX STATISTIQUES.

Didacticiel réalisé par les départements statistiques de **l'ACTA**, de l'**ENITA** de Bordeaux, de l'**INA-PG**, de l'**Institut de l'Élevage**, du **SCEES**, du **CFPPA** de Pixérécourt.

OBJECTIFS ; PRÉ-REQUIS ; MÉTHODES :

* C'est une introduction, aux statistiques descriptives, aux probabilités, et à l'inférence, conventionnelles et parfois un peu moins.
* Il a été conçu et réalisé par une équipe d'enseignents et de praticiens des statistiques.
* Il ne nécessite pas de pré-requis en statistique mais un niveau 2nde scientifique pour faire les initiations, et un niveau terminale scientifique pour utiliser l'ensemble du didacticiel.
* Il procède de la pédagogie par l'exemple, la formalisation étant présentée en fin de modules.
* Il est interactif pendant les cessions, et des exercices corrigés sont proposés en fin de modules.
* Il peut être utilisé comme support de cours avec des élèves, ou bien comme outils d'autoformation pour les enseignants. Les exemples présentés sont originaux, les analyses et commentaires pratiques sont très fournis.

CONTENU :

* Il sera composé de 7 modules indépendants, représentant entre 30 et 60 heures de formation :
  1- Graphisme et statistiques descriptives,
  2- Probabilités,
  3- Régression linéaire,
  4- Estimation et échantillonnage,
  5- Analyse d'un problème et analyse de la variance,
  6- Lois usuelles,
  7-Tests d'hypothèses.
  La version actuelle comprend les modules 1, 2, 3, 4 et 5. les modules 6 et 7 seront disponibles au début du 2ème semestre 1993.
* Chaque module est composé de nombreuses étapes qui peuvent constituer des points d'arrêt et de reprise des cessions.

MATÉRIEL ; LICENCE :

* Il nécessite au minimum un I386SX-VGA-couleur-souris.
* Il est en licence mixte au Ministère de l'Agriculture pour 1500FTTC au CNERTA à DIJON ; (à l'étude pour l'E.N.).
* Il est commercialisé par
  3P Informatique ; 4 rue R. Barthélémy, 92120 MONTROUGE ; (1)40 92 08 07.
* Tous renseignements complémentaires et démonstrations peuvent être demandés auprès des personnes suivantes :

  J.P. Desécures ; ACTA, 149, rue de Bercy 75595 PARIS CEDEX 12 ; (1)40 04 50 00
  C. Lopez ; Institut de l'Elevage, 149, rue de Bercy 75595 PARIS CEDEX 12 ; (1)40 04 52 69
  H. Raymondaud ; CFPPA de Pixérécourt, BP 10, 54220 MALZEVILLE ; 83 21 65 22.

ANNEXES

BIBLIOGRAPHIE

L'objectif de cette partie est de présenter un certain nombre d'ouvrages, la plupart en langue française, qui peuvent servir de référence théorique ou servir de source d'exemples et d'exercices, les niveaux mathématiques étant assez variés ; ces ouvrages se trouvent pour la plupart dans les bibliothèques scientifiques universitaires. Bien sûr, cette liste ne prétend pas être exhaustive. Quelques articles comportent une bibliographie, parfois commentée, dont les références ne sont pas toutes indiquées ci-dessous.

La mise en forme est de Pichard J.F. à partir des commentaires de MM. Boyera H., Fredon D., Pichard J.F., Raymondaud H.

Abboud N. et Audroing J.F. : Probabilités et Inférence Statistique, Coll. Supérieur/Economie, Nathan, 1989.

Ce livre de 352 p. est particulièrement destiné aux étudiants de Sciences Economiques, DEUG et licence. Il traite de l'introduction au calcul des probabilités : lois discrètes, continues et convergences stochastiques, puis de l'estimation et des tests d'hypothèse. Chaque chapitre est suivi d'exercices et de leurs solutions. La présentation est classique.

AFNOR : Recueil de normes françaises, statistiques, 6e éd. 1993.

tome 1 : vocabulaire, estimation et tests statistiques,
tome 2 : contrôles de statistiques de fabrication et d'acceptation,
tome 3 : traitement des résultats de mesures.

C'est le recueil de normes en statistique utilisées dans les milieux industriels. Les explications sont en général succinctes, c'est essentiellement l'algorithme de calcul qui est donné, avec quelques indications pour l'interprétation. Mais il y a toujours au moins un exemple traité pour éclairer la démarche.

A.P.M.E.P. : Analyse des données, 2 tomes, 1980.

Ces deux volumes (240 p. et 320 p.) sont formés d'articles écrits par différents auteurs. La plupart de ces articles est consacré aux méthodes de description statistique : l'analyse des correspondances, l'analyse en composantes principales et la classification. Deux articles concernent la régression. L'article de Pontier sur l'analyse discriminante commence par un exemple de test qui pose bien le problème.

Baillargeon G. : Probabilités, statistique et techniques de régression, éd. SMG, Québec, 1989 ; distributeur : Ellipses-Edition Marketing.

Ce gros ouvrage de 631 pages commence par une présentation classique de la statistique descriptive, du calcul des probabilités et des lois usuelles (un tiers du volume). Les 4 chapitres suivants portent sur l'estimation et les tests d'hypothèse. Les 5 derniers chapitres sont consacrés à la corrélation linéaire, la régression linéaire simple et multiple. Chaque chapitre est illustré d'exemples traités et est complété par de nombreux exercices. Le texte n'est pas très concis ; il y de nombreuses redites. Est-ce pour faire un bon poids ? Néanmoins, ce livre est assez facile à lire et peut être utile pour ses nombreux exemples.

Bertin J. : Le graphique et le traitement graphique de l'information, Flammarion, 1977.

Ce livre de 277 pp. a un peu vieilli, mais il contient beaucoup d'idées intéressantes et simples à mettre en oeuvre.

Bigot B. et Verlant B. : Mathématiques, statistique et probabilités, Fouchet, 1990.

Ce livre de 219 pages est destiné aux élèves en BTS "Comptabilité et gestion" et "Informatique de gestion". Pour chaque chapitre, une partie cours énonce sans formalisme les notions de base, suivie d'une partie TP et de nombreux exercices et problèmes qui sont, pour la plupart, corrigés.

Cailliez F. et Pages J.P. : Introduction à l'analyse des données, SMASH, 1976.

Cet ouvrage de 616 p. commence par une première partie (chap. 1 à 5) de rappels de mathématiques : espaces vectoriels, applications linéaires et matrices, distances, théorie générale des espaces euclidiens et des applications linéaires, en particulier les projecteurs. Puis vient un chapitre sur la statistique descriptive dans $\mathbf{R}^n$.
Les chapitres suivants portent sur différentes méthodes d'analyse des données qui utilisent l'appareil mathématique rappelé au début : analyse en composantes principales, régression linéaire, analyse canonique, analyse factorielle discriminante, des correspondances, classification automatique.

Céhessat Ronald : Exercices commentés de statistique et informatique appliquée, Dunod, 1976.

Livre d'exercices de 460 pages (avec corrigés) associé au traité de Lebart,... (cf. ci-dessous).

Cétama : Statistique appliquée à l'exploitation des mesures, Masson, 1e éd., 1976 ; 2e éd. 1986.

Cet ouvrage du Commissariat à l'Energie Atomique présente, à partir de nombreux exemples issus de données de laboratoire, l'utilisation de nombreuses méthodes statistiques : estimation, tests, régression linéaire, analyse de la variance,... (les explications théoriques sont un peu succinctes).
Cet ouvrage est complété par une partie tables statistiques assez complète portant sur la plupart des distributions utilisées pour les tests, ainsi que des abaques pour les intervalles de confiance et les risques de seconde espèce des tests.

Couty F., Debord J., Fredon D. : Probabilités et statistiques pour biologistes, Flash U, A.Colin, 1990.

Ce livre de 208 p., destiné aux Deug B, IUT, BTS biologistes et agricoles, présente d'abord quelques notions de calcul des probabilités (lois discrètes et continues) puis les principales méthodes statistiques que l'on peut aborder avec le public visé : estimation par intervalle, tests d'hypothèse, régression linéaire, analyse de variance et tests non paramétriques. Chaque chapitre commence par un bref résumé de cours décrivant l'objectif et l'algorithme de la méthode (pas de théorie), puis des énoncés d'exercices et de problèmes, dont beaucoup issus de textes d'examen, suivis des solutions.

Dagnélie Pierre, professeur à la Faculté des Sciences Agronomiques de Gembloux (Belgique), a publié plusieurs ouvrages tournés vers les applications agronomiques.

Dagnélie Pierre : Théorie et méthodes statistiques (2 volumes), Presses Agronomiques de Gembloux (Belgique), 1973, 1975.

Le tome 1 (378 p.) commence par la statistique descriptive (à 1, 2 et 3 dimensions) puis présente les notions de probabilité et de distributions théoriques. Il se termine par les principes de l'inférence

statistique : distributions d'échantillonnage, estimation et tests d'hypothèses.

Le tome 2 (463 p.) couvre les tests classiques, l'analyse de la variance (jusqu'à 3 facteurs contrôlés), les méthodes relatives à la régression et à la corrélation (dim 2 et 3), les transformations de variables et quelques méthodes non paramétriques.
l'ensemble fourmille d'exemples agronomiques et il existe aussi un fascicule d'exercices :

Dagnélie Pierre : Théorie et méthodes statistiques, exercices, (186 p.), 1981.
Et pour ceux qui s'intéressent à l'expérimentation réelle, un fascicule complémentaire intitulé :
Dagnélie Pierre : Principes d'expérimentation, (182 p.) présente les divers problèmes qui se posent à l'expérimentateur : choix des unités expérimentales, des observations, du plan d'expérience. Il présente divers protocoles expérimentaux et leur analyse statistique complète.
Dagnélie Pierre : Analyse statistique à plusieurs variables (362 p., 1975, 2e éd. 1986), traite dans le même esprit que T.M.S. des distributions multidimensionnelles, de la régression multiple, de la corrélation multiple et partielle, de l'analyse des composantes, de l'analyse factorielle, des problèmes de classification et de classement et de l'analyse de la variance à plusieurs variables.

Dagnélie Pierre : Statistique théorique et appliquée, 1992.

Diday E., Lemaire J., Pouget J., Testu F. : Eléments d'analyse des données, Dunod, 1982.

Ce traité de 460 p. porte sur différentes méthodes de l'analyse des données : classification automatique par hiérarchie et par partitions, les méthodes linéaires - régression linéaire et analyse factorielle -, l'analyse discriminante et les méthodes ordinales.

Droesbeke J.J. : Eléments de statistique, Ellipses-Marketing, 446 p., 1988.

Erickson B.H., Nosanchuk T.A. : Understanding data: An introduction to exploratory and confirmatory data analysis for students in the Social Sciences, Milton Keynes, Open University Press, 1977.

Un ouvrage introductif à la statistique exploratoire et confirmatoire, d'une qualité pédagogique exceptionnelle (pourquoi de tels ouvrages n'existent pas en français?!).

Feller, William : An introduction to probability theory and its applications, Wiley, New-York, tome I, 1e éd. 1950, 2e éd. 1957 ; tome II, 1966.

Malgré son âge déjà respectable, ce traité reste un des livres de référence sur la théorie des probabilités. On y trouve les démonstrations sous différentes conditions des théorèmes de convergence : loi faible et forte des grands nombres, théorème central-limit et extensions. De nombreux exemples illustrent la plupart des sujets du calcul des probabilités : conditionnement, chaînes de Markov, marches aléatoires, processus de Poisson, distributions infiniment divisibles,...
Pour beaucoup de résultats, des références bibliographiques sont indiquées, ce qui donne des points de repère historiques.
Les chapitres sont suivis de nombreux exercices et problèmes (quelques-uns assez difficiles), certains avec solutions à la fin du livre.

Foucart Thierry : Introduction aux tests statistiques : enseignement assisté par ordinateur, Technip, 1991.

Ce livre de 176 p. présente, de façon simple, les tests d'hypothèse les plus courants. Il est accompagné de disquettes pour mettre en oeuvre ces tests.

Fourgeaud C. et Fuchs A. : Statistique, Dunod, 2e éd. 1972.

Cet ouvrage traite de statistique mathématique, à un niveau assez élevé. Il introduit l'estimation et les tests à partir de la théorie de la décision.

Jaffard Paul : Initiation aux méthodes de la statistique et du calcul des probabilités, Masson, Paris, 1973.

Ce livre de la collection du CNAM présente, dans une première partie, le calcul des probabilités d'une façon simple. L'étude de la statistique s'appuie sur cette première partie et aborde l'estimation, les tests paramétriques et non paramétriques, le modèle linéaire : régression et analyse de la variance à 1 et 2 facteurs.

Jambu M. et al : L'exploration informatique et statistique des données, Dunod, Paris, 1989.

Le seul ouvrage en français qui, sur 505 pages, en consacre 60 à la présentation de quelques méthodes graphiques univariées et multivariées de l'E.D.A. (*). Le reste est de l'analyse des données, présentée assez pédagogiquement. Il aborde aussi le problème de la structuration des données et de leur gestion informatique. Il traite essentiellement de l'analyse des données : analyse conjointe, ACP, analyse des correspondances, classification et fait le lien avec le traitement informatique.

Hoel P.G. : Statistique mathématique, t. 1, Coll. U, A. Colin, 1991.

Ce livre de 304 pages est le tome 1 de la traduction de la 5e édition du traité *Introduction to Mathematical Statistics*, éd. J. Wiley & Sons, 1984. La première partie (chap. 2 et 3) donne les principaux résultats du calcul des probabilités et les lois usuelles discrètes et continues. La présentation est classique, essentiellement avec les jeux (pièces, dés, cartes) et les urnes.
La deuxième partie introduit les méthodes statistiques : estimation, tests, théorie de l'échantillonnage et les méthodes empiriques pour la corrélation et la régression. Les justifications théoriques et autres développements sont reportés au 2e tome.
Les différents chapitres sont émaillés de nombreux exemples complètement traités et chacun est suivi de plusieurs dizaines d'exercices qui ont presque tous une solution numérique en appendice. Cependant, la plupart de ces exercices sont des applications directes des résultats donnés dans le chapitre. Les plus intéressants sont ceux de la 2e partie (chapitres 4 à 7) sur les méthodes statistiques ; on peut y trouver des idées de sujets pour les élèves.
Le texte semble par moment peu précis et un peu flou ; peut-être est-ce dû à la traduction.

Lebart L., Morineau A., Fénelon J.P. : Traitement des données statistiques, méthodes et programmes, Dunod, 1982.

Ouvrage de 518 pages, très clair, agréable à lire, où chaque notion est illustrée par un exemple. Il couvre le calcul des probabilités (jusqu'aux théorèmes de convergence), son application au raisonnement statistique, les méthodes non paramétriques, le modèle linéaire (régression, analyse de la variance et de la covariance), les méthodes factorielles et de classification automatique.

---

*) E.D.A. = Exploratory Data Analysis.

Chaque chapitre a une notice bibliographique.
Seule critique : les programmes sont en Fortran et APL, langages peu utilisés en micro-informatique.

Montfort A. : Cours de probabilités, Economica, 1980.

Exposé complet et de bon niveau basé sur la théorie de la mesure.

Mosteller F., Tukey J.W. : Data Analysis and Regression, a second course in Statistics, Addison-Weysley, Reading, 1977.

Ouvrage impressionnant, passant en revue dans une perspective exploratoire les principaux outils et l'histoire de la statistique ; livre critique et très riche. Le deuxième cours en statistique - partie du titre - s'adresse certainement autant aux statisticiens pratiquant la statistique "classique" qu'à l'étudiant. Une réorientation statistique... après la statistique classique. Ce livre est, comme E.D.A. de Tukey, un classique.

Sanders D.H., Murph A.F., ENG R.J. : Les statistiques, une approche nouvelle, McGraw-Hill, Québec, 1984.

Ce livre de 450 pages présente de façon simple et humoristique les principales méthodes statistiques : estimation, tests d'hypothèse, analyse de variance, séries chronologiques, régression, méthodes non paramétriques. Chaque chapitre se termine par des questions de compréhension et des problèmes. De nombreuses remarques judicieuses.

Saporta Gilbert : Probabilités, analyse des données et statistique, Technip, 1990.

Cet ouvrage de 492 p. présente un panorama assez complet des méthodes de la statistique, aussi bien descriptive qu'inférentielle.
La première partie donne les outils probabilistes : variables et vecteurs aléatoires avec de nombreuses lois, et une introduction aux processus aléatoires.
La deuxième partie, intitulée "La statistique exploratoire", étudie la description à une variable, puis à 2 variables avec les mesures de liaison entre variables, enfin la description multidimensionnelle avec l'analyse en composantes principales, l'analyse canonique, l'analyse des correspondances, --- multiples et des méthodes de classification.
La troisième partie est intitulée "la statistique inférentielle" et commence par la théorie de l'échantillonnage. Le chapitre suivant traite de l'estimation : exhaustivité, estimation sans biais de variance minimale, méthode du maximum de vraisemblance, intervalle de confiance. Ensuite un chapitre sur les tests d'hypothèse, d'ajustement et de comparaison d'échantillons paramétriques ou non, puis l'analyse de la variance à 1 et 2 facteurs. Les 2 chapitres suivants portent sur la régression linéaire, simple et multiple. Cette 3e partie se termine par un chapitre sur l'analyse discriminante.
Cet ouvrage est complété par des tables statistiques pour les principaux tests.
On peut cependant regretter que ce livre contienne si peu d'exemples traités pour éclairer les différentes méthodes.

Tassi Philippe : Méthodes statistiques, Economica, Paris, 1985.

Ce livre porte essentiellement sur la statistique mathématique.
Après des rappels de calcul des probabilités, il présente le cadre formel de la décision statistique : règle de décision, principe minimax et principe bayésien, puis l'exhaustivité.
La 2e partie porte sur l'estimation : propriétés d'un estimateur et méthodes d'estimation. La 3e partie

traite des tests, avec l'optique décisionnelle et l'optique de Neyman et Pearson, puis présente les principaux tests classiques.
Quelques exemples éclairent les sujets abordés.

Tomassone R., Lesquoy E., Miller C. : La Régression, nouveaux regards sur une ancienne méthode statistique, Masson, 1983.

Un exposé de référence à partir d'exemples complètement traités.

Tufte E.R. : The Visual Display of Quantitative Information, Graphics Press, Chesshire, 1983.

Un "tour de force" (Tukey dixit) ; un ouvrage plein d'idées sur les graphiques, leurs qualités et défauts. De plus, l'ouvrage est beau à voir et n'est pas du tout technique. Toute personne intéressée par les graphiques doit connaître ce livre.

Tukey J.W. : Exploratory Data Analysis (E.D.A.), Addison-Weysley, Reading, 1977.

Le classique ! L'ouvrage est une collection de techniques très variées, des propositions d'analyse. Un livre à lire et à relire. Certaines parties sont très accessibles, d'autres se découvrent après la 3è, 10è... lecture.

Wonnacot Th.H., Wonnacot R.J. : Statistique: économie, gestion, sciences, médecine. Economica, 3è éd. 1988.

Une présentation simple et claire de la plupart des méthodes statistiques. De nombreux exemples traités, surtout orientés vers l'économie et les sciences humaines.

Wonnacot Th.H., Wonnacot R.J. : Statistique, exercices d'application, Economica, 4è éd. 1991.

Une mine d'exemples : plus de 900 p.

Brochures des IREM

Seules sont répertoriées ici les brochures récentes portant sur la statistique et le calcul des probabilités.

IREM de Besançon

- L'enseignement des statistiques et des probabilités en STS, 1990.

Cette brochure de 96 pages présente, sous forme de cours, les principaux éléments de statistique : calcul des probabilités et les lois usuelles, lois limites, estimation, tests d'hypothèses, fiabilité. Chaque chapitre est illustré d'exemples et d'exercices corrigés.

IREM de Clermont-Ferrand

- Simulation de quelques variables aléatoires réelles, Fleury G., 1984.

Quelques utilisations du générateur de nombres pseudo-aléatoires d'un micro-ordinateur pour la simulation de diverses lois.

IREM de Grenoble

- Probabilités et statistiques en Première, Ghesquière et Pariselle, février 1993.

IREM de Lorraine

- Statistiques, Bac Pro : Enel J., Leiritz F., 1990.

Ce fascicule, pour 1ère et Terminale Pro, traite des graphiques et résumés statistiques à 1 et 2 variables à partir de nombreux exemples. Un reproche : pas de table des matières ni de pagination.

IREM de Montpellier

- Un exemple d'application de la loi binomiale, collectif, 1984.

Utilisation de la loi binomiale pour des plans de contrôle statistique de réception avec courbe d'efficacité, développée à partir d'exemples.

IREM de Paris-Nord

Le groupe inter-IREM Lycées Techniques publie depuis plusieurs années des recueils de sujets avec corrigés des BTS de différentes sections. En particulier :

- Fiabilité (n° 48), document support de stages "Statistique et Probabilités pour les STS".

S'adresser : Université PARIS-NORD
C.S.P.- IREM, Groupe Lycées Techniques
Avenue Jean-Baptiste Clément, 93430 VILLETANEUSE.

IREM de Paris-Sud

- Statistiques, Parzysz B. et Szwed T., 1985.

Organisation des données ; séries temporelles, indices des prix ; statistiques à 1 et 2 variables, ajustement linéaire.

IREM de Rouen

- Ars Conjectandi de Jacques Bernoulli, traduction de N. Meusnier, 1987.
Traduction de la 4e partie de l'Ars Conjectandi où est établi la loi des grands nombres. Un texte historique de référence.
- Les enquêtes à questions nominales, Lannuzel B., Orange G., Pichard J.F., 1989.
Etude descriptive d'une enquête à questions nominales : tableau de contingence, tri-à-plat et tri croisé, mesures d'association entre caractères. Approche d'un traitement global par une typologie des caractères et une classification.
- Graphiques au collège, groupe Statistique, 1991.
Ce document soulève le problème de l'utilisation des graphiques en statistique. On propose un essai de classification des graphiques selon la nature de la population et du caractère. Quoique ce document soit orienté vers le collège, les réflexions qui y sont faites, en particulier sur l'histogramme, peuvent servir pour des niveaux ultérieurs.
- Les probabilités pour le lycée - 1, Sinègre L., 1992.
Compte rendu d'un stage destiné aux professeurs pour leur permettre de mieux dominer ce qu'ils ont à enseigner en Terminale ou en Prépa. Ce premier fascicule de 34 p. porte sur la combinatoire et l'utilisation des arbres, puis la formule de Bayes. Un 2ème fascicule est en préparation.

IREM de Strasbourg

- Enseigner les probabilités en classe de 1ère, octobre 1992.
Brochure de 111 p. conçu dans l'esprit du programme de 1991, avec plusieurs activités d'introduction aux probabilités, des exercices avec solutions abrégées et des réflexions sur la notion de probabilité.

Cassettes vidéo relatives aux probabilités et aux statistiques

Commentaires de D. Fredon, IREM de Limoges

Chroniques du hasard I et II, CNDP

Niveau : lycée
Contenu I : Pascal et le chevalier de Méré regardent la télévision en jouant au passe-dix. Que d'événements depuis qu'ils sont morts ! A l'évidence, le chevalier de Méré a besoin de quelques leçons de dénombrement.
Contenu II : Leçons de probabilité à Port-Royal où l'on se divertit en jouant à la marelle aléatoire. L'homme du XXe siècle est-il en train de faire le pari inverse de celui de Pascal ?
I : cassette VHS, 16 min, Réf. 002A5977 ; II : cassette VHS, 14 min, Réf. 002P5978
CNDP : 29, rue d'Ulm, 75230 PARIS Cédex 05.

La statistique et le vivant (le jeu de la science et du hasard), entretiens avec le professeur Daniel SCHWARTZ

Le contenu de cette cassette retient de la statistique, non pas les aspects techniques, mais la philosophie de ce qui est une nouvelle façon de penser... Pas de formules, pas de calculs.
INSERM, 101, rue de Tolbiac, 75654 PARIS Cédex 13.

APPROXIMATION BINOMIALE - NORMALE

UNE ILLUSTRATION DU THEOREME DE MOIVRE - LAPLACE

PICHARD JEAN FRANÇOIS
IREM DE ROUEN

On utilise souvent l'approximation de la loi binomiale par la loi normale (nom maintenant consacré donné à la loi de Laplace-Gauss, étudiée pourtant en premier par Daniel Bernoulli). Ceci est justifié par le théorème de Moivre-Laplace, qui est un cas particulier du théorème central-limit.

Cependant, quelques difficultés demeurent pour les étudiants, et quelquefois pour les utilisateurs de la statistique, qui se posent en particulier les questions suivantes :

- comment établir le lien entre la loi binomiale, qui est discrète, et la loi normale qui est continue ?
- à partir de quelle taille d'échantillon peut-on valablement utiliser cette approximation ?

Posée sous cette forme, cette dernière question est imprécise puisque la loi binomiale dépend de 2 paramètres, notés habituellement n et p, qui vont influer tous les deux sur les valeurs de probabilité. Il faut de plus préciser ce qu'on entend par approximation valable. La réponse sera différente suivant l'utilisation que l'on en fait : probabilité d'événement, test, intervalle de confiance.

L'application du théorème de Moivre-Laplace à une v.a. binomiale $X \in \mathcal{B}(n, p)$ permet d'obtenir des valeurs approchées pour les probabilités :

$$P[X=k] \approx P\left[\frac{k-1/2}{\sqrt{npq}} \le U \le \frac{k+1/2}{\sqrt{npq}}\right], \text{ pour n grand, } k \in \mathbf{N}, \text{ et où } U \in \mathcal{N}(0,1).$$

Pour établir le lien entre loi binomiale et loi normale, on va représenter la distribution de la v.a. X de loi $\mathcal{B}$ (n,p) par un histogramme. Pour cela, on associe à chaque valeur entière k de X un intervalle [k-1/2, k+1/2] de largeur 1 centré en cette valeur, la hauteur du rectangle correspondant étant P[X=k]. L'aire de tous ces rectangles est alors égale à 1, puisque c'est la somme des probabilités. L'association d'un intervalle de largeur 1 à une valeur entière est appelée correction de continuité.

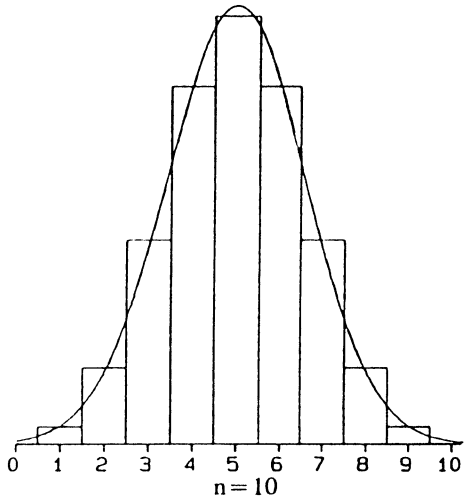
La comparaison de ces deux lois se fera en superposant le graphe de la densité de la loi normale à l'histogramme ainsi construit. L'aire totale sous cette courbe est aussi égale à 1. Les probabilités associées à une valeur de X sont d'une part l'aire du rectangle correspondant, et pour l'approximation normale d'autre part, l'aire de la surface sous la courbe densité ayant pour base l'intervalle considéré.

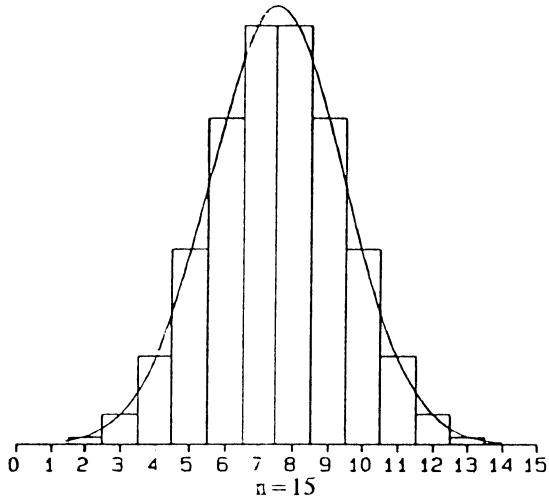
Afin de visualiser l'évolution de l'écart entre la loi binomiale et la loi normale, lorsque la taille de l'échantillon augmente, à valeur de p constante, les graphiques suivants sont tracés par rapport à la variable centrée réduite ; cela élimine l'effet taille.

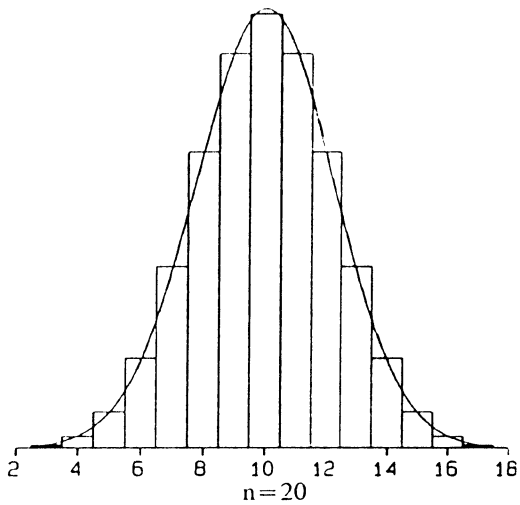
Pour caractériser le degré d'approximation, on a indiqué l'écart maximum entre les probabilités pour les valeurs de X et celles correspondantes de U, ainsi que la valeur de X où ce maximum est atteint. Cependant, pour le calcul approché de la fonction de répartition de X (les probabilités cumulées) à l'aide de celle de U, l'écart est plus grand pour certaines valeurs que l'écart maximum sur les probabilités individuelles.

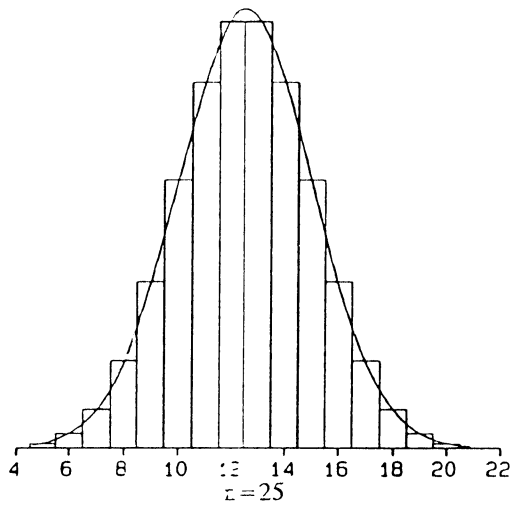
On voit ainsi que l'approximation est très bonne lorsque la loi est symétrique (p=1/2), même pour de petites valeurs de n ; alors que pour une loi fortement dissymétrique (p petit ou p voisin de 1), la taille doit être grande afin d'obtenir le même degré d'approximation.

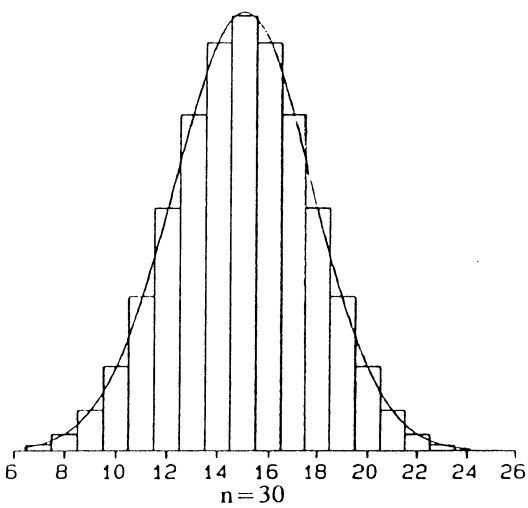
La convergence (en loi) d'une v.a. suivant une loi de Poisson $\mathcal{P}(\lambda)$, centrée réduite, vers la loi normale quand $\lambda \to \infty$ se visualise de la même façon en superposant l'histogramme correspondant à la loi de Poisson et le graphe de la densité de la loi normale.

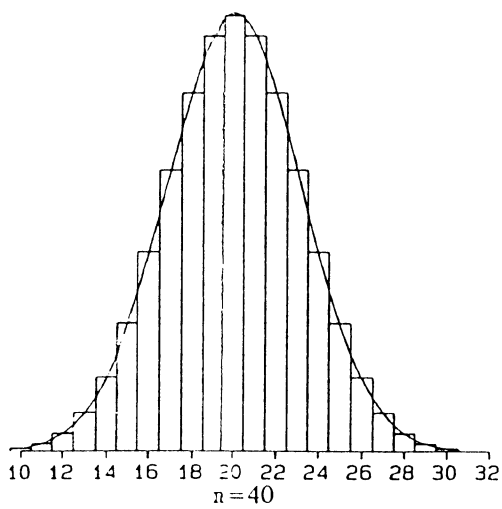




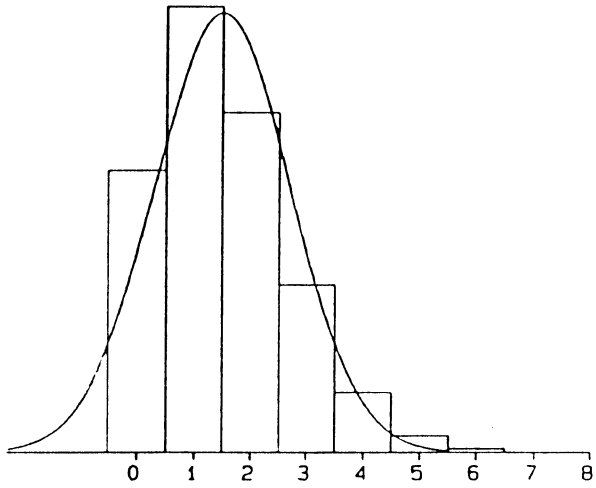




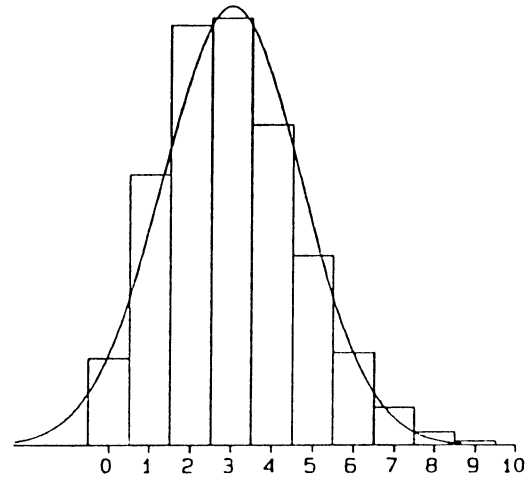




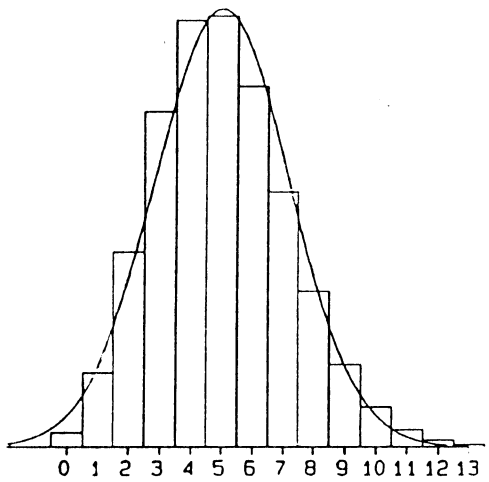
Approximation Binomiale-Normale p=0.5



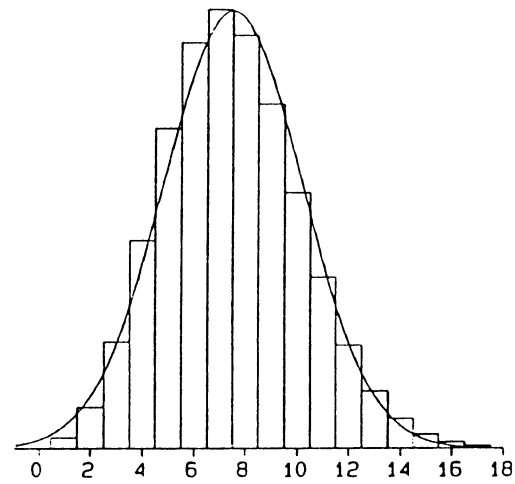
n=30, ecart max=6.05% pour k=0



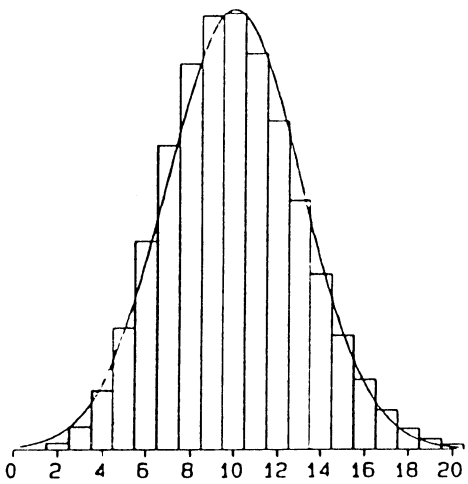
n=60, ecart max=2.95% pour k=2



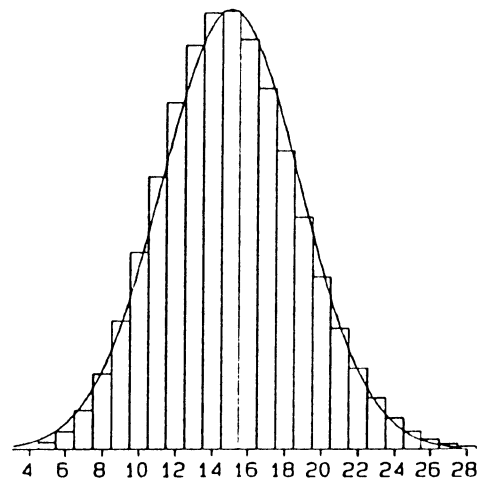
n=100, ecart max=1.96% pour k=3



n=150, ecart max=1.25% pour k=5

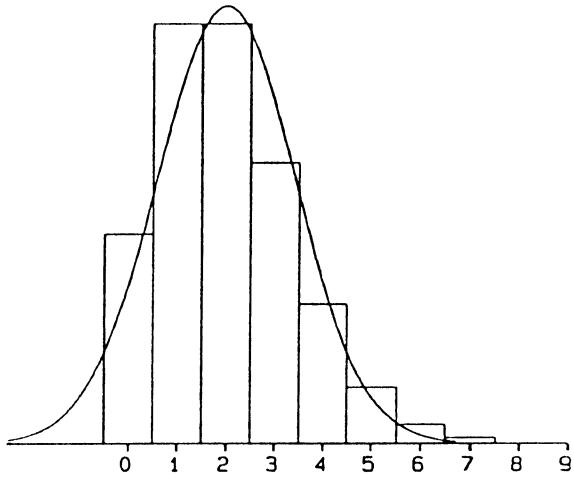


n=200, ecart max=0.91% pour k=8

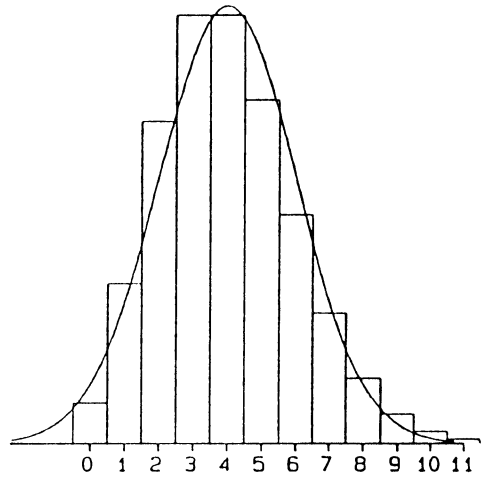


n=300, ecart max=0.63% pour k=12

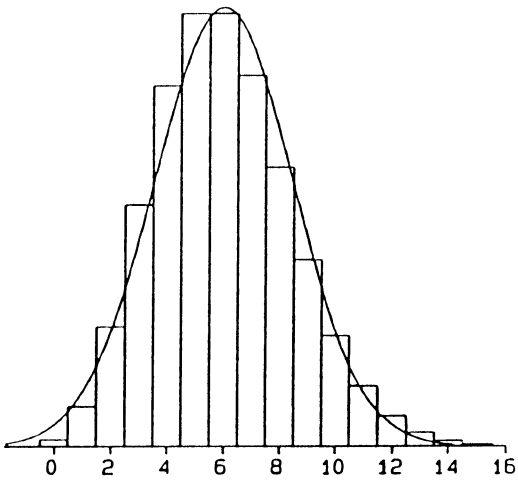
Approximation Binomiale-Normale : p=0.05



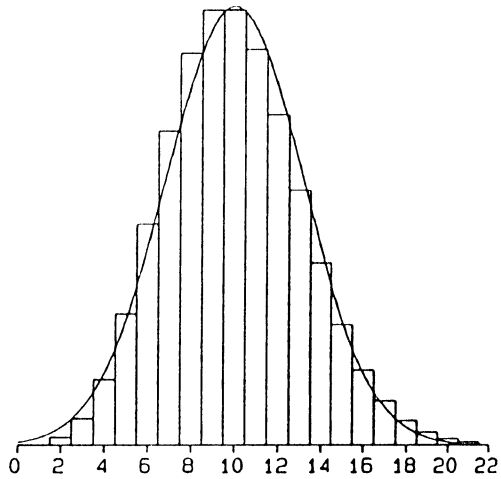
mu =2.0, ecart max=5.33% pour k=1



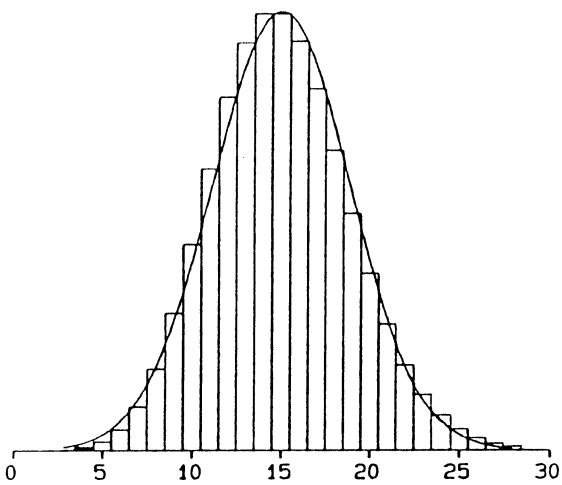
mu =4.0, ecart max=2.55% pour k=2



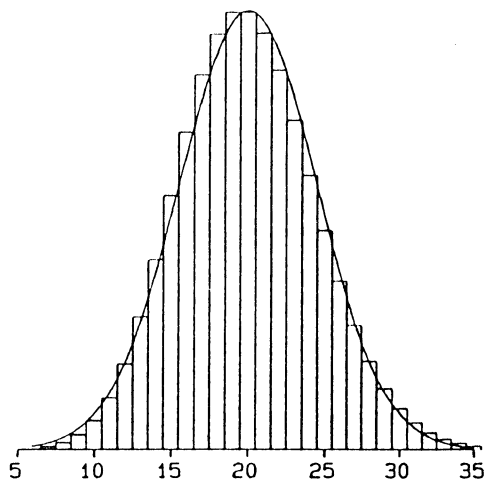
mu =6.0, ecart max=1.74% pour k=4



mu =10.0, ecart max=0.97% pour k=7



mu =15.0, ecart max=0.66% pour k=12



mu =20.0, ecart max=0.48% pour k=17

Approximation Poisson-Normale

EVALUATION DE L'UNIVERSITE D'ETE

PICHARD J.F.

L'université d'été "Statistique dans les formations technologiques", d'une durée de plus de 30 heures sur 5 jours, a abordé différents sujets concernant la statistique. Il s'agissait, en fin de période, de déterminer si cette session de formation avait rempli, vu du côté des participants, tout ou partie des objectifs que s'étaient assigné les membres du comité d'organisation et de la commission inter-IREM "Enseignement de la Statistique et du Calcul des Probabilités".

Pour ce faire, un questionnaire a été élaboré pendant la durée de l'université d'été, afin de déterminer les impressions immédiates des participants.

Le questionnaire, dont le texte se trouve dans les pages suivantes, comporte plusieurs parties, subdivisées elles-mêmes en plusieurs questions élémentaires. Tout d'abord des questions (1 à 5) de description du participant, puis des questions portant sur l'organisation matérielle et pédagogique de l'université d'été (6 et 7), ensuite des questions (8 et 9) sur la façon dont ont été appréciés les conférences et ateliers, enfin une question (10) sur l'opinion globale et sur les suites éventuelles à donner à cette université d'été.

L'étude qui suit (*) a uniquement pour ambition de donner une description des participants à l'université d'été, ayant répondu au questionnaire, et de leurs opinions concernant le déroulement de celle-ci. Le groupe des répondants ne constituait pas un échantillon aléatoire issu d'un ensemble plus vaste (même de l'ensemble des participants, quoiqu'il reproduit fidèlement la structure de cet ensemble par rapport aux questions sexe, âge et type d'enseignement) et toute extension à une population plus grande peut conduire à des conclusions aventureuses, voire erronées.

On a recueilli 42 réponses sur 47 participants : presque tous les participants ont répondu. En raison de la faible taille de l'échantillon, les relations entre variables sont difficiles à discerner, d'autant que pour certaines questions, une des réponses était fortement prépondérante. Le niveau de liaison entre variables n'est pas ici très fiable ; j'ai utilisé pour les croisements entre variables les profils lignes et la contribution au lien (I.C. Lerman).

Le groupe (des répondants) était formé pour 3/4 d'hommes, 20% d'enseignants de lycée agricole et la moitié dans la tranche d'âge de plus de 45 ans. La question sexe est non ou peu corrélée avec l'ensemble des autres. Les enseignants de lycée agricole sont plus jeunes que ceux de l'Education Nationale : 56% dans la tranche 40-45 ans et 33% de plus de 45 ans contre 21% de 40-45 ans et 54% de plus de 45 ans dans l'enseignement général.

Une première constatation intéressante : ceux qui ont étudié la statistique dans leur cursus universitaire (ils sont peu, 14%) ont aussi étudié le calcul des probabilités (plus de la moitié des participants), et plus les personnes sont jeunes, plus grande est la proportion de ceux qui ont étudié le calcul des probabilités et la statistique.

---

*) Cette étude a été réalisée avec le logiciel ANAKEST (auteur Pichard) et est basée sur les idées développées dans la brochure : Enquêtes à questions nominales par Lannuzel B., Orange G., Pichard J.F. ; IREM de ROUEN, 1989.

Par rapport à l'ensemble des réponses, on peut partager le groupe en 3 sous-groupes ayant des comportements spécifiques : celui des 23 professeurs de l'enseignement général ayant enseigné en S.T.S. (**), groupe noté TSG, celui formé des 10 participants n'ayant jamais enseigné en S.T.S. (noté TS0) et celui des 9 enseignants de lycées agricoles (noté TSA), qui ont tous en charge des S.T.S. depuis plus de 3 ans. En effet, en combinant les questions 4 et 5, on obtient une nouvelle variable, dont les modalités correspondent aux 3 groupes cités, qui devient la plus corrélée avec l'ensemble des autres.

Cependant, les opinions concernant l'organisation matérielle (6) et pédagogique (7) sont à peu près en même proportion parmi les sous-groupes indiqués ; c'est-à-dire que cette classe de questions est peu ou non corrélée avec les autres, avec quelques nuances : pour l'organisation de l'hébergement, 70% du groupe TSA étaient très satisfaits, mais seulement le tiers de TS0 ; pour l'organisation en ateliers, 70% de très satisfaits dans TSA et TS0, mais 40% de TSG. L'organisation en ateliers parallèles a recueilli entre 20% et 30% de très satisfaits, le choix était peut-être dur à faire et il faudrait prévoir de répéter les ateliers. Globalement, il y a eu peu de mécontents, la proportion des plutôt et très satisfaits est supérieure à 85% pour ces différentes questions, 74% pour ateliers en parallèle.

Le niveau théorique des ateliers est presque unanimement jugé convenable, mais le niveau des conférences (de certaines) a été jugé difficile par 40% de TS0 et TSG. Il faut noter à ce propos que 90% de TS0 ont étudié les probabilités dans leur cursus universitaire, contre 2/3 pour TSA et 1/3 pour TSG. Peut-être est-ce une question d'âge car 60% de TSG sont dans la tranche > 45 ans, contre 40% pour TS0 et le tiers pour TSA.

L'intérêt des thèmes statistiques abordés lors de cette université d'été était apprécié par les questions 8 et 9. La proportion des plutôt et très satisfaits varie entre 76% (régression linéaire) et 88 à 92% pour les autres thèmes, avec des variations suivant les sous-groupes. Ainsi, les méthodes bayésiennes ont été très appréciées par 80% de TSA et TS0, mais 65% de TSG ; la régression linéaire a été plutôt ou très appréciée par 65% de TSA et TS0, mais par 87% de TSG.

La distinction des sous-groupes indiqués est très nette dans la question 9 qui se décompose en 2 questions élémentaires 9.1 et 9.2

La question 9.1 sur l'exposé marquant pour la formation de TS a recueilli 80% de non-réponses dans TS0 ; les participants de TSG ont choisi l'exposé sur les tests à 52% et celui sur l'estimation à 26%, ceux de TSA ont choisi l'analyse de variance à 44% et les tests à 33%.

Pour la question 9.2 sur l'intérêt pour la formation personnelle, les participants de TS0 ont choisi les méthodes bayésiennes à 70%, puis l'analyse de variance à 20% ; ceux de TSG ont choisi les méthodes bayésiennes à 48%, puis l'analyse de variance à 30% ; ceux de TSA ont choisi les méthodes bayésiennes à 44% et se sont dispersés entre les autres exposés.

Pour la question 10.1 sur les attentes, les participants de TS0 et TSG s'estiment très ou plutôt satisfaits à 90%, ceux de TSA à 66% en raison des non-réponses (33%).

Je compléterais cette étude du questionnaire par des remarques que m'ont communiqué certains des participants. Cette université d'été portait sur l'enseignement de la statistique. L'apport théorique sur les méthodes statistiques a été jugé très satisfaisant par presque tous les participants, mais certains ont trouvé qu'il y avait trop de thèmes différents. Quand à la partie enseignement, elle n'a été abordée que dans quelques ateliers et certains auraient aimé lui voir consacrer une place plus importante.

**) S.T.S. = Section de Techniciens Supérieurs.

UNIVERSITE D'ETE
STATISTIQUE DANS LES FORMATIONS TECHNOLOGIQUES
LA ROCHELLE du 1.09.92 au 5.09.92

QUESTIONNAIRE AUX PARTICIPANTS

Entourez la réponse vous convenant

1 – SEXE : homme femme

2 – AGE : moins de 35 ans
de 35 à 40 ans
de 40+ à 45 ans
plus de 45 ans

3 – DANS VOTRE CURSUS UNIVERSITAIRE, AVEZ VOUS ETUDIE :

– Les probabilités oui non

– La statistique oui non

4 – TYPE D'ENSEIGNEMENT ASSURE:

général agricole

5 – AVEZ–VOUS DEJA ENSEIGNE DANS DES SECTIONS DE TS :

	jamais	1an	2ans	3ans	4ans et plus.
EN 89–90 ?	OUI	NON			
EN 90–91 ?	OUI	NON			
EN 91–92 ?	OUI	NON			

6 – ORGANISATION MATERIELLE : ETES–VOUS SATISFAIT DE :

6.1 L'HEBERGEMENT	très	plutôt	peu	pas du tout
6.2 REPAS	très	plutôt	peu	pas du tout
6.3 SALLES DE COURS ET ATELIERS	très	plutôt	peu	pas du tout
6.4 L'AFFICHAGE D'INFORMATION	très	plutôt	peu	pas du tout
6.5 LE SECRETARIAT	très	plutôt	peu	pas du tout
6.6 LA DEMI–JOURNEE DETENTE	très	plutôt	peu	pas du tout

6.7 AVEZ–VOUS DES PROPOSITIONS A FAIRE SUR CES SUJETS

7 - ORGANISATION PEDAGOGIQUE : ETES-VOUS SATISFAIT DE :.....

7.1 L'EMPLOI DU TEMPS :

durée des conférences	très	plutôt	peu	pas du tout
durée des ateliers	très	plutôt	peu	pas du tout
durée des pauses	très	plutôt	peu	pas du tout

7.2 L'ORGANISATION DE L'ENSEIGNEMENT :

conférences	très	plutôt	peu	pas du tout
ateliers	très	plutôt	peu	pas du tout

7.3 L'ORGANISATION EN ATELIERS PARALLELES :

très plutôt peu pas du tout

7.4 LE NIVEAU THEORIQUE ETAIT-IL :

pour conférences et exposés	trop difficile	convenable	simple
pour ateliers	trop difficile	convenable	simple

8 - THEMES : AVEZ-VOUS ETE INTERESSE PAR LES THEMES SUIVANTS :

8.1 ESTIMATION	très	plutôt	peu	pas
8.2 TESTS D'HYPOTHESES	très	plutôt	peu	pas
8.3 METHODES BAYESIENNES	très	plutôt	peu	pas
8.4 REGRESSION LINEAIRE	très	plutôt	peu	pas
8.5 ANALYSE DE LA VARIANCE	très	plutôt	peu	pas
8.6 SUJETS D'ATELIERS	très	plutôt	peu	pas

9 - QUEL EST L'EXPOSE OU LA CONFERENCE QUI VOUS A LE PLUS MARQUE :

9.1 POUR LA FORMATION DE TS.....

9.2 POUR VOTRE FORMATION PERSONNELLE .....

COMMENTAIRES :

10 - FINALEMENT :

10.1 PAR RAPPORT AU LIBELLE DU B.O., VOS ATTENTES RELATIVES A CETTE UNIVERSITE D'ETE ONT ETE SATISFAITES : très plutôt peu pas

10.2 S'IL Y AVAIT UNE UNIVERSITE D'ETE EN 1994, QUELS SUJETS AIMERIEZ-VOUS VOIR TRAITER.....?

10.3 SERIEZ-VOUS, DANS CE CAS, INTERESSE A ANIMER :

UNE CONFERENCE : .....?

UN ATELIER :.....?

ADRESSE ADMINISTRATIVE DES INTERVENANTS (enseignement supérieur)

BENINEL F.	IUT de NIORT, quai Dugueselin – 79000 NIORT
CELLIER Dominique	Dépt.de Maths, U.F.R. Sciences, BP 118 – 76135 MONT SAINT AIGNAN
COURCOUX Philippe	ENITIAA. Chemin de la Gérandière – 44072 NANTES Cedex 2
FOUCART Thierry	IUT d'Orléans, domaine universitaire, rue d'Issoudun – BP 6729 45067 ORLEANS Cedex
FREDON Daniel	Directeur de l'IREM de Limoges, Université de Limoges, 123, avenue Albert Thomas – 87060 LIMOGES Cedex
GRAS Régis	IRMAR, Université de Rennes I, Campus de Beaulieu – 35042 RENNES Cedex
HENRY Michel	IREM de BESANÇON, Faculté des Sciences, La Bouloie 25030 BESANÇON Cedex
JANVIER Michel	IREM de MONTPELLIER, Université Sciences et Techniques du Languedoc, Place E. Bataillon – 34095 MONTPELLIER Cedex
MERIGOT Michel	IREM de NICE, Parc Valrose, avenue Valrose – 06034 NICE Cedex
PICHARD Jean François	responsable pédagogique de l'université d'été, responsable de la commission inter-IREM Statistique et Probabilités directeur de l'IREM de Rouen, Université de Rouen, 1 rue Thomas Beckett, 76135 MONT SAINT AIGNAN
PIEDNOIR Jean Louis	Inspecteur Général, 110 rue de Grenelle 75007 PARIS

Secrétariat

Mme LAMARCHE D.	IREM de Rouen, 1 rue Thomas Beckett – 76135 MONT SAINT AIGNAN

ADRESSE ADMINISTRATIVE DES PARTICIPANTS

NOM	ETABLISSEMENT
ANGELIQUE F.	LEGTA de NANCY-Pixérécourt – 54220 MALZEVILLE
BARBE D.	Lycée Jean Jaures – 71200 LE CREUSOT
BENOIST M.	Lycée E. Branly, 33 rue du Petit Bois – 94000 CRETEIL
BIGOT B.	Lycée Roosevelt, 10 rue Roosevelt – 51100 REIMS
BLANDIN J.P.	Lycée René-Josué Valin, rue Henri Barbusse – 17023 LA ROCHELLE Cedex
BOYERA H.	Lycée J. Audiberti, Boulevard Wilson – B.P. 218 – 06604 ANTIBES
BRIN P.	Lycée Technique E. Branly, 33 rue du Petit Bois – 94000 CRETEIL

BRUNET D.	L.T. Rouvière, Quartier St Musse – B.P. 1205 – 83000 TOULON
BURG P.	LEGTA d'OBERNAI, Bd Europe – 67210 OBERNAI
CASTRES L.	Lycée d'Altitude – 05105 BRIANÇON
CHAPUT B.	Lycée Edouard Herriot, rue de la Maladière – 10300 SAINTE SAVINE
CHAVIGNY G.	Lycée Jules Haag, 1 rue Labbé – 25000 BESANÇON
CHAVIGNY M.	Lycée Jules Haag, 1 rue Labbé – 25000 BESANÇON
COUTY-FREDON F.	Lycée Jean Giraudoux – 87 BELLAC
DANGLETERRE D.	LEGTA de DOUAI, 158 rue Motte Julien – 59500 DOUAI
DELZONGLE F.	L.T.I. Gustave Eiffel, 101 avenue du Président Wilson – 94 CACHAN
DUVET G.	Lycée des Eucalyptus, avenue des Eucalyptus – 06000 NICE
FAURE J.C.	LEGTA, route de st Hilaire – 11000 CARCASSONNE
FOURNIER	IUT de LA ROCHELLE, rue de Roux – 17026 LA ROCHELLE Cedex
GAUMET J.P.	LEGTA Le Robillard, Lieury – 14170 SAINT PIERRE sur DIVES
GIRARD J.C	IUFM de Lyon, 90 rue de la Richelandière – 42100 SAINT ETIENNE
GUILLEMOT M.	Lycée L. de Vinci, rue E. Branly – 77000 MELUN
JACQUOT Y.	Lycée Technique du Génie Civil, rue Laugier – 06600 ANTIBES
LABROUE F.	Lycée F. Villon, 10 avenue Marc Sanguier, 75004 PARIS
LE BERRE P.	Lycée Le Corbusier, 44 rue Réchossière – 93300 AUBERVILLIERS
LE COZLEER J.L.	LEGTA Le Robillard, Lieury – 14170 SAINT PIERRE sur DIVES
LE MENN A.	I.U.T. La Rochelle – Dépt Biologie, rue de Roux 17026 LA ROCHELLE Cedex
LEGRAND G.	Lycée Privé St Joseph, 39 rue du Transvaal – 21000 DIJON
MAILLES A.	Université de Provence, avenue R. Schuman – 13621 AIX-en-PROVENCE Cedex
MARIANI R.	Lycée Technique Ph. de Girard, Route de Tarascon – 84000 AVIGNON
MELLAN A.	EIL-LEGTA, BP 141 – 74805 LA ROCHE SUR FORON
NICAULT R.	Lycée Polyvalent, Route de St Pair – 50400 GRANVILLE
NOEL A.	Rectorat, 25 rue de Fontenelle – 76000 ROUEN
PARNAUDEAU J.M.	LEGTA de Venours – 86480 ROUILLE
PAVY J.	LEGTA Le Robillard, Lieury – 14170 ST PIERRE-sur-Dives
PHILBERT M.	L.T. L. Armand, 173 Bd de Strasbourg – 94130 NOGENT SUR MARNE
PRADIN J.	LEGTA de Moulins – 03000 NEUVY
RAYMONDAUD H.	C.F.P.P.A. de Pixérécourt – 54220 MALZEVILLE
REBORD C.	Lycée C. DUPUY – 43003 LE PUY EN VELAY
RIVES M.F.	Lycée Technique, Bd Berthelot – 13200 ARLES
ROCHETAING G.	Lycée Classique de Cocody – ABIDJAN – COTE D'IVOIRE
SAINSOT J.M.	Lycée Raoul Dantry – 87000 LIMOGES
SAINT PIERRE G.	Lycée Edouard Branly, rue du Petit Bois – 94000 CRETEIL
STAREK M.	Rectorat SAIA, 20 Bd d'Alsace-Lorraine, B.P. 2609 – 80026 AMIENS Cedex
SUPRIN Y.	Lycée Vallée du Cailly, rue du Petit Aulnay – 76250 DEVILLE-lès-ROUEN
URDAMPILLETA V.	E. N. d'Industrie Laitière et Agro-alimentaires, B.P. 49 – 17700 SURGERES
VARLOT C.	LEGTA de CHALONS SUR MARNE – 51460 SOMME VESLE
VERON D.	Lycée Edouard Branly, rue du Petit Bois – 94000 CRETEIL

Titre : Actes de l'Université d'été de Statistique

La Rochelle, 1-5 septembre 1992.

Auteurs : intervenants de l'université d'été

Résumé :

Ces Actes rassemblent les textes des conférences et les comptes rendus des ateliers qui se sont déroulés lors de cette session. Ils s'adressent surtout aux professeurs enseignant dans les Sections de Techniciens Supérieurs, aussi bien à finalité professionnelle que ceux de l'enseignement agricole.

Les sujets traités sont essentiellement de la Statistique inférentielle : méthodes d'estimation classique et bayésienne, tests d'hypothèse, régression linéaire, analyse de variance.

Mots clés :

statistique inférentielle
estimation ponctuelle et par intervalle
tests d'hypothèse
statistique bayésienne
régression linéaire
analyse de variance

Editeur : J.F. PICHARD, IREM de ROUEN

Directeur de publication : J.F. PICHARD

Date : avril 1993 ; 240 pages, format A4, prix : 95 F

n° ISBN : 2-86239-046-1 ; dépôt légal : 2ème trimestre 1993