REVUE DE STATISTIQUE APPLIQUÉE

E. ROY
J. P. PAGES
F. FAGNANI

Quelques aspects élémentaires de la pratique statistique en biologie clinique : problèmes posés par la possibilité de stockage et de traitement de données nombreuses

Revue de statistique appliquée, tome 24, n° 2 (1976), p. 5-18
<http://www.numdam.org/item?id=RSA_1976__24_2_5_0>

QUELQUES ASPECTS ÉLÉMENTAIRES DE LA PRATIQUE STATISTIQUE EN BIOLOGIE CLINIQUE : PROBLÈMES POSÉS PAR LA POSSIBILITÉ DE STOCKAGE ET DE TRAITEMENT DE DONNÉES NOMBREUSES (1)

E. ROY *, J. P. PAGES * et F. FAGNANI **

INTRODUCTION

Il faut reconnaître que le statisticien face à des problèmes à résoudre – comme ceux qui se posent actuellement en biologie clinique – ne dispose pas a priori d'une stratégie "idéale", lui permettant d'évoluer à coup sûr et en dépensant le moins d'énergie possible vers "la solution" ou la certitude d'une absence de solutions. Il ne dispose que d'un certain nombre de principes généraux issus de son expérience, de ses connaissances techniques et des possibilités matérielles qui lui sont offertes. Il est enfin largement soumis à l'influence du milieu professionnel dans lequel sa pratique s'insère et dont il doit assimiler le mode de pensée et apprécier le degré d'achèvement des connaissances afin de définir au mieux son action.

La biologie clinique, qui regroupe l'ensemble des utilisations médicales des examens de laboratoire, constitue un domaine très large d'application des méthodes statistiques. L'automatisation des examens de laboratoire a entraîné un essor considérable de ces méthodes comme instrument de diagnostic, de surveillance, et à présent, de dépistage des maladies. Mais, comme il arrive souvent en pareil cas, l'existence de l'instrument a, en quelque sorte, précédé la réflexion et les études sur ses modes d'utilisation socialement souhaitables. En effet, dans la pratique courante de la médecine, les résultats des examens de laboratoire sont pratiquement ininterprétables en dehors des déviations très importantes dues à la présence de pathologies avérées. Or, on veut souvent leur faire dire beaucoup plus : qui ne connaît les infortunés, affligés de la "maladie du cholestérol" auxquels on a imposé de draconiens régimes ?

La notion de "valeurs de référence" a été introduite de façon récente afin d'engager une réflexion et une recherche enfin sérieuses sur ce qu'on appelait auparavant les "valeurs normales". Dans cette approche, on se

* Laboratoire de Statistique et d'Etudes Economiques et Sociales – Département de Protection – Commissariat à l'Energie Atomique.

** Chargé de recherche INSERM

(1) Article remis en Janvier 1974, révisé en Janvier 1976.

Mots-Clés : **Valeur de référence. Nombre de données/hypothèses. Pratique de l'analyse de données.**

propose de définir un ensemble de valeurs-limite, de seuils (le problème étant de même nature que celui du contrôle de qualité) à partir desquels, pour un individu donné défini par toutes ses caractéristiques jugées pertinentes, pour une certaine technique de mesure et pour un objectif médical donné (diagnostic, dépistage, etc.), on peut décider d'une attitude médicale précise.

Pour définir ces valeurs de référence en médecine préventive, on assiste actuellement à la constitution, ou plutôt à la tentative de constitution, de sortes de banques de données individuelles ; celles-ci sont souvent conçues comme un moyen de différer la réalisation de travaux d'ordre épidémiologique impossibles à mettre en œuvre a priori faute de moyens (ou d'idées). Outre le problème de l'intérêt intrinsèque que présente l'existence de ces fichiers rassemblés sans objectifs préalables très clairs, d'autres questions se posent au niveau du traitement lui-même.

En effet, les données dont on dispose sont nombreuses et il n'est pas rare que les échantillons soient grands voire très grands, les effectifs dépassant souvent le millier ; or la pratique de la statistique classique, en présence de grands échantillons, n'est presque jamais enseignée et cela particulièrement aux biologistes et aux médecins.

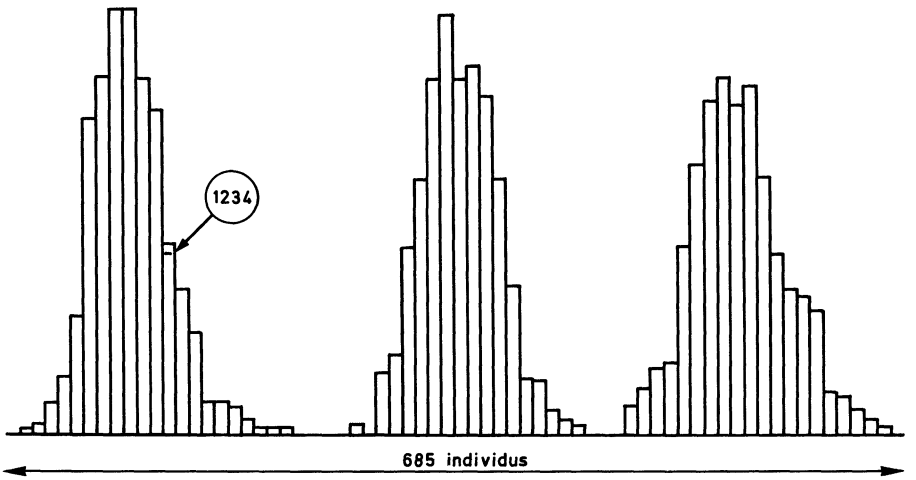
L'objet du présent article n'est pas de résoudre les problèmes posés par la biologie clinique mais, en reprenant les thèmes les plus élémentaires, de rappeler les liaisons existant au niveau de la pratique statistique entre : objectif de l'étude, hypothèses faites pour traiter le problème et taille des échantillons ; on espère ainsi provoquer chez certains une remise en cause des réflexes conduisant, en présence de données nombreuses, à une pratique beaucoup trop dépendante encore des méthodes instituées par des statisticiens qui avaient pour souci principal l'économie de moyens.

Se plaçant dans le cadre de l'étude des paramètres biologiques, nous discuterons dans une première partie quelques thèmes très simples de la démarche classique en univers unidimensionnel ; puis, dans une deuxième partie, abandonnant résolument le modèle probabiliste, nous rappellerons, en passant en revue quelques techniques d'analyse de données multidimensionnelles faisant appel à l'algèbre linéaire, dans quelle mesure la pratique de ces techniques est autant fonction des objectifs poursuivis que du volume et de la nature même des données.

1 – APPROCHE UNIDIMENSIONNELLE

Le premier outil à la disposition du statisticien est l'histogramme ; si le nombre de données n'est pas trop grand, on a intérêt, pour conserver le maximum d'information, à le construire en accumulant les observations repérées par leur numéro, dans des classes de largeur égale (cf. fig. 1) : cette technique de présentation permet d'isoler rapidement les données douteuses. Rappelons qu'un tel histogramme fournit en premier lieu un classement des observations et, en second lieu seulement, une idée de la distribution statistique du paramètre. On donne dans les *figures 1, 2 et 3* les histogrammes obtenus au Centre de Médecine Préventive de Vandœuvre-les-Nancy, respectivement pour le phosphore (fig. 2), l'acide urique (fig. 1) et le cholestérol (fig. 3) (classe d'âge 30-50, sexe masculin). Ces histogrammes ont un aspect "normal" ; à l'œil, rien à signaler de particulier ; qu'en faire donc ?

ACIDE URIQUE PHOSPHORE CHOLESTEROL



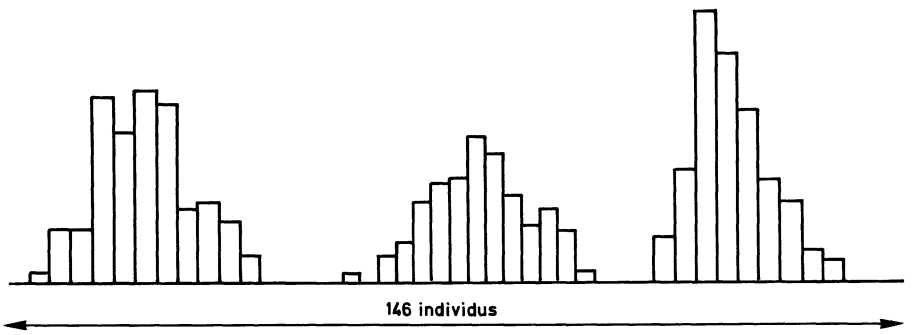


Figure : 1 Figure : 2 Figure : 3

Le réflexe malheureusement classique consiste, sans se soucier des objectifs, à se dire : nous sommes en présence d'échantillons tirés de lois gaussiennes ; le couple "moyenne – écart-type" formant donc un résumé exhaustif de l'information, nous effectuerons toutes les tâches futures à l'aide de ces seuls paramètres.

Dans le tableau de la figure 4 sont décrits les résultats obtenus par un test de khi-deux effectué pour juger de la qualité de l'hypothèse "la distribution considérée est normale" ; les probabilités de dépasser le khi-deux obtenu sont données pour onze paramètres biologiques, dont les trois paramètres précédents : phosphore, acide urique et cholestérol, en considérant successivement des échantillons de taille moyenne et de grande taille.

Ces résultats sont-ils surprenants ? Le seul qui le soit pour le statisticien, dont l'attitude est ici à opposer à celle qu'il aurait dans le cas d'un petit échantillon, concerne le cholestérol pour lequel l'hypothèse de normalité

n'est pas rejetée dans le cas du grand échantillon : en effet, les lois statistiques, et en particulier la loi de LAPLACE-GAUSS, sont des êtres abstraits, des modèles, des approximations de la réalité : si les échantillons sont grands, le test est puissant et l'approximation devient flagrante (test significatif) ; nous dirons ici que l'échantillon n'était pas encore assez grand pour rejeter l'hypothèse de normalité pour le cholestérol, ou que le modèle "normal" est meilleur pour le cholestérol que pour les autres paramètres [7].

Test du khi-deux

	Taille de l'échantillon 685		Taille de l'échantillon 146	
	Normalité	Log Normalité	Normalité	Log Normalité
Protéines totale	$p < \epsilon$	$p < \epsilon$	$p = 10^{-5}$	$p = 10^{-5}$
Calcium	$p < \epsilon$	$p < \epsilon$	$p < \epsilon$	$p < \epsilon$
phosphore	$p < \epsilon$	$p < \epsilon$	*p = 0,43*	*p = 0,075*
Cholestérol	*p = 0,04*	*p = 0,05*	*p = 0,15*	*p = 0,43*
Glucose	$p < \epsilon$	$p < \epsilon$	$p < \epsilon$	p = 0,007
Urée	$p < \epsilon$	$p < \epsilon$	*p = 0,16*	$p = 10^{-6}$
Acide urique	$p < \epsilon$	$p < \epsilon$	*p = 0,16*	*p = 0,35*
Créatinine	$p < \epsilon$	$p < \epsilon$	$p < \epsilon$	$p < \epsilon$
Phosphatase alcaline	$p < \epsilon$	$p < \epsilon$	p = 0,0067	*p = 0,26*
G.O.T.	$p < \epsilon$	$p < \epsilon$	$p < \epsilon$	p = 0,002
G.P.T.	$p < \epsilon$	$p < \epsilon$	$p < \epsilon$	p = 0,006
p = probabilité que le khi-deux théorique dépasse le khi-deux calculé ; $\epsilon = 10^{-6}$				

Figure 4.

Est-ce pour cela que nous dirons que le cholestérol est un "meilleur" paramètre que le phosphore ou l'acide urique ?

Nous serons tentés de dire le contraire, nous souvenant que la distribution normale avait été introduite en astronomie par GAUSS comme étant celle des erreurs [9] ; les erreurs qui interviennent ici dans le cholestérol (variabilités analytique et intra-individuelle) sont certainement très importantes en regard de la variabilité "inter-individuelle" due à des facteurs tels que l'alimentation, le tabac, la morphologie,. . .

Une autre hypothèse fréquemment utilisée, la "log-normalité", aurait pu être envisagée ; on jugera de la qualité de l'hypothèse "la distribution considérée est log-normale" en consultant à nouveau le tableau de la figure 4 où les résultats d'un test de khi-deux sont présentés de la même façon que précédemment. Les résultats sont presque identiques : l'hypothèse n'est pas rejetée pour les petits échantillons ; par contre, elle est toujours rejetée, sauf pour le cholestérol, dans le cas d'un grand échantillon.

Faut-il conclure à la vanité de toute hypothèse de loi et en particulier dans le cas de données nombreuses ?

Evidemment non, mais l'analyste, compte tenu de l'approximation, doit connaître les avantages et les inconvénients de l'hypothèse qu'il retient : même une hypothèse dénoncée comme fausse, par un test statistique par exemple, peut se révéler efficace pour effectuer certaines tâches ; par contre, cette même hypothèse peut conduire dans d'autres cas à des résultats suspects.

a) Réduction

Le premier avantage de l'hypothèse de normalité est de permettre une "réduction" de l'information (gain au niveau du coût) – cette réduction peut être même considérée souvent comme une "amélioration" : Dans le cas de grands échantillons par exemple, on peut améliorer la définition d'une valeur de référence en utilisant le modèle normal, même reconnu statistiquement comme non satisfaisant par un test (KOLMOGOROV, khi-deux etc.) [1]. En effet, la précision de l'estimateur u du fractile 97,5 par exemple sera moins bonne, si l'écart à la normalité n'est pas trop grand, en procédant de façon non paramétrique, plutôt que de procéder plus simplement en utilisant la moyenne m et l'écart-type s de la distribution à l'aide de la formule $u = m + 1{,}96\ s$; cette propriété résulte d'une part du caractère exhaustif de la réduction (m, s) sous hypothèse de normalité et d'autre part du fait que l'on travaille aux extrêmités de la distribution où la précision des probabilités estimées sans faire d'hypothèse de loi est mauvaise [10].

Cette réduction peut se faire à l'aide d'autres modèles que le modèle normal, et il est évident qu'un meilleur modèle (au sens d'un khi-deux par exemple) permettra de gagner encore au niveau de la précision dans la définition d'une valeur de référence (diminution du biais) : compte tenu les multiples autres avantages statistiques du modèle normal, parmi les autres modèles possibles les plus prisés sont surtout ceux qui, par une transformation simple sur la variable considérée, permettent de retrouver la normalité ; on peut par exemple associer à la variable x une variable plus "normale" y par une transformation du type :

$$y = a + b \cdot g\left(\frac{x - c}{d}\right) \quad \text{(JOHNSON)} \qquad [11]$$

où g est une fonction monotone,
et a, b, c et d sont des paramètres à déterminer.

Parmi les transformations du type "Johnson" la transformation log (modèle log-normal) est la plus pratiquée.

Rappelons que les transformations n'ont pas toujours pour unique but de retrouver la normalité ; elles sont utilisées aussi très souvent pour que des hypothèses d'une autre nature, préalables à l'application d'une technique, soient suffisamment respectées (additivité des effets et homogénéité des variances en analyse de la variance par exemple).

b) Comparaison

Si deux échantillons sont considérés, ils peuvent être issus de distributions statistiques parfois très différentes tout en ayant des moyennes pratiquement semblables ; ces différences au niveau des formes globales des distributions peuvent conduire le biologiste à des réflexions pertinentes et compliqueront parfois sa tâche quand il aura à définir par exemple les frontières entre le normal et le pathologique. Ces différences apparaîtront comme flagrantes parfois en comparant à l'œil les deux histogrammes ; pour plus de certitude, on aura recours à un test non paramétrique, le test de KOLMOGOROV-SMIRNOV par exemple : là encore, on se méfiera des conclusions hâtives dans le cas où les échantillons sont grands, le test ayant alors tendance à rejeter l'hypothèse d'identité des distributions, même si les différences entre les lois sont minimes.

Pour juger de la différence des moyennes de deux échantillons, rien de tel qu'un test t (STUDENT). Les conditions d'utilisation du t de STUDENT sont a priori très restrictives, les lois "parentes" devant être normales et de même écart-type. Mais le test est robuste. En effet, même si les deux variables ne suivent pas des lois normales – on supposera alors leurs écarts-types de même ordre de grandeur – leurs moyennes, quand la taille n de l'échantillon est grande, sont approximativement distribuées normalement si les conditions du "théorème central limite" sont vérifiées [6]. Ces conditions sont en général respectées (l'exemple le plus classique de loi n'obéissant pas aux conditions précédentes est la loi de CAUCHY). Toutefois, si les distributions parentes sont fortement asymétriques, l'approximation n'est bonne que pour n très grand, car la convergence de la loi des moyennes vers la normalité est lente [19].

Rappel technique :

Sous l'hypothèse de normalité, moyenne et écart-type sont indépendants ; cette propriété est d'ailleurs caractéristique de la loi normale (théorème de GEARY) [6]. Le t de STUDENT est donc le rapport de deux quantités indépendantes s'il y a normalité. Voici l'expression asymptotique du cœfficient de corrélation ρ entre la moyenne et la variance s^2 dans le cas d'une loi quelconque [12] :

$$\rho = \frac{K_3}{[K_2\,(K_4 + K_2^2)]^{1/2}} \quad \frac{(n_1\ n_2)^{1/2}}{n_1 + n_2 - 2} \quad \left(\frac{1}{n_2} - \frac{1}{n_1}\right)$$

où K_i est le cumulant d'ordre i, et n_1 et n_2 les effectifs des deux échantillons.

On constate donc que cette corrélation est nulle (ce qui ne garantit d'ailleurs pas l'indépendance des deux quantités) si la loi est symétrique ($K_3 = \mu_3 = 0$ et $K_3 = \mu_3$) ; on remarquera de plus que cette corrélation est d'autant plus forte que n_1 est différent de n_2 et que la loi est plus dyssymétrique.

Le test non paramétrique U de MANN et WITHNEY est couramment utilisé pour mettre en évidence des différences significatives entre des distributions que l'on suppose en général identiques à une translation près ; la conclusion s'exprime alors naturellement en termes de différences entre

moyennes. En général, si les échantillons sont petits, le test U met bien en évidence les différences entre moyennes. Mais si les échantillons sont grands, si le U est significatif – et il l'est d'ailleurs quasi-automatiquement si les échantillons sont très grands –, on aura du mal à interpréter les différences constatées entre ces deux distributions [15] [21].

Compte tenu du fait que les tests non paramétriques sont sensibles aux différences entre les distributions (ils ne sont pas "monospécifiques") on ne les utilisera, en général, pour comparer deux statistiques (les moyennes par exemple) que pour des échantillons de petite taille.

Voici les résultats des tests de STUDENT, et de MANN et WITHNEY appliqués à onze paramètres biologiques, d'une part pour une population "à jeun" et une population "non à jeun", et d'autre part pour une population de "fumeurs" et une population de "non fumeurs". [cf. fig. 5].

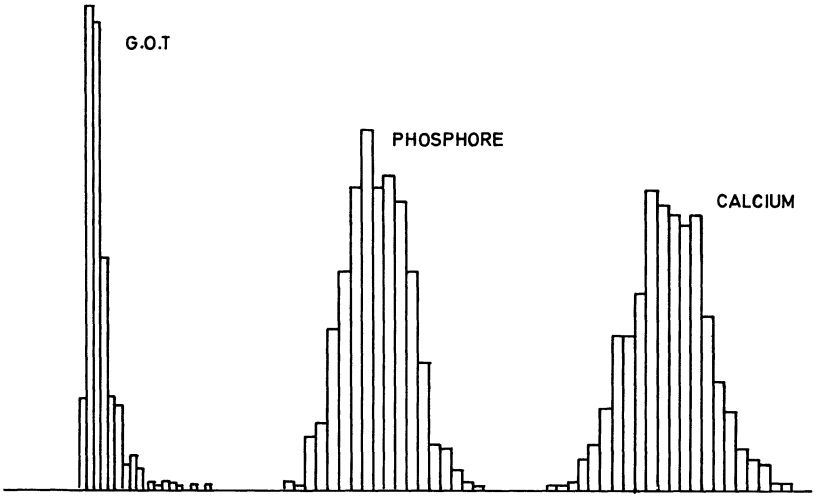
Test des Moyennes

	390 Individus à jeun 295 Individus non à jeun		365 Individus fumeurs 320 Individus non fumeurs	
	Student	Mann Withney	Student	Mann Withney
Protéines totales	$p = 0{,}018$	$p = 0{,}009$	$p = 0{,}121$	$p = 0{,}184$
Calcium	$p = 0{,}73$	$p = 0{,}76$	$p = 0{,}156$	$p = 0{,}146$
Phosphore	$p = 10^{-9}$	$p = 10^{-9}$	$p = 10^{-5}$	$p = 10^{-4}$
Cholestérol	$p = 0{,}78$	$p = 0{,}96$	$p = 0{,}967$	$p = 0{,}992$
Glucose	$p = 10^{-8}$	$p = 10^{-8}$	$p = 0{,}982$	$p = 0{,}38$
Urée	$p = 0{,}242$	$p = 0{,}07$	$p = 10^{-7}$	$p = 10^{-7}$
Acide urique	$p = 0{,}008$	$p = 0{,}007$	$p = 0{,}0238$	$p = 0{,}026$
Créatinine	$p = 0{,}5$	$p = 0{,}4$	$p = 0{,}468$	$p = 0{,}382$
Phosphatase alcaline	$p = 0{,}928$	$p = 0{,}68$	$p = 10^{-4}$	$p = 10^{-4}$
G.O.T.	$p = 0{,}682$	$p = 0{,}07$	$p = 0{,}138$	$p = 0{,}79$
G.P.T.	$p = 0{,}865$	$p = 0{,}66$	$p = 0{,}8$	$p = 0{,}598$
p = probabilité que le t de STUDENT ou le U de MANN et WITHNEY dépasse la valeur calculée				

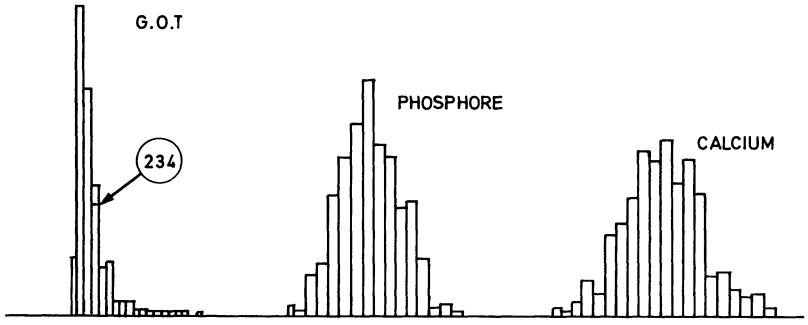
Figure 5.

Les résultats obtenus sont sensiblement différents lorsque les distributions sont fortement asymétriques, ce qui est le cas par exemple pour l'enzyme G.O.T. (cf. fig. 6, 7 et 8). Le test t est alors douteux (convergence lente) mais l'information apportée par le test U est-elle intéressante (grand échantillon) ?

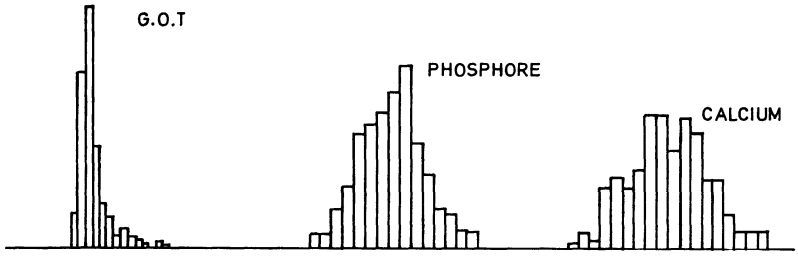
On remarque que pour le cholestérol qui suivait de très près une loi de LAPLACE-GAUSS – c'est un mauvais paramètre ! – aucune différence significative n'est mise en évidence entre les deux populations considérées.



Population mélangée 685



390 individus à jeun



295 individus non à jeun

Figure : 6 Figure : 7 Figure : 8

2 – APPROCHE MULTIDIMENSIONNELLE

La notion de variable isolée dont la distribution immuable serait la résultante d'aléas incontrôlables est une notion abstraite très éloignée de la réalité en biologie.

Toute variable est en réalité toujours à considérer parmi tout un paquet d'autres variables qui conditionnent son comportement ; le tout forme un écheveau qu'il s'agit de démêler ; la tâche n'est pas facile : les variables sont de natures très diverses ; elles sont quantitatives ou qualitatives, précises ou imprécises ; les liaisons sont complexes et, vu l'interdépendance des variables de l'écheveau, elles ne peuvent être mises en évidence en général uniquement en regardant les variables deux à deux.

Il en est ainsi des paramètres biologiques rencontrés au paragraphe 1 qui non seulement présentent entre eux certaines liaisons mais dépendent aussi de l'âge, du sexe, des habitudes alimentaires, de l'état psychologique, de pathologies éventuelles, de la morphologie, etc.

Certes, le biologiste a déjà des idées ; il a pu vérifier, en ayant recours parfois à un test statistique, l'influence de certains facteurs comme l'âge, le sexe ou le tabac ; mais il est encore très loin d'en savoir assez pour définir sans ambiguité des valeurs ou des frontières de référence. Les techniques d'analyse de données multidimensionnelles lui permettront parfois grâce aux documents très clairs qu'elles fournissent [1] – et qui sont souvent des images très peu déformées de la réalité – d'avoir une vision plus globale des phénomènes ; ces techniques qui sont descriptives, donc modestes, ont pour principal objet, rappelons-le, de permettre un classement des observations en tenant compte de nombreux points de vue (les variables) et d'étudier, en dégageant les grands traits, les liaisons entre paquets de variables. La connaissance de ces techniques – il ne s'agit pas de les exercer sur n'importe quoi et une réflexion poussée au niveau de la définition et de la qualité des paramètres est un préalable essentiel – permet de découvrir un cadre de pensée, complémentaire mais d'esprit tout à fait différent de celui de la statistique classique, où la stratégie à utiliser est pourtant, comme précédemment, entièrement dépendante des objectifs et du nombre d'observations. Les résultats fournis par une technique d'analyse des données seront considérés comme bons si l'information initiale n'a pas été trop déformée (ceci sera mesuré par la part d'inertie expliquée), si les résultats ne dépendant pas trop des aléas de l'échantillonnage (stabilité), si enfin les documents statistiques obtenus sont riches de faits (interprétabilité). La stabilité dépend des liaisons existant entre les variables et du nombre d'observations considérées. Signalons que de plus en plus les documents fournis par les analyses factorielles sont utilisés comme éléments de langage [1] ; synthétiques et percutants, ils valent souvent mieux qu'un long discours.

On rappelle ici, en insistant sur certains points, les techniques d'analyse de données "linéaire" les plus connues que l'on peut utiliser à l'heure actuelle sur bon nombre d'ordinateurs [4].

2.1. LES TECHNIQUES EN ANALYSE DE DONNEES

2-1.1. Objectif n° 1 – Classer un ensemble d'observations multidimensionnelles

La technique "reine" pour classer des observations multidimensionnelles est l'analyse en composantes principales. Cette technique repose principalement sur l'arbitraire qu'est le choix d'une distance euclidienne pour

mesurer la proximité entre individus. Que les variables soient qualitatives ou quantitatives, on peut pratiquer l'analyse en composantes principales – le rôle de la matrice de corrélation étant joué dans le premier cas par la matrice des coefficients de TSCHUPROW [12, 17, 20]. Si les données sont dichotomiques (variables qualitatives à deux modalités), les coefficients de TSCHUPROW ne sont autres que les $\phi^2 \left(= \frac{\chi^2}{n}\right)$ associés aux différents tableaux de contingence.

Les documents fournis par l'analyse, simples à interpréter, permettent de regrouper les observations en classes d'observations proches relativement à la distance choisie (qui tient compte de l'ensemble des caractères) et de regrouper les caractères en classes de caractères bien corrélés entre eux. Si l'interprétation des "corrélations" entre caractères est claire quand il s'agit de variables qualitatives – puisqu'on a alors affaire en gros a des Khi-deux – on sera prudent quand il s'agira d'interpréter des corrélations entre variables quantitatives.

L'analyse en composantes principales peut être utilisée éventuellement en présence de données manquantes, mais cela exige de l'analyste de la prudence et du "doigté" [13]. Suite aux travaux de Yves ESCOUFIER, l'analyse en composantes principales est à l'heure actuelle utilisée pour juger des proximités entre tableaux de données [8, 16, 17, 20].

2-1.2. Objectif n° 2 – Etudier la dépendance entre deux paquets de caractères

Les techniques principales d'analyse de données permettant d'étudier la dépendance entre deux paquets de caractères sont classées suivant le nombre et la nature des caractères considérés.

1 – Cas de deux variables quantitatives

Soient x et y deux variables quantitatives mesurées sur n individus.

a) Lorsque l'on a *peu d'observations,* la liaison la plus simple que l'on puisse imaginer est la liaison linéaire ; on mesurera l'intensité de cette liaison éventuelle par un cœfficient de corrélation linéaire.

b) *Si n est grand,* on a parfois intérêt à rendre qualitative l'une des variables en séparant l'intervalle de variations en k intervalles qui sont alors considérés comme les k modalités d'une variable qualitative. On peut mettre ainsi en évidence des liaisons fonctionnelles plus générales que les liaisons linéaires. L'indice mesurant l'intensité de cette liaison est un *rapport de corrélation.*

c) Si *n est très grand,* on rend parfois qualitatives les deux variables, ce qui permet d'étudier des liaisons plus générales que les liaisons fonctionnelles (relations) ; l'indice mesurant la dépendance entre les deux variables utilisé est ici un khi-deux. Les graphiques fournis par l'analyse factorielle des correspondances permettent alors très simplement, en exploitant la dépendance entre ces deux variables qualitatives considérées, de regrouper en modalités "équivalentes" les modalités de chacune des deux variables.

2 – Cas de plusieurs variables quantitatives

a) *Régression multiple*

Il s'agit ici de mettre en évidence la liaison entre :

– une variable quantitative (à expliquer)
et
– p variables quantitatives (explicatives).

La pratique de la régression n'est pas aussi simple que le laisse supposer le modèle linéaire considéré :

– Le nombre de variables explicatives ne doit pas être trop grand par rapport au nombre d'observations (fiabilité du modèle).

– Les variables explicatives, si elles sont corrélées exigent d'utiliser des procédures particulières (régression pas à pas, régression sur variables orthogonales, Ridge Regression) [5, 13].

– Les "coefficients de régression" fournis par l'analyse peuvent être exigés positifs (régression sous contraintes) [5].

b) *Analyse canonique*

L'objectif est ici de dresser un bilan des corrélations (linéaires) existant entre deux paquets de variables quantitatives.

c) *Analyse factorielle discriminante* [18]

Les documents très clairs fournis par une analyse factorielle discriminante permettent théoriquement de faire le point sur la liaison existant entre :

– une variable qualitative
et
– p variables quantitatives.

Exemple : Etude de la variabilité de paramètres biologiques en fonction de diverses pathologies.

Mais, si le nombre d'observations le permet, ayant rendu "qualitatif", par une analyse en composantes principales par exemple, l'un des paquets de caractères quantitatifs (c'est-à-dire ayant rangé les individus en classes d'individus proches relativement à l'ensemble des variables du paquet) on peut utiliser l'analyse discriminante à la place de l'analyse canonique ; on ne se restreint plus alors à l'étude des liaisons "linéaires" ; des liaisons "fonctionnelles" beaucoup plus générales peuvent être mises en évidence. Ceci revient à travailler non pas sur les coefficients de corrélation linéaire, mais sur les rapports de corrélation.

d) *Analyse factorielle des correspondances*

Si l'on veut préciser la forme de la liaison existant entre deux variables qualitatives, c'est l'analyse factorielle des correspondances qu'on utilise en général.

L'information à analyser est ici présentée sous la forme d'un "tableau de contingence".

Mais, là encore, si le nombre d'observations le permet, ayant rendu qualitatifs les deux paquets de caractères quantitatifs (par une analyse en composantes principales par exemple), on peut utiliser l'analyse factorielle des correspondances à la place de l'analyse canonique ; on étudie alors des liaisons de type beaucoup plus général que les fonctions (relations).

Remarque

Quand les individus sont décrits par plus de deux variables qualitatives, si on accorde à ces variables qualitatives des poids proportionnels aux nombres de modalités, on obtiendra une bonne description des individus et des proximités entre les modalités des différentes variables en effectuant une analyse factorielle des correspondances sur le tableau logique dont les colonnes sont les individus et dont les lignes correspondent aux variables indicatrices associées aux modalités des différents caractères [14]. On effectue alors une analyse des correspondances généralisée, analogue à celle proposée par BURT en 1950 [3].

CONCLUSION

La pratique des statisticiens, au début du siècle, reflétait en partie la faiblesse des moyens de mise en œuvre (données rares, calculs manuels) qu'ils avaient à leur dispcsition.

Si cette faiblesse a contribué largement au développement de théories où les statisticiens ont fait preuve d'une grande ingéniosité (théorie de l'échantillonnage, théorie de la décision, . . .) elle est devenue avec le temps un frein ; l'imagination étant fortement bridée par le souci d'économiser.

Aussi, l'apparition des moyens de calcul et de stockage puissants a-t-elle introduit une révolution dans la recherche et la pratique de la statistique.

Devant le foisonnement d'idées et de techniques nouvelles, le chercheur en biologie, face à des données nombreuses, est souvent désorienté : il ne sait quoi utiliser dans la panoplie de l'analyse des données, constatant que les techniques sont souvent semblables en ce qui concerne les objectifs énoncés. Il ne sait plus où s'arrête la description et où commence l'inférence. Il ne fait plus confiance au modèle probabiliste, car il sait que les distributions multidimensionnelles qu'il rencontre ne sont pas d'un type standard et il est gêné par le fait que les tests d'adéquation qu'il utilise en univers unidimensionnel conduisent presque toujours à un résultat significatif, lorsqu'il a des données nombreuses. Pourtant, le modèle probabiliste a toujours de l'intérêt, dans le cas où le nombre d'observations est grand, en univers unidimensionnel où il est fréquent qu'un modèle a priori contraignant (technique paramétrique), jugé comme très approximatif par un test d'adéquation, permette d'obtenir de meilleurs résultats qu'en opérant sans modèle. Le modèle en quelque sorte permet d'enrichir les données. En univers multidimensionnel, on ne peut juger de la stabilité des résultats de l'analyse qu'en faisant référence au modèle probabiliste.

La construction de programmes a permis, et c'est un bien, le développement de méthodes descriptives multimensionnelles qui sont devenues des outils faciles à manier pour des biologistes non mathématiciens, mais certaines techniques doivent leur succès plus à l'existence d'un bon programme qu'à leurs mérites propres. De plus, se limiter à la description est une attitude qui consiste à éviter d'approfondir plus avant les véritables utilisations qui seront faites des résultats accumulés – (qu'on soit dans l'impossibilité de prévoir ces utilisations ou qu'on préfère les ignorer !). Or, le point sur

lequel il faut bien insister est le suivant : la simple description, dans une population même supposée homogène, de la distribution d'un certain paramètre ne fournit en aucune façon l'information nécessaire pour résoudre complètement un problème particulier de dépistage, de diagnostic, de recherche... Connaître par exemple la distribution d'un paramètre biologique dans une population de malades correspondant à une certaine pathologie ne peut, en aucun cas, être considéré comme suffisant, en vue d'établir un diagnostic pour un nouveau cas qui se présente ou une stratégie de dépistage. Il faudrait en effet posséder une information de même type et supposée comparable dans une population saine ou exposée à d'autres pathologies. Le fait général qui est sous-jacent est bien que les objectifs d'ordre décisionnel, implicites à la plupart des mesures biologiques effectuées, ne peuvent être atteints à partir de simples accumulations de données descriptives, même supposées de bonne qualité. En particulier, la définition des "valeurs de référence" qui sont des seuils (ou frontières en univers multidimensionnel) opérationnels, c'est-à-dire permettant de prendre des décisions, ne peut résulter que d'une analyse menée en fonction des objectifs bien précis qui sont poursuivis, et ne peut en aucun cas être assimilée à la donnée de "valeurs caractéristiques" des distributions observées.

BIBLIOGRAPHIE

[1] J. ALIA – Le prix d'un Français – *Le Nouvel Observateur* n° 463 (sept. 1973)

[2] J. BRENOT, P. CAZES, N. LACOURLY – Pratique de la régression : qualité et protection – *Cahier du BURO* n° 23 1975.

[3] C. BURT – The factorial analysis of qualitative data. *British Journal of Statistical psychology.* Vol. 3, *3*, p. 166-195 (1950).

[4] F. CAILLIEZ et J.P. PAGES – Introduction à l'analyse des données – S.M.A.S.H. (1976)

[5] P. CAZES – Protection de la régression. Régression sous contraintes. Document 74117. Laboratoire de Biométrie du C.N.R.Z. (1974)

[6] D. DUGUE – Traité de statistique théorique et appliquée. Masson Edit. 1958.

[7] L.R. ELVEBACK et coll. – Health, Normality and the ghost of GAUSS. *J.A.M.A.*, janvier 1970.

[8] Y. ESCOUFIER – Echantillonnage dans une population de variables aléatoires réelles. Thèse de doctorat, Montpellier (1970).

[9] GAUSS – Theoria motus corporum coelestium in sectionibus conicis solem ambi entium, publié à Hamburg, 1809.

[10] E.K. HARRIS – Review of statistical methods for calculating reference intervals in clinical laboratory medicine. Second Congress of automation and prospective. Pont-à-Mousson (1972).

[11] M.H. HOYLE – Transformations. An introduction and a bibliography. *Inst. Stat. Rev.* vol. 41. 2, (1973).

[12] M.G. KENDALL et A. STUART – The advanced theory of statistics. Tome 2, p. 466-487 et p. 557 Griffin (1961).

[13] N. LACOURLY – Problèmes statistiques posés par le dépouillement d'enquêtes alimentaires. Thèse de 3e cycle – Paris VI (1974).

[14] L. LEBART – L'orientation du dépouillement de certaines enquêtes par l'analyse des correspondances multiples. *Annales du C.R.E.D.O.C.* n° 2 (1975).

[15] H.R. NEAVE, C.W.J. GRANGER – A Monte-carlo study comparing various two samples tests for differences in mean. *Technometrics,* vol. 10, *3,* août (1968).

[16] J.P. PAGES – A propos des opérateurs d'Y. ESCOUFIER. Séminaire I.R.I.A. – Classification automatique et perception par ordinateur (1974).

[17] J.P. PAGES, Y. ESCOUFIER, P. CAZES – Opérateurs et analyse des tableaux à plus de deux dimensions *Cahiers du Buro* n° 2 – 1976

[18] J.M. ROMEDER – Méthodes et programmes d'analyse discriminante. Dunod, Paris (1973)

[19] H. ROUANET et B. LECLERC – Le rôle de la distribution normale en statistique *Mathématique et Sciences Humaines* n° 32 (1970)

[20] G. SAPORTA – Liaisons entre plusieurs ensembles de variables et codage de données qualitatives. Thèse de 3e cycle. Paris VI (1975).

[21] S. SIEGEL – Non parametric statistics. Mac Graw Hill (1956).