Epidémiol. Santé anim., 1984, n° 6, 17-25

TECHNIQUES COURANTES UTILISEES DANS LE TRAITEMENT STATISTIQUE DES DONNEES EN BIOLOGIE

*J.C. FAYET**

RESUME : En s'appuyant sur des exemples, l'auteur présente les notions de base relatives aux tests statistiques usuels : analyse d'un tableau de fréquences, analyse de variance, régression linéaire.

SUMMARY : With the help of examples, the author gives basical informations about common statistical tests : frequency tables analysis, variance analysis, linear regression model.

* * *

I. INTRODUCTION

Toutes les personnes travaillant dans le domaine biologique sont amenées à acquérir des quantités, parfois impressionnantes, de données de nature variée. Il est évident qu'en présence d'un nombre de données dépassant quelques centaines, il faut avoir recours à des moyens informatiques plus ou moins puissants (du microordinateur jusqu'à un gros système). En revanche, lorsque le nombre de données à traiter reste raisonnable, on peut envisager de faire un traitement qualifié de "manuel". En fait, même dans ce cas, on a intérêt à utiliser du petit matériel de calcul, éventuellement programmable ou disposant de fonctions statistiques, le boulier s'avérant par trop insuffisant. Dans les deux cas, il faut savoir quel traitement appliquer à ses données.

Mon propos n'est pas de donner une liste exhaustive des problèmes à traiter en biologie mais plutôt d'aborder les plus fréquents et les plus simples à résoudre, ne nécessitant pas de moyens informatiques très puissants.

II. LES DONNEES EN BIOLOGIE

A. Echantillons et populations

Les données sont généralement basées sur des observations individuelles. Mais le terme "observations individuelles" peut recouvrir des entités différentes. L'individu peut être la vache, le prélèvement sanguin. Il peut aussi être l'exploitation, un pool de plasmas (lorsque l'analyse nécessite un volume important qu'un seul donneur ne peut fournir (rat ou souris, par exemple)).

---

* *Laboratoire d'Eco-pathologie, I.N.R.A. - C.R.Z.V. de Theix, 63122 Ceyrat.*

Les propriétés mesurées sur des observations individuelles sont appelées caractères ou plus habituellement variables.

Ainsi, le contrôle laitier mesure sur les vaches (individus) au moins la production, le taux butyreux, le taux azoté (3 variables).

En statistique, on entend par population, l'ensemble des observations individuelles existant quelque part dans le monde (ou au moins dans une zone d'échantillonnage limité dans le temps et l'espace), sur lequel on veut tirer des conclusions.

Si vous prenez 5 vaches sur lesquelles seront réalisées des formules leucocytaires et que vous désirez, à partir de cet échantillon, tirer des conclusions sur l'ensemble des vaches, alors la population, d'où est issu l'échantillon, est la formule leucocytaire de tous les individus femelles appartenant à l'espèces Bos taurus.

Si vous vous limitez volontairement à un échantillon plus étroit : les vaches Salers en 1ère lactation au 2ème mois de lactation et que les conclusions sont restreintes à ce sous ensemble particulier, alors la population sera la formule leucocytaire des Salers en 1ère lactation, pendant le 2ème mois de lactation.

B. Nature des variables

On distingue habituellement deux types de données : qualitatives et quantitatives.

1. Données qualitatives

Les données qualitatives sont celles qui, par définition, ne peuvent pas être quantifiées. Par exemple la couleur, l'odeur, la race, le sexe, vivant ou mort.

Encore est-il possible, dans certains cas, de quantifier des variables qualitatives. Par exemple une couleur peut être définie par sa longueur d'onde.

2. Données quantitatives

Les données quantitatives peuvent être divisées en deux classes.

a. Données continues

Ce sont celles pour lesquelles on estime suffisament petit, l'intervalle le plus faible entre deux données, par rapport à la grandeur mesurée.

Ex. : la taille d'un homme mesurée en cm,

Ex. : le poids d'une vache en kg (pas celui d'une souris, dans la même unité),

Ex. : la natrémie en mMoles/l.

b. Données discontinues ou discrètes

Ce sont celles qui sont nécessairement représentées par des valeurs entières.

Ex. : dénombrement, ou un âge exprimé en années.

Mais là encore, il est possible de changer le type de données. Ainsi, la quantité de paille sur une litière exprimée en kg/m² peut être regroupée en classes de type qualitatif :

$Q = 0$ kg	pas du tout
$0 < Q \leq 2$ kg	un peu
$2 < Q \leq 4$ kg	beaucoup
$4 < Q \leq 6$ kg	passionnément
$Q > 6$ kg	à la folie

III. TRAITEMENT DES DONNEES

Le tableau I recense quelques-une des techniques courantes utilisées dans le traitement des données biologiques.

Nous allons simplement, pour trois d'entre elles, essayer de donner des exemples illustrant leur domaine d'application. Une précision doit être fournie : l'utilisation de ces méthodes statistiques est subordonnée à un certain nombre de conditions qu'il n'est pas possible d'aborder ici. Il faut néanmoins connaître ces contraintes. Elles sont décrites dans de nombreux ouvrages, généralement accompagnées d'exemples ; il est indispensable de leur consacrer un minimum de temps de lecture. Pour rester dans un cadre limité, on pourra consulter les ouvrages de Dagnelie, de Snedecor et Cochran, de Tomassone et al., ou de Tranchefort qui ont le mérite d'avoir été écrits ou traduits en français. L'initiation à l'analyse multidimensionnelle peut se faire grâce aux livres de Fenelon ou de Lefebvre.

Tableau I : Techniques courantes utilisées dans le traitement des données.

NOM	VARIABLES A EXPLIQUER	VARIABLES EXPLICATIVES
Analyse d'un tableau de fréquences (X^2)	1 quantitative	P qualitatives
Régression linéaire simple	1 quantitative	1 quantitative
Régression linéaire multiple	1 quantitative	P quantitatives
Analyse de variance	1 quantitative	P qualitatives
Analyse de covariance	1 quantitative	P quantitatives Q qualitatives
Analyse discriminante	1 qualitative	P quantitative

Rappel de quelques définitions de base :

Moyenne : $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

Somme des carrés des écarts : $SCE = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$

Variance : $s^2 = \frac{SCE}{n-1} \rightarrow$ nombre de degrés de liberté (ddl)

Ecart-type : $s = \sqrt{s^2}$

Somme des produits des écarts par rapport aux moyennes de deux variables : $SPE = \sum_{i=1}^{n} \left[(x_i - \bar{x})(y_i - \bar{y})\right] = \Sigma XY - \Sigma X \frac{\Sigma Y}{N}$

A. Analyse d'un tableau de fréquences - Test d'indépendance

Exemple d'application :

Influence du délai d'ingestion du colostrum sur la diarrhée des veaux nouveau-nés.

Diarrhée \ Délai	$\leqslant 1$ h	$\geqslant 6$ h	
non	a ($\hat{a}$)	b ($\hat{b}$)	tl_1
oui	c ($\hat{c}$)	d ($\hat{d}$)	tl_2
	t c 1	t c 2	tg

On a une variable quantitative (délai) qui a été regroupée en 2 niveaux ($\leqslant 1$ h ; $\geqslant 6$ h). On note s'il y a ou non apparition de diarrhée pendant une période déterminée (par exemple pendant les 72 heures après la naissance).

On obtient, pour chaque case du tableau, des fréquences absolues a, b, c, d, nommées fo (fréquences observées).

La question est de savoir si, quel que soit le délai, le pourcentage d'animaux ayant eu une diarrhée est le même.

Méthode : On effectue le test du X^2 selon le principe suivant :

Pour chaque case, on calcule l'effectif théorique $\hat{f}$ que l'on devrait rencontrer s'il n'y avait pas d'effet de la variable explicative (délai).

$$\hat{f} = \frac{\text{total de la ligne x Total de la colonne}}{\text{Total général}}$$

et on obtient 4 nouvelles valeurs $\hat{a}$, $\hat{b}$, $\hat{c}$, $\hat{d}$.

Le test du X^2 consiste à savoir si l'écart entre les fréquences observées fo et les fréquences théoriques $\hat{f}$ est suffisamment petit pour être simplement dû au hasard de l'échantillonnage.

Pour chaque case, on calcule :

$$k = \frac{(fo - \hat{f})^2}{\hat{f}}$$

La valeur du X^2 est égale à la somme de toutes les valeurs de k ainsi calculées. On la compare à une valeur théorique, en tenant compte d'une probabilité (généralement 5 %) et du nombre de degrés de liberté (le nombre de degrés de liberté est :

(nombre de lignes - 1) x (nombre de colonnes - 1).

Dans le cas envisagé, il serait donc égal à 1).

Si la valeur observée du X^2 est supérieure ou égale à la valeur théorique on rejette l'hypothèse que les fréquences observées sont dues au hasard.

On admet alors que la variable explicative prise en compte a exercé une influence sur la fréquence d'apparition du phénomène observé.

Si la valeur du X^2 observé est inférieure à sa valeur théorique, on accepte l'hypothèse que les fréquences observées n'ont pas été influencées par la variable explicative.

Autres exemples :

Vaccinés, non vaccinés, indemnes, malades.
Traités, non traités, survivants, morts...

B. Comparaison de moyennes entre p échantillons ($p \geqslant 2$). Analyse de variance

Dans ce cas, on dispose d'une variable quantitative mesurée, dont on veut savoir si la variabilité est explicable par un (plusieurs) facteur (s).

Dans un souci de simplification, on va considérer le cas d'un seul facteur (variable qualitative explicative) à p niveaux.

Il n'est pas possible, dans le cadre de cet exposé, de rentrer dans le détail de l'analyse de variance. Aussi, vais-je me contenter de donner quelques exemples possibles d'application et d'en tirer la philosophie.

Il faut également savoir que, dans le cas de deux échantillons, le test "t" pour échantillons indépendants et l'analyse de variance constituent des tests interchangeables.

Exemples d'application :

Variable à expliquer	Variables explicative (facteur de variation)
Gain de poids moyen quotidien chez le porc (GMQ)	Trois niveaux de vitamine B12 incorporée à l'alimentation
Mesure de pH d'une solution	Quatre pH mètres différents
Nombre de vers adultes trouvés chez des rats infestés avec une dose unique de larves	1 vermifuge à tester par rapport à un placebo
Production de lait à même alimentation et même n° de lactation	Quatre races laitières

Principes de l'analyse de variance à un facteur fixe

Pour fixer les idées, prenons le premier exemple cité :

Supposons qu'il y ait dans chacun des 3 niveaux de vitamine B12 (A, B, C) 5 porcelets participant à l'expérimentation.

On dispose donc, à la fin, de 15 valeurs de GMQ sur lesquelles on peut calculer une SCE totale.

Le problème consiste à décomposer cette SCE totale en ses deux parties. En effet, à l'intérieur de chacun des lots on va observer une certaine variabilité qui peut aussi s'exprimer par une SCE appelée SCE intra ou SCE résiduelle. Cette SCE est ainsi nommée, car elle exprime la variation à l'intérieur (intra) des lots (ou niveaux du facteur) ou encore la partie de variation non expliquée (résiduelle) par le facteur.

L'autre composante de la SCE totale est la SCE inter ou SCE factorielle. Elle est ainsi nommée, car elle rend compte de la variabilité entre les lots ou encore de la variabilité expliquée par le facteur.

Dans tous les cas, SCE totale = SCE inter + SCE intra
= SCE factorielle + SCE résiduelle

Connaissant les différentes SCE, on peut calculer les variances correspondantes :

$$S^2 \text{ inter} = \frac{\text{SCE inter}}{\text{ddl inter}}$$

$$S^2 \text{ intra} = \frac{\text{SCE intra}}{\text{ddl intra}}$$

Le test F de l'analyse de variance ne fait que comparer ces deux valeurs.

$$F_{obs} = \frac{S^2 \text{ inter}}{S^2 \text{ intra}}$$

Si la valeur ainsi obtenue F_{obs} est supérieure ou égale à F théorique (à un seuil de probabilité choisi et avec le nombre de degrés de liberté correspondant), on rejette l'hypothèse impliquant qu'il n'y a pas de différences entre les niveaux du facteur. Il faut alors rechercher où se situent les différences.

Si la valeur de F_{obs} est inférieure à F théorique, on accepte l'hypothèse qu'il n'existe pas de différence entre les lots.

En résumé, il est facile de comprendre que si la variation entre lots est du même ordre de grandeur que la variation à l'intérieur des lots le facteur étudié a été sans influence sur la variable mesurée.

C. La régression linéaire

Définition : "La régression est un modèle mathématique linéaire reliant une variable aléatoire Y (à expliquer) à k variables ($k \geqslant 1$) explicatives, aléatoires ou non, ce modèle étant construit pour prédire ultérieurement Y" (Tranchefort).

Nous n'envisagerons que le cas le plus simple, où il y a une seule variable explicative notée X.

Ainsi, le problème consiste à estimer les paramètres a et b de l'équation Y = a + bX.

Exemples d'application :

Variables à expliquer (dépendante)	Variable explicative (indépendante)
. Nombre de pathologies décrites par exploitation en un an	. Production laitière moyenne par exploitation
. Pourcentage de lipides chez des animaux vivants	. Volume relatif de distribution de l'eau lourde
. Poids vif	. Age
. pH d'une solution	. Température de la solution

Principe du calcul de la droite de régression par la méthode des moindres carrés

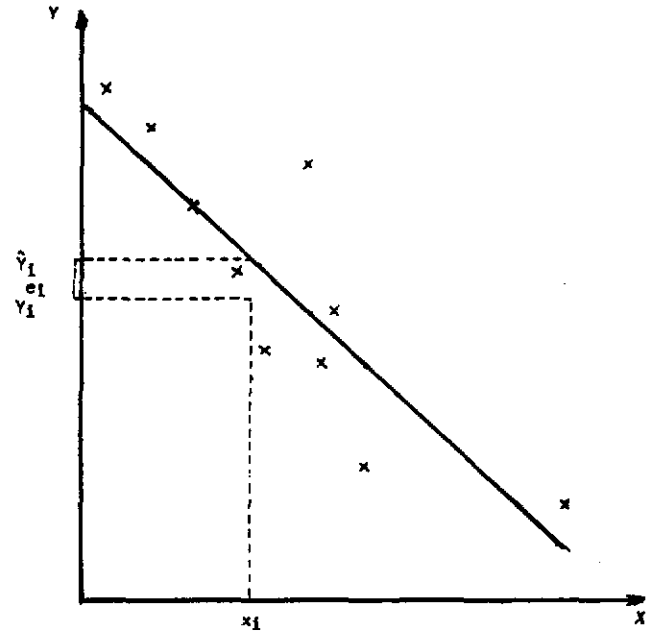
Il est évident qu'en biologie, le modèle mathématique Y = a + bX est généralement insuffisant pour décrire le phénomène observé. Il ne s'applique qu'au cas où les variables sont parfaitement liées (ex. : le périmètre et la longueur du côté d'un carré).

En réalité, en biologie, il subsiste généralement une différence entre la valeur du point observée Yi et sa valeur estimée sur la droite de régression $\hat{Y}i$ (figure 1).

La différence e entre ces deux valeurs est appelé résidu.

$$e_i = Yi - \hat{Y}i$$

Figure 1 : La régression linéaire simple



Pour chaque x_i, il existe une valeur y_i observée et une valeur $\hat{y}_i$ estimée (point sur la droite.

La différence $e_i = y_i - \hat{y}_i$ est appelée résidu.

La somme de tous les e_i doit être nulle et la somme de leurs carrés est minimale, d'où le nom de la méthode des "moindres carrés". On pourrait, empiriquement, trouver le minimum de cette somme des carrés des écarts résiduels, en faisant pivoter la droite autour du point défini par $\bar{X}$, $\bar{Y}$ et en calculant les résidus pour chacune des positions de la droite (!). Gros travail.

Il existe une méthode mathématique, relativement facile à mettre en oeuvre, qu'on trouvera dans l'un ou l'autre des ouvrages cités. En outre, la plupart des calculettes modernes sont munies de touches préprogrammées qui rendent aisé le calcul des paramètres d'une régression. Il ne faut pas croire que cette facilité dispense l'utilisateur de la connaissance des conditions d'application. Par ailleurs, un graphique des données s'avère toujours instructif avant d'effectuer les calculs.

Le paramètre b estimateur de la pente vaut : $b = \frac{SPE}{SCEx}$

Le paramètre a estimateur de l'ordonnée à l'origine vaut = $a = \bar{Y} - b\bar{X}$

IV. CONCLUSIONS

Les trois techniques de traitement statistique des données que nous avons survolées sont très classiques et les programmes correspondants sont faciles à se procurer et à mettre en oeuvre.

Mais, pour conclure, il convient d'insister sur un point que l'utilisateur doit garder présent à l'esprit.

> Aucun calcul statistique, aussi sophistiqué soit-il, ne permet d'établir une relation de cause à effet.

Cette relation dépend de l'utilisateur et de la théorie qu'il a échafaudée. Le traitement statistique ne fait qu'éprouver cette théorie à partir des résultats expérimentaux.

Aussi, c'est l'utilisateur qui imagine et suppose que le rendement d'une variété de blé est explicable, au moins en partie, par la pluviométrie du mois de mai, par exemple, et non l'inverse. C'est sa connaissance des phénomènes biologiques qui l'autorise à juger cette hypothèse comme vraisemblable et non pas la moindre considération statistique.

REFERENCES BIBLIOGRAPHIQUES

DAGNELIE (P.), 1973.- Théorie et méthodes statistiques. Vol. I. La statistique descriptive et les fondements de l'inférence statistique. Presses Agronomiques de Gembloux, 378 p.

DAGNELIE (P.), 1975.- Théorie et méthodes statistiques. Vol. II. Les méthodes de l'inférence statistique. Presses Agronomiques de Gembloux, 463 p.

DAGNELIE (P.), 1977.- Analyse statistique à plusieurs variables. Presses Agronomiques de Gembloux, 362 p.

FENELON (J.P.), 1981.- Qu'est-ce que l'analyse des données ? Lefonen, 311 p.

LEFEBVRE (J.), 1976.- Introduction aux analyses statistiques multidimensionnelles. Masson, 219 p.

TOMASSONE (R.), LESQUOY (E.) et MILLIER (C.), 1983.- La régression : nouveaux regards sur une ancienne méthode statistique. INRA. Actualités scientifiques et agronomiques. Vol. 13. Masson, 180 p.

TRANCHEFORT (J.), 1974.- La régression : application à l'agronomie. I.T.C.F., 178 p.

SNEDECOR (G.W.) et COCHRAN (W.G.), 1971.- Méthodes statistiques. 6ème édition traduit par Boelle H. et Camhaji E. A.C.T.A., 649 p.

*
* *