Ann. Zootech., 1964, **13** (4), 379-387.

NOTE TECHNIQUE

RÉALISATION ET UTILISATION DE PROGRAMMES SUR ORDINATEUR QUI PERMETTENT LES CALCULS STATISTIQUES DE BASE SUR DES DONNÉES BIOMÉTRIQUES

R. ROUVIER, J.-C. CANONGE

Station centrale de Génétique animale,
Centre national de Recherches zootechniques, Jouy-en-Josas (Seine-et-Oise)

SOMMAIRE

Nous avons réalisé un groupe de programmes permettant d'effectuer sur ordinateur les opérations simples de la statistique : calcul des paramètres de séries statistiques, détection des données extrêmes des distributions, établissement de distributions de fréquences, test de linéarité des liaisons avec calcul des coefficients de corrélation et de régression. Ces calculs peuvent être faits sur les valeurs brutes ou sur les écarts à la droite de régression. Une partie des calculs (test de linéarité) peut se faire après transformation logarithmique des données introduites.

Nous indiquons, à la suite d'une description rapide de ces programmes, les possibilités de leur utilisation et leur enchaînement, dans une étude biométrique élémentaire préalable à un traitement statistique plus élaboré des données.

INTRODUCTION

Le traitement statistique élaboré des données quantitatives d'observation ou d'expérimentation suppose en général que certaines hypothèses soient réalisées. Ces hypothèses concernent :

— La distribution des valeurs observées des caractères (distributions marginales et liées).

— La nature des liaisons statistiques entre certaines variables prises deux à deux.

La vérification de ces hypothèses, qui semble être un travail préliminaire important, peut être une opération très longue (surtout pour des lois à plusieurs variables) lorsqu'on utilise des moyens de calcul électromécaniques. Il a paru utile de concevoir et de réaliser un ensemble de programmes utilisables sur ordinateur qui permettent de rendre systématique ce travail. Ces programmes de calcul de la Station de Génétique animale, désignés sous les rubriques 62 013, 62 014, 62 015, 62 016, permettent :

— le calcul de paramètres statistiques : moyenne, variance, écart-type, coefficient de variation ;

— la recherche des valeurs se trouvant aux extrémités des distributions, et qui correspondent soit à des animaux (ou des groupes d'animaux) exceptionnels, soit à des erreurs de mesure ;

— l'établissement de distributions de fréquences absolues et relatives, cumulées et non cumulées, qui permettent la comparaison de distributions observées à des lois de probabilité théoriques ;

— l'étude de la liaison statistique entre une variable considérée comme indépendante et une ou plusieurs variables dépendantes, par :

● observation de l'évolution des moyennes de la variable dépendante par classes de la variable indépendante ;

● calcul des paramètres statistiques simples (dont la variance), dans chacune de ces classes, ce qui donne la possibilité d'effectuer un test d'homogénéité des variances de la variable dépendante dans les différentes classes ;

● test de la linéarité des liaisons estimées par la droite de régression, entre une ou plusieurs variables dépendantes et une variable indépendante. Ces calculs peuvent se faire à partir des valeurs observées des variables ou après transformation logarithmique.

— Le calcul des écarts à la droite de régression, comptés parallèlement à l'axe des ordonnées. Ces écarts constituent la distribution résiduelle lorsque la variable indépendante est maintenue constante. Dans le cas d'une distribution normale à deux ou plusieurs variables, ces écarts suivent également une loi normale à une ou plusieurs variables. Ils se prêtent donc aux calculs statistiques qui nécessitent l'hypothèse d'une loi normale. Cela ne serait pas le cas des rapports (variable dépendante divisée par la variable indépendante) que l'on calcule parfois.

DESCRIPTION DES PROGRAMMES

Cette description est donnée en détail dans les notices techniques figurant à la bibliothèque de la Station centrale de Génétique animale. Nous indiquons les éléments de cette description dont la connaissance est nécessaire pour une bonne utilisation scientifique des programmes.

Tous ces programmes, utilisables sur ordinateur IBM 1620 (20 000 positions de mémoire), considèrent simultanément 16 variables au maximum. Ils ont été conçus en vue d'être utilisés en séquence (les résultats de l'un fournissant des données pour le suivant), et de façon automatique.

Programme 62 013 : Calcul de paramètres de séries statistiques

Le programme calcule simultanément pour chaque variable X considérée un certain nombre d'éléments et en particulier l'effectif N, la moyenne arithmétique $\bar{X}$, la variance s^2, l'écart-type s, et le coefficient de variation.

Programme 62 014 : Établissement de distributions et de courbes de fréquence

Un certain nombre de constantes sont introduites :

● $\overline{X}$ et s (moyenne et écart-type pour chaque variable considérée, ces quantités étant calculées par le programme 62 013 ou fixées par l'utilisateur).

● Une valeur de grande borne (G. B.) et une valeur de petite borne (P. B.).

● Un paramètre p qui fixe l'amplitude de classe de départ, égale à $\frac{s}{4}$ ($p = 1$) ou $\frac{s}{2}$ ($p = 2$).

● Un paramètre q qui a son utilité en rapport avec l'utilisation du programme 62 015.

Les calculs sont effectués en deux stades :

1° *Premier stade*

— Élimination des calculs ultérieurs et tabulation de toute carte dont une variable au moins présente une valeur X telle que :

$$X \leqslant \overline{X} - \text{G. B.} \times s \text{ ou } X \geqslant \overline{X} + \text{G. B.} \times s$$

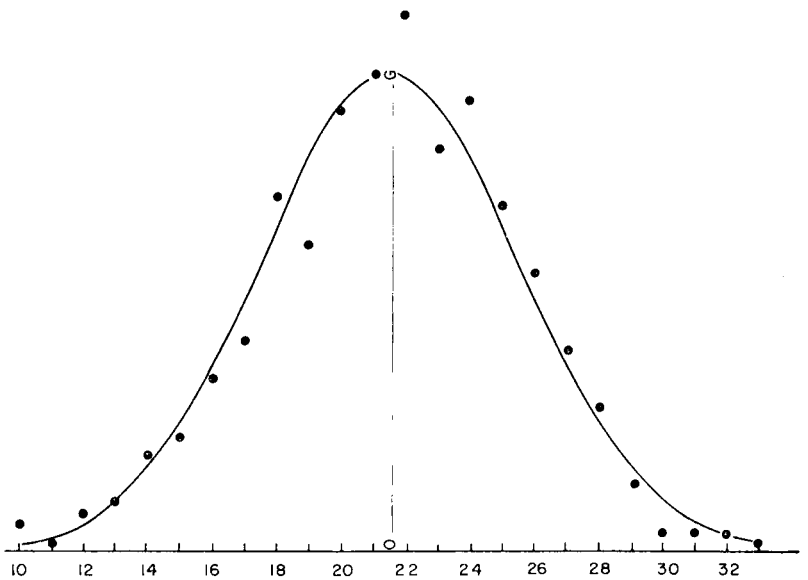
— Recherche des valeurs X, avec tabulation des cartes correspondantes, telles que :

$$\overline{X} - \text{G. B.} \times s < X \leqslant \overline{X} - \text{P. B.} \times s$$
$$\overline{X} + \text{P. B.} \times s \leqslant X < \overline{X} + \text{G. B.} \times s$$

2° *Deuxième stade.*

Sur les données restantes,

2.1. Calcul des paramètres statistiques.



GRAPHIQUE 1 — *Exemple de tracé de courbe de fréquences par la machine à écrire (programme* 62 104). *La machine à écrire frappe :*

en abscisse : les numéros de classes (10 à 33) ;

en ordonnée : l'axe vertical O G dont la longueur représente la probabilité 0,9950 ; *des signes (représentés ici par des points) dont les ordonnées représentent les fréquences observées dans les classes.*

La courbe tracée représente la variation de la fonction densité de probabilité de la loi normale centrée réduite.

La distribution observée qui est indiquée correspond à des données fournies par la Station expérimentale d'Aviculture du Magneraud (poids de coquelets âgés de 8 semaines).

2.2 Établissement des distributions de fréquence, dans l'intervalle $\overline{X} - 5\,s$ (borne supérieure de la première classe comprise), $\overline{X} + 5\,s$ (borne inférieure de la dernière classe non comprise).

Si $p = 1$, on a 42 classes. Les bornes supérieures des classes de départ, de la forme $\overline{X} + K\frac{s}{4}$ (K variant de — 20 à + 20) sont en général des fractions des valeurs observables de la variable X. Pour chaque classe K, on indique le nombre d'observations dans la classe, les fréquences (par classe, cumulées avec celles des classes précédentes), la borne supérieure de la classe exprimée en unités physiques de la variable (X_k est la valeur observable de la variable immédiatement inférieure à $\overline{X} + K\frac{s}{4}$).

Lors de la sortie des résultats, la machine à écrire indique les courbes de fréquence dans l'intervalle $\overline{X} - 3\,s$, $\overline{X} + 3\,s$. Ces courbes peuvent être superposées à une courbe théorique de Laplace-Gauss tracée à la même échelle (graph. 1). On peut ainsi apprécier graphiquement la normalité de l'échantillon, dans la mesure ou l'unité physique de mesure utilisée est très faible par rapport à l'écart-type s.

Programme 62 015 : Test de linéarité. Corrélation et régression par classe

On introduit des bornes de classes d'une variable considérée comme variable indépendante. Les valeurs de ces bornes, sont soit fixées par l'utilisateur, soit déterminées par le programme 62 014. Dans ce dernier cas, les effectifs N_i par classe sont tels que $N_i \geqslant \frac{N}{20}$, avec la condition $N_i \geqslant q$ (q est le paramètre précédemment indiqué).

Le programme calcule et donne pour chacune des variables dépendantes considérées simultanément :

- les moyennes des variables dépendantes et indépendante, dans chaque classe, ainsi que leurs moyennes générales ;
- les éléments permettant de faire le test de linéarité, c'est-à-dire la valeur de F :

$$F = \frac{\text{variance des moyennes de classes à partir de la droite de régression}}{\text{variance résiduelle à partir des moyennes de classes}}$$

et les degrés de liberté correspondants ;

- les caractéristiques η^2 (carré du rapport de corrélation), r et r^2 (coefficient de corrélation linéaire et son carré) ;
- le coefficient de régression par rapport à la variable indépendante ;
- les variances résiduelles de la variable dépendante à partir de ses moyennes de classes et à partir de la droite de régression.

Les formules de calcul utilisées sont celles décrites par MORICE et CHARTIER (1954).

Il convient de remarquer que le test de linéarité effectué par ce programme est, dans la plupart des cas, approché du fait qu'on peut avoir dans chaque classe plusieurs valeurs différentes de la variable indépendante.

Ce programme peut aussi être utilisé pour obtenir, parmi les résultats précédents, ceux qui sont calculables sur les données relatives à chaque classe particulière de la variable indépendante (la valeur de F ne peut évidemment pas être calculée dans ce cas).

Programme 62 016 : Test de linéarité : cas général.
Calcul des écarts à une droite de régression.

Ce programme permet dans une première partie, pour une variable indépendante X et une ou plusieurs variables dépendantes Y, de tester la linéarité de la liaison entre Y (ou Log Y) et X (ou log X)

Chaque valeur observée de X (ou de log X) constitue une classe. La méthode de calcul est celle décrite par OSTLE (1954).

Les résultats donnés sont les mêmes que ceux du programme 62 015, moins les moyennes des variables par classe.

Dans une deuxième partie, il calcule les écarts à la droite de régression de Y (ou Log Y) en X (ou Log X). Soient Y la valeur observée correspondant à X observée, $\hat{Y}$ la valeur de Y estimée par la

droite de régression, $\overline{Y}$ la moyenne générale des Y, s_t l'écart type résiduel de Y à partir de la droite de régression. Ces écarts sont calculés suivant l'une des trois formes :

$$Y - \hat{Y} + \overline{Y}$$

$$\frac{Y - \hat{Y}}{s_t}$$

$$\frac{Y - \hat{Y}}{s_t} a + b$$

(a et b étant des constantes données).

UTILISATION SCIENTIFIQUE DES PROGRAMMES

Premier problème

On dispose d'un fichier important de données quantitatives (résultats de contrôle dans les fermes ou en stations) et l'on désire déceler les observations (animaux par exemple) qui présentent des valeurs extrêmes pour un ou plusieurs caractères mesurés. Ces valeurs correspondent :

— soit à des erreurs de mesures ou de transcription (valeurs aberrantes ou douteuses),

— soit à des animaux exceptionnels.

La solution.

Elle consiste à :

1° Utiliser le programme 62 013 qui donne une première estimation de la moyenne ($\overline{X}$) et de l'écart-type (s) de chaque variable. Utiliser ensuite le programme 62 014 en précisant de ne pas effectuer la partie distribution de fréquence de ce programme. Les valeurs de la grande borne (G.B.) et de la petite borne (P.B.) sont fixées par l'utilisateur. On indiquera, par exemple :

G.B. = 4,0 P.B. = 3,0

L'utilisateur pourra étudier le cas particulier de chaque animal dont les résultats ont été « tabulés ».

2° L'utilisateur peut, s'il a déjà une certaine connaissance biométrique du matériel étudié, fixer lui-même des valeurs de $\overline{X}$ et s (au lieu de les faire calculer par le programme 62 013).

Deuxième problème

L'utilisateur après avoir exploré son fichier (1er problème), éliminé ou non certains animaux, corrigé certaines valeurs, désire étudier la forme de la distribution de chaque caractère mesuré, ce qui lui permettra d'en déduire des transformations éventuelles des variables.

La solution.

Le fichier de base aura probablement été modifié depuis la dernière opération mécanographique effectuée. On peut alors :

1° Utiliser le programme 62 013.

Utiliser ensuite le programme 62 014 en demandant la distribution de fréquence : les valeurs $\overline{X}$ et s (calculées par le programme 62 013) de chaque variable sont utilisées pour déterminer les bornes des 42 classes de travail qui ont pour amplitude $\frac{s}{4}$ dans l'intervalle $(\overline{X} - 5\ s, \overline{X} + 5\ s)$. Les bornes réelles X_k des classes effectives (valeurs observables de la variable) sont tabulées (voir la description du programme 62 014).

Les nombres d'observations (absolus, relatifs) par classes qui sont tabulés permettent (connaissant $\overline{X}$ et s) de tester l'ajustement de la distribution observée à une distribution théorique (test du χ^2).

Les fréquences relatives cumulées permettent un test graphique de la normalité de la distribution des valeurs brutes ou logarithmiques (droite de HENRI, avec papier gausso-arithmétique ou gausso-logarithmique).

Les fréquences relatives par classes (regroupées ou non) permettent de tracer les histogrammes.

2° Le programme 62 014 peut être utilisé dans le même but que précédemment, l'utilisateur fixant lui-même les valeurs des paramètres $\overline{X}$ et s, soit des valeurs $\overline{X}^1$ et s^1. Cela présente un intérêt lorsque l'utilisateur a une connaissance approximative de $\overline{X}$ et s, ou lorsqu'il préfère que s^1 soit un multiple de l'unité de mesure, ou que $\overline{X}^1$ ait une valeur particulière.

Le programme 62 014 donne aussi les valeurs d'une variable particulière correspondant à des fractiles déterminés de cette variable, ce qui fournit des indications sur la forme de la distribution (symétrie en particulier).

3° Un certain nombre n de paquets de cartes séparés peuvent être traités par le programme 62 013, ce qui donnera pour chaque variable les valeurs de n couples moyenne-écart type $(\overline{X}, s)$. L'étude de liaison entre $\overline{X}$ et s permettra éventuellement de conclure sur la normalité de la distribution par application du théorème de GEARY[1] (« si les variables parentes d'un échantillon ont un second moment et si les statistiques $\overline{X}$ et s sont indépendantes, les variables parentes sont de LAPLACE-GAUSS »).

La forme de la liaison entre $\overline{X}$ et s (ou $\overline{X}$ et s^2), permet par ailleurs de prévoir le type de transformation approprié de la variable (BARTLETT, 1947).

Troisième problème

3.1. — On veut connaître la forme de la liaison des variables dépendantes et indépendante, et faire un test approché de linéarité des liaisons.

3.2. — On veut tester l'homogénéité des variances dans des classes de la va-

(1) En toute rigeur, ce théorème est énoncé par DUGUE (1958) avec :

$$s^2 = \frac{\sum_{i=1}^{i=N} (X_i - \overline{X})^2}{N}$$ alors que le programme 62 013 calcule

$$s^2 = \frac{\sum_{i=1}^{i=N} (X_i - \overline{X})^2}{N-1}$$

riable indépendante, et connaître les coefficients de corrélation et régression dans chacune de ces classes.

Solution.

Le programme 62 015 donne les éléments de réponse à la question 31. Les bornes de classes de la variable indépendante ont été déterminées par le programme 62 014 (d'après la valeur du paramètre q). Lors du même passage, le programme donne en plus du test approché de la linéarité des liaisons, les coefficients de corrélation et de régression (voir la partie description).

Les paquets de cartes correspondant à différentes classes d'une variable (classes déterminées par le programme 62 014 comme précédemment, où toutes autres classes faites par l'utilisateur) peuvent être traités séparément et en séquence par ce programme, ce qui donne les éléments de réponse à la question 32.

On possède aussi, dans ce dernier cas, les éléments pour étudier la relation moyenne-écart-type ou moyenne-variance.

Quatrième problème

Test exact de la linéarité des liaisons entre une variable indépendante et une ou plusieurs variables dépendantes.

Solution.

Le programme 62 016 (1re partie) permet ce test (voir la description du programme). L'opportunité de la transformation logarithmique pour la variable indépendante ou les variables dépendantes est indiquée par les études préalables (forme des distributions, évolution des moyennes par classes, relation moyenne-écart-type) ou par ce que l'on sait de la nature biologique des données (cas de l'étude de l'évolution du poids en fonction de l'âge par exemple).

Rappelons que le programme 62 016 donne les variances résiduelles (la variable indépendante étant fixée) réelles (à partir des moyennes de classe) et estimées à partir de la droite de régression.

Cinquième problème

Étude de distributions résiduelles de variables dépendantes, la variable indépendante étant fixée.

Solution.

Le programme 62 016 (2e partie) donne les écarts par rapport à la droite de régression. Lorsque la liaison peut être considérée comme linéaire, l'étude de la distribution de ces écarts présente un très grand intérêt. Ils pourront représenter, par exemple :

— le poids d'une partie corporelle des animaux à poids corporel total constant ;

— la production de matière grasse à production totale de lait constante ;

— le poids corporel à âge fixé. Par exemple, les écarts liés du poids par rapport à la droite de régression du poids sur l'âge peuvent être exprimés en notes permettant de porter directement un jugement de valeur sur l'individu.

Tous les problèmes évoqués jusqu'à présent, et leurs solutions, peuvent être appliqués à ces écarts.

Il est particulièrement intéressant de remarquer que lorsque la distribution marginale des valeurs brutes n'est pas de LAPLACE-GAUSS, il se peut que la distribution des écarts le soit. On peut donc dans ce cas faire sur ces écarts tous les calculs et tests qui supposent que la distribution est normale.

SYNTHÈSE DES POSSIBILITÉS D'UTILISATION DES PROGRAMMES

Le tableau 1 résume les possibilités d'utilisation de l'ensemble des programmes. Chaque ligne du tableau représente une possibilité, + indique l'utilisation du programme, (+) indique une utilisation facultative.

Possibilité 1 : Solution du 1er problème : exploitation d'un fichier. On utilise le programme 62 013 et le programme 62 014 sans la partie distributions de fréquence.

Possibilité 2 : Solution du 2e problème : étude de la forme des distributions et recherche de transformations permettant de normaliser les distributions (l'utilisation du programme 62 013 est facultative).

TABLEAU 1

Résumé des possibilités d'utilisation de l'ensemble des programmes : chaque ligne du tableau représente une possibilité, + indique l'utilisation du programme, (+) indique une utilisation facultative.

Programmes→ Possibilités ↓	62 013	62 014	62 015	62 016
1	+	+		
2	(+)	+		
3	(+)	(+)	+	
4	(+)	(+)		+
5	(+)	(+)		+
6	+	+		(+)

Possibilité 3 : Solution du 3e problème : forme de la liaison des variables dépendantes et indépendantes, calcul des variances, coefficients de régression et corrélation par classes (de la variable indépendante ou pour toutes autres classes). Dans la plupart des cas il y aura eu un passage préalable par les programmes 62 013 et 62 014.

Possibilité 4 : Solution du 4e problème : test de linéarité des liaisons, sur les variables brutes ou après une transformation logarithmique (programme 62 016, 1re partie seulement).

Possibilité 5 : Solution du 5e problème : calcul des écarts à la droite de régression (programme 62 016, 1re et 2e parties).

Possibilité 6 : Solution du 5e problème : étude des écarts à la droite de régression. S'il y a au moins deux variables dépendantes on peut éventuellement utiliser de nouveau le programme 62 016 pour ces écarts.

Les écarts calculés par le programme 62 016 peuvent être étudiés de la même façon que les valeurs initiales (possibilités 1 à 6).

Reçu pour publication en mai 1964.

REMERCIEMENTS

Les auteurs tiennent à remercier M. B. VISSAC de la Station de Génétique animale pour l'aide qu'il leur a apportée dans la conception de l'enchaînement des calculs que les programmes précédemment décrits permettent de faire.

SUMMARY

DRAWING UP AND USE OF ELECTRONIC COMPUTER PROGRAMS FOR BASIC STATISTICAL CALCULATIONS WITH BIOMETRICAL DATA

The possibilities offered by electronic computer calculation led us to the realisation of a group of programs which make it possible to perform certain elementary statistical operations : calculation of the parameters of statistical series ; detection of the extreme values of the distributions ; establishment of frequency distributions; linearity tests and calculation of correlation and regression coefficients. The calculation can be made on the original data or on their deviations from the regression line. Some of the calculations (linearity tests) can be performed after logarithmic transformation of the data.

After a brief description of the programs, the possibilities of their use and application in an elementary biometrical study, preliminary to a more elaborate statistical treatment of the data, are indicated.

RÉFÉRENCES BIBLIOGRAPHIQUES

BARTLETT M. S., 1947. L'utilisation des transformations. *Biometrics*, **3**, 39-53.

DUGUÉ D., 1958. *Traité de statistique théorique et appliquée. Analyse aléatoire. Algèbre aléatoire*, p. 186, Masson et Cie, Paris.

MORICE E., CHARTIER F., 1954. *Méthode statistique, deuxième partie, Analyse statistique*, 344-351 et 376-377, Imprimerie nationale, Paris.

OSTLE B., 1954. *Statistics in Research. Basic concepts and techniques for Research workers*, 154-155, the Iowa State College Press, Ames, Iowa.