REVUE DE STATISTIQUE APPLIQUÉE

MARIE-FRANÇOISE MONCHOUX BOULARD

Une application des tests séquentiels à la sélection en génétique animale

Revue de statistique appliquée, tome 17, n° 2 (1969), p. 47-68

<http://www.numdam.org/item?id=RSA_1969__17_2_47_0>

UNE APPLICATION DES TESTS SÉQUENTIELS A LA SÉLECTION EN GÉNÉTIQUE ANIMALE

Marie-Françoise MONCHOUX BOULARD

Assistante à la Faculté des Sciences de Toulouse

Le but de cet article est de décrire une expérience de tests séquentiels appliqués à la sélection en génétique animale. Ce travail a fait l'objet d'une thèse de 3ème cycle (cf. BOULARD [1] Paris 1964). L'expérience a porté sur la sélection des vaches laitières. En conséquence, on rappelle rapidement les facteurs génétiques et non génétiques qui influent sur la production laitière. Le test séquentiel est décrit. Ses résultats sont ensuite discutés et comparés aux résultats obtenus par la méthode actuelle de sélection.

I - INTRODUCTION

En sélection animale, l'efficacité d'un choix est fonction du nombre de descendants possibles pour l'animal. Une vache peut en avoir 2 ou 3. Un taureau peut en avoir 10 000 à 30 000, d'où la sélection sur les pères. On considère donc un lot de descendantes de chaque taureau mis à l'épreuve et on compare les résultats des productions moyennes de chaque lot afin de mesurer les différences héréditaires entre les pères. Cette expérience, appelée testage des taureaux, demande environ 4 à 6 ans. Elle est longue et coûteuse (entretien des taureaux, nourriture, etc...) d'où les choix économiques interviennent.

Les problèmes qui se posent alors sont moins des problèmes de connaissance que des problèmes de décision. Il s'agit d'avoir un classement aussi précoce que possible des taureaux afin d'abattre les plus mauvais ou, au contraire, d'utiliser les meilleurs pour la reproduction. D'autre part, on ne peut pas arrêter l'expérience comme dans certains contrôles industriels en éliminant les lots inférieurs. En effet, on produit N filles par taureau. Les résultats des lactations de ces filles arrivent échelonnés dans le temps, de sorte que la décision prise à l'instant t sur n filles ($n < N$) sera vérifiée ultérieurement sur l'ensemble des N filles. Par conséquent, toute décision qui sera reconnue fausse par la suite aura des répercussions psychologiques graves.

Le fichier sur lequel porte notre étude a été établi par les services de la Coopérative du Centre Nord (centre de CHARMOY, dans l'Yonne). Il comporte les résultats des premières lactations de 5 000 vaches environ, filles d'une cinquantaine de taureaux de race normande (données recueillies de 1955 à 1961 inclus). Ces données sont obtenues au cours des contrôles

laitiers : tous les mois environ, on mesure différents facteurs concernant une traite. On calcule ensuite la production laitière totale, c'est-à-dire la production au cours de la lactation en intégrant par la méthode des trapèzes.

II - HYPOTHESES

Nous supposons dans cette étude que :

1/ Les données satisfont aux conditions d'échantillonnage : indépendance, répartition au hasard...

2/ Les facteurs de variation font varier la moyenne et non l'écart-type.

Nous devons insister sur le fait que nos données sont plutôt des résultats d'enquêtes que des résultats d'expérimentation rigoureuse au sens statistique :

- beaucoup de causes de variation échappent au vouloir humain : maladie, accident des animaux, conditions climatiques, etc...

- les facteurs humains interviennent : le fermier peut vendre ses animaux ; de lui dépendent aussi la nourriture et l'hygiène des animaux.

III - CAUSES DE VARIATIONS NON GENETIQUES DE LA PRODUCTION LAITIERE

La production laitière varie sous l'influence de facteurs génétiques et de facteurs non génétiques. Pour augmenter la précision du testage, on cherche à estimer et corriger, par des méthodes statistiques, les facteurs non liés au taureau. Nous citerons les principaux de ces facteurs d'après les études de MM. POLY et VISSAC (1959) [7] et de MM. POLY, POUTOUS et FREBLING (1965) [8].

a/ Année

Nous essayons de l'éliminer en partie en considérant les écarts des productions à la moyenne de l'année précédente (cf. plus loin au § V-1).

b/ N° de lactation

Nous ne considérons ici que la première lactation, ce qui élimine ce facteur.

c/ Durée de lactation

La corrélation entre la production laitière totale et la durée est de l'ordre de 0,70. Dans nos calculs, nous avons éliminé les lactations que les zootechniciens ont l'habitude de considérer comme anormalement longues ou courtes.

d/ Milieu de production

La production laitière des vaches est fortement influencée par les conditions d'entretien, d'alimentation, d'habitat et les méthodes de traite employées. Tous ces facteurs sont liés au troupeau, donc à l'étable dont fait

partie l'animal. Les services de recherche de la Station Centrale de Génétique Animale (Jouy-en-Josas) établissent en fin d'année un classement des étables suivant leur niveau de production moyen. On définit six classes d'étables. Le fichier que nous avons ne comporte que les résultats des classes de 2 à 6, les étables de classe 1 étant trop mauvaises pour qu'on en tienne compte dans des questions touchant à la sélection.

Nous avons cherché à estimer les composantes de variance dues à l'influence du milieu. Le modèle est celui des composantes de la variance (cf. GRAYBILL [3] p. 337, KEMPTHORNE [5] p. 103) dans le cas particulier d'une classification à une voie avec nombre inégal de données dans les groupes.

Nous avons trouvé :

1/ que l'effet influence du milieu sur la production laitière est hautement significatif à 0,001 ;

2/ que la composante de variance due à cet effet est de l'ordre de 10 à 15 % de la variance totale - comme le montre le tableau d'analyse de variance ci-après - bien que cette influence soit en partie masquée par les irrégularités dues à la durée de lactation. Des calculs ont montré que si on élimine l'effet durée, l'influence du milieu atteint 25 à 30 % de la variance (valeur bibliographique la plus fréquente).

Tableau des estimations de la variance due à l'effet milieu de production.

Année	1955	1956	1957	1958	1959	1960	1961
Variance étable (1)	156290	137494	121786	75275	89783	148027	105915
Variance résiduelle (2)	784354	806849	914353	1 054078	752805	826058	926347
Rapport $\frac{(1)}{(1)+(2)}$	0,166	0,146	0,118	0,067	0,106	0,152	0,103

IV - EFFET GENETIQUE DU AU PERE

Les calculs faits à la Station de Génétique Animale du C.N.R.Z. montrent que l'effet génétique dû au père est très faible, de l'ordre de 5 à 10 % de la variance totale, d'où la difficulté de la sélection. Néanmoins, on constate de grandes différences génétiques entre les pères, ce qui justifie la sélection.

Les généticiens ont défini un index qui permet d'estimer la valeur génétique du père. Cet index est établi d'après les résultats de lactations du lot de testage (cf. [2], [7], [8]).

Une autre méthode de sélection consiste à considérer les résultats des premiers contrôles de la lactation. En effet, la corrélation entre la production laitière totale P_t et la production au cours des k premiers contrôles : P_k, est forte dès le 5ème contrôle et augmente relativement peu entre le

5ème et le 8ème contrôle. Si on pratique la sélection d'après les résultats des 4 premiers contrôles, le gain de temps est en moyenne de 6 mois.

Corrélation entre P_t et P_k ; k = 1, 2, ...8. (Résultats tirés de [7]).

P_1 0,65

P_2 0,78

P_3 0,80

P_4 0,89

P_5 0,88

P_6 0,89

P_7 0,92

P_8 0,97

V - METHODE SEQUENTIELLE

V-1 - Définitions et notations.

Nous appellerons lot l'ensemble des filles d'un taureau. Ce lot a pour numéro le numéro mécanographique du taureau, père de ce lot.

Nous ferons les calculs séquentiels lot par lot, au fur et à mesure qu'arriveront les données du lot.

Y_{iact} = Production laitière totale au cours de la première lactation d'une vache i, fille d'un taureau t, entretenue dans une étable de classe c et ayant terminé sa lactation au cours de l'année a.

$Y_{.ac.}$ = Moyenne des lactations dans les étables de classe c pour l'année a.

Une étable est dite de classe c pour l'année a si, à la fin de l'année a-1, elle a été classée dans la classe c, d'après ses performances laitières au cours de l'année a-1.

Dans tous les calculs qui suivront nous utiliserons la variable X_{it} définie par :

$$X_{it} = Y_{iact} - Y_{.a-1\ c.}$$

Nous retranchons à chaque donnée la moyenne de classe de l'année précédente, ne connaissant pas encore la moyenne de l'année en cours quand nous faisons les calculs. Ce faisant, nous éliminons en partie l'influence de l'année et du milieu de production.

Nous voulons comparer les moyennes des lots $x_{.t}$, aussi nous utilisons le test séquentiel de moyenne d'une loi normale d'écart-type connu décrit par Abraham WALD (1947) [11] (p. 117). Les calculs étant faits lot par lot, nous noterons par la suite la variable X_{it}, x_i ou x quand il n'y aura pas d'ambiguïté.

V-2 - Test séquentiel de moyenne d'une loi normale (σ connu).

Soit x une variable aléatoire, normale, de moyenne m inconnue, d'écart type σ connu. Nous voulons tester que $m > m'$ (hypothèse H_0).

On peut en général trouver deux valeurs m_0 et m_1 telles que $m_0 < m' < m_1$; m_0 et m_1 définissent trois zones :

- une zone de préférence pour le rejet de H_0 quand $m \leqslant m_0$.
- une zone de préférence pour l'acceptation de H_0 quand $m \geqslant m_1$.
- une zone d'indifférence quand $m_0 < m < m_1$.

Soient :

α erreur de 1ère espèce : rejeter H_0 sachant que H_0 est vraie ,

β erreur de 2ème espèce : accepter H_0 sachant que H_0 n'est pas vraie,

$x_1, x_2 \ldots x_n$ des observations successives sur x.

La probabilité de l'échantillon $x_1 \ldots x_n$ est

$$p_{0n} = \frac{1}{(2\pi)^{n/2}\,\sigma^n}\, e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m_0)^2} \qquad \text{si } m = m_0$$

$$p_{1n} = \frac{1}{(2\pi)^{n/2}\,\sigma^n}\, e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m_1)^2} \qquad \text{si } m = m_1$$

On calcule $\frac{p_{1n}}{p_{0n}}$ à chaque nouvelle observation x_n du lot considéré. Vu la forme exponentielle du rapport, on considère son logarithme.

(1) Si

$$\log \frac{\beta}{1-\alpha} < \log \frac{p_{1n}}{p_{0n}} < \log \frac{1-\beta}{\alpha}$$

on continue l'expérience

(2) Si

$$\log \frac{1-\beta}{\alpha} \leqslant \operatorname{Log} \frac{p_{1n}}{p_{0n}}$$

on accepte H_0 : $m > m'$

(3) Si

$$\log \frac{p_{1n}}{p_{0n}} \leqslant \log \frac{\beta}{1-\alpha} ,$$

on rejette H_0, on accepte $m < m'$

Après quelques simplifications, on a :

$$\log \frac{p_{1n}}{p_{0n}} = \left(\frac{m_1 - m_0}{\sigma^2}\right)\left[\sum_{i=1}^{n} x_i - n\left(\frac{m_0 + m_1}{2}\right)\right]$$

Si on multiplie chaque membre des inégalités (1) (2) et (3) par $\frac{\sigma^2}{m_1 - m_0}$ puis ajoutant à chaque membre la quantité $n\left(\frac{m_0 + m_1}{2}\right)$, on voit apparaitre à gauche et à droite de (1) les nombres r_n et a_n.

$$r_n = \frac{\sigma^2}{m_1 - m_0} \log \frac{\beta}{1 - \alpha} + n\left(\frac{m_0 + m_1}{2}\right)$$

$$a_n = \frac{\sigma^2}{m_1 - m_0} \log \frac{1 - \beta}{\alpha} + n\left(\frac{m_0 + m_1}{2}\right)$$

Les inégalités (1) (2) et (3) deviennent :

(4)

$$r_n < \sum_{i=1}^{n} x_i < a_n$$

on continue l'expérience

(5)

$$a_n \leqslant \sum_{i=1}^{n} x_i$$

on accepte H_0 $m > m'$

(6)

$$\sum_{i=1}^{n} x_i \leqslant r_n$$

on rejette H_0 , on accepte $m < m'$

V-3 - Application pratique

A chaque stade de l'expérience et pour chaque lot séparément, on calcule

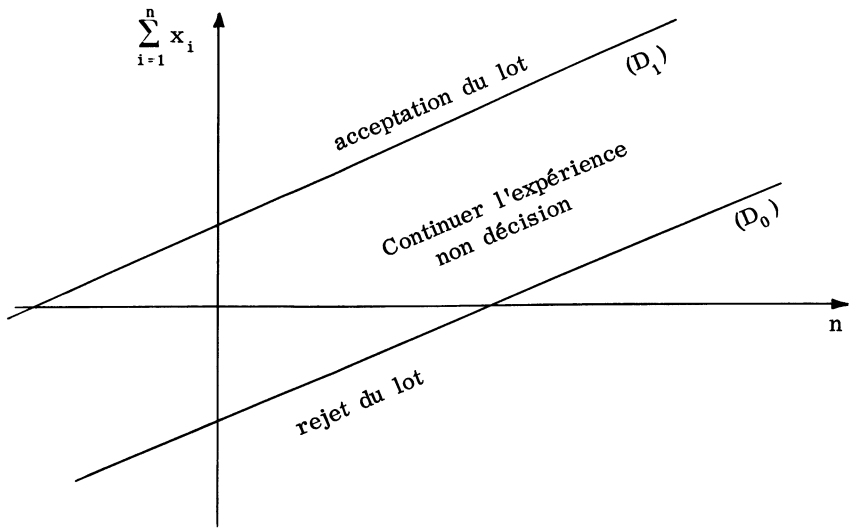
$$\sum_{i=1}^{n} x_i \ , \ a_n \quad \text{et} \quad r_n$$

et on prend les décisions données par (4) (5) et (6). Ceci peut se faire aisément sur ordinateur.

Méthode graphique

a_n et r_n sont des fonctions linéaires de n qu'on peut représenter par des droites D_1 et D_0 qui définissent 3 zones : acceptation, rejet, continuer l'expérience (cf. graphique).

On reporte à chaque stade sur ce graphique le point $\left(n, \sum_{i=1}^{n} x_i\right)$. Tant que ces points se trouvent entre D_0 et D_1, on continue l'expérience. Dès que ce point ne se situe plus entre ces lignes, on prend la décision correspondant à la zone atteinte.



V-4 - Définition du test séquentiel utilisé dans l'expérience

Le test décrit précédemment est défini par 4 paramètres α, β, m_0 et m_1 et une constante σ^2.

$\sigma^2 = 490\,000$ kg de lait d'après les calculs faits à la Station de Génétique Animale.

Les valeurs de α, β, m_0 et m_1 vont être fixées par des considérations opérationnelles. Il s'agit en effet de réduire le nombre d'observations nécessaires au choix ou au rejet d'un taureau.

Soit E(n/m) l'espérance mathématique du nombre d'observations n demandé par le test quand m est la vraie valeur de la moyenne. La théorie de WALD donne [11] (p. 123) :

$$E(n/m_0) = 2\,\sigma^2 \frac{(1-\alpha)\log\left(\frac{1-\alpha}{\beta}\right) - \alpha \log\left(\frac{1-\beta}{\alpha}\right)}{(m_1 - m_0)^2}$$

$$E(n/m_1) = 2\,\sigma^2 \frac{(1-\beta)\log\left(\frac{1-\beta}{\alpha}\right) - \beta \log\left(\frac{1-\alpha}{\beta}\right)}{(m_1 - m_0)^2}$$

Nous avons calculé les valeurs de $E(n/m_0)$ et $E(n/m_1)$ sur ordinateur IBM 7044 pour différentes valeurs des paramètres. Nous ne reproduisons ici que quelques valeurs avec risques égaux $\alpha = \beta$. Dans ce cas, $E(n/m_0) = E(n/m_1)$ que nous notons E(n) sur le tableau de la page suivante.

E(n/m) est inversement proportionnel à $(m_1 - m_0)^2$. Nous avons donc à chercher un optimum entre deux solutions extrêmes : sélection très fine ($m_1 - m_0$ petit) et un grand nombre d'observations ou sélection plus grossière ($m_1 - m_0$ grand) et peu d'observations. D'autre part, E(n/m) croît quand les risques α et β décroissent.

Nombre moyen d'expériences nécessaires à l'acceptation ou au rejet de l'hypothèse

$m_1 - m_0$ KG	E(n) $\alpha = 0,01$ $\beta = 0,01$	E(n) $\alpha = 0,05$ $\beta = 0,05$	E(n) $\alpha = 0,1$ $\beta = 0,1$
50	1 765	1039	689
100	441	260	172
150	196	115	77
200	110	65	43
250	71	42	28
300	49	29	19
350	36	21	14
400	28	16	11
450	22	13	9
500	18	10	7

Dans une première étude [1] nous avons pris des risques symétriques $\alpha = 0,05$; $\beta = 0,05$. Actuellement, la sélection demande environ 40 résultats. Aussi avons-nous choisi pour E(n) une valeur inférieure à 40, E(n) voisin de 30, ce qui détermine $m_1 - m_0 = 300$ d'après le tableau précédent.

V-5 - Résultats du premier test séquentiel

$$\alpha = \beta = 0,05 \ ; \ m_0 = -100 \text{ kg} \ ; \ m_1 = +200 \text{ kg}$$

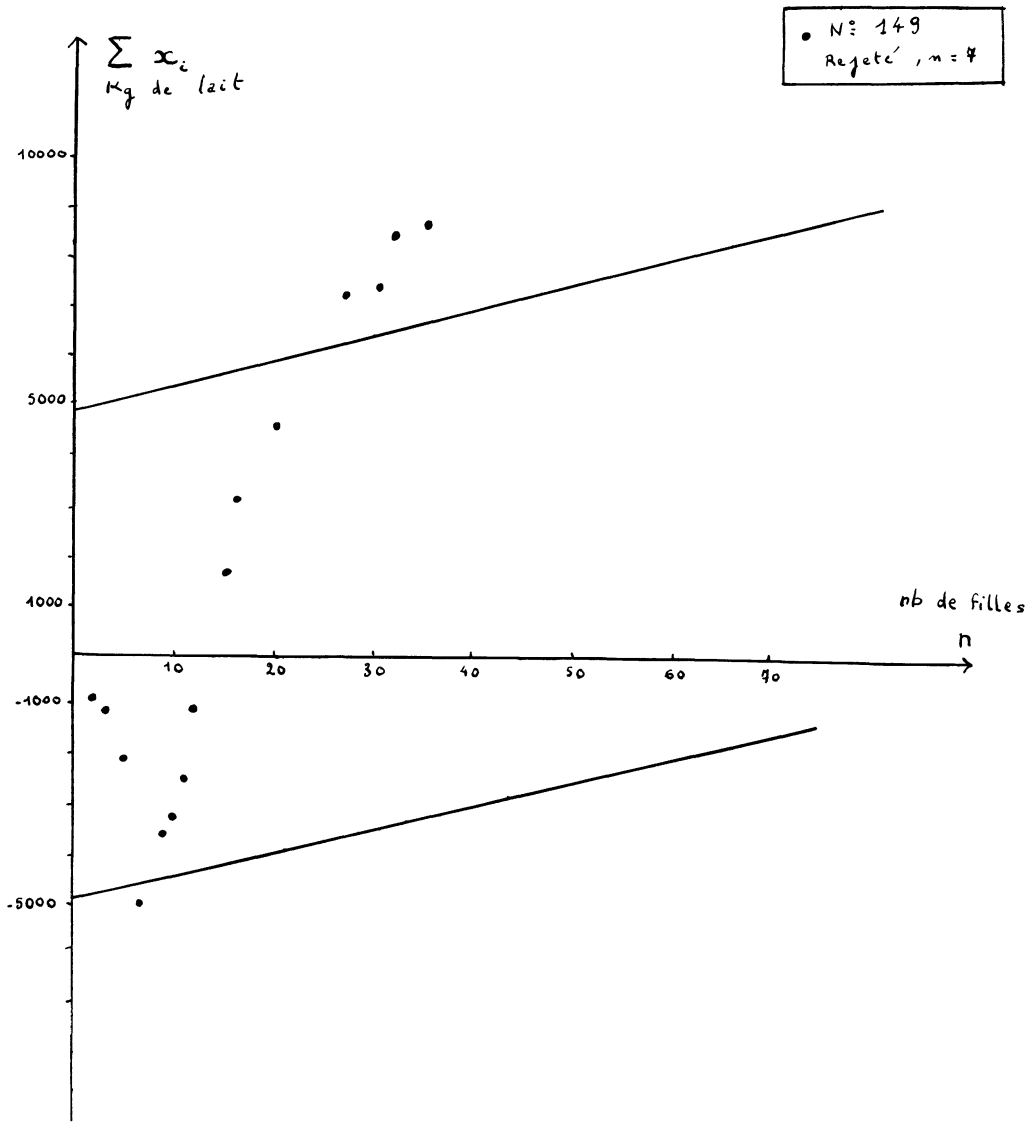
Ordonnée à l'origine des droites de WALD : ± 4900 kg
Pente des droites : + 50 kg

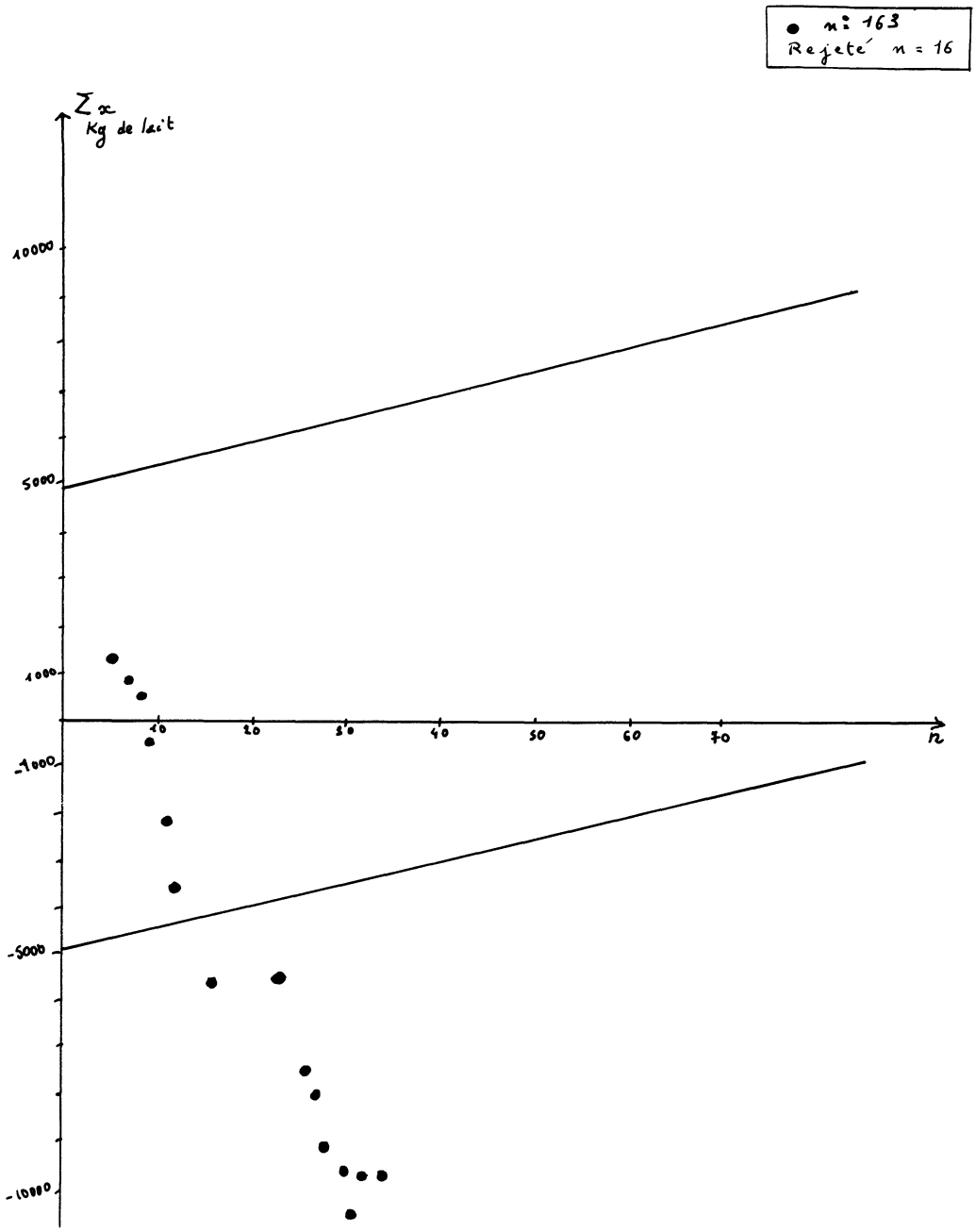
Nous reproduisons ci-après quelques courbes séquentielles parmi les plus significatives.

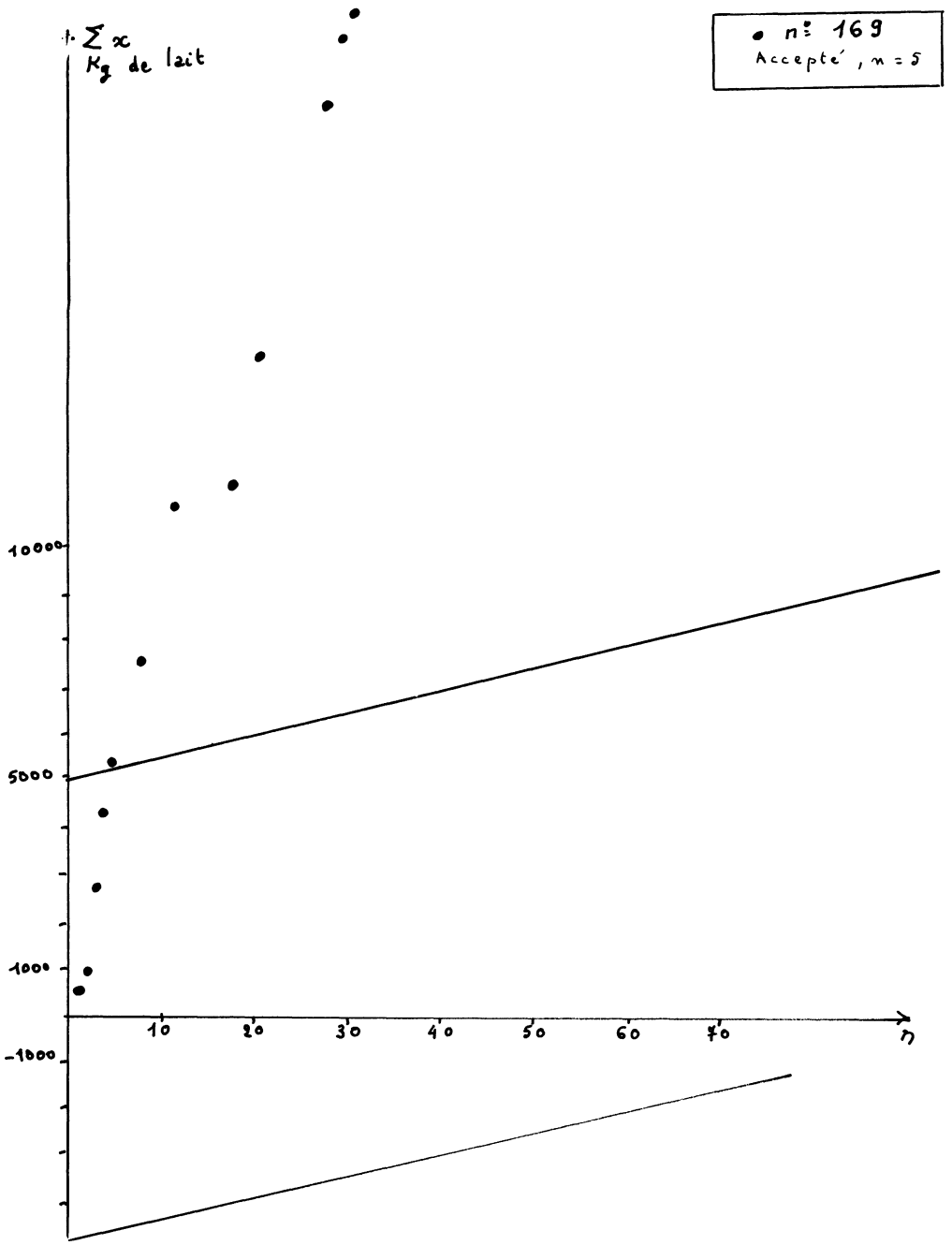
Le nombre de points du graphique ne correspond pas au nombre de résultats. En effet, nous avons eu les données mois par mois. Il peut se faire, par exemple, qu'il y ait trois résultats de lactation arrivant au cours du même mois. Dans ce cas on passera sur le graphique du point d'abscisse n au point d'abscisse n + 3 sans point intermédiaire. Nous joignons à titre explicatif la feuille de calcul correspondant à une des courbes citées (n° 169).

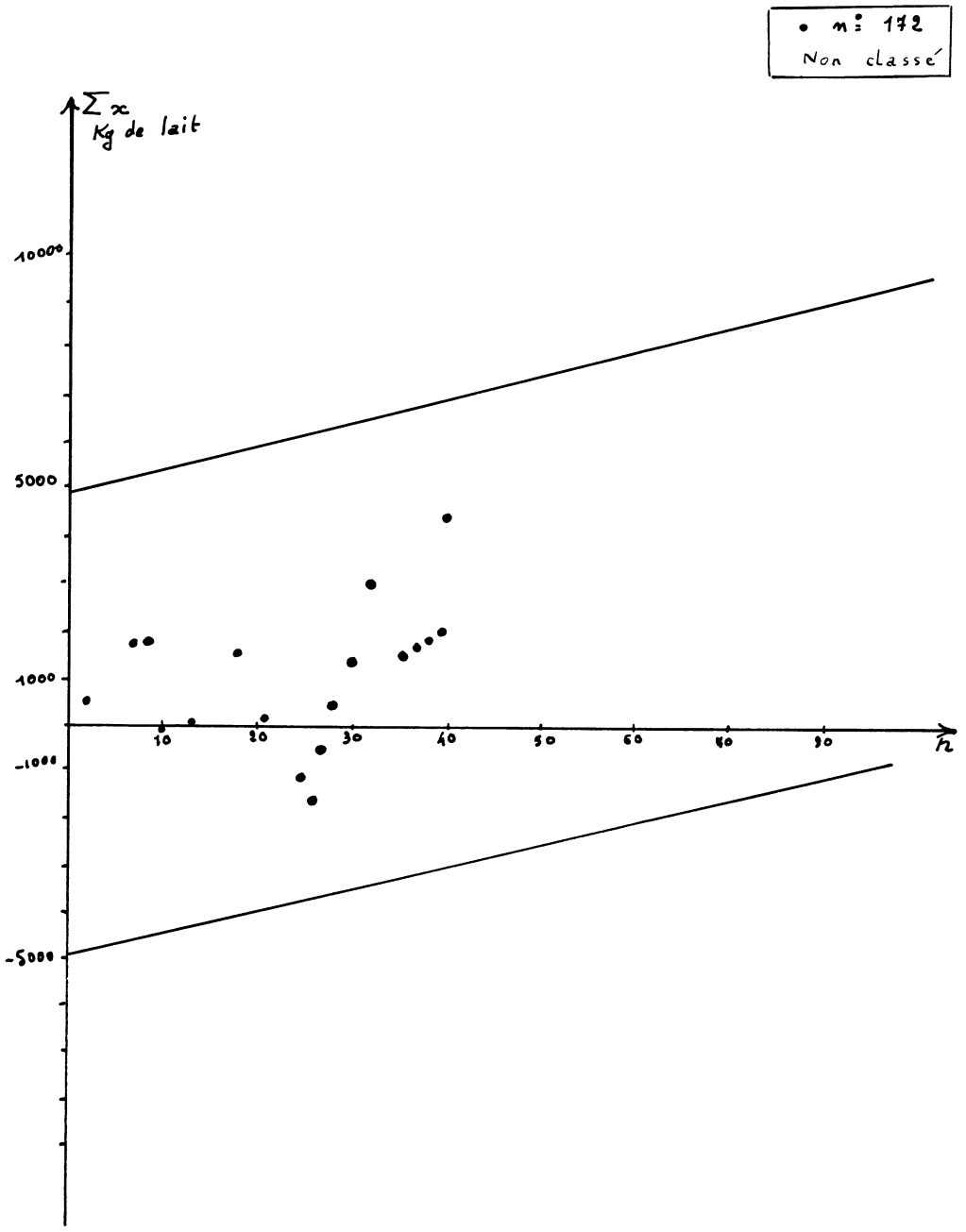
Nous avons fait ce test indépendamment des connaissances à postériori que nous avions. Dès qu'une courbe séquentielle franchit une des droites de WALD, nous prenons la décision correspondante, quel que soit le chemin suivi par la suite (cas du n° 149, rejeté pour n = 7, n'a qu'un seul point dans la zone de rejet et atteint ensuite la zone d'acceptation pour n = 27). Les cas semblables sont signalés à la rubrique "Observations et anomalies".

La courbe du n° 163 montre un cas type de rejet pour n = 16 ; le n° 169, un cas type d'acceptation pour n = 8, le n° 172 un lot non classé.


N° 149
Rejeté, n = 4
Σ xi
Kg de lait
10000
5000
1000
-1000
-5000
nb de filles
n
10
20
30
40
50
60
70


● n = 163
Rejeté n = 16
Σx
Kg de lait
10000
5000
1000
-1000
-5000
-10000
10
20
30
40
50
60
70
n


$\sum x$
Kg de lait
n° 169
Accepté, n = 5
10000
5000
1000
-1000
10
20
30
40
50
60
70
n


• m = 172
Non classé
Σx
Kg de lait
10000
5000
1000
-1000
-5000
10
20
30
40
50
60
70
80
n

TEST SEQUENTIEL

Lot n° : 169

Décision : Accepté, n = 5

Année	Mois	Nb de résultats par mois	N	$\sum_{i=1}^{N} X_i$
1959	1	1	1	529
	2			
	3	1	2	920
	4			
	5			
	6			
	7			
	8			
	9			
	10			
	11			
	12			
1960	1	1	3	2759
	2	1	4	4334
	3	1	5	5333
	4			
	5	3	8	7489
	6			
	7	4	12	10955
	8	6	18	11296
	9	3	21	14067
	10	7	28	19495
	11	2	30	20898
	12			
1961	1			
	2			
	3			
	4	1	31	21460

Lots acceptés par le test séquentiel

N° mécanographique du lot	Nombre d'observations nécessaires
073	34
081	37
098	39
116	27
118	19
127	53
169	5
194	19
196	17
193	31
197	10
227	11
199	32
202	14
208	18
221	17
135	27
209	23
225	13

Observations et anomalies

- Acceptation puis indécision, puis acceptation : 3 cas (n° 116, 118, 135).
- Aucune erreur du genre : acceptation, puis rejet.

Lots rejetés

N° mécanographique du lot	Nombre d'observations nécessaires
076	5
080	33
108	19
113	16
115	11
145	22
149	7
150	18
163	16
166	22
174	22
191	13
211	22
147	7

Observations et anomalies

- Rejet puis indécision :
5 cas (n° 108, 113, 115, 174, 147).
- Rejet, indécision puis rejet :
2 cas (n° 076, 191).
- Rejet, puis acceptation :
2 cas (n° 149, 150).

Lots non classés par le test

N° mécanographique du lot	Nombre total d'observations au moment du test
091	24
099	30
106	2
128	23
117	48
165	38
120	29
171	31
172	40
173	33
183	12
198	36
195	35
226	20
228	12
230	11
232	13

Observations et anomalies

- Au moment du test, nous n'avions pas tous les résultats de certains taureaux qui étaient en cours de testage (n° 226, 228, 230, 232).

- n° 106 (deux résultats).

Si N est le nombre moyen expérimental d'observations nécessaires pour atteindre la décision, le premier test séquentiel donne :

19 lots acceptés
N = 24
14 lots rejetés
N = 17
17 cas non classés par le test.

On remarquera le petit nombre d'observations demandé par ce test, ce qui semble répondre aux souhaits des généticiens et des utilisateurs.

On peut espérer, en fin d'expérience que le test se prononcera sur un plus grand nombre de cas puisque nous n'avions pas tous les résultats du testage au moment de l'expérience (cf. [1]).

En étudiant de près les données et les courbes, il nous a paru évident que bon nombre d'anomalies étaient dues à des résultats de lactations ou trop longues ou trop courtes, d'où l'intérêt par la suite de faire ce test après correction statistique sur la durée ou sur la somme des premiers contrôles.

V-6 - Résultats du 2ème test séquentiel

$$\alpha = \beta = 0{,}05 \quad m_1 = 250 \text{ kg} \quad m_0 = -50 \text{ kg}$$

Ordonnée à l'origine des droites = ± 4900 kg
Pente des droites = + 100 kg, plus raide que dans le test I.

Lots acceptés

N° mécanographique du lot	Nombre d'observations nécessaires
73	34
098	43
116	29
127	58
118	28
169	8
194	19
196	17
193	31
197	10
202	14
208	18
221	17
225	13
135	34
209	23

Observations et anomalies

- Acceptation puis non décision : 1 cas (n° 135).
- Acceptation, non décision, acceptation : 3 cas (n° 116, 118, 209).

Lots rejetés

N° mécanographique du lot	Nombre d'observations nécessaires
080	23
076	5
108	19
115	6
113	15
145	21
150	11
149	7
163	16
166	12
171	18
174	17
191	13
198	18
211	11
232	13
147	7

Observations et anomalies

- Rejet, puis acceptation :
1 cas (n° 149)

- Rejet, puis indécision :
5 cas (n° 150, 171, 174, 198, 147).

- Rejet, indécision, rejet :
6 cas (n° 076, 108, 115, 113, 191, 211).

Lots non classés par le test

N° mécanographique du lot	Nombre d'observations total au moment du test
081	38
091	24
099	30
106	2
128	23
117	48
165	38
120	29
172	40
173	33
183	12
195	35
227	11
199	32
226	20
228	12
230	11

Observations et anomalies (Voir test I).

En résumé, le 2ème test séquentiel donne :

- 16 lots acceptés
N = 25

- 17 lots rejetés
N = 14

- 17 cas non classés.

Si on compare le test II au test I, on voit que le test II est plus sévère que le test I : moins de choix, rejet plus rapide mais moins net.

Les remarques faites pour le test I s'appliquent au test II qui est assez voisin du test I.

VI - CONCORDANCE DE LA METHODE SEQUENTIELLE AVEC LA METHODE ACTUELLE DE SELECTION

Actuellement, on considère que l'index génotypique donne une indication valable sur les taureaux pour la sélection, comme nous l'avons vu précédemment au § IV.

Nous allons voir si les 19 taureaux dont l'index est le plus fort sont les 19 choisis par le test I, si les 17 taureaux dont l'index est le plus faible sont les 17 rejetés par le test I, si les cas d'indécision coïncident avec la zone intermédiaire des index.

VI-1 - Classement des taureaux par index décroissant

On note + si les décisions coïncident,
- si les décisions ne coïncident pas.

Nous avons séparé par un trait horizontal dans le tableau les 19 premiers, puis les 17 suivants, puis les 17 derniers.

Classement des taureaux par index décroissants

Numéro mécanographique du taureau	Index final	Concordance avec les résultats séquentiels
169	791	+
197	652	+
202	522	+
196	309	+
116	300	+
127	291	+
128	268	- non classé
073	263	+
225	244	+
194	240	+
098	190	+
081	183	+
172	175	- non classé
208	128	+
209	126	+
227	112	+
149	109	- rejeté
118	103	+
174	43	- rejeté
183	30	+
199	3	- accepté
091	1	+
173	-68	+
221	-76	- accepté
226	-97	+
193	-108	- accepté
099	-145	+
147	-159	- rejeté
106	-167	+
150	-169	- rejeté
171	-178	+
135	-181	- accepté
120	-193	+
198	-205	+
195	-216	+
165	-229	+

Numéro mécano-graphique du taureau	Index final	Concordance avec les résultats séquentiels
117	-253	- non classé
228	-285	- non classé
076	-345	+
108	-348	+
080	-362	+
230	-404	- non classé
115	-417	+
113	-475	+
166	-483	+
145	-525	+
232	-555	- non classé
211	-593	+
191	-646	+
163	-666	+

VI-2 - Table de contingence

Définitions :

Une classification A divise la population en c classes A_1, A_2 ... A_c (ici c = 3, A est le test séquentiel).

Une classification B divise la pupulation en r classes B_1, B_2 ... B_r (ici r = 3, B est le classement d'après l'index génotypique).

On représente ceci dans un tableau r x c :

n_{ij} est l'effectif qui appartient à la classe A_i et à la classe B_j

$n_{i.}$ est l'effectif de la population de la classe A_i

$n_{.j}$ est l'effectif de la population de la classe B_j

n est l'effectif total

f_{ij}, $f_{i.}$, $f_{.j}$ sont les fréquences correspondantes.

Dans notre expérience, on a le tableau suivant :

	Acceptation par le test séquentiel	Non classement par le test séquentiel	Rejet par le test séquentiel	Total
Meilleurs index	15	2	2	19
Index intermédiaires	4	11	2	17
Index les plus mauvais	0	4	10	14
Total	19	17	14	50

Etude des résultats non concordants : 14 cas

- jugement sur trop peu de données (N < 25) : 5 cas (n° 128, 221, 228, 230, 232).

- erreurs dues au test même : 2 cas (n° 149, 150).

On teste l'indépendance des deux variables "catégorisées" en calculant le carré de contingence (cf. KENDALL II, p. 536 [6] 1943).

$$X^2 = n \left(\sum_{ij} \frac{n_{ij}^2}{n_{i.}\, n_{.j}} - 1 \right)$$

X^2 étant asymptotiquement distribué comme un χ^2 à (r - 1) (c - 1) degrés de liberté

Ici, $X^2 = 35,657$

X^2 est hautement significatif à 0,005 de la dépendance entre ces variables.

Nous allons essayer de mesurer l'association entre les deux méthodes (GOODMAN et KRUSKAL [4] p. 757).

VI-3 - Modèle mesurant la fidélité entre les deux méthodes.

Les classifications A et B sont les mêmes. Ce qui diffère c'est la manière de répartir les individus dans les classes : par exemple, aptitudes d'un individu jugées par deux tests différents, concordance entre les résultats de deux expériences différentes, etc...

Hypothèses :

- Les classifications A et B sont les mêmes

- Les classifications peuvent ou non provenir d'une distribution continue sous-jacente.

- Il existe un ordre sous-jacent entre les classes.

- Nous retenons les mesures proposées par GOODMAN et KRUSKAL [4](1954) à cause de leur interprétation probabiliste évidente.

$m_1 = \sum_a f_{aa}$ = Probabilité de concordance des résultats

$0 \leqslant m_1 \leqslant 1$

$m_1 = 0$: aucun des résultats ne concordent.

$m_1 = 1$: tous les résultats sont concordants. Seule la diagonale principale du tableau est occupée. C'est le cas d'association complète.

$m_2 = \sum_{|a-b|<1} f_{ab}$ = Probabilité de concordance avec une classe voisine

Si on veut, on peut pondérer ces probabilités et définir une mesure du type :

$$m_3 = \sum_a f_{aa} + k \sum_{|a-b|<1} f_{ab}$$

k pouvant être déterminé par des fonctions de perte définies par les impératifs économiques de l'expérience.

Dans notre expérience (cf. tableau), nous avons :

Probabilité de concordance des choix 0,72
Probabilité de choix opposés 0,04
Probabilité de concordance avec une classe voisine..... 0,96

La concordance entre les deux méthodes de sélection est donc assez forte dans notre expérience.

VII - CONCLUSION

D'après les résultats expérimentaux que nous avons obtenus [1] (1964), l'application des tests séquentiels semble justifiée dans le cas de la sélection laitière.

En effet, certaines décisions sont prises plus tôt, la concordance est bonne avec la méthode habituelle et les calculs sont simples à réaliser.

Cette technique de sélection sur tests séquentiels est actuellement appliquée à grande échelle aux résultats des premiers contrôles laitiers (cf. POUTOUS [10] 1966).

Je tiens à remercier Monsieur le Professeur DUGUE, Directeur de l'Institut de Statistique de l'Université de Paris, Monsieur le professeur LAUDET, Directeur du Centre d'Informatique de Toulouse, Monsieur POLY, Directeur de recherches à l'I.N.R.A. (Station centrale de génétique animale à Jouy-en-Josas 78) et son équipe de recherche, spécialement Monsieur POUTOUS (chargé de recherches à l'I.N.R.A.).

BIBLIOGRAPHIE

[1] BOULARD M.F. - Application des tests séquentiels à la sélection en génétique animale. Thèse 3ème cycle. Faculté des Sciences. Paris (1964).

[2] FREBLING J., POUTOUS M. et CALOMITTI S. - Méthode de calcul des index de production laitière. Bull. Techn. d'Inf. 208 (Avril 1966).

[3] GRAYBILL F.A. - An introduction to linear statistical models. Tome I Mac Graw Hill Book Company (1961).

[4] GOODMANN et KRUSKAL - Measures of association for cross classifications. Journal of the American Statistical Association n° 49 (p. 732) 1954 ; n° 53 (p. 814) 1958 ; n° 54 (p. 123) 1959 ; n° 58 (p. 310) 1963.

[5] KEMPTHORNE O. - The Design and analysis of experiments. John WILEY New York (1952).

[6] KENDALL M.G. and STUART A. - The advanced theory of statistics. Charles Griffin London (1943).

[7] POLY J. et VISSAC B. - Interprétation des résultats du contrôle laitier en vue du testage des taureaux.Bulletin Technique d'Information des Ingénieurs des Services Agricoles n° 145 (1959).

[8] POLY J., POUTOUS M. et FREBLING J. - Méthode de calcul d'index de production laitière. Bulletin Technique d'Information des Ingénieurs des Services Agricoles n° 205 (Décembre 1965).

[9] POUTOUS M. et VISSAC B. - Recherche théorique des conditions de rentabilité maximum de l'épreuve de descendance des taureaux d'insémination artificielle. Annales Zootechnique 11 (4) p. 233 (1962).

[10] POUTOUS M. - Communication personnelle (1966). Station Centrale de Génétique Animale C.N.R.Z. 78 Jouy-en-Josas.

[11] WALD A. - Sequential Analysis. John WILEY New York (1947).