Ann. Génét. Sél. anim., 1976, **8** (4), 501-509.

NOMBRE D'OBSERVATIONS A FAIRE POUR UNE SÉLECTION EFFICACE SUR UN CARACTÈRE DICHOTOMIQUE

Paule RENAUD

avec la collaboration technique de S. BACH

Institut National Agronomique Paris-Grignon, 16, rue Claude-Bernard Paris (5e)

RÉSUMÉ

Dans une expérimentation destinée à la sélection sur un caractère présentant deux modalités, l'une favorable, l'autre défavorable, on propose de mesurer l'efficacité du dispositif expérimental par la probabilité qu'il a de permettre de détecter le meilleur individu de la population à sélectionner. A l'aide d'une distribution de probabilité reflétant la connaissance qu'a le sélectionneur de la fréquence de la modalité recherchée dans sa population, la probabilité de détecter le meilleur est calculée. C'est une fonction croissante de la taille de l'expérimentation. Dès lors, cette taille peut être choisie de manière à garantir une efficacité requise au départ.

INTRODUCTION

Dans une étude de sélection animale, on est fréquemment confronté à des caractères dichotomiques. Le cas le plus trivial est celui qui, naturellement ne présente que deux modalités par exemples, présence ou absence d'un gène, nouveau né mort ou vivant, animal fécondé ou non. Il peut arriver aussi qu'un caractère qualitatif présente plus de deux modalités mais que, l'une d'entre elles étant particulièrement intéressante, l'alternative soit créée en l'opposant à la réunion de toutes les autres. Enfin, même un caractère quantitatif discret ou continu est parfois rendu dichotomique par la définition d'un seuil ; les deux modalités étant alors « inférieur au seuil » et « supérieur au seuil » ; par exemple, chez le lapin, en dessous d'un certain nombre de descendants sevrés par portée une femelle peut être éliminée parce qu'insuffisamment prolifique; au dessus elle sera considérée comme bonne.

Il est proposé ici d'étudier un dispositif expérimental destiné à la sélection sur un caractère présentant deux modalités, l'une considérée comme favorable, l'autre non. Dans le cas où la population à sélectionner est relativement petite on calcule la

probabilité de détecter le meilleur individu, c'est-à-dire celui qui correspond à la plus grande probabilité de la modalité favorable. Cette probabilité P* de détecter le meilleur est une fonction croissante de la taille des lots expérimentaux ce qui permet, réciproquement, de déterminer la taille de l'expérimentation assurant d'une probabilité prédéterminée P* de détecter le meilleur. Les calculs s'appliquent de manière symétrique à la détection du plus mauvais. Cette étude fait naturellement suite à une recherche analogue pour une sélection sur un caractère quantitatif gaussien (Paule RENAUD, 1976).

CHOIX DE LA MÉTHODE

Des calculs de probabilité de garder le meilleur pour un caractère dichotomique menés par NEBENZAHL et SOBEL (1972) et SOBEL et HUYETT (1957) aboutissaient à des résultats que les utilisateurs de statistique ont souvent peine à exploiter. La méthode employée y est celle de la « zone d'indifférence » dont voici le principe : la probabilité P* de détecter le meilleur dépend des probabilités de la modalité favorable A attachées à tous les individus c'est-à-dire d'autant de paramètres p_1, p_2,... p_r que d'individus dans la population. Pour exploiter le calcul, il fallait éliminer tous ces paramètres. En un premier temps un nombre Δ était choisi ($0 \leqslant \Delta \leqslant 1$) et on admettait que si la différence des probabilités de la modalité A entre le meilleur et le suivant était inférieure à Δ, le sélectionneur considérait qu'il lui était indifférent de détecter ou non le meilleur puisque le second lui était presque équivalent. Ensuite, hors de la zone d'indifférence P* dépendant toujours de tous les paramètres p_i, les auteurs en cherchaient le minimum et proposaient donc de travailler sur la quantité :

$$\min_{p_1, p_2, \ldots p_r} P^*(p_1, p_2, \ldots p_r)$$

ce minimum étant pris dans le domaine où $(\forall i)$, $\sup_i p_i - p_i \geqslant \Delta$.

L'avènement des méthodes bayesiennes habitue les utilisateurs de probabilités et statistique à mettre, sur l'espace des paramètres inconnus, une distribution de probabilité *a priori* et c'est ce qui est proposé ici. Une loi de probabilité sur les paramètres p_1, p_2... p_r qui rend compte de leur distribution dans la population permet de calculer la probabilité de trouver le meilleur en éliminant les paramètres p_1, p_2,... p_r par intégration. On prend l'espérance de P* (p_1, p_2,... p_r) relativement à la distribution *a priori* de p_1, p_2,... p_r.

Le choix de la distribution *a priori* n'est en rien plus compliqué que celui de la zone d'indifférence Δ, il faut dans l'un et l'autre cas que l'expérimentateur fasse appel à sa connaissance pratique du caractère sur lequel il sélectionne. De plus, la loi *a priori* permet un calcul exact de P* évitant ainsi la perte de puissance due à l'utilisation d'un minimum.

MODÈLE, CALCULS ET RÉSULTATS

Appelons A la modalité favorable du caractère et $\bar{A}$ l'autre. Une épreuve étant définie, l'individu π subissant cette épreuve répond A avec la probabilité p (par exemple, dans une population animale l'épreuve pour π peut consister à considérer l'un de ses descendants à la naissance, A est la réponse né-vivant; $\bar{A}$ la réponse né-mort; p est la probabilité attachée à π que l'un, au hasard, de ses descendants directs soit né-vivant).

p est considérée comme une variable aléatoire dont la loi de probabilité est celle de sa répartition dans la population globale. Nous admettons que l'expérimentateur est capable de se donner cette loi de probabilité. En effet, il est rarissime que l'on s'engage dans un processus de sélection sur une modalité A d'un caractère sans avoir une bonne idée de ce qu'est la fréquence moyenne de A dans la population.

On prend pour espérance de p cette fréquence. On a souvent aussi une estimation (parfois grossière) de la variance de p; il s'agit d'une variance inter-individus. Sur cette espérance et cette variance on peut ajuster une distribution type : la distribution d'une loi bêta.

En effet, des raisonnements logiques (RAIFFA et SCHAIFER, 1961 et FERGUSSON, 1967) plaident en faveur d'une distribution *a priori* bêta pour une probabilité lorsque les observations sont binomiales. On dit que les distributions binomiale et bêta sont conjuguées. De plus, une loi de probabilité bêta dépendant de deux paramètres peut s'ajuster sur des distributions assez variées définies sur (0,1) et donner satisfaction dans la plupart des cas.

Rappelons donc qu'une variable aléatoire de loi bêta a une fonction de densité dépendant des paramètres α et β qui s'écrit :

$$f(x) = \frac{(\alpha + \beta - 1)!}{(\alpha - 1)!\,(\beta - 1)!}\, x^{\alpha - 1}\,(1 - x)^{\beta - 1} \qquad \text{pour } 0 \leqslant x \leqslant 1$$
$$= 0 \quad \text{ailleurs}$$

son espérance est $\dfrac{\alpha}{\alpha + \beta}$

sa variance : $\dfrac{\alpha\beta}{(\alpha + \beta)^2\,(\alpha + \beta + 1)}$

Les r individus $\pi_1, \pi_2, \ldots \pi_r$ parmi lesquels la sélection opère sont supposés tirés indépendamment les uns des autres d'une population globale dans laquelle la probabilité p de réaliser A est régie par la loi bêta choisie. Ceci implique que les probabilités qui leurs sont attachées $p_1, p_2, \ldots p_r$ sont des variables aléatoires indépendantes et de même loi bêta. En conséquences pratiques de cette hypothèse d'indépendance apparaissent certaines contraintes sur les individus $\pi_1, \pi_2, \ldots \pi_r$ pour lesquels ce modèle peut être appliqué. Ils doivent, par exemple, être non apparentés.

Il faut maintenant acquérir de l'information sur les probabilités inconnues $p_1, p_2, \ldots p_r$ attachées respectivement aux individus $\pi_1, \pi_2, \ldots \pi_r$ de manière à pouvoir procéder à la sélection. Choisissons un mode d'expérimentation qui consiste à observer chaque individu m fois avec toutes les conditions d'indépendance qui permettent de résumer l'information par r variables binomiales :

> Z_i est l'aléatoire égale au nombre de fois où A se réalise parmi m essais sur π_i $(i = 1, 2, \ldots r)$.
>
> Z_i a une loi binomiale d'effectif m et de probabilité p_i (p_i étant considéré comme une aléatoire il s'agit d'une loi conditionnée par p_i).
>
> Les couples aléatoires (Z^i, p_i), $(i = 1, 2, \ldots r)$ sont indépendants puisque les individus $\pi_1, \pi_2, \ldots \pi_r$ le sont.

Le meilleur parmi $\pi_1, \pi_2, \ldots \pi_r$ est celui qui a la probabilité la plus grande de réaliser A. Nous nous proposons de calculer la probabilité de le détecter grâce à une telle expérimentation. Il faut peut-être commencer par se demander si c'est bien là une mesure de l'efficacité de l'expérimentation quant à la sélection.

Répondons d'abord ceci : Si nous travaillons sur une très grande population ($r = 500$ par exemple), peu importe de connaître le meilleur. En effet, on a de fortes chances pour que le meilleur et les suivants soient très rapprochés et alors, savoir lequel exactement est le meilleur est peu important. Par contre, si l'on a deux individus ($r = 2$) dont l'un doit être éliminé il est fondamental de savoir lequel est le

meilleur. Si, entre ces deux extrêmes on a un petit nombre r d'individus il est important d'avoir une bonne probabilité de mettre en évidence une vedette éventuelle.

Nous calculerons donc la probabilité de détecter le meilleur pour r variant de 2 à 9. Pour une population plus nombreuse il faudra changer de critère et se contenter de garder un lot d'entre eux qui, en moyenne, correspondra à des valeurs de p plus grandes que celles de la population initiale.

Nous nous plaçons donc dans le cas où r est petit et calculons la probabilité de détecter le meilleur c'est-à-dire l'individu $\pi_{(i)}$ correspondant à la probabilité $p_{(i)} = \sup_i p$.

($\pi_{(i)}$ a une probabilité 1 d'être unique puisque, la loi de $p_1, p_2, \ldots p_r$ étant continue, la probabilité d'avoir deux individus strictement équivalents est nulle.)

Le meilleur est détecté si la variable aléatoire $Z_{(i)}$ correspondant à $\pi_{(i)}$ prend une valeur supérieure à celles de toutes les autres variables Z_i [$i \neq (i)$]. Il faut en effet que celui auquel est attachée la plus grande valeur de p réalise la modalité A plus souvent que les autres. La probabilité P* de cet événement est donnée par la suite de calculs suivants :

On pose :

$$M(z_1, p_1) = \frac{1}{m+\alpha+\beta-1} \sum_{z_2=0}^{z_1-1} \frac{\binom{m}{z_2}}{\binom{m+\alpha+\beta-2}{z_2+\alpha-1}} \times \sum_{k=z_1+\alpha}^{m+\alpha+\beta-1} \binom{m+\alpha+\beta-1}{k} p_1^{\,k} (1-p_1)^{m+\alpha+\beta-1-k}$$

et alors :

$$P^* = r \left[\frac{(\alpha+\beta-1)!}{(\alpha-1)!\,(\beta-1)!} \right]^r \sum_{z_1=1}^{m} \binom{m}{z_1} \times \int_0^1 p_1^{\,z_1+\alpha-1} (1-p_1)^{m-z_1+\beta-1} \left[M(z_1, p_1) \right]^{r-1} dp_1$$

P* a été calculée par des méthodes d'intégration numérique (sur le CII, MITRA 15 de l'Institut National Agronomique Paris-Grignon) pour toutes les valeurs entières de α et β comprises entre 1 et 20 et pour $m = 1, 2, 3, 5, 10, 20, 30, 50$; r variant de 2 à 9.

Les résultats sont très partiellement consignés sous forme d'abaques dans les figures 1 et 2.

EXPLOITATION DES ABAQUES (fig. 1 et 2)

— Du point de vue technique :

1. Elles sont présentées sur du papier semi-logarithmique. Leur extrême régularité autorise les interpolations tenant compte de l'échelle logarithmique sur m.

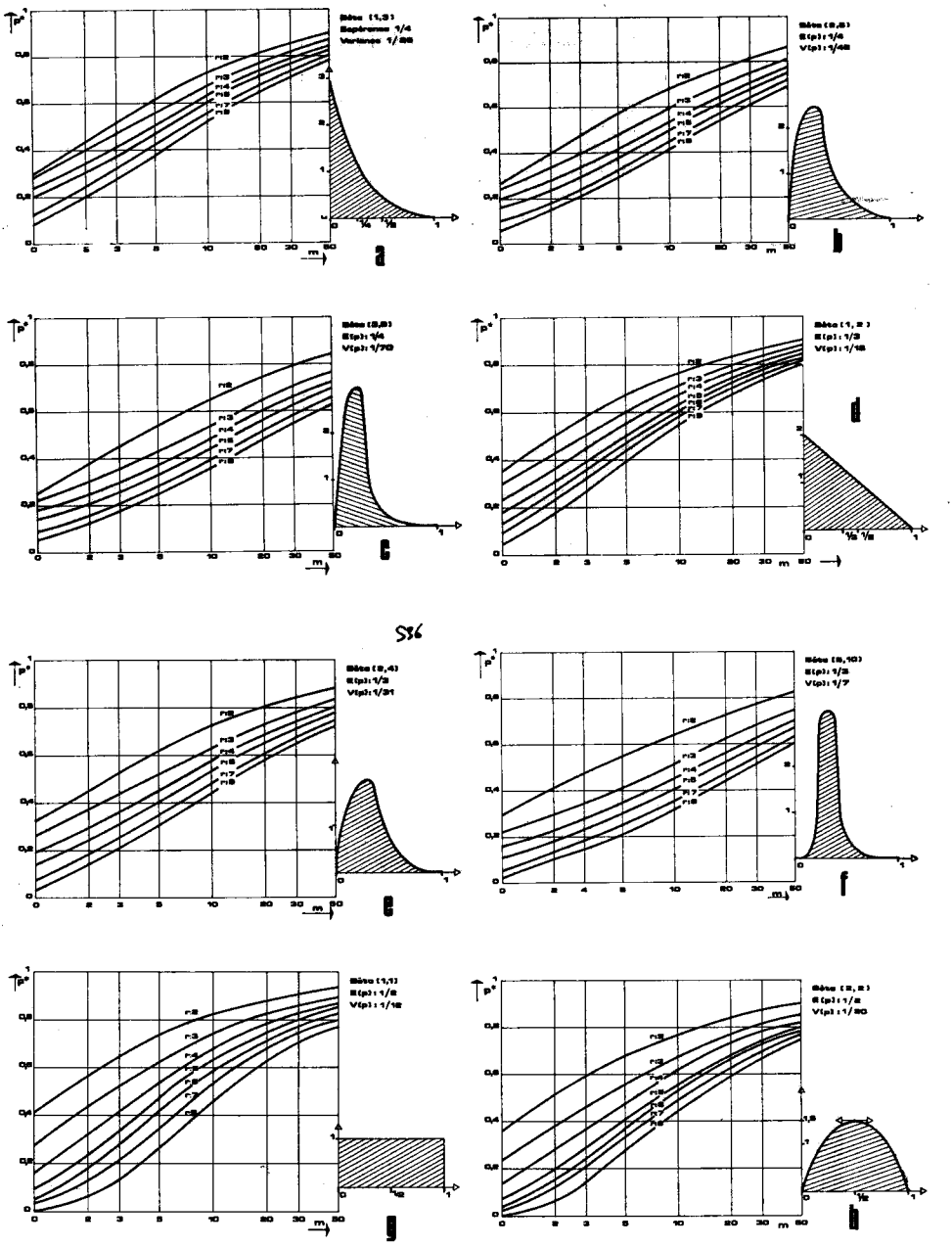

FIG. 1. — *Probabilité P* de détecter le meilleur d'une population d'effectif r (r = 2, 3, 4, 5, 7, 9) en fonction de la taille m des lots expérimentaux selon la loi a priori bêta (α, β) de la probabilité p de la qualité sélectionnée*

a, b, c pour une loi bêta centrée sur 1/4
d, e, f pour une loi bêta centrée sur 1/3
g, h pour une loi bêta centrée sur 1/2

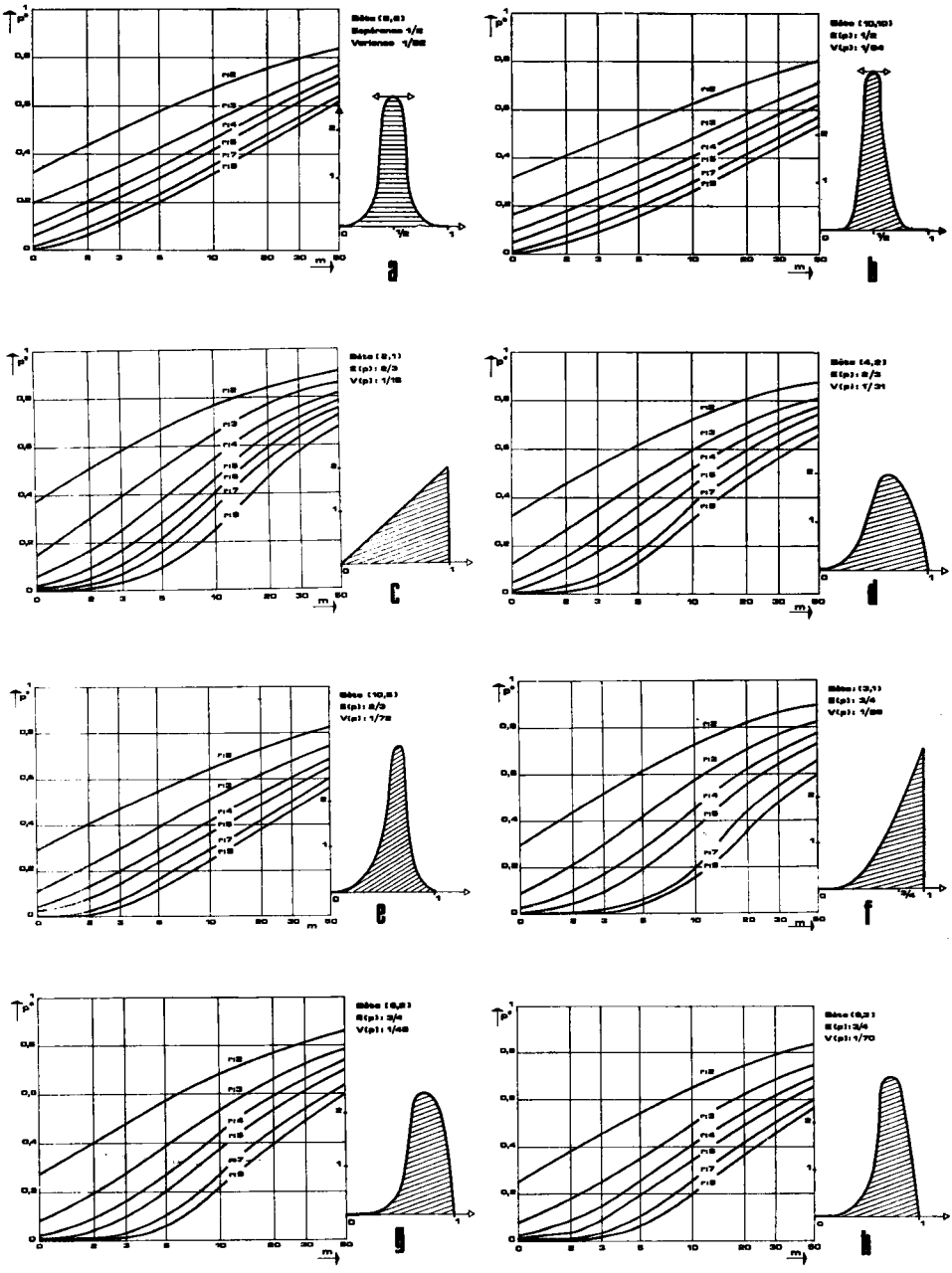

FIG. 2. — *Probabilité P* de détecter le meilleur d'une population d'effectif r (r = 2, 3, 4, 5, 7, 9) en fonction de la taille m des lots expérimentaux selon la loi a priori bêta (α, β) de la probabilité p de la qualité sélectionnée*

a, b pour une loi bêta centrée sur 1/2
c, d, e pour une loi bêta centrée sur 2/3
f, g, h pour une loi bêta centrée sur 3/4

2. Pour $r = 2$, la valeur de P* est la même à m fixé pour le couple de paramètres (α, β) et pour le couple (β, α). Un raisonnement probabiliste considérant les rôles symétriques joués par α et β et remarquant que « détecter le meilleur » est mathématiquement le même problème que « détecter le plus mauvais » si on n'a que deux individus permet de le démontrer.

— Du point de vue du choix de la distribution *a priori* :

1. Il faut remarquer que plus grande est la variabilité de p dans sa distribution *a priori*, plus grande est la probabilité P* (il est évident que le meilleur a plus de chances de s'écarter des suivants si p est très variable).

2. Par contre, à variances égales, la probabilité de détecter le meilleur avec une loi *a priori* bêta (α, β), $(\alpha < \beta)$ est supérieure à la probabilité P* correspondant à une loi *a priori* (β, α). Ceci est dû à la dissymétrie de la distribution bêta.

3. Plus la population dans laquelle s'exerce la sélection est grande (r) ; plus petite est la probabilité de détecter le meilleur (le meilleur et le suivant sont en général plus proches dans une grande population).

En conséquence, la méthode la plus sérieuse pour choisir les paramètres de la distribution bêta est l'ajustement des deux premiers moments centrés. Il faut néanmoins retenir que puisque P* croît lorsque la variance de la loi *a priori* croît, pour ne pas risquer de surestimer P*, il vaut mieux légèrement sous-estimer la variance. Il est, en particulier déconseillé de choisir une loi *a priori* « non informative » de grande variance, le calcul perdrait sa signification.

Le choix de l'espérance est assez facile puisqu'il correspond en somme à la fréquence de A dans la population. On aura souvent une valeur à attribuer à $\dfrac{\alpha}{\alpha + \beta}$

Le choix de la variance est plus scabreux. L'utilisateur considérant ce qu'il connaît en ce qui concerne la fréquence de A, doit faire la part de la variance inter-individus. Seule cette variance entre individu doit être ajustée à $\dfrac{\alpha\beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)}$ achevant ainsi la détermination des deux paramètres.

La méthode traditionnelle d'ajustement de la loi *a priori* bêta qui consisterait à prendre pour α le nombre de fois où A a été observé antérieurement dans un échantillon de taille $\alpha + \beta$ est mauvaise ici. Elle répond à une théorie qui ne s'applique pas à notre problème puisqu'elle néglige le fait que, au départ, nous avons affaire à des individus différents.

CONCLUSION

On peut constater que, même dans des populations de petites tailles, pour un nombre d'observations usuel, la probabilité de détecter le meilleur n'est pas bien grande. (Il en est de même par symétrie pour la probabilité de détecter le plus mauvais.)

Pour éclairer cette conclusion, donnons en exemple le tableau des tailles m des 4 lots expérimentaux qui assurent d'une probabilité $P^* = 0{,}7$ de détecter le meilleur parmi 4 individus selons des valeurs de α et β variant de 1 à 5.

α \ β	1	2	3	4	5
1	12	13	17	20	23
2	20	18	20	23	26
3	30	23	25	27	31
4	41	29	30	30	34
5	50	35	34	36	39

Le choix de m est évidemment postérieur à celui des paramètres α et β et l'importance du choix de la distribution *a priori* est illustrée dans ce tableau par la diversité des valeurs de m trouvées.

Cette étude permet donc de déterminer la taille des lots expérimentaux et n'envisage que le cas d'expériences équilibrées (m est l'effectif commun à tous les individus). En effet, on a tout intérêt, lorsqu'on met un dispositif expérimental à essais simultanés en place à le choisir équilibré (même si, dans le cours de sa réalisation un déséquilibre risque de s'établir). Cependant il ne faut pas perdre de vue que le plan d'expérience le plus efficace pour la recherche du meilleur individu ou d'un meilleur traitement est celui qui s'appelle en anglais « play the winner » (SOBEL et WEISS, 1970 et 1971; HOEL *et al.*, 1972; SIMON *et al.*, 1975). Ce plan est applicable à une étude séquentielle où, avant d'introduire un nouvel élément expérimental, le résultat du précédent est connu. Il permet ainsi de limiter le nombre d'essais sur les moins bons et est donc particulièrement adapté à la recherche du meilleur parmi plusieurs traitements médicaux. En expérimentation animale où l'on rechercherait le meilleur reproducteur ce dispositif est inapplicable car son aspect séquentiel induirait une expérience de trop longue durée.

On s'est placé ici avant l'expérience mais, une fois le dispositif réalisé les variables Z_1, Z_2,... Z_r ont pris des valeurs expérimentales et a posteriori, la probabilité d'avoir détecté le meilleur vaudrait d'être calculée afin d'éviter des décisions de sélection dont les résultats seraient trop peu sûrs.

Enfin, les résultats numériques ne figurant pas sur les abaques sont tenus à la disposition du lecteur à l'Institut National Agronomique Paris-Grignon.

Reçu pour publication en décembre 1976.

SUMMARY

NUMBER OF OBSERVATIONS FOR AN EFFICIENT SELECTION ON A DICHOTOMIC CHARACTER

In an experimentation in view of a selection on a qualitative character with two modalities, one being considered as good, the other worst, it is proposed to measure the efficiency of the experimental design by its probability of allowing to detect the best element of the population.

With the help of a probability distribution reflecting the selector's knowledge about the frequency of the good modality in his population, the probability of detecting the best one of the population is reckoned. It is an increasing function of the size of the experimental design.

Conversely, the number of observations can be chosen to guarantee a predetermined efficiency.

RÉFÉRENCES BIBLIOGRAPHIQUES

FERGUSSON, T. S., 1967. *Mathematical Statistics. A Decision Theoretic Approach.* Academic Press. New York.

HOEL, D. G. SOBEL, M. and WEISS, G. H., 1972. A two Stage procedure for choosing the better of two binomial populations. *Biometrika*, **59**, 317-22.

NEBENZAHL, E. and SOBEL, M., 1972. Play-the-winner sampling for a fixed sample size binomial selection. *Biometrika*, **59**, 1-8.

RAIFFA, H. and SCHLAIFER, R., 1961. *Applied Statistical Decision Theory. Division of research Harward University*, Boston.

RENAUD P., 1976. Probabilité de garder le meilleur lors d'une sélection. *Revue de Statist. Appliquée*, **24**, 5-23.

SIMON, R., WEISS, G. H. and HOEL, D. G., 1975. Sequential analysis of binomial clinical trials. *Biometrika* **62**, 195-200.

SOBEL, M. and HUYETT, M., 1957. Selecting the best one of several binomial populations. *Bell system tech. J.*, **36**, 537-576.

SOBEL, M. and WEISS, G. H., 1970. Play-the-winner sampling for selecting the better of two binomial populations. *Biometrika*, **57**, 357-365.

SOBEL, M. and WEISS, G. H., 1971 *b*. Play-the-winner rule and inverse sampling in selecting the better of two binomial populations. *J. Am. Statist. Ass.*, **66**, 545-551.