REVUE DE STATISTIQUE APPLIQUÉE

PAULE RENAUD

Probabilité de garder le meilleur lors d'une sélection

Revue de statistique appliquée, tome 24, n° 1 (1976), p. 5-23

<http://www.numdam.org/item?id=RSA_1976__24_1_5_0>

PROBABILITÉ DE GARDER LE MEILLEUR LORS D'UNE SÉLECTION [1]

Paule RENAUD
INA 16, rue Claude Bernard - Paris 5e

Il s'agit d'une sélection dans une population de taille N sur laquelle existe un critère qui permet un classement des individus. On se propose de ne garder qu'un sous-ensemble d'éléments contenant les meilleurs et notamment le meilleur.

Le problème serait simple si l'ordre était connu dans la population mais il ne l'est pas. On recueille une information partielle par une mesure indirecte et, selon la qualité de l'information, la probabilité d'avoir le meilleur parmi les individus sélectionnés est plus ou moins grande. Cette probabilité est calculée dans la première partie de ce travail en utilisant une mesure à priori sur les paramètres du modèle. En un deuxième temps, on étudie les plans d'expérience susceptibles de donner sur les individus une information suffisante pour que la sélection réalise l'objectif fixé (garder le meilleur).

I – INTRODUCTION

La situation est celle où, dans une population de taille N on sélectionne un sous-ensemble.

On appelle p la probabilité qu'un individu quelconque soit gardé et la taille du sous-ensemble sélectionné a pour espérance Np. La population globale est ordonnée par une relation du style "meilleur que" définie en affectant à chaque élément la valeur d'un paramètre réel θ et en considérant l'ordre induit, sur la population, par l'ordre des nombres réels. La sélection a pour but de garder les meilleurs c'est-à-dire ceux auxquels sont attachés les plus grands paramètres.

Cependant, le paramètre de chaque individu est inconnu et par conséquent, l'ordre aussi est inconnu. On recueille, sur θ une information partielle sous forme d'estimation par une méthode de régression. La précision de l'information est résumée, en hypothèse de normalité, grâce à la valeur d'un coefficient de corrélation ρ et selon la précision, la sélection a une probabilité plus ou moins grande de garder le meilleur. Si on appelle P^* cette probabilité, P^* s'exprime en fonction de ρ et c'est la probabilité $P^*(\rho)$ qui est étudiée dans la première partie sous les hypothèses de loi à priori normale du paramètre et d'indépendance entre les individus. Dans la seconde partie on cherche des plans d'expérience qui assurent une information suffisante c'est-à-dire une valeur de ρ assez grande pour que la probabilité de garder le meilleur soit égale à un nombre donné à l'avance.

---

(1) Article remis en juin 1974, révisé en Décembre 1974.

Mots-clés : Sélection, Distribution a priori, Puissance

II – MODELE MATHEMATIQUE

Pour définir la relation d'ordre on attache à chaque élément de la population globale étudiée, un paramètre réel θ. Aux N individus correspondent donc N valeurs réelles $\theta_1, \theta_2, \ldots, \theta_N$. On dit que l'individu d'indice i est meilleur que celui d'indice i' si et seulement si $\theta_i > \theta_{i'}$. Le meilleur correspond au plus grand des paramètres θ_i (le cas d'ex-aequo sera éliminé par la suite).

Mettons une loi à priori sur θ. Le paramètre θ attaché à un individu est supposé être une valeur prise par une aléatoire A qui, dans la population globale, a une loi de probabilité normale centrée. Notons ceci :

$$A : N(0, \sigma^2)$$

A est une aléatoire dont on ne peut observer les valeurs. Elle est prédite grâce à une aléatoire Z par une technique de régression et on suppose le couple (A, Z) bi-normal. On écrit donc :

$$A = Z + E$$

Z et E sont des aléatoires indépendantes

$$\sigma^2 = \sigma'^2 + \sigma''^2 > \sigma'^2 \qquad \text{où} \quad \sigma'^2 \text{ est la variance de Z}$$

$$\text{et} \quad \sigma''^2 \text{ la variance de E}$$

$$\text{cov}\,(A, Z) = \sigma'^2$$

$$\rho = \frac{\sigma'}{\sigma} \quad \text{est le coefficient de corrélation entre A et Z.}$$

La sélection s'effectue sur les valeurs prises par la variable observable Z. Ecrivons sa loi de probabilité. La régression de Z sur A donne :

$$Z = \rho^2 A + G$$

A et G sont indépendantes et la variance de G est $\sigma'^2 (1 - \rho^2)$. La loi de Z conditionnée par $A = \theta$ est donc normale d'espérance $\rho^2 \theta$ et de variance $\sigma'^2 (1 - \rho^2)$.

La loi à priori de Z est normale d'espérance zéro et de variance σ'^2.

La procédure de sélection est définie par la règle suivante :
Un élément i est gardé si la variable Z_i qui lui est attachée prend une valeur supérieure à un certain seuil λ déterminé par :

$$P(Z > \lambda) = p \qquad (1)$$

où p est une proportion choisie à l'avance.

On appelle F la fonction de répartition de l'aléatoire normale réduite et t le point tel que $F(t) = 1 - p$ et le seuil λ est déterminé par :

$$\lambda = \sigma' t$$

III – CALCULS DE LA PROBABILITE DE GARDER LE MEILLEUR

Il nous faut supposer que les N éléments parmi lesquels on sélectionne sont obtenus grâce à un échantillonnage correct dans leur population, c'est-à-dire qu'ils ont été choisis au hasard indépendamment les uns des autres. Cet échantillon correspond à N aléatoires $A_1, A_2, \ldots A_N$ indépendantes et de loi $N(0, \sigma^2)$.

Le meilleur individu correspond à la plus grande valeur des A_i soit à l'aléatoire $S = \sup A_i$.

(La probabilité d'avoir des ex-aequo est nulle : il y a donc un meilleur).

Appelons i* l'indice du meilleur et Z_{i*} l'aléatoire Z qui lui correspond. Conditionnellement au fait que l'aléatoire S prend la valeur s, la loi de Z_{i*} est:

$$N[\rho^2 s, \sigma'^2 (1 - \rho^2)]$$

Si g(s) est la densité de probabilité de S on peut écrire la probabilité P* de garder le meilleur :

$$P^* = \int_{-\infty}^{\infty} [P_s (Z_{i*} > \lambda)]\, g(s)\, ds$$

où $P_s(Z_{i*} > \lambda)$ est la probabilité, conditionnée par $S = s$, que le meilleur soit gardé.

La fonction qui est sous le signe d'intégration s'exprime grâce à la fonction de répartition F de l'aléatoire normale réduite :

$$P_s[Z_{i*} > \lambda] = F\left(\frac{\rho^2 s - \lambda}{\sigma'\sqrt{1 - \rho^2}}\right)$$

g(s) s'obtient en écrivant que :

$$P(S < s) = P\{A_i < s\ ;\ i = 1, 2, \ldots . N\} = F^N\left(\frac{s}{\sigma}\right)$$

d'où
$$g(s)\, ds = d\left[F^N\left(\frac{s}{\sigma}\right)\right]$$

et
$$P^* = \int_{-\infty}^{\infty} F\left(\frac{\rho^2 s - \lambda}{\sigma'\sqrt{1 - \rho^2}}\right) d\left[F^N\left(\frac{s}{\sigma}\right)\right]$$

Par intégration par parties évidentes :

$$P^* = 1 - \frac{1}{\sqrt{2\pi}} \frac{\rho^2}{\sigma'\sqrt{1 - \rho^2}} \int_{-\infty}^{+\infty} F^N\left(\frac{s}{\sigma}\right) e^{-\frac{1}{2}\left[\frac{\rho^2 s - \lambda}{\sigma'\sqrt{1 - \rho^2}}\right]^2} ds$$

On peut simplifier l'intégrale afin que n'apparaissent qu'un nombre minimum de paramètres :

$$\lambda = \sigma' t\ ; \frac{\sigma'}{\sigma} = \rho\ ; \quad \text{et pour} \quad \rho \neq 0 \quad \text{et} \quad \rho \neq 1$$

on fait le changement de variables :

$$\frac{\rho}{\sqrt{1-\rho^2}}\left(\frac{s}{\sigma}-\frac{t}{\rho}\right)=u$$

On obtient ainsi :

$$P^* = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} F^N \left[\frac{1}{\rho} (\sqrt{1-\rho^2} \, . \, u + t) \right] . \, e^{-\frac{u^2}{2}} \, du$$

Cette probabilité ne dépend que de ρ, N et p. Les calculs ont été faits pour ρ variant de 0,05 en 0,05 entre 0,05 et 0,95 ; pour N variant de 10 en 10 de 10 à 100 et pour p = 0,05 ; 0,10 ; 0,15 ; 0,20 ; 0,25 ; 0,30 ; 0,40 ; 0,50.

Les abaques correspondantes suivent.

Remarque :

– L'utilisateur peut préférer une autre procédure de sélection car celle qui est proposée ici garde un nombre aléatoire d'individus dont l'espérance est Np. Elle présente donc l'inconvénient de risquer de ne garder aucun élément, il suffirait pour cela qu'aucune valeur de Z ne soit supérieure à λ.

Pour le cas où on désirerait garder un nombre fixé K d'éléments, on pourrait considérer les calculs faits ici comme approximatifs en remplaçant p par K/N et en gardant les K individus correspondant aux K plus grandes valeurs expérimentales de Z.

Les courbes ci-dessous n'ont pas été construites pour toutes les valeurs de N et p et dans la plupart des cas, il faudrait avoir recours à des interpolations doubles en N et p. En outre, l'utilisation d'abaques est toujours lourde et c'est pourquoi il va être proposé une formule approximative donnant P* en fonction de ρ de manière facile à calculer.

On peut approcher P* par une expression polynomiale ρ. L'existence d'un point d'inflexion nous force à choisir un polynôme dont le degré est au minimum 3 mais l'extrême régularité des abaques nous porte à essayer exactement le degré 3. Ce polynôme sera déterminé par deux points et les dérivées en ces points.

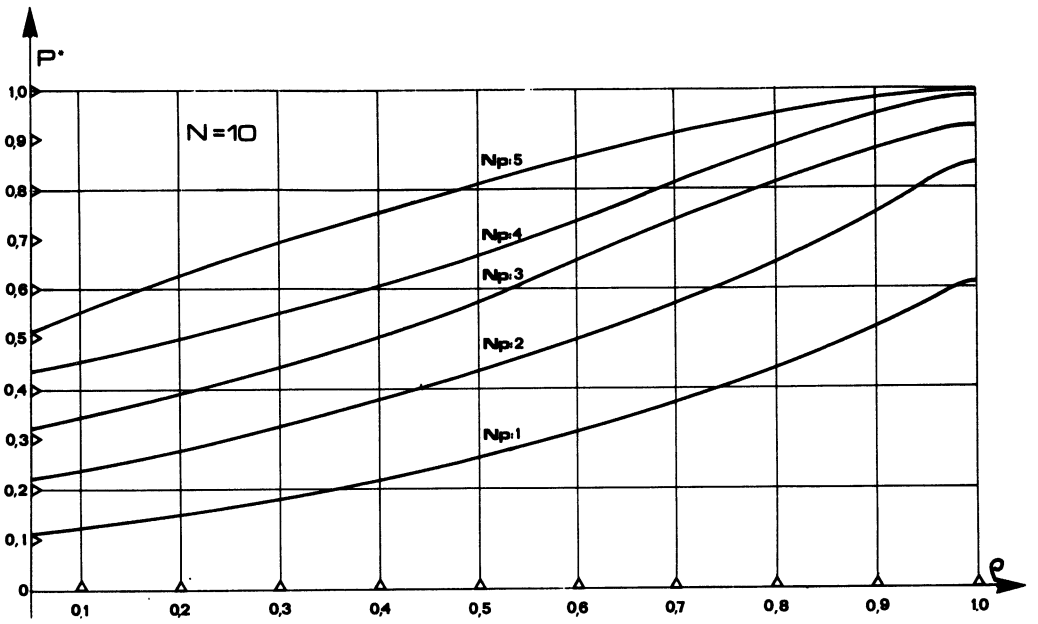
En effets, le calcul nous donne les résultats suivants :

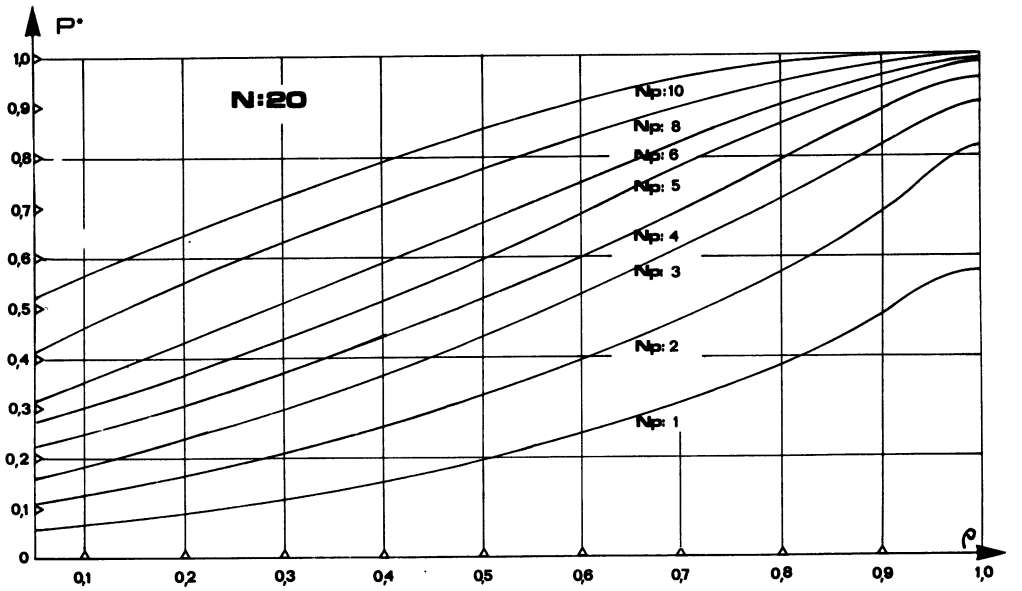
pour $\rho = 0 \; ; P^* = p$

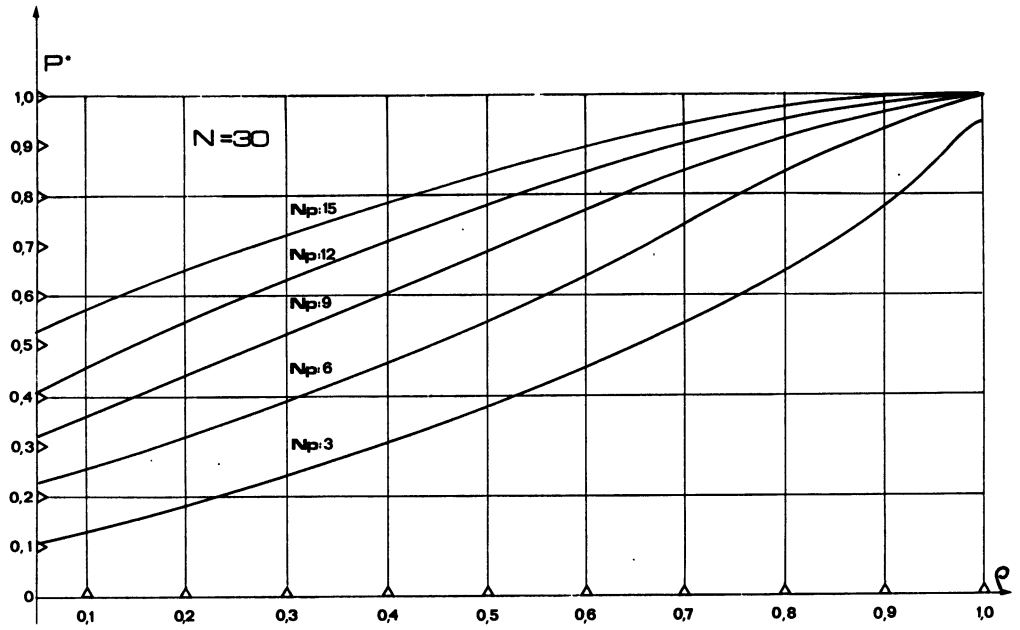
pour $\rho = 1 \; ; P^* = 1 - (1-p)^N$

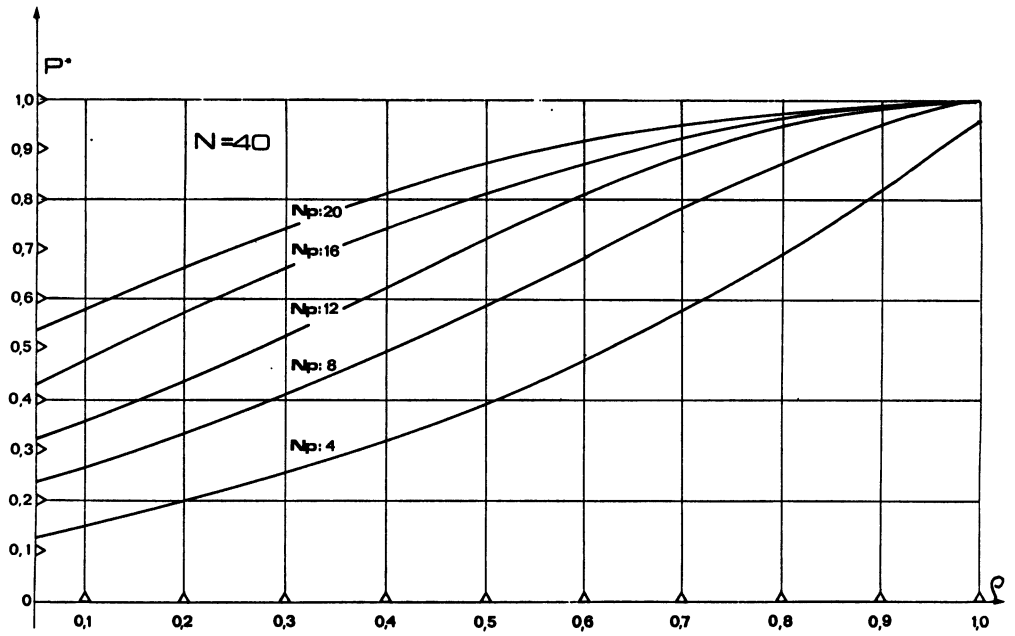
La dérivée de P* par rapport à ρ peut être calculée. Au point $\rho = 1$ elle prend la valeur 0 ; au point $\rho = 0$ elle prend la valeur :

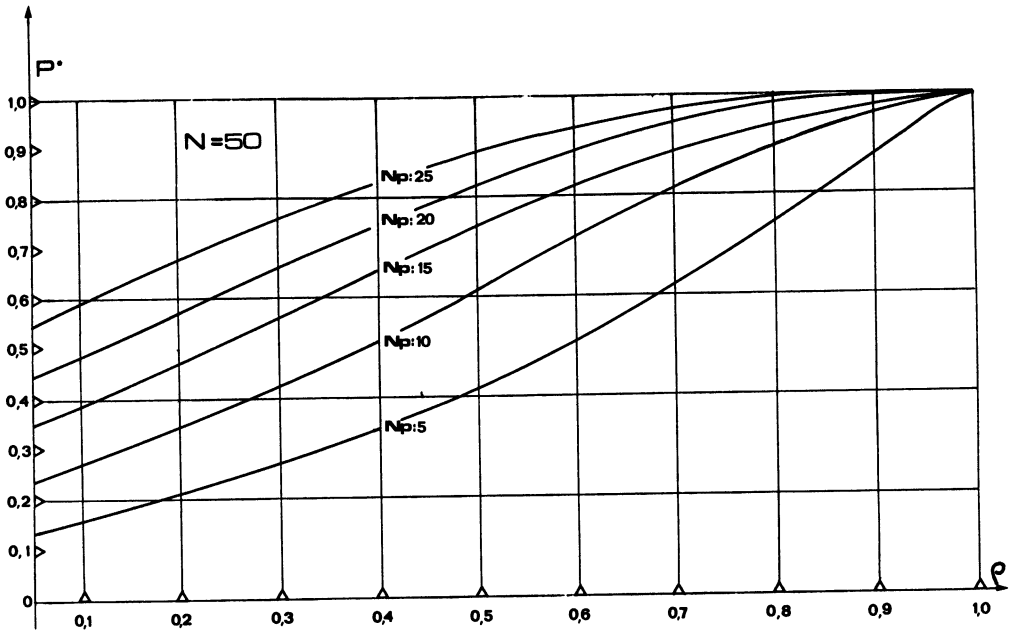
$$f_0 (N, p) = \frac{e^{-\frac{1}{2}t^2}}{2\pi} \int_{-\infty}^{+\infty} N \, F^{N-1}(s) \, e^{-\frac{1}{2}s^2} \, s \, ds$$

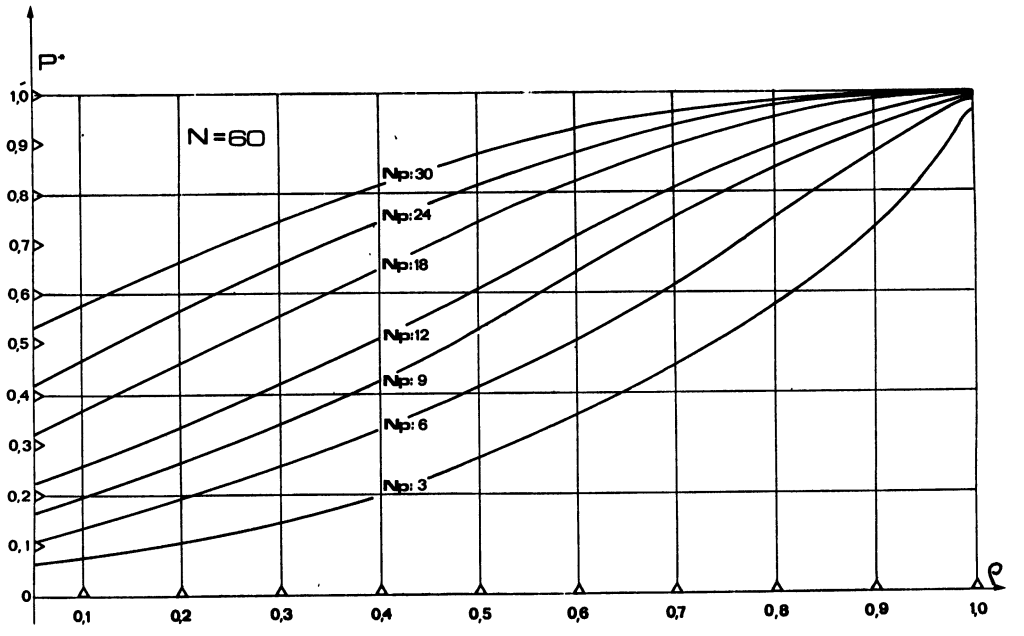

P*
N=10
Np:5
Np:4
Np:3
Np:2
Np:1
ρ
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1,0

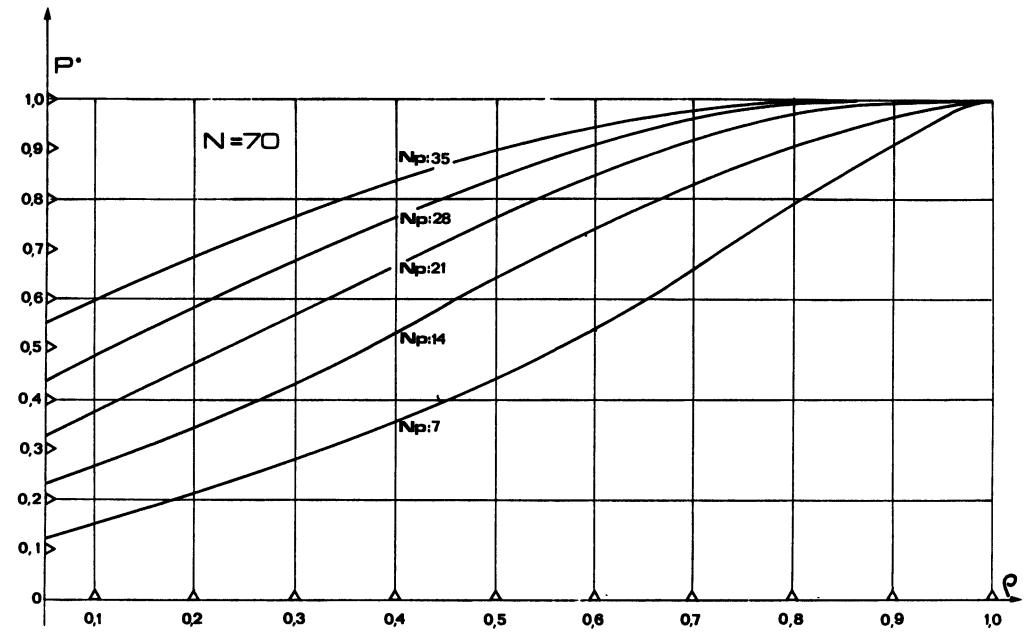

P*
N:20
Np:10
Np: 8
Np: 6
Np: 5
Np: 4
Np: 3
Np: 2
Np: 1
ρ
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1,0

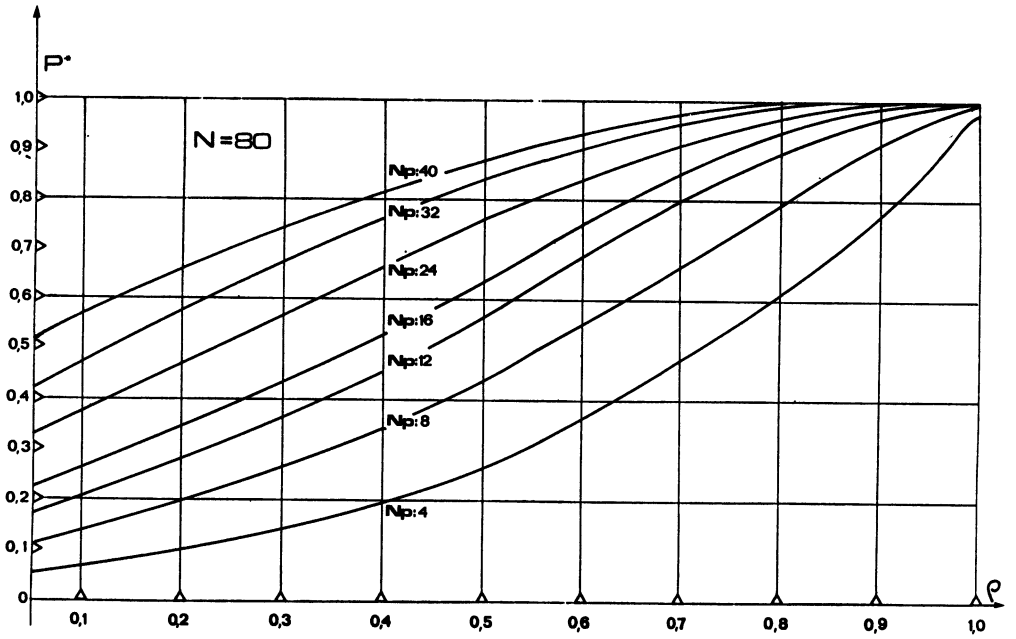

P*
N=30
Np:15
Np:12
Np:9
Np:6
Np:3
ρ
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1,0

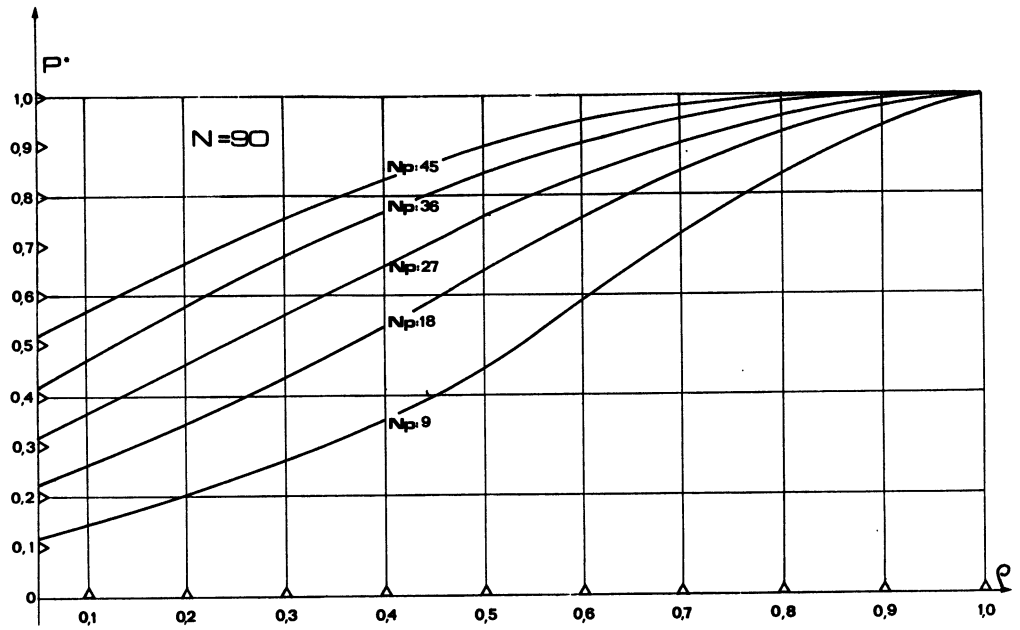

P*
N=40
Np:20
Np:16
Np:12
Np:8
Np:4
ρ
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1,0

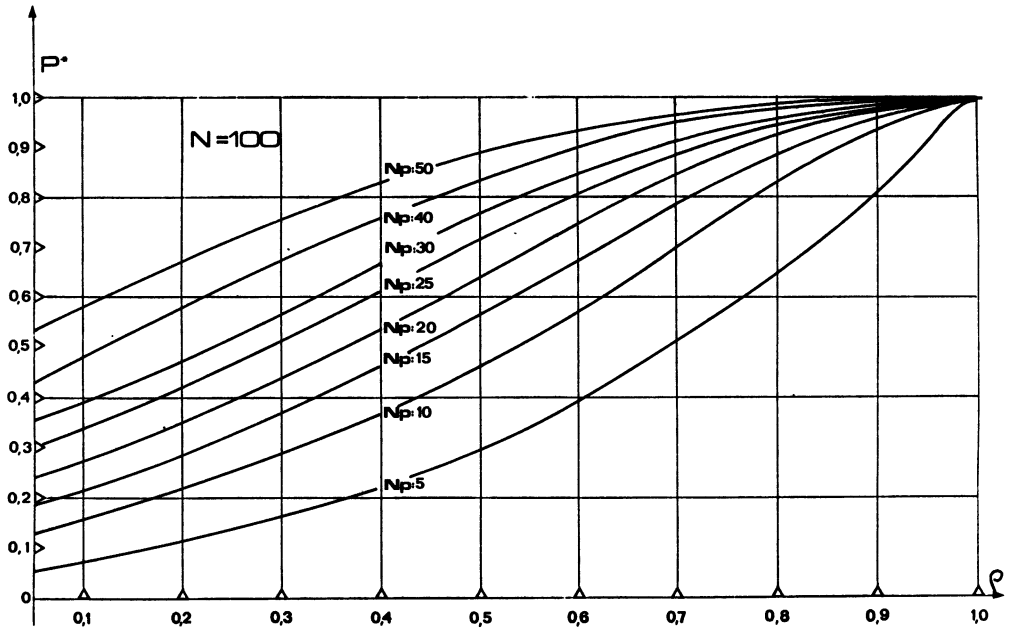

P*
N=50
Np:25
Np:20
Np:15
Np:10
Np:5
ρ
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1,0


P*
N=60
Np:30
Np:24
Np:18
Np:12
Np:9
Np:6
Np:3
ρ
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1,0


P*
N=70
Np:35
Np:28
Np:21
Np:14
Np:7
ρ


P*
N=80
Np:40
Np:32
Np:24
Np:16
Np:12
Np:8
Np:4
ρ


P*
N=90
Np:45
Np:36
Np:27
Np:18
Np:9
ρ
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1,0


P*
N=100
Np:50
Np:40
Np:30
Np:25
Np:20
Np:15
Np:10
Np:5
ρ
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1,0

Cette expression $f_0(p, N)$ n'est pas simple. Son calcul a été programmé pour des valeurs de p variant de 0,05 à 0,50 de 0,05 en 0,05 et pour les valeurs de N allant de 10 en 10 à partir de 10 jusqu'à 100. Sur les valeurs trouvées on ajuste un polynôme du second degré en p et N ce qui permet d'écrire :

$$f_0(N, p) \# - 0{,}087 + 0{,}0068\ N + 3{,}128\ p - 0{,}000047\ N^2 - 3{,}59\ p^2 + 0{,}0061\ pN$$

Il est facile alors de trouver le polynôme du troisième degré qui passe par les deux points :

$$(\rho = 0\ ,\ P^* = p) \quad \text{et} \quad (\rho = 1\ ,\ P^* = 1 - (1 - p)^N)$$

et tel que $\left(\frac{dP^*}{d\rho}\right)_{\rho=0} = f_0(N, p)$ et $\left(\frac{dP^*}{d\rho}\right)_{\rho=1} = 0$

c'est :

$$P^* = p + f_0(N, p)\ .\ \rho + [3(1 - p)(1 - (1 - p)^{N-1}) - 2f_0(N, p)]\ \rho^2 + [-2(1 - p)(1 - (1 - p)^{N-1}) + f_0(N, p)]\ \rho^3$$

Cette formule donne une approximation d'erreur inférieure à 0,03 pour les valeurs de p comprises entre 0,25 et 0,50 et pour les valeurs de N au moins égales à 20.

L'ajustement n'est pas très bon aux valeurs extrêmes, c'est-à-dire pour $p < 0{,}25$ et $N < 20$. L'erreur est inférieure à 0,05.

IV – APPLICATION A L'ANALYSE DE VARIANCE A UN FACTEUR DE VARIATION DONT ON VEUT SELECTIONNER CERTAINS NIVEAUX.

Etudions trois types de dispositifs expérimentaux : les lots, les blocs complets, les blocs incomplets équilibrés.

A) Le dispositif expérimental permet d'avoir, pour chacun des N niveaux, un lot de répétitions.

Dans une sélection d'une proportion p parmi N niveaux, exiger que la probabilité de garder le meilleur soit un nombre donné P* permet de déterminer le nombre M de répétitions par niveau.

Le modèle peut s'écrire :

$$Y_{ij} = \mu + (\theta_i - \theta_.) + E_{ij} \qquad \begin{matrix} i = 1, 2 \ldots N \\ j = 1, 2 \ldots M \end{matrix}$$

où $\theta_. = \sum_{i=1}^{N} \frac{\theta_i}{N}$ en supposant le dispositif équilibré.

Les aléatoires E_{ij} sont toutes indépendantes et de loi $N(0, \sigma_1^2)$. On fait l'hypothèse que θ_i est la valeur d'une variable aléatoire A_i de loi $N(0, \sigma^2)$ et

on suppose les aléatoires $A_1, A_2 \ldots A_N$ indépendantes. (On a donc mis sur le paramètre multidimensionnel $\theta' = (\theta_1, \theta_2 \ldots \theta_N)$ la distribution à priori produit des distributions $N(0, \sigma^2)$.). Proposons nous de ne garder en moyenne qu'une proportion p de niveaux parmi les N niveaux du facteur de variation.

Posons $$Y_{i.} = \sum_{j=1}^{M} \frac{Y_{ij}}{M} \quad \text{et} \quad Y_{..} = \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{Y_{ij}}{NM}$$

L'estimateur de θ_i est traditionnellement $Y_{i.} - Y_{..}$ ce qui conduit à essayer de prédire A_i grâce à une régression sur $Y_{i.} - Y_{..}$ afin d'appliquer la théorie précédente.

Le modèle permet d'écrire que, conditionnellement à

$$(A_1 = \theta_1, \ldots, A_i = \theta_i, \ldots A_N = \theta_N),$$

la loi de $Y_{i.} - Y_{..}$ s'obtient en écrivant :

$$Y_{i.} - Y_{..} = \theta_i - \theta_. + E_{i.} - E_{..}$$

où $$E_{i.} = \sum_{j=1}^{M} \frac{E_{ij}}{M} \quad \text{et} \quad E_{..} = \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{E_{ij}}{NM}$$

La loi conditionnelle de $E_{i.} - E_{..}$ ne dépend pas des θ_i et est normale, la loi à priori de $E_{i.} - E_{..}$ est donc indépendante des aléatoires A_i. Par suite on obtient la loi à priori de $Y_{i.} - Y_{..}$ grâce au modèle "à priori" :

$$Y_{i.} - Y_{..} = A_i - A_. + E_{i.} - E_{..}$$

où les aléatoires $E_{i.} - E_{..}$ sont indépendantes des aléatoires $A_{i'}$ quel que soit le couple (i, i').

On voit que, pour sa distribution à priori, $Y_{i.} - Y_{..}$ est une variable d'espérance nulle et par conséquent, pour le calcul de la régression, il suffit de connaître la covariance entre A_i et $Y_{i.} - Y_{..}$ et la variance de $Y_{i.} - Y_{..}$

Or : $$\text{var}(Y_{i.} - Y_{..}) = \frac{N-1}{N}\left(\sigma^2 + \frac{\sigma_1^2}{M}\right)$$

$$\text{cov}(A_i, Y_{i.} - Y_{..}) = \frac{N-1}{N} \cdot \sigma^2$$

On a donc pour estimer A_i l'aléatoire :

$$Z_i = \frac{\sigma^2}{\sigma^2 + \frac{\sigma_1^2}{M}} (Y_{i.} - Y..)$$

La variance de Z est : $\sigma'^2 = \frac{N-1}{N} \cdot \frac{\sigma^4}{\sigma^2 + \frac{\sigma_1^2}{M}}$

Le coefficient de corrélation ρ entre A et Z a pour carré :

$$\rho^2 = \frac{\sigma'^2}{\sigma^2} = \frac{N-1}{N} \cdot \frac{1}{1 + \frac{\sigma_1^2}{\sigma^2 \cdot M}}$$

Le calcul de la probabilité de garder le meilleur en fonction de ρ , p et N est applicable. Si on exige la probabilité P* de garder le meilleur on peut tirer, des abaques une information sur M.

En effet, lorsque les contraintes de l'expérimentateur fixent le nombre de niveaux et la proportion de gardés : $N = N_0$ et $p = p_0$, les abaques donnent le coefficient ρ_0 qui assure la probabilité P* de garder le meilleur.

De la formule ci-dessus, on tire alors M

$$M \# \frac{\rho_0^2 \, \sigma_1^2/\sigma^2}{1 - \rho_0^2 - \frac{1}{N}}$$

On prend évidemment pour M, la valeur entière la plus proche du rapport qui se trouve au second membre.

Pour simplifier l'utilisation de ce résultat, on peut chercher une fois de plus une approximation de la fonction qui donne ρ quand P*, N et p sont connus. Des considérations analogues à celles qui ont permis l'approximation de P* en fonction de ρ, N et p permettent de proposer, une valeur approximative de ρ. Grâce à la suite de calculs ci-dessous :

$$\beta = 1 - (1 - p)^N$$

$$g(N , p) = 4{,}461 - 0{,}0293.N - 12{,}48.p + 0{,}00015N^2 + 13{,}93p^2 + 0{,}0168p.N$$

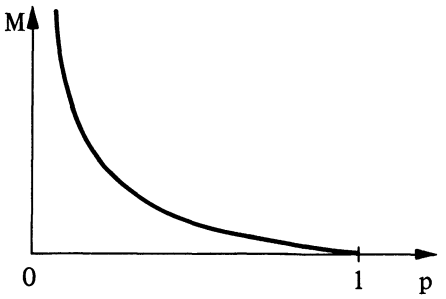
$$\alpha = \frac{2p\,(p - \beta)\, g(N , p) + 3p - 3\beta}{2\,(p - \beta)\, g(N , p) + 1}$$

Pour P* inférieur simultanément à α et β on a :

$$\rho = 1 - \frac{(P^* - \alpha)\sqrt{\beta - P^*}}{(p - \alpha)\sqrt{\beta - p}}$$

Ces calculs donnent ρ avec une erreur qui est au maximum de 0,03 lorsque p est compris entre 0,25 et 0,50 et N compris entre 20 et 100. La valeur de ρ étant ainsi trouvée permet de calculer M.

Si maintenant, la seule contrainte de l'expérimentateur est $N = N_0$ il a la liberté de choisir la proportion p de gardés ainsi que le nombre M, l'étude précédente nous donne l'ensemble des points (p , M) qui nous assurent la probabilité P* de garder le meilleur :



P* fixés
N_0

L'utilisateur choisira d'augmenter p pour diminuer M, ou d'augmenter M pour diminuer p, selon ses contraintes pratiques. Il pourra aussi, dans une optique décisionnelle, se donner une fonction de perte L, fonction du nombre de gardés p N et de la taille du dispositif expérimental N M. Il choisira alors le couple (p , M) qui minimise la fonction de perte sous la contrainte exprimée par la figure ci-dessus.

Remarque 1

Le cas où N ne serait pas fixé serait inintéressant car, le meilleur parmi N n'est évidemment pas équivalent au meilleur parmi N' (N' $\neq$ N).

Remarque 2

La difficulté essentielle de cette méthode est celle que l'on rencontre couramment dans les problèmes de puissance des tests. Elle réside dans le fait que σ^2 et σ_1^2 sont des variances dont, en général, on ignore les valeurs.

On peut, dans ce cas procéder en deux étapes :

– dans une première étape, se donner une valeur approximative de σ_1^2/σ^2 et de la proportion p d'individus que l'on désire garder ; en déduire M donc le dispositif expérimental.

– le deuxième stade intervient après l'expérience. M est alors connu mais on a une estimation $\hat{\sigma}_1^2/\hat{\sigma}^2$ de σ_1^2/σ^2 .

Si cette estimation est très différente de la valeur choisie à la première étape et si on tient à s'assurer la probabilité P* de garder le meilleur, on peut encore jouer sur la proportion p de gardés et sélectionner la proportion p_0 qui assurera la probabilité P* pour la valeur $\hat{\sigma}_1^2/\hat{\sigma}^2$ du rapport σ_1^2/σ^2

B) Le dispositif expérimental est celui des blocs complets :

Le modèle peut s'écrire sous la forme :

$$Y_{ij} = \mu + (\theta_i - \theta_.) + \beta_j + E'_{ij}$$

$$i = 1 , 2 \ldots N$$

$$j = 1 , 2 \ldots M$$

i est l'indice du niveau du traitement

j est l'indice du bloc

Les aléatoires E'_{ij} sont toutes indépendantes et de distribution $N(0, \sigma_1^2)$. Avec les mêmes hypothèses sur la distribution à priori du paramètre multidimensionnel $\theta' = (\theta_1, \theta_2 \ldots \theta_n)$ que dans le cas de l'analyse de variance avec répétitions on a, pour la loi à priori de $Y_{i.} - Y_{..}$ la relation :

$$Y_{i.} - Y_{..} = A_i - A_. + E'_{i.} - E'_{..}$$

avec indépendance des variables A et E.

Tous les calculs restent valables avec les mêmes notations. On détermine le nombre M de blocs nécessaires afin de s'assurer de la probabilité P* de garder le meilleur des niveaux en en sélectionnant une proportion p.

C) Le dispositif expérimental est celui des blocs incomplets équilibrés

On est amené à choisir un tel dispositif expérimental lorsque les blocs sont trop petits pour contenir les N niveaux du traitement que l'on veut étudier. Se posent alors les problèmes suivants :

- Combien mettre d'unités expérimentales par bloc :(k) ?
- Combien de fois implanter chacun des niveaux : (M) ?
- Combien de blocs nous faut-il ; (b) ?

(Ces problèmes n'étant évidemment pas indépendants puisque on a $N \cdot M = b \cdot k$).

Le critère de choix du dispositif expérimental est toujours celui qui consiste à s'assurer la probabilité P* de garder le meilleur dans une sélection d'une proportion p parmi les N niveaux. Il nous faut écrire le modèle, la loi à priori, chercher la régression et le coefficient ρ associé en fonction des paramètres du dispositif expérimental.

Le modèle est :

$$Y_{ij} = \mu + (\theta_i - \theta_.) + \beta_j + E_{ij} \qquad \theta_. = \sum_{i=1}^{N} \frac{\theta_i}{N}$$

Les aléatoires E_{ij} sont indépendantes et $N(0, \sigma_1^2)$

β_j est une constante inconnue.

$i = 1, 2 \ldots N$

$j = 1, 2 \ldots b$

$n_{ij} = \begin{cases} 1 \text{ si le } i^e \text{ niveau apparaît dans le bloc } j \\ 0 \text{ si i non dans j} \end{cases}$

On a :

$$\sum_{j=1}^{b} n_{ij} = M \; ; \; \sum_{i=1}^{N} n_{ij} = k \; ; \; \sum_{j=1}^{b} n_{ij}\, n_{i'j} = \lambda = \frac{M(k-1)}{N-1}$$

Rappelons que l'estimateur traditionnel de θ_i est défini comme suit :

On pose :

$$\left| \begin{array}{l} Y_{i0} = \sum_{j=1}^{b} n_{ij}\, Y_{ij} \\ \\ Y_{0j} = \sum_{i=1}^{N} n_{ij}\, Y_{ij} \end{array} \right.$$

et

$$\hat{\theta}_i = \frac{k}{(M\,k - M + \lambda)} \left[Y_{i0} - \frac{1}{k} \sum_{j=1}^{b} n_{ij}\, Y_{0j} \right]$$

Supposons maintenant que les N niveaux des traitements ont été choisis au hasard dans leur population et indépendamment les uns des autres et que l'on peut prendre pour hypothèse que le paramètre $\theta' = (\theta_1 \ldots \theta_n)$ est une valeur expérimentale du vecteur aléatoire $A' = (A_1 , A_2 \ldots A_N)$ dont la distribution est le produit des distributions marginales $N(0 , \sigma^2)$

Le meilleur des niveaux est celui qui correspond à l'aléatoire Sup A_i. Nous pouvons appliquer la théorie précédente si nous prédisons A_i grâce à une variable Z_i obtenue par régression. La forme de l'estimateur traditionnel de θ_i nous porte à prendre pour variable Z_i la régression de A_i sur :

$$\left[Y_{i0} - \frac{1}{k} \sum_{j=1}^{b} n_{ij}\, Y_{0j} \right]$$

Le modèle à priori s'écrit :

$$Y_{ij} = \mu + A_i - A_. + \beta_j + E_{ij} \qquad \text{où} \qquad A_. = \sum_i \frac{A_i}{N}$$

et on déduit :

$$Y_{i0} - \frac{1}{k} \sum_{j=1}^{b} n_{ij}\, Y_{0j} = A_i \left(\frac{k\,M - M + \lambda}{k} \right) - \frac{\lambda}{k} A_0 + \sum_{j=1}^{b} n_{ij}\, E_{ij}$$

$$- \frac{1}{k} \sum_{i'=1}^{N} \sum_{j=1}^{b} n_{ij}\, n_{i'j}\, E_{i'j}$$

où $$A_0 = \sum_i A_i = N\, A_.$$

On voit que, à priori, la variable $Y_{i0} - \frac{1}{k} \sum_{j=1}^{b} n_{ij}\, Y_{0j}$ est centrée. Pour calculer la régression il suffit d'étudier la covariance entre

$$A_i \quad \text{et} \quad Y_{i0} - \frac{1}{k} \sum_{j=1}^{b} n_{ij}\, Y_{0j} \quad \text{et la variance de} \quad Y_{i0} - \frac{1}{k} \sum_{j=1}^{b} n_{ij}\, Y_{0j}$$

Le calcul donne :

$$\text{Cov}\left(A_i\,,\,Y_{i0}-\frac{1}{k}\sum_{j=1}^{b} n_{ij}\,Y_{0j}\right)=\sigma^2\,.\,\frac{k\,M-M}{k}$$

et

$$\text{Var}\left(Y_{i0}-\frac{1}{k}\sum_{j=1}^{b} n_{ij}\,Y_{0j}\right)=\frac{M(k-1)}{k^2}\left[\frac{M(k-1)\,N}{N-1}\,\sigma^2+k\,\sigma_1^2\right]$$

On peut donc écrire la régression de A_i sur

$$Y_{i0}-\frac{1}{k}\sum_{j=1}^{b} n_{ij}\,Y_{0j}$$

sous la forme :

$$Z_i=\beta\left(Y_{i0}-\frac{1}{k}\sum_{j=1}^{b} n_{ij}\,Y_{0j}\right)$$

où

$$\beta=\frac{k(N-1)}{M(k-1)\,N+k(N-1)\,\dfrac{\sigma_1^2}{\sigma^2}}$$

Il faut maintenant calculer $\rho^2=\dfrac{\sigma'^2}{\sigma^2}$

On a

$$\sigma'^2=\beta^2\,.\,\text{Var}\left(Y_{i0}-\frac{1}{k}\sum_{j=1}^{b} n_{ij}\,Y_{0j}\right)$$

et on en déduit :

$$\boxed{\rho^2=\frac{M(k-1)(N-1)}{M(k-1)\,N+k\,(N-1)\,\dfrac{\sigma_1^2}{\sigma^2}}}$$

En supposant que N et p soient fixés, le sélectionneur désirant la probabilité P* de garder le meilleur va déterminer une valeur ρ_0^2 grâce aux abaques ou aux approximations déjà vues. Il est raisonnable de penser que lorsque l'on choisit un dispositif en blocs incomplets équilibrés on a le nombre k d'unités par bloc qui est fixé soit k_0. Le résultat ci-dessus permet de calculer le nombre de réalisations de chaque niveau.

$$\boxed{M=\frac{\rho_0^2\;\sigma_1^2/\sigma^2}{\dfrac{(k_0-1)}{k_0}\times\dfrac{N}{N-1}\times\left(1-\rho_0^2-\dfrac{1}{N}\right)}}$$

(En posant k = N on retrouve bien la formule des blocs complets). Le dispositif expérimental est alors parfaitement déterminé car le nombre de blocs est $b = \frac{N \cdot M}{k_0}$.

Les remarques faites au sujet de l'analyse de variance avec répétitions restent valables pour le cas de ce dispositif expérimental.

V – APPLICATION AU CAS MULTIVARIATE

Imaginons une population dans laquelle à chaque individu est attaché un vecteur G (une composante de G peut être une mesure de la valeur de l'individu relativement à un caractère donné. Le vecteur G représente alors la valeur de l'individu sur plusieurs caractères simultanés). On veut sélectionner les meilleurs. Il faut d'abord les définir c'est-à-dire prendre un critère unique qui nous donne une relation d'ordre sur la population. Une manière de faire consiste à prendre une moyenne pondérée des composantes de G : $A = b' G$ (où b est un vecteur de même dimension que G dont les composantes sont des paramètres donnés. b' signifie transposée de b).

A chaque individu on attache donc une valeur de A et on peut dire qu'un élément est meilleur qu'un autre s'il lui correspond une plus grande valeur de A.

Si on suppose que le vecteur G n'est pas directement observable mais qu'on peut avoir des valeurs d'un vecteur P avec les hypothèses suivantes :

Le vecteur $\left(\frac{G}{P}\right)$ est multinormal de matrice des espérances $\left(\frac{0}{0}\right)$ et de matrice des covariance connue $\Gamma = \left(\begin{array}{c|c} \Gamma_{gg} & \Gamma_{gp} \\ \hline \Gamma'_{gp} & \Gamma_{pp} \end{array}\right)$

(si la matrice des espérances n'est pas nulle, on la suppose connue et on peut faire une translation sur le vecteur P pour se ramener à ce cas).

On définit une variable Z pour prédire A par une technique de régression ; en appelant $\hat{G}$ la régression de G sur P on a :

(3)
$$Z = b' \hat{G} = b' \Gamma_{gp} \Gamma_{pp}^{-1} P$$

(On peut écrire ceci si Γ_{pp} est de plein rang ; si elle ne l'était pas il faudrait s'y ramener).

Le coefficient de corrélation ρ entre A et Z peut alors être calculé. Il est donné par la formule :

$$\rho^2 = \frac{b' \Gamma_{gp} \Gamma_{pp}^{-1} \Gamma'_{gp} b}{b' \Gamma_{gg} b}$$

Si maintenant, sur un échantillon de taille N issu de la population, on veut effectuer une sélection et ne garder qu'une proportion p d'individus et si on peut imaginer que les N éléments de l'échantillon ont été choisis par tirage au

sort indépendamment les uns des autres, on peut appliquer la théorie précédente. Le choix des éléments gardés se fait sur les valeurs prises par la variable Z. Le calcul de ρ ci-dessus permet de trouver la probabilité P* de garder le meilleur en fonction de ρ, N et p.

VI – GAIN ASSOCIE A UNE SELECTION – CONCLUSION

Par définition nous appelons gain l'espérance de A dans la population sélectionnée soit E(A/Après sélection) = $\mathcal{G}$.

Calculons ce gain.

On a A = Z + E avec Z et E qui sont des variables indépendantes. La sélection s'effectuant sur Z, l'espérance de E après sélection est l'espérance de E à priori, c'est-à-dire zéro.

Donc E(A/Après sélection) = E(Z/Après sélection).

Or la loi de Z après sélection a pour densité :

$$\begin{cases} \dfrac{1}{p}\dfrac{1}{\sqrt{2\pi}\,\sigma'} \exp\left\{-\dfrac{1}{2}\dfrac{z^2}{\sigma'^2}\right\} & \text{pour } \dfrac{z}{\sigma'} > t \\ 0 \quad \text{pour } \dfrac{z}{\sigma'} < t \quad \text{où} \quad t \text{ est tel que } F(t) = 1 - p \end{cases}$$

On en déduit l'espérance Z après sélection qui nous donne pour le gain la formule suivante :

$$\mathcal{G} = \frac{\sigma'}{p} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$$

Le gain ne dépend que de p et σ'.

Il est remarquable qu'il ne dépende pas de σ et même de N. Si on garde 2 individus parmi 10 le gain est le même que lorsqu'on en garde 20 parmi 100. Or la probabilité de garder le meilleur est beaucoup plus grande dans le second cas et en plus, le meilleur parmi 100 n'est pas équivalent au meilleur parmi 10.

Coefficient de corrélation	Probabilité de garder le meilleur avec p = 0,2	
	N = 10 ; Np = 2	N = 100 ; Np = 20
ρ = 0,05	0,2233	0,2319
ρ = 0,50	0,4564	0,6664
ρ = 0,85	0,7178	0,9772

La différence entre les deux critères gain et P* tient dans le fait que le gain est calculé en espérance. Un gain peut être obtenu en gardant le 2e et le 3e même si on élimine le meilleur. C'est un gain moyen.

On peut se demander alors quel est le bon critère lorsqu'on effectue une sélection.

Cette question appelle une réponse circonstanciée. En effet, s'il s'agit de sélectionner la composition d'un engrais parmi une infinité de compositions possibles, le gain paraît un critère adéquat. L'important est d'obtenir, en moyenne, un meilleur rendement.

Si, par contre, on se place en sélection animale sur des taureaux par exemple où, grâce à l'insémination artificielle, il suffit de garder 5 ou 6 taureaux reproducteurs, il est important de ne pas laisser échapper les meilleurs car ceux qui ne seront pas gardés ne se reproduiront pas. Leurs gènes seront perdus à jamais.

Le critère "du meilleur" prend alors tout son sens.

BIBLIOGRAPHIE :

Mahamunulu DESU et Milton SOBEL
"A fixed subset-size approach to the selection problem" *Revue Biometrika* (1968) 55 2 – p. 401.

GUPTA (1963 a) Table 1
"Probability integrals of the multivariate normal" *The annals of Mathematical Statistics* n° 34 p. 792.

GUPTA SS and SOBEL M 1957
"On a statistic which arises in selection and ranking problems" *The annals of Mathematical Statistics* n° 28 p. 957.

Shanti S GUPTA
"On some multiple décision (selection and ranking) rules" *Technometrics* vo. 7 n° 2 (mai 65)

Mahamunulu DESU
"A selection problem" *Annals of Mathematical Statistics* 1970 vol. 41 n° 5 p. 1596.

J.S. WILLIAMS
"The evaluation of a selection index" *Biometrics* sept. 62.