REVUE DE STATISTIQUE APPLIQUÉE

A. ZVENIGOROSKY
J. MILLIARD

Approximation de la distribution binomiale par la distribution normale ou par la distribution de Poisson

Revue de statistique appliquée, tome 31, n° 1 (1983), p. 63-73
<http://www.numdam.org/item?id=RSA_1983__31_1_63_0>

APPROXIMATION DE LA DISTRIBUTION BINOMIALE PAR LA DISTRIBUTION NORMALE OU PAR LA DISTRIBUTION DE POISSON

ZVENIGOROSKY A. (*) et MILLIARD J. (*)

Groupe d'Etudes et de Recherches en Sciences de l'Ingénieur (Caen) (**)

RESUME

Nous déterminons les domaines de validité de l'approximation d'une loi de Bernouilli par une loi de Poisson ou par une loi de Gauss, lorsque l'erreur relative tolérée est inférieure à 1 %. Les résultats sont présentés pour 3 valeurs moyennes de la variable aléatoire considérée 10, 100 et 1 000 et pour une large gamme de valeurs des probabilités de l'événement élémentaire considéré.

Nous déterminons également les domaines où les probabilités cumulées données par les lois de Bernouilli et de Poisson dans des intervalles de 1, 2 ou 3 écart-types autour de la valeur moyenne, peuvent être calculées au moyen de la loi de Gauss.

INTRODUCTION

Dans les calculs statistiques, la distribution binômiale ou distribution de Bernouilli joue un rôle important car elle s'adapte théoriquement à de nombreux cas rencontrés dans la pratique. Malheureusement, cette distribution est d'un maniement difficile dès que le nombre des épreuves croît beaucoup, les calculs devenant longs et fastidieux. Il est alors plus commode d'utiliser la distribution normale ou la distribution de Poisson comme distributions limites.

Nous nous proposons de préciser ici, dans quelques cas, les domaines dans lesquels ces distributions limites sont utilisables, en calculant l'erreur relative commise.

1. PRESENTATION DU PROBLEME ET NOTATIONS

1.1. Une variable aléatoire discontinue X suit une loi de Bernouilli si elle peut prendre une des n + 1 valeurs entières :

$$0, 1, 2, \ldots k \ldots n,$$

(*) Enseignants à l'I.U.T. de Caen, Département Mesures physiques.

(**) Ce groupe est rattaché à l'Institut des Sciences de la Matière et du Rayonnement de Caen (I.S.M.Ra).

Mots-Clés : **Distribution binômiale.**

avec une probabilité :

$$\text{Prob}(X = k) = B(k) = \binom{n}{k} p^k q^{n-k} \qquad (1)$$

où $0 < p < 1$ et $q = 1 - p$. La moyenne et l'écart-type de la variable aléatoire X sont :

$$\bar{k} = np \text{ et } \sigma = \sqrt{\bar{k}(1 - p)} \qquad (2)$$

La distribution de Bernouilli s'introduit tout naturellement dans un grand nombre de cas rencontrés dans la pratique et notamment dans deux domaines que nous prendrons comme exemples, la radioactivité et la fiabilité des dispositifs électroniques pendant leur durée de vie utile.

En radioactivité, la probabilité de désintégration d'un noyau intact à l'instant t, entre les instants t et t + dt est égale à λ dt, λ étant la constante radioactive du noyau considéré. En fiabilité, la probabilité de défaillance d'un dispositif intact à l'instant t, entre les instants t et t + dt est égale à λ dt si λ désigne alors le taux instantané de défaillance. La probabilité de désintégration d'un noyau et la probabilité de défaillance d'un disposif entre les instants 0 et t sont égales à :

$$p = 1 - e^{-\lambda t}$$

La probabilité d'obtenir k désintégrations ou défaillances à l'instant t à partir d'un échantillon de n éléments intacts à l'instant t = 0 est donnée par la loi de BERNOUILLI [1].

Les valeurs de n peuvent varier de quelques unités (en radioactivité et en fiabilité) à des nombres très élevés (dans un milligramme de radium il y a $2{,}7.10^{18}$ atomes). Quant aux valeurs de p, elles peuvent prendre toutes les valeurs possibles entre 0 et 1. En radioactivité les constantes radioactives varient entre $2{,}28\ 10^{6}\ s^{-1}$(*) pour le polonium 212 et $1{,}57\ 10^{-25}\ s^{-1}$ pour le plomb 204. En fiabilité le taux instantané de défaillance varie de $10^{-8}\ s^{-1}$ pour les composants grands publics à $10^{-13}\ s^{-1}$ pour les composants de très haute fiabilité utilisés dans les satellites et les répéteurs sous-marins.

Or la loi de BERNOUILLI devient d'un maniement difficile dès que n est grand car il faut alors calculer les factorielles n ! et (n – k) ! Les calculs sont encore plus longs et fastidieux lorsque l'on désire calculer des probabilités cumulées du type :

$$\text{Prob}(k_1 \leqslant X \leqslant k_2) = \sum_{i=k_1}^{k_2} \binom{n}{i} p^i q^{n-i} \qquad (3)$$

On trouve dans les ouvrages spécialisés des tables de probabilités et de probabilités cumulées pour la distribution de BERNOUILLI [1, 2, 3]. Ces tables ne concernent généralement que des valeurs de n limitées (à 1 000 dans les cas les plus favorables) et des valeurs de p en nombre réduit, typiquement de 0,1 en 0,1 de 0 à 1. On peut également utiliser la relation entre la loi binômiale et la loi Beta pour calculer les probabilités cumulées [4]. Cependant, sauf cas particulier, on ne dispose pas de ces tables dans la pratique.

(*) La probabilité étant telle que $p = \lambda$ dt (λ constante radioactive ou taux de défaillance dans les exemples cités) a nécessairement la dimension de l'inverse d'un temps. λ s'exprime donc en s^{-1}.

1.2. Fort heureusement, on peut souvent utiliser comme approximation de la loi binômiale, soit la distribution normale :

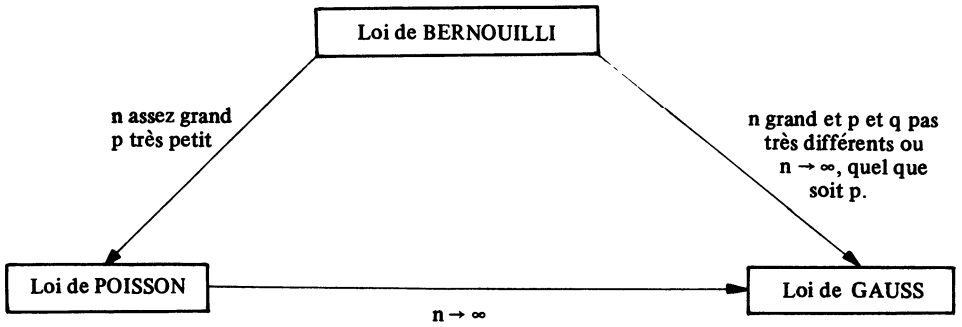
$$\text{Prob}\,(X = k) = G(k) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(k-\bar{k})^2}{2\sigma^2}\right) \qquad (4)$$

pour laquelle on peut se référer à la loi normale réduite qui, elle, est d'un usage très répandu, soit à la distribution de Poisson :

$$\text{Prob}\,(X = k) = \pi(k) = \frac{(\bar{k})^k\, e^{-\bar{k}}}{k\,!} \qquad (5)$$

pour laquelle les calculs sont plus aisés.

Grossièrement définis, les domaines de validité de ces approximations sont les suivants :



On trouve dans certains articles [5, 6, 7, 8] et dans les ouvrages consacrés aux variables aléatoires [9, 10, 11] des indications sur les domaines de validité de ces approximations. Malheureusement, ces indications sont souvent parcellaires et difficiles à relier pour couvrir le champ des besoins.

1.3. Il est pourtant fort utile de connaître avec précision les domaines de validité de ces approximations pour justifier l'emploi de G(k) à la place de B(k) notamment quand on désire connaître la probabilité d'obtenir une valeur X comprise entre deux limites (détermination des barres d'erreurs). On emploie alors les résultats bien connus pour la loi normale :

$$P_1^G = \text{Prob}\,(\bar{k} - \sigma < k < \bar{k} + \sigma) = 68{,}26\,\%$$

$$P_2^G = \text{Prob}\,(\bar{k} - 2\sigma < k < \bar{k} + 2\sigma) = 95{,}46\,\% \qquad (6)$$

$$P_3^G = \text{Prob}\,(\bar{k} - 3\sigma < k < \bar{k} + 3\sigma) = 99{,}74\,\%$$

1.4. Justifier l'approximation d'une loi de BERNOUILLI par une loi de Poisson est également nécessaire lorsqu'on désire employer le modèle théorique dénommé processus de Poisson pour représenter les phénomènes aléatoires qui sont décrits en toute rigueur par une loi de BERNOUILLI [12, 13].

2. COMPARAISON DES PROBABILITES DONNEES PAR LES LOIS DE BERNOUILLI, GAUSS ET POISSON

Nous procédons de la manière suivante :

a) Le nombre moyen d'événements $\bar{k}$ est tout d'abord fixé. Nous avons effectué les calculs pour 3 valeurs de $\bar{k}$: 10, 100 et 1000.

b) Pour chaque valeur de $\bar{k}$ nous choisissons ensuite une série de valeurs de n :

$\bar{k}$	n
10	11 à 10^8
100	110 à 10^8
1000	1100 à 10^9

n et $\bar{k}$ étant fixées nous calculons alors les paramètres p et σ en utilisant les relations (2). Pour les valeurs de $\bar{k}$ choisies la probabilité est comprise entre :

$$10^{-7} \quad \text{et} \quad 0{,}9091$$

c) Nous calculons ensuite les probabilités données par la loi de Gauss (G) pour les valeurs de k suivantes :

$$\bar{k} - 3\sigma, \bar{k} - 2\sigma, \bar{k} - \sigma, \bar{k}, \bar{k} + \sigma, \bar{k} + 2\sigma, \bar{k} + 3\sigma .$$

et les probabilités données par les lois de POISSON (π) et BERNOUILLI (B) pour les valeurs entières k' les plus voisines des valeurs de k précédentes.

d) Enfin, nous calculons les écarts relatifs $(G - B)/B$ et $(\pi - B)/B$ en pourcent.

A partir des résultats de ces calculs, nous avons tracé les courbes donnant ces écarts en fonction de p pour k = 10, 100 et 1000 en nous limitant chaque fois à $k = \bar{k}$ (courbes 1, 2, 3, 4, 5, 6).

Les tableaux 1 et 2 présentent les domaines de probabilité à l'intérieur desquels les écarts relatifs restent constamment inférieurs à 1 %. En dehors de ces domaines, les écarts relatifs fluctuent avec des amplitudes supérieures à ± 1 %.

N.B. : Les calculs ont été faits sur un micro-ordinateur Z 89 de Zenith Data System. Le programme rédigé en BASIC a été écrit en utilisant au mieux les ressources de la double précision. Les résultats ont ensuite été tronqués pour ne garder que sept chiffres significatifs au maximum.

3. PROBABILITE POUR QUE LA VARIABLE SOIT COMPRISE ENTRE CERTAINES LIMITES

Pour chaque couple de valeurs $\bar{k}$ et n nous calculons les probabilités

$$P_i = \text{Prob}\ (\bar{k} - i\sigma \leqslant k \leqslant \bar{k} + i\sigma) \quad (i = 1, 2, 3)$$

pour les lois de BERNOUILLI P_i^B et de POISSON P_i^π en arrondissant les valeurs limites aux entiers k' les plus voisins. Nous comparons ces probabilités cumulées à celles qui sont données par la loi de GAUSS (6) en calculant les écarts en pourcent :

$$\frac{P_i^B - P_i^G}{P_i^G} \quad \text{et} \quad \frac{P_i^\pi - P_i^G}{P_i^G}$$

Les valeurs de $\bar{k}$ et n choisies sont les suivantes :

$\bar{k}$	n
10	11 à 10^7
100	111 à 10^8
1000	1111 à 10^8

ce qui conduit à des probabilités comprises entre : 10^{-6} et 0,9001.

Le tableau 3 résume les résultats obtenus en présentant les domaines des probabilités à l'intérieur desquels les écarts relatifs restent constamment inférieurs à ± 1 %.

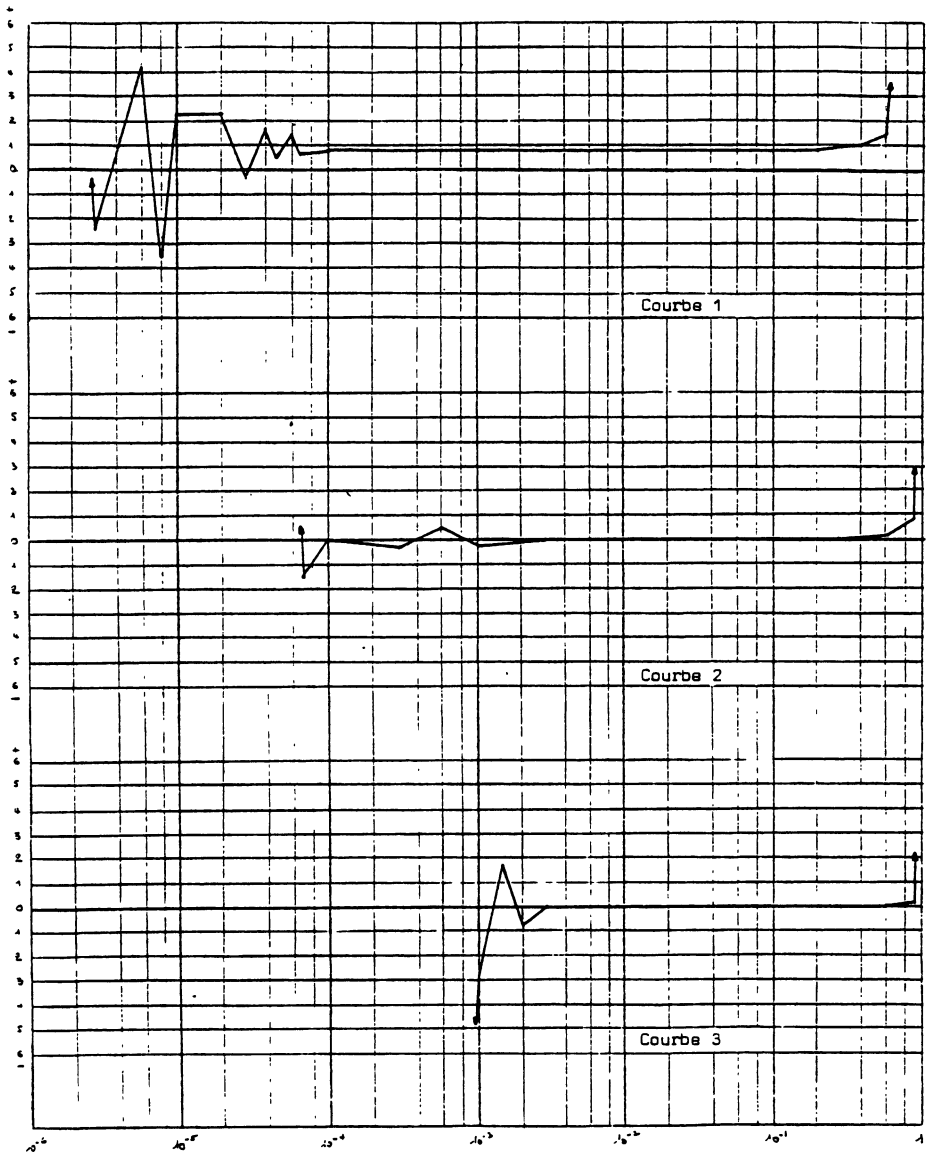


Figure 1. – Courbes de $\frac{G - B}{B}$ en % pour 1) $\bar{k} = 10$ 2) $\bar{k} = 100$ et pour $k = \bar{k}$ 3) $\bar{k} = 1000$

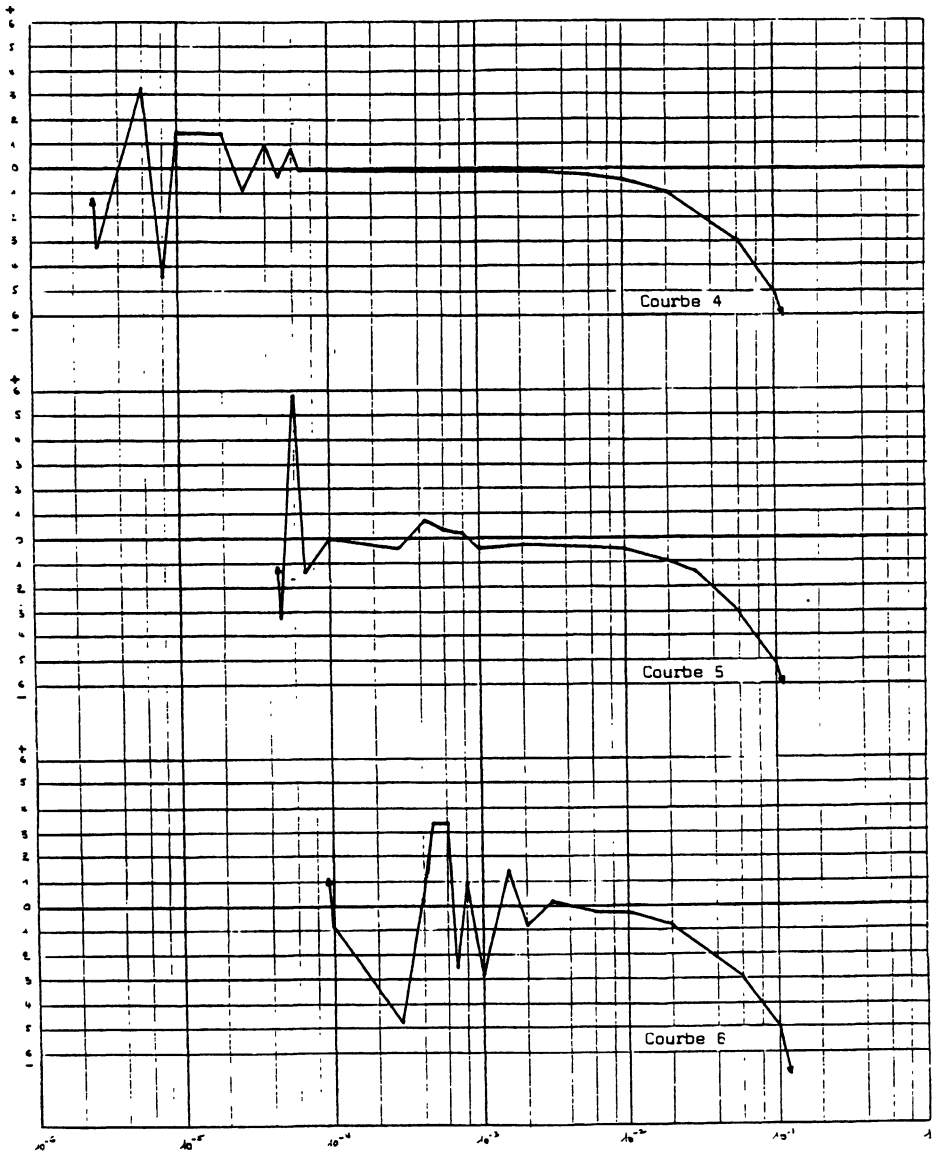


Figure 2. – Courbes de $\frac{\pi - B}{B}$ en % pour 4) $\bar{k} = 10$ 5) $\bar{k} = 100$ et pour $k = \bar{k}$ 6) $\bar{k} = 1000$

TABLEAU 1

Domaines à l'intérieur desquels l'écart relatif $\frac{G - B}{B}$ %

reste constamment inférieur à 1 % pour différentes valeurs de k

	$\bar{k} = 10$				$\bar{k} = 100$				$\bar{k} = 1000$			
p \ $k = \bar{k} \pm i\sigma$ i	0	$\pm\sigma$	$\pm 2\sigma$	$\pm 3\sigma$	0	$\pm\sigma$	$\pm 2\sigma$	$\pm 3\sigma$	0	$\pm\sigma$	$\pm 2\sigma$	$\pm 3\sigma$
0,9	↓	> 1%	> 1%	> 1%	↓	> 1%	> 1%	> 1%	↓	> 1%	> 1%	> 1%
0,6												
0,3												
0,1												
0,06												
0,03												
0,01												
0,006												
0,003												
0,001												
0,0006												
0,0003												
0,0001												
0,00006												
0,00003												
0,00001												
0,000006												
0,000003												
0,000001												

TABLEAU 2

Domaines à l'intérieur desquels l'écart relatif $\frac{\pi - B}{B}$ %
reste constamment inférieur à 1 % pour différentes valeurs de k

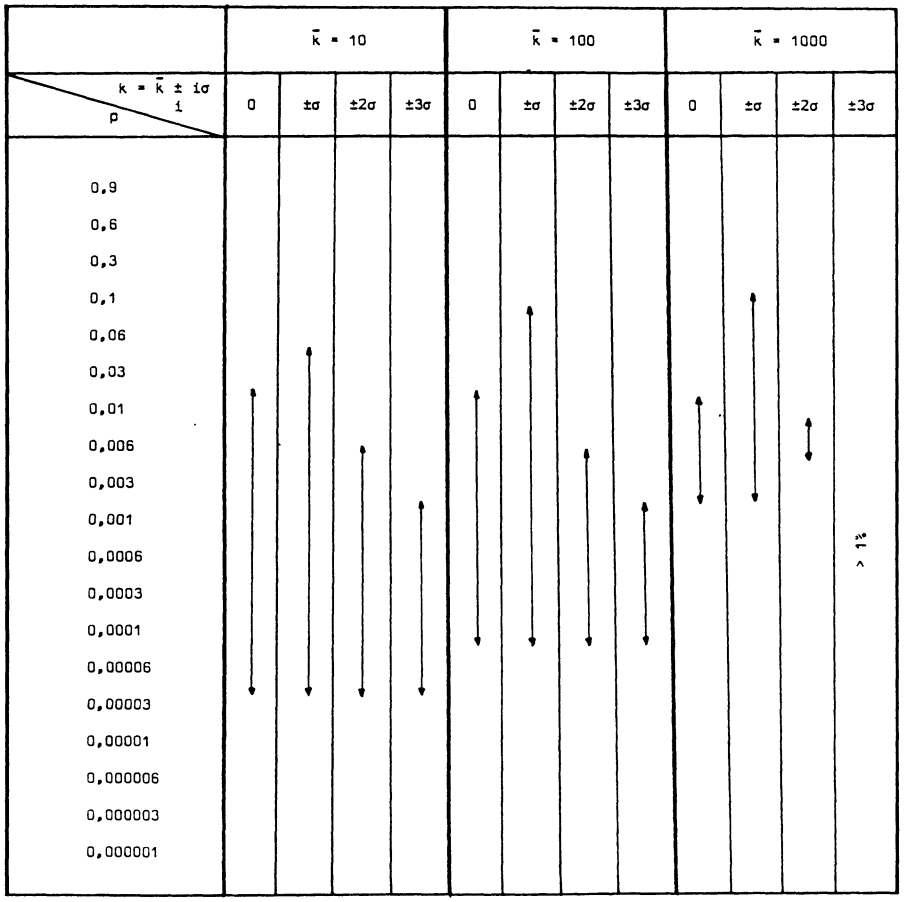
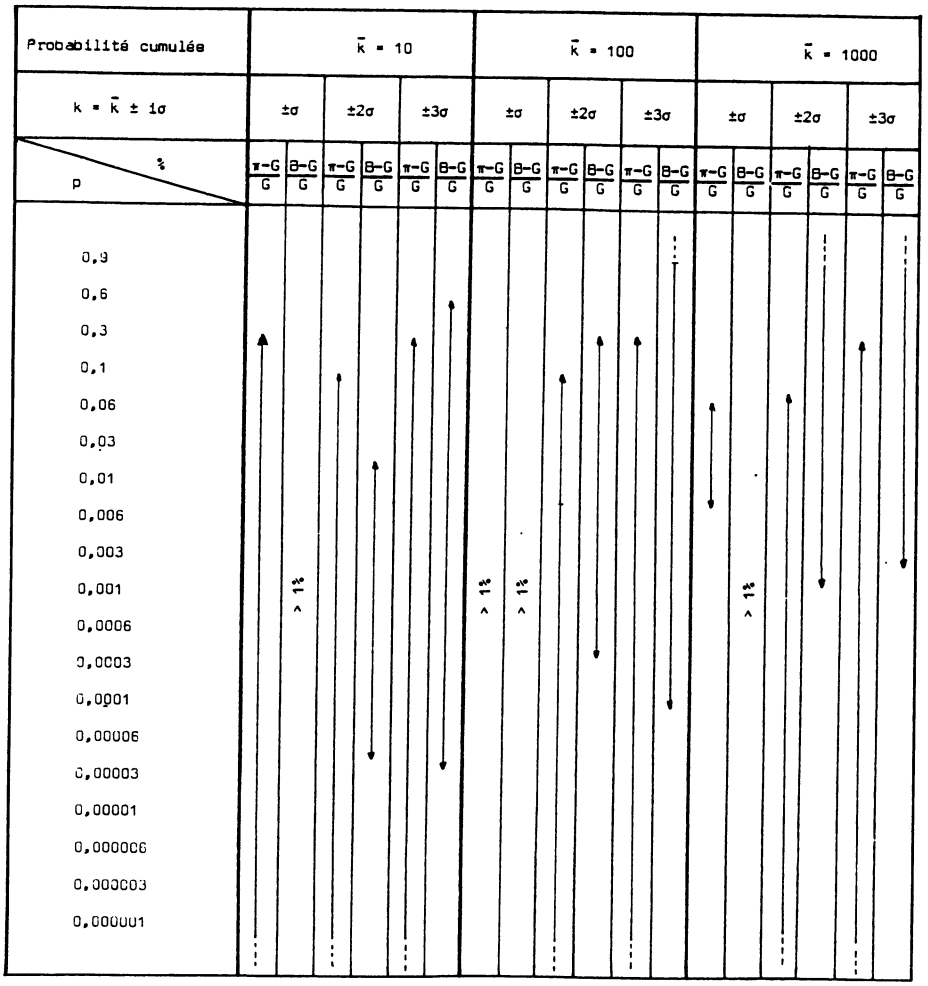


TABLEAU 3

Domaines à l'intérieur desquels les écarts relatifs $\frac{P_i^B - P_i^G}{P_i^G}$ % et $\frac{P_i^\pi - P_i^G}{P_i^G}$ % restent constamment inférieurs à 1 % pour différentes valeurs de k



CONCLUSION

Dans tous les résultats présentés ici, la limite supérieure de l'erreur admise a été fixée à 1 %.

Pour calculer la probabilité d'obtenir un nombre d'événements donné, la loi de BERNOUILLI peut être remplacée par une loi de GAUSS, seulement dans le cas où ce nombre est très voisin de la valeur moyenne (tableau 1). Par contre, la loi de POISSON peut remplacer la loi de BERNOUILLI pour un nombre d'événements différents de 1, 2 ou 3 écart-types de la valeur moyenne, avec cependant des gammes de valeurs des probabilités élémentaires qui vont en se restreignant. Il convient ici de souligner que l'approximation n'est pas obligatoirement valable pour les événements très rares (tableau 2).

Pour calculer les probabilités cumulées, l'approximation des lois de BERNOUILLI et de Poisson par une loi de GAUSS est justifiée, lorsque l'intervalle est de 3 écart-types au-dessus et au-dessous de la valeur moyenne du nombre d'événements. Par contre, dans des intervalles plus réduits, il faut être plus prudent (tableau 3).

BIBLIOGRAPHIE

[1] *The advanced theory of statistics.* Vol. I Distribution theory Ch. Griffin et Cie, London 1963. – Listes de tables de la distribution binômiale, p. 124.

[2] *Handbook of mathematical functions.* Dover publications. – Listes de tables de la distribution binômiale, p. 963.

[3] Tables statistiques. *RFSA*, n° spécial 1973.

[4] N.L. JOHNSON, S. KOTZ. – *Discrete distributions.* Vol. 2 Hougton Miffin Company, Boston 1970.

[5] W. FELLER. – On the normal approximation to the binomial distribution. *Ann. Math. Stat.* 1945, Vol. 16, p. 319.

[6] M.S. RAFF. – On approximating the binomial point. *J. Am. Statist. Ass.*, 51 (293) 1956.

[7] J.H.C. LISMAN. – Comparaison entre la distribution binômiale symétrique et la distribution "normale discontinue". *RFSA.* 1972, vol. 20, n° 3, p. 85.

[8] B. SORIN. – Sur l'approximation gaussienne des queues de distribution binomiale. *RFSA.* 1972, vol. 20, n° 3, p. 89.

[9] P.L. MEYER. – *Introductory probability and statistical applications.* Addison – Wesley P.C. 1970.

[10] S. BERMAN et R. BEZARD. – *Statistique et probabilité.* Editions Chiron, Paris 1973.

[11] G. PARREINS. – *Techniques statistiques.* Dunod Techniques, 1974.

[12] L. JANOSSY. – *Theory and practice of the evaluation of measurements.* Oxford University Press. London, 1965.

[13] A. ANGOT. – Compléments de mathématiques. *Editions de la Revue d'Optique.* Paris, 1957.

Nota – Les auteurs peuvent fournir à tout lecteur intéressé des photocopies des tableaux numériques correspondant aux calculs du § 2 de cet article.