Exemplaire corrigé par l'auteur

J. exp. mar. Biol. Ecol., 1972, Vol. 9, pp. 121–132; © North-Holland Publishing Company

CALCUL DE L'ERREUR SUR UN COMPTAGE DE ZOOPLANCTON

SERGE FRONTIER

Centre ORSTOM de Nosy Be, Madagascar

Abstract: The counts from a wide variety of plankton collections show a stochastic relation between mean and variance. The ~~cubes~~ of the values (except for very small numbers which are Poisson in their distribution) are Gaussian in their distribution with a stable variance and therefore the error at say the 95 % probability level may be determined. The error for a given value, N, is a function of N and the relative error decreases rapidly as N increases, then very slowly so that it is not necessary to count more than 100 individuals in a given aliquot of the sample. At this limit the error (at the 95 % level) of the total sample is about 31 % of that of the fraction counted.

Résumé: Des comptages répétés intéressant des groupes planctoniques très divers font apparaître une liaison stochastique entre moyennes et variances. Les racines cubiques des nombres observés (à l'exception des plus petits, qui sont distribués suivant la loi de Poisson) ont une distribution gaussienne, avec une variance stable dont on déduit l'erreur statistique au niveau de confiance 95 %. L'erreur sur un effectif N est fonction de N; l'erreur relative diminue d'abord vite quand N augmente, puis très lentement, de sorte qu'il n'est pas rentable de compter effectivement plus d'une centaine d'individus, ce comptage s'appliquant à une fraction aliquote de l'échantillon. En ramenant ce nombre à l'échantillon total on fait une erreur relative (au niveau 95 %) au plus égale à 31 % quelle que soit la fraction objet du dénombrement.

ETUDE EMPIRIQUE DE LA DISTRIBUTION DES ERREURS

MÉTHODE DE COMPTAGE

Un échantillon de plancton fixé, d'un volume sédimenté compris entre 10 et 300 cm^3, est mis en suspension dans un volume de liquide filtreur 10 fois supérieur, homogénéisé à l'aide d'une poire de caoutchouc remplie et vidée une dizaine de fois, et sous-échantillonné par un prélèvement à l'aide de cette poire. Une poire de caoutchouc moyennement rigide (fournisseur: Thomas, Philadelphia, U.S.A.) manipulée toujours de la même façon permet de réaliser des prélèvements avec une erreur inférieure à 2 % sur le volume. Le contenu de cette dernière, représentant une fraction volumétrique de la suspension connue et comprise entre 1/10 et 1/40, est déposé dans une cuvette de Dolfuss: ~~C'est~~ cuve de verre rectangulaire de 5 cm sur 10 dont le fond est partagé en 200 carrés de 5 mm de côté par un quadrillage en relief (fournisseur: Leune, Paris). Les organismes planctoniques sont alors dénombrés sous stéréomicroscope, soit dans la cuvette entière soit dans une partie aliquote facile à déterminer grâce au quadrillage. La fraction de récolte examinée varie alors entre 1/10 et 1/400 (le plus fréquemment 1/20): on peut alors considérer que les effectifs de points matériels dispersés au hasard dans la phase liquide se distribueraient suivant les lois de Poisson. En fait, les particules organiques dénombrées ont des propriétés physiques

qui impliquent des interactions: écartement mutuel de particules d'une certaine masse dès qu'elles sont nombreuses dans la phase; tendance des particules à s'agglomérer entre elles par des phénomènes de surface. Ces deux types d'interaction, qui tendent respectivement à sous-disperser et à sur-disperser les distributions, coexistent avec des modalités qui varient suivant les caractères physiques de l'échantillon en suspension et le type d'organisme considéré. A priori, une étude empirique semble s'imposer pour chaque type d'organisme et chaque aspect physique de la suspension.

N'existe-t-il pas, cependant, une allure moyenne de la distribution des écarts que l'on pourrait appliquer à l'ensemble des catégories planctoniques pour en déduire un ordre de grandeur, valable dans la majorité des cas, de l'erreur statistique? Pour y répondre, des comptages répétés ont été effectués sur des organismes très divers dans une série de 28 récoltes.

Les organismes dénombrés appartiennent aux 32 catégories suivantes (allant de l'ordre à l'unité sub-spécifique):

Solmundella bitentaculata	*Acartia amboinensis*
Siphonophores calycophores	Copépodes
Cténaires cydippoïdes	Cumacés
Sagitta enflata	Tanaïdacés
Sagitta, autres espèces	Amphipodes gammariens
Atlanta gaudichaudi véligères	*Hyperia* spp.
A. gaudichaudi jeunes et adultes	Stomatopodes larves
Creseis acicula véligères	Anomoures larves
C. acicula jeunes et adultes	Brachyoures larves
C. chierchiae véligères	*Lucifer* spp. protozoés
C. chierchiae jeunes et adultes	*Lucifer* sp. mysis
Cavolinia longirostris jeunes	*Lucifer* spp. mastigopus
Clionina longicaudata	*Lucifer* spp. adultes
Penilia avirostris	Salpes
Evadne tergestina	Dolioles
Ostracodes	Appendiculaires

Chaque récolte a fait l'objet de trois sous-échantillonnages successifs dans lesquels ont été dénombrés les 32 taxa. Le sous-échantillon est replacé après comptage dans la récolte.

Chaque ensemble de trois comptages permet de calculer une moyenne et une variance, celle-ci avec deux degrés de liberté. Nous rechercherons dans une première étape une liaison stochastique entre moyennes et variances en considérant les 1773 résultats de comptage dans leur ensemble. Afin d'alléger les calculs, nous grouperons les comptages triplés ayant donné une même moyenne, ou des moyennes voisines, de façon à obtenir des ensembles d'au moins 45 effectifs dénombrés (30 degrés de liberté). Nous obtenons ainsi 29 points (m, s^2).

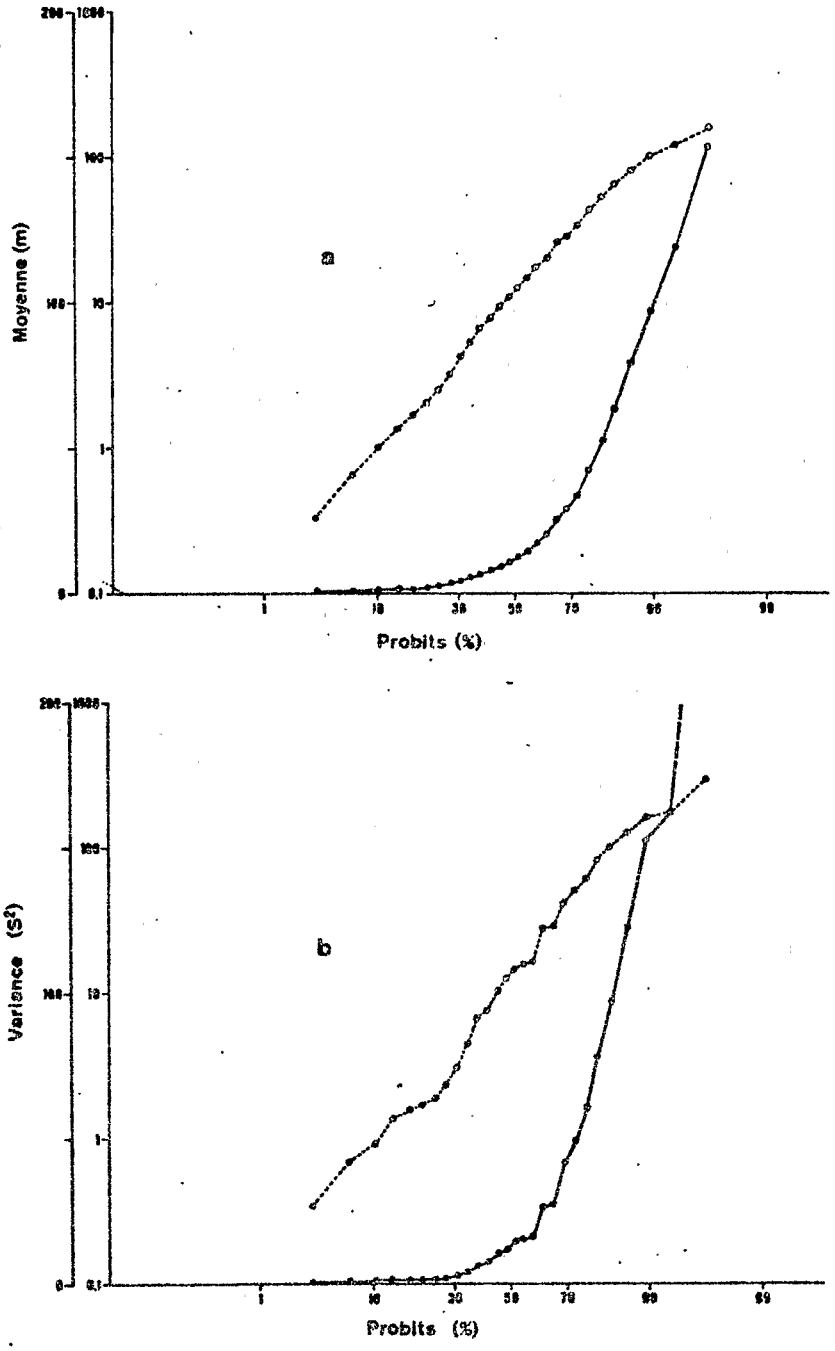

Fig. 1. Normalisation approximative des distributions des moyennes (a) et des variances (b) des effectifs non transformés: test Probit. Trait plein: distributions des paramètres; trait interrompu: distributions de leurs logarithmes.

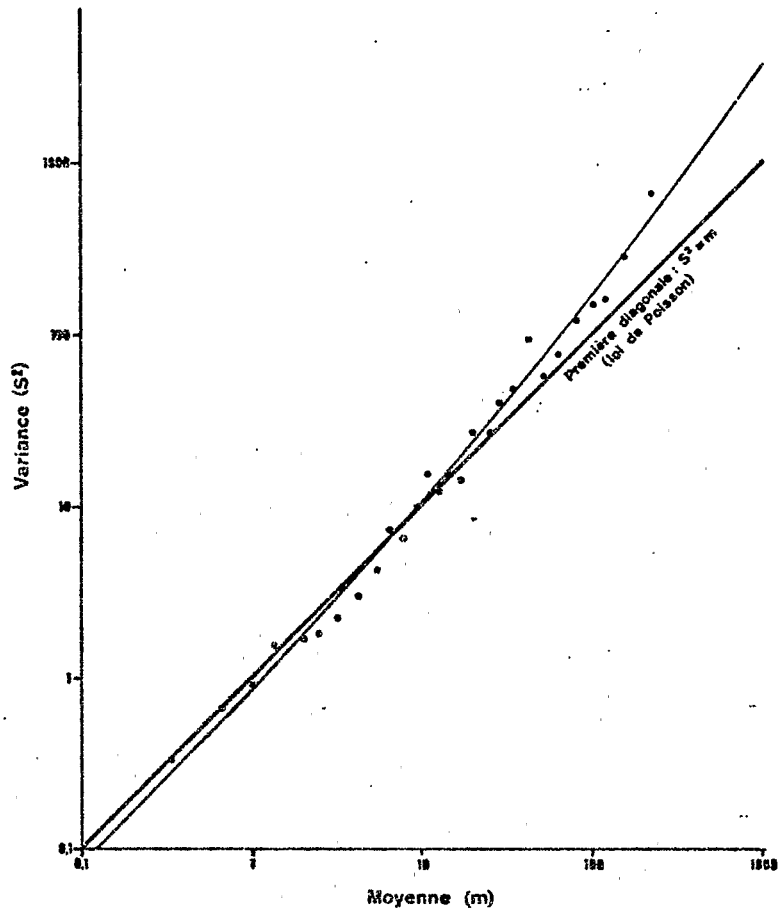

Fig. 2. Relation entre moyennes et variances dans l'ensemble des effectifs non transformés (échelle log-log). Trait épais: première diagonale; trait fin: arc de parabole (voir texte).

Un artefact apparaît du fait qu'il a été, arbitrairement, compté beaucoup plus de petits effectifs que de grands: il s'ensuit des distributions très dissymétriques de l'ensemble des 29 moyennes et de celui des 29 variances. On considère alors les logarithmes de ces moyennes et variances, ce qui normalise approximativement les deux distributions (test probit: Fig. 1). On constate que les 19 points (m, s^2) se rangent régulièrement, sur un graphique log-log, le long d'une ligne légèrement concave vers les s^2 supérieurs (Fig. 2) qui peut être approchée par l'arc de parabole d'équation:

$$y = 0{,}0598x^2 + 1{,}0391x - 0{,}0757$$

(avec $x = \log_{10} m$, $y = \log_{10} s^2$). Le sommet de la parabole a pour coordonnées $(-8{,}6881, -4{,}5896)$. Si l'on pose $X = (x+8{,}6881)^2$ on obtient une corrélation linéaire entre X et y, avec un coefficient de corrélation égal à 0,9946.

La loi reliant moyenne et variance dans des comptages réitérés semble donc bien établie. Cependant, sa forme analytique dérivée de l'équation précédente est difficilement utilisable. Nous simplifierons la représentation en décomposant l'arc de parabole trouvé en deux segments de droite contigus. On remarque que les points correspondants à des moyennes inférieures à 10 se placent au voisinage de la première diagonale (distribution de Poisson). Pour $m > 10$ on observe un décollement du nuage de points d'abord au-dessous de la diagonale (légère sous-dispersion) puis au-dessus (surdispersion) – faits qui peuvent s'interpréter mécaniquement.

La distribution de Poisson, à condition que la moyenne ne soit pas trop petite, est normalisée par la transformation $\sqrt{(N+\frac{3}{8})}$, la nouvelle variable ayant une variance stable et voisine de 0,25 (Anscombe 1948).

Pour $m > 10$ on ajustera le nuage de points à un segment de droite donné par la régression de y en x;

$$y = 1{,}2265x - 0{,}2165$$

(coefficient de corrélation: 0,9793). En revenant aux variables initiales:

$$\log s^2 = 1{,}2265 \log m - 0{,}2165$$
$$s^2 = 0{,}61 m^{1{,}23}$$

On sait dès lors stabiliser la variance. En effet (Kendall, 1967) si l'on a une relation de la forme $\sigma^2 = g(\mu)$, la transformation à utiliser est

$$f(N) = \int^N \frac{dt}{\sqrt{g(t)}}$$

(la constante d'intégration étant arbitraire) ~~et choisie de manière à normaliser la distribution~~. Dans le cas qui nous intéresse la relation est de la forme

$$\sigma^2 = a\mu^b$$

d'où,

$$f(N) = \frac{(N+C)^{1-b/2}}{(N+C)^{0.385}} \quad N^{1-\frac{b}{2}} + C$$

C étant ~~à déterminer empiriquement~~ arbitraire.

Notre objectif est le calcul d'un ordre de grandeur de l'erreur. Remarquant que 0,385 est peu différent de 1/3, posons $C = 0$ et testons la transformation $N^{\frac{1}{3}}$. On constate qu'après transformation racine cubique sur les effectifs comptés le coefficient de corrélation entre moyennes et variances devient égal à $-0{,}3124$, valeur non significative c'est-à-dire pouvant n'être que fortuitement supérieure à celle correspondant à un calcul plus rigoureux.

A titre indicatif testons de la même façon l'effet des transformations rencontrées dans la littérature. La transformation logarithmique surcorrige considérablement, donnant un coefficient de corrélation égal à $-0{,}9163$. La transformation racine

carrée donne un coefficient de corrélation de $-0{,}4983$ et la transformation $\log^2$ intermédiaire entre les deux précédentes (Frontier, 1969) un coefficient de $-0{,}5421$. Ces deux dernières valeurs sont tout juste significatives au seuil 5 %.

Nous adopterons donc pour $m > 10$ la transformation racine cubique. Les 17 nouvelles variances sont:

0,07333	0.03512	0.02553
0,05740	0,06330	0,03779
0,04473	0,03520	0,03960
0,02686	0,08082	0,03388
0,06461	0,06979	0,03644
		0,03765
		0,04800

La variance globale, calculée avec $17 \times 30 = 510$ degrés de liberté, est égale à 0,04765. L'homogénéité de l'ensemble des 17 variances considérées comme estimations de la variance globale est vérifiée par le test de Bartlett (1937): si $s_1^2, s_2^2, \ldots, s_k^2$ sont k estimations de la variance obtenues avec $n_1, n_2, \ldots n_k$ degrés de liberté, s^2 l'estimation globale obtenue avec n degrés de liberté, et si ces estimations sont homogènes, la quantité

$$B = \frac{n \log s^2 - \sum_1^k n_i \log s_i^2}{0{,}4343 \left[1 + \frac{1}{3}(k-1) \cdot \left(\sum_1^k \frac{1}{n_i} - \frac{1}{n}\right)\right]}$$

(logarithmes de base 10) est approximativement distribuée comme un χ^2 à $(k-1)$ degrés de liberté. On a ici $k = 17$, $n = 510$, $n_i = 30$. On trouve $B = 29{,}83$, valeur significativement trop élevée au seuil 5 %, non au seuil 1 %. L'indépendance de la variance par rapport à la moyenne ne supprime donc pas une certaine variabilité, qui se trouve à la limite de la signification. On pourrait reconnaître ici soit l'hétérogénéité introduite en considérant en un seul ensemble des dispersions d'organismes très divers, soit l'erreur introduite en substituant la transformation $\sqrt[3]{N}$ à la transformation $(N+C)^{0{,}365}$ / $N^{0{,}365} + C$.

Nous ferons néanmoins l'hypothèse que les erreurs de comptage sont, après transformation racine cubique sur les données brutes, distribuées avec une variance constante voisine de 0,04765. Le test probit appliqué à cette distribution des erreurs conclut à une normalité très satisfaisante (Fig. 3b). La normalité de la distribution des écarts est assez bien réalisée après transformation racine carrée ou $\log^2$, en dépit d'une stabilisation insuffisante de la variance. Par contre la transformation logarithmique (Fig. 3a) donne une courbe sigmoïde. On trouve graphiquement, en construisant l'intervalle (15,87 %, 84,13 %) correspondant à $\pm\sigma$, un écart-type égal à 0,18 d'où une variance de 0,0324. En réalité cette estimation est entachée d'un biais

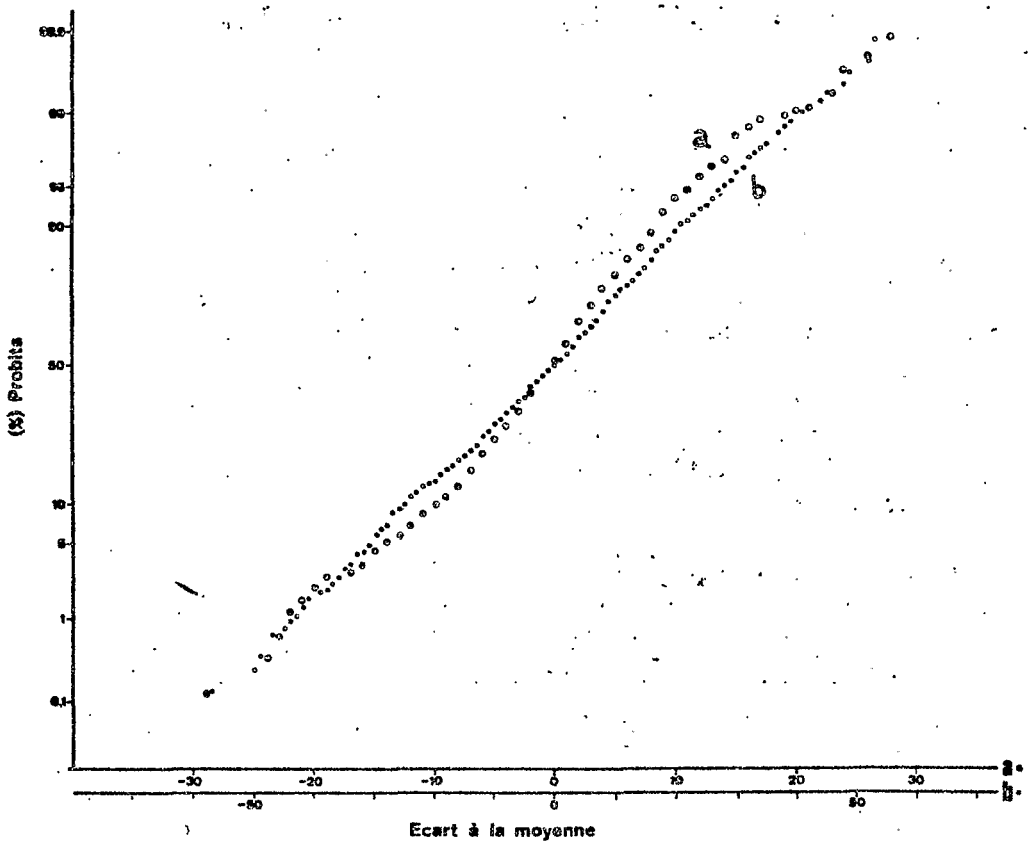

Fig. 3. Distribution des écarts après transformation logarithme (a), et racine cubique (b) sur les données test probit. Ecarts ×100

important: elle est en effet déterminée en utilisant la totalité des données comme si elles étaient indépendantes, c'est-à-dire à l'aide de 765 degrés de liberté au lieu de 510. On obtiendra une estimation correcte en multipliant par 3/2 l'estimation graphique, ce qui donne une variance de 0,0486 voisine de celle calculée ci-dessus: 0,04765. Nous retiendrons l'estimation déterminée graphiquement, obtenue à l'aide de la partie centrale $(m \pm \sigma)$ de la courbe «probit» qui s'ajuste très finement à un segment de droite. Nous poserons en conclusion que les écarts sont, après transformation racine cubique effectuée sur les données brutes, distribués normalement autour de zéro avec une variance égale à 0,0486. Cette règle va nous permettre de calculer l'erreur statistique.

Calcul de l'Erreur Statistique

Pour les moyennes inférieures à 10 nous avons admis que les distributions sont de Poisson, donc normalisables par la transformation $\sqrt{(N+\frac{3}{8})}$, la nouvelle variance

étant stable et égale à 0,25. L'intervalle de confiance au niveau 95 % sera donc, pour les variables transformées,

$$(\sqrt{N+\tfrac{3}{8}}-1,\ \sqrt{N+\tfrac{3}{8}}+1).$$

L'introduction de transformations non linéaires destinées à normaliser les distributions et stabiliser les variances introduit la question du biais: la moyenne et l'écart-type de la transformée ne sont pas les transformés de la moyenne et de l'écart-type. Cette question se rattache au problème de la définition d'une «valeur la plus probable» et d'un «intervalle de confiance». Si ces quantités sont données comme des fonctions des paramètres de la distribution, elles sont biaisées dans une transformation non linéaire, et il peut paraître paradoxal que la valeur la plus probable et les bornes de l'intervalle de confiance dépendent de la façon de mesurer l'objet, (différent si l'on mesure une grandeur ou sa racine cubique). On contournera cette difficulté en admettant que ces quantités classiques ont trait exclusivement à des distributions normales, et qu'en cas de distribution non normale elles sont définies par référence aux quantités correspondantes de la distribution normalisée par transformation adéquate. Cela revient à les définir en tant que quantiles (5 %, 50 %, 95 %) puisque ceux-ci restent stables dans une transformation continue et monotone. L'intervalle de confiance dans les nombres originaux aura donc par définition pour bornes les transformées inverses des bornes de l'intervalle de confiance dans la distribution normalisée. Ainsi l'intervalle de confiance au niveau 95 % pour la loi de Poisson sera de la forme,

$$(\sqrt{N+\tfrac{3}{8}}-1)^2-\tfrac{3}{8},\ (\sqrt{N+\tfrac{3}{8}}+1)^2-\tfrac{3}{8}$$

Le Tableau I indique cet intervalle pour des moyennes comprises entre 4 et 10.

TABLEAU I

m	Intervalles de confiance au seuil 95 %
4	0,8– 9,2
5	1,4–10,7
6	1,9–12,1
7	2,6–13,5
8	3,2–14,8
9	3,8–16,0
10	4,6–17,4

Pour les moyennes supérieures à 10 la transformation racine cubique normalise, la nouvelle variance étant égale à 0,0486, d'où une erreur au niveau de confiance 95 % égale à 0,44. L'intervalle de confiance sur les effectifs non transformés sera de la forme

$$(\sqrt[3]{N}-0{,}44)^3,\ (\sqrt[3]{N}+0{,}44)^3$$

Le Tableau II donne cet intervalle pour les valeurs comprises entre 10 et 350. On peut

TABLEAU II

m	Intervalle de confiance au niveau 95 %	Erreur relative ×100 Inférieure	Supérieure
10	5– 17	50	74
20	12– 31	42	57
30	19– 45	37	49
40	27– 58	34	44
50	34– 70	32	40
60	42– 82	30	37
70	50– 95	29	35
80	58–107	28	34
90	66–119	27	32
100	74–131	26	31
110	82–143	25	30
120	91–155	25	29
130	99–167	24	28
140	107–179	23	28
150	116–190	23	27
160	124–202	22	26
170	133–214	22	26
180	141–226	22	26
190	150–237	21	25
200	158–249	21	25
210	166–260	21	24
220	175–272	20	24
230	184–284	20	23
240	192–294	20	23
250	201–305	20	22
260	210–317	19	22
270	218–329	19	22
280	227–340	19	21
290	236–352	19	21
300	244–363	19	21
310	254–375	18	21
320	262–386	18	21
330	271–397	18	20
340	280–409	18	20
350	289–420	17	20

également construire un abaque donnant directement l'intervalle de confiance pour toute valeur de N en portant en abscisses les valeurs de N et en ordonnées les valeurs des bornes supérieures et inférieures des intervalles correspondants (Fig. 4 et abaque hors-texte). On obtient une courbe d'allure parabolique, symétrique par rapport à la première diagonale, ce qui a une importance pour la suite du raisonnement.

Cette symétrie est une propriété indépendante de la transformation. En effet, soit $f(N)$ la transformation normalisante, et appelons $f(N)^{\pm}\alpha$ l'intervalle de confiance dans les nouvelles variables. Un point A, borne inférieure d'un intervalle de confiance dans les données non transformées a pour coordonnées sur l'abaque: $x = N$,

$y = f^{-1}(f(N) - a)$. Le point A' symétrique de A par rapport à la première diagonale a comme coordonnées: $x' = f^{-1}(f(N) - a)$, $y' = N$. La borne supérieure de l'intervalle de confiance dans les nombres transformés, correspondant à l'effectif $N' = f^{-1}(f(N) - a)$ est $f(N') + a = f \circ f^{-1}(f(N) - a) + a = f(N)$. En revenant aux nombres non transformés: $f^{-1} \circ f(N) = N$: la borne supérieure est donc A'. On démontrerait de même que tout point de la branche supérieure de la courbe a pour symétrique par rapport à la première diagonale un point de la branche inférieure.

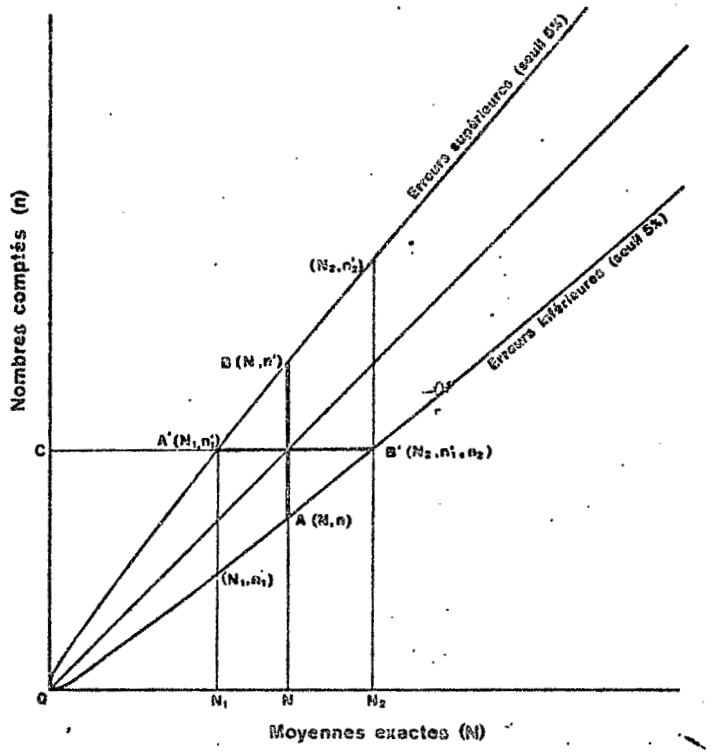

Fig. 4. Courbe fournissant les valeurs des bornes supérieure et inférieure de l'intervalle de confiance au niveau 95 %, en fonction de l'effectif. AB: intervalle de confiance pour une moyenne exacte N; A'B': intervalle d'estimation à partir d'un comptage C (= N).

Pour une valeur exacte N l'effectif compté C sera compris dans 95 % des cas entre deux valeurs n et n', ordonnées des points A et B (Fig. 4). Or nous sommes dans la situation inverse, à savoir que nous estimons la valeur vraie à partir d'un comptage C. Dans 95 % des cas la valeur vraie se trouvera dans un intervalle (N_1, N_2) défini de la façon suivante: C est égal à la borne supérieure n'_1 de l'intervalle de confiance de N_1, et à la borne inférieure n_2 de l'intervalle de confiance de N_2. La droite d'ordonnée $C = n'_1 = n_2$ rencontre la courbe en A' et B' et, en raison de la symétrie démontrée ci-dessus, AB et A'B' sont symétriques par rapport à la première bissectrice. On peut donc indifféremment considérer AB comme intervalle de confiance d'une moyenne

vraie égale à N, ou comme intervalle d'estimation d'un comptage égal à N.

L'abaque (p. 132) permet en outre de déterminer simplement l'erreur statistique relative. Il suffit de remarquer que, pour raison d'homothétie, l'erreur relative est la même tout le long d'une droite passant par l'origine. A toute valeur K de l'erreur relative correspond une droite D_K (de pente $1+K$). L'intersection de D_K avec la courbe fournit la valeur de N pour laquelle l'erreur relative au seuil 95 % est égale à K. Quelques valeurs de l'erreur relative ont été portées le long de la courbe de l'abaque hors-texte.

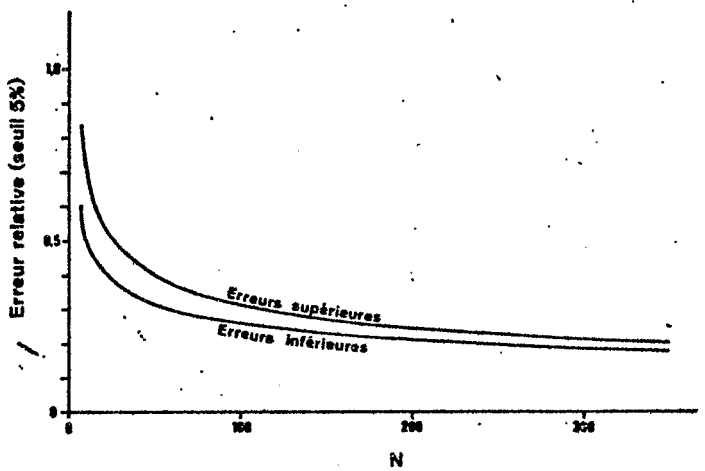

Fig. 5. Variations de l'erreur relative au niveau de confiance 95 % en fonction du nombre compté.

La Fig. 5 représente les variations en fonction de N de l'erreur relative au niveau 95 %. On remarque que l'erreur supérieure est plus grande que l'erreur inférieure. La courbe supérieure indique l'effectif qu'il convient de compter réellement pour obtenir une erreur relative au plus égale à $|K|$. On constate que si pour les petits nombres l'erreur relative diminue très vite quand N augmente, la décroissance devient ensuite très lente: le comptage de 50 organismes assure une erreur d'au plus 40 %, celui de 100 organismes une erreur d'au plus 31 %, celui de 200 organismes une erreur d'au plus 25 %, et il faudrait – à supposer que la loi puisse être extrapolée – compter environ 2500 individus pour être assuré d'une erreur inférieure ou égale à 10 %.

Etant donnée l'hétérogénéité spatiale des peuplements naturels, il semble qu'une précision de 31 % suffise largement à caractériser un échantillonage. Comme cette précision n'est que faiblement améliorée par le doublement de l'effectif compté, il semble judicieux de conclure qu'une centaine d'individus, pour chaque catégorie d'organismes étudiée, doit être effectivement dénombrée, dans une partie aliquote connue. Ce nombre est ensuite ramené à la récolte totale et l'intervalle d'estimation sera, en valeur relative: (−26 %, +31 %).

Bibliographie

Anscombe, F. J., 1948. The transformation of Poisson, binomial and negative binomial data. *Biometrika*, Vol. 35, pp. 246–254.

BARTLETT, M. S., 1937. Properties of sufficiency and statistical tests. *Proc. Roy. Soc., Ser. A*, Vol. 160, pp. 268.

BARTLETT, M. S., 1947. The use of transformations. *Biometrics*, Vol. 3, pp. 39–52.

FRONTIER, S., 1969. Sur une méthode d'analyse faunistique rapide du zooplancton. *J. exp. mar. Biol. Ecol.*, Vol. 3, pp. 18–26.

KENDALL, M. G., 1968. *Advanced theory of statistics*. Hafner, London.