Estimation du paramétre de distribution de la distribution binomiale négative: A priori, effort d'échantillonnage, et information

Lise Vaudor

ESTIMATION DU PARAMÈTRE DE DISPERSION DE LA DISTRIBUTION BINOMIALE NEGATIVE: A PRIORI, EFFORT D'ÉCHANTILLONNAGE, ET INFORMATION.

Lise Vaudor
CEMAGREF MALY, 3 bis quai Chauveau, CP 220, F-69336 Lyon Cedex 09, France.

Mots-clé: biais, binomiale négative, entropie, estimation, échantillonnage

Résumé

La distribution binomiale négative est fréquemment utilisée pour modéliser la distribution des données d'abondance surdispersées. De ce fait, l'ajustement de la binomiale négative aux données d'abondance et l'estimation de ses paramètres est un enjeu majeur pour de très nombreuses études en écologie, épidémiologie, actuariat, etc. qui reposent sur des données de ce type. L'estimation des paramètres de la binomiale négative, et en particulier de son paramètre de dispersion est néanmoins problématique, car les estimateurs disponibles sont biaisés et peu efficaces, notamment lorsque les échantillons sont petits, que l'espérance de l'abondance est faible, et que sa variance est forte. Dans cette étude, nous tentons d'identifier la source de ces problèmes d'estimation. Pour ce faire, nous utilisons l'entropie (Shannon, 1949) pour quantifier l'information qu'apportent les échantillons sur la dispersion, en fonction des caractéristiques de ces échantillons (taille, moyenne), et des valeurs réelles des paramètres de la binomiale négative. On montre ainsi que le nombre total d'individus observés est un facteur clé de la qualité de l'estimation, et que pour un effort d'échantillonnage donné l'estimation de la dispersion est vraisemblablement très peu fiable si la dispersion réelle se situe dans des gammes de valeurs extrêmes (notamment des valeurs fortes).

Abstract

The negative binomial distribiution is frequently used to model overdispersed count data. Adjusting the negative binomial and estimating its parameters is thus a major concern in ecology, epidemiology, actuaries, etc. that often deal with that type of data. However, the estimation of the parameters of the negative binomial, and the estimation of its dispersion parameter in particular, are often a problematic issue because available estimators are biased and have low efficiency. In particular, if samples are small, if the expected value of the distribution is low, or if its variance is high, the estimators are particularly unreliable. In this study, we try to identify the reasons behind these estimation problems. To do this, we use entropy (Shannon, 1949) to quantify the information about dispersion that the samples provide. Entropy depends on the characteristics of the samples (size, mean) and of the real values of the parameters of the negative binomial. We show that the total number of individuals observed in a sample is a key driver of estimation quality. We also highlight the fact that for a given sampling effort the estimation of dispersion is likely to be really unreliable if the real value of dispersion lies in extremes ranges of values (notably, high values).

Le modèle de distribution le plus classiquement utilisé pour des données d'abondance surdispersées (soit une part majeure des données étudiées en écologie, épidémiologie, actuariat, etc.) est la loi binomiale négative (BN). La BN peut être paramétrisée de la façon suivante, avec μ comme paramètre de moyenne et α comme paramètre de dispersion :

$$\forall y \in \mathrm{N} \quad P(X=y) = \left(\frac{\Gamma(y+\alpha^{-1})}{y!\,\Gamma(\alpha^{-1})} \right) (\mu\alpha)^{y} (1+\mu\alpha)^{-(y+\alpha^{-1})} \qquad \mu \in R^{+*},\ \alpha \in R^{+*}$$

L'étude de données d'abondance, si elle repose sur une hypothèse de distribution BN, requiert un ajustement du modèle de distribution et donc une estimation de ses paramètres. La qualité de l'estimation à la fois de la moyenne et de la dispersion est donc un point essentiel de nombreuses études.

L'estimation du paramètre de dispersion de la BN est pourtant reconnu comme problématique. En effet les estimateurs de ce paramètre (par exemple l'estimateur par la méthode des moments ou par maximum de vraisemblance) sont biaisés : l'estimation est en moyenne fortement différente de la valeur réelle. Lloyd-Smith (2007) illustre l'importance du biais selon les gammes de valeurs de paramètres de la BN, et selon la taille d'échantillon. Il montre ainsi que le biais est d'autant plus important que le paramètre de moyenne et la taille d'échantillon sont faibles, et qu'il correspond à une sous-estimation de la dispersion. Ce phénomène a été étudié par divers auteurs notamment dans le but de comparer les différents estimateurs existants et d'en proposer de nouveaux (voir par exemple Clark et Perry 1989 ou Saha et Paul 2005). Les estimateurs proposés demeurent cependant biaisés (Wang 1996) et relativement peu efficaces, de sorte que dans certaines conditions (i.e. pour certaines valeurs des paramètres et certaines tailles d'échantillon) les estimateurs classiques demeurent plus satisfaisants.

Nous nous attachons quant à nous à expliquer le phénomène de biais d'estimation, en fournissant pour chaque échantillon observable une mesure de l'information qu'il recèle quant à la dispersion, et en caractérisant les conditions engendrant des observations peu informatives. Nous montrons que l'effectif (somme des abondances) des échantillons est le facteur clé du biais d'estimation de la dispersion : le biais est d'autant plus fort que la probabilité d'observer des échantillons de faible abondance totale est importante. C'est le cas notamment pour une moyenne de distribution ou une taille d'échantillon faibles, ou pour un paramètre de dispersion fort. Par exemple, si l'effectif pour un échantillon est de deux individus seulement, seules deux dispersions possibles sont observables : soit les deux individus se trouvent dans le même point d'échantillonnage, soit ils se trouvent dans deux points différents. Dans ce cas, l'échantillon est en fait très peu informatif quand à la dispersion réelle des données. De manière plus générale, le nombre de dispersions observables est d'autant plus faible que l'effectif est faible.

Nous utilisons l'entropie (Shannon, 1948) pour mesurer l'information fournie par l'estimateur du paramètre de la dispersion pour un échantillon de taille N (N=20,50), d'effectif T (T≤40) et issu d'une BN avec pour valeur de paramètre de dispersion α.
L'entropie H associée à l'estimateur A du paramètre α conditionnellement à l'effectif T, et aux valeurs de paramètres μ et α, est d'univers $\Omega = \{\alpha_1^*, ..., \alpha_n^*\}$ fini et correspond à :

$$\begin{aligned} H(A/T,\mu,\alpha) &= -\sum_{i=1}^{n} p(\alpha_i^*/T,\mu,\alpha) \log p(\alpha_i^*/T,\mu,\alpha) \\ &= -\sum_{i=1}^{n} p(\alpha_i^*/T,\alpha) \log p(\alpha_i^*/T,\alpha) \\ &= H(A/T,\alpha) \end{aligned}$$

Pour calculer $H(A/T,\mu,\alpha)$, nous listons toutes les répartitions possibles de T individus dans N points, calculons les estimations par maximum de vraisemblance $\alpha_1^*, ..., \alpha_n^*$ correspondant à l'ensemble des n échantillons ainsi listés, et calculons la probabilité $p(\alpha_i^*/T,\mu,\alpha)$.

La distribution des dispersions observables (μ*,α*) dans l'espace des paramètres, ainsi que le profil d'entropie pour chaque effectif en fonction du paramètre α, illustrent le fait que pour un effort d'échantillonnage donné et pour (μ,α) dans certaines gammes de valeurs (notamment μ faible et α fort), l'estimation de la dispersion est vraisemblablement très peu fiable. L'effort et la méthode d'échantillonnage doivent être ajustées en conséquence. En particulier, échantillonner jusqu'à avoir un effectif supérieur à un certain seuil (échantillonnage inverse) semble souhaitable.

Bibliographie

[1] Clark, S. J. and Perry, J. N. (1989) Estimation of the negative binomial parameter κ by maximum quasi-likelihood. *Biometrics,* 45, 309-316.

[2] Lloyd-Smith, J. O. (2007) Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. *PLoS ONE*, 2, e180.

[3] Saha, K. and Paul, S. (2005) Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics* 61, 179-185.

[4] Shannon, C. E. (1948) A mathematical theory of communication. *Bell Syst. Tech.* J. 27, 379-423.

[5] Wang, Y. N. (1996) Estimation problems for the two-parameter negative binomial distribution. *Statistics & Probability Letters 26*, 113-114.